REDUCING RESPONDENT BURDEN ON AN AGRICULTURAL SURVEY

Barry L. Ford, U.S. Department of Agriculture
R. R. Hocking, Mississippi State University
Anne Coleman, Mississippi State University

PURPOSE:

The purpose of this paper is to discuss certain sampling plans in the context of respondent burden and their application to a large scale agricultural survey. These survey procedures attempt to lower the respondent burden by utilizing multivariate regression.

INTRODUCTION:

Surveys sometimes use questionnaires which require a great deal of time for the respondent to complete because of a large number of questions. This situtation occurs not only because the subject may be complex but also because the additional cost of asking 100 questions, for example, instead of 10 questions is relatively small. Once a survey is designed and implemented, budget constraints often dictate obtaining as much information as possible from each respondent.

For example, the U.S. Department of Agriculture (USDA) runs a survey to estimate the expenditures of farmers when producing their crops and livestock. The questionnaire for this survey contains over 600 questions--questions ranging from expenditures on seeds to expenditures for diesel oil to expenditures for wire. The amount of detail needed on farm expenditures is so great that the ideal solution is to use a set of independent surveys, each of which obtains information on a different part of a farmer's expenditures. However, the number of farmers contacted would then be enourmous, and thus, the cost of getting expenditure information would also be enourmous. Therefore, the USDA carries out one survey with a questionnaire which asks about all expenditures but may require five or six hours to complete.

Recently the burden placed on respondents by surveys has received more attention because of complaints by the public about the large number of surveys (both in government and in private industry) and the inconvenience of these surveys. These complaints cause the USDA to be more aware of the respondent burden caused by its surveys. Respondent burden now has no exact definition in the literature but is often used in a general sense to signify the time and other costs to each respondent answering a survey questionnaire. When the USDA obtains expenditure information by avoiding a set of surveys with a small number of questions on each survey and by using one survey with a large number of questions, the USDA changes a small respondent burden on many people to a large respondent burden on a few people. One could also view this trade-off as an exchange of monetary cost to the government for an increase in respondent burden. Since everyone benefits from a reduction in government costs, everyone benefits from this exchange except that sample of farmers who must give the time and effort to answer over 600 questions.

Can some survey procedure reduce the burden placed on this sample of farmers without greatly increasing costs to the survey organization? Under certain conditions the answer is yes. The usual circumstance is that the questions on a survey all relate to one specific area (eg. expenditures of farmers) so that the variables obtained from each respondent are highly correlated with each other. Therefore, the correlation structure of the variables can be used to avoid asking all of the respondents every question and still achieve a satisfactory level on the estimates.

SAMPLE DESIGNS:

Suppose the goal is to make estimates on a large set of variables which are denoted by the vector $\underline{w}$. If $\underline{w}$ is composed of two sets of variables, $\underline{w} = \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix}$, where $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ and $\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_q \end{bmatrix}$, then let plan A be the use of two independent samples to measure $\underline{x}$ and $\underline{y}$. If the size of each sample size is n, then the data would be of the form:

$\underline{x}_1$	
$\underline{x}_2$	
$\vdots$	
$\underline{x}_n$	
	$\underline{y}_{n+1}$
	$\underline{y}_{n+2}$
	$\vdots$
	$\underline{y}_{2n}$

where the subscript refers to data collected on a particular observation (i.e. respondent).

To minimize survey cost an organization might obtain all of the variables on each respondent. One can call this strategy plan B:

$\underline{x}_1$	$\underline{y}_1$
$\underline{x}_2$	$\underline{y}_2$
$\vdots$	$\vdots$
$\underline{x}_n$	$\underline{y}_n$

A simple method of using the correlation structure to avoid asking all n respondents every question is to use a multivariate regression estimator in a double sampling approach. The variables in $\underline{x}$ should be correlated with the variables composing $\underline{y}$, and thus $\underline{x}$ can fairly well

predict $\underline{y}$ by using a multivariate linear regression. This sample design, plan C, would obtain values for both $\underline{x}$ and $\underline{y}$ for the first n_1 observations and then obtain values of only $\underline{x}$ on the remaining $(n-n_1)$ observations:

$\underline{x}_1$	$\underline{y}_1$
$\underline{x}_2$	$\underline{y}_2$
⋮	
$\underline{x}_{n_1}$	$\underline{y}_{n_1}$
$\underline{x}_{n_1+1}$	
$\underline{x}_{n_1+2}$	
⋮	
$\underline{x}_n$	

To estimate the mean vector of $\underline{w}$, $\underline{\mu}_w$; one estimates the two components--$\underline{\mu}_w = \begin{bmatrix} \underline{\mu}_x \\ \underline{\mu}_y \end{bmatrix}$. Of course, the estimator of $\underline{\mu}_x$ is the average of n vectors, i.e.:

$$\hat{\underline{\mu}}_x = \frac{\sum_{i=1}^{n} \underline{x}_i}{n} .$$

To estimate $\underline{\mu}_y$, however, one must apply a multivariate linear regression estimator: If $\overset{*}{\underline{x}}$ and $\overset{*}{\underline{y}}$ are averages using only the first n respondents, then:

$$\hat{\underline{\mu}}_y = \overset{*}{\underline{y}} + B' (\hat{\underline{\mu}}_x - \overset{*}{\underline{x}})$$

where B is a matrix of regression coefficient given by $B = \Sigma_{xx}^{-1} \Sigma_{xy}$ and:

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma'_{xy} & \Sigma_{yy} \end{bmatrix}$$

is the variance - covariance matrix of $\underline{w}$. The variance of $\underline{\mu}_x$ is, of course, :

$$\text{Var } (\hat{\underline{\mu}}_x) = \frac{1}{n} \Sigma_{xx}.$$

Given a known B, the variance of $\hat{\underline{\mu}}_y$ is [1]:

$$\text{Var } (\hat{\underline{\mu}}_y) = \frac{1}{n} \Sigma_{yy} - \frac{(n-n_1)}{nn_1} B' \Sigma_{xx} B.$$

Obviously, plan C will cause the standard errors of the variables in $\underline{y}$ to be larger than those in plan A or B. A re-expression [1] of Var $(\hat{\underline{\mu}}_y)$ is:

$$\text{Var } (\hat{\mu}_i) = \left[1 - \frac{(n-n_1)}{n} R_i^2 \right] \sigma_i^2/n_1$$

for i = 1, 2, ...q

where $\hat{\mu}_i$ is the estimate of $E(y_i)$. One can easily observe that the change in the level of the variance of the estimates is closely related to R_i, the multiple correlation coefficient for variable i. Thus, if $\underline{x}$ is a fairly good predicator of $\underline{y}$, there is little loss in accuracy. The R_i^2 (for i = 1, 2, ..., q) should determine whether one should adopt plan C over plan B.

What is the respondent burden in plans A, B, and C? There are several different ways one might wish to measure respondent burden, but for the purposes of this paper we simply adopt the number of questions per respondent. If $\mathcal{B}$ represents respondent burden, one has:

$$\mathcal{B}_A = \frac{\frac{1}{2}(np + nq)}{n}$$

$$\mathcal{B}_B = \frac{np + nq}{n}$$

$$\mathcal{B}_C = \frac{np + n_1 q}{n}$$

The enumerators of the burdens are written so that obviously $\mathcal{B}_A < \mathcal{B}_B$ if p and q are non-zero. Of course, $\mathcal{B}_A$ would be even smaller and the cost of plan A even larger if more than two samples were used. Although plan C would not differ much in cost from plan B, its respondent burden is less than plan B. However, plan C does suffer a loss in the accuracy of estimating $\underline{\mu}_y$ -- a loss which depends on R_i (i = 1, 2, ..., q). Because n_1 can be smaller when the R_i's are larger, *the respondent burden can be decreased when the* R_i *are larger.*

Given a large number of variables, one might wish to elaborate on this double sampling scheme of plan C by separating the dependent variables into several closely related groups. For example, in the USDA expenditure survey there might be two groups of dependent variables:

group I : dependent variables relating to seed and plant expenses.

group II: dependent variables relating to fertilizers and pesticides.

The variables in $\underline{x}$ should be highly related to the variables in both groups I and II. For instance, some of the variables in $\underline{x}$ might be crop acreages. Groups I and II can be obtained from different respondents in the sample. Rather than just a $\underline{y}$ vector, there is a $\underline{y}$ vector and a $\underline{z}$ vector, and the sample design (called plan D) is:

$\underline{x}_1$	$\underline{y}_1$
$\underline{x}_2$	$\underline{y}_2$
$\vdots$	$\vdots$
$\underline{x}_{n_1}$	$\underline{y}_{n_1}$
$\underline{x}_{n_1+1}$	$\underline{z}_{n+1}$
$\underline{x}_{n_1+2}$	$\underline{z}_{n+2}$
$\vdots$	$\vdots$
$\underline{x}_n$	$\underline{z}_n$

Plan D is still a double sampling scheme, but there are now two sets of dependent variables. One has $\underline{w} = \begin{bmatrix} \underline{x} \\ \underline{y} \\ \underline{z} \end{bmatrix}$ and $\underline{\mu}_w = \begin{bmatrix} \underline{\mu}_x \\ \underline{\mu}_y \\ \underline{\mu}_z \end{bmatrix}$.

Also, Σ, the variance - covariance matrix of $\underline{w}$; is now:

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{xz} \\ \Sigma'_{xy} & \Sigma_{yy} & \Sigma_{yz} \\ \Sigma'_{xz} & \Sigma'_{yz} & \Sigma_{zz} \end{bmatrix}$$

The estimators are of the same form as those in plan C:

$$\hat{\underline{\mu}}_x = \sum_{i=1}^{n} \underline{x}_i / n$$

$$\hat{\underline{\mu}}_y = \overset{*}{\underline{y}} + B'_y (\hat{\underline{\mu}}_x - \overset{*}{\underline{x}}_y)$$

$$\hat{\underline{\mu}}_z = \overset{*}{\underline{z}} + B'_z (\hat{\underline{\mu}}_x - \overset{*}{\underline{x}}_z)$$

where:

$$B_y = \Sigma_{xx}^{-1} \Sigma_{xy} \qquad \overset{*}{\underline{x}}_y = \frac{\sum_{i=1}^{n_1} \underline{x}_i}{n_1}$$

$$B_z = \Sigma_{xx}^{-1} \Sigma_{xz} \qquad \overset{*}{\underline{z}} = \frac{\sum_{i=n_1}^{n} \underline{z}_i}{(n-n_1)}$$

$$\overset{*}{\underline{y}} = \frac{\sum_{i=1}^{n_1} \underline{y}_i}{n_1} \qquad \overset{*}{\underline{x}}_z = \frac{\sum_{i=n_1+1}^{n} \underline{x}_i}{(n-n_1)}$$

The variances of $\hat{\underline{\mu}}_x$, $\hat{\underline{\mu}}_y$ and $\hat{\underline{\mu}}_z$ are also of the same form as in plan C.

If $n_2 = (n-n_1)$, $\underline{x}$ has p variables, $\underline{y}$ has q variables, and $\underline{z}$ has r variables, then:

$$\mathcal{B}_D = \frac{n_1 (p+q) + n_2 (p+r)}{n} = \frac{np + n_1 q + n_2 r}{n}.$$

(When $n_1 = n_2$, then $\mathcal{B}_D = \mathcal{B}_C$.) The respondent burden in plan D does not differ greatly from plan C. However, the respondent burden in plan D is spread more evenly throughout the sample. Plan C has n_1 respondents answering all the questions and the rest of the sample answering a few questions while plan D has all respondents answering about the same number of questions.

Studying plan D, one might consider whether correlations exist between $\underline{y}$ and $\underline{z}$ which might be useful. If they do exist, one might want to ask a very small subgroup in the sample all of the questions in order to estimate Σ and use the information in Σ_{yz}. Thus, one has plan E:

Group 1	$\underline{x}_1$	$\underline{y}_1$	$\underline{z}_1$
	$\underline{x}_2$	$\underline{y}_2$	$\underline{z}_2$
	$\vdots$	$\vdots$	$\vdots$
	$\underline{x}_{n_1}$	$\underline{y}_{n_1}$	$\underline{z}_{n_1}$
Group 2	$\underline{x}_1$	$\underline{y}_1$	
	$\underline{x}_2$	$\underline{y}_2$	
	$\vdots$	$\vdots$	
	$\underline{x}_{n_2}$	$\underline{y}_{n_2}$	
Group 3	$\underline{x}_1$		$\underline{z}_1$
	$\underline{x}_2$		$\underline{z}_2$
	$\vdots$		$\vdots$
	$\underline{x}_{n_3}$		$\underline{z}_{n_3}$

To discuss the estimators for plan E, one must discuss a general class of sample designs studied by Hocking, Hartley, et al. [1, 2, 3]. The USDA calls sample designs in this class pattern samples because they depend on a pattern of missing values for certain observations. The observations in a pattern sample can be divided into groups based upon which variables are observed and which are not. These groups are called pattern groups. The first pattern group in plan E is a complete group--all variables are observed. The second group has only $\underline{x}$ and $\underline{y}$ observed, and third group has only $\underline{x}$ and $\underline{z}$ observed. A pattern design is called nested when the pattern groups can be arranged so that the variables in each pattern group are a subset of the preceding group. Plan C is a nested design, but plans D and E are not.

Each pattern group can be represented by a design matrix. A design matirx is composed of 0's and 1's to signify which variables are being observed. For example, if pattern group 1 is composed of four variables, its design matrix, D_1, is the identity matrix:

$$D_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

If the second pattern group only contains variables 2 and 4, then:

$$D_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Pattern groups may contain variables which are totals of other variables. For instance, the third group might contain variable 1 and the sum of variables 3 and 4:

$$D_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

(Note that the rank of a design matrix can not be less than the number of rows.)

If the data are separated into pattern groups, then in the t^{th} group there are n_t observations which yield an estimate of a mean vector, $\hat{\underline{\mu}}_t$, of the variables in that group and $\hat{\Sigma}_t$, the estimated variance - covariance matrix among those variables. Under the assumption that the first group is a complete group, i.e. $D_1 = I$, the maximum likelihood estimators are [1]:

$$\overset{*}{\underline{\mu}} = \left[I + \sum_{t=2}^{T} \frac{n_t}{n_1} \overset{*}{\Sigma} D_t' \overset{*}{\Sigma}_t^{-1} D_t \right]^{-1} * \left[\hat{\underline{\mu}}_1 + \sum_{t=2}^{T} \frac{n_t}{n_1} \overset{*}{\Sigma} D_t' \overset{*}{\Sigma}_t^{-1} \hat{\underline{\mu}}_t \right]$$

$$\overset{*}{\Sigma} = \left[I + \sum_{t=2}^{T} \frac{n_t}{n_1} \overset{*}{\Sigma} D_t' \overset{*}{\Sigma}_t^{-1} D_t \right]^{-1} * \left[\frac{1}{n} M_1 + \sum_{t=2}^{T} \frac{1}{n_1} \overset{*}{\Sigma} D_t' \overset{*}{\Sigma}_t^{-1} M_t \overset{*}{\Sigma}_t^{-1} D_t \overset{*}{\Sigma} \right]$$

where:

$$\overset{*}{\underline{\mu}}_t = D_t \overset{*}{\underline{\mu}}$$

$\overset{*}{\Sigma}_t$ = the sub-matrix of $\overset{*}{\Sigma}$ pertaining to the variables in the t^{th} group

$$H_t = (\hat{\underline{\mu}}_t - \overset{*}{\underline{\mu}}_t)(\hat{\underline{\mu}}_t - \overset{*}{\underline{\mu}}_t)'$$

$$M_t = n_t(\hat{\Sigma}_t + H_t).$$

Equations I and II must be solved iteratively:

1: calculate $\overset{*}{\underline{\mu}}$ by using the variance - covariance matrix from the first pattern group as $\overset{*}{\Sigma}$.

2: use the $\overset{*}{\underline{\mu}}$ from step 1 to calculate H_t and M_t in order to form a new $\overset{*}{\Sigma}$.

3: return to step 1 and use the new estimate of $\overset{*}{\Sigma}$ to make a new estimate of $\overset{*}{\underline{\mu}}$.

4: keep cycling through steps 1-3 until a tolerance level is reached.

The maximum likelihood equations actually have a more general form which is not dependent on $D_1 = I$, but because of the additional notation involved, this form is not discussed in this paper. The computer programs which perform the iterations (written by Anne Coleman [1, Attachment 2]) rely on $D_1 = I$.

Covergence in the iteration process is not assured in general, but in practical applications convergence has taken less than ten iterations. The variance - covariance matrix of $\overset{*}{\underline{\mu}}$ is estimated by:

$$W = \sum_{t=1}^{T} D_t' W_t D_t$$

where

$$W_t = n_t \overset{*}{\Sigma}_t^{-1}.$$

Since there is no explicit formulation for the standard error of a particular variable, probably, the best method of seeing whether gains are likely in using plan E over plan D is to compare the multiple correlation coefficients. For the variables in y, for example, one can compare the multiple correlation coefficient when the variables in $\underline{x}$ are used as independent variables to the multiple correlation coefficient when $\underline{x}$ and $\underline{z}$ are used as independent variables. Although not exactly correct, this approach should be a good indication of the improvement one might expect in adopting plan E over plan D.

The respondent burden of plan E is a slight increase of plan D if the percentage of respondents in the first pattern group is small. The burden is:

$$\mathcal{B}_E = \frac{n_1 (p+q+r) + n_2 (p+q) + n_3 (p+r)}{n}$$

$$= \frac{np + (n_1+n_2) q + (n_1+n_3) r}{n}.$$

APPLICATION:

The purpose of this section is to compare a measure of respondent burden and the coefficients of variation resulting from sampling plans A, B, C, D, and E by using data from an agricultural survey. A complete data set from this survey was obtained under sampling plan B, but the values of certain variables in different sample units were simulated as missing in order to produce plans D and plan E.

The agricultural survey used as an illustration is a survey which estimates the various costs to farmers of growing agricultural products. This survey uses an elaborate questionnaire to obtain values for over 600 variables. This paper will actually restrict itself to a subset of 60 variables because of the computer cost involved. Although the agricultural survey is a complex sample design in practice, for illustrative purposes the data set is used as though it were from a simple random sample. Further work is planned in extending results to more complex sample designs.

Table 1 compares the respondent burden and the average coefficient of variation in plans A, B, C, D, and E. The measure of the respondent burden is the average number of variables obtained

from each respondent, i.e. sample unit.

Plan A is obviously the optimum plan except that it costs approximately twice as much as the other plans. Thus, plan B is better than plan A, but the respondent burden is double. Plan C is better than plan B because there is only a small increase in the average coefficient of variation but the respondent is reduced by 35%. Plan D is better than plan C because it spreads the respondent burden more evenly across the entire sample. Plan E does decrease the average coefficient of variation slightly, but at a slight increase in both respondent burden and its spread. When one also consideres the more complicated estimation process involved with plan E, plan D is the most attractive sampling scheme.

Table 1: A comparison of five sampling plans in relation to data from an agricultural survey.

Plan	Respondent Burden	Spread of the Respondent Burden (Standard Deviation)	Average Coefficient of Variation	Approx. Cost
A	30	0	0.17	2($k)
B	60	0	0.17	$k
C	39	21	0.21	$k
D	39	1	0.21	$k
E	41	7	0.20	$k

BIBLIOGRAPHY:

[1] Hocking, R. R. Final Report on Design of Sample Surveys to Reduce Respondent Burden. 1977. (A report written by contract with Research Division Economics, Statistics, and Cooperatives Service, Room 4818, South Building, U.S. Department of Agriculture, Washington, D.C. 20250.)

[2] Hartley, H. O. and Hocking, R. R. "The Analysis of Incomplete Data," Biometrics, Volume 27, pages 783 - 824. 1971.

[3] Hocking, R. R., Huddleston, H. F., and Hunt, H. H. "A Procedure for Editing Survey Data," Journal of the Royal Statistical Society - Series C, Volume 23, No. 2 pages 121 - 133. 1974.