MULTIPLE IMPUTATION OF INDIVIDUAL SOCIAL SECURITY BENEFIT AMOUNTS--PART I

Thomas N. Herzog, Department of Housing and Urban Development
Clarise Lancaster, Social Security Administration

The problem of nonresponse in sample surveys is presently of considerable interest. For example, the National Academy of Sciences recently formed a panel on Incomplete Data and there has been an increase in the number of papers on this subject. Rubin [1978] suggests three possible reasons for this upsurge in interest in the nonresponse problem:

1. Current surveys appear to be suffering serious problems of nonresponse.
2. There exists a growing awareness that standard methods of handling nonresponse may not be entirely satisfactory.
3. Missing data problems form a fertile area for statistical research.

Our work entails the application of a procedure suggested by Rubin [1978 and 1979]. We will describe a two-stage imputation protocol which we have used to predict the social security benefit amounts of a number of individuals surveyed during the March 1973 Current Population Survey (CPS). We have considered this survey because we have administrative data for the nonrespondents and so are able to validate our results. We will compare the results of 100 applications of the two-stage protocol to those of the Census Bureau's 1973 hot-deck procedure. Finally, we will argue that--

1. at least two values should be imputed for each missing item (regardless of the type of imputation protocol employed), and
2. although an explicit modeling scheme may be much more expensive than a generalized hot-deck procedure applicable to a wide range of variables, such a scheme should sometimes be employed to impute the missing values of one or more key variables.

This work has been partitioned into two parts. Part I consists of three sections; Part II has two sections as well as the list of references. The initial section contains a brief description of the CPS as well as an outline of Rubin's basic approach. In the last two sections of Part I, we describe our attempts to construct an imputation protocol for one particular variable. The first section of Part II discusses our procedure for randomly generating 100 sets of imputed values. Finally, in the last section, we compare our results to those of the Census Bureau's 1973 hot-deck procedure.

BACKGROUND AND INTRODUCTION

The Current Population Survey (CPS).--The Census Bureau's Current Population Survey (CPS) has a multi-stage, stratified clustered sample design. It is a monthly household survey of over 50,000 households and about 150,000 individuals within these households. The Current Population Survey's principal goal is to estimate the labor-force status of non-institutionalized civilians at least 14 years of age. Every March CPS interview includes a series of (supplementary) income questions designed to ascertain all of the respondent's sources of income during the preceding calendar year.

As already mentioned, we have restricted our attention to CPS data, specifically to the CPS-SSA-IRS Exact Match File (see Aziz, Kilss and Scheuren [1978]). This file is based on the March 1973 CPS. In order to improve the reliability of the income data, the CPS database was combined with SSA and IRS administrative records. Essentially, this process consisted of "matching" individual CPS respondents to their 1972 tax return as well as their SSA earnings and benefit information. The "Match File," then, combines the probability sampling aspects of the CPS with the more reliable administrative records of the IRS and SSA.

At present the Census Bureau (see, for example, Coder [1978]) is using a modified hot-deck procedure to impute missing data items in the CPS. This procedure assigns the value of an item from a complete record to the record having the corresponding value absent. The complete record chosen is identical or nearly identical to the incomplete record as far as certain respondent-supplied characteristics are concerned.

Anticipated Improvements.--Imputation protocols of the type proposed here should improve the quality of the database resulting from the sample survey. The imputation of at least two values for each missing datum gives the individual analyzing the resulting public-use file the opportunity to calculate the statistics of his choice without having to model the missingness, and to examine the extra variation in these statistics resulting from the nonresponse.

Potential Improvements.--Relative to the Census Bureau's extant imputation scheme, we feel that our method could also result in the following improvements:

1. a reduction of the bias of the estimates of interest,
2. a reduction in the true sampling variance of these estimates,
3. a more accurate method for estimating the variances of interest, and
4. the ability to display the sensitivity of the answers to the assumed similarities between the respondents and the nonrespondents.

Improvements 1 and 4 are possible because we can incorporate into our imputation protocols our notions of the known or suspected differences between respondents and nonrespondents. A more detailed explanation is required concerning improvements 2 and 3.

The true sampling variance may be considered, intuitively, to have two components. One is due to the variation of the answers of the respondents. The other is due to the variation associated with the distributions used to generate the imputed values for each missing item. Frequently, the first component is the only one taken into account when sampling variances are calculated. As Rubin [1979] points out, this results in an underestimation of the true variance. Rubin [1979] also shows that the use of two or more imputations per missing value will usually lead to a reduction in the true sampling variances of interest.

Rubin's Approach.--At this point, it is appro-

priate to present a brief outline of Rubin's basic approach. The first step is to construct a model of the distribution of the respondent data for an individual income item. Such models may, for instance, be of the form $Y = a + bX + e$, where Y represents the income item we are trying to impute for the nonrespondents, X represents the available background data, "a" and "b" are unknown parameters, and "e" is a random error term.

The second step is to calculate the posterior distribution of the parameters of the model chosen above. For example, in the regression model above, we would calculate the distribution of the parameters "a" and "b."

The distribution may then be modified before being applied to the nonrespondents' data. This modification reflects our external notions of the similarities between respondents and nonrespondents. By "external," we mean, for example, evidence obtained from related studies. Rubin [1977] presents a specific example of this type of adjustment procedure.

Since Rubin [1979] has already described his proposed imputation scheme in full detail, we would prefer not to say too much more about it now. However, a few additional highlights are in order. First, Rubin [1979] suggests carrying out this step-by-step procedure for several "reasonable" protocols rather than just one. This is important because it will enable us to display the sensitivity of answers (i.e., summary statistics based on imputed values) to a range of protocols. For example, we should be able to see whether the variation in answers, as we try a variety of reasonable imputation protocols, swamps the usual standard errors that would be associated with the answers. At the other extreme, we should be able to see whether the usual standard errors that are associated with the answers swamp the variability that we see in the answers as we move from one reasonable imputation protocol to another. Performing repeated data analyses on the original dataset with different imputed values is a natural way to display this sensitivity.

In the work that we will discuss in this paper, we have only considered a single protocol. Because the nonrespondent data are quite similar to those of the respondents, we felt that it was not necessary to modify the posterior distribution of the respondent parameters.

CONSTRUCTION OF MULTIPLE LINEAR REGRESSION MODEL

In this section and the next, we will describe our attempts to use some data on CPS respondents to construct an imputation model for predicting individual OASDI (old age, survivor and disability income) payments for calendar year 1972.

The Basic Data.--We considered only the 1128 individuals surveyed during the March 1973 CPS who had all of the following characteristics:

1. at least 62 years of age as of December 1972,
2. male,
3. in panels 1, 2, or 5 of the March 1973 CPS,
4. responded for themselves, and
5. had a usable administrative record. 1/

Characteristics (1) - (5) above were chosen for the following reasons:

1. We wanted to consider only those people who met the minimum age-eligibility criterion for an SSA annuity.
2. We felt it was difficult (at this time) to determine the amount that individual females (especially, those married, widowed or divorced) would receive from SSA; whereas, for males, this determination is relatively straightforward.
3. Most of the interviews in panels 1, 2, and 5 are conducted in person; in addition, the data from one or more of the five remaining panels can be used later to validate the models constructed using the data of panels 1, 2, and 5.
4. The type of missingness among self-respondents may be substantially different from that among proxy-respondents.
5. Since we wish to predict individual administrative OASDI benefit amounts using the individual's CPS responses, we only want to consider those cases in which the survey data were "matched" with the appropriate administrative OASDI benefit amount.

In all, 1128 individuals satisfied the above criteria. Of these, 999 had no missing (CPS) income amounts of any kind. It is these 999 "respondents" whose data we will use to construct our imputation protocol. Of the remaining 129 individuals (whom we will refer to as "nonrespondents"), 59 failed to answer the CPS question on OASDI benefits and so had their OASDI benefit amounts imputed by the CPS hot-deck. Our goal is to impute the OASDI benefits of these 59 individuals. For ease of terminology, we will use the term "beneficiaries" (or "recipients") to refer to those whose administrative data indicated that they received some payments during calendar year 1972. The entire group of 1128 individuals may be represented as follows:

Exhibit 1.-- Number of Individuals by Response and Beneficiary Status

OASDI Beneficiary Status	Response Status			
	Total	Respondents	Non-respondents*	
Total.....	1128	999	129	(59)
Nonrecipients....	227	200	27	(10)
Recipients.....	901	799	102	(49)
Average Administrative Benefits.	$1827	$1820	$1888	($1886)

*The numbers in parentheses pertain to those who failed to answer the March CPS question on OASDI benefits.

The description of our imputation protocol has been divided into three parts. In the remainder of this section, we will describe the construction of a multiple linear regression model to predict individual recipient OASDI benefit amounts. In the ensuing section, we present a log-linear model to be used to predict OASDI recipiency status. Finally, in the first section of Part II we describe our attempt to use our two-stage imputation protocol to predict the missing OASDI benefit amounts of 59 surveyed individuals.

The Regression Variables.--In this sub-section we will define the variables used in the regression model to predict the benefits of recipient nonrespondents. Table 1 contains a list of all the variables used as well as their item numbers on the Match File. (The reader who does not desire

to wade through all the definitions of these variables is advised to skip this section.)

We attempted to use the data on recipient respondents to predict the benefits of recipient nonrespondents. In all, there were 799 recipient respondents and 102 recipient nonrespondents (of whom 49 did not respond to the CPS questions on OASDI benefits). In order to get a "good" fit using a linear model, we attempted to predict the natural logarithm of the amount of individual (administrative) SSA benefits multiplied by 100,000. In order to keep the data in integer form and thereby conserve computer storage space, we rounded the product of 100,000 and the natural logarithm of the individual (administrative) SSA benefits to the nearest integer.

In all, 14 variables (or characteristics) were used as independent (or predictor) variables in our regression models. The following numerical variables were employed in their original form:

1. age (62, 63, ...)
2. square of age
3. reported income of rest of family
4. reported individual income other than SSA benefits and earned income (defined below)
5. D1 (defined below)
6. D2 (defined below)
7. number of years of school (0, 1,..., 16, $\geq$17)

In order to define D1 and D2, we let

X = CPS reported wages for 1972

and

Y = the sum of reported farm and nonfarm self-employment income.

We then define "earned income of an individual" as Z = X + max (0,Y).

For those individuals aged 72 and over we define D1=D2=0. For the rest of the individuals who are between 62 and 71, we define

$$D1 = \begin{cases} 0 & Z \leq \$1,680 \\ Z - \$1,680 & \$1,680 < Z \leq \$2,880 \\ \$1200 & Z > \$2,880 \end{cases}$$

and D2 = min [max (0, Z - $2,880), $6,152] where Z is as defined above.

The variables D1 and D2 were employed to aid in the prediction of benefits for those ages 62-71. The motivation for these particular definitions is as follows: In 1972, OASDI beneficiaries under age 72 could earn up to $1,680 without having their OASDI benefits reduced. Benefits were reduced by half a dollar for each of the first $1,200 earned in excess of $1,680 and by one dollar for each dollar earned in excess of $2,880. Those earning at least $6,152 received no OASDI benefits. We hoped that the D1 and D2 variables would help us to incorporate the reduced benefit features into our model. Since those aged 72 or older were exempt from the reduced benefit provisions of the Social Security regulations, we set D1=D2=0 for these individuals.

We will now discuss the seven "qualitative" independent variables used in our regression models. These variables were constructed so that those categories observed most frequently were generally assigned a value of 0.

1. race and ethnicity status (0=white non-Spanish, 1=other)
2. veteran's status (0=non-veteran, 1=veteran)
3. interview type (0=interview conducted in person, 1=otherwise)
4. insured status 2/:
   - 0 if "not insured"
   - 1 if "fully insured and eligible for disability"
   - 2 if "fully insured but not eligible for disabiity"
   - 3 if "currently insured only."
5. location (i.e, central city, ring of SMSA (Standard Metropolitan Statistical Area), urban non-SMSA, rural non-farm, and rural farm). Four 0-1 indicator variables are employed. All four indicator variables are set equal to zero for individuals residing in a central city; otherwise, three of the indicators are set equal to zero and the fourth, corresponding to the type of location of the individual, is assigned a value of one.
6. number of weeks worked during calendar year 1972. This variable was partitioned in the Match File as: 0, 1-13, 14-26, 27-39, 40-47, 48-49, and 50-52 weeks. Six 0-1 indicator variables were formed based on the number of weeks worked. All were set equal to zero for individuals working at least 50 weeks; otherwise a single indicator, corresponding to the number of weeks worked, was set equal to 1.
7. marital and household status. Six indicator variables were formed. All six were set equal to zero for individuals who were "married with spouse present." The other indicator variables corresponded to the following classifications:
   1. single and head of household,
   2. single but not head of household,
   3. widower and head of household,
   4. widower but not head of household,
   5. other marital status and head of household, and
   6. other marital status but not head of household.

The Regression Analysis.--We first performed the regression analysis separately for each of panels 1, 2, and 5. The coefficients of determination, R^2, and the square roots of the residual mean squares are shown in Exhibit 2. The coefficient estimates are exhibited in Table 1.

Exhibit 2.-- Summary Statistics for Panels 1, 2, and 5 Before Deletion of Outliers

Panel Number(s)	No. of Observations	R^2 (log of benefits)	R^2 (benefits)	Residual Root Mean Square (of logs)
1	278	.4458	.3326	32,445.
2	260	.3322	.2735	39,924.
5	261	.2645	.2085	38,970.
1,2,5	799	.2840	.2218	37,815.

The combined model consisted of the same 27 independent variables as the separate panel models. We did not add an indicator variable for the panels. We gave each of the 799 observations equal weight in the combined model just as we

did in the separate panel models.

Plots of the predicted values versus the residual values revealed a small number of outliers. We adopted a naive procedure for eliminating outliers. For each panel, we first constructed a regression model using all of the recipient values. We then deleted all those cases whose residual values exceeded 2.5 times the square root of the residual mean square 3/ (in absolute value). We repeated the above procedure (separately for each panel) until all of the residual values were less than 2.5 times the square root of the residual mean square (in absolute value).

This process resulted in the elimination of three outliers from panel 1, nine from panel 2, and four from panel 5. All of the outlier values were well below $1,888, the average observed (nonzero) administrative amount. The smallest administrative OASDI values on the 799 recipients are presented below. The values of the outliers have been underlined. The ellipses represent omitted intermediate values.

Panel 1 (3 deleted items)
215 245 421 525 602...875...

Panel 2 (9 deleted items)
135 204 276 430 551 605
609 609 609 609 661 713
715 728... 774...854...

Panel 5 (4 deleted items)
85 241 299 420 645...

Since the minimum monthly annuity payment was $56.32, in December of 1972, the majority of the outlier values probably represent individuals who only received OASDI benefits for a few months of 1972 and, therefore, had an "artificially" low benefit level.

The summary statistics for each of the regression models formed after the deletion of outliers are shown in Exhibit 3. It is interesting to note that the coefficient of determination continues to decrease as the number of months in the sample increases. Moreover, this pattern is more pronounced after the deletion of outliers than before. The regression coefficient estimates are displayed in Table 1.

Exhibit 3.--Summary Statistics for Panels 1, 2, and 5 After Deletion of Outliers

Panel Number(s)	No. of Original Observations	No. of Points Deleted	R^2 (of logs of benefits)	$\bar{R}^2$ (of benefits)	Residual Root Mean Square (of logs)
1	278	3	.4266	.3463	29,984.
2	260	9	.3795	.3094	32,127.
5	261	4	.2055	.1898	33,702.
1,2,5	799	16	.2877	.2514	32,043.

While many of the predictor variables were not significant (as evidenced by their t-statistics), we left them in the final regression model because we felt they would probably improve the prediction.

Finally, we used the combined model (without the outliers) to predict the individual benefit amounts of each of the 102 recipient nonrespondents. We found that the average predicted value was $1,834 compared to an actual average administrative value value of $1,888--a difference of about $54 (compared to a difference of $112 before the deletion of outliers).

CONSTRUCTION OF LOG-LINEAR MODELS

In this section, we will describe some log-linear models used to predict the OASDI beneficiary status of each member of a group of individuals interviewed during the 1973 Current Population Survey. The individuals whose beneficiary status we wish to predict are those defined previously to be non-respondents. We will use the data on the 999 respondents to construct our basic prediction model.

During calendar year 1972, almost all U. S. citizens at least 72 years of age were entitled to OASDI benefits (irrespective of the extent of their prior contributions to the Social Security System). For this reason, we decided to construct separate models for those ages 62-71 and those at least 72 years of age.

Modeling Those Age 62-71.--The model employed consisted of three predictor variables: earned income, predicted OASDI benefit amounts, and age. For the reasons given above in the definitions of D1 and D2, earned income was partitioned into two levels: (1) less than $1,680 and (2) greater than or equal to $1,680. The predicted OASDI benefit variable (obtained from the basic combined regression model already defined) was partitioned into the five intervals shown in Exhibit 4. The age variable was partitioned into 10 categories--one for each of the individual ages 62-71. The basic frequency count data (summed over the age categories) are displayed in Exhibit 4. In order to ensure that none of the one-way marginal totals was equal to zero, we added a small amount--specifically 0.05--to each of the cells whose observed frequency count was equal to zero. This was done because our computer program does not work if any of the marginal totals is zero; however, it might have been more judicious here to add an amount smaller than 0.05 to each zero cell. The model involving the earned income-predicted benefit amount two-way interaction variables together with the one-way age variable was the best of those considered. The estimated parameter values of this model are displayed in Table 2.

Exhibit 4.-- Observed Frequency Counts for Those Age 62-71

	Beneficiary Status			
	Non-Recipient		Recipient	
Predicted OASDI	Earned	Income	Earned	Income
Benefit Amount	< $1680	≥ $1680	< $1680	≥ $1680
< $1,250	5	96	8	23
$1,251-$1,500	1	20	48	8
$1,501-$1,750	4	7	91	15
$1,751-$2,000	11	3	107	11
> $2,000	20	4	127	4

Modeling Those At Least 72 Years of Age.--The proposed model for this group contained two predictor variables--earned income and predicted OASDI benefits--partitioned as before. Because all of the higher earned-income non-recipient cells had an observed cell frequency of zero, we decided to add 0.05 to each of these 5 cells, as well as the other zero cell. The choice of 0.05 was made because the addition of such a small amount had

virtually no effect on the final model. The basic frequency count data are shown in Exhibit 5. The best model involved both the earned income and predicted OASDI benefit one-way marginals. The estimated parameter values of this model are shown in Table 2.

Exhibit 5.-- Observed Frequency Counts for Those at Least 72 Years of Age

	Beneficiary Status			
	Non-Recipient		Recipient	
Predicted OASDI	Earned	Income	Earned	Income
Benefit Amount	< $1680	≥ $1680	< $1680	≥ $1680
≤ $1,250	5	0	13	0
$1,251-$1,500	7	0	58	5
$1,501-$1,750	5	0	113	5
$1,751-$2,000	1	0	101	5
> $2,000	11	0	35	22

ACKNOWLEDGEMENTS

We would like to thank Stephen Chilton, Graham Kalton, Daniel Kasprzyk and especially Donald Rubin and Fritz Scheuren for their many helpful comments and suggestions. We are also most grateful to Mrs. Beverly Scott for typing the innumerable drafts of this work.

FOOTNOTES

1/In the CPS, there are eight rotation groups or panels, the number of the individual panels corresponding to the number of months an address in the sample has been included in the survey.

2/See pages 242-249 of Aziz, Kilss, and Scheuren [1978], for the definition of this term. The values of this variable were unfortunately entered into the regression model in the numeric coding scheme of the Match File rather than being converted to indicator variables. This was due to a computer programming error. However, since the vast majority of those surveyed were fully insured most of the values used in the regression were either 1 or 2. Consequently, we feel that this error should not affect the results very much.

3/These are calculated under the simple random sampling assumption.

Table 1

Item Numbers and Coefficient Estimates for the Combined Model

VARIABLES	*MATCH FILE ITEM NUMBER(S)*	*COEFFICIENTS OF THE COMBINED MODEL BEFORE DELETION OF OUTLIERS*	*AFTER DELETION OF OUTLIERS*
CONSTANT		.01212	.28758
AGE	1.33	1.38584	1.27725
VET STATUS	1.44	1.05687	1.06119
NUMBER OF WEEKS WORKED:	1.61		
0		1.05370	1.07173
1-13		1.01291	1.02550
14-26		1.06564	1.08550
27-39		.79221	.79242
40-47		.73134	.93517
48-9		1.01115	1.00522
NUMBER OF YEARS OF SCHOOL	1.68	1.00875	1.00285
MARITAL/HEAD OF HOUSEHOLD STATUS:	1.41 *AND* 1.40		
SINGLE/HEAD OF PRIMARY HOUSEHOLD		.99458	.95052
SINGLE/NOT HEAD OF PRIMARY HOUSEHOLD		.92415	.92831
WIDOWER/HEAD OF PRIMARY HOUSEHOLD		.97879	.94594
WIDOWER/NOT HEAD OF PRIMARY HOUSEHOLD		.86673	.87255
OTHER/HEAD OF PRIMARY HOUSEHOLD		.99825	.96484
OTHER/NOT HEAD OF PRIMARY HOUSEHOLD		.93415	.93324
INCOME OF OTHER FAMILY MEMBERS	1.88 *LESS* 1.86	1.00000	1.00000
RACE/ETHNICITY STATUS	1.40 *AND* 1.45	.84685	.85938
INSURED STATUS	2.24	.81267	.79449
AGE SQUARED	1.33	.99783	.99836
D1	1.70+1.72+1.74	.99975	.99987
LOCATION:	1.23 *AND* 1.24		
RING OF SMSA		.95696	.96706
URBAN NON-SMSA		.93707	.91613
RURAL NON-FARM		.84947	.84396
RURAL FARM		.73727	.76382
OTHER INCOME OF INDIVIDUAL	1.86	1.00001	1.00001
D2	1.70+1.72+1.74	.99993	.99992
INTERVIEW TYPE	1.12	.98876	1.03190

THE DEPENDENT VARIABLE, ADMINISTRATIVE SOCIAL SECURITY BENEFITS, IS ITEM NUMBER 6.5.
OTHER INCOME OF INDIVIDUAL IS MORE ACCURATELY DEFINED AS 1.86-(1.70+1.72+1.74+1.76)
USABLE MATCHES ARE DEFINED TO BE THOSE RECORDS FOR WHICH THE VALUE OF ITEM 7.18 *IS UNEQUAL TO* 4.

Table 2

Parameter Estimates for Contingency Table Models

Variable	Model for those 72+	Model for those 62 - 71
Constant	-2.046	-0.620
Earned Income:		
Less than $1,680	0.855	-0.592
Predicted OASDI Benefits:		
$0 -1,250	0.720	0.582
$1,251-1,500	0.136	-0.151
$1,501-1,750	-0.364	-0.336
$1,751-2,000	-1.09	-0.330
Age:		
62		0.681
63		0.364
64		0.394
65		-0.251
66		-0.317
67		0.135
68		-0.409
69		-0.428
70		-0.0157
Earned Income-OASDI Benefit Interaction Terms:		
$0 -1,250		0.125
$1,251-1,500		-0.599
$1,501-1,750		-0.0320
$1,751-2,000		0.374