AN EMPIRICAL INVESTIGATION OF AN ALTERNATIVE NONRESPONSE MODEL FOR THE ESTIMATION OF HOG TOTALS

Matt Fetter, National Agricultural Statistics Service
USDA/NASS/Research Division/3251 Old Lee Hwy., Room 305, Fairfax, VA. 22030

KEY WORDS: Nonresponse model, reweighting, bias

1. THE SURVEY

The National Agriculture Statistics Service (NASS) conducts the Quarterly Agriculture Survey (QAS) to collect data on cropland acreage, grain storage, and various livestock items including hogs. The QAS is a "multiple frame" survey. Two independent frames are sampled, the "list" frame and the "area" frame. The "list" frame is a list of farm operations across the U.S. that NASS maintains. The "area" frame is composed of all land in the contiguous U.S. The QAS estimate for an item of interest is constructed by adding the estimate obtained from the list frame sample with the estimate obtained from the farm operations in the area frame sample that are not on the list frame. All estimators of interest in this paper are list frame estimators, thus the area frame portion of the multiple frame estimate will not be discussed further.

The QAS list frame is sampled using a stratified simple random sample design. Stratification is based primarily on each unit's control data for hogs, grain storage capacity, and acreage. A priority scheme is used to place each unit into exactly one stratum. The resulting stratification is not optimal for any one particular item of interest. For example, one stratum might be composed of units having similar grain storage capacity, but their hog characteristics might be quite different. Another stratum might be composed of units having similar hog characteristics but very different cropland acreage, and so forth.

Total nonresponse for QAS list frame samples typically range from 10 to 20 percent. Even for total nonrespondents there is often some information known about the sampled unit. For example, the interviewer or enumerator may be able to determine that the sampled unit is in business. Sometimes the presence or absence of hogs can be determined even though the actual number of hogs may be unknown. This partial information can be used to reduce nonresponse bias.

With the exception of certain self-representing strata, NASS currently uses sampling weight adjustment procedures (reweighting) to reduce nonresponse bias in its estimate of list frame hog totals. Two different estimators are used to model nonresponse. The first assumes that the nonrespondents can be reasonably well represented by the respondents. This is a strong assumption and its validity is seriously questioned. The estimator that is based on this assumption is not of significant interest and will not be formally discussed here. The other estimator is based on a model that uses the hog presence/absence information that is available on some nonrespondents. This estimator will be referred to as the Adjusted Estimator.

The Adjusted Estimator, developed by Crank (1979), was designed to take advantage of all partial information that was available on nonrespondents. At that time, the QAS was only designed to capture information regarding the presence or absence of hogs for nonrespondents. Currently, the QAS captures information regarding in/out of business status (ag-status) for nonresponding units. Cox (1993) described an alternative estimator that incorporates this additional information into the nonresponse model. The purpose of this paper is to describe this alternative estimator (referred to henceforth as the Revised Estimator) and to investigate the effect it has on the level of the estimates produced by the Adjusted Estimator.

The Revised Estimator, was applied to historic data so that a direct comparison of the two estimators could be made. Five major hog producing states (Georgia, Illinois, Indiana, Iowa, and North Carolina) were chosen for this purpose. The Revised Estimator was applied to 15 consecutive QAS surveys (June 88 - December 91) for each state. A comparison of the two estimates could then be made for each state in each quarter.

2. WEIGHTING CELL FORMATION

A weighting cell is defined as a group of sampled units within which nonresponse adjustments are computed and applied to the sampling weights. If the propensity to respond is linked to certain hog characteristics of the sampled units, it is desirable that weighting cells be composed of units that are similar in these characteristics. Under these conditions, all units within a weighting cell would be equally likely to respond. Thus the respondents would be representative of the nonrespondents and nonresponse bias would be minimal.

For the Adjusted Estimator, the weighting cells are the design strata. Because the stratification of the list frame is not optimal for hog estimation, design strata are not the most efficient cells for computing and applying nonresponse adjustments. Thus respondents are less likely to be representative of the nonrespondents within these cells. Through the use of poststratification, it is possible that improved weighting cells can be defined.

3. THE NONRESPONSE MODELS

In order to claim that a reweighted estimator is unbiased in the presence of nonresponse, some assumptions must be made about the nonrespondents. If all other factors are considered equal, the estimator based on the most sound set of assumptions would be judged as the estimator of choice for the reduction of nonresponse bias.

When considering the form of these estimators, it will be helpful to think of the estimation procedure as consisting of a sequence of three specific steps. For each sampled unit, three determinations need to be made. These are:

1) the sampled unit's status as an agricultural operation (ag-status). [(Is the unit in business or out of business)? This determination is only applicable in the case of the Revised Estimator.]

2) the sampled unit's status as a hog operation (hog-status). (Does the sampled unit raise hogs or not)?

3) the sampled unit's status as a hog-total respondent (hog-total status). (Is the number of hogs associated with the sampled unit known)?

A complete respondent will be defined as a sampled unit for which the number of hogs associated with that unit is known. A nonrespondent will be defined as any sampled unit for which any one of the above determinations can not be made.

In order to compare the nonresponse models implied by the estimators considered here, the underlying assumptions must be understood. At each modeled level of nonresponse, a valid assumption concerning the nonrespondents is required to claim that the estimator is unbiased in the presence of nonresponse.

The Adjusted Estimator adjusts for nonresponse at two levels, the hog-status level and the hog-total status level. Therefore, one assumption concerning the nonrespondents at each level must be valid. For the hog-status level, the required assumption is:

Assumption 1A. The probability that hog-status will be determined is the same for all sampled units in a particular stratum. This implies that hog-status nonrespondents represent a simple random sample of the stratum population.

For the hog-total status level the required assumption is:

Assumption 2A. Within a stratum, amongst all units which have been determined to be hog operations, the probability that the number of hogs associated with that unit will be obtained is the same for each unit. This implies that within a stratum, hog operations that are complete respondents represent a simple random sample of all sampled units which have been determined to be hog operations.

If N(h) represents the stratum h population size and n(h) represents the stratum h sample size, the Adjusted Estimator can be expressed in the following form at the stratum level:

$$\hat{Y}(h) = W_{samp}(h)\, A_{hog-st}(h) \sum_{e=1}^{2} A_{hog-tot}(he) \sum_{i=1}^{n(he)} y(hei) \quad (1)$$

where:

$\hat{Y}(h)$ represents the estimated number of hogs in stratum h.

$W_{samp}(h) = N(h) / n(h)$.

$A_{hog\text{-}st}(h) = n(h) / n_{hog\text{-}st\ resp}(h)$, the hog-status nonresponse adjustment for stratum h,

where:

$n_{hog\text{-}st\ resp}(h)$ represents the number of hog-status respondents in stratum h.

$A_{hog\text{-}tot}(he) = n_{hog\text{-}st\ resp}(he) / n_{comp\text{-}resp}(he)$, the hog-total status nonresponse adjustment for weighting class e in stratum h,

where:

$n_{hog\text{-}st\ resp}(he)$ represents the number of hog-status respondents in weighting class e within strataum h and,

$n_{comp\text{-}resp}(he)$ represents the number of complete respondents in weighting class e within stratum h.

y(hei) represents the number of hogs reported by complete respondent i in weighting class e within stratum h.

n(he) represents the number of units in class e in stratum h.

The subscript e denotes two distinct sets (classes) of hog-status respondents in stratum h; hog operations and non-hog operations. Once a sampled unit is identified as a non-hog unit, the number of hogs associated with that unit is immediately known to be zero. Thus all identified non-hog units are complete respondents. Let e= 1 denote this class. For this class, there is no nonresponse at the hog-total status level. Thus:

$A_{hog\text{-}tot}(h1) = 1$ since:
$n_{hog\text{-}st\ resp}(h1) = n_{comp\text{-}resp}(h1)$.

For the hog operation units (e=2), $A_{hog\text{-}tot}(h2)$ must be expressed in the general form stated above.

The Revised Estimator adjusts for nonresponse at three levels, the additional level being the ag-status level. For each of the three levels, one valid assumption is required for the estimator to be unbiased. These assumptions are:

Assumption 1R. The probability that ag-status will be determined is the same for all sampled units in a particular weighting cell. This implies that ag-status nonrespondents can be thought of as a random sample of the cell population.

Assumption 2R. Within a particular weighting cell composed of identified ag-operations, the probability that hog-status will be determined is the same for all units comprising that cell. This implies that hog-status nonrespondents can be thought of as a random sample of the units composing the cell.

Assumption 3R. Within a particular weighting cell composed of identified hog operations, the probability that hog-total status will be determined is the same for each unit in that cell. This implies that the hog-total status nonrespondents can be thought of as a random sample of the units composing the cell.

The assumptions on which these estimators are based are likely to be invalid unless the weighting cells are judiciously defined. In order to increase the likely validity of the underlying assumptions of the Revised Estimator, it was desirable to define the weighting cells in such a way that they would be composed of units having similar hog characteristics. Poststratification based on each unit's hog control data was used to form weighting cells. Thus the weighting cells were defined similarly to the way that design strata would be defined for a hog-specific survey. In order to further increase efficiency, the weighting cells (post-strata) were defined to insure that approximately 20 complete respondents would be contained in each cell. (The Adjusted estimator is not implemented in such a way as to insure reasonably high numbers of complete respondents).

Because the weighting cells cut across design strata, the Revised Estimator will be expressed at the final nonresponse adjustment cell level, e, e= 1,...,E. The general form of the Revised Estimator is:

$$\hat{Y}(e) = A_{hog\text{-}tot}(e) \sum_{i}^{n_e} W_{samp}(ei)\ A_{ps}(ei)\ A_{ag\text{-}st}(ei)\ A_{hog\text{-}st}(ei)\ y(ei) \qquad (2)$$

where:

$\hat{Y}(e)$ represents the estimate of the total for hog-total status weighting cell e,

y(ei) represents the number of hogs reported by unit i in weighting cell e.

n_e represents the number of sampled units in weighting cell e,

$W_{samp}(ei)$ represents the sampling weight for the ith unit in weighting cell e,

$A_{ps}(ei)$ represents the poststratification adjustment for the ith unit in weighting cell e,

$A_{ag\text{-}st}(ei)$ represents the ag-status nonresponse adjustment for the ith unit in weighting cell e,

$A_{hog\text{-}st}(ei)$ represents the hog-status nonresponse adjustment for the ith unit in weighting cell e, and

$A_{hog\text{-}tot}(e)$ represents the hog-total status nonresponse adjustment for the ith unit in weighting cell e. (Note all hog-total status respondents have the same hog-total status adjustment within class e).

All of the nonresponse adjustments have the usual form:

$$\frac{\sum_{\text{all sampled units } \in \text{ cell}} W^*_{samp}}{\sum_{\text{all responding units } \in \text{ cell}} W^*_{samp}}$$

where W^* represents the sampling weight or an adjusted sampling weight, depending on the level of the adjustment. All nonrespondents have a nonresponse adjustment of zero by definition.

The poststratification adjustment has the following form:

$$A_{ps}(ei) = \frac{N(g)}{\sum_{i \in g} W_{samp}(gi)}$$

where N(g) represents the number of units on the list frame that fall in poststratum g and $W_{samp}(gi)$ is the sampling weight for the ith sampled unit in poststratum g.

4. THE VALIDITY OF THE ASSUMPTIONS

Although the assumptions implied by the Adjusted Estimator are reasonable, they are not beyond justifiable criticism. As stated earlier, assumption 1A asserts that within a stratum, all sampled units are equally likely to be hog-status respondents. However, if the partial information concerning ag-status is considered valid, then the original sample can be divided into three mutually exclusive groups: 1) those units for which ag-status is not determined, 2) those units identified as non-ag units, and, 3) those units identified as ag units. All units in the first group have a zero probability of having hog-status determined because hog-status determination implies ag-status determination. Clearly, hog-status determination is certain for all units in the second group because all non-ag units have zero hogs. Therefore, one could argue that it would be desirable to augment the nonresponse model so that the probability of determining ag-status is the same for all sampled units, while the probability of determining hog-status is the same for all sampled units which are known to be ag-operations. If valid, this argument would imply that the Adjusted Estimator is based on a misspecified model.

If the Adjusted Estimator is based on a misspecified nonresponse model, it is of interest to understand the effect that this misspecification is having on the estimates of hog totals. First, an argument for the nature of the misspecification will be presented. Second, the effect of this misspecification on the level of the estimate will be described.

All ag-status nonrespondents are either: 1) non-ag units (out of business), 2) non-hog ag-operations, or 3) hog operations. Because every unit in the population must be one of these types, it is reasonable to assume that ag-status nonrespondents represent a random sample of the cell (stratum) population. However, the Adjusted Estimator is based on the stronger assumption that the hog-status nonrespondents as a whole represent a random sample of the cell (stratum) population (see figure 1). For a moment, let us assume that this assumption is valid. If we adopt as a premise that a subset of this set-- ag-status nonrespondents, represents a random sample of the cell population, then the compliment of this subset-- identified ag-operations that are hog-status nonrespondents, must also represent a random sample of the cell population. It will now be argued that the Adjusted Estimator's assumption is not reasonable under the

adopted premise.

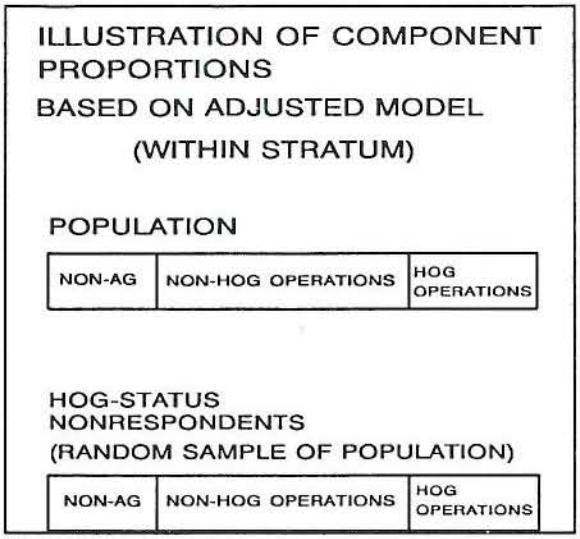


Figure 1

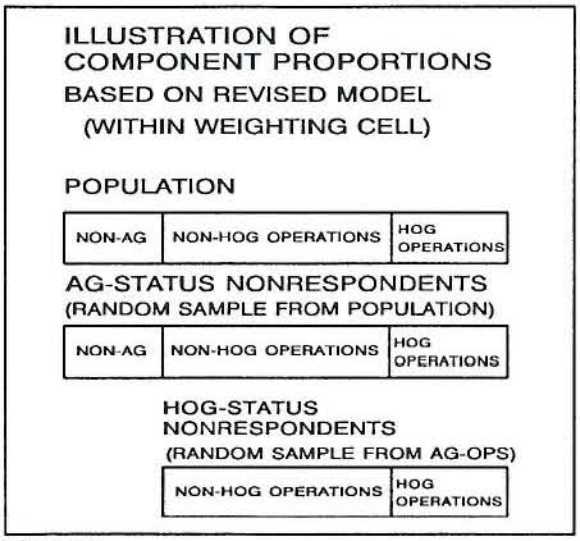


Figure 2

All identified ag-operations that are hog-status nonrespondents must either be non-hog ag-operations or hog operations. Because non-ag operations are missing from this group (all non-ag units are hog-status respondents-- they are non hog units), it is difficult to argue that identified ag-operations that are hog-status nonrespondents can be thought of as a random sample of the cell population (see figure 2). The effect of this misspecification is to bias the estimate downward. This can be explained as follows:

Identified ag-operations that are hog-status nonrespondents have only one source of zeros-- non-hog ag-operations, whereas ag-status nonrespondents have two sources of zeros-- non-ag units and non-hog ag-operations (see figure 2). It therefore seems reasonable to assume that identified ag-operations that are hog-status nonrespondents are more likely to be hog operations than ag-status nonrespondents. It is thus argued that the Adjusted Estimator essentially underestimates the proportion of hog operations in the population. It gives an unbiased estimate of this proportion for the ag-status nonrespondents but gives a downward biased estimate for those identified ag-operations that are hog-status nonrespondents.

The Revised Estimator is based on the augmented model referred to earlier. The difference between the underlying models of the Revised and Adjusted Estimators is that the Adjusted Estimator models all hog-status nonrespondents the same way (Assumption 1A). The Revised Estimator models ag-status nonrespondents as if they are a random sample of the cell population (Assumption 1R), and models identified ag-operations that are hog-status nonrespondents as if they represent a random sample of those units identified to be ag-operations (Assumption 2R).

Note that both estimators model identified hog operations that are hog-total nonrespondents as though they represent a random sample of those records identified to be hog-operations. (Assumption 2A is essentially the same as assumption 3R.)

5. RESULTS AND CONCLUSIONS

The main focus of the research was to observe how estimates obtained from the Revised Estimator would compare to those produced by the Adjusted Estimator using historical QAS data files. The observed effect of applying the Revised Estimator to historical data is an increase in the estimated total number of hogs. This supports the argument that the Adjusted Estimator is biased downwards. The average percentage increase relative to the Adjusted Estimator ranged from a low of 0.64 percent in Iowa to a high of 2.96 percent in Georgia. Across the five states studied, the increase averaged 1.53 percent over all quarters. There were several quarters for which the Revised Estimator produced a lower estimate than the Adjusted Estimator. This was not due to the nonresponse model, but was caused by the poststratification crossing design strata. The Revised Estimator tracked well with the

other estimators for all states. Figure 3 shows the relationship between the estimators for Illinois.

The structure of this Revised Estimator is appealing because it provides separate assumptions for each of the three stages of nonresponse. A logical argument has been made that the distribution of the nonrespondent population is different between the ag-status and hog-status stages. The assumptions that nonrespondents are random samples at each stage serve as a reasonable baseline approach, but as yet have not been validated by empirical evidence. Further study is needed to determine the appropriateness of these assumptions.

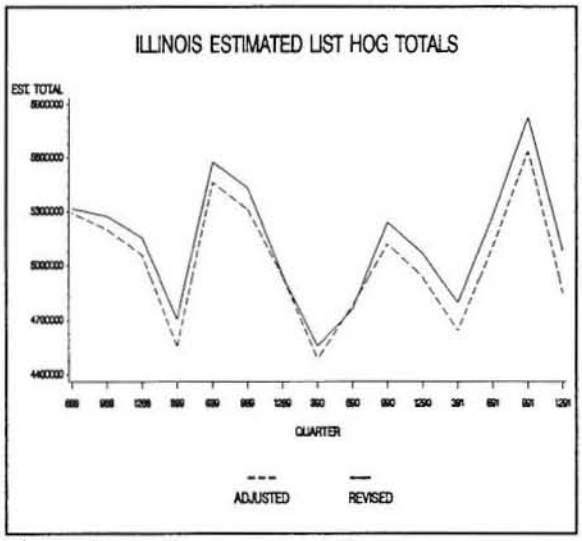


Figure 3

REFERENCES

[1] Agricultural Statistics Board (1991), "Hogs and Pigs" (September 1991 Report), National Agricultural Statistics Service.

[2] Cox, B. G. (1993), "Weighting Class Adjustments for Nonresponse in Integrated Surveys: Framework for Hog Estimation," National Agricultural Statistics Service.

[3] Cox, B. G. (Unpublished), "Weighting Survey Data for Analysis," National Agricultural Statistics Service.

[4] Crank, K. N. (1979), "The Use of Partial Information to Adjust for Nonresponse," Agriculture, Economics, Statistics, and Cooperatives Service.

[5] Fuchs, D. R., and Bass, R. T. (1990), "1989 Agricultural Surveys Survey Administration Analysis," National Agricultural Statistics Service.

[6] Kott, P. S. (1990), "Nonresponse Adjustments in NASS Agricultural Surveys," National Agricultural Statistics Service.

[7] Kott, P. S. and Thorson, J. (1989), "Improving Variance Estimates for Livestock Surveys," National Agricultural Statistics Service.

[8] Garribay, R. and Huffman, J. E. Jr. (1991), "1990 Agricultural Survey Administration Analysis," National Agricultural Statistics Service.

IDENTIFYING AND CLASSIFYING REASONS FOR NONRESPONSE ON THE 1991 FARM COSTS AND RETURNS SURVEY

Terry P. O'Connor, USDA/NASS
Research Division/3251 Old Lee Hwy., Fairfax, VA 22030

ABSTRACT

A research study was conducted during the 1991 Farm Costs and Returns Survey (FCRS) to identify and classify the reasons given to field interviewers by potential respondents for refusing to participate in the survey. The reasons given by field interviewers for coding a sampled unit as inaccessible during the survey were also identified and classified.

The research was conducted in all 48 surveyed states, and included 6 FCRS questionnaire versions. Upon receiving a refusal, interviewers were instructed to record the reason given on the face page of the questionnaire. If no reason was given, or in cases where more than one reason was given, the interviewers were instructed to discuss the concerns of the respondent in regards to completing an interview, and identify the main reason for refusing. When a sampled unit was coded as a inaccessible, interviewers were instructed to explain the reason for the inaccessible.

During the survey statistician's manual edit of the questionnaires, the reasons for refusal or inaccessible were reviewed and compared to a coded list of reasons for nonresponse compiled from previous research into this topic on the FCRS. Statisticians could consider the comments from the interviewers as a match to a pre-coded response, or add additional codes for unique comments.

The nonresponse rate on FCRS averages 30% per year. The reasons behind the nonresponse have been a source of speculation for many years, and previously only anecdotal evidence was available on which to base efforts to maximize response. This research shows the anecdotal evidence to have been on the mark in some cases an off in others.

INTRODUCTION

The Farm Costs and Returns Survey (FCRS) is a face to face interview survey conducted annually during February and March by the National Agricultural Statistics Service (NASS). It is a survey of the agricultural sector, and is conducted in the 48 conterminous states to collect detailed information on farm expenditures and income, costs of production and demographic data. The FCRS has a multiple frame design utilizing a list sample of medium and large ranches and farms, and an area nonoverlap sample of Resident Farm Operators (RFOs) not represented by the list, most of whom operate small farms (Rutz, 1991).

While all 48 FCRS states utilize the same survey procedures, the FCRS includes several questionnaire versions used in different combinations across the country. The versions used in a particular state for a given year depend upon the agriculture in that state and the areas of agricultural specialization being studied. Costs of producing the various agricultural commodities are studied on a year-to-year rotating basis. There are variations in geography, sample sizes, farm or ranch types and sizes, economic conditions and respondent attitudes about the survey across the country; therefore, many factors must be considered when making direct state to state comparisons of the survey results (Rutz, 1991).

The 1991 FCRS national response rate was 67.9 percent, with a refusal rate of 24.9 percent and an inaccessible rate of 7.2 percent. Response rates on the survey have declined slightly over time, despite extensive efforts to limit nonresponse. While NASS uses farm expense data from the FCRS in its reports, the primary user of the FCRS dataset is the Economic Research Service (ERS), which utilizes all of the FCRS data in producing economic analyses and cost of production reports (Rutz, 1991).

A benefit of collecting this type of information is that survey managers can make adjustments to the public's perception of a too long interview by testing a shortened version of the questionnaire (as is being planned for the 1992 FCRS). Headquarters can prepare materials to aid survey statisticians in training their interviewers to meet the challenges of the refusal types common across states. Survey statisticians should develop materials for use in their state workshops to prepare interviewers for situations common to their state. Experienced interviewers who have had success in converting refusals into respondents should share their techniques through panel presentations or group discussions. In this way, interviewers will maximize

response rates on the initial contact by being prepared to discuss concerns and grievances brought up by the respondents, thus avoiding the additional time and money costs of a re-contact.

BACKGROUND

The research project to identify and classify nonresponse on the FCRS stems from four years of preliminary work which the author completed while on staff in the South Carolina and Indiana State Statistical Offices (SSOs).

Beginning with the 1985 FCRS, the author required that the South Carolina interviewers document the reasons given by respondents who refused to participate in the survey. Previously, interviewers were likely to simply write "refusal" across the questionnaire, and the comments the interviewer received from a refusal were discussed second or third hand if at all, and were sketchy at best.

Then on the 1986 FCRS, South Carolina was selected as one of six states to take part in a refusal conversion research project. All respondents who refused to participate in the survey during the initial contact were to be re-contacted with the purpose of convincing them to complete an interview. It was apparent that interviewers selected to re-contact a refusal in the current survey had an advantage if they were aware of the reason the respondent gave when initially refusing.

The information on "reasons for refusing" gathered during 1985 were discussed during the training workshop for the 1986 FCRS, and responses to the reasons were developed by the interviewers. To prepare for the re-contact required by the research, interviewers were again required to write on the questionnaire the exact reason or circumstances behind each refusal received on the FCRS. In this way, subsequent interviewers were made aware of the events of the initial contact.

The primary benefit of identifying the refusal types was that the interviewers could PREPARE for common situations before encountering them in interview situations. According to interviewer comments, this preparation improved their confidence in approaching interviews, and even when they could not prevent a refusal, they were able to set the stage for the respondent's cooperation on other upcoming surveys. The second benefit was that, when approaching a re-contact on the refusal conversion project, the subsequent interviewer could prepare for a specific situation. A third benefit was that interviewers (with their supervisor's approval) could eliminate re-contacts of certain refusal types (violent refusals, death in the family, etc.), saving money and time during the critical data collection period.

Perhaps because the refusal conversion project was new and received much attention, or perhaps because the refusal identification preparation worked, the FCRS response rate in South Carolina for 1986 was 17 percent higher than in 1985 (Dillard, 1987). The author attributes most of this increase to interviewer preparation on the initial contact since only a small number of refusal conversions were obtained.

Upon transferring to the Indiana SSO, the author again instructed the field interviewers to document the reasons given by refusals. While the refusal identification and interviewer preparation led to an initial decrease from 35 percent to 31 percent in the refusal rate in Indiana, no additional gains have been evident, with the refusal rate averaging 31 percent over the past five years. The list of refusal types compiled during this time served as the basis of the refusal list utilized for the nonresponse identification project on the 1990 FCRS.

This research was conducted during February and March, 1991. The six test states included two states that averaged high nonresponse rates, two states that averaged mid-level nonresponse rates, and two states that averaged low nonresponse rates on the FCRS. Comments from the FCRS post-survey evaluations completed by survey statisticians around the country alluded to problems with certain refusal types, but with only anecdotal information to support their impressions. Evaluations included the following comments:

* "Some farmers feel it's none of our business."
* "Many farm operators refused due to the length of the questionnaire."
* "Most of the second time contacts were refusals and didn't want to be contacted again."

Some...many...most. The 1990 FCRS nonresponse identification project was expanded to all surveyed states for 1991 in order to put some numbers on these valid concerns and to better determine what NASS is up against when trying to minimize FCRS nonresponse.

RESULTS

The results of the 1991 refusal identification and classification research are listed in Appendix A.

Refusal types coded 01 - 53 were provided in the survey instructions; codes 200 - 409 were initially left blank for state use, and states added refusal types based upon their data collection experiences with the survey.

The most frequent reason given by the farmers when refusing to participate in the survey was "Would not take the time / too busy". This response was given by 1,395 of the 5,663 refusals encountered (24.6%), and was recorded nearly twice as often as the next most frequent response. This seems to be strong evidence for those involved with the survey who believe that farmers perceive the interview to take too long.

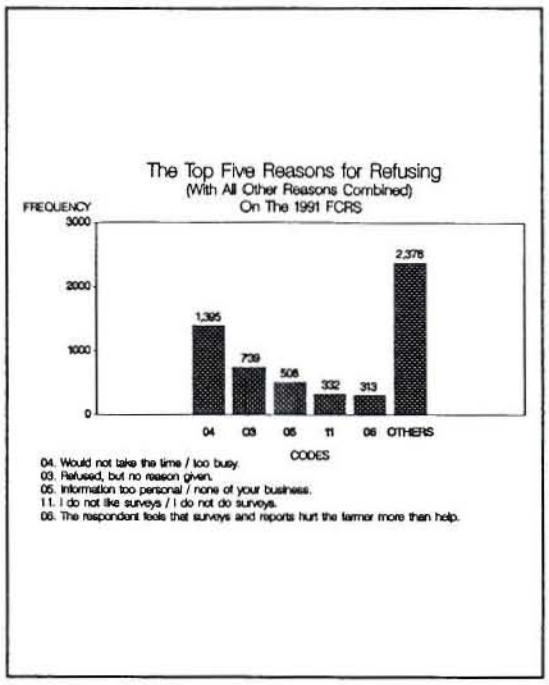
The second most frequent reason recorded was "Refused, but no reason given", mentioned 739 times, or 13.0 percent of the total refusals received. This category represents a difficult type of refusal to convert to a respondent: they just say NO. They may understand what NASS is and its mission, and may even recognize the interviewer from previous contacts, but cut off any attempt at an interview before their concerns can be identified and addressed.

The third most frequent reason recorded was "Information too personal / none of your business", mentioned 508 times, or 9.0 percent of the total refusals received. Together these first three reasons account for 46.7 percent of the total refusals received, and the top five reasons account for 58 percent, even though 52 different reasons for refusing were mentioned during this research.

Refusal reasons mentioned as frequently and as widespread as these five should be addressed on a national level. However, SSOs must review their state specific data to determine which less frequently mentioned reasons are important to their state.

This research also involved identifying and classifying the reasons given by an interviewer when coding a sampled unit inaccessible, shown in Appendix B. Inaccessible types coded 75 - 150 were provided in the survey instructions; codes 500 - 709 were initially left blank for state use, and the states added inaccessible types based upon their data collection experiences with the survey. While basically separate from the refusal identification, certain respondent situations (such as "Family illness / death") could be coded either as a refusal, an inaccessible or a valid zero out-of-business depending upon the circumstances encountered.

One benefit of this research is that the number of incomplete questionnaires, that is, those questionnaires



for which the respondent could not or would not provide enough information for the interview to be completed, is evident for the first time. For the 1991 survey, 263 questionnaires were coded as incomplete and were not summarized. This amounts to 3.6 percent of the nonresponse, but is only 1.2 percent of the total survey contacts.

The most frequent inaccessible reason recorded by the interviewers was "Tried several times; could not reach anyone for an appointment. Just an extremely busy person.", given for 455 of the 1,653 inaccessibles encountered (27.5%). This is a surprising finding in light of the six week data collection period.

The second most frequent inaccessible reason recorded was "Illness / death in the family prevents the operator from responding", mentioned 182 times, representing 11.0 percent of the total. This is a difficult situation for an interviewer to encounter, and setting the stage to see a respondent under better circumstances in the future is the best that can be accomplished.

The third most frequent reason recorded was "Farm records are not available until after the survey period closes", mentioned 172 times, representing 10.4 percent of the total. Together these first three reasons account for 48.9 percent of the total inaccessibles recorded, with 23 different reasons for coding an inaccessible mentioned during this research.

SSOs must review their state specific data to determine

which additional reasons are important to their state. For instance, "The operator is away on an extended vacation", normally thought to be a Midwest or Northern situation for escaping the snow, was also mentioned in California, Florida and other warm weather states.

DISCUSSION AND RECOMMENDATIONS

Data analysts, survey managers, statisticians and interviewers are concerned about the levels of nonresponse on the FCRS. Being close to the survey, they develop impressions about what factors are "driving" the nonresponse. The purpose of this research is to identify the reasons for nonresponse, and to attach some numbers to them in order to rank their relative importance. Considering the nature of the FCRS, that it is a long, detailed interview of a respondent's operating procedures, income and expenses, assets and liabilities and demographic information, many survey organizations would be thrilled to have a national response rate exceeding 70 percent. Rather than defend this position, the survey managers at NASS and ERS continually strive to improve the response rate on the survey.

Following a discussion of the preliminary results of this study and from previous consideration of the subject, NASS and ERS have agreed to test a shortened version of the questionnaire for the 1992 survey year. A detailed discussion of the benefits of a shortened questionnaire version can be found in Dillard (1991). NASS will provide training and materials to the survey statisticians at the regional workshops in January, 1993, to aid in training their field interviewers during state workshops. Additionally, the information is useful in the weighting of survey results and summarization.

According to Turner (1992) the FCRS nonresponse adjustment factor is based on an assumption that all nonrespondents are operating farms; that is, they would provide positive records if interviewed. Miss-coding valid zero reports as nonrespondents will positively bias the expanded indications. Turner (1992) states that, "Identifying these (nonresponse) reasons will enable enumerators to improve classification of cases where no farm appears to exist as a valid zero. Continued emphasis should be given to classifying only positives as refusals and inaccessibles. Those nonrespondents that have no indication of being in business should be coded as out of business."

Look at the pattern of nonresponse across the data collection period, and an interesting picture appears. In five of the seven survey weeks, more refusals occurred on Mondays than on any other single day, and during the other two weeks, the number of Monday refusals is near the peak for the week. This is probably a function of more interviews being attempted on Mondays, but it may also indicate that Mondays are not the best day to attempt a long interview without a prior appointment. Otherwise, the distribution of refusals seems normally spread throughout the survey period.

As might be expected, the number of inaccessibles peaks near the end of the data collection period when time constraints force the interviewers to begin to give up on respondents who either cannot be located or who continue to put off the interview when contacted. In general, the incomplete interviews seem normally spread throughout the data collection period.

As the results from the six test states in the 1990 research served as an excellent predictor of the 1991 results, there does not appear to be enough yearly variation to justify transferring this research into an operational aspect of the survey. I recommend that this research be repeated in three years. In this way, each SSO can be updated on the causes of nonresponse likely to be encountered, and patterns of nonresponse can be compared.

REFERENCES

Dillard, Dave and T. Gregory (1987). 1986 FCRS Analysis, Report II. Response Rates, Interview Times, and Data Collection Costs. United States Department of Agriculture, National Agricultural Statistics Service.

Gregory, Thomas L. (1990). 1989 Farm Costs and Returns Survey, Survey Administration Analysis. United States Department of Agriculture, National Agricultural Statistics Service, Agricultural Statistics Board, NASS Staff Report SMB Number 90-03.

Rutz, Jack L. and C.L. Cadwallader (1991). 1990 Farm Costs and Returns Survey, Survey Administration Analysis. United States Department of Agriculture, National Agricultural Statistics Service, Research and Applications Division, NASS Staff Report SMB Number 91-04.

Turner, Kay (1992). Modification of FCRS Nonresponse Adjustment Procedures. United States Department of Agriculture, National Agricultural Statistics Service, Research Division, SRB Research Report Number SRB-92-08.

APPENDIX A: Reasons Given By Respondents When Refusing To Participate on the 1991 Farm Costs and Returns Survey, All States and Versions Combined.

FREQUENCY	CODE	REASON
1,395	04.	Would not take the time / too busy.
739	03.	Refused, but no reason given.
508	05.	Information too personal / none of your business.
332	11.	"I do not like surveys / I do not do surveys."
313	06.	The respondent feels that surveys and reports hurt the farmer more than help.
255	02.	Contact attempted, but respondent refuses on all surveys, and refused on this one.
253	10.	"I will have nothing to do with the Government."
195	34.	Respondent will do other surveys, but not financial surveys.
135	20.	Family illness / death.
134	12.	Respondent only does compulsory surveys.
128	18.	The respondent feels the operation's records are inadequate to complete the interview.
120	16.	"My farm is too small to count / too small to be representative."
120	17.	"You contact me too often."
105	21.	Operator would not keep appointments.
97	19.	Farm records are at the tax advisors / lawyers.
95	07.	"I did this survey before, but not again."
89	01.	Known refusal, no contact attempted.
72	32.	"This is not a farm."
64	24.	Violent / threatening refusals.
58	52.	Questionnaire not sent to the field to avoid jeopardizing cooperation on other surveys.
56	27.	Respondent is quitting farming.
48	28.	Out of business now, will not answer for the previous year.
46	23.	Wants to be paid for interview time and effort.
42	08.	"I just did a different survey for your office."
40	22.	Spouse / secretary / etc. will not let the enumerator see the operator.
36	13.	The respondent does not think the information is kept confidential.
36	26.	Respondent does not want to report due to legal / financial problems.
30	25.	Respondent does not want to talk about farming.
29	14.	The respondent mentions a specific grievance with the SSO or NASS (other than confidentiality).
22	29.	Figures for the previous year were not typical.
18	09.	"I just did a survey for someone else."
18	53.	Would not answer the door even though they were home.
5	365.	Operator called the office after receiving the pre-survey letter, asked not to be contacted.
5	366.	The operator does not believe in statistics, so will not complete an interview.
4	15.	The respondent mentions a specific grievance with the state cooperator.
2	240.	Needed partner to provide some information; partner refused.
2	260.	Getting divorced, too upset to respond.
2	265.	Operator has a grievance with the IRS.
2	267.	Fed up.
1	215.	Water rights curtailed, will not cooperate.
1	250.	"The government is broke, how can we afford to send these people out?"
1	255.	NASS data is not accurate. Too political.
1	256.	Doing well financially -- does not want to respond.
1	257.	Operator has several operations and could not separate records for the sampled unit.
1	258.	Upset with the government -- has to spend $20,000 to dig up fuel tanks.
1	262.	Farmhouse and records lost in a fire, January, 1992.
1	269.	This survey is not needed.
1	270.	Responded previously on this survey, and asked to be excused this year.
1	335.	The respondent feels the operation is too complex for our survey.
1	340.	The respondent has a specific grievance with ASCS.
1	341.	The farm operation is in a blind trust for a national politician.
1	367.	His father would not do surveys, so neither will the son.

5,663	Total Responses	

* Code numbers not listed were not used.

APPENDIX B: Reasons Given By Enumerators When Coding a Sample Unit as Inaccessible/Incomplete on the 1991 Farm Costs and Returns Survey, All States and Versions Combined.

FREQUENCY	CODE	REASON
455	116.	Tried several times; could not reach anyone for an appointment. Just an extremely busy person.
263	150.	INCOMPLETE -- Respondent provided partial information, but would not or could not provide enough information to make the questionnaire complete.
182	84.	Illness / death in the family prevents the operator from responding.
172	85.	Farm records are not available until after the survey period closes.
169	86.	Respondent postponed the interview beyond the end of the survey period.
142	79.	The operator is away on an extended vacation.
80	81.	The operator is away on business.
67	80.	The operator is away on a brief vacation.
27	76.	No respondent, as listed on the label, could be found.
26	94.	Inaccessible, but no reason given.
18	82.	The address on the label is summer-seasonal housing.
12	75.	No operation, as listed on the label, could be found.
9	83.	Access to the address on the label was denied by a gate / guard / etc.
7	78.	The address on the label is vacant / burned out / no structure exists.
7	87.	Enumerator workload prevented this operation from being contacted during the survey period.
5	591.	The operator moved away during 1991.
3	667.	The questionnaire was returned too late to be included in the summary.
2	92.	Non-English speaking respondent; interpreter not available.
2	119.	Enumerator mistake; caught it too late to complete an interview within the survey period.
1	120.	Operator has several operations and could not separate records for the sampled unit.
1	540.	Questionnaire from the enumerator lost in the mail.
1	561.	Operator had just gotten out of jail and would not talk with anyone from the government.
1	565.	Enumerator did not contact sufficiently; gave up too soon.
1	580.	Enumerator error, should not have collected the data.

1,653 Total Responses		

* Code numbers not listed were not used.

AN EVALUATION OF NONRESPONSE ADJUSTMENT WITHIN WEIGHTING CLASS CELLS FOR THE FARM COSTS AND RETURNS SURVEY

Kay Turner, USDA/NASS
Research Division Room 305, 3251 Old Lee Hwy. Fairfax, VA 22030

INTRODUCTION and OBJECTIVES

The Farm Costs and Returns Survey (FCRS) is conducted by the National Agricultural Statistics Service (NASS) during February and March of each year. The data are collected in the 48 contiguous States from farm operators/managers for the preceding year via personal interviews. Various versions of the FCRS collect detailed and aggregate expenditure, income, asset, liability and cost of production data. The data from the FCRS are used to ascertain the financial status of the agriculture sector by supplying information such as: farmers' net income, costs of producing commodities, financial situation of farm operators, debt held by farm operators, and importance of production expense items. Farm organizations, agribusinesses, Congress, the Department of Agriculture, farmers, and ranchers are some of the groups that utilize FCRS data (NASS, 1989). Each year a sample is drawn for the FCRS using both list and area frames. The list frame includes mainly large and specialty operations. The area frame includes small operations not on the list frame, or nonoverlap (NOL) (NASS, 1991).

Nonresponse exists because all sampled farm operators do not respond to the survey. The two types of nonrespondents are refusals (the farm operator declines the interview) and inaccessibles (the farm operator cannot be contacted). Kalton and Maligalig (1991) note,

> "When total nonresponse occurs, the survey analysis may simply be carried out on the data provided by the responding elements. However, since responding and nonresponding elements may differ systematically in their survey characteristics, there is a risk with this approach that the survey estimators will be biased. It is therefore a common practice to attempt to compensate for the missing data arising from total nonresponse by some form of weighting adjustment".

Previous analysis (Turner, 1992) has indicated that FCRS direct estimates at the U.S. level for five major variables over the years 1987-1990 are biased downward as follows: three major expense items are biased downward about 10%, while land in farms and number of farms are biased downward about 20%. An inappropriate nonresponse adjustment for the list frame portion of the multiple frame (MF) estimate and undercoverage of farms are major causes of this bias. The 1990 FCRS nonresponse adjustment procedures will be referred to as the current procedure. Currently, FCRS data are collected under the following assumption.

> Assumption a: All nonrespondents would qualify for an interview and would have some positive responses to the survey (i.e., are positive records).

In the supervising and editing manual, field enumerators are instructed to code all out of business (zero) records, who would not qualify for an interview, as respondents. These instructions are intended to ensure that all nonrespondents would qualify for an interview, i.e., have an agricultural operation. Since all interviews are face to face, it is possible to determine if a record is in business or not. The underlying assumption of the current list frame nonresponse adjustment factor, which assumes nonrespondents are similar to all respondents, conflicts with Assumption a because the adjustment assumes nonrespondents can include positive and zero records. The current area frame nonoverlap (NOL) nonresponse adjustment factor, which is applied at the State level, assumes nonrespondents are all positive records and is consistent with Assumption a.

Objectives 1 and 2 of this study involved the application of a simple adjustment (which is consistent with Assumption a) to list frame sample records using the following weighting classes: 1) the design strata, and 2) type/size cells over strata. Objective 3 examined the effect of applying the adjustment at a type/size cell level to area frame NOL records. Weighting classes or cells based on farm type and economic size are intended to provide more homogeneity within weighting classes and heterogeneity across weighting classes than the current classes (strata for the list and States for the area NOL) provide. If the weighting classes are effective in capturing this homogeneity within and heterogeneity across classes with respect to response probabilities, they will help reduce nonresponse bias. Previously

reported control data were used to place nonrespondents into appropriate type/size cells.

EXPANSION FACTORS

The area frame sampling unit is a segment of land, usually about one square mile in area, within a land use stratum. Area frame reporting units are residents of the sampled segments who reported agricultural activity on the previous June Agricultural Survey (JAS), and who are NOL with respect to the FCRS list. The list frame sampling unit is a name on the list sampling frame (LSF). The reporting units are all operating arrangements associated with the sampled names. In the following notation, let h denote a sampling stratum; c denote a type/size weighting cell within a State; and s denote a State.

Furthermore, let
t = h, c, or s as appropriate,
N(t) = number of sampling units in the population denoted by t,
n(t) = number of sampling units sampled from the population denoted by t,
g(t) = number of positive respondent reporting units in t,
f(t) = number of zero respondent reporting units in t,
r(t) = g(t) + f(t) = number of respondent reporting units in t,
e(t) = number of positive nonrespondent reporting units in t,
j(t) = number of zero nonrespondent reporting units in t, and
m(t) = e(t) + j(t) = number of nonrespondent reporting units in t.

Finally, let
$r^*(t)$ = the number of respondent sampling units in t, and
$m^*(t)$ = the number of nonrespondent sampling units in t.

For a sampling unit of the area frame to be classified as nonrespondent, the interviews of all qualifying residents in a land segment must be coded as refusals and inaccessibles. For the list frame, there is usually one reporting unit per sampling unit. If the reporting unit refuses or is inaccessible, then it is a nonrespondent sampling unit. When there is more than one reporting unit associated with a list frame sampling unit, these operating arrangements are referred to as multiple operations. A nonrespondent sampling unit exists in the case of multiple operations when all of the questionnaires corresponding to the sampled name are classified as refusals and inaccessibles.

The current list frame expansion factor is

$$EF = \frac{N(h)}{n(h)} * \frac{n(h)}{r^*(h)}. \qquad (1)$$

The FCRS summary currently has two methods for adjusting the list and area frames for nonresponse due to refusals and inaccessibles. Both procedures are described below. Each sampled unit is initially assigned an original expansion factor that would be applicable if there were no nonresponse, that is, if a usable report was obtained from each reporting unit. For both the area and list frames, the original expansion factor is the first term of Equation (1). The corresponding assumption of this term is the following.

> Assumption b: the n(h) sampled units in stratum h are a simple random sample of sampling units from the N(h) population units in the stratum.

This assumption is clearly true. Since all reporting units do not respond, the original expansion factor is multiplied by an adjustment factor to account for the nonrespondent reporting units. The second term of Equation (1) is based on the following assumption.

> Assumption c: the $r^*(h)$ respondent sampling units in stratum h are a simple random sample from the n(h) sampled units.

If Assumption c were true, then the $m^*(h)$ nonrespondent sampling units would also be a simple random sample of the n(h) sampled units in stratum h. This contradicts Assumption a, where all nonrespondents are assumed to be positive.

The following expansion factor is designed to be consistent with Assumption a and meet Objectives 1, 2, and 3 of this study. The level at which the nonresponse adjustment is calculated, which is represented by x, varies and will be described below.

$$EF = \frac{N(h)}{n(h)} * \frac{g(x)+e(x)}{g(x)}. \qquad (2)$$

The modified list frame expansion factors for Objectives 1 and 2 each have the form of Equation (2) where the nonresponse adjustment factor (term two) is **calculated at the stratum level (x = h) for Objective 1 and at the type/size cell level (x = c) for Objective 2.** These nonresponse adjustment factors are consistent with FCRS Assumption a (i.e. all nonrespondents are positives) since they are based entirely on positive records. The nonresponse adjustment factors of Objectives (1) and (2) are based on the following assumption.

> Assumption d: the positive respondent reporting units {g(h), g(c)} are a simple random sample from the positive reporting units in the stratum or weighting cell.

The current area frame expansion factor has the same form of Equation (2) where the nonresponse adjustment factor (term two) is calculated at the State level (x = s). The nonresponse adjustment factor (term two) of Equation (2) is **applied at the type/size cell level (x = c) for Objective 3.** The nonresponse adjustment factors of the current area frame expansion factor and Objective 3 are both based on Assumption d above.

DATA DESCRIPTION

For this project, 1990 FCRS data were used. The variables that were examined in the analysis are: total expenses, livestock expenses, labor expenses, land in farms, and number of farms. Nine States (Arizona, Colorado, Georgia, Illinois, Kansas, Montana, New York, North Carolina, and Wisconsin) could not be included for the list frame type/size cell analysis because the control data for size were missing. Control data for list records were obtained from the list sampling frame. For area NOL records, control information was collected on the previous June Agricultural Survey. Type categories were collapsed into two classes: crops and livestock. The following five size cells were chosen with respect to annual total gross value of sales: 1} 1 to 9,999, 2} 10,000 to 39,999, 3} 40,000 to 99,999, 4} 100,000 to 249,999, and 5} 250,000 plus. Since variance inflation can result when adjustment factors are not based upon adequate sample sizes, a goal of at least 20 positive respondent records with control data per weighting class was set. (Cox, 1991). To ensure uniform collapsing of cells, a priority scheme and logic flowchart were followed.

RESULTS

Objective 1

Expansions and CV's were obtained for the five variables using the current list frame nonresponse adjustment factor, term two of Equation (1), applied to each of the 281 strata in the 39 States. The modified nonresponse adjustment factor, term two of Equation (2), was applied to each of the 281 strata in the 39 States for Objective 1. The modified nonresponse adjustment factor by stratum produced expansions approximately 9% to 10% higher than the current expansions. Four of the CV's are slightly greater than those of the current method and one CV is the same.

Objective 2

Term two of Equation (2) was applied by type/size cell within State to evaluate Objective 2 for list frame estimates. A total of 212 cells were used over the 39 States. These estimates are 10% to 17% greater than the current estimates. The CV's tend to be slightly larger than those for the unadjusted expansions or for adjusted expansions at the stratum level.

Objective 3

To evaluate Objective 3, records were assigned to area frame NOL type/size cells within State using the same logic used for the list frame records. A total of 68 cells were used for Objective 3. The estimates of Objective 3 are very near the current NOL estimates. Three of the CV's are less than those of the current method and two are greater. Since the percentage change in the estimates is small for Objective 3, these results indicate that application of the nonresponse adjustment for the area frame NOL within cells has negligible effect.

MULTIPLE FRAME RESULTS

List and area NOL results have been considered separately. Multiple frame results show the effect of the list and area NOL results together. Agricultural Statistics Board numbers, which are considered to be truth, exist for number of farms and land in farms. For the three expense items, "Pseudo Board" values (Turner, 1992) were calculated that adjust somewhat for the FCRS undercoverage of farms. The Pseudo Board values represent a minimum value of truth since there are other factors that also contribute to the downward bias. Nonresponse adjusted MF estimates were calculated at the 48 State level using type/size cells

within State for both the list and area NOL indications. The list data for the nine States with unknown size control data were expanded by stratum using the modified nonresponse adjustment factor, since type/size cells could not be created. The probable effect of using the modified nonresponse adjustment by stratum for these nine States on the 48 State MF indications, instead of using the modified nonresponse adjustment by type/size cell within each State, is to bring the indications downward. These nonresponse adjusted MF estimates as well as the current MF estimates arecompared to the Board and Pseudo Board estimates in Table 1. The nonresponse adjusted MF estimates for the expense items closely match their Pseudo Board values, ranging from 3.7% below to 1.3% above. Land in farms adjusted for nonresponse is still biased downward by about 13%. This bias is probably due in part to the tendency of farm operators to underreport total farm acreage (McClung, 1988). However, this bias is about 8 percentage points smaller than the current bias of 21%. This reduction in bias, represented by the last column in Table 1, for land in farms is comparable to the reduction for the expense items, indicating about an 8 to 11 percentage point effect on the MF estimates for these items. One important characteristic of these four items is that approximately 23% of the MF estimates are from the area frame NOL. The reduction in bias for number of farms is only about 4 percentage points, but approximately 58% of the MF estimate is from the area frame NOL. Since the nonresponse adjustment had negligible effect on the area frame NOL, the bias reduction for the MF estimate is also small.

Table 1: 1990 Current MF Estimates and Nonresponse Adjusted MF Estimates Using Type/Size Cells Within State at 48 State Level Compared to 1990 Board and Pseudo Board Estimates.

Item	1990 Board & Pseudo Board Estimates (mil.)	Current MF (% of Board)	Nonresponse Adjusted Type/Size Cells MF (% of Board)	Nonrs. Adjstd. (% of Board) - Current MF (% of Board)
Total Expenses	150,269	87.9%	96.3%	8.4%
Livestock Expenses	16,864	88.9%	97.1%	8.2%
Labor Expenses	14,828	90.1%	101.3%	11.2%
Land in Farms	985	78.8%	86.6%	7.8%
No. of Farms	2.1352	82.1%	85.8%	3.7%

CONCLUSIONS

Results indicated that the largest bias reduction for the list frame portion of the estimate occurred using type/size cells over strata. Evidently, these cells do a more effective job of grouping homogeneous records together than the current design strata. There was little effect, however, from using type/size cells for area frame NOL records primarily because cells could only be created in 17 of the 48 States because of the goal of at least 20 records per cell. A major factor to the remaining downward bias on all five items is the undercoverage of farms by FCRS. The CV's of the nonresponse adjusted estimates increased slightly as compared to the current CV's. This probably reflects more the failure of the variance approximation procedure than the nonresponse adjustment procedures.

RECOMMENDATIONS

Analysis of 1990 data indicated the adjustment should be made using type/size weighting classes within each State for the list frame records. The recommendation for Objective 3 was optional, since the impact of type/size cells within State was negligible on the area side. It was recommended that analysis be conducted on the 1991 FCRS data to determine if type/size weighting classes within each State were needed, or if the list frame strata were adequate weighting classes. The 1992 FCRS used the modified nonresponse adjustment at the design stratum level, since 1991 list frame stratification changes were expected to better account for type and size of farm and since the creation of type/size cells would have added complexity to the summary process. Since the nonresponse adjustment is based on the assumption that all nonrespondents have operating farms, survey training materials and instructions should continue to emphasize that refusal and inaccessible sampling units must be farm operators.

REFERENCES

(1) Cox, Brenda G. (1993), "Weighting Class Adjustments for Nonresponse in Integrated Surveys: Framework for Hog Estimation," SRB Research Report Number SRB-93-03, National Agricultural Statistics Service.

(2) Kalton, Graham and Dalisay Maligalig (1991), "A Comparison Of Methods Of Weighting Adjustment For Nonresponse," 1991 Annual Research Conference Proceedings, Bureau of the Census, pp. 409-428.

(3) McClung, Gretchen (1988), "A Commodity Weighted Estimator," Staff Report No. SRB-88-02, National Agricultural Statistics Service.

(4) National Agricultural Statistics Service (1989), "Interviewer's Manual 1989 Farm Costs & Returns Survey (FCRS)," Author.

(5) National Agricultural Statistics Service (1991), "1990 Farm Costs and Returns Supervising and Editing Manual," Author.

(6) Turner, Kay (1992), "Modification of FCRS Nonresponse Adjustment Procedures," SRB Research Report Number SRB-92-08, National Agricultural Statistics Service.

ACKNOWLEDGMENTS

The author appreciates the significant contributions of time and effort by Bill Iwig and Dave Dillard. Thanks to Fatu Wesley for her valued support and advice, to Jim Burt for variance analysis, to Fred Warren and Martin Ozga for downloading the data, to Dick Clark and Dan Ledbury for supplying FCRS and Board numbers, and to Phil Kott for his valuable review of the original NASS staff report (Turner, 1992). Sincere thanks to everyone who provided assistance on this project.

AN ANALYSIS OF VARIANCE APPROACH TO DEFINING IMPUTATION CELLS FOR A COMPLEX AGRICULTURAL SURVEY

John Amrhein, National Agricultural Statistics Service
NASS/RD, 3251 Old Lee Hwy, Fairfax, VA 22030

KEYWORDS: Analysis of Variance, Wald Statistic, Complex Survey, Imputation

Abstract

Conventional analysis of variance techniques, such as F tests, that are used to determine if subpopulation means are significantly different, rely on the assumption that the observations are independent and identically distributed. It is often the case that survey data, collected using a complex sampling design, violate this assumption. This paper demonstrates the use of an analysis of variance model, adjusted for a complex sampling design, as an effective tool to define imputation cells when adjusting for nonresponse in a complex agricultural survey. A solution to the normal equations is derived in the usual manner. However, a resampling technique is used to obtain a consistent estimate of the covariance matrix. This estimate is then used to calculate a Wald statistic for conducting F tests on testable hypotheses. Examples for several survey items are discussed.

Introduction

The use of models in the analysis of sample survey data continues to be an important area of study. Skinner et al. consolidate much of the work that has been accomplished concerning this issue. This paper presents an application of the analysis of variance model to complex survey data. Although the discussion focuses on employing analysis of variance to define imputation cells, the general method described is appropriate for many survey applications involving an analysis of variance.

The first section of this paper discusses two broad approaches to modelling complex survey data and introduces a general linear model to be considered. The Wald statistic is presented as the conventional test statistic for an analysis of variance. An adjustment to the conventional technique that allows the application of an analysis of variance to non-iid observations is discussed. The next section describes, through an application to a complex survey, how an analysis of variance model can be a useful tool to test the effectiveness of imputation cells that are often used when imputing for survey nonresponse.

The Model and the Test Statistic

Complex survey designs are implemented to increase the precision of estimates when there is knowledge about the underlying structure of the population of interest or to facilitate data collection. In such cases the assumption of independent and identically distributed (iid) observations underlying conventional data analysis techniques, found in many computer analysis packages, is invalid. Ignoring the complex sampling design in favor of the iid assumption during data analysis results in biased estimation of sampling variances. Therefore, an improved analysis can be realized by accounting for the complex design.

Skinner et al. discuss aggregated and disaggregated approaches to modelling complex survey data. The aggregated approach models the survey variables of interest at the population level and accounts for the survey design through adjustments to standard analysis procedures under iid assumptions. The disaggregated approach includes the survey design in the specification of the model. For example, columns of binary variables defining membership in strata or clusters can be included in the matrix of independent regressor variables. A solution to the normal equations can be obtained and linear combinations of the coefficients can be used to test for significant differences of the cluster (or subpopulation) means.

Consider the fixed effects analysis of variance model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \qquad (1)$$

where $\mathbf{y}$ is a vector of values for a measured or observed survey item; $\mathbf{X}$ is a known design matrix; $\beta' = (\mu, \alpha_1, \ldots \alpha_d)$, where d is the number of effects, is a vector of coefficients of unknown value; and $\mathbf{e}$ is a vector of residuals such that $E(\mathbf{e})=\mathbf{0}$ and $V(\mathbf{e})=\mathbf{V}$. Assuming that the residuals are iid Normal random variables, the standard analysis of variance can be conducted. Under these assumptions, the Wald statistic:

$$Q_k=(K'\hat{\beta}-m)'[K'\hat{V}_{\hat{\beta}}K]^{-1}(K'\hat{\beta}-m) \qquad (2)$$

where $\hat{\beta}$ represents a solution to the normal equations, has a central chi-squared distribution with degrees of freedom equal to the rank of K' under the null

hypothesis H_o: $K'\beta=m$, where K' is a contrast matrix whose rows represent testable linear combinations of the coefficients. By correctly specifying K' and setting $m=0$, the significance of the effects in partitioning the variance of the dependent variable can be tested using conventional F tests. Koch et al. present an approach to adjusting conventional analysis of variance techniques for data from complex survey designs; that is, for cases in which the iid assumption is not valid. A further review and example are presented by Freeman where he labels this method the KFF method. In these examples the method of weighted least squares is used to estimate the vector β and a consistent estimator (such as a resampling technique) is used to estimate the covariance matrix for β.

Similarly, Skinner et al. discuss the following procedure, which is followed in this study. Let Π be the diagonal matrix of inclusion probabilities for the sampled units and consider again the model in (1). The weighted least squares estimator for the parameter vector, β, for a model of full rank is

$$\hat{\beta}=(X'\Pi^{-1}X)^{-1}X'\Pi^{-1}Y. \qquad (3)$$

This is the product of the design unbiased, Horvitz-Thompson estimators for $X'X$ and $X'Y$ (Shah et al.). The estimator in (3) is model unbiased under the model in (1) (Skinner et al.) and has true variance

$$V(\hat{\beta}|X)=V_{\hat{\beta}}=(X'\Pi^{-1}X)^{-1}X'\Pi^{-1}V\Pi^{-1}X(X'\Pi^{-1}X)^{-1}. \qquad (4)$$

If the estimator for the covariance matrix given in (4) is consistent, then the techniques described for the iid case can be used for tests of significance. In this case, the Wald statistic in (2) is approximately distributed chi-squared. Standard statistical software can be used to conduct the necessary regression, and various resampling techniques are available for consistent estimation of (4). Rao et al. present some recent work concerning the use of the jackknife, balanced repeated replication and the bootstrap methods for inference with complex survey data. A discussion concerning which method is most appropriate is beyond the scope of this paper. This is one of several areas requiring further investigation.

Application to Mean Imputation

Kalton and Kasprzyk explain how most models underlying imputation or reweighting adjustments for nonresponse fit the form of the general linear model in (1) where the design matrix **X** defines the inclusion of an observation in a reweighting or imputation cell and may also include auxiliary variables, and β is an effects vector. One method of adjusting for nonresponse involves defining population cells in which nonresponse is assumed or known to be ignorable; that is, we want to form cells in which the nonrespondents are considered to be a simple random sample of the original sampled units in that cell and the within-cell variance is small. By conducting an analysis of variance, the contribution of a variable in defining homogeneous groups can be measured by testing if its coefficient is significantly different from zero.

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture collects crop, livestock and grain stock inventory data through a series of sample surveys. NASS draws samples from a list frame and an area frame and conducts concurrent surveys each June. NASS' area frame consists of the land area of the United States. Each county within each state is stratified based on land use and, in the event of agricultural land, percent cultivation. Substratification is performed based on the type of agricultural activity (crop, livestock, etc.). A two (or sometimes three) stage sampling process selects segments of land for enumeration in June. A segment is a cluster of tracts. Tracts are defined as areas of land within a segment under one operating arrangement. Each tract is associated with an operator so that it can be matched against the list frame. Tracts from the area frame sample that were found to be not eligible for sampling from the list frame, labeled non-overlap (NOL) tracts, are identified and used as a measure of the incompleteness of the list. The expanded (weighted) data from NOL tracts are combined with expanded (weighted) data from the list sample to derive multiple frame totals. NOL tracts are restratified based on data collected in the June survey. Based on the new strata, a stratified simple random subsample of the NOL tracts is drawn for each subsequent, or "follow-on", survey in the twelve month survey cycle following the June survey. As in June, the totals from the area subsample are added to totals from the list sample for full population multiple frame totals. Thus, the overall sample design for selecting NOL tracts for follow-on surveys is a two phase stratified design.

NASS uses agricultural statistics districts (geographical delineations) within each state to define imputation cells to adjust for item nonresponse in the NOL sample of the follow-on surveys. This study was initiated to determine the appropriateness of these cell definitions. Because of data abundance and the one phase design, June NOL observations were used in this study rather than follow-on survey observations. The one phase design is easier to mimic when resampling.

The model analyzed in this study is the following:

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (5)$$

where: y_{ij} = value of survey item for unit ij
$i = 1,\ldots,d$ agricultural statistics districts where $1 < d < 9$ for a given state
$j = 1,\ldots,n_i$ observations
μ = the population mean for the state
α_i = the effect of the i^{th} ASD
e_{ij} = the residual for the ij^{th} observation where $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma^2_i$

Although the above model groups the observations into subpopulations (districts), these groups are aggregates of survey design clusters or strata. Therefore, the aggregated approach as described by Skinner et al. was adopted for this study. The objective was to determine if separating the responses into districts aids in partitioning the variance. Results indicating that districts significantly partition the variance of reported values for the survey item (e.g. cropland acres, total number of hogs, etc.) would provide support for the assumption that the cells are homogeneous groups of iid observations and that nonrespondents are a simple random sample within each cell. It is not an objective to derive a predictive model for the dependent variable. Indeed, the analysis of variance models in this study are not full rank. Therefore, the solutions to the normal equations are not unique and cannot be used for prediction.

The bootstrap technique described by Rao et al. was used, with 250 iterations, to estimate the variance given in (4). The null hypothesis was H_o: $K'\beta = m$, where $K' = [\mathbf{0}_{d-1}\ \mathbf{1}_{d-1}\ -\mathbf{1} \cdot \mathbf{I}_{d-1}]$ and $\mathbf{m} = \mathbf{0}$. Index d is the number of districts in the state; i.e. for the fixed effects, α_i, i ranges from 1 to d. Defining K' in this manner tests that all α_i's are equal or, effectually, that all α_i's equal zero. For a good explanation of why this is so, the reader is referred to Searle.

The cell means were tested for significant differences by calculating a p-value which is defined as the probability that a random variable that is distributed $F_{d-1,\, n-(d+1)}$ is greater than the observed value of the Wald statistic divided by its degrees of freedom. That is:

$$\text{p-value} = \text{Prob}[\, F_{d-1,\, n-(d+1)} > Q_k/(d-1)\,].$$

When the denominator degrees of freedom, n-(d+1), is low, using the F distribution as described above offers a refinement over comparing Q_k to a χ^2_{d-1} value (Skinner et al. p. 79). For the hypothesis described above, a small p-value would indicate that the district means are significantly different.

Results obtained in this study indicate that districts aid in the separation of the variance of the agricultural survey items in this study. The analysis was conducted using only positive data. The objective was to test if any difference existed between district means for those farms that had the item of interest. Therefore, for example, when analyzing cropland acres, only those farms with reported cropland acres greater than zero were included. The four survey items that were tested are as follows, with the number of observed p-values that were less than .1 over the number of states included in the analysis given in parentheses: cropland acres (36/48), number of hogs (13/30), on-farm grain storage capacity (29/36) and winter wheat harvested acres (6/15). States were excluded from a given analysis if, as in the cases of Alaska and Hawaii, they are not included in the survey program, or there were too few observations for the survey item. For example, Rhode Island has a small number of hog operations and, therefore, was not included in the analysis of total number of hogs. Also, NOL sampled units are used as a measure of list incompleteness. Therefore, we are dealing with fewer observations than if data from the list frame sample were used.

From these findings, it can be concluded that defining imputation cells based on agricultural statistics districts partitions the population into more homogeneous groups than if the cells were defined at the state level. Cell means were found to be significantly different in enough states that NASS should not collapse imputation cells to the state level. It cannot, however, be concluded from this study that districts partition the population into the most homogeneous groups. There may be other auxiliary variables available that partition the population into more homogeneous groups.

Discussion

A natural question to ask is if any improvement in analysis was realized by accounting for the sample design. Therefore, an analysis of variance was also conducted ignoring the sample design and assuming iid observations. The number of states with an observed p-value of less than .1 over the number of states tested, given in parentheses, is as follows: for cropland acres (21/48), number of hogs (3/30), on-farm grain storage capacity (14/36) and winter wheat harvested acres (8/15). The lower occurrence of significance, at a .1 level, (except for winter wheat harvested acres) suggests that the stratification resulted in more precise estimates. The analysis of variance that accounted for the stratification detected differences that

the analysis assuming independent observations did not. The difference in results is most apparent with the number of hogs survey item. With only three of thirty states showing significance, one would conclude that districts do not aid in partitioning the variance. The estimated design effects in these cases would be less than one.

In conclusion, the strategy outlined here for conducting an analysis of variance is an effective tool that can be used to determine the effectiveness of the imputation cells rather than relying solely on expert opinion, as is often done.

References

Freeman, D. H. Jr. (1988), "Sample Survey Analysis: Analysis of Variance and Contingency Tables," Handbook of Statistics 6, eds. P.R. Krishnaiah and C.R. Rao, New York: North Holland, 415-426.

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

Koch, G. G., Freeman, D. H. Jr. and Freeman, J. L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys," Int. Stat. Rev., 43:1, 59-78.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992), "Some Recent Work on Resampling Methods for Complex Surveys," Survey Methodology, 18:2, 209-217.

Searle, S. R. (1971), Linear Models, New York: John Wiley and Sons, p. 240.

Shah, Babubhai V., Holt, M. M. and Folsom, R. E. (1977), "Inference About Regression Models From Sample Survey Data," Bulletin of the International Statistical Institute, 41:3, 43-57.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989), Analysis of Complex Surveys, New York: John Wiley and Sons.