A COMPARISON OF THREE ALTERNATIVE TECHNIQUES FOR ALLOCATING UNREPORTED SOCIAL SECURITY INCOME ON THE SURVEY OF THE LOW-INCOME AGED AND DISABLED

Sylvester J. Schieber - Social Security Administration

Survey data gathered from the general public invariably contain some incomplete detail. This frequently results from the respondents' inability or unwillingness to provide the information sought, but, of course, it can also be caused by other facets of the survey process. In dealing with incomplete information on a set of survey items three options are commonly considered:

1. Cases with incomplete information can be discarded from a particular analysis with no adjustments being made to account for the change in structure of the remaining portions of the sample.

2. Incomplete records can be discarded and the remaining records reweighted in an attempt to make the respondents representative of the population originally sampled.

3. Missing data can be imputed or allocated to render the records complete.

Over the years a considerable literature has evolved concerning the impact of missing data on various statistical techniques. A fairly extensive survey and bibliography of this body of literature have been provided by Hartley and Hocking (1971). Rubin (1976) points out that most of the work in this area makes either an implicit or explicit assumption that the process causing missing data can be ignored. However, it is fairly well-known that missing data are frequently not missing at random and that the process by which they arise might have serious implications for any inferences derived from the data set in which they exist.This being the case, the techniques used to allocate missing data can have potentially serious effects on analytical issues for which the data are used.

Recently there has been an increasing interest in the missing data problem and the various implications of the techniques available to handle it. At these meetings, for example, there were over fifteen papers given on this subject. Some of these approach the issue primarily from a theoretical or conceptual perspective. Others report work that is more empirical in nature. Within both of these approaches some authors evaluate the impact of a single technique on various measures (e.g., means, variances, bias, etc.) while others provide a comparative review of several allocation procedures. The present effort is comparative and empirical in nature.

The original intent was to compare three methods for imputing missing Social Security Income Data. As the analysis evolved two variants of one of the initial techniques were also considered. Since the processes themselves were being analyzed the data treated as unknown were actually provided in an interview setting. The results from the various imputation techniques are compared with the original responses and the implications for aggregate measures of Social Security Income are considered.

In Section 1 of the paper there is a brief discussion of the nature of the data that are used for this analysis; the next section describes the allocation techniques being scrutinized; the third section presents the comparisons of the allocated data; and the final section describes some limitations and further research that may be warranted with these data and techniques.

1. THE VEHICLE FOR THE ANALYSIS

The data source on which this analysis is based is the Survey of the Low-Income Aged and Disabled (SLIAD) conducted by the Bureau of Census for the Social Security Administration.[1] The intent of the survey was to establish a baseline on the populations both eligible and potentially eligible for the Supplemental Security Income (SSI) program. SSI is a cash assistance program for aged, blind, and disabled persons, begun in January 1974 and administered by the Social Security Administration. SLIAD elicited a wide range of social and economic data pertinent to an evaluation of the populations to be served by SSI. SLIAD included four national samples selected in 1973 comprising approximately 18,000 noninstitutionalized adults. The data used in this analysis were collected during the period from October to December 1973.

The nature and sizes of the four samples and the noninstitutionalized populations that are represented in 1973 are shown in table 1. There were two separate sets of aged and disabled samples. One of each of these was a sample of individuals who received welfare benefits from one of the adult assistance programs during 1973. They are referred to as the Welfare Samples. The Aged Welfare Sample represents the 1973 Old Age Assistance recipient population and the Disabled Welfare Sample represents the 1973 combined Aid

Table 1: PERSONS INTERVIEWED AND REPRESENTED BY THE SOCIAL SECURITY ADMINISTRATION SURVEY OF THE LOW-INCOME AGED AND DISABLED for 1973

Persons	1973 Aged	1973 Disabled
Adult Assistance- SSI Conversion Cases:		
Sampled	5,192	6,167
Represented (000's)	1,665.2	1,157.9
CPS Retired Rotation Sample Persons:		
Sampled	3,402	2,790
Represented (000's)	15,445.0	4,726.

to the Blind and Aid to the Permanently and Totally Disabled populations. The remaining aged and disabled samples were each a generalized sample of individuals derived from retired rotation groups from the Bureau of Census' Current Population Survey Samples. These latter groups met certain income and categorical screening criteria in order to qualify for sample selection.

The decision to restrict the analysis to Social Security income was based on several factors. An income item was chosen because income data are frequently harder to elicit in a voluntary survey framework than demographic or attitudinal information. Income is also critically important in most of the analyses for which public survey data are used. Given the nature of the four samples available, SSA income was the most prevalent form of income reported. Also, the fact that a match had been made of the interview records with SSA administrative payment data weighed heavily on the decision to select Social Security income as the focus of the analysis.

Typically, the process of imputing data is only invoked in instances where the interview record is incomplete. Information provided in complete records is used to allocate missing data. However, the representative nature of the imputed data vis-a-vis the interview records to which they are attributed cannot be known because comparisons cannot be made with the actual values that the respondents would have provided if they had reported the information. In this instance the missing data are"manufactured" so the allocated data can be compared to data actually reported. The study universe includes only records that originally provided complete information on all of the items utilized in the imputation and analytical process.

The number of cases in the study universe from each of the four samples represents 80 to 90 percent of each of the original samples (see table 2). From each of these four samples a 15 percent random subsample of cases was selected. These cases were treated as if they had provided no information on the survey regarding Social Security retirement, survivors, or disability insurance benefits. The imputation of benefits was done for the respondent and other family members separately. The other family members include the spouse, and/or minor children of the respondent, if present.[2]

2. IMPUTING THE MISSING DATA

The three imputation techniques that were to be tested were a hot deck, an administrative record match, and a two-stage estimation procedure. The first stage of this latter procedure was modified as the analysis evolved. In the discussion this is referred to as the modified two-stage procedure. Another variant, the hybrid two-stage, combines facets of both the hot deck and modified two-stage procedures.

Hot Deck.--The first procedure for imputation used here was a sequential hot deck. The records were classified in a 128 cell matrix: four categories for the parent samples, two race categories (white and nonwhite), two age classifications (18-44 and 45+ for the disabled samples and 65-74 and 75+ for the aged), two sex, two marital status, and two urban-rural classes; the last processed record for whom the Social Security income data were reported had that information stored in the appropriate cell of the matrix. The matrix was updated by each record which contained the reported Social Security information, i.e., it was current or "hot." A record that needed allocation was imputed data from the appropriate cell of the matrix.

Table 2: NUMBER OF RECORDS FROM EACH SAMPLE INCLUDED AND THE ASSIGNMENT OF INCOMPLETE RESPONSES FOR THE ALLOCATION ANALYSIS

Sample	Complete records	Assigned Missing Social Security Income	Remaining records
Welfare Aged	4,499	679	3,820
Welfare Disabled	5,563	858	4,705
CPS Aged	2,712	387	2,325
CPS Disabled	2,417	378	2,039

Administrative Record Match.--The second set of imputations used the SSA Master Beneficiary Record (MBR), the administrative record of beneficiary payments, matched to the survey files on the basis of Social Security numbers. The MBR data were available only for family members of the respondent because Social Security numbers were not provided in the interview on anyone other than the respondent. The MBR data could be compared to nonrespondent members of the family because all benefits paid under the respondent's entitlement were included in the record information. The administrative record provided a separate accounting of benefits accruing to the respondent, the respondent's spouse, and minor children under the entitlement.

Prior to the actual imputation of the interview data several adjustments were made to the administrative record to guarantee direct comparability with the interview file. For example, MBR annual amounts had to be adjusted to reflect the number of months that benefits were received by those who did not receive them for the full 12 months prior to interview because the administrative record calendar 1973 accounting period did not exactly correspond with the survey accounting period of 12 months prior to date of interview.[3] One other important consideration was the automatic deduction made from the monthly Social Security benefit to cover the supplemental portion of the Medicare Insurance (SMI). The monthly deduction amount during 1973 was $6.30 per month. Adjustment for this was made only in the case of the Aged CPS sample, however. Most of the States paid the monthly SMI contribution for eligible adult assistance recipients under their medical assistance (MEDICAID) programs. In fact, the MBR indicated that less than 7 percent of the SLIAD welfare sample respondents were making their own contribution for SMI at the end of 1973. The CPS Disabled did not receive any adjustment in SSA benefits for SMI deductions because they only be-came eligible for the program in July 1973 and only a minority were taking advantage of it by the end of that year. After the family units, time periods, etc., were adjusted for correspondence the missing SSA data were imputed by using the MBR reported benefits as the survey amount.

Multivariate,Two-stage Estimation.--In this case the imputation of Social Security benefits was done by means of a two-stage probability model. The first stage of the model involved the utilization of LOGIT analysis to estimate the probability of receiving Social Security income. The model was estimated using the 85 percent portion of the samples which contained no missing data. The estimated coefficients were then applied to the 15 percent portions of the samples which contained the missing data elements. A decision rule was applied so that a case with a probability of greater than or equal .5 was considered to be a recipient of Social Security and any with computed probabilities below that was not.

The second stage of the model was estimated using stepwise ordinary least squares regression. In this instance the model was obtained by employing only those who reported receiving Social Security income. The estimated coefficients were applied to the cases with missing data which were predicted to be recipients of Social Security benefits in the first stage.

For the "respondent" the probability and receipt of Social Security income were considered to be a function of current employment status,demographic characteristics, education, family composition, region and urban-rural place of residence, and from the respondent's work history profile, career occupation, industry of employment, and attachment to the labor force. For "others in the family" the receipt of SSA income was dependent on the "respondent's" receipt of Social Security and the age and relationship of these other persons to the "respondent." If there was no spouse present, his or her employment status was important. Each of the four samples was estimated separately.

The actual rates of recipiency of SSA income by the two aged samples were quite high (88 percent for the CPS and 65 percent for the Welfare group). This complicated the process of finding a model that would generate a reasonable number of cases to be treated in the imputation as having no SSA income. The LOGIT models estimated in the process of this endeavor consistently showed several variables to be significant in explaining the receipt of SSA, but generally resulted in very high probabilities of receiving Social Security in most cases for the two aged samples. Several alternative specifications of the models were tested for these two samples. Linear versions of the models were also tested. In each instance, virtually all of the CPS aged respondents were attributed with the receipt of SSA income by this technique. Recipiency of SSA benefits was also greatly overestimated for the Welfare aged respondents by this method. None of the modeling variations improved the results. This led to the first variant technique.

The Modified Two-Stage.--In the first stage of this procedure recipiency was assigned randomly on the basis of the group probability of receiving Social Security (see Herzog, 1978, for details). Operationally this entailed generating a stream of random numbers in the interval (0,1) and determining recipiency on the basis of whether or not the random number generated was greater than the group rate of nonreceipt of SSA benefits. The second stage of the procedure again used the OLS regression results from reporting SSA beneficiaries to assign benefit levels once the recipiency issue was resolved.

The Hybrid Two-Stage.--This procedure combined elements of the hot deck and the modified two-stage techniques described above. Recipiency was determined by using the same first-stage process as in the modified two-stage approach. Once recipiency was established, the level of benefits received was calculated by using a combined regression and hot-decking procedure. The hot-deck classification structure was used to add residuals to calculated benefits for "recipients" whose SSA income was being allocated (see Scheuren, 1976, for details). The residuals were calculated for each recipient of Social Security among the 85 percent subsample after the estimation of the regression equation. Then a ratio of the residual to the estimated benefit level was calculated and put into the appropriate hot-deck cell to be used to adjust the calculated benefit levels for those cases being allocated Social Security income. The purpose of incorporating this "residual adjustment" procedure was to protect against the possibility that some part of the response surface fit badly and also to preserve the population variance among the cases being allocated. In order to prevent outliers among the reporting group from causing outlandish adjustments to the calculated benefit levels of those whose information was missing, the "residual adjustment" was limited by the size of the standard error of the regression estimate.[4]

3. COMPARING REPORTED AND ALLOCATED SOCIAL SECURITY INCOME

The hot-deck procedure did a good job of estimating the correct number of SSA recipients in the four samples (see table 3). This occurred for both categories: sample "respondent" and "others in the family" (henceforth: "others"). The administrative record match was also generally in accordance with the interview records with regard to the number of "respondents" receiving Social Security. For the "others" however, the number of recipients based on the administrative record was consistently greater than that from the interview reports. It was only in the case of the Welfare Disabled that the MBR allocations varied to any great extent from the interview records.

The multivariate two-stage model substantially overestimated the SSA income recipiency levels for both aged samples. For the disabled this technique also overestimated recipiency for the CPS group but underestimated it for the Welfare contingent. Allocation of recipiency at random on the basis of the various group probabilities

Table 3.--Number of Recipients of 1973 Social Security Income as Reported by Selected SLIAD Respondents and the Allocated SSA Income by Various Techniques

	Respondent				Others in Family			
	CPS		Welfare		CPS		Welfare	
	Aged	Disabled	Aged	Disabled	Aged	Disabled	Aged	Disabled
Total number of cases	387	378	679	858	114	210	191	218
Receiving Social Security according to:								
Interview Record	342	162	442	258	87	78	124	67
Hot Deck Allocation	338	159	419	246	87	63	120	73
Administrative Record Match Allocation	344	165	448	288	91	85	134	79
Multivariate Two-Stage	386	211	655	175	75	134	153	144
Modified two-stage & two-stage hybrid[1]	348	159	440	234	87	73	110	66

[1] The modified two-stage and two-stage hybrid techniques yield the same distribution in this table because the two-stage hybrid model utilized the first-stage of the modified two-stage approach to establish recipiency.

TABLE 4.--Percent of selected SLIAD respondents correctly classified as receiving or not receiving SSA income during 1973 by various allocation techniques

	Respondent				Others in Family			
	CPS		Welfare		CPS		Welfare	
	Aged	Disabled	Aged	Disabled	Aged	Disabled	Aged	Disabled
Total number of cases	387	378	679	858	114	210	191	218
Percent Correctly Classified by:								
Hot Deck Allocation	80.1	54.8	56.8	60.3	71.9	48.1	57.1	61.5
Administrative Record Match Allocation	98.2	94.5	86.5	93.9	96.3	92.8	91.7	91.8
Multivariate Two-Stage	87.9	58.5	65.1	64.8	59.7	49.5	58.6	44.5
Modified two-stage & two-stage hybrid [1]	79.6	55.8	54.9	58.6	57.9	53.8	45.6	53.7

[1] The modified two-stage and two-stage hybrid techniques yield the same distribution in this table because the two-stage hybrid model utilized the first-stage of the modified two-stage approach to establish recipiency.

TABLE 5.--Means and standard deviations of 1973 social security income as reported by selected SLIAD respondents and allocated by various techniques [1]

	Respondent				Other in Family			
	CPS		Welfare		CPS		Welfare	
	Aged	Disabled	Aged	Disabled	Aged	Disabled	Aged	Disabled
Interview Record								
Mean ($\bar{X}_1$)	1,692.57	1,655.75	1,187.23	1,136.91	1,360.25	1,741.33	1,023.73	1,295.57
σ_{x_1}	721.73	884.94	496.70	580.44	697.25	812.23	584.17	476.62
Hot Deck Allocation								
Mean ($\bar{X}_2$)	1,699.23	1,686.51	1,198.35	1,263.91	1,378.59	1,679.22	950.03	1,288.85
$\sigma_{\bar{x}_2}$	668.83	755.92	491.51	603.18	809.96	723.31	529.41	693.07
$\bar{X}_1-\bar{X}_2$	-6.66	-30.76	-11.11	-126.99	-18.33	62.11	73.70	6.72
$\sigma_{\bar{x}_1-\bar{x}_2}$	53.40	91.80	32.80	52.82	114.58	129.47	68.98	99.85
Administrative Record Match Allocation								
Mean ($\bar{X}_3$)	1,679.69	1,628.50	1,140.59	1,149.26	1,468.06	1,642.00	986.44	1,280.28
$\sigma_{\bar{x}_3}$	650.47	780.86	445.67	561.08	639.70	782.57	555.97	525.90
$\bar{X}_1-\bar{X}_3$	12.87	27.24	46.64	-12.35	-107.81	99.33	37.29	15.29
$\sigma_{\bar{x}_1-\bar{x}_3}$	52.51	92.36	30.70	49.03	104.00	125.15	68.78	83.01
Multivariate two-stage								
Mean ($\bar{X}_4$)	1,918.90	1,129.45	1,322.05	1,022.51	1,588.85	1,052.97	842.16	734.03
$\sigma_{\bar{x}_4}$	394.26	416.47	246.95	456.66	594.60	475.43	369.64	411.25
$\bar{X}_1-\bar{X}_4$	-226.34	526.30	-134.81	114.40	-228.60	688.36	181.57	561.54
$\sigma_{\bar{x}_1-\bar{x}_4}$	43.93	75.21	24.34	50.03	101.50	100.72	57.59	67.56
Modified two-stage								
Mean($\bar{X}_5$)	1,909.66	1,052.44	1,306.95	825.27	1,534.49	955.16	848.33	658.27
$\sigma_{\bar{x}_5}$	393.18	400.49	252.47	396.96	609.12	464.09	343.15	386.56
$\bar{X}_1-\bar{X}_5$	-217.09	603.31	-119.72	311.65	-174.24	786.17	175.41	673.29
$\sigma_{\bar{x}_1-\bar{x}_5}$	44.40	76.44	25.38	44.55	99.26	106.81	59.11	75.20
Two-stage hybrid								
Mean ($\bar{X}_6$)	1,775.71	1,338.43	1,136.46	1,021.49	1,296.09	1,044.51	820.58	1,053.80
$\sigma_{\bar{x}_6}$	646.80	713.87	379.24	602.68	711.99	619.77	412.85	1,043.76
$\bar{X}_1-\bar{X}_6$	-83.14	317.31	50.77	115.43	64.16	696.82	203.15	241.76
$\sigma_{\bar{x}_1-\bar{x}_6}$	52.25	89.66	28.74	53.51	106.84	117.13	63.03	145.77

[1] Includes only records with nonzero amounts in each instance.

of receiving SSA income was generally consistent with reported recipiency.

Two methods of evaluating the effectiveness of the alternative procedures for determining recipiency were employed here. The first focused on aggregate recipiency levels and the other on individual assignments.

In the first instance there is no concern with the proper classification of a specific individual. The model is considered to do a good job if each recipient classified as a nonrecipient is offset by a nonrecipient being classified as a recipient. In other words, the concern is that the model results in the right number of SSA beneficiaries being allocated. The effectiveness of the procedures, in this sense, can be tested by means of the McNemar χ^2 test (Siegel, 1956).

Based on the McNemar test, the first-stage LOGIT procedure outlined, resulted in improper levels of Social Security recipiency in every instance. The deterministic application of the first-stage model as it is currently conceptualized consistently resulted in significantly biased recipiency rates. As indicated earlier, the results of the LOGIT estimations consistently showed several variables to be significant in explaining the receipt of SSA income. Without doubt, using some other decision rule than the 50 percent rule applied in this case could improve the results obtained here. However, in this instance there was no derivation of a systematic, defensible process for establishing such variable decision rules as would be needed for samples with varying characteristics.

The hot deck procedure allocated correct rates of recipiency for both sets of Welfare and CPS "respondents." The administrative record provided a higher than expected recipiency rate for the disabled welfare "respondents." This was also the case for the "others" category for both welfare samples.[5] From the perspective of the McNemar test the random assignment of SSA income recipiency worked extremely well. In the aggregate, 4 out of 5 of the techniques used here allocated generally reliable levels of Social Security income recipiency. The McNemar test does not identify problems of misallocation, however, as long as they tend to offset each other.

The percentage of each of the samples correctly classified by recipiency status is presented in table 4. Without applying any rigorous statistical tests it seems safe to conclude that the administrative record match provides the most consistent set of "correct" classifications. Comparing the percentage of cases in each of the sample categories that were correctly classified as receiving SSA income, the administrative record equalled or surpassed the other techniques in every instance. Beyond that, there is no strong discernible pattern that allows a clear-cut ranking of the other three techniques on their ability to identify correctly Social Security beneficiary status across the four samples. If the procedures are ranked by column in table 4, then some consistency emerges, at least for the "respondent" category. For all four samples the multivariate two-stage procedure ranks second and the hot deck technique ranks third in three-of-four samples in correctly classifying the "respondents'" Social Security recipiency status. While these rankings were consistent for the "respondent" category, the differences in portions of each of the samples correctly classified by the hot deck, LOGIT, and random allocation techniques were quite small. In the "others" category there was no pattern beyond the dominant effectiveness of the administrative record match.

Thus far the discussion of the effects of the various allocations has dealt only with the recipiency of Social Security income. Of equal importance is the amount of income allocated by the various procedures. Comparisons of the mean reported-and-allocated benefit levels are presented in table 5. The mean SSA income allocated by the hot deck procedure was significantly different from that reported by the CPS disabled "respondents." None of the other means (at the two sigma level) from the hot-deck allocation was significantly different from the mean reported amount. Differences in the mean amount of SSA income allocated from the MBR when compared to reported benefit levels were not significantly different from zero in a single instance.

The multivariate two-stage model consistently produced means that were significantly different from the reported amounts. The modified two-stage model rendered essentially the same set of means as the original two-stage model. This is hardly surprising since the regressions were identical, only the cases to which they applied differed.

It was expected that the two-stage hybrid model would generate distributions more like the reported distributions than either of the other two-stage models. In fact it was expected that the two-stage hybrid distributions would be very similar to the hot deck results. The results were somewhat mixed.

In the "respondent" category the mean and population variance estimates were better from the two-stage hybrid than for either of the other two-stage models. In three of the four samples the means and variances from the hybrid were within similar ranges from the reported distributions as was the hot deck from the reported information. The exception was the CPS Disabled Sample. This procedure worked very well for both sets of aged "respondents." The differences in the effectiveness between the aged and disabled may be attributable to the differences in the number of cases being allocated (refer to the bottom line in table 3). In fact the CPS Disabled group where this procedure was least effective contained the smallest number of nonrespondents allocated among the four samples in the "respondent" category. If one considers the hot-deck allocation procedure, it also did better for aged "respondents" than for the disabled on both mean and variance estimation.

In the "others" category the problems appear to be more serious for two reasons. First, the small number of cases is more critical for the "others" category than for "respondents." Secondly, one of the most important predictors of the "others" Social Security benefit levels was the "respondents" benefit level. This made the specification and operationalization of the "respondent" portion of the model very critical. Any error in the allocated "respondent" benefit is compounded in the estimation of the "others" Social Security benefits. This would suggest that the results for the two aged samples should be better than those for the disabled since this procedure worked better for the aged samples. This is generally the case. The estimated mean for the CPS Disabled "others" is much lower by this technique than the mean reported benefit level. For the CPS Disabled this procedure generates a distribution with a much greater population variance than existed for reported benefits.[6]

While this two-stage hybrid model provided mixed results it holds a great deal of promise. The problems encountered here could probably be overcome by sorting the file to insure that no complete record is used in the hot deck to update more than one incomplete record. The results here suggest that this problem might eliminate itself in larger samples.

As a final comparison between the allocated and reported income amounts a series of correlations was run for each of the samples, for "respondents" and "others" separately. As expected, administrative record allocations were the most closely correlated with reported amounts, with the coefficients mostly around .9. None of the other methodologies resulted in coefficients even approaching this and none of the procedures established a clear dominance over the others in this regard.

4. LIMITATIONS AND FURTHER CONSIDERATIONS

Before one becomes a committed advocate of one or the other of these procedures certain reservations should be mentioned. The income nonresponse that was evaluated here was designed to be random. Nonresponse in actual survey settings does not occur randomly in many instances. This has potentially serious implications for two of the procedures used here. If nonresponse is not random then stratification for purposes of hot-deck allocation becomes quite important. In a similar vein, if random allocation were the method of choice, there would need to be some sort of stratification and then random allocation within cells if such a procedure were to be employed. In the research reported here Social Security recipiency was randomly assigned. In a real setting, however, if benefits or nonresponse were structurally associated with the characteristics of the persons being sampled then allocations by this technique should be made within groups as they relate to recipiency and nonresponse. It should be clear that this issue does not preclude the use of these techniques, only complicates them.

The administrative match procedure described here was tremendously effective from every perspective, but the deck was loaded, so to speak. The administrative data used here were payment records generated mechanically and probably precisely. Administrative file matches will not be so effective in instances where the records themselves are dependent on respondent reporting. Then the administrative record is also highly subject to errors of omission and commission compounded by problems of recall and arithmetic. Frequently the items that are present in the administrative record do not correspond definitionally with survey counterparts. In many cases, the unit across which a survey measures variables is quite different from that of the programmatic record. Indeed there are a great many problems that must be reconciled before an administrative record file can be used to allocate missing survey data. The process of gathering and merging the files themselves is very complex. In addition, there are legal and social concerns that must be dealt with.

While all these issues are quite serious the potential rewards from record matching are great. The definitional and unit problems will probably never be totally resolved. Nevertheless, administrative files provide much information that could be employed in the imputation process. While the imputations may not be direct, administrative data could well include critical predictors for generating allocations.

Rubin (1978) suggests that several allocations should be provided for a particular piece of missing information. It would then be the researcher's obligation to choose the optimal allocation based on the research goals subject to the underlying assumptions in the various allocations. This would place the responsibility of selecting the proper allocation technique on the substantive researcher where it should be, but there is no guarantee that this responsibility will be shouldered wisely.

If a series of allocations is to be provided with data sets then it is quite conceivable that identical research techniques can result in different conclusions depending on the set of allocations used. So it becomes important to look not only at the impact of allocation on the items directly affected, but also on research outcomes. For example, what would be the impact of four separate allocations by procedures outlined in this paper, or others, on welfare reform estimates as generated from the 1976 Survey of Income and Education? An enormous number of alternative issues of this nature could be formulated.

ACKNOWLEDGEMENTS

I wish to acknowledge the computational help provided by Stephen Dawes and Willard Edwards of the Social Security Administration (SSA). Beatrice J. Young typed each of the drafts of the paper. Sandy Lancaster of SSA provided editorial assistance on the final version of the paper.

This paper has been substantially revised subsequent to the presentation at the 1978 American Statistical Association Meetings. Special thanks must be given to Thomas Herzog (now at the Department of Housing and Urban Development) who suggested the random allocation of recipiency discussed in the paper. I also to wish to thank John Bailar of the National Institutes of Health for his comments at the presentation of the paper. The greatest contribution to this effort came from Frederick J. Scheuren of SSA who suggested the residual hot-deck procedure. He also read several drafts of the paper and provided many helpful comments and insights that would not have occurred to the author. Any errors are solely the responsibility of the author.

FOOTNOTES

1. For a complete discussion of the general purpose and design of the Social Security Administration Survey of the Low-Income Aged and Disabled see Tom Tissue, "The Survey of the Low-Income Aged and Disabled: An Introduction," Social Security Bulletin, vol. 40, February 1977, pp. 3-11, and Erma Barron, "The Survey of Low-Income Aged and Disabled (SLIAD): Survey Design, Estimation Procedures, and Sampling Variability," Survey Report No. 5 (forthcoming).

2. In the discussion the sampled person is referred to as the respondent. The interviewer was to attempt to interview the actual person selected in the sample when it was possible. If it were impossible for that person to respond then a proxy was accepted. Throughout this analysis, however, the respondent is considered to be the sampled person regardless of proxy or self-response to the questions. In any event, questions pertaining to other family members' income, characteristics, etc. were directed to the person answering the questions. (If these other persons were present during the interview they were not discouraged from helping provide the information being elicited.)

3. There were no cost-of-living increases in SSA benefits during the period covered by the 1973 interview obviating the need for any adjustment of this type.

4. The last step is somewhat arbitrary but is defensible. In several instances here a very small number of cases were being allocated so the implications of outliers were quite serious. Because of the small number of cases in these instances the residual values used in the adjustment were limited to absolute values no larger than the standard error of the regression estimate. In larger samples it would probably be better to allow a larger range of residual variation (e.g., 1.960 for 95 percent or 1.645 for 90 percent inclusion) to account for the population variance in the values being allocated.

5. Part of the difference in the reported and administrative record recipiency rates described may be attributable to a problem some respondents have in correctly classifying certain types of income within a survey framework. A relevant discussion of this problem is presented by Vaughan (1978).

6. The regressions could only be expected to do better than the hot deck if the extra prediction items do add to the explanatory power of the model. The stepwise models included here did not force the hot deck stratifiers into the regressions so the marginal effect of the added variables has not been computed thus far. It is worthy of note, however, that for the "others" category the R^2 for the regression was .50 for the CPS Aged Sample where this allocation procedure worked well but only .30 in the case of the Welfare Disabled where the variance of the allocated variable was much greater than in the distribution of reported benefits.

REFERENCES

[1] Barron, E., "The Survey of the Low-Income Aged and Disabled (SLIAD): Survey Design, Estimation Procedure and Sampling Variability," Survey Report No. 5, Social Security Administration, forthcoming.

[2] Berkson, J., "Maximum Likelihood and Minimum χ^2 Estimates of the Logistic Function," Journal of the American Statistical Association, Vol. 50, (1955): 130-162.

[3] Hartley, H.O. and Hocking, R.R., "The Analysis of Incomplete Data," Biometrics, Vol. 27, No. 4, (1971): 783-823.

[4] Herzog, T., "Imputation Models for Large Scale Surveys," invited paper, National Center for Health Statistics, July 1978.

[5] Rubin, D.B., "Inference and Missing Data," Biometrika, Vol. 63, No. 3 (1976):581-592.

[6] Rubin, D.B., "Multiple Imputations in Sample Surveys - A Phenomenonological Bayesian Approach to Nonresponse," 1978 American Statistical Association Proceedings, Social Statistics Session.

[7] Scheuren, F.J., "Preliminary Notes on the Partially Missing Data Problem--Some (Very) Elementary Considerations," Social Security Administration, Unpublished, 1976.

[8] Siegel, S., Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company, Inc., 1956

[9] Tissue, T.L., "The Survey of the Low-Income Aged and Disabled: An Introduction," Social Security Bulletin, Vol. 40, February 1977: 3-11.

[10] Vaughan, D.R., "Errors in Reporting Supplemental Security Income Recipiency in a Pilot Household Survey," 1978 American Statistical Association Proceedings, Social Statistics Session.