MODEL-BASED SAMPLE SELECTION AND PREDICTION PROCEDURES FOR TWO-STAGE SAMPLING

Nancy J. Carter, California State University, Chico
and G. David Faulkenberry, Oregon State University
Nancy J. Carter, Dept. of Math and Statistics, CSUC, Chico, CA 95929-0525

Key Words and Phrases: Superpopulation, Variate values, Simulation

ABSTRACT

The problem of sample selection when predicting variate values for all individual units in a finite population based on a sample of some of the units is investigated. A superpopulation model-based prediction approach is proposed for the two-stage sampling problem. An iterative sample-selection procedure influenced by the consideration of minimizing the prediction errors with respect to the model is discussed. An example illustrates and evaluates the proposed procedures. Further research subjects and possible problems are discussed.

1. INTRODUCTION

The use of a superpopulation model-based prediction approach to two-stage sampling is not new. For example, Royall (1976) used a linear least-squares prediction approach to two-stage sampling where the goal was to predict the population total over all units. What is different about the work presented in this paper is that, while past work has concentrated on predicting one value (say a mean or total) over all units, the goal of this research is to derive a prediction for each individual population unit based on a sample of only some of the units. The approach used is to assume an underlying superpopulation model and to select the sample units and derive the individual population unit predictors based on this model.

2. MODEL AND DEVELOPMENT OF PREDICTORS

2.1 Model, Definitions, and Notation

Consider a finite population of N identifiable units (assume N is a known integer). Associated with unit i is the random variable Y_i. The joint distribution of $Y_1,\dots,Y_N$ will be denoted by ξ. Also associated with unit i are p known auxiliary variables $X_{i1},\dots,X_{ip}$. Let $\underset{\sim}{X}_i' = (X_{i1},\dots,X_{ip})$ denote the vector of auxiliary variables for unit i. The independent random variables $Y_1,\dots,Y_N$ are assumed to be distributed so that $E_\xi(Y_i) = \underset{\sim}{X}_i'\underset{\sim}{\beta}$ and $\mathrm{Var}_\xi(Y_i) = \sigma^2$ (assume homogeneous variances). It is assumed that σ^2 and $\underset{\sim}{\beta}' = (\beta_1,\dots,\beta_p)$ are unknown constants.

Contained within unit i are M_i subunits $i1,\dots,iM_i$ (M_i is a known integer). Associated with subunit ij is the random variable Z_{ij} where Z_{ij} represents some characteristic of interest for $i=1,\dots,N$ and $j=1,\dots,M_i$. The independent random variables Z_{ij} are distributed so that $E_\xi(Z_{ij}) = \mu_i$ and $\mathrm{Var}_\xi(Z_{ij}) = \sigma_i^2$.

It is assumed that $Y_i = \sum_{j=1}^{M_i} Z_{ij}$. Thus, the random variable Y_i represents the total over unit i for some characteristic of interest. Since $Y_i = \sum_{j=1}^{M_i} Z_{ij}$, it follows that $E_\xi(Y_i) = M_i\mu_i$. But $E_\xi(Y_i) = \underset{\sim}{X}_i'\underset{\sim}{\beta}$ which implies $M_i\mu_i = \underset{\sim}{X}_i'\underset{\sim}{\beta}$ and $\mu_i = \frac{\underset{\sim}{X}_i'\underset{\sim}{\beta}}{M_i}$. Similarly, $\mathrm{Var}_\xi(Y_i) = M_i\sigma_i^2$ and $\mathrm{Var}_\xi(Y_i) = \sigma^2$. Thus, $\sigma_i^2 = \frac{\sigma^2}{M_i}$.

In drawing the sample, elements are selected for observation in a two-stage procedure. First, a sample s of n units from the N is selected. Next, from the M_i subunits in first stage unit i, a subsample s_i of size m_i is chosen at random. Without loss of generality, assume the units are arranged so that the first n are sample units and the remaining N-n are nonsample units. Further assume there is a fixed total sample size, m. Then $\sum_{i=1}^{n} m_i = m$.

Since units 1,...,n are subsampled, it is necessary to estimate Y_i for these units based on the sample. The estimator chosen for use is $\hat{Y}_i = \sum_{j\in s_i} Z_{ij} + (M_i - m_i)\overline{Z}_i = M_i\overline{Z}_i$ where $\overline{Z}_i = \frac{\sum_{s_i} Z_{ij}}{m_i}$. Under the assumed model, $E_\xi(\hat{Y}_i) = \underset{\sim}{X}_i'\underset{\sim}{\beta}$ and $\mathrm{Var}_\xi(\hat{Y}_i) = \frac{M_i\sigma^2}{m_i}$.

Following the notation of Royall (1976), let $\underset{\sim}{X}_I$ denote the n x p matrix of auxiliary variables and $\underset{\sim}{V}_I$ the n x n covariance matrix associated with the n sample units. Similarly, denote by $\underset{\sim}{X}_{II}$ and $\underset{\sim}{V}_{II}$ the corresponding matrices for the N-n nonsample units. Let $\underset{\sim}{V}_{II,I}$ be the (N-n) x n matrix of covariances between nonsample and sample units. Denote by $\underset{\sim}{Y}$ the N x 1 vector of random variables $Y_1,\dots Y_N$.

If $\underset{\sim}{Y}$ is arranged so that the first n units are $Y_1,\dots,Y_n$ and the remaining N-n are $Y_{n+1},\dots Y_N$, the model states that $E_\xi(Y) = \underset{\sim}{X}\underset{\sim}{\beta}$ and $\mathrm{Cov}_\xi(Y) = \underset{\sim}{V}$ where

$$\underset{\sim}{Y} = \begin{bmatrix} \underset{\sim}{Y}_I \\ \underset{\sim}{Y}_{II} \end{bmatrix}, \quad \underset{\sim}{X} = \begin{bmatrix} \underset{\sim}{X}_I \\ \underset{\sim}{X}_{II} \end{bmatrix},$$

$$\utilde{V} = \begin{bmatrix} \utilde{V}_I & \utilde{V}'_{II,I} \\ \utilde{V}_{II,I} & \utilde{V}_{II} \end{bmatrix}$$

and $\utilde{\beta}$ is defined above. That is,

$$\utilde{Y}_I = \begin{bmatrix} Y_1 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} \quad \text{and}$$

$$\utilde{X}_I = \begin{bmatrix} X_{11} & X_{12} & \cdot & \cdot & \cdot & X_{1p} \\ X_{21} & X_{22} & \cdot & \cdot & \cdot & X_{2p} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ X_{n1} & X_{n2} & \cdot & \cdot & \cdot & X_{np} \end{bmatrix}$$

For this situation,

$$\utilde{V}_I = \sigma^2 \begin{bmatrix} \frac{M_1}{m_1} & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \frac{M_2}{m_2} & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ 0 & & & & & \frac{M_n}{m_n} \end{bmatrix}$$

$= \sigma^2 \utilde{A}_I$.

Since it is assumed $\text{Cov}_\xi(\hat{Y}_i, Y_j) = 0$ for all $i = 1,\ldots,n$ and $j= n+1,\ldots,N$, $\utilde{V}_{II,I}$ is a (N-n) x n matrix of zeros. Therefore,

$$\utilde{V} = \sigma^2 \begin{bmatrix} \frac{M_1}{m_1} & 0 & & & & & 0 \\ 0 & \frac{M_2}{m_2} & 0 & & & & 0 \\ & & \cdot & & & & \\ \cdot & & & \frac{M_n}{m_n} & & & \\ & & & & 1 & & \\ & & & & & \cdot & \\ 0 & & & & & & 1 \end{bmatrix}$$

$= \sigma^2 \utilde{A}$.

Therefore, under model ξ, the weighted least-squares estimate of $\utilde{\beta}$ is $\hat{\utilde{\beta}} = (\utilde{X}'_I \utilde{V}_I^{-1} \utilde{X}_I)^{-1} \utilde{X}'_I \utilde{V}_I^{-1} \utilde{Y}_I$ or equivalently, $\hat{\utilde{\beta}} = (\utilde{X}'_I \utilde{A}_I^{-1} \utilde{X}_I)^{-1} \utilde{X}'_I \utilde{A}_I^{-1} \utilde{Y}_I$.

2.2 Development of Predictors and Error Variance of Predictors for the Nonsample Units

The totals for the nonsample units are $Y_j (j=n+1,\ldots,N)$. Since $E_\xi(Y_j) = \utilde{X}'_j \utilde{\beta}$ and $\hat{\utilde{\beta}}$ is ξ-unbiased for $\utilde{\beta}$, the natural predictor for Y_j is $\utilde{X}'_j \hat{\utilde{\beta}}$.

The model-based approach taken suggests the appropriate measure of uncertainty for the predictors is the error variance with respect to ξ. Therefore, it is assumed the samples s and s_i are fixed for $i=1,\ldots n$ and, for $j = n+1,\ldots,N$ the following quantity is considered:

$$E_\xi(Y_j - \utilde{X}'_j \hat{\utilde{\beta}})^2 = \text{Var}_\xi(Y_j) + 2\text{Cov}_\xi(Y_j, \utilde{X}'_j \hat{\utilde{\beta}}) + \text{Var}_\xi(\utilde{X}'_j \hat{\utilde{\beta}}).$$

Since j is a nonsample unit and $\hat{\utilde{\beta}}$ is computed from the sample units, it must be true that $\text{Cov}_\xi(Y_j,\ \utilde{X}'_j \hat{\utilde{\beta}}) = 0$. Therefore

$$E_\xi(Y_j - \utilde{X}'_j \hat{\utilde{\beta}})^2 = \sigma^2 [1 + \utilde{X}'_j (\utilde{X}' \utilde{A}_I^{-1} \utilde{X}_I)^{-1} \utilde{X}_j]$$

2.3 Development of Predictors and Error Variance of Predictors for the Sample Units.

For the sample units the total is

$$Y_k = \sum_{j=1}^{M_k} Z_{kj} \qquad (k=1,\ldots,n).$$

For unit k, the sample estimate of Y_k is $\hat{Y}_k$. The $\hat{Y}_k$ values are used to construct $\hat{\beta}$, the model-based estimate of $\underset{\sim}{\beta}$. Note that the sample units are composed of two kinds of subunits: those that are sampled and those that are not sampled. To take advantage of the assumed model relationships, a weighted predictor for sample unit k, $\hat{Y}_k^*$, is used where $\hat{Y}_k^* = w_k\hat{Y}_k + (1 - w_k)\underset{\sim}{X}'_k\hat{\underset{\sim}{\beta}}$ and where w_k is the weight given Y_k. The weight, w_k, that minimizes $E_\xi(Y_k - \hat{Y}_k^*)^2$ is $w_k = \frac{m_k}{M_k}$.

Substituting $w_k = \frac{m_k}{M_k}$ into the $\hat{Y}_k^*$ formula gives $\hat{Y}_k^*$ $= \left(\frac{m_k}{M_k}\right)Y_k + \left(1 - \frac{m_k}{M_k}\right)\underset{\sim}{X}'_k\hat{\underset{\sim}{\beta}} = \sum_{s_k} Z_{jk} + \left(1 - \frac{m_k}{M_k}\right)\underset{\sim}{X}'_k\hat{\underset{\sim}{\beta}}$.

When $w_k = \frac{m_k}{M_k}$ is substituted into $E_\xi(Y_k - \hat{Y}_k^*)^2$ the result is $E_\xi(Y_k - \hat{Y}_k^*)^2 =$

$$\sigma^2\left(1 - \frac{m_k}{M_k}\right)[1 + \left(1 - \frac{m_k}{M_k}\right)\underset{\sim}{X}'_k(\underset{\sim}{X}'_I\underset{\sim}{A}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_k].$$

3. SAMPLE SELECTION PROCEDURE

In theory, it would be desirable to select the sample in such a way as to give small prediction errors

$$E_\xi(Y_j - \underset{\sim}{X}'_j\hat{\underset{\sim}{\beta}})^2 \text{ for } j = n+1,\ldots,N \text{ and } E_\xi(Y_k - \hat{Y}_k^*)^2$$

for i=1,...,n.

However, looking at these error variances, it is seen that

$$E_\xi(Y_j - \underset{\sim}{X}'_j\hat{\beta})^2 = \sigma^2[1 + \underset{\sim}{X}'_j(\underset{\sim}{X}'_I\underset{\sim}{A}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_j] \qquad (3.1)$$

and $E_\xi(Y_k - \hat{Y}_k^*)^2 =$

$$\sigma^2\left(1 - \frac{m_k}{M_k}\right)[1 + \left(1 - \frac{m_k}{M_k}\right)\underset{\sim}{X}'_k(\underset{\sim}{X}'_I\underset{\sim}{A}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_k] \qquad (3.2)$$

where $\underset{\sim}{A}_I^{-1}$ depends on the m_k for k=1,...,n. Since both equations(3.1) and (3.2) depend on the m_k the following sample selection scheme is proposed.

First choose some starting size for the number of units selected, say n_1. Next use the procedure described in the paper titled "A Model-Based Sample Selection Procedure For One-Stage Sampling" by Carter and Faulkenberry (1989). When using the sample selection procedure for one-stage sampling to pick the n_1 units, assume no subsampling (i.e., $m_k = M_k$ for $k=1,\ldots,n_1$). The next step is to determine how to choose the m_k such that $E_\xi(Y_k - \hat{Y}_k)^2$ is minimized for $k=1,\ldots,n_1$ with the restriction that $\sum_{k=1}^{n_1} m_k = m$ (where m is the fixed total sample size). The method of Lagrange multipliers was used to accomplish this goal and the result obtained was

$$m_k = \frac{\sqrt{M_k}}{\sum_{k=1}^{n_1}\sqrt{M_k}}\, m$$

Now equations (3.1) and (3.2), the error variances of the predictors, can be computed. Once the error variances are computed, the maximum over the $N - n_1$ nonsample unit error variances can be determined. These maximum error variances are used to "judge" the sampling plan. The goal now is to minimize the maximum over these equations with perhaps the added constraint of a minimum allowable error on the nonsample units. To achieve this goal, the entire sample selection process is repeated beginning with n_2 units where $n_1 < n_2 < N$. By continuing in this way a sequence of sampling plans, $P_1, P_2, \ldots$ is derived. This sequence is continued until the maximum over the N equations (3.1) and (3.2) is no longer being reduced or, if there are other constraints, until all of these are met. It should be noted that since the total sample size is fixed, when n is increased, the number of subunits sampled is decreased. Hence, increasing n increases the (3.2) values.

The amount of increase in n_i each time is a matter of interpretation. If the maximum over (3.1) and (3.2) is changing a lot from P_{i-1} to P_i then a larger change in n_i should be made than if there is little reduction taking place. If it happens that n_1 was chosen so that at plan P_2 (using n_2) the maximum is increasing, reduce n_1 to n'_2 and repeat the procedure suggested above only reducing n at each step until the condition in step 6 is met.

This procedure does require a lot of computation but it is all relatively easy to do. All of the computations mentioned above are simple and straightforward.

4. EXAMPLE: AN EVALUATION OF THE SAMPLE SECTION AND PREDICTION PROCEDURES USING ACTUAL DATA

Actual data was used to examine how well the sample selection process and the prediction procedures worked. The data set that was used came from the U.S. Bureau of the Census - County and City Data Book (1983).

The ultimate goal was to predict the number of physicians (column 39) for each state. States were the first-stage units. The second-stage units were counties within states. The auxiliary variables used were total persons per state (column 2) and total income in the state (column 116).

Even though data were provided on the number of physicians per state, this information was treated as unknown. It was later used as a check on the predictions derived by the method described in this paper.

Hence, using the procedures of this paper with two auxiliary variables per state, states were chosen to be sampled. Within the sampled states, counties were chosen to be sampled. For all counties, data were available on the number of physicians per county (hence, these data were available for the sampled counties). The prediction procedures proposed were then applied to the sample data and using the model, predictions were derived for the number of physcians per state.

In order to evaluate these predictions, the following statistic was computed for the sampled states:

$$\frac{Y_i - \hat{Y}_i^*}{Y_i} \qquad (4.1)$$

where Y_i is the true number of physicians for state i (based on the Census data) and $\hat{Y}_i^*$ is the model-based estimator for state i (i=1,...,n where n is the number of sampled states). For the nonsample states, a similar statistic was computed:

$$\frac{Y_j - \hat{Y}_j}{Y_j} \qquad (4.2)$$

where Y_j is the number of physicians per state j based on the Census data and $\hat{Y}_j = \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}}$ is the model-based predictor (j = n+1,...,N). Thus equation (4.1) measures the relationship of the estimated value to the true value in units of the true value for the sampled states. Equation (4.1) measures the proportion of error in the predicted values. If (4.1) is zero, that implies the predicted value and the true value are equal which means zero error. Hence, values of (4.1) "close" to zero indicate that the predicted value was "close" to the true value. Conversely, "large" absolute values of (4.1) indicate the predicted values are not good. A similar explanation holds for equation (4.2) except that (4.2) measures the error in the predictions for the non-sampled states. Here N=48 since Alaska and Virginia did not have county data available.

The total number of counties in the 48 states used for this example is 2936. Three different total sample sizes were selected in order to see what effect total sample size had on the quality of the predictions. The three sample sizes were m = 200, 300 and 600. That is, m = 200 is approximately a 6.8% sample, m = 350 is approximately a 10.2% sample and m = 600 is roughly a 20.4% sample of the total number of counties. For each sample size, the first step was to select the states to be used at the first-stage of the process. This was done using the Carter-Faulkenberry technique described in "A Model-Based Sample Selection Procedure For One-Stage Sampling" (1989). The one-stage Carter-Faulkenberry sample selection process determined both the number of states to use and which states to use. For m = 200, it turned out that the optimum number of states to sample was n = 5 (California, New York, Pennsylvania, Ohio, North Carolina). Finally, when m = 300, the optimum number of states to sample was n = 9 (California, New York, Pennsylvania, Ohio, North Carolina, Texas, Tennessee, New Jersey, Illinois) and for m = 600, the optimum number of states to sample was n = 13 (California, New York, Pennsylvania, Ohio, North Carolina, Texas, Tennessee, New Jersey, Illinois, Georgia, Connecticut, Alabama, Florida).

In order to evaluate the prediction and sample selection processes, 1000 samples of size m were chosen for each of the fixed sample sizes: m = 200, 300 and 600. For each of these samples, the n values for equation (4.1) and the N - n values for equation (4.2) were computed. The mean of the 1000 (4.1) and the mean of the 1000 (4.2) values were next computed for each state in order to see how the prediction procedures behaved in the long run. That is, the mean of the (4.1) values was computed for each sample state where the mean is

$$\sum_{k=1}^{1000} \frac{Y_i - \hat{Y}_{ik}^*}{Y_i} \Big/ 1000 \qquad (4.3)$$

for fixed i = 1,⋯, n and where $\hat{Y}_{ik}^*$ is the predicted value for state i for simulation number k where k = 1, ..., 1000.

Similarly, the mean of the (4.2) values,

$$\sum_{k=1}^{1000} \frac{Y_j - \hat{Y}_{jk}}{Y_j} \Big/ 1000 \qquad (4.4)$$

was computed for each nonsample state j=n+1,...,N and where $\hat{Y}_{jk}$ is the predicted value for state j for simulation number k with k = 1, ..., 1000.

In addition to the means given by (4.3) and (4.4), standard deviations, ranges, maximums, and minimums were computed for each state where the computations were done over the 1000 values of either (4.1) for sample states or (4.2) for nonsample states. These statistics were computed for m = 200, 300 and 600.

Summary statistics for the 1000 (4.1) and (4.2) values were computed for each of the three cases (i.e., m=200, 300 and 600). Generally, the procedure appeared to work well for the sampled states and varied some for the nonsampled states. When m=200 and n=5, the mean of the 1000 (4.1) numbers for each of the 5 states is only off by a maximum of 6.2% For the nonsampled states, N-n = 43 and the mean of the 1000 (4.2) numbers for each of the 43 states is off by a maximum of 117%. That maximum occurred in Wyoming and it is probably due to the fact that Wyoming has only 512 doctors, most of whom are located in a couple of counties. If those counties were not chosen for the sample, the estimate would vary quite a bit from the true value. When m=300 n=9 and N-n=39, it was seen that the means for the sampled states are a little more variable (now off by a maximum mean of 22.8%) while the nonsampled states have means which are off by a maximum of 97.7% (again in Wyoming). Finally, when m=600, n=13 and N-n=35, the maximum means for both sampled and nonsampled states are smaller than those when m=300, n=9 and N-n=39. Overall, the means are good for the sampled states and vary for the nonsampled states. The geographic distribution of the doctors seems to be a very important factor in how well the process performs.

5. COMMENTS AND FURTHER QUESTIONS

A number of subjects need to be investigated before this model-based sample selection and prediction process becomes clearly practical. One subject is that the proposed procedures are model-based. The sensitivity of the procedures to deviations from model assumptions is a critical point. The second-stage units are randomly selected but the first-stage units are chosen with purposive and not random sampling. Comments about the possible problems encountered are discussed in Carter and Faulkenberry (1989).

A further important subject for examination is the case when there are multiple predictions desired for each unit. Multiple predictions (for example: number of physicians per state and number of serious crimes per state) may cause the auxiliary variables to change. This would effect the sample selection and prediction processes but in unknown ways. It is not clear how to select the "best" set of one-stage sample units in order to predict variate values for two or more characteristics of interest per unit.

Also, this paper proposed a procedure which required homogeneous variances of the first-stage units. Nonhomogeneous variances will require some adjustment to these procedures.

In conclusion, the sample selection and prediction processes proposed in this paper may be useful in deter-mining predictors for two-stage sampling when a good model relationship exists.

REFERENCES

Carter, N.J. (1981), "Predicting Unit Variate Values in a Finite Population," Ph.D. Thesis, Oregon State University.

Carter, N.J. and Faulkenberry, G.D. (1989), "A Model-Based Sample Selection Procedure For One-Stage Sampling," Communications in Statistics, Vol. 18, No. 8, pp 3135-3147.

Royall, R.M. (1976), "The Linear Least-Squares Prediction Approach to Two-Stage Sampling," Journal of the American Statistical Association, 71, pp. 657-664.

U.S. Bureau of the Census, "County and City Data Book, 1983," U.S. Government Printing Office: 1983.