IMPUTATION FOR THE INCOME AND ASSETS MODULE OF THE MEDICARE CURRENT BENEFICIARIES SURVEY (MCBS)

Marianne Winglee, Lana Ryaboy, and David Judkins, Westat, Inc.
Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Keyword: Item nonresponse, hot-deck imputation

1. Introduction

This paper discusses imputation procedures used to assign values to item nonresponses in the income and assets (I&A) module of the Medicare Current Beneficiaries Survey (MCBS), and examines the effects of imputation on estimates of total income and total assets. A form of hot-deck imputation was implemented through WESDECK. The properties of WESDECK are discussed.

2. Data Items Subject to Imputation

The MCBS is an ongoing national survey of current Medicare beneficiaries, and the I&A module is a supplement administered during Round 3 of the MCBS. Participants were current Medicare beneficiaries who lived in the community and were not institutionalized in nursing homes at the time of the survey. Over 10,000 people were involved and they were asked to provide detailed information about their income, assets, and debts in 1991.

Participants were visited at home by an interviewer who administered the questionnaire through a computer assisted personal interview (CAPI) system. Questions were asked about income from various sources (including social security, employment, supplemental social security, public assistance, food stamp, retirement funds such as IRA or Keogh accounts, rents from properties, interest from bank accounts, and dividends from stocks or other investments); the values of assets (including the values of own home, cars, savings, stocks, life insurance, retirement funds, property, and other business investments); and the amount of debts (from home mortgage, car payments, other properties, business, and other debts including debts on medical bills).

The questions about each financial resource consisted of several parts. In general, participants were asked a probe question regarding whether they held a resource. Those who responded "yes" were asked further questions about the person holding the resource (whether it was self, spouse, self and spouse jointly, or self and spouse separately), the value of the resource, as well as any debts or income associated with the resource.

The items subject to imputation are the probe items, the items on the values of financial resources, debts and income. No imputation is required on the "who received" items because there are relatively few missing data, most of which can be deduced through logical edits. The amount of missing data on the items subject to imputation are summarized in Table 1. The nonresponse rates for the probe items are relatively small. In most cases, the missing rate for individual probe items is less than 5 percent, and over 80 percent of the participants provided complete data for all probe items. Nonresponses to the items on the value of assets, and the amounts of debts and income, are more prevalent. For these items, the rate of nonresponse is computed as a percent of those participants who held the particular resource.

Table 1. Percent of missing data for I&A items

		Percent missing		
Variable	Resource	Probe	Value	Inc/Debt
SSRR_	Social security, railroad retirement	0.5	-	17
WAGE_	Job earnings, business for pay	0.6	-	33
SSI_	Supplemental Social Security	1.0	-	14
PUBLIC_	Public Assistance	0.5	-	30
FOOD_	Food Stamps	0.5	-	12
PENSION_	Disability, retirement, annuity	1.7	-	17
IRA_	IRA, or Keogh account	3.9	49	-
HOME_	Own home	1.0	30	30
CAR_	Car, truck, or recreational vehicles	1.2	33	32
BANK_	Checking and savings bank accounts	5.5	73	53
STCK_	Stocks, mutual funds, investments	5.4	75	40
LINS_	Life insurance policy	5.1	53	-
PROP_	Other real estate property	1.9	29	25
BUSI_	Miscellaneous assets: business, farm	3.0	49	52
BONDS_	Bonds	4.8	20	-
OTHER_	Other debt	2.9	-	61

3. Method of Imputation

A form of hot-deck procedure was used to assign values to item nonresponses in the I&A module of the MCBS. There is a large variety of hot-deck imputation methods (Bailar and Bailar,1978; Ford, 1983; and Sande, 1983). Basically, this procedure sorts

respondents and nonrespondents into imputation classes, and the values of the respondents are assigned to the nonrespondents within the same class. The assumption is that after controlling for the classification factors, the distribution of responses for the nonrespondents resembles that of the respondents. An attraction of this procedure is that since all imputed values are taken from the respondents, it cannot assign an impossible value (i.e., a negative income), a feature that does not necessarily hold for all imputation schemes.

Westat has developed an imputation software, WESDECK, that performs hot-deck imputation. WESDECK has two main advantages. The first advantage is that it avoids multiple use of donors of the imputed values. This feature is important because the multiple use of donors leads to an increase in the variance of survey estimates (Kalton and Kish, 1984). For univariate analyses of the variable subject to imputation, the hot-deck imputation scheme is equivalent to weighting adjustments. The same univariate estimates are obtained by the imputation as by a scheme that drops the nonrespondents from the analysis and compensates for them by adding their weights to those of their matched respondents. Therefore the multiple use of donors of the imputed value has the same harmful effect on the precision of survey estimates as a high variability in weights.

Another advantage of WESDECK is that the number of imputation classes can be large. WESDECK recognizes two types of boundaries between imputation classes, "soft" and "hard" boundaries, corresponding to boundaries that can or cannot be crossed when there is a need to search for donors.

In the imputation process, WESDECK begins by sorting respondents and nonrespondents in the order of the imputation classes. Then for each class, the ratio of respondents to nonrespondents is used to determine whether there are sufficient donors within the class. Where possible, imputations are performed within classes; but where necessary, a donor may be located from a neighboring class, across a soft boundary. Donors may be sought outside a class either because there is no donor within the class, or because donors within the class have already been used to donate their values the maximum number of times that has been specified. The search for donors looks forward and backward in the data file until sufficient donors are found or until a hard boundary is encountered. In the imputation step, the exact number of donors is located for each cell that requires imputation and the donors are randomly matched with the recipients.

The donor search feature is attractive because users can use many auxiliary variables to define a large number of imputation classes. A concern with imputation is that it can distort the relationship between variables. In general, if a variable x is used as an auxiliary variable in imputing for y, then the association between y and x is preserved. However, if x is not used as an auxiliary variable in the imputation of y, the association between y and x may be attenuated (Santos, 1981, Kalton and Kasprzyk, 1986). When many auxiliary variables are used to define imputation classes, more relationships can be preserved.

4. Imputation of the Probe items

Using WESDECK, the probe items in the I&A module were imputed in sequence following the order of the variables listed in Table 1. The items were arranged such that items with less amount of missing data were imputed first, and related items were imputed in succession. For example, the first item for imputation is the probe item about income from social security, the item with the smallest amount of missing data. This was followed by other related income items such as employment income, supplemental social security income, etc.

The auxiliary variables used to define imputation classes are all the probe items, except for the item being imputed. The classification of the probe items created classes with specific patterns of financial profile. Using a complete classification of these variables, the underlying model used in the imputation contained all the main effects and all interactions of the auxiliary variables (see Kalton and Kasprzyk, 1986).

For example, in order to impute for the probe SSRR_PRB, social security and railroad retirement income, the other probe items down the list WAGE_PRB through BUSI_PRB were classified to form imputation classes. The 13 auxiliary variables, each with 3 categories ("yes", "no", or "missing"), formed a total of over 1,400 imputation classes. The first two variables, WAGE_PRB and SSI_PRB, were used to form hard boundaries classes such that no donors were sought across these classes. In principle, people should not be receiving social security income, and payment for employment, and supplement social security income together. However, combinations of any two components of these incomes are possible. The other variables formed soft boundary classes, the boundaries of which could be crossed when necessary. Donors within imputation classes were used to a maximum of three times before WESDECK sought donors outside a class.

The imputation of the other probe items followed the same strategy. To impute for the item WAGE_PRB, the imputation classes were formed by classifying SSRR_PRB and SSI_PRB through BUSI_PRB. This time, there were only two categories in SSRR_PRB because the missing data have already been assigned a value. This sequential approach has the attraction of preserving covariance.

The probe items on assets were imputed after the imputations for the probe items on income and the amount of income. The imputation classes in these imputations were created by classifying the other asset probes, and income decile classes. The amount of income from the sources social security through pension income were summed to provide total income. As asset holding patterns are known to be related to total income, it was necessary to include the level of income in the imputation classes.

5. Imputation of the Dollar Amount Items

The items on the value of assets, and the amounts of income and debt, were imputed using the same hot-deck procedure. For these items, data may be missing because a person answered "yes" to the probe item but failed to answer the subsequent question about the amount, or the person did not answer the probe item and his/her response was imputed to be a "yes" response. In the latter case, the person was never asked the item.

The auxiliary variables used to form imputation classes were carefully chosen to encompass the variables that were significant predictors of the dollar amount subject to imputation and the person's propensity to respond. These variables included: the person's age, gender, race/ethnicity, education level, marital status, family size, region, whether he/she had Medicaid, other public medical plan, other private coverage, and whether he/she is a proxy respondent or reporting for self or spouse. The association between these variables and the dollar values subject to imputation were examined through a series of cross-tabulations and tests of associations. Depending on the strength of the associations, different auxiliary variables were selected for the imputation of the different items. The imputation classes for each item are discussed in the technical report to HCFA (Judkins, Winglee, and Ryaboy, 1993).

For example, in imputing the item on the amount of social security income, the auxiliary variables used to form imputation classes were: status (spouse in household and both received social security income, spouse in household and income for either self or spouse alone, and no spouse and income for self alone), Round 1 income (above or below $25,000, or missing), receipt of public assistance income (yes, no), self or proxy respondent, held private health insurance (yes, no), had Medicaid (yes, no), region (east, west, south, central), age group (65 years and older, less than 65), educational level (less than high school, high school, college), race/ethnicity (black, Hispanic, other), and gender (male, female). The first three variables: status, Round 1 income, and public assistance income were used to define classes with hard boundaries. The expectation is that people who received public assistance income would have a small social security income. The other variables defined classes with soft boundaries that could be ignored when necessary. Donors within classes were used to a maximum of three times before WESDECK located additional donors from neighboring classes.

In general, the maximum number of donations per donor was set at three for most items. However, for some assets items with a large amount of missing data, the number of donations had to be increased to four times. For example, to impute the items on the value of stocks and the amount of savings in banks, items with nonresponse rates of almost 75 percent, donors could be used up to four times.

For the assets items that generate income, or are associated with debts, (i.e., home value and home mortgage; stock values and dividends; and business value, liabilities and income), the sets of related items were imputed together. A person with missing data on all components of the set of related items were assigned the values from a matched donor with observed data on all components. For cases with missing data on one item of the set, (e.g. missing the stock value but reported the amount of dividend), donors were used to provide a ratio of stock value over dividend. This ratio was multiplied by the reported dividend to attain a value for the missing response. Similarly, for cases with missing dividend but reported stock value, the same strategy was applied to impute the missing component.

6. Effects of Imputation on Estimates of Total Income and Total Assets

For the I&A module of the MCBS, data were collected for individual financial resources, and imputations were conducted for each item separately. However, the ultimate goal of the I&A supplement is to produce estimates of total income and total assets for Medicare beneficiaries. Without imputation, the estimation of these totals would have to depend on cases with complete responses on all components of the totals, and cases with partially complete data would be discarded. This approach is wasteful of data. In the case of MCBS, less than a third of all sampled persons responded to all items regarding the values of income and assets.

The characteristics of the complete respondents as compared with those of the total sample are shown in Table 2. When compared with the total sample, a higher percentage of the complete respondents had less education, larger families, lower reported income in Round 1, received Medicaid, and were younger than 65 years of age. It is possible that for this particular survey, people with limited financial resources were more capable of responding to the data items regarding their finances because there were fewer items to report. People with many sources of income and assets had to respond to many items and therefore were more likely to have omitted one or more of the items.

Table 2. Percent of complete respondents and all sampled persons with characteristics

Characteristics		Percent based on: Complete respondents	Total sample
Education	< High school	56	47
	High school	26	30
	College	16	22
Family size	One	31	31
	Two	38	49
	Three +	31	21
Round 1 income	$25,000 +	9	21
	< $25,000	89	75
Medicaid	Yes	30	14
	No	70	86
Age	< 64 yrs	32	18
	64 yrs +	68	82
Gender	Male	47	44
	Female	53	56
Sample size		2,878	10,066

Since the characteristics of the complete respondents are different from those of the full sample, estimates of total income and total assets based on the complete respondents alone are likely to be biased. Table 3 shows the percentages of people in different income categories as estimated using data from the complete respondents and data with imputation. The income categories are constructed using the imputed data to define the quintile cutoff points. With the imputed data, the estimated percentage of sampled persons with a total income less than $6,000 is 20 percent, the corresponding estimate from the complete respondents is 35 percent. Similarly for total assets, the imputed data estimates that about 20 percent of sampled people had less than $1,200 worth of assets. The corresponding estimate from the complete respondents is 48 percent. Given the characteristics of the complete respondents, it is fairly evident that the statistics based on the complete respondents overestimate the proportion of people in the low income and low assets categories.

Table 3. Estimated percentage by total income and total asset category

Category	Percent based on: Complete respondents	Imputed data
Total income		
0 - $6,000	35	20
$6,000 - $10,000	25	20
$10,000 - $15,000	17	20
$15,000 - $25,000	12	20
$25,000 +	11	20
Total assets		
0 - $1,200	48	20
$1,200 - $33,000	20	20
$33,000 - $78,000	14	20
$78,000 - $160,000	10	20
$160,000 +	8	20

Tabulations of total income and total assets categories with background characteristics variables have shown that the imputation conducted in this study is effective in preserving the relationships between these variables. Table 4 a and b shows the coefficients of association between total income and total assets with selected variables of background characteristics. The values shown on the table are the Cramer's V, which has an attainable upper bound of 1 and a range of $-1 \leq V \leq 1$. Since the background variables are of different dimensions, the Cramer's V is considered a better choice than other measures of association (Freeman, 1987). In most cases, the level of association for the reported data and the imputed data are comparable suggesting that the imputation for the I&A module have largely preserved the bivariate relationships.

In conclusion, for the I&A module of the MCBS, estimates of total income and total assets based only on complete respondents are biased. Hot-deck imputation, WESDECK, was conducted to help reduce the risk of bias in survey estimates. The advantages of WESDECK are: (1) it controls the multiple use of donors, (2) it is flexible in expanding the donor pool, and (3) the impute values are always plausible values. Investigations show that the bivariate relationships between total income and total assets and selected background variables are largely preserved.

Table 4a. Relationship between total income and background variables

Variable	Cramer's V based on:	
	Complete respondents	Imputed data
Total asset	0.37	0.35
Education level	0.22	0.23
Gender	0.19	0.18
Age	0.15	0.15
Race	0.14	0.14
Spouse in household	0.56	0.53
Medicaid	0.45	0.44
Private insurance	0.48	0.46
Received home help	0.12	0.12
Proxy respondent	0.30	0.28
Census division	0.14	0.11

Table 4b: Relationship between total assets and background variables

Variable	Cramer's V based on:	
	Complete respondents	Imputed data
Education level	0.21	0.22
Gender	0.14	0.10
Age	0.16	0.18
Race	0.13	0.15
Spouse in household	0.47	0.42
Medicaid	0.44	0.49
Private insurance	0.47	0.49
Received home help	0.16	0.18
Proxy respondent	0.32	0.31
Census division	0.15	0.12

ACKNOWLEDGMENTS:

The authors would like to thank Dr. J. Michael Brick of Westat for his reviews and comments. This work was supported by the Office of the Actuary within HCFA.

REFERENCES:

Bailar, B.A., and Bailar, J.C.,III (1978). Comparison of two procedures for imputing missing survey values. In *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Ford, B.L. (1983). An Overview of Hot-deck Procedures. In *Incomplete Data in Sample Surveys. Vol. 3: Proceedings of a Symposium.* Madow, W.G. and Olkin, I. (eds.), New York: Academic Press.

Freeman, D.H., Jr. (1987). *Applied Categorical Data Analysis.* Marcel Dekker, Inc.

Judkins, D., Winglee, M.W., Ryaboy, L. (1993) MCBS: Report on Imputation for the First Income and Assets Supplement.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics - Theory and Methods*, 19(16), 1919-1939.

Kalton, G., and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, 1-16.

Sande, I.G. (1983). Hot-deck Imputation Procedures. In *Incomplete Data in Sample Surveys. Vol. 3: Proceedings of a Symposium.* Madow, W.G. and Olkin, I. (eds.), New York: Academic Press.

Santos, R.L. (1981), *"Effects of Imputation on regression coefficients, "Proceedings of the section on Survey Research Methods*, American Statistical Association, 140-145