SRD WORKING PAPER

A GENERAL OVERVIEW OF THE MISSING DATA PROBLEM

by

BARRY L. FORD

Sample Survey Research Branch
Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture
Washington, D.C.

August, 1978

PURPOSE

Two technical reports have been released which provide a detailed comparison of various missing data procedures. In fact, the details may have obscured general philosophies and concepts. The purpose of this working paper is to discuss these philosophies and concepts. After two reports the time seems appropriate to reflect upon the trends in the work which is already done and to outline the future work which is needed. Thus, this paper has a more general, discursive and often subjective tone than the two previous reports, but this tone is needed in order to give an overview without digressing into statistical technicalities.

DISCUSSION

Introduction:

There is little need to emphasize the problem of missing data. The alarming increase in refusals and inaccessibles in some states has underlined the importance of the problem for ESCS. For example, nonresponse for cattle and hog surveys has surpassed 20 percent in some midwestern states and exceeds 30 percent in some strata of large operators. Other statistical agencies have also noted the increase in survey nonresponse.

The nonresponse rate is a valid indicator of survey quality--as valid as coefficients of variation and standard errors. What do 5 percent coefficients of variation mean when the nonresponse rate is 25 percent? They probably do not mean very much. The Office of Management and Budget, for example, has recently stressed the importance of response rates to survey quality by setting a minimum bound of 50 percent on the response rates of government surveys. In the future one might expect OMB to provide more extensive guidelines on response rates.

Each agency is responsible for making its own adjustments for missing data. For example, ESCS list frame surveys that estimate numbers of livestock have a very simple adjustment. The current procedure is to:

1: delete refusals and inaccessibles (which reduces the sample size).

2: have a statistician impute for single, missing items.

This procedure reveals the fact that there are two types of missing data:

1: missing records -- all of the values for a sample unit are missing except for a control number

2: partially complete records -- only a few values are missing for a sample unit.

Since ESCS research on missing data has concentrated almost entirely on missing records, the discussion will first be confined to that particular area.

Missing Records--The Problem of Information

The basic problem with missing records--refusals and inaccessibles--is an information problem. *What information does one have on missing records?* By deleting the missing records from the sample the operational procedure assumes that there is no useful additional information. Thus, the assumption is implicitly made that the missing records are distributed the same as the reported records. When the multiple frame methodology was first planned, this assumption seemed reasonable to make because nonresponse rates were low. Therefore, the impact on survey estimates when the assumption did not hold was minimal. However, the developers of multiple frame methodology could not foresee the large increase in nonresponse rates which ESCS has recently experienced. Assumptions that were reasonable when a few records were missing are no longer reasonable as the nonresponse rate increases.

If it is unreasonable to assume that the missing records are distributed the same as the reported records, what is the best assumption ESCS can make? This question is really based on the more fundamental question of what information does ESCS have on the missing records. With regard to list surveys for livestock estimates, there are two types of information common to all states:

1: a control variable used to stratify the list

2: geographical information from the mailing address.

Because the control variable is the most important information available for a missing record, control data of a high quality is necessary to improve upon the assumption that reported and missing records have the same distributions. Logically, procedures which adjust for missing records are highly dependent on good control data. This dependency is so strong that before deciding which procedure is the best, one must answer "Is the quality of the control data good enough to warrant the adoption of *any* procedure over the operational one?"

The Statistical Research Division was interested in the quality of the control data before the release of two reports on missing records, but those studies have intensified the desire to examine the control data in the multiple frame states. A few states have been examined already, and the correlations within each stratum between the control variable and reported variables were usually below 0.30. These low correlations do not necessarily mean that the control variable is inadequate for stratification. However, they do restrict the effectiveness any missing record procedure *might* have in compensating for for nonrespondents. The second research report on missing record procedures indicates that at least a 0.60 correlation within each stratum between the control and reported variables may be needed before *any* missing record procedure improves upon the current operational procedure.

It is probably unfair to expect current control information to be adequate to the needs of missing record procedures. Its main purpose is to divide the list population into four to six strata in order to minimize the standard errors. This purpose requires much less powerful information than the needs of missing record procedures. Thus, the whole process of creating control information and its purpose needs to be evaluated.

Procedures to Adjust for Missing Records

When the Generalized Edit System was created, plans called for a module to be added which would impute values for any missing records. However, there is another option to consider. This option is to modify the summarization procedure so that the estimates are adjusted directly to reflect the effect of missing data. For example, one might directly increase the estimates by 20 percent rather than indirectly increasing them by going through the data set and imputing individual values for the missing records.

Almost any missing record procedure may be an imputation or a summarization procedure depending on its use. For instance, once a regression has produced an equation representing the relationship between a control variable and a survey variable, this equation may then be applied to the estimate (a summarization process) or to each missing unit in the sample (an imputation process).

Before one should decide on a missing data procedure, one should decide if a summarization or an imputation procedure is desired. Summarization procedures are usually the more direct approach and, therefore, easier to apply-- especially when the variables are quantitative and the sample design is as uncomplicated as

in a stratified simple random sample. On the other hand an imputation procedure produces a "clean" data set (i.e. data with no errors or gaps) and this facilitates further analysis. However, summarization may be ineffective in multi-stage sampling (such as the JES), and imputation procedures usually provoke the accusation of "making up" data. Statistics Canada, for example, uses an imputation procedure because one of its primary functions is to produce "clean" data sets which other government agencies use for their own analysis. Although ESCS is not currently in this position, some type of imputation process may still be desired.

During the research of procedures which adjust for missing records, two types of procedures emerged:

1: hot deck procedures which rely on a post-stratification of the reported data in order to substitute values from "similar" records

2: regression procedures which use regression relationships among the variables to adjust the estimates.

Hot deck procedures are imputation methods while regression procedures can be summarization *or* imputation methods.

Imputation methods can cause underestimates of standard errors, but replication is a useful tool to correct this defect. If the sample design is complex, even a regression procedure must often be used as an imputation method, and thus, the sample design must be replicated. Although yielding unbiased estimates of standard errors, replication does complicate the sample design. Therefore, statisticians should be aware that in many situations where a missing record procedure is desired, replication may also be required.

Additional Information for Missing Records:

All of the previous discussion (like the two ESCS reports on missing record procedures) has been strictly concerned with using existing information to adjust for missing records. However, there is the alternative of collecting additional information.

A good example of this technique is currently being tested by the Statistical Research Division and has already been the subject of one working paper, "A Study of Nonrespondents in Nebraska March Hogs Survey, 1978". This paper suggested using an estimator which only requires knowledge of whether the nonrespondent had *any* hogs or not. This estimator recognizes that a larger proportion of nonrespondents have hogs than the respondents. Thus, the mean of respondents having hogs is applied by stratum to the nonrespondents having hogs while nonrespondents without hogs receive zeros. Often nonrespondents will give this information in spite of refusing to give specific hog numbers. The main problem is that there are still a subgroup of nonrespondents for whom one might not find out even that much information.

Observational data is another example of additional information. On surveys where only personal interviews are used, the enumerators can observe whether livestock or livestock equipment (thus indirectly indicating livestock) are present. In fact, observational data is currently used for nonrespondents on the June Enumerative Survey. While farm and weighted estimators can be greatly biased by nonresponse, the tract estimator has great strength against a nonresponse bias because of observational data by the enumerators.

Missing Items:

Other organizations have done much more research on missing items than on missing records although the problem of missing items is not usually considered as serious. The main reason for this imbalance in research is that missing items are a more tractable problem, i.e. solutions are usually easier to derive and more

effective when applied. After all, there is often useful information available on missing items because the variables collected on a sampling unit are often highly correlated with one another.

Although missing items are not a very threatening problem to livestock list surveys, they could be to other surveys -- for example, labor surveys. Even the livestock list survey would benefit from a computerized procedure to adjust for missing items. A computerized procedure is not only more systematic and consistent than a statistician's edit but also probably decreases the time spent on edits.

PROPOSED RESEARCH

Missing Records:

The Statistical Research Division has completed two studies on records missing from the list sample of a multiple frame survey. The goal of these two studies was to find the missing record procedure which yielded the most accurate estimates. Achievement of this goal was obstructed by the low correlations between the survey data and the control data. Thus, past research points to a need for assessment of the control variable in terms of monitoring, construction and function.

The first step is simply to monitor the present quality of the control data. Although the Statistical Research Division has noticed the poor quality of control data in a few test states, the overall picture for multiple frame states is unknown. For each name in the selected sample of a list survey, the value of the control variable should be placed with the sample data. This action should become part of the operational procedure.

The second step is to examine the methods of constructing the control variable. There has never been any formal analysis of the construction of the control variable. Currently most states seem to follow a peak number approach to prevent large operators from being in the strata with large expansion factors. However, this approach needs to be re-evaluated to determine costs in terms of the standard errors of the estimates and the effects of missing data. Also, the idea of one control variable should be reviewed--perhaps two or three control variables for each livestock specie is a better approach.

The third step is to re-evaluate list stratification which is a function of the control information. The sampling rule that more than five or six strata do not substantially decrease the standard error is generally true, but the standard error is not the only factor that should be considered. The

effects of nonresponse should also be considered. Protection against missing records implies accurate control information, and accurate control information implies the possibility of a large number of strata. Given good control data, ten, twenty or a hundred strata may be a better idea than four to six strata. The concept of dividing the population into a large number of strata and then sampling a few units in each is used quite often in government surveys outside of the USDA.

Also, different methods of stratifying when two or more stratification variables are used has been the subject of current research outside the agency. If a geographical variable is important to the sample data, a geographical variable should be added to the stratification process. Another example might be the type of livestock operation, eg. farrowing operation v.s. feeder operation or consistent operator v.s. in-and-out operator.

One should recognize that all previous testing of missing record procedures was made using current survey data. If control variables, stratification principles, etc. are changed, then the testing must be done again. This restriction is necessary because more accurate control numbers and small stratum widths will improve estimates using the current procedure as well as other missing record procedures. Thus, all missing record procedures will improve, but whether the other procedures wil improve *more* than the operational procedure is a question to be answered by future research.

Missing Items:

Considering missing items as a problem by themselves has never been the subject of testing by the Statistical Research Division. The purpose of applying a computerized procedure on missing items is to improve the estimating program through a consistent, systematic and time-saving process in place of the current manual procedure. This approach may be particularly important to comparability

between survey results for different time periods and to avoid the subjectivity involved in a manual edit. Research on this problem (which is highly related to previous research) could proceed relatively quickly depending on the priority the agency wishes to attach to developing a procedure for missing items.