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

Thomas N. Herzog, Department of Housing and Urban Development

In the first part of this work described on the preceding pages, we discussed a two-stage imputation protocol for predicting individual OASDI benefit amounts. In Part II, we will discuss an implementation of this procedure and then compare the results to those of the CPS hot-deck. This paper is extracted from a larger work 1/ (Herzog [1980]) which contains a more detailed description as well as additional tables.

MULTIPLE IMPUTATION OF OASDI BENEFITS

In this section, we describe a procedure used to generate 100 imputed benefit amounts (possibly $0) for each of the 59 individuals under consideration who had OASDI benefits allocated by the March 1973 CPS. The reader will recall that of the 129 "nonrespondents" who had at least one CPS income item allocated, 59 had OASDI benefits allocated. We also present the results of an implementation of this multiple imputation scheme.

The Prediction of Recipiency Status.--We begin by describing the procedure used to predict whether or not an individual was an OASDI recipient. The process employed consists of two basic steps carried out independently 100 times to produce 100 distinct sets of imputed values. These steps are:

1. The generation of 100 estimates for each of two sets of parameters. Each set of parameters is drawn from the posterior distribution of the parameters of one of the two final models: the first for those at least 72 years of age and the second for those 62-71 years of age. In this generation process then, we are just drawing values from (or simulating) the posterior distribution of the parameters, given the observed respondent values.
2. The generation, for each of our 59 individuals, of 100 "imputed" benefit statuses using the 72+ or 62-71 model of step 1, each value corresponding to a drawn parameter in step 1. Thus, each such "imputed" value will indicate whether or not an individual nonrespondent is an OASDI recipient.

The first step is carried out by assuming that (for each of the two models) the set of (independent) parameters may be considered to be a random vector having a multivariate normal distribution. 2/ For each model, the mean of the distribution is the vector of parameter estimates of the model. We used the procedure described in Appendix B of Herzog [1980] to estimate the required variance-covariance matrix. We then adjust this estimator so that the effective sample size was approximately that of the full 1980 CPS.

The second step of the procedure is carried out by first calculating the probability of being a recipient for each income-benefit-age cell. We then generated the appropriate number of uniform random numbers (18 for the group over 72 years of age and 41 for the 62-71 year-old group) and counted the number of recipients in each group. For each of the two models, we discarded all sets of imputed values for which the number of imputed recipients was outside a symmetric 90 percent confidence interval centered at the number of recipients expected under the model. Since the number of imputed recipients was necessarily an integer value, we employed the customary 3/ randomization scheme at the end values to ensure that we actually had a 90 percent confidence interval. In those instances in which the expected number of recipients was close to 100 percent of the number of recipients and it was not possible to construct a symmetric 90 percent confidence interval about the mean, we used the (shorter) symmetric confidence interval whose upper end was 100 percent of the number of recipients. Again, a randomization scheme was employed at the lower end of the interval. The above scheme was employed to avoid the prediction of values substantially different from those expected, while preserving the expected value and symmetry of the prediction process.

The Prediction of Individual OASDI Benefit Amounts.--We next describe the procedure used to predict OASDI benefit amounts for those individuals predicted to be OASDI recipients. Again, the process employed consists of two basic steps carried out independently 100 times to produce 100 sets of imputed values. These steps are:

1. The generation of 100 estimates for each of the 28 coefficients of the second combined regression model described in Part I. Each set of parameters is drawn from the posterior distribution of the parameters of the basic model. Here again we are merely drawing values from (or simulating) the posterior distribution of the parameters, given the observed recipient respondent values.
2. The generation, for each of the 100 regression models of step (1), of an "imputed" benefit amount for each of the individuals predicted (by the log-linear model) to be recipients.

We now present a detailed description of this two-step procedure.

The first step is carried out by assuming that the set of 28 regression coefficients of the basic model may be considered to be a random vector having a multivariate normal distribution. The mean vector of this distribution is simply the vector of estimated coefficients of the basic combined model. The variance-covariance matrix, V, is

$$V = 0.25\sigma^2 (X'X)^{-1}$$

where σ^2 is the residual mean square from the basic model and X is the 783 by 28 matrix of observations. The factor of 0.25 was employed because the effective sample size considered here was roughly one-fourth that of the full 1980 CPS.

Essentially, the generation of each set of 28 regression coefficients is accomplished via a 28-stage sequential scheme. Since estimates of the first coefficient's mean, μ_1, and standard deviation, σ_1, are now readily available, (from the basic model and the variance-covariance matrix, respectively) we just randomly draw a value from the normal distribution having mean, μ_1, and standard deviation, σ_1. This provides an

estimate of the first regression coefficient. Using the procedure discussed in Chapter 2 of Anderson [1958] to compute the conditional mean and variance of the n-th estimated coefficient given the first n-1 (where $2 \leq n \leq 28$), we are then able to obtain a complete vector of 28 estimated coefficients.

The second step of the procedure is carried out by first calculating the (mean) value predicted by the regression model for each of the individuals predicted to be recipients by the log-linear model. Each of these values is considered to be the mean of a normal distribution; while the residual mean square of the original regression model is employed as the variance. The value assigned to each of the individuals considered is obtained by randomly drawing a value from the appropriate normal distribution, truncated by removing the 5 percent tail at each end of the distribution. This truncation scheme was employed for the same reasons as given in the previous section.

COMPARING OUR RESULTS TO THOSE OF CPS HOT-DECK

In the table which appears at the end of this section, we exhibit the basic data of our larger paper. The table consists of two parts--one for those at least 72 years of age, the other for those 62-71.

Since our two-stage (or Rubin) model is attempting to predict administrative values while the CPS hot-deck attempts to predict the answers that nonrespondents would have given had they responded, we adjusted the hot-deck so that it too predicted administrative values. Using Rubin's scheme we generated 100 imputed values for each missing item whereas we only generated 2 sets of imputed values using the modified hot-deck. We were then able to compare the results of both procedures to the actual administrative values.

The administrative values, the two "modified" hot-deck values, the first two Rubin values, and the mean of all 100 Rubin values are displayed in the table for each individual nonrespondent. In Herzog [1980],we constructed a number of tables to enable us to compare the results of the two imputation procedures to the actual administrative values. Here we will only mention a few of the highlights.

Using the data at the bottom of both parts of the attached table, we compared the average imputed value under each imputation procedure to the corresponding average administrative value. We did this separately for each age group. For the eighteen individuals at least 72 years of age, the average administrative value of $1,832 was higher than both the average hot-deck value of $1,579 (over both sets of imputations) and the average (over all 100 sets of) Rubin values of $1,697. The principal reason that the mean Rubin value was lower than the average administrative value was that only 89.56 percent of the Rubin values were predicted to be non-zero whereas the true proportion was 17/18 or 94.44 percent.

For the forty-one individuals ages 62-71, the average administrative value of $1,450 was lower than the average hot-deck value of $1,547 but slightly higher than the average Rubin value of $1,417. Again, a large part of the difference between the Rubin and administrative values may be explained by the corresponding percents of recipiency: 32/41=77.05 percent for the actual values versus 74.39 percent for the Rubin values.

The mean absolute deviation of $739 resulting from the use of the mean of all 100 Rubin values for all ages was lower than the mean absolute deviation of $832 resulting from the use of the mean of the two adjusted hot-deck values. Most of this difference is attributable to the data on those at least 72 years of age. In terms of the square roots of the mean squared deviation the verdict was $1,039 for the hot-deck to $913 for Rubin--Rubin doing somewhat better here too. The use of the mean of all 100 Rubin values rather than just the first two Rubin values, also produced lower mean deviations for the combined group (i.e., all those 62+) in both metrics.

We next determined on a case by case basis whether the mean of the two adjusted hot-deck values was closer to the administrative value than (1) the mean of the first two Rubin values or (2) the mean of all 100 Rubin values. We found that for those at least 72 years of age, the mean of 100 Rubin values was closer in 12 cases out of 18. Considering all 59 cases, we found that the mean of 100 Rubin values was closer in 34 instances or 57.6 percent of the time.

We next examined the standard errors of the mean benefit amounts for various collections of imputed values. Our best estimate of the standard error (i.e., that based on all 100 Rubin values) of the mean benefit amount (for all ages) is $202. This compares to a value of $128 obtained by taking the arithmetic average of the standard errors of the mean of each of the 100 individual sets of imputed values. The last average only includes the within-component of the variance (i.e., within each set of 59 imputed values) and does not include the variance between components. Our best estimate, then, is that on the average, the use of only the within-component of the variance results in an estimated standard error that is only about 64 percent of the actual standard error. Since (in the absence of a replicated design in which the imputation process is carried out independently across replicates) the between-component of the variance can only be estimated if there are at least two values imputed for each missing item, we strongly recommend that those imputing missing data items in complex sample surveys impute at least two values for each missing item.

Unfortunately, even the imputation of two values for each missing item may not be enough if there is a lot of variation from one set of imputations to the next. For example, using only the first two sets of Rubin imputed values we obtained an estimated standard error of only $129 compared to a value of $202 for all 100 sets. The reason for the lack of stability of the estimates here may be that we have a very small sample which is highly sensitive to the prediction of beneficiary status.

We next considered the standard errors of the two sets of values formed by appending the administrative values of the 1,069 individuals (who had reported OASDI benefits, possibly for $0) to each set of imputed hot-deck values. Because only slightly more than 5 percent of those under consideration had imputed OASDI benefit amounts, there is not much difference among the estimated standard errors of the expanded dataset. In

particular, the estimated within-component of the standard error based on the two sets of hot-deck values is virtually equal to the corresponding estimated standard error for both variance components (i.e., $27.06 versus $27.08). This is because the hot-deck procedure is not performed independently across the two sets of imputed values. Consequently, the between-component of the variance is practically zero.

Finally, we considered the standard errors of the 100 sets of values formed by appending the administrative values of the 1,069 individuals (who had reported OASDI benefits, possibly for $0) to each of the 100 sets of imputed Rubin values. Again, the range of values is quite narrow. The estimated within-component of the standard error of $27.30 based on all 100 Rubin values is quite close to the administrative estimate of $27.18. The total estimated standard error (including the between component of the standard error) is only $28.49. Again, this difference is relatively small because only a small percentage of values had to be imputed. Nevertheless, even in this case, the omission of the between-component (i.e., that due to the imputation process) would result in an almost 5 percent underestimate of the standard error.

FOOTNOTES

1/This may be obtained by writing to the author at Room 6280, HUD, 451 7th Street, S. W., Washington, D. C. 20410.

2/Although the natural conjugate family of prior distributions for the multinomial likelihood's function is the Dirichlet, we chose to use a multinormal prior because (1) we felt it would be too difficult to estimate the parameters of the Dirichlet distribution and (2) we thought that the multinormal would not give particularly unreasonable results.

3/Randomization procedures of the type employed here are discussed in Section 20.22 of Kendall and Stuart [1979].

REFERENCES

[1] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, New York, 1958.

[2] Aziz, F., Kilss, B., and Scheuren, F. 1973 Current Population Survey--Administrative Record Exact Match File Codebook, Report No. 4, Studies from Interagency Data Linkages, Social Security Administration, Washington, D. C., 1978.

[3] Coder, J., "Income Data Collection and Processing for the March Income Supplement to the Current Population Survey," Proceedings of the Survey of Income and Program Participation Data Processing Workshop, DHEW, 1978.

[4] Herzog, T. N., Multiple Imputation of Individual Social Security Benefit Amounts, U. S. Department of Housing and Urban Development, 1980.

[5] Kendall, M. and Stuart, A., The Advanced Theory of Statistics, 4th Edition, Vol. 2, Macmillan, New York, 1979.

[6] Knuth, D. E., The Art of Computer Programming, Vol. 2, Seminumerical Algorithms, Addison-Wesley, Reading, Mass., 1969.

[7] Lancaster, C., Determining Social Security recipiency for imputation of nonresponse in the CPS, Social Security Administration Memorandum SRV-53, Washington, D. C., January 26, 1979.

[8] Rubin, D. B., "Formalizing Subjective Notions About the Effect of Non-Respondents in Sample Surveys," The Journal of the American Statistical Association, 72, 359, 538-543, 1977.

[9] Rubin, D. B., "Multiple Imputations in Sample Surveys -- A Phenomenological Bayesian Approach to Nonresponse," Imputation and Editing of Faulty or Missing Survey Data, U. S. Department of Commerce, Social Security Administration, Washington, D. C., 1978.

[10] Rubin, D. B., (1979), "Illustrating the Use of Multiple Imputations to Handle Nonresponse in Sample Surveys," Proceedings of the 1979 International Statistics Institute, Manila, to appear.

TABLE - PART I

Multiple Imputations for Those With Allocated
OASDI Benefits and Over 72 Years of Age

		HOTDECK VALUES		IMPUTATIONS		COLLECTION OF IMPUTED VALUES		
NONRES NUMBER	ADMIN VALUE	FIRST VALUE	SECOND VALUE	FIRST VALUE	SECOND VALUE	MEAN OF ALL 100	PERCENT OF RECIPIENCY	STANDARD ERROR
1	1489	2085	2257	1400.50	1181.20	1555.01	99	420.56
2	2103	609	645	1650.17	1416.98	1719.53	95	596.43
3	2027	1968	2142	974.56	1019.94	1348.02	100	338.31
4	993	2059	2375	1858.02	2223.14	1596.13	96	516.88
5	0	2161	2025	3359.50	2656.24	2496.81	81	1444.67
6	2199	2296	2432	.00	1270.47	851.61	68	636.72
7	2487	1968	1839	2276.47	.00	1796.86	77	1120.00
8	2375	2271	2363	1885.66	.00	1949.05	98	560.80
9	1419	609	666	2781.50	3343.01	1773.27	72	1257.68
10	1854	0	2085	1692.57	2001.36	1938.99	100	468.77
11	2161	1603	1864	1497.99	2771.64	1873.20	98	566.98
12	2957	605	799	2255.92	.00	1725.20	73	1169.76
13	1187	2774	2348	1102.49	1778.75	1591.05	100	373.26
14	1835	791	797	1859.52	1585.78	1811.95	82	1000.65
15	887	1344	1609	2090.98	1316.64	1546.06	95	543.92
16	2375	1864	1603	2803.75	2391.59	1949.52	99	527.61
17	2384	0	2354	1688.80	1699.49	1756.03	84	920.13
18	2239	887	735	1164.78	1073.65	1260.98	95	429.37
MEAN:	1831.72	1438.56	1718.78	1796.84	1540.55	1696.63	89.56	789.26

NOTE: EACH OF THE FIRST SIX MEANS IS THE ARITHMETIC AVERAGE OF THE CORRESPONDING COLUMN OF BENEFIT AMOUNTS, WHILE THE SEVENTH IS THE ARITHMETIC AVERAGE OF THE PERCENT OF RECIPIENCY STATUS. THE MEAN OF THE LAST COLUMN IS THE SQUARE ROOT OF THE ARITHMETIC AVERAGE OF THE SQUARES OF THE STANDARD ERRORS(I.E., THE VARIANCES).

TABLE - PART II

Multiple Imputations for Those With Allocated
OASDI Benefits and 62 - 71 Years of Age

		HOTDECK VALUES		IMPUTATIONS		COLLECTION OF IMPUTED VALUES		
NONRES NUMBER	ADMIN VALUE	FIRST VALUE	SECOND VALUE	FIRST VALUE	SECOND VALUE	MEAN OF ALL 100	PERCENT OF RECIPIENCY	STANDARD ERROR
1	2027	1845	2613	.00	2438.27	1396.03	72	967.22
2	1047	1352	1548	.00	.00	834.67	51	866.56
3	2095	2613	2239	1692.40	1697.60	768.74	67	604.12
4	2112	2703	2685	1594.82	1753.41	1687.87	87	815.79
5	1947	2314	0	2142.06	1944.56	1065.70	60	929.66
6	2242	1870	815	2221.26	1941.71	1951.93	89	860.24
7	2341	2161	2025	.00	2711.00	1710.57	79	1025.43
8	2631	1877	1451	2852.47	2252.23	1939.90	84	1007.03
9	2597	0	2314	2530.61	1511.14	2144.34	85	1066.42
10	0	0	0	.00	2696.77	1737.49	73	1209.99
11	0	0	0	884.12	916.98	701.76	66	547.03
12	561	735	887	1355.60	1092.85	1463.35	78	898.53
13	2662	1778	666	1095.82	735.48	676.30	73	469.93
14	992	1546	993	1449.68	2090.34	1848.20	87	913.23
15	2204	1274	1035	1612.41	1536.66	1562.36	83	844.44
16	0	0	0	821.36	.00	415.71	63	346.64
17	1928	1378	1010	1452.45	2691.01	1934.42	90	811.03
18	2392	1718	2660	1319.26	.00	1101.59	61	957.19
19	2217	2018	1994	2491.58	1399.53	1868.86	81	1057.10
20	2685	2085	1920	2937.78	2679.56	2021.47	79	1184.73
21	1491	1279	1102	.00	1008.68	897.39	55	875.05
22	0	2604	0	1535.00	.00	712.36	64	577.65
23	1056	2058	1757	1411.18	1743.09	787.13	51	815.11
24	2343	2027	965	1471.62	1468.23	1477.34	80	861.49
25	0	2651	609	1781.91	1872.94	1437.02	87	686.19
26	1930	3103	2161	1845.74	1476.57	892.60	67	708.26
27	0	0	0	.00	1006.59	824.31	73	558.65
28	2169	2210	2314	2464.07	2355.40	1830.42	79	1065.30
29	2597	965	2027	1850.93	2270.65	1690.76	78	1013.98
30	848	1718	2315	1693.63	.00	1917.46	85	945.39
31	0	2411	2703	2476.77	2592.03	2157.13	89	941.29
32	2450	1438	1313	2073.52	1971.63	1945.88	84	1034.23
33	0	1568	1597	1819.68	2263.29	2124.97	81	1202.97
34	887	1031	1687	2724.71	1572.47	1713.04	81	951.13
35	410	1641	1990	.00	1027.45	765.01	62	657.56
36	0	2631	2597	2110.41	1588.93	984.78	57	924.42
37	1016	0	0	.00	919.83	547.86	70	398.13
38	1445	2199	2203	2110.66	1520.73	1118.53	57	1030.91
39	1774	2090	2239	1628.41	2537.47	1633.98	74	1090.76
40	2505	1313	1438	.00	2683.38	1828.96	84	935.59
41	1846	2523	2218	.00	2882.89	1968.68	84	1047.04
MEAN:	1449.93	1627.49	1465.61	1401.27	1630.52	1416.75	74.39	895.82

NOTE: EACH OF THE FIRST SIX MEANS IS THE ARITHMETIC AVERAGE OF THE CORRESPONDING COLUMN OF BENEFIT AMOUNTS, WHILE THE SEVENTH IS THE ARITHMETIC AVERAGE OF THE PERCENT OF RECIPIENCY STATUS. THE MEAN OF THE LAST COLUMN IS THE SQUARE ROOT OF THE ARITHMETIC AVERAGE OF THE SQUARES OF THE STANDARD ERRORS(I.E., THE VARIANCES).