WEIGHTING TO ADJUST FOR PARTIAL NONRESPONSE

Lawrence R. Ernst, U.S. Bureau of the Census

I. Introduction

Partial nonresponse is one of the more common sources of nonsampling error in sample surveys. Several procedures have been devised to adjust for it. In this paper we will be concerned with one particular survey, the March Current Population Survey of the Bureau of the Census, and one type of partial noninterview, in which all of the income items on the supplement page need imputation. It will not be assumed that any of the other questions on this page have been answered either, but it will be assumed in general that responses are available to all of the questions on the "control card" and basic labor force page. It has been found from previous studies that the distribution of values for the respondents and nonrespondents on the income questions is quite different in character, nonrespondents having a much higher mean total money income for example, and hence some procedure to impute for these missing items is necessary [3].

We will focus on the implementation for this particular situation , one of the more common nonresponse adjustment procedures, weighting. We will also compare this procedure with the method currently being used by the Census Bureau for this adjustment, the "hot deck." The latter procedure consists of matching records with missing income items to complete records with similar characteristics (same age group, race, sex, education level, etc.). When a match is made, the responses from the completed record on those items which were omitted on the partial nonrespondent's record are duplicated on the latter record, thereby creating a completed record for the partial nonrespondent. Although there are several advantages to this procedure, which will be discussed in section 5, there are also some drawbacks. Among the drawbacks is the fact that "hot deck" imputation procedures in general result in higher variance than if no imputation had been done [2, pp. 139-141].

Consequently it was decided to explore the alternative of using a weighting scheme for imputation of the missing income items. In this type of procedure an estimator for the sample mean for a given item is found by first partitioning the survey population into a number of subclasses, called either weighting classes, strata, or cells, such that the members of a given cell all share certain characteristics as, for example, those mentioned in the previous paragraph. In a sample of size n, if n_h, m_h, and y_h denote respectively the number of respondents, number of nonrespondents, and the sample mean for the respondents in the h-th stratum, then $w_h = (m_h + n_h)/n$ is the weight associated with this stratum and $\bar{y} = \Sigma\, w_h \bar{y}_h$, where the summation is over all the strata, is taken as the estimator of the sample mean. The use of this estimator is motivated by the belief that if the characteristics used to form the cells are well chosen, then the true mean for the nonrespondents in a given cell would be approximately equal to the true mean for the respondents, and this estimator would consequently be less biased than would the ordinary sample mean unadjusted for nonresponse.

In practice, a key problem with the use of weighting procedures is the determination of which characteristics to use in forming the cells in order to obtain an estimator with minimum mean square error (MSE). Since only the variance component of the MSE can be estimated from a sample alone, it was decided to concentrate first on obtaining a set of cells which would minimize the variance of the estimated mean and later determine how satisfactorily the weighting scheme lowered the bias. In this connection it was necessary to obtain a formula for $\bar{y}$ as an estimator of the population mean. This is presented in section 2. The procedure followed to determine a set of cells which minimizes variance using this formula is given in section 3, where it is also explained why it was necessary to revise significantly our weighting procedure to take bias into account, and how this was done. In section 4 we discuss the testing of our final weighting scheme and compare it with two different "hot deck" procedures. Finally, our conclusions are presented in section 5.

2. A Variance Formula for Weighting

2.1 Notation

Suppose that we have a simple random sample of n units chosen from a population of N units, which are classified into L weighting classes. Consider any positive integer $h \le L$. Let n_h, m_h denote the number of respondents and nonrespondents respectively in the h-th weighting class, and take N_h, M_h respectively to be the total number of units in the response and nonresponse population of this stratum. Define $\bar{n}_h = nN_h/N$, $\bar{m}_h = nM_h/N$. Denote by y_{hi} the value for the i-th unit in the stratum of which the first n_h units will be assumed to be the respondents.

Let

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$$

$$\sigma_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h} .$$

i.e., $\bar{y}_h$, $\bar{Y}_h$, σ_h^2 are respectively the sample mean, the population mean and the population variance for the respondent population of the h-th weighting class. Define $\bar{y}$, the adjusted sample mean by

$$\bar{y} = \sum_{h=1}^{L} \left(\frac{m_h + n_h}{n}\right)\bar{y}_h$$

Finally, let $\bar{Y} = E(\bar{y})$ and $\sigma_{\bar{y}}^2 = E(\bar{y} - \bar{Y})^2$.

2.2 The Variance Formula and Explanation of its Terms

It can be shown that an approximate expression for $\sigma_{\bar{y}}^2$ is given by

$$\sigma_{\bar{y}}^2 \doteq \left[\frac{1}{n^2} \Sigma(\bar{m}_h+\bar{n}_h)(\sigma_h^2+(\bar{Y}_h-\bar{Y})^2)\right]$$

$$+ \left[\frac{1}{n^2} \Sigma \frac{\bar{m}_h}{\bar{n}_h}(\bar{m}_h+\bar{n}_h)\sigma_h^2\right]$$

$$+ \left[\frac{1}{n^2} \Sigma \left(\frac{\bar{m}_h}{\bar{n}_h} + \frac{n\bar{m}_h}{(n-\bar{n}_h)\bar{n}_h^2} + \frac{n^2\bar{m}_h^2}{(n-\bar{n}_h)^2\bar{n}_h^2}\right)\sigma_h^2\right] .$$

For i=1, 2, 3, let S_i denote the i-th bracketed term in the above formula. We will explain each of these terms.

We observe that by [1, p. 99] the expression S_1 is, for the special case when in each cell the population mean and variance are the same for the respondents and the nonrespondents, simply the sample variance of the mean of the population for a random sample with replacement of size n . If our choice of weighting cells is a good one, S_1 should be approximately equal to this value.

S_2 arises from the fact that in general a sample does not consist entirely of respondents. Note that the addition of S_2 to S_1 increases the within cell variance component of S_1 but does not effect the between cell component. This is because the sample variance within any cell is determined only by the respondents in that cell while the between cell variance term depends on the total number of sample units in each of the cells, that is, respondents plus nonrespondents.

The presence of S_3 occurs from the fact that

$$E\left(\frac{m_h^2}{n_h}\right) \neq \frac{\bar{m}_h^2}{\bar{n}_h} .$$

2.3 General Procedure for Use of Variance Formula in Determination of Choice Cells

We proceed to present a general procedure for choosing a set of cells for which the value of $\sigma_{\bar{y}}^2$ is nearly minimal. This procedure will be used in the next section to arrive at an initial weighting scheme.

Roughly, the procedure that will be used for cell choice is a multistage one at every stage of which each cell formed in the previous stage is subdivided if and only if the component of $\sigma_{\bar{y}}^2$ for the cells formed by the subdivision is lower than that for the original cell. If at any stage, say the k-th, this selective subdivision does not appreciably decrease the total value of $\sigma_{\bar{y}}^2$, when compared to stage k-1, then we stop and take the cells formed at stage k-1 to be our final choice.

We now explain this method in more detail. A number of variables which are candidates for use in cell division are first selected. In stage 1 we compute separately $\sigma_{\bar{y}}^2$ for the cells determined by each of the variables, and choose a variable which results in a set of cells for which $\sigma_{\bar{y}}^2$ is minimum. If this minimum value of $\sigma_{\bar{y}}^2$ is appreciably lower than the value of $\sigma_{\bar{y}}^2$ for the case when all respondents and nonrespondents are in a single cell we then take for our stage 1 set of cells the set determined by this variable and proceed to stage 2. Otherwise we stop and do not use weighting at all.

In general at stage k of our subdivision, for k > 1, we subdivide the cells formed at stage k-1 for each of the variables not chosen in stages 1 through k-1. For each such variable we subdivide only those cells for which the value of the component of $\sigma_{\bar{y}}^2$, for the cells formed by the subdivision, is lower than that for the original cell. (Clearly, not subdividing a cell is equivalent to subdividing and then collapsing back to the original cell.) We then compute $\sigma_{\bar{y}}^2$ for the set of cells determined by each of these variables using this selective subdivision process, and choose a variable for which $\sigma_{\bar{y}}^2$ is minimum. If this minimum value of $\sigma_{\bar{y}}^2$ is appreciably lower than the value obtained at stage k-1, then the set of cells formed by the selective subdivision using this variable is taken as the stage k set of cells and we proceed to stage k+1. Otherwise we take as the final set of cells the set formed at stage k-1.

3. The Construction of the Weighting Scheme

In our initial construction of a set of weighting classes, using the method just outlined, the 18 variables given in table 1 were used as candidates for cell division. The information concerning these characterstics is in general available for everyone in the survey from either the basic labor force page or the "control card," although, in some cases, the answers had been obtained by a previous imputation.

The program was run using as data the March CPS with the income supplement of 1975-77. The item whose mean was estimated was total money income The program was run through 9 stages.

The results are given in table 2, in which stage 0 indicates the unweighted case. From this table and other information the following problems were noted.

A. The decrease in the variance from the unweighted case to the final stage involving weighting using nine characteristics was only 10.3. The principal reason for this small decrease was the large size of our sample, an average of approximately 110,000 records per survey. Consequently, since the variance component of the MSE was changed such a small amount in absolute terms by this weighting scheme, the usefulness of this procedure was nearly completely dependent on its effectiveness in bias reduction. (It should be noted that in surveys using much smaller samples, variance would be a much more significant component of MSE and using a scheme such as this, which concentrated on variance reduction, would be more reasonable. In fact, our weighting scheme did reduce the variance by 1.7 percent which I believe is significant in relative terms, in view of the fact that the nonresponse rate was only about 5 percent.)

B. It appeared that the characteristics that decreased the variance most were not those that formed cells which best separated people of different income levels. For example, occupation did not enter the weighting scheme until the final two stages and "Activity Last Week" was not present at all. The apparent reason that characteristics such as these were not the best for variance reduction was that some of the cells formed by such characteristics principally contained people with very high incomes and such cells also tended to have both relatively large within-cell variances and high nonresponse ratios, a combination that results in large contributions to the overall variance which are not offset by the smaller contributions of the other cells. However, it would appear intuitively that such characteristics do well at bias reduction.

C. A great deal of collapsing was taking place. For example after four stages only 127 cells existed compared to the 256 cells that there would have been if collapsing had not been allowed. This collapsing was caused by the same problem mentioned in the last paragraph, namely the formation of cells that had large within cell variances coupled with high nonresponse ratios, resulting, in many instances, in a larger contribution to the total variance from the subdivided cell than from the original cell. It seemed, however, that the small increase in variance which would have resulted from not collapsing would have been more than offset by a bias reduction resulting from the formation of cells which better separated people of different income levels.

The above observations led to the conclusion that this weighting scheme, although effective in reducing variance, did not work particularly well on the much more significant bias component of MSE. Consequently it was decided to rerun our program to obtain a new weighting scheme which would be constructed in the opposite manner from the first, that is, now we would find a scheme which was hopefully optimal in reducing bias and then check how well it worked on the relatively insignficant variance. Our program was similar to the previous one except that at each stage, instead of choosing the characteristic which decreased variance the most, we picked the one which resulted in a mean whose difference in absolute value from the mean of the previous stage was maximum. It was hoped that choosing the characteristics in this manner would result in the estimated sample mean converging after several stages to a close approximation to the true mean. This new program was run using only the data from the 1976 surey because of budget constraints. The results are presented in table 3. It was decided to stop at stage 7 since further stages would not have changed the mean significantly.

Comparing the results of this scheme with the previous one we first observed that in the second scheme the decrease in the variance from the beginning to the end was 3.5, significantly different in relative terms from the 10.3 decrease in the first scheme, but not in absolute terms. On the other hand, there was quite a difference in both relative and absolute terms in the effect that the two schemes had on the estimate of the mean. In the first program, weighting resulted in an increase of $89 or 1.5 percent in the mean while in the second program the increase was $171 or 2.9 percent. Because the characteristics chosen and the lack of collapsing resulted in cells which better separated individuals of different income levels, it was felt that the bias was smaller in the second scheme. Consequently, it was this scheme, with a single modification, that we decided to test and compare with the "hot deck", as will be described in the next section. The modification consisted of substituting XVII - "Middle Division of Occupation Groups" for VIII - "Activity Last Week" for those individuals who had worked the previous week.

4. Test of Weighting Procedure and Comparison with the "Hot Deck"

It was decided that the best way of testing both our weighting procedure and the "hot deck" and comparing the two was to use the only recent set of records for which the income of both the respondents and the nonresponents was available from an independent source, namely the "1973 CPS - Internal Revenue Service (IRS) - Social Security Administration (SSA) Exact Match File," which matched the majority of the participants in the CPS with their IRS records.

Only wage and salary earnings were used in this test because the definitions of this item given by IRS and the Census Bureu are quite similar, which is not the case for other types of income. Unfortunately, it had been found in a previous study by Herriot and Spiers [3], which compared the hot deck procedure then in use with IRS figures from the same match file, that even for the respondents the ratio of CPS to IRS earnings was .981, a significant deviation from 1 which resulted from a number of factors. For nonrespondents the ratio was .911 according to that study. In our program we found that the mean earnings of the 65,714 individuals, both respondents and nonrespondents, who were matched to IRS records, was $5,265 by IRS figures, $5,160 using our weighting scheme, and $5,098 by the "hot deck" procedure in use in 1973. Since the ratio of the mean earnings by the weighting scheme to the IRS mean is thus .980 while the corresponding "hot deck" to IRS ratio is .968, it appears that the weighting scheme does not have the significant downward bias in the imputed mean for the nonrespondents that the "hot deck" procedure had. However, it should be noted that comparisons between the results in [3] and our study may be misleading for several reasons. First, in the former paper, mean earnings were computed only for people who had both positive IRS and CPS earnings, while in our paper all matched records were used for that purpose. Furthermore, in [3] a couple filing a joint return was treated as a single unit in computing means and if either person was a nonrespondent then that unit was put in the nonrespondent category since IRS data were not available for the husband and wife individually. It was because of this problem with joint returns that we did not attempt to analyze our data separately for respondents and nonrespondents.

We next used the matched records to perform another test of our weighting scheme which would not be affected by the discrepancy between CPS and IRS earnings figures for respondents and the complications described above caused by joint returns. In this new test we considered only the 17,202 matched records of people who filed individual tax returns. These records consisted of 15,744 respondents and 1,458 nonrespondents on the CPS wage and salary earnings question. We then used the IRS wage figures for the respondents and our weighting scheme to find an imputed mean for all the matched individual taxpayers and a mean for the nonrespondents alone. These means together with the corresponding means obtained from the actual IRS figures are given in table 4. From the table it follows that the bias of the unweighted estimator is -$36, while for the weighted estimator it is -$12, a relative bias reduction of 66.7 percent. Furthermore, for the nonrespondents the ratio of the mean by weighting to the IRS mean is 96.8 percent while for all matched individual taxpayers this ratio is 99.7 percent.

The "hot deck" in use in 1973 [5] is the one in which we have compared our weighting scheme to using the match file data. In 1975 a revision was made of this "hot deck." The new procedure uses vastly more cells than the previous one, nearly two billion versus only about three thousand [6]. We compared our weighting scheme with the new "hot deck" procedure with respect to total income data from the March 1976 CPS. The results are given in table 5. Since no valid data are available for the nonrespondents in that survey we made no attempt to draw any conclusions from this comparision

5. Conclusions

It would appear from the testing involving the match files that our weighting scheme produces smaller bias than the "hot deck" procedure in use in 1973. However, the weighting scheme uses many more cells than that "hot deck" and, therefore, one might either conclude that the weighting scheme worked better in this test because of inherent advantages it possesses over all "hot deck" procedures, or alternatively conclude that this particular weighting scheme worked better than this particular "hot deck" because of the difference in the number of cells. I personnally lean to the latter point of view since bias is the dominant component of mean square error in the particular situation we have been studying, and in general the larger the number of cells, provided they are well chosen, the smaller one might expect the bias to be. If this point of view is accepted one might also then conclude that the revised "hot deck" works best of all since it uses many more cells than our weighting scheme. A weighting scheme would be nearly impossible to use with so many cells because there would be many more cells than the sample size. Although there certainly is some question whether two billion cells are necessary to obtain a good estimate of the mean, the Census Bureau publishes tables of income versus many characteristics and it would seem that, for these tables to be most accurate, as many of these variables as possible should be used in cell formation.

One characteristic of the "hot deck" not shared by weighting is that a complete record is created for each partial nonrespondent. Although this has the advantage of simplicity, it also has disadvantages. An example of a disadvantage of creating complete records for the partial nonrespondents is the problem caused by the need to insure consistency in the answers on such records in order to guarantee that individuals as unlikely as physicians with only a grade school education are not created. In order to avoid such occurrences, some individuals who have failed to answer only the income question on the supplement page have had their responses to the work experience questions and/or job questions deleted and then imputed for. This does not appear to be a desirable situation.

Another problem with the current "hot deck", which is caused by the use of such a large number of cells, is that sometimes a nonrespondent cannot be matched with a respondent on the first try. In such cases, the number of cells is reduced in several stages until a match is found. In the process, completed information is sometimes removed for the reason mentioned in the previous paragraph. Some nonrespondents are finally matched with respondents on very few characteristics, as little as age, sex, and educational attainment in 0.5 percent of the imputations. The several stages of matching necessary also increase the cost of using this procedure.

Overall, though, I believe that the current "hot deck" procedure is preferable to weighting for use in the CPS because of the large sample size and the need for accuracy with respect to so many different relationships. On the other hand, in any survey with a much smaller sample size I believe that weighting would be preferable since in such a survey variance would become more significant, and weighting is more effective in controlling variance than the "hot deck."

REFERENCES

[1] Cochran, W. G. Sampling Techniques. John Wiley and Sons, Inc., New York, 1963.

[2] Hansen, M. H., Hurwitz, W. N., and Madow, W. G. Sample Survey Methods and Theory, Volume II, John Wiley and Sons, Inc., New York, 1953.

[3] Herriot, R. A. and Spiers, E. F. "Measuring the Impact on Income Statistics of Reporting Differences Between the Current Population Survey and Administrative Sources," Proceedings of the Social Statistics Section, American Statistical Association, 1975, pp 147-158.

[4] Thomsen, Ib. "A Note on the Efficiency of Weighting When Analyzing Survey Data," Abreidsnotater, Statistisk Sentralbyra, Oslo, 1973.

[5] Spiers, E.F. and Knott, J. J. "Computer Method to Process Missing Income and Work Experience Information in the Current Population Survey," 1969 Proceedings of the American Statistical Association, Social Statistics Section, pp. 289-297.

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

Table 1. Variables Used as Candidates for Cell Division

	Characteristics	Number of levels
I.	Age	4
II.	Highest Grade Completed	4
III.	Relation to Family Head	4
IV.	Sex and Marital Status of Family Head and Labor Force Status of Wife	4
V.	Own or Rent Residence	3
VI.	Sex	2
VII.	Class of Worker	4
VIII.	Activity Last Week	4
IX.	Region	4
X.	Race and Ethnicity	3
XI.	Hours Worked Last Week	4
XII.	Household Income on "Control Card"	4
XIII.	Number of Children in Family Under 18	4
XIV.	Location of Residence 1	4
XV.	Location of Residence 2	4
XVI.	Broad Division of Occupation Groups	4
XVII.	Middle Division of Occupation Groups	13
XVIII.	Fine Division of Occupation Groups	49

Table 2. Results of First Weighting Scheme Program, March 1975-77 CPS

Stage	Characteristic Yielding Lowest Variance	Mean	Variance
0		$6108	616.7
1	III - Relation to Family Head	6143	611.4
2	II - Highest Grade Completed	6163	610.0
3	XI - Hours Worked Last Week	6196	608.9
4	XII - Sex and Marital Status of Family Head	6196	608.3
5	IX - Region	6198	607.9
6	XII - Household Income on "Control Card"	6197	607.4
7	XV - Location of Residence 2	6198	607.1
8	XVI - Broad Division of Occupation Groups	6198	606.7
9	XVII - Middle Division of Occupation Groups	6197	606.4

Table 3. Results of Second Weighting Scheme Program, March 1976 CPS

Stage	Chosen Characteristics	Mean	Variance
0		$5980	646.6
1	XVI - Broad Division of Occupation Groups	6113	648.6
2	XI - Hours Worked Last Week	6131	645.0
3	I - Age	6139	645.4
4	XV - Location of Residence 2	6147	645.2
5	VIII - Activity Last Week	6155	646.1
6	III - Relation to Family Head	6149	643.5
7	II - Highest Grade Completed	6151	643.1

Table 4. Comparison of Weighting for Imputation with Actual IRS Wage Figures for Matched Taxpayers Filing Nonjoint Individual Returns, March 1973 CPS

Item	Respondents	Nonrespondents	Total
Number of Records	15,744	1,458	17,202
Mean IRS	$ 4,051	$4,475	$ 4,087
Mean by Weighting	4,051	4,334	4,075

Note: Nonrespondents were all persons for whom CPS wages had to be imputed.

Table 5. Comparison of Weighting and New "Hot Deck" Procedure for March 1976 CPS (persons 14 years or older)

Item	Respondents	Nonrespondents	Total
Number of Records	96,726	5,796	102,522
Mean by "Hot Deck"	$ 5,980	$9,748	$ 6,193
Mean by Weighting	5,980	9,005	6,151

Note: Nonrespondents were persons for whom no income information was provided in the survey. Everyone else was treated as a respondent even though in some cases they may have had missing income items which were imputed.