A MODEL-BASED SAMPLE SELECTION PROCEDURE FOR ONE-STAGE SAMPLING

Nancy J. Carter, California State University, Chico, G. David Faulkenberry, Oregon State University
Nancy J. Carter, Department of Mathematics & Statistics, CSUC, Chico, Chico, CA 95929

KEY WORDS: Superpopulation; Prediction; Variate Values

ABSTRACT

The problem of sample selection, when a one-stage superpopulation model-based approach is used to predict individual variate values for each unit in a finite population based on a sample of only some of the units, is investigated. The model framework is discussed and a sample selection scheme based on the model is derived. The sample selection scheme is evaluated using actual data. Future research topics including multiple predictions per unit are suggested.

1. INTRODUCTION

The problem considered here is as follows: Assume a one-stage superpopulation model-based approach is used to predict individual variate values for each unit in the population, based on a sample of some of the units. How should the sample units be selected? That is, if we are using a model-based approach to predict variate values for all units in a finite population, based on auxiliary variables and a sample of some of the units, what is the best way to select the sampled units? The purpose of this paper is to examine a sample selection procedure based on the model.

Problem Formulation and Inference Model

Consider a finite population of units, $\{u_1, u_2, \ldots, u_N\}$ which have associated with them known auxiliary variables, $\{\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_N\}$, where $\underset{\sim}{x}_i' = (1, x_{i1}, \ldots, x_{ik})$. Let $(Y_1, \ldots, Y_N)$ be independent random variables and assume that $Y_i \sim \xi(\underset{\sim}{x}_i'\underset{\sim}{\beta}, \sigma^2)$ where $\underset{\sim}{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k)$ and σ^2 are unknown parameters. Think of $(y_1, y_2, \ldots, y_N)$ as a particular realization of this random vector. A sample of n of the variate values is observed and on the basis of this sample together with the auxiliary variables, an estimate the variable values of each of the remaining $N-n$ units will be derived.

To simplify the notation, let the population of units be arranged so that the first n units correspond to the sample units. Denote the observed variate vector by $y' = (y_1, \ldots, y_N)$ and let

$$X = \begin{bmatrix} \underset{\sim}{x}_1' \\ \underset{\sim}{x}_2' \\ \vdots \\ \underset{\sim}{x}_n' \end{bmatrix}$$

As estimators for the $N-n$ units not in the sample, the work of Royall (1971) will be used and the best linear unbiased predictor (with respect to the distribution ξ)

$$\hat{y}_i = \underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}} \qquad \text{for } i = n+1, \ldots, N$$

where $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'y$ will be the estimators. At this point only the homogeneous variance case will be considered. Some discussion of this case will be given in the last section.

The problem thus formulated is one of multiple prediciton in a linear regression context. The uniqueness of the problem is that there is a prespecified finite number of units such that a sample of size n of these may be selected to observe, then predictors must be derived for the unspecified ones. Of course, this prediciton is only with respect to ξ. The question of randomization must still be dealt with. However, first inference with respect to ξ will be considered and then probability sampling will be discussed. The main question with which this paper is concerned, which units to take in the sample, will now be considered.

Optimal Design with Respect to ξ

The expected mean square error of an individual prediction is, for $i = n+1, \ldots, N$,

$$E_\xi(\hat{y}_i - Y_i)^2 = E_\xi(\underset{\sim}{x}_i\hat{\underset{\sim}{\beta}} - Y_i)^2.$$

Now for $i=n+1, \ldots, N$, $\hat{\underset{\sim}{\beta}}$ and Y_i are independent, so

$$\begin{aligned} E_\xi(\underset{\sim}{x}_i\hat{\underset{\sim}{\beta}} - Y_i)^2 &= V(\underset{\sim}{x}_i\hat{\underset{\sim}{\beta}}) + V(Y_i) \\ &= \underset{\sim}{x}_i'(X'X)^{-1}\underset{\sim}{x}_i\sigma^2 + \sigma^2 \\ &= \sigma^2[\underset{\sim}{x}_i'(X'X)^{-1}\underset{\sim}{x}_i + 1] \end{aligned}$$

In order to express this expected mean square error in terms of deviations about the mean, let

$$\underset{\sim}{d}_i = \begin{bmatrix} x_{i1} - \bar{x}.1 \\ \vdots \\ x_{ik} - \bar{x}.k \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} \underset{\sim}{d}_1' \\ \vdots \\ \underset{\sim}{d}_n' \end{bmatrix}$$

and it will be seen from Searle (1971) that,

$$E_\xi(\underset{\sim}{X}_i\hat{\underset{\sim}{\beta}} - Y_i)^2 = \sigma^2[1 + \frac{1}{n} + \underset{\sim}{d}_i'(D'D)^{-1}\underset{\sim}{d}_i].$$

So, for given n, the MSE of a particular estimate is determined by $\underset{\sim}{d}_i'(D'D)^{-1}\underset{\sim}{d}_i$. It also happens that in the normal theory context, multiple prediciton intervals have width determined by the quantity $\underset{\sim}{d}_i'(D'D)^{-1}\underset{\sim}{d}_i$. This quantity will be used as a basis for designing the sample. The problem however is still complicated since there are $N-n$ estimates. There will be a value $\underset{\sim}{d}_i'(D'D)^{-1}\underset{\sim}{d}_i$ for each of the $N-n$ units. With the typical population size it is not practical to consider $\binom{N}{n}$ sets and evaluate the $N-n$ predictions for each set. Even if it were practical, there is still the problem of choosing a criterion for deciding when one set of $N-n$ predictions is better than another set. Alternatives that might be used include dealing with the maximum error, the total error, or to consider each prediction in the sense that, for a particular sample, s, the question could be asked, is there a unit in s that could be replaced by a unit not in s, such that every prediction would be as good or better than the predictions with s.

To attack the problem directly by deciding which n units should be put in the sample seems to be particularly cumbersome. Consequently the proposed procedure starts with all N units in the population, then uses a stepwise procedure to decide which one of these units to eliminate. A unit is eliminated at each step by deciding which one of those remaining would be easiest to predict. That is, omit the one with the smallest expected mean square error at each step.

To express this more precisely, let $\bar{\underset{\sim}{x}}' = (x_{\cdot 1},\ldots,x_{\cdot k})$, where $\bar{\underset{\sim}{x}}'$ is calculated using all N population units, and also calculate D using all N units in the population. In addition let $\underset{\sim}{x}'_{(j)}$ and $D_{(j)}$ denote that unit j is omitted from the calculations. That is, $\bar{\underset{\sim}{x}}_{(j)}$ and $D_{(j)}$ are calculated using $N-1$ units. The question then is, if one unit were left out to predict, which unit would be easiest to predict, i.e., which unit has smallest $\underset{\sim}{d}'_{(j)}(D'_{(j)}D_{(j)})^{-1}\underset{\sim}{d}_{(j)}$ where $\underset{\sim}{d}_{(j)}$ denotes deviations about $\bar{\underset{\sim}{x}}_{(j)}$? It turns out that this is equivalent to finding the minimum of $\underset{\sim}{d}'_j(D'D)^{-1}\underset{\sim}{d}_j$ (the proof of this result is omitted for the sake of brevity). This implies that the unit to omit is the one closest to $\bar{\underset{\sim}{x}}$ using the metric $\underset{\sim}{d}'_j(D'D)^{-1}\underset{\sim}{d}_j$.

Once one unit, say k, has been eliminated, from the sample, the previous process is repeated, only now the "population" consists of the $N-1$ remaining units. That is, use $\bar{\underset{\sim}{x}}_{(k)}$ and $D_{(k)}$ in place of $\bar{\underset{\sim}{x}}$ and D and repeat the procedure. Eliminate another unit and repeat again with the new "population" of $N-2$ units. This process continues until only n units are left. These n units will be the sample units.

This selection method requires $N-n$ steps with $\underset{\sim}{d}'_j(D'D)^{-1}\underset{\sim}{d}_j$ evaluated $N - i + 1$ times at step i. This will generally be fast and easy to compute. A major strength of this method is that it is practical and applicable. However, by using a stepwise approach it will not necessarily select the same n units that would have been selected if all possible sets of n were considered, and the set that satisfied some optimality criterion were chosen.

Example: Evaluation of the Procedure Using Actual Data

In order to examine how well this sample selection process works, actual data was used to evaluate it. The data set that was used provided information for 141 large standard metropolitan statistical areas (SMSAs) in the United States. This data set is located on pages 1109 - 1113 of Neter, Wasserman and Kutner (1985). The ultimate goal was to predict the number of physicians (variable 6) based on the auxiliary variables land area (var 2), total population (var 3) and total personal income (var 10). A multiple regression using the entire set of 141 with the previously mentioned variables, yielded a multiple R^2 of 96.4%. Therefore it appeared there was a good relationship between the three predictor (auxiliary) variables and the response variable.

As a method of evaluation, for the sample selection procedure it was decided to compare the proposed procedure with two possible optimality criteria. The optimality criteria chosen were:

1. $\min \Sigma E_\xi(\hat{y}_i - Y_i)^2$ and
2. $\min \max E_\xi(\hat{y}_i - Y_i)^2$,

where the sum and the max are taken over all non-sample units. The idea, of course, is that the proposed sampling plan might be justified as giving a sample which is approximately optimal according to one of these criteria.

Since $N = 141$ is too large to work with when we are looking at all $\binom{N}{n}$ samples, the decision was made to use $N = 10$ and $n = 4$. That is $N = 10$ of the SMSAs were randomly chosen as the 'population'. Then all $\binom{10}{4} = 210$ samples of size 4 from this population were examined. For each sample $\Sigma E(\hat{y}_i - Y_i)^2$ was computed and the max of $E_\xi(\hat{y}_i - Y_i)^2$ was determined. In addition, the proposed sample selection procedure was run to determine which of the 210 samples would have been selected by this procedure. It was determined which samples would have satisfied criteria 1 and 2 and then the values obtained for criteria 1 and 2 were compared with the $\Sigma E_\xi(\hat{y}_i - Y_i)^2$ and max $E_\xi(\hat{y}_i - Y_i)^2$ obtained for the sample chosen by the proposed procedure. This procedure was repeated 30 times, i.e., thirty 'populations' of size ten were drawn from the data set and this evaluation procedure was repeated for each of these 'populations'. The following ratio was then computed:

$$\frac{\Sigma E_\xi(\hat{y}_i - Y_i)^2}{\min \Sigma E_\xi(\hat{y}_i - Y_i)^2}$$

← sum for sample chosen by the proposed procedure

← minimum sum for all 210 possible samples

If this ratio is 1, the proposed sample selection process chooses the optimal sample according to this criterion. A second ratio was computed, namely:

$$\frac{\max E_\xi(\hat{y}_i - Y_i)^2}{\min \max E_\xi(\hat{y}_i - Y_i)^2}$$

← max for sample chosen by the proposed procedure

← min max for all 210 samples

Once again, a ratio of 1 means the sample chosen by the proposed procedure was optimal according to this criterion.

Summary statistics for the comparisons are given in Table 1. The medians of the ratios are quite good for both optimality criteria. They indicate the proposed procedure produced $\Sigma E_\xi(\hat{y}_i - Y_i)^2$ and max $E_\xi(\hat{y}_i - Y_i)^2$ values which differed from the optimal values by less than 4% and 8%, respectively, 50% of the time. The mean ratios are also good. In fact, 25 of the thirty sum ratios are less than 1.25 and 20 of the max ratios are also less than 1.25.

Table 1

	Sum Ratio	Max Ratio
mean	1.194	1.311
median	1.035	1.075
standard deviation	.326	.529
Percent exact match	26.6	33.3
Percent differ by 1	66.7	56.7
Percent differ by 2	6.7	10

The evaluation of the sample selection procedure up to this point seems to show that the proposed sample selection scheme performs very well and often gives samples "close" to "optimal". However, a population of size 10 and a sample of size 4 are very small numbers and are not realistic in practical terms. Since one goal of this research is to produce a sample selection scheme which is useful for actual data sets, it was decided to see how well this sample selection process performed when used with the entire data set. That is, for $N = 141$ and with $n = 47$ (chosen to be 1/3 of the data set), the sample selection process was used and predictors were derived for the nonsample units. Since the true values for the predictor variable (number of physicians) were available, the prediction errors were also computed. Table 2 gives summary values for $Y_i - \underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}}$ and $(Y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta})^2$ for the non-sample values.

Table 2

	$Y_i - \underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}}$	$(Y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta})^2$
min	-594.76(at unit 53)	11.80 (at unit 70)
max	1369.62(at unit 68)	1875847.4(at unit 68)
mean	-57.38	88036.57
median	-99.18	30864.487

Table 2 shows that "on the average", the predictions produced based on this sample overestimated the number of physicians per SMSA by approximately 58 (using the mean) and with a median overestimate of 99 per SMSA. The actual number of physicians per SMSA ranged from 140 to 25,627 with a median of 769. Thus, the predictions derived in this manner should be "close" to those achieved by "optimal" sampling and they tend on average to slightly overestimate the number of physicians.

Comments and Further Questions

There are a number of topics that require investigation before this model-based sample selection scheme becomes truly practical. The sensitivity of the model to deviations from the model assumptions is of critical importance in deciding whether to use randomization or purposive sampling. One of the goals of this research was to determine what role the model could play in the sample design. Toward that end the purposive sampling scheme described in this paper was developed. However, there are many things to consider before performing nonrandom sampling. Hájek (1981, pp. 14 and 20) has stated that the sample selection is affected by many factors including population size, assumption of independence, sample size, number of auxiliary variables, and how well the population is mixed. Further research should be done on the sample selection process and how it affects the predictors and error variances before this procedure is routinely applied. However, it should be noted that a relatively small sample size may justify using the proposed purposive sampling scheme. As pointed out in Hansen, Madow, and Tepping (1983, p. 791), when dealing with a practical problem where the sample size is relatively small, model-dependent inferences may be preferable to strictly random sample selection. This topic definitely requires further study.

Another important topic for study is the situation where there are multiple predictions per unit. Even if there are still homogeneous variances, the auxiliary variables may change. What effect does this have on the sample selection process? If purposive sampling were to be used, how could one select the "best" set of sample units in order to predict variate values for two or more characteristics of interest per unit? These questions need examination.

One consequence of the proposed allocation scheme deserves special mention. By ordering the sample units in the proposed way, any "outlier" will definitely be included in the sample. This may not be desirable in estimating $\underset{\sim}{\beta}$. One possible alternative to the proposed plan would be to use our method to order the population units and then to oversample the units with the larger values from this ordered set. Other types of sampling schemes may also be appropriate. This is another topic worth studying.

It should also be noted that the sample selection scheme proposed in this paper will work for homogeneous variances. A different approach must be taken for nonhomogeneous variances.

To conclude, when individual unit predictors are required for each unit in a finite population but due to restrictions it is not possible to sample each unit, the sampling scheme described here is a possible procedure for selecting the "best" sample.

Acknowledgements

This research was supported by EPA Grant Number R807226-01. The authors wish to thank Dr. Ross A. Gollan of Deakin University, Geelong, Australia for reviewing this paper. Any remaining errors are our responsibility, and we gratefully acknowledge the improvements made by Dr. Gollan.

References

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

Fedorov, V.V. (1972), Theory of Optimal Experiments. Translated and edited by W.J. Studden and E.M. Klimko, New York: Academic Press, Inc.

Hájek, J. (1981), Sampling from a Finite Population, Edited by V. Dupac, New York: Marcel Dekker.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," Journal of the American Statistical Association, 78, pp. 776-793.

Neter, J., Wasserman, W., and Kutner, M.H. (1985), Applied Linear Statistical Models, 2nd ed., Homewood, Illinois: Richard D. Irwin, Inc.

Royall, R.M. (1971), "Linear Regression Models in Finite Population Sampling Theory", in Foundations of Statistical Inference, eds., V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart & Winston, pp. 259-274.