IMPUTATION OF ITEM MISSING DATA IN THE HEALTH AND RETIREMENT SURVEY

Steven G. Heeringa, Survey Research Center
University of Michigan, Ann Arbor, Michigan 48106-1248

The Health and Retirement Survey (HRS) is a National Institute of Aging (NIA) sponsored study of retirement planning and decision-making in the U.S. household population. The objective of this paper is to describe the item missing data problems that affect the first wave of this major new survey program and to discuss alternative methods for the imputation of missing values in the general public use data set that is scheduled for release in the spring of 1994. The paper will briefly scan the range of HRS item missing data and imputation problems but will focus most directly on the difficult problem of item missing data for household asset and liability amounts.

1. Introduction

1.A. Survey Population

The HRS is a study of U.S. households that include one or more persons born during the period 1931 to 1941. Households in Alaska and Hawaii, housing units on military bases and group quarters units are excluded from the survey population. The HRS sampling unit is the household. The unit of observation is a household "financial unit" -- either a married couple household where one or both spouses fall in the eligible age range or a single unmarried adult who is 51 to 61 years old. In households where there is more than one such financial unit, one is chosen at random for interview. The interview with each household financial unit consists of at least two parts -- a household interview designed to collect household level data and a person interview. In married couple households, both spouses[1] complete the person interview and the household interview is administered to the individual most knowledgeable about household financial matters.

1.B. HRS Data Base Conventions for Imputed Variables

Based on the general imputation plan outlined in this paper, imputed values for item missing data will be included in the HRS Wave 1 public use data set. Researchers who request the public use data file will be informed of exactly which items in the data set are imputed values. A multiply imputed version of the HRS Wave 1 public use data set will also be available. In each data set imputed items will be identified to researchers by indicator flag variables, one indicator per survey variable.

1.C. Outline of the Paper

Section 2 provides a review of imputation developments and objectives as they apply to the HRS and similar public use data sets for major survey programs. A general discussion of HRS item missing data problems and corresponding imputation approaches is contained in Section 3. Section 4 opens the discussion of survey responses to HRS income and asset amounts, focusing on special HRS questionnaire design and interviewing conventions that provide added information for imputation and estimation when actual value amounts are not reported. Patterns and rates of missing data for HRS net worth components (assets and liabilities) are described in detail in Section 5. Section 6 of the paper discusses models for the imputation of missing data for HRS net worth components. A Generalized Iterative Bayesian Simulation (GIBS) technique for multiple imputation of HRS net worth components is presented in Section 7. Section 8 describes statistical and practical problems that restrict the general applicability of these methods to the HRS missing data problem. The paper concludes in Section 9 with a summary.

2. Review of Imputation Development and Objectives

2.A. Development of Imputation Methods and Theory

Item missing data has long been recognized as a problem for survey data analysts. Prior to the late 1970s or early 1980s, the accepted procedure among social science data analysts was to explicitly record values as missing but to take no corrective steps in analysis. Formal recognition of imputation as a statistical technique for dealing with item missing data may have originated with the establishment of the National Research Council's (NRC) Panel on Incomplete Data. Many of the earliest papers on imputation concepts and theory appear in the three-volume publication of the proceedings of the symposium (Madow *et al.*, 1983) which was organized by the NRC Panel. Through the late 1970s and early 1980s, survey statisticians continued to conduct research and to publish on imputation methods (Rubin, 1980; Sande, 1981; Kalton, 1983). Intuitive and ad hoc procedures were recast and labeled as specific methods, each based on a statistical model (mean value imputation, Hot Deck and Cold Deck procedures, regression-based imputation, multiple imputation). The introduction of imputation methods to survey practice was a slow process and by no means a universal one. During the 1980s, major federal survey data programs in the U.S. and Canada took the lead in the development and application of basic imputation methods such as the Hot Deck. In the United States, developments in imputation methods were promoted by survey programs such as The Survey of Income and Program Participation (SIPP), programs which require the collection of many variables that are subject to significant amounts of item nonresponse. Imputation theory and application were greatly extended by the introduction of the multiple imputation technique (Rubin, 1987). By decade's end, large-scale, general purpose imputation of item missing values in major survey data sets had become a common and accepted practice.

Coincident with the development of formal theory for imputation methods, there have been major developments in the theory and practical tools for making inference from multivariate data sets where missing data follow structured patterns (Rubin, 1974; Little and Rubin, 1987). Introduction of the EM (expectation, maximization) algorithm (Dempster, Laird and Rubin, 1977) provides analysts with a procedure to iteratively derive maximum likelihood estimates (MLEs) when the multivariate pattern of missing data is a more arbitrary one. The most recent advances in theory and application are in the area of generalized iterative Bayesian simulation (GIBS) methods (Meng and Rubin, 1992).

2.B. Why Impute?

The HRS is a complex survey with numerous variables collected across many areas of inquiry. Imputation of item missing data on the scale encountered in the HRS and other major survey programs does not permit consideration of models for each individual variable. Variables will be grouped and a common model may be used to impute missing values for each variable in the class. The imputation procedures will require that the data are missing at random (MAR).

Before moving ahead with wholesale imputation of item missing data in the HRS public use data set, it is important to answer the question: why impute? Why not leave the handling of missing data to the data analyst? The HRS is a national data resource and will be used extensively by researchers with widely varying substantive or policy interests and equally wide-ranging levels of statistical sophistication. For analysts with advanced statistical training, the proposed public use data file with its companion set of imputation indicator variables permits the data to be analyzed in its original form. However, most HRS data analysts will not have the training, tools and time to create their own imputations or to employ sophisticated estimation techniques. With these researchers in mind, the responsible choice is to make available a fully imputed data set.

3. HRS Missing Data Problems

Selection of an imputation procedure starts with the choice of a probability model that describes the distribution of the variable of interest (missing or not) and its relationship to a vector of known covariates. Under the assumption that responses to items are MAR, the chosen model is not used to simply predict the "correct" value for the missing response but also to preserve the stochastic properties of the completed (imputed) data vector -- the true sampling variance and covariance properties for the chosen model. The choice of the imputation procedure (and the underlying model) depends on the properties of the variable of interest. To discuss the general approach to imputation of the HRS Wave 1 data set we may consider three classes of variables/imputation problems:

3.A. Imputation is Not Needed or Needed Only for Some Forms of Item Missing Data

It is important to recognize that some HRS item missing data may best be left as is. Obvious examples are speculative questions or questions of knowledge where a "Don't Know" response indicates a state of thought or knowledge and not simply a failure to comply with the request for factual information. One example of this class of variable can be found in the response to survey item K15, "When you [and your (husband/wife/partner)] do retire, are you likely to move to a different location, stay where you are, or what?"

Many researchers would prefer to treat the "Don't Know" replies not as missing data but as a response state that indicates uncertainty about future plans, etc.

3.B. Categorical Response Item Imputation

HRS Wave 1 categorical items requiring imputation take several forms. The simplest form is the dichotomous (yes/no) response:

Item B29 "*Are you often troubled with pain?*"

A somewhat more complicated form is a variable with three or more nominal categories:

Item B29(f) "*Is your back pain due to a slipped disk, is it due to arthritis, or is it due to some other condition?*"

Finally, there may be a true ordinal relationship among the response categories:

Item B29(g) "*When the pain is at its worst, is it mild, moderate or severe?*"

Review of HRS Wave 1 frequency runs suggests that item missing data rates on most categorical response questions are less than 1%. However, there are categorical items in the health section (e.g., source of back pain) and elsewhere, for which item missing data is in the 2%-10% range. There are several methods that we can consider for imputation of categorical responses. For many categorical variables, item missing data rates will be small and available data models for the response distribution will be weak. The practical decision in the majority of such cases will be to use simple Hot Deck methods or some form of stochastic imputation based on the observed distribution of sample responses. However, some key categorical variables will warrant a more careful investigation of log-linear/logit linear models of response values. For example, a logit linear model may be used to estimate the expected value of the (0,1) response as a function of observed covariates for each case. For missing data cases, the predicted value from the logit model is used to randomly determine the assignment of a "0" or a "1" code to the case. Similar randomized imputation based on predicted response probabilities from log-linear or cumulative logit models can be used to make imputations to multi-category nominal and ordinal categorical variables.

3.C. Continuous Response Items

The highest rates of HRS Wave 1 item missing data are found in the continuous variable items, particularly those involving income, asset and liability amounts. By and large, the majority of HRS Wave 1 imputations for continuous response items will be performed using models for the conditional distribution of the survey variable. In general, regression models are used to express the conditional relationship of a single continuous variable of interest to a set of categorical and continuous covariates. Simultaneous imputation of multiple variables using multivariate probability models is one way to ensure that distributional consistency is maintained. [See Section 6.]

4. "Bracketed" and Range Card Responses

Due to sensitivity/privacy concerns or poor respondent knowledge/recall, survey questions that request amounts -- income, assets, liabilities, transfers -- are expected to have a relatively high rate of item missing data. The HRS took several steps to address this problem at both the questionnaire design stage and during the interview process itself.

4.A. Bracketing of Amounts

For key asset items, if a respondent could not recall or refused to report the exact value for the item, the HRS Wave 1 questionnaire followed up with a short sequence of questions designed to "bracket" the true response value. The question sequences open by asking the household respondent if the household owns the asset in question (e.g., a business). If the asset is owned, its exact value is requested. If the exact value is not reported, the questionnaire routes the respondent through a series of dichotomous response questions which attempt to "bracket" the value of the asset. Taking the business asset and IRA/KEOGH account value question sequences as examples, the finest level of bracketing attainable through the questions is shown in Table 1 below.

Routing the respondent through the nested series of bracketing questions does not guarantee that a specific bracket will be identified for the unreported amount. In some cases, no additional information will be obtained. In other cases, the responses will indicate that the true value lies in one of three brackets, but not precisely which of the three brackets. By example, a respondent may indicate that the value of their IRA or Keogh account is > = $25,000 but cannot/will not indicate if it is $25,000-$49,999, $50,000-$99,999, or $100,000+.

Table 1
Examples of Response Bracket Ranges for HRS Asset Items[2]

Bracket	Business Value Response	IRA, KEOGH Response
1	$1 - $9,999	$1-4,999
2	$10,000 - $49,999	$5,000 - $24,999
3	$50,000 - $499,999	$25,000 - $49,999
4	$500,000 +	$50,000 - $99,999
5	Inapplicable	$100,000 +

4.B. Range Card

Question sequences designed to bracket the unreported value for a household asset or income item require valuable interview time to administer, and their use was reserved only for income and asset items which experience had shown to have the most serious item missing data problems. For all other income and asset items, the questionnaire was designed to ask only for the actual amount with no follow-up series of questions. If the respondent did not know or refused to provide the actual amount, the interviewer presented them with a range card, as illustrated in Figure 1. A single range card provided the respondent with a choice of ten dollar amount categories.

Figure 1
HRS Range Card Categories

A.	LESS THAN $1,000
B.	$500 - $1,000
C.	$1,001 - 2,500
D.	$2,501 - $10,000
E.	$10,001 - $50,000
F.	$50,001 - $250,000
G.	$250,001 - $999,999
H.	$1 MILLION - $9,999,999
I.	$10 MILLION - $100 MILLION
J.	MORE THAN $100 MILLION

5. Rates of Missing Data for Components of Household Net Worth

The highest rates of item missing data occur in the measurement of assets and liabilities which comprise a household's net worth. At the same time, reliable measurement of household net worth or "wealth" is critical for HRS researchers whose goal is to study the relationship of health and financial well-being to retirement planning and post-retirement needs. The following three sections focus on the HRS measures of net worth components -- rates and patterns of missing data and potential approaches to the imputation of missing amounts.

5.A. Patterns of Response to Net Worth Component Items

The HRS Wave 1 interview measures household net worth as a series of fifteen asset and six liability question items. Table 2 summarizes the missing data problem for each of the asset and liability components of the composite HRS net worth variable. The left-hand panel of Table 2 identifies the individual asset (A) and liability (L) components which may figure in the computation of a household's net worth. The central panel, labeled "Does item apply?", provides estimates of the percentage of HRS sample households (unweighted) that reported having each asset or liability. For example, of the n=7078[3] respondent households included in this summary, 23.2% report owning real estate other than their personal residence. For households that report owning a particular asset or having a particular type of debt, the right-hand panel of Table 2 describes the distribution of response types: actual value, bracketed value[4], range card value, or missing data value. Bracketing question sequences were provided only for the first nine asset items. The remaining asset questions and all liability questions relied solely on the range card as means of obtaining bounds when the true value was not reported.

Among financial assets, the percentage of actual value reports ranges from 67.7% for stocks and mutual funds to 87.9% for combined value of vehicles and other personal property. Depending on the asset, the percentage of bracketed responses ranges from 8% for property to 21.3% for stocks and mutual funds. Even though a bracketing question sequence was provided for these asset items, from 2.2% to 6.0% of bounded response values were recorded as choices from the range card. The rates of completely missing data -- proportions of cases where no real information on bounding values is available -- range from 1.9% of responses for the vehicle and property question to 10.2% for value of bonds.

The second tier of asset items may be labeled "property assets" and includes the values of homes, farms, and mobile homes, the most prevalent assets being personal homes and second homes. Actual values for personal homes were reported by 95.6% of HRS sample home-owners. The remaining 4.4% of responses to the home value question are divided between range card (bounded) values (1.0%) and completely missing data (3.4%). Completely missing data rates for liability components of household net worth range from 3.4% for amount owed on a home equity credit line to 11.6% for amount due on a home equity credit loan. The completely missing data rate for the amount of homeowners' first mortgage is 6.8%.

5.B. Patterns of Bracketing

As noted in the previous section, only bracketing or bounding value information is available for from 8% to 21% of responses to key net worth component items. From Table 2, we assume that 26.9% of HRS Wave 1 respondent households report stock or mutual fund ownership. Of these cases, 67.7% report actual amounts of stock and mutual fund values; 5.8% report no information on the value of these assets; and 26.5% report some bracketing or range card information. Table 3 describes in more detail the distribution of the bracketed cases. A total of 81 cases or 4.4% of Table 2 households owning these assets report that the value of stocks and mutual funds lies in the range $1 - $4999. At the other end of the distribution, 39 cases report a bracketed amount of $100,000 - $499,999 and 8 cases report that the value of stock is > =$500,000. Sometimes respondents provided some but not all of the bracketing information that the HRS interview requested. A total of 12 respondents (code 35) reported that the value of stock was >$25,000 but provided no information concerning possible upper boundary values. As a second example, the right-hand panel of Table 3 provides the distribution for responses to questions concerning the value of checking and savings account balances.

5.C. Loss of Information for Individual Households

The HRS net worth variable and its major components -- total assets, total debt, financial assets, etc. -- will be constructed by aggregating the component variables listed in Table 2. The data on response distributions given in Tables 2 and 3 provide a univariate and cross-sectional view of the missing data problem. This is certainly a valuable perspective, but it is also important to recognize that the analytic purpose of these data is predominantly multivariate. Therefore, another way to look at the missing data problem is across items within responding household units.

Total net worth for each household is the sum of up to 21 asset and liability (negatively signed) components. The column margin of Table 4 shows the distribution of HRS households according to the actual number of components that figure into the households' net worth computations. Recognizing that all assets and liabilities are not equally valued,[5] the rows of Table 4 identify how many of the required components have actual reported values. For example, a total of n=925 HRS Wave 1 households require exactly four asset and liability components for the households' net worth computations. Of these, 58.8% reported actual values for all four items, 22.7% reported 3 of 4, 11.7% reported 2, 4.9% reported one, and 2.0% did not provide actual values for any of the four required items. Table 5 repeats the presentation of Table 4 with the simple difference that the row margin counts both actual value and bracketed value responses. Of the n=925 households that require four components for the net worth calculation, 84.5% now provide either actual or bracketed amount values for each item -- only 0.1% provide no information at all.

6. Imputation Models for HRS Net Worth Components

Ideally, the choice of the imputation model/method for the HRS net worth component variables will meet five criteria or objectives: a) timeliness; b) efficient use of all available data; 3) multivariate consistency; 4) preservation of stochastic properties in the complete data set including covariances, conditional and marginal distributions under the model; and 5) sample or "data-based" estimation of variance properties of completed data estimates.

6.A. Review of the Data Structure for the Problem

Regardless of the chosen model and imputation or estimation method, handling of missing data for HRS net worth components is a challenging problem. As noted above, the designers of the HRS questionnaire recognized the problem and took steps to obtain information that could be used to address the missing data problem. In the HRS, measurement of a net worth component can be decomposed into three nested data elements:

1) Screening/indicator variables that tell us that the household does/does not have the asset or liability in question;

2) Variables or item responses that allow us to determine boundary values for non-zero amounts; and finally

3) Actual amounts for assets and liabilities.

6.B. A Simple Model

Taking a univariate perspective on the problem, there is a very simple, three-step hierarchical approach to imputation and the creation of a "completed" data set:

1) Determine from actual data or impute whether or not the household has the asset or liability in question;

2) Depending on the outcome of step 1, use actual values or imputation to determine the bracket or boundary for all non-zero asset and liability amounts; and

3) Conditional on observed data for bounded intervals (step 2), impute the actual amount for missing data cases.

For simplicity, imputations required in steps 1 and 2 could be performed randomly within respondent classes, by Hot Deck methods or by similar ANOVA type models. If desired, multivariate models for categorical data (logistic regression, log-linear models) could be used for the step 1 and 2 imputations. Subject to specified boundary constraints, assignment of actual amounts to missing data cases in step 3 could be performed by mean value or random imputation, by Hot Deck methods, or by a regression model with random residual. This simple model certainly meets the timeliness and practicality criteria. The major deficiency of this simple model is that it is not built on a truly multivariate framework.

Table 2
HRS Wave 1 Net Worth Components
Distribution of Responses by Response Type (n = 7078 respondent households)

Asset or Liability Item	Does Item Apply?				If Item Applies To Household						
	Total	Yes	No	DK	Total		Actual Value	Bracketed Value	Range Card Value	Missing Value
					n	%				
A: Real Estate (not home)	100%	23.2%	76.8%	0.5%	1605	100%	75.0%	16.1%	5.7%	3.2%
A: Vehicles, Personal Property	100%	-	-	-	7078	100%	87.9%	8.0%	2.2%	1.9%
A: Business	100%	16.1%	83.9%	0.4%	1110	100%	68.3%	21.1%	4.6%	6.0%
A: IRA, KEOGH	100%	37.0%	63.0%	0.6%	2578	100%	74.3%	15.9%	4.8%	5.0%
A: Stock, Mutual Funds	100%	26.9%	73.1%	0.8%	1844	100%	67.7%	21.3%	5.2%	5.8%
A: Checking, Savings	100%	78.1%	21.9%	1.0%	5458	100%	73.8%	16.2%	4.8%	5.2%
A: CDs, Savings Bonds, T-Bills	100%	25.3%	74.7%	1.0%	1715	100%	71.2%	16.1%	6.0%	6.7%
A: Bonds	100%	6.8%	93.2%	1.0%	410	100%	70.5%	13.7%	5.6%	10.2%
A: Other Assets	100%	15.8%	84.2%	1.0%	1047	100%	74.7%	4.1%	4.6%	6.6%
A: Farm, Ranch - fully owned	100%	3.2%	96.8%	-	221	100%	88.7%	-	7.2%	4.1%
A: Farm, Ranch - partly owned	100%	0.3%	99.7%	-	22	100%	81.8%	-	13.6%	4.6%
A: Mobile Home - site only	100%	0.1%	99.9%		1	100%	100%	-	0.0%	0.0%
A: Mobile Home - home only	100%	2.5%	97.5%	-	175	100%	88.0%	-	5.1%	6.9%
A: Home, Apartment, Land	100%	67.6%	32.4%	-	4783	100%	95.5%	-	1.0%	3.4%
A: Home (2nd)	100%	12.7%	87.3%	-	902	100%	92.9%	-	1.1%	6.0%
L: 1st Mortgage	100%	43.1%	56.9%	-	3052	100%	92.5%	-	0.7%	6.8%
L: 2nd Mortgage	100%	4.9%	95.1%	-	346	100%	91.9%	-	0.9%	7.2%
L: Home Equity Loan	100%	2.3%	97.7%	-	164	100%	87.8%	-	0.6%	11.6%
L: Home Equity Credit line	100%	7.1%	92.9%	-	505	100%	95.8%	-	0.8%	3.4%
L: Owed on Second Home	100%	5.4%	94.6%	-	382	100%	92.7%	-	0.8%	6.5%
L: Any Other Debt	100%	38.0%	62.0%	-	2600	100%	90.8%	-	3.4%	5.8%

Table 3
HRS Distribution of Respondents by Type and Value of Asset

ASSET: Stock, Mutual Funds					
Code	Lower Bound	Upper Bound	Cases	% Total Cases	% Asset Cases
0	No Asset		5,175	73.11	
9	DK Asset		59	0.83	
1	Actual		1,249	17.65	67.73
11	1	4,999	81	1.14	4.39
22	5,000	24,999	120	1.70	6.51
33	25,000	99,999	108	1.53	5.86
44	100,000	499,999	39	0.55	2.11
55	500,000	OPEN	8	0.11	0.44
12	1	24,999	13	0.18	0.70
15	1	OPEN	106	1.50	5.75
35	25,000	OPEN	12	0.17	0.65
45	100,000	OPEN	12	0.17	0.65
6	RANGE CARD		96	1.36	5.21
TOTAL			7,078	100.00	100.00

ASSET: Checking, Savings					
Code	Lower Bound	Upper Bound	Cases	% Total Cases	% Asset Cases
0	No Asset		1,548	21.87	
9	DK Asset		72	1.02	
1	Actual		4,025	56.88	73.76
11	1	999	125	1.77	2.29
22	1,000	4,999	211	2.98	3.87
33	5,000	9,999	176	2.49	3.22
44	10,000	49,999	192	2.71	3.52
55	50,000	OPEN	84	1.19	1.54
12	1	4,999	21	0.30	0.38
15	1	OPEN	286	4.03	5.22
35	5,000	OPEN	51	0.72	0.93
45	10,000	OPEN	26	0.37	0.48
6	RANGE CARD		261	3.69	4.78
TOTAL			7,078	100.00	100.00

Table 4
HRS Wave 1 Net Worth Components
Total Net Worth Components for Household

TOTAL ACTUAL VALUE RESPONSES	1	2	3	4	5	6	7	8	9	10+
0	6.0 %	4.2%	2.8%	2.0%	1.2%	1.6%	0.6%	1.0%	2.2%	8.1%
1	94.0 %	19.8%	10.2%	4.9%	2.9%	2.4%	2.3%	2.1%	1.7%	1.5%
2		75.9%	22.0%	11.7%	7.7%	4.3%	4.5%	4.5%	3.2%	3.7%
3			65%	22.7%	12.0%	8.0%	5.4%	5.7%	4.5%	2.7%
4				58.8%	21.5%	10.0%	7.6%	5.9%	2.7%	2.8%
5					54.8%	17.7%	8.8%	6.2%	3.7%	4.0%
6						55.9%	16.1%	8.8%	6.5%	2.9%
7							54.7%	18.3%	8.2%	11.0 %
8								47.4%	14.9%	3.7%
9									52.2%	6.4%
10 +										53.2%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
n=	520	615	792	925	1127	996	795	578	402	328

Table 5
HRS Wave 1 Net Worth Components
Total Net Worth Components for Household

TOTAL ITEMS WITH ACTUAL OR BRACKETED VALUES	1	2	3	4	5	6	7	8	9	10+
0	1.7%	0.5%	0.1%	0.1%	0.0%	0.1%	0.0 %	0.0 %	0.0%	0.0%
1	98.3 %	8.3%	3.9%	1.1%	0.2 %	0.3%	0.1%	0.0%	0.3%	0.1%
2		91.2%	10.5%	3.0%	1.8%	1.0%	0.6%	0.7 %	0.3%	0.0%
3			85.5%	11.2 %	3.2%	2.2%	1.0%	0.5%	1.2%	2.0%
4				84.5%	10.7%	2.1%	1.4%	1.6%	0.3%	0.3%
5					84.1%	8.9%	2.5%	2.3%	0.5%	2.3%
6						85.3%	8.3%	2.6%	2.5%	0.4%
7							86.0%	10.7%	1.7%	0.7%
8								81.7%	11.0%	2.0%
9									82.3%	.5%
10+										88.2%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
n =	520	615	792	925	1127	996	795	578	402	328

Instead, the imputation of net worth components is reduced to a set of more or less independent processes. If multivariate models (e.g., regression) are used in each process, there is no assurance that imputed values will produce logical, let alone statistically consistent, outcomes. Logical consistency and some degree of statistical consistency can be imposed on the process by establishing a priority for individual variables and proceeding systematically through the ordered list, each time conditioning the imputation on the outcome for the previous step.

A truly multivariate framework for imputation is more than an independent or even sequenced collection of univariate regression models. The multivariate framework incorporates a model for the joint distribution of all variables -- both the dependent variables of primary interest and ancillary or covariate variables.[6] For example, to perform multivariate imputation of item missing data for the value of household stocks, bonds, mutual funds, IRAs and KEOGH retirement accounts, we might consider the joint distribution of these assets along with selected covariates such as household income, years of employment, salary and type of job held, etc.

6.C. General Location Model for the HRS Net Worth Component Data

Advances in multivariate imputation theory and program applications (Rubin and Schafer, 1990; Schafer,1992; Kennickell, 1992) lead us to consider recasting the HRS net worth imputation problem in a truly multivariate mold. Specifically, in this subsection we present a multivariate general location model (Olkin and Tate, 1961; Little and Schluchter, 1985; Little and Rubin, 1987) for the HRS net worth component data.[7] The following sections provide a simple theoretical introduction to the general location model and the adaptation of the HRS net worth component data to its multivariate framework for categorical and continuous data. Section 7 explores the possibility that the general location model may be used in combination with GIBS methodology to perform multiple imputation of missing data on HRS net worth components. The reader will notice that there are a number of problems -- both statistical and practical -- that restrict the application of the general location model and GIBS methods to the HRS item missing data problem. These problems along with possible solutions are discussed in Section 8 of this paper.

6.C.1. HRS Data Structure for the General Location Model

As described in Section 6.B, the data for the problem consist of a mixture of categorical and continuous variables. For each asset or liability component, the categorical variables include a dichotomous indicator of asset/liability holding and a constructed bracket or boundary value variable for non-zero value amounts. Respondent reports of actual dollar value amounts for assets and liabilities comprise the continuous variable set. To place the missing data problem for HRS net worth components into a simpler structure, we can create for each asset and liability component a single ordered categorical variable by:

1) Assigning all households who actually report that they do not hold the asset or have the liability in question to the "$0" category;[8]

2) Assigning households where the actual non-zero value or bounding values are known to an ordered amount category. Amount ranges for the individual categories could be based on the bracket categories that apply to the individual item or in lieu of brackets, the universal range card categories (see Figure 1);

3) Assigning households which report neither actual amounts nor boundary values to a missing data code. For these cases, the category of the ordered variable itself will be imputed before imputation of an actual amount value can take place.[9]

Each category of this constructed variable will contain a mixture of households where actual values are known (approx. 80%-90%) and households where actual values must be imputed (approx. 10%-20%). A total of Q=22 such ordered categorical variables would be constructed, one for each asset and liability component of net worth. Corresponding to each of the Q categorical variables is the continuous amount variable for the item: 1) $0 valued; 2) non-zero valued and known; or 3) non-zero valued and not known.

The following section describes a general location model for these data that may be used for multivariate model estimation problems and for imputation in a multivariate context.

6.C.2. The General Location Model

Following Little and Rubin (1987), for the complete data case we assume a random sample of n observations on Q categorical variables and P continuous variables. [In our special case, n=7078 there are Q=P pairs of categorical and continuous variables, and there is a one-to-one relationship between the value of the jth (j=1,...,P) continuous variable and the categories of the jth categorical variable.] Fixed by the boundary values for bracketing questions and range card[10] bounds, each of the j=1,...,Q categorical variables has I_j ordered categories (including the zero value and the open ended category). The complete set of categorical variables defines a Q-way contingency table with $C = \sum_j I_j$ cells. The vector of categorical variables for the ith household is the (1 x Q) dimensional vector Z_i. The corresponding vector of continuous variables for the ith respondent household is $\underset{\sim}{x}_i$. If we array the C cells of the contingency table as a 1 x C vector, a household's vector of categorical variables, Z_i can be used to construct its 1 x C indicator vector, W_i, which contains a 1 in the cell position for the household case and zeros elsewhere. In theory there are m=1,....C such indicator vectors, E_m, each with a one value in the mth position and zeros elsewhere.

Again, following the presentation by Little and Rubin (1987), the general location model for the joint distribution of the response data ($\underset{\sim}{x}_i$, $\underset{\sim}{w}_i$) is specified as follows:

For the marginal distribution of the cell indicator variable, $\underset{\sim}{w}_i$, we assume a multinomial probability model

$$Pr(w_i = E_m) = \pi_m, \quad m = 1,\ldots,C; \Sigma \pi_m = 1.$$

Conditional on the case belonging to the mth multinomial cell (i.e., asset and liability configuration), $W_i = E_m$. the distribution of the continuous values is assumed to be multivariate normal with mean vector, μ_m, and common variance covariance matrix, Ω.

$$Pr(x_i / w_i \cdot = E_m) \overset{iid}{\sim} N_p(\mu_m, \Omega)$$

For the full model, the complete parameter vector consists of:

$$\theta = (\Pi, \Gamma, \Omega)$$

where

$\underset{1\tilde{}c}{\Pi} = (\pi, \ldots, \Pi_c)$ *the* 1*xC vector of multinomial cell probabilities;*

$\underset{C\tilde{}p}{\Gamma} = (\mu_{mj})$ *the CxP matrix of means,* 1 *mean value for each of p multivariate normal variables in each of m=*1,...,*C cells.*

$\underset{p\tilde{}p}{\Omega}$ = *the PxP variance/covariance matrix that is common to the multivariate normal distribution of each of the m=*1,...,*C multinomial cells.*

7. Multiple Imputation of HRS Net Worth by GIBS Methods

Under the assumption of an ignorable missing data mechanism, the method of multiple imputation (Rubin, 1987) provides analysts with a tool to explicitly estimate and incorporate the variance of the imputation process and thereby draw correct inference in their analysis of survey data. GIBS methods or iterative posterior simulation techniques (Meng and Rubin) are iterative algorithms designed to yield

Bayesian point estimates or make Bayesian inference for parameters of multivariate distributions.

A highly detailed discussion of inference under multiple imputation or posterior simulation methods is beyond the scope of this paper.[11] Of importance here is the fact that recent theoretical developments have identified a natural link between Bayesian posterior simulation and multiple imputation of item missing data. Software for the necessary computations is now becoming available (Kennickell, 1992; Schafer, 1991).

7.A. Overview of the GIBS Sampling Cycle

A GIBS sampling approach to posterior simulation and multiple imputation (Raghunathan, 1993) applicable to the HRS involves a sequence of many iterations. As in Section 6.D.2, the complete household data vector of categorical and continuous variables is denoted as Y_i. Collectively the observed values for all households will be represented by Y_{obs} and the missing values by Y_{mis}. Each iteration involves a "P" step and an "I" step (Tanner and Wong, 1987).

In the "P" step, k independent sample draws of parameter values are made from the posterior density:

$$\Theta^{(t)} \sim P\left(\Theta / Y_{obs}, Y_{mis}^{(t)}\right)$$

In the "I" step, independent draws for all missing values are made from the k conditional predictive distributions of the missing values corresponding to the k parameter draws of the preceding P step:

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} / Y_{obs}, \Theta^{(t)}\right)$$

After a large (infinite) number of iterations of the P and I step cycle, the following distributional approximations are expected to hold (Schafer, 1991):

$$P\left(\Theta / Y_{obs}, Y_{mis}\right) \rightarrow P\left(\Theta / Y_{obs}\right)$$

$$P\left(Y_{mis} / Y_{obs}, \Theta\right) \rightarrow P\left(Y_{mis} / Y_{obs}\right)$$

At the last iteration of the I step, the k independent draws of missing data values for the data vector Y provide a multiple imputation of k independent replicates.

7.B. Applying the GIBS Procedure to HRS Net Worth Imputation

For data structures that can be represented in the form of a general location model, Raghunathan (1993) describes the application of GIBS sampling to multiple imputation of item missing data. Here we summarize the procedure as it applies to the general location model for the HRS net worth components.

<u>**Initializing the Process**</u>

1. First, to simplify the estimation of the multinomial distribution parameters of the model, a loglinear model for the cell probabilities, π_m, is specified for the data.

2. Next, the I step of the GIBS sampling cycle is initialized. Simple imputation procedures such as the Hot Deck Method or random imputation can be used to supply starting values for missing observations in each household's vector of net worth component values.

3. Recall that conditional on the multinomial cell, the distribution of actual values for asset and liability components is presumed to be multivariate normal with mean vector, μ_m, and common covariance matrix, Ω. From the initial "completed" data vector produced in (2), Raghunathan's suggestion is to apply bootstrap sampling to the observations in each of the m=1,...,C multinomial cells and to use the mean of these samples as the initial draw of the mean value parameter vector.

<u>**The P Step**</u>:

4. Based on the starting values obtained in (2 and 3), the posterior distributions of the parameters are estimated. Assuming a flat prior distribution, Raghunathan (1993) provides the posterior distributions for all model parameters (including log-linear model parameters). New values of the parameters are then drawn from the estimated posterior distributions.

<u>**The I-Step**</u>:

The I step of the GIBS cycle for the HRS net worth data actually would consist of two steps: i) a draw of a multinomial cell for cases where this information is missing; ii) conditional on cell assignments, a sequence of draws of missing data values for the j=1,...,Q net worth component values.

5. I-Step 1: The P-step will supply current values for the log-linear model parameters, which in turn supply estimates of each of the cell probabilities, π_m. If actual or bracketing values are missing for household assets or liability components, the exact cell assignment for that household is not known. Conditional on the current vector of cell probabilities, $\pi_m^{(t)}$, each such household is therefore assigned to a cell with probability proportionate to the cell probability.

6. I-Step 2: For the current cycle, each household is now assigned to one of the m=1,...,C multinomial cells of the model. Conditional on the assigned cell, missing values in the asset and liability response vector, Y_{mis}, are imputed by sampling from the posterior predictive distribution. In theory, the posterior predictive distribution of the Y_{mis} is a multivariate normal distribution. In practice, the draw from the multivariate predictive posterior distribution is approximated by a sequence of draws from the univariate distributions of the asset and liability variables conditional (generally by a regression model) on the most current draw of each other variable.

At this point, the process returns to step 4 where the posterior distribution of the parameters are reestimated, draws are made from these posterior distributions, and the "I" two-step sequence is repeated. As noted above, for multiple imputations (or posterior simulation), the P and I step are replicated k=1,...,K times at each cycle.

8. Special Problems: Statistical and Practical

The preceding subsections have described a general location model framework for HRS net worth component variables and introduced the possibility that GIBS methods may be used for multiple imputation of item missing data. If this model and imputation method are to be applied to the HRS data, a number of statistical and practical problems must be addressed. Several such problems warrant a note here.

8.A. Categorical Variable Dimensions: Propagation of Cells

Under the general location model, survey response data (or imputation as needed) would be used to assign each HRS household to 1 of C indexed cells in the Q-way contingency table. If the general location model were applied to the full set of 22 components of household net worth and if six ordered amount categories were defined for each component, the maximum number of possible categories would be $C=6^{22} = 1.316 \times 10^{17}$! Of course, all but a small fraction of these cells would be either sampling or structural zeros. To reduce the categorical dimension of the model, we can: restrict the number of variables that are considered; reduce the number of ordered categories per variable; or both. In the case of the HRS net worth missing data problem, the full set of net worth components could be grouped into subsets with high intra-group and low inter-group partial correlations. The number of ordered categories per asset or liability variable could also be reduced. The intended outcome of the variable subsetting and category grouping would be a cross-classification of manageable dimension. The final categorical dimension of the model could then be reduced further by collapsing any remaining sparse cells in the full cross-classification.

The extreme in variable subsetting would be to consider a single net worth component variable at a time which returns us to the simple independent or sequenced regression approaches described in Section 6.C. Extreme collapsing of the ordered categories for net worth component amounts would lead us to consider simpler multivariate models for strictly continuous data. [See Section 8.B.]

8.B. Zero-Value Amount Categories

The continuous univariate distribution of each HRS net worth component amount is truncated at a lower value of zero. Most HRS respondent households will have a $0 value for the majority of the 22 assets and liabilities measured in the survey questionnaire. The adaptation of the multivariate data general location model to the HRS net worth component variables partitions the full range of each continuous amount variable into bounded intervals (ordered amount categories). Even if the number of cells (categorical dimension) of the general location model is reduced to manageable levels by subsetting variables and combining interval categories (Section 8.A.), many households will still have one or more zero-valued variables in the subset. The model assumption of multivariate normality clearly breaks down for any cell that explicitly includes the "zero-value" category for one or more of the net worth component variables. Within these cells, the values of those amount items will be exactly zero and not a random variable over some bounded interval -- and variance and covariances for such items will be zero.

One suggested approach to this problem is to bypass the assignment of households to cells based on ordered amount categories and to approach the problem using a simple multivariate model for strictly continuous data. The GIBS methods presented in Section 7 could be used to estimate model parameters and impute missing values from the full range of each continuous variable. In the I step of each GIBS cycle, item missing data would be imputed by a sample draw from the posterior predictive distribution for missing data values. For those cases where bracketing or range card data is available, the drawn value would be accepted only if it lies in the bracket or range card interval.

This approach might be considered for a general location model in which the range of each variable was divided into a limited number of broad categories. For example, two ordered categories might be considered for each component: 1) zero and low-value amounts; 2) medium-value and high-value amounts. Logarithmic or similar transformations of the amount values would further improve the approximation to multivariate normality within the model cells defined by the cross-classification of these broad categories.

8.C. The Open-Ended Categories

As outlined in Section 6.D.1, for each net worth component the model includes a single variable with ordered categories that represent ranges for actual amount values. The last category of each such variable represents a range that is bounded below but not above -- i.e., it is "open-ended." These categories present several problems. First, the distributions of the untransformed continuous value responses within each of these categories are highly skew. This represents a departure from the general location model assumption of multivariate normality for continuous variables within the $m=1,\ldots,C$ multinomial cells.

Compounding the distributional problem in the open ended categories is the limited number of observations that are available for estimating parameters. Estimation and imputation outcomes will be highly sensitive to model misspecification in these cells. If the sparse data problem is too extreme, it may be necessary to look to other approaches for the imputation of assets and liabilities in high wealth households. One alternative would to be to use Cold Deck imputation (Kalton, 1983). Cold deck donors would be obtained from other survey data sources such as the 1989 Survey of Consumer Finances (SCF) (Heeringa, Woodburn and Juster, 1990), a survey that contains large numbers of very wealthy households with complex mixtures of assets and liabilities.

8.D. Keeping the Process Manageable

The necessary statistical theory and program applications to employ GIBS methods for multiple imputation are available now and will be tested on the HRS net worth missing data problem. However, the sheer size and scope of the problem raise a practical concern related to the time required to design, test, implement and evaluate the imputation procedure. In applying GIBS sampling methods in multiple imputation of item missing data on the 1989 SCF data set, Kennickell (1992) describes problems with slow convergence of the GIBS algorithm and general time-to-completion when the imputation procedure is run on current generation UNIX machines. It is unclear to what extent the 1989 SCF experience could be improved upon with alternative algorithms or streamlining of program steps.

9. Summary

Nonresponse adjustment and imputation methods have been developed to attenuate the potential biases from these sources of nonsampling error. Multiple imputation methods provide the tool to estimate the error (variance) that is added to the data in attempting to adjust for missing data biases. By their nature, most nonresponse adjustment and imputation methods have been approached as post-hoc procedures. However, through survey experience and theoretical and empirical research, a point has been reached where samples, designs, questionnaires and survey procedures can better address the problems of nonresponse and item missing data.

The use of bracketing follow-up questions and range card response options in the HRS is one example where questionnaire design and interviewing procedures have been specifically designed to address the item missing data problems for financial variables. As discussed in Section 5, the additional information collected by these methods can play a valuable role in the model estimation and imputation of item missing data for the components of household net worth. The general location model (Section 6) for the resulting mixed categorical and continuous data provides a useful framework for model estimation and for imputation of item missing values. Provided several statistical and practical problems with this multivariate model can be overcome, GIBS methods are a promising approach to multiple imputation of item missing data for household net worth components and other household financial variables.

References

David, M., Little, R. J. A., Samuhal, M. E., and Triest, R. K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, **81**, 29-41.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **B39**, 1-38.

Geman, D., and Geman, S. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian reconstruction of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

Heeringa, S., Juster, F. T., and Woodburn, L. (1991). The 1989 Survey of Consumer Finances: A survey design for wealth estimation, forthcoming in *The Review of Income and Wealth*.

Kalton, G. K. (1983). *Compensating for Missing Survey Data*. Ann Arbor, MI: Survey Research Center, Institute for Social Research.

Kennickell, A. B. (1992). Imputation of the 1989 Survey of Consumer Finances: Stochastic Relation and Multiple Imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 57-77.

Li, K. H. (1988). Imputation using Markov Chains, *Journal of Statistical Computing Simulation*, **30**, 57-79.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. A., and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497(492?)-512.

Madow, W. G., Nisselson, H., and Olkin, I. (eds.), (1983). *Incomplete Data in Sample Surveys*. New York: Academic Press.

Meng, X. L., and Rubin, D. B. (1991). Recent extensions to the EM algorithm, *Fourth Valencia International Meetings on Bayesian Statistics*, Peñiscola, Spain.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 448-465.

Raghunathan, T. E. (1993). A split questionnaire survey design. Manuscript submitted for review to the *Journal of the American Statistical Association*.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, **69**, 467-474.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Rubin, D. B. (1980). Handling nonresponse in sample surveys by multiple imputations. U.S. Bureau of the Census, Washington, D.C.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceedings of the Statistical Computing Section of the American Statistical Association*.

Sande, I. G. (1981). Imputation in surveys: coping with reality. *Survey Methodology*, **7**, 21-43.

Sande, I. G. (1983). Hot-deck imputation procedures. In *Incomplete Data in Sample Surveys, Volume, 3, Proceedings of the Symposium*. W. G. Madow and I. Olkin (eds.). New York: Academic Press, pp. 334-350.

Schafer, J. L. (1991). Algorithms for multiple imputation and posterior simulation from incomplete multivariate data with ignorable nonresponse. Doctoral theses. Department of Statistics, Harvard University, Cambridge, MA.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, No. 398, 528-550.

Footnotes

[1] The eligibility criteria require that at least one member of the married couple belong to the 1931-1941 birth-year cohorts. If a married couple meets this basic eligibility definition, both spouses are administered the person interview regardless of their individual ages.

[2] The number of brackets and the associated dollar amounts vary to reflect differences in the range of the underlying asset distribution.

[3] The n=7078 case data set used to develop Table 2 is a preliminary version of the full HRS Wave 1 data set of n=7703 cooperating households.

[4] The bracketed value category includes cases in which, due to nonresponse or uncertainty, the boundary values for the amount may span two or three of the actual bracket ranges for the item question.

[5] One alternative would be to create an index of "missingness" which is the ratio of the sum of reported amount to total household net worth. For such a computation, amounts for missing items could be imputed or assigned median values for bracket ranges.

[6] In theory one could pursue the joint distribution of all survey variables; however, this is practically impossible and unnecessary. In practice, we will work with small sets of dependent variables (say 3 to 6) that are well correlated with each other and a selected set of covariates. See Section 6.D.

[7] The author wishes to thank T.E. Raghunathan and J.L Schafer for their willingness to share ideas and software.

[8] It is possible for an HRS household to hold an asset that has $0 dollar value. However, for purposes of the imputation procedure and net worth computations, there is no penalty to combining these $0 value asset holders with households who do not hold the asset in question.

[9] The small number (<1%) of cases where the respondent did not report holding or not holding an asset or liability should be distinguished from cases where there is complete missing data for a non-zero amount.

[10] For the HRS net worth component problem, there are few true structural zeros in the multi-way contingency table. Structural zeros could arise in pairings of assets and liabilities, e.g., it is unlikely that a household would have a first mortgage on a home if they did not in fact own the home.

[11] For a discussion of Gibbs Sampling, one GIBS technique that applies to the general location model of Section 6.D, the reader is referred to Geman and Geman (1984), Li (1988) or Schafer (1991).