PREDICTING UNIT VARIATE VALUES IN A FINITE POPULATION

Nancy J. Carter, California State University, Chico

Introduction

The topic investigated in this paper is that of predicting variate values for all individual units in a finite population based on a sample of the units. It was suggested by the following problem. The Environmental Protection Agency(E.P.A.) wanted to conduct a survey to predict the annual number of recreation days by activity for each county in Oregon, Washington, and Idaho. There were four activities of interest and results were needed for each county in the three-state area. Restrictions made it impossible to sample each county; and it was necessary, therefore, to design a survey that would provide predictions for each county without actually sampling each county. In general, the problem is to design a survey to predict (estimate) variate values for each unit in a finite population based on a model derived from sample units and auxiliary variables. A superpopulation model-based approach was used to solve this problem.

Model, Definitions and Notation

There are a number of possible variations on the prediction problem. The linear least-squares prediction approach was the one used for this situation. The notation will follow that of Royall (1976).

Consider a finite population of N identifiable units (where N is a known integer). Associated with unit i is some value $y_i (i=1, \ldots, N)$. We assume y_i is some observed value of the random variable Y_i, where Y_i represents some characteristic of interest over unit i. The joint distribution of $Y_1, \ldots, Y_N$ will be denoted by ξ. Also associated with unit i are p known auxiliary variables $X_{i1}, X_{i2}, \ldots, X_{ip}$. Let $\underset{\sim}{X}_i' = (X_{i1}, \ldots, X_{ip})$ be the vector of auxiliary variables for unit i. Therefore, $y_1, y_2, \ldots y_N$ are realized values of the independent random variables $Y_1, \ldots, Y_N$, where $E_\xi(Y_i) = \underset{\sim}{X}_i' \underset{\sim}{\beta}$ and $\mathrm{Var}_\xi(Y_i) = \sigma_i^2 = \sigma^2 g(\underset{\sim}{X}_i)$. The function $g(.)$ is known with $g(\underset{\sim}{X}_i) > 0$ for all $\underset{\sim}{X}_i$ not identically zero. We assume σ^2 and $\underset{\sim}{\beta}' = (\beta_1, \ldots, \beta_p)$ are unknown constants.

We draw a sample of n $(p \leq n < N)$ units. The choice of n and the specific units chosen for the sample will not be discussed at this time. Without loss of generality, assume the units are arranged so that the first n are sample units and the remaining $N - n$ are not sampled.

Denote by $\underset{\sim}{X}_I$ the $n \times p$ matrix of auxiliary variables and by $\underset{\sim}{V}_I$ the $n \times n$ 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 $(N - n) \times n$ matrix of covariances between nonsample and sample units. Denote by $\underset{\sim}{Y}$ the $N \times 1$ vector of random variables $Y_1, \ldots, Y_N$.

If $\underset{\sim}{Y}$ is arranged so that the first n units are those in the sample, the model states that $E_\xi(\underset{\sim}{Y}) = \underset{\sim}{X}\underset{\sim}{\beta}$ and $\mathrm{Cov}_\xi(\underset{\sim}{Y}) = \underset{\sim}{V}$ where

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

and $\underset{\sim}{\beta}$ is defined as above. That is,

$$\underset{n \times 1}{\underset{\sim}{Y}_I} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \text{and} \quad \underset{n \times p}{\underset{\sim}{X}_I} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

Also,

$$\underset{n \times n}{\underset{\sim}{V}_I} = \sigma^2 \begin{bmatrix} g(\underset{\sim}{X}_1) & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & g(\underset{\sim}{X}_2) & & & & 0 \\ \vdots & & \cdot & & & \vdots \\ & & & \cdot & & \\ 0 & 0 & \cdot & \cdot & \cdot & g(\underset{\sim}{X})_n \end{bmatrix} = \sigma^2 \underset{\sim}{W}_I$$

and

$$\underset{(N-n) \times (N-n)}{\underset{\sim}{V}_{II}} = \sigma^2 \begin{bmatrix} g(\underset{\sim}{X}_{n+1}) & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & \cdot & & \\ 0 & \cdot & \cdot & \cdot & \cdot & g(\underset{\sim}{X}_N) \end{bmatrix} = \sigma^2 \underset{\sim}{W}_{II}.$$

Since it is assumed $\mathrm{Cov}_\xi(Y_i, Y_j) = 0$ for all $i \neq j$ $(i, j = 1, \ldots, N)$, $\underset{\sim}{V}_{II,I}$ is a $(N-n) \times n$ matrix of zeros. Therefore,

$$\underset{N \times N}{\underset{\sim}{V}} = \sigma^2 \begin{bmatrix} g(\underset{\sim}{X}_1) & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdot & & & & & \cdot \\ \cdot & & \cdot & & & & \\ \cdot & & & & & & \\ 0 & 0 & \cdot & \cdot & \cdot & & g(\underset{\sim}{X}_N) \end{bmatrix} = \sigma^2 \underset{\sim}{W}.$$

By elementary matrix properties:

$$\underset{n \times n}{\underset{\sim}{V}_I^{-1}} = \frac{1}{\sigma^2} \begin{bmatrix} \frac{1}{g(\underset{\sim}{X}_1)} & \cdot & \cdot & \cdot & \cdot & 0 \\ & & & & & \cdot \\ 0 & \frac{1}{g(\underset{\sim}{X}_2)} & & & & 0 \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & \frac{1}{g(\underset{\sim}{X}_n)} \end{bmatrix} = \frac{1}{\sigma^2} \underset{\sim}{W}_I^{-1}.$$

Under model ξ, if $\hat{\underset{\sim}{\beta}}$ is the weighted least-squares estimate of $\underset{\sim}{\beta}$, $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}_I' \underset{\sim}{V}_I^{-1} \underset{\sim}{X}_I)^{-1} \underset{\sim}{X}_I' \underset{\sim}{V}_I^{-1} \underset{\sim}{Y}_I$.

Predictors

Recall the goal is to predict the value of some characteristic of interest for each unit in the population. For the sample units this value is the observed $y_i (i=1), \ldots, n)$. For the nonsample units the value is $Y_j (j=n+1,\ldots,N)$. Since $E_\xi(Y_j) = \underset{\sim}{X}_j'\underset{\sim}{\beta}$ and $\hat{\beta}$ is ξ-unbiased for $\underset{\sim}{\beta}$, i.e., $E_\xi(\hat{\underset{\sim}{\beta}}) = \underset{\sim}{\beta}$, the natural predictor for Y_j is $\underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}}$.

Error Variance of Predictors

The model based approach which has been used suggests the appropriate measure of uncertainty of the predictors is the mean-square error or error variance with respect to ξ, the joint distribution of the units in $\underset{\sim}{Y}$. Therefore, assume the sample $\underset{\sim}{s}$ of units is fixed and consider $E_\xi(Y_j - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})^2$ for $j=n+1,\ldots,N$.

$$E_\xi(Y_j - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})^2 = E_\xi(Y_j - \underset{\sim}{X}_j'\underset{\sim}{\beta} + \underset{\sim}{X}_j'\underset{\sim}{\beta} - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})^2$$

$$= E_\xi\left[(Y_j - \underset{\sim}{X}_j'\underset{\sim}{\beta}) + (\underset{\sim}{X}_j'\underset{\sim}{\beta} - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})\right]^2$$

$$= E_\xi(Y_j - \underset{\sim}{X}_j'\underset{\sim}{\beta})^2 + 2E_\xi(Y_j - \underset{\sim}{X}_j\underset{\sim}{\beta})(\underset{\sim}{X}_j'\underset{\sim}{\beta} - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})$$

$$+ E_\xi(\underset{\sim}{X}_j'\underset{\sim}{\beta} - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})^2$$

$$= \mathrm{Var}_\xi(Y_j) + 2\mathrm{Cov}_\xi(Y_j, \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}}) + \mathrm{Var}_\xi(\underset{\sim}{X}_j\hat{\underset{\sim}{\beta}}).$$

Since Y_j is a nonsample unit, independent of the sample random variables from which $\hat{\underset{\sim}{\beta}}$ is computed, $\mathrm{Cov}_\xi(Y_j, \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}}) = 0$.

Therefore,

$$E_\xi(Y_j - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})^2 = \mathrm{Var}_\xi(Y_j) + 2\mathrm{Cov}_\xi(Y_j, \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}}) +$$

$$+ \mathrm{Var}_\xi(\underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}}) = \sigma^2\, g(\underset{\sim}{X}_j) + 0 + \underset{\sim}{X}_j'\mathrm{Cov}_\xi(\hat{\underset{\sim}{\beta}})\underset{\sim}{X}_j$$

$$= \sigma^2\, g(\underset{\sim}{X}_j) + \underset{\sim}{X}_j'\mathrm{Cov}_\xi(\hat{\underset{\sim}{\beta}})\underset{\sim}{X}_j \ .$$

$$\mathrm{Cov}_\xi(\hat{\underset{\sim}{\beta}}) = \mathrm{Cov}_\xi\left[(\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I)^{-1}\ \underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{Y}_I\right]$$

$$= (\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}_I\underset{\sim}{V}_I^{-1}\mathrm{Cov}_\xi\ (\underset{\sim}{Y}_I)\ \left[(\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\right]'$$

$$= (\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{V}_I\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I(\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I)^{-1}$$

$$= (\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I)^{-1}$$

$$= \sigma^2\ (\underset{\sim}{X}_I\underset{\sim}{W}_I^{-1}\underset{\sim}{X}_I)^{-1} \ .$$

Hence,

$$E_\xi(Y_j - \underset{\sim}{X}_j'\hat{\underset{\sim}{\beta}})^2 = \sigma^2\, g(\underset{\sim}{X}_j) + \underset{\sim}{X}_j'\mathrm{Cov}_\xi(\hat{\underset{\sim}{\beta}})\underset{\sim}{X}_j$$

$$= \sigma^2\, g(\underset{\sim}{X}_j) + \underset{\sim}{X}_j'(\underset{\sim}{X}_I'\underset{\sim}{V}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_j$$

$$= \sigma^2\, g(\underset{\sim}{X}_j) + \sigma^2\underset{\sim}{X}_j'(\underset{\sim}{X}_I'\underset{\sim}{W}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_j$$

$$= \sigma^2\left[g(\underset{\sim}{X}_j) + \underset{\sim}{X}_j'(\underset{\sim}{X}_I'\underset{\sim}{W}_I^{-1}\underset{\sim}{X}_I)^{-1}\underset{\sim}{X}_j\right] \ .$$

At this point also note that a ξ-unbiased estimate of σ^2 is given by:

$$\hat{\sigma}^2 = (n-p)^{-1} \sum_{i=1}^{n} \frac{(Y_i - \underset{\sim}{X}_i'\hat{\underset{\sim}{\beta}})^2}{g(X_i)} \ .$$

Comments and Topics For Future Study

A major problem for the one-stage sample situation is encountered in determining which specific n units should be chosen for the sample. That is, given n, which units are the best ones to sample in order to minimize the error variances of the predictors. This depends on how best and minimize are defined. There are N-n error variances to consider and minimizing these could be done a number of ways. One possible solution is to examine all N units and then do a stepwise elimination until only N-n are left; see Carter (1981). The n eliminated units are those used for the sample. This problem illustrates one of the differences between predicting individual variate values for all units and predicting a total over all units. When predicting a total over all units, there is only one error variance to minimize.

Other problems arise in extending the model to the two-stage prediction situation. That is, the situation where the y_i are not directly observable when unit i is selected but instead must be estimated by a subsample of the subunits within unit i. For example, how many units n and subunits $m_i (i=1,\ldots,n)$ should be selected when the total sample size $m = \sum_{i=1}^{n} m_i$ is fixed but $m_i (i=1,\ldots,n)$ and n are random? What are the predictors and error variances of the predictors for sample and nonsample units in this two-stage problem? These questions are investigated in my Ph.D. thesis; see Carter(1982).

Topics which I have not investigated include the robustness of the models for the one-and two-stage problems, multiple predictions per unit (for example, how this affects the sample selection), and fixed cost sample allocation. All of these topics should be studied to determine the usefulness of this particular prediction technique.

References

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

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