MISSING DATA PROCEDURES: A COMPARATIVE STUDY

Barry L. Ford, Statistical Reporting Service, USDA

PURPOSE

The purpose of this paper is to discuss an investigation of the missing data problem in a list frame survey which has a simple stratified design. A description of the missing data procedures is given, and a general theoretical comparison is made among them. The main thrust of this report is a simulation experiment with these missing data procedures. All of the examples and data are from agricultural surveys by the Statistical Reporting Service (SRS) of the United States Department of Agriculture (USDA).

INTRODUCTION

In the area of survey design, the missing data problem is one of increasing concern. Nonresponse rates of 10% are not unusual for SRS surveys, and there is a fear that these rates may increase.

Why worry about missing data? If there is little difference between the missing data and the reported data in a simple stratified design, the only consequence of missing data is the reduction in sample size. This reduction can easily be offset by increasing the initial sample size. However, in many cases it is probable that the missing data and the reported data re not alike.

A difference between missing and reported data leads to biases in the survey estimates. The size of these biases depends on: 1. the magnitude of the difference between the missing and reported data. 2. the percentage of nonresponse.

Let: p = the percentage of the population which would respond
q = the percentage of the population which would not respond
μ = the population mean
μ_1 = the mean of the part of the population which would respond
μ_2 = the mean of the part of the population which would not respond

Of course, $\mu = p\mu_1 + q\mu_2$. Also, let

$$D = \mu_1 - \mu_2.$$

Then, the relative difference between the data which would be reported and the data which would not be reported is:

$$D' = \frac{\mu_1 - \mu_2}{\mu}.$$

The bias in only using the reported data to estimate μ is:

$$B = \mu_1 - \mu = q(\mu_1 - \mu_2) = qD.$$

Thus, the relationship between B, q and D is linear. Undoubtedly, these potential biases are the real cause of concern about missing data.

Similarly, the relationship between B' (the relative bias), q and D' is also linear:

$$B' = \frac{\mu_1 - \mu_2}{\mu} = q\frac{(\mu_1 - \mu_2)}{\mu} = qD'.$$

The causes of missing data are complex and varied, but the emphasis in any survey should be on eliminating or minimizing the likelihood of missing data before the survey starts. Procedures to estimate for missing data are a stopgap measure -- they are techniques to use after the survey is over when no other alternative is possible. Obviously, no procedure can be as good as not having any missing data. Furthermore, when the percentage of missing data is extremely high, there is probably no procedure that can estimate the missing data efficiently enough to make the survey worthwhile. With moderate and low missing data rates, perhaps some missing data procedures can minimize the bias to a tolerable level.

The six missing data procedures discussed in this investigation are the double sampling ratio procedure, the double sampling regression procedure, and four variations of the hot deck procedure. Some general advantages and disadvantages of each one are outlined.

The Double Sampling Ratio Procedure

Often there is an auxiliary variable associated with each sampling unit. This auxiliary variable may be a variable that is used to stratify the population, an observed variable, or any other additional variable that can be obtained for the whole sample. There should also be a reasonable correlation between the primary variable and the auxiliary variable.

The double sampling ratio design is well-known. In this experiment the first sample is the selected sample, including missing and reported data. Then the second sample is only the reported data. The ratio estimator and its approximate variance for a simple random sample (1, pg. 340) are:

$$\text{(I)}\quad \bar{y}_{Ratio} = \frac{\bar{y}}{\bar{x}}\,\bar{x}'$$

$$\text{(II)}\quad VAR(\bar{y}_{Ratio}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)(2R\,S_yS_x - R^2S_x^2)$$

where:

$\bar{x}'$ = average of the auxiliary variable over the whole sample
$\bar{x}$ = average of the auxiliary variable over the part of the sample that reported data
$\bar{y}$ = average of the primary variable over the part of the sample that reported data
$\bar{X}$ = the average of the auxiliary variable over the whole population
$\bar{Y}$ = the average of the primary variable over the whole population
$R = \frac{\bar{Y}}{\bar{X}}$
S_x^2 = the variance of the auxiliary variable
S_y^2 = the variance of the primary variable

ρ = the correlation between x and y
n' = size of the entire sample
n = size of the sample that reported data
N = size of the population
(Note that the variance was multiplied by the finite population correction factor.)

Although the double sampling ratio estimator is almost always a biased estimator, it is easy to compute even for complex samples. In this report the design is a simple stratified sample so the above formulas are applied in each stratum. Usually S_y^2, S_x^2, ρ, and R are unknown, but their corresponding sample estimates can be substituted into the previous two equations (I and II). As Cochran points out (1, pg. 341), the resulting estimate of variance is not unbiased but appears to be a good approximation.

This ratio estimator makes two assumptions: 1. the initial sample is a random sample 2. the missing data comprise a *random* subsample of the initial sample. This second assumption is probably violated in most surveys; to what degree it is violated depends of course, on the particular situation. One hopes that the ratio estimate and its variance are fairly insensitive to a violation of the second assumption.

In essence the ratio estimator is a linear regression estimator with the intercept assumed to be zero. If the population does not follow the assumption of a linear model, then the ratio estimator (or any regression estimator) becomes a biased estimator. Researchers rarely accept the linear population model as completely realistic, but approximate analytical results and empirical studies show the bias is usually small (3, pg. 23-25; 7, pg. 208-209; 9).

One should remember that in a stratified design there also exists a combined ratio estimator. This estimator is used when the ratio $R = \frac{\bar{Y}}{\bar{X}}$ is equal in all strata. For the data in this study the idea that the ratios in all strata are equal is believed to be false. Thus, a separate ratio estimator is used in each stratum. However, the separate ratio estimator has an inherent danger of accumulating a serious bias across all strata. This accumulation is more likely to be serious when the stratum biases are in the same direction (1, pg. 168-173).

The Double Sampling Regression Procedure

The double sampling regression procedure is also quite common. Like the ratio procedure one has an auxiliary variable in addition to the primary variable. The formulas are (1, pg. 336-339):

(III) $\bar{y}_{Reg} = \bar{y} + b(\bar{x}' - \bar{x})$

(IV) $VAR(\bar{y}_{Reg}) = \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'}$.

We will estimate $VAR(\bar{y}_{Reg})$ with:

$$var(\bar{y}_{Reg}) = \frac{s_{y.x}^2}{n} + \frac{s_y^2 - s_{y.x}^2}{n'}$$

Adjusting $var(\bar{y}_{Reg})$ by a finite population correction factor of $1 - \frac{n}{N}$, one obtains:

$$(V)\quad var'(\bar{y}_{Reg}) = (1 - \frac{n}{N})\left[\frac{s_{y.x}^2}{n} + \frac{s_y^2 - s_{y.x}^2}{n'}\right]$$

as an estimate of the variance of $\bar{y}_{Reg}$ in a finite population where:

$\bar{x}'$, S_y^2, ρ^2, n', n, N are the same as for the ratio estimator

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

$$s_{y.x}^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n} (y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2\right].$$

As noted for the ratio estimator, the assumption of a linear population model is ignored. The estimators may then be biased, but one empirical study (3, pg. 22-25) lends support to the conjecture that these biases are small. Also, in the stratified data of this study a separate regression estimator is used as opposed to a combined regression estimator. Therefore, one again has the danger of accumulating a large bias across all strata, especially when each stratum bias is in the same direction (1, pg. 200-202).

Hot Deck Procedure

The hot deck is probably the most common missing data procedure in use at the present time, especially in complex surveys. The Bureau of the Census, the Statistical Reporting Service, Statistics Canada, and many otherscurrently employ this missing data procedure. In spite of this wide use little testing or theoretical analysis on the impact of the hot deck procedure has ever been published (8). This situation is really not too surprising because, although the hot deck procedure is intuitively satisfying and extremely flexible; its flexibility and lack of a strong theoretical development deter anything but broad generalizations of its effects.

A basic outline of the hot deck procedure is:

1: Separate the sample into I classes based on k variables.
2: If an item is missing in a certain class, then randomly select a reported item from the same class.
3: Substitute the selected item for the missing item.
4: Compute sample estimates as if there are no missing values.

The most obvious consequence of this procedure one would challenge is the fact that the estimated variances of the sample mean are almost certainly biased below their actual values. Step

4 above allows one to use a sample size that includes the number of missing values. Thus, the loss of information due to missing data is not reflected in the sampling errors. For example, suppose two surveys cover the same population and have the same sample size. Furthermore, one survey has 30% missing data, and the other survey has no missing data. After applying the hot deck procedure, the errors of the estimates of these two surveys would probably be about equal. *Obviously, the standard errors from the survey that used the hot deck procedure should reflect the fact that 30% of the information is missing.*

One should also note that the *sample elements are no longer independent.* The hot deck procedure is essentially a duplicating process with reported values substituting for missing values. The covariance that results from this duplication is ignored in the hot deck procedure. Ignoring this covariance can be a serious error.

Probably the greatest attraction of the hot deck procedure is its operational simplicity. The classification of the data items into I classes is an extremely adaptable method. The classification variables may be cardinal, ordinal, categorical, etc. In fact, the whole classification method may vary from the subjective to the mathematically rigorous. In addition, many complex surveys will use the hot deck procedure because of the pressure to retain the planned sample design (eg. self-weighing designs, survey designs using balanced repeated replications, etc.). However, the looseness of the classification method has tended also to obstruct theoretical evaluations of the hot deck procedure and thus to impede any theoretical comparisons between it and other missing data procedures.

The hot deck does have some simple qualities to recommend it. For example, let $E(\bar{x}-\mu) = B$ be the bias associated with nonresponse when estimating the population mean, μ, with the mean of a simple random sample. To estimate μ using the hot deck procedure one divides the sample into I classes. Let $E(\bar{x}_i - \mu_i)$ be the bias in class i, i = 1, 2, ..., I. If p_i^* is the proportion of the population in class i, then the bias, B_{HD}, associated with the estimated mean, $\bar{x}_{HD}$, of the sample data after applying the hot deck procedure is simply:

$$B_{HD} = E(\bar{x}_{HD} - \mu) = \sum_{i=1}^{I} p_i^* B_i.$$

To prove this equation one notes:

$$E[\bar{x}_{HD}] = E\left[\sum_{i=1}^{I} p_i \bar{x}_i\right] = E_{n_i}\left[E\left\{\sum_{i=1}^{I} p_i \bar{x}_i \,\middle|\, n_i\right\}\right]$$

where n_i is the number of sample units that fell in class i, n is the total sample size, $p_i = \frac{n_i}{n}$, and $\bar{x}_i$ is the sample average for class i. The expected value inside the braces is over fixed n_i, and the expected value outside the braces is then over all possible values of n_i. Obviously,

$$E_{n_i}\left[E\left\{\sum_{i=1}^{I} p_i \bar{x}_i \,\middle|\, n_i\right\}\right] = E_{n_i}\left[\sum_{i=1}^{I} p_i \mu_i\right] = \sum_{i=1}^{I} p_i^* \mu_i .$$

In spite of the fact that $\bar{x}_i$ and n_i are not independent, they are uncorrelated. Now, if $|B_i| < |B|$, for each i then:

$$(I)\quad |B_{HD}| = \left|\sum_{i=1}^{I} p_i^* B_i\right| < \sum_{i=1}^{I} p_i^* |B_i| < \sum_{i=1}^{I} p_i^* |B| = |B|.$$

Thus, one can see that the bias using the hot deck procedure is less than the bias caused by omission data *on the condition that* $|B_i| < |B|$ for each i. This condition should hold in most cases, but there is no guarantee because it is a function of the quality of the classification method. A good classification method should decrease the absolute value of the bias below $|B|$ in each of the I classes. However, the hot deck allows any classification. The goodness of the classification process is left to the integrity of the statistician.

The "Closest" Procedure

One possible alternative to the random substitution of the hot deck procedure is to substitute the "closest" reported item for each of the missing items. With one auxiliary variable, the "closest" value to a missing item is simply the value for which the absolute difference between the auxiliary variable of the missing item and the auxiliary variable of the reported item is minimized. In the case of ties for the "closest" auxiliary value a random selection of one of the tied values is made.

This procedure should have the same effect as assigning the population to many strata and, selecting a few units from each stratum (since the stratification is based on the auxiliary variable). Thus, suppositions that the hot deck method improves with more narrowly defined strata can be examined with the results of the "closest" procedure.

Given a good range coverage, this procedure is fairly robust to very curved relationships between the auxiliary and primary variables. The data in this investigation is not curved enough to reveal this robust property of the "closest" procedure.

The "Two Closest" Procedure

This procedure is another variation of the hot deck procedure. Instead of substituting the "closest" reported item for each missing item, one substitutes the average of the "closest" value whose auxiliary value is smaller than the reported item and the "closest" value whose value is larger than the reported item.

The "Class" Mean Procedure

This last variation of the hot deck procedure substitutes the average of the reported units in a class for each missing unit in that class. It is the simplest and probably the

cheapest of the procedures presented in this paper

THE SIMULATION EXPERIMENT

Why Use Simulation?

The need for simulation in this investigation is to compare the estimated variance of the estimated means. Possibly one might be able to compare how differences in the missing and reported data theoretically affect the estimated means using these six missing data procedures. However, the problem of analytically comparing the estimated variances of the estimated means is unreasonable. The fact that some assumptions fail in each procedure ties a knot in the analytical work.

For example, one should recall the estimated mean of the hot deck procedure, $\bar{x}_{HD}$. Assuming there are differences in the missing and reported data, one can not explicitly write the expected value of the estimated variance of $\bar{x}_{HD}$, $E[\hat{Var}(\bar{x}_{HD})]$. In fact, it is not known if $E[\hat{Var}(\bar{x}_{HD})] = Var(\bar{x}_{HD})$, and the author strongly doubts that it does. However, this paper will provide no evidence to support that supposition because the structure of the simulation of this experiment does not allow an estimate of $Var(\bar{x}_{HD})$ But does allow an estimate of $E[\hat{Var}(\bar{x}_{HD})]$. If $E[\hat{Var}(\bar{x}_{HD})] \neq Var(\bar{x}_{HD})$, then there is quite a weakness in the hot deck procedure. The costs of a simulation experiment providing this type of evidence would be much greater than the simulation actually used. This investigation contents itself with comparing the estimated variance of the estimated mean for each procedure with the estimated variance if the sample had no missing data. These comparisons will serve the purpose of revealing certain key qualities of each procedure.

One should note that the double sampling ratio and regression procedures also have variance estimates that involve assumptions and approximations that may be tenuous. For example, the assumption of a linear model is usually invalid in the regression and ratio procedures, and the ratio procedure simply uses substitution as an approximation to variance estimation. On the basis of two important studies (3;9) and practical experience one does not expect these biases in the variance estimates to be substantial for large samples. However, the comparisons among the procedures may be sensitive enough that these biases would be large enough to affect the comparisons.

Analysis

The primary point in the comparison of these procedures will be the minimization of the biases caused by missing data in the estimated means. Secondary importance is given to the comparisons of the estimated variances of the estimated means.

An important aspect of this study is the fact that the comparisons are based *on an experimental design where each observation is a result of a simulation of a procedure.* This situation is quite different from a simulation study where there may be a thousand or more simulations in order to narrow the confidence interval of an estimate almost to a point. Requests should be sent to the author for full details of the experimental design and details of the data used.

The correlations of the auxiliary variable and primary variable in the data range from 0.0 to 0.43. One may think that with larger correlations between the primary and auxiliary variable the estimates from the regression procedure would improve dramatically compared to the other procedures. However, the data in this study prevent evidence for or against this hypothesis. Surely, larger correlations would improve estimates resulting from all the procedures. Whether this improvement is equal for all the procedures is the question which can not be answered in this report.

Results

From the analysis of variance the six missing data procedures do not yield significantly different estimates of the mean. The average improvement in the estimated mean using the six missing data procedures is shown in Table 1. The relative bias reduction ranges from 8% to 26%.

Table 1: Average improvement in the estimated mean from the simulation of six missing data procedures.

B=Bias	*A*=Average estimate of mean Minus the true sample mean	$\frac{B-A}{B} \cdot 100\%$
-1.9	-1.40	26%
-2.9	-2.66	8%
-4.0	-3.63	9%
-6.2	-5.51	11%
-9.0	-7.65	15%
-14.0	-11.27	20%

Since there is no significant difference in the estimated means among the six procedures, the focus of interest becomes the estimated variances of the estimated means. The analysis of variance for the six missing data procedures with the estimated variance as the dependent variable was performed. Obviously, there is a significant difference among the estimated variances because the test statistic is so large:

$$\frac{MS_T}{MS_{TxPlots[Within\ AxB]}} = \frac{162.245}{0.391} = 414.95 .$$

Performing Duncan's multiple comparison test at a 95% significance level separates the procedures into the following groups:

t_1: double sampling ratio procedure
t_2: double sampling regression procedure

t_3: hot deck procedure with random substitution
t_4: "closest" procedure
t_5: "two closest" procedure

t_6: "class" mean procedure.

Perhaps it is more revealing to examine the estimated variances within each A x B cell, i.e. at different levels of bias. Table 2 shows the estimated variance using each procedure minus 9.685, the estimated variance if the sample has no missing data. Table 2 is on the next page.

The usual criterion for judging the variances of the estimated means resulting from the missing data procedures is that the smallest variance is the best. However, a procedure may result in a small estimated variance simply because of a large negative bias. By comparing the estimated variances resulting from the procedures with 9.685, the estimated variance if the sample has no missing data; one can judge if there are any large negative biases in the estimated variances. For example, if the estimated variance resulting from a missing data procedure is 7.20; then obviously, there is a large negative bias in estimating the variance.

One first notices that in Table 2, as in all the results, there is little difference between the regression and ratio estimates (procedures 1 and 2). One then notes that in the first three cases there is zero bias, and the hot deck estimator with random substitution (procedure 3) yields variances close to 9.685. The ratio and regression procedures have larger differences because they depend on the weak correlations of the primary and auxiliary variable. The "closest" procedure, the "two closest" procedure, and the "class" mean procedure (4, 5 and 6) have negative values. The negative values indicate *that their estimated variances are even less than the estimated variance if the sample has no missing data.*

CONCLUSIONS

The most important aspect in comparing these missing data procedures is to protect against biases in the estimated means (or totals). A split plot analysis of variance shows no significant differences among the estimated means which result in using these procedures. All the procedures reduce the relative bias that results from accepting the mean of the reported data as an estimate of the population mean. This reduction in relative bias is studied considering various non-response rates and considering various differences in the respondents and non-respondents. Varying from 8% to 26%, the reduction in relative bias averaged 15%. Considering the low correlations between the auxiliary and primary variables, this reduction is reasonable.

An important, though secondary, importance is attached to the estimated variances of the estimated means. All of these estimated variances except those from the ratio and regression procedures are generally less than the estimated variance that result with no missing data in the sample. Furthermore, this discrepancy *increases* as the relative bias increases. This part of the investigation clearly reveals why the hot deck, the "closest", the "two closest", and the "class" mean procedures may be undesirable. One is in the peculiar circumstance that, although all the procedures perform equally in reducing the bias of the mean, one can not choose a procedure on the basis of the smallest variance. The variances are deceiving. For example, using the hot deck procedure with the "class" mean substitution yields the smallest estimated variance, but this procedure is probably the worst. The results of this report indicate a large negative bias in the estimated variance resulting from the hot deck procedure because this estimated variance is much less than the sample variance when there is no missing data.

One should remember, however, that in many cases the only additional information on missing data is of a non-numerical nature. For example, data may be missing for a certain firm, but the only additional information is that it is a small insurance firm in Richmond, Virginia. In these cases, the hot deck procedure represents the only missing data procedure available. In the agricultural data of this investigation the additional information *is* numerical with the result that the ratio or regression procedure can be used and is better than the other procedures.

The final result of this investigation is a recommendation of the ratio or regression procedures (the effects of these two procedures being indistinquishable). These two procedures have been more theoretically explored than the other procedures. This estimated variances of the estimated means from the ratio or regression procedure reflect better than the other procedures the true quality of the data. The costs involved with the ratio regression computer program averaged about $5.00 simulation while the program for the other procedures averaged about $30.00 per simulation. Thus, the computer costs of implementing the ratio or regression procedure are probably much lower.

BIBLIOGRAPHY

1. Cochran, William G., Sampling Techniques, John Wiley and Sons, Inc. 1963.
2. Dear, R.E., "A Principle Component Missing Data Method for Multiple Regression Models" Systems Development Corporation, Santa Monica, California, SP - 86, 1959.
3. Frankel, Martain R., Inference From Survey Samples: An Empirical Investigation. Institute for Social Research, University of Michigan, 1971.
4. Hartley, H.O. and Hocking, R.R., "The Analysis of Incomplete Data", Biometrics, Volume 27, pages 783-824, 1971.
5. Hocking, R.R. and Smith, W.B., "Estimation of Parameters in the Multiple Normal Distribution with Missing Observation", Journal of the American Statistical Association, Volume 63, pages 159-173, 1968.
6. Kirk, Roger E., Experimental Design: Procedures for the Behavioral Sciences, Brooks Cole Publishing Company, 1968.
7. Kish, Leslie, Survey Sampling, John Wiley and Sons, Inc. 1965.
8. Rockwell Richard C., "An Investigation of Imputation and Differential Quality of Data in the 1970 Census", Journal of the American Statistical Association, Volume 70, pg. 39-42, 1975.

Table 2 -- The difference between the estimated variance of the estimated mean resulting from a simulation of the missing data and 9.685, the estimated variance of the estimated mean if the sample has no missing data.

B=Bias	Estimated Variance of the Estimated Mean Minus 9.685					
	Double Sampling Ratio Procedure	Double Sampling Regression Procedure	Hot Deck Procedures			
			Random Substitution	"Closest"	"Two Closest"	"Class" Mean
0.0	0.556	0.556	0.126	-0.024	-0.045	-0.528
0.0	0.885	0.883	0.055	-0.701	-0.295	-1.238
0.0	2.478	2.478	0.580	-0.785	-0.415	-1.971
-1.9	0.263	0.263	-0.095	-0.416	-0.454	-0.781
-2.9	-0.223	-0.224	-0.654	-0.970	-0.906	-1.204
-4.0	0.355	0.353	-0.884	-0.973	-1.094	-1.685
-6.2	0.096	0.094	-0.950	-1.277	-1.385	-1.872
-9.0	1.087	1.084	-1.205	-1.434	-1.700	-2.928
-14.0	-0.086	-0.090	-1.928	-2.579	-2.750	-3.571
Average Over All Levels of Bias	0.601	0.600	-0.551	-1.018	-1.005	-1.753

FOR FULL DETAILS OF THIS RESEARCH ONE SHOULD WRITE TO BARRY L. FORD, ROOM 2920, SOUTH BUILDING, USDA, WASHINGTON, D.C., 20250