MULTIPLE REGRESSION ESTIMATION USING CLASSIFIED LANDSAT DATA

by

George A. Hanuschak
USDA/ESCS

Manuel Cárdenas
ASA/USDA Fellow
New Mexico State University

New Techniques Section
Research and Development Branch
Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture

April 1978

Multiple Regression Estimation Using Classified LANDSAT Data

INTRODUCTION

This study was undertaken to investigate the degree that the precision of the single independent variable regression estimator for crop hectarage using classified LANDSAT data could be improved by the addition of one or more independent variables. For example, we can consider the addition of the variable representing the number of pixels [1/] classified as soybeans to our present regression of corn hectares (ground sample data) on pixels classified as corn. Huddleston and Ray [2/] have recently investigated the use of LANDSAT data in a multiple regression estimator for average corn yields and reported information gains over current objective yield estimates on the order of 1.27 to 1.42.

SINGLE INDEPENDENT VARIABLES CASE

The form of the regression estimator [3/] used by the Statistical Research Division of the Economics, Statistics, and Cooperatives Service within a land use stratum for a particular crop (corn, for example) is given by:

$$\bar{y}_r = \bar{y} + \hat{b} \, (\bar{X}_1 - \bar{x}_1)$$

where

$\bar{y}_r$ = the estimated mean hectarage of the particular crop in question per segment.

$\bar{y}$ = the mean hectarage of the particular crop per segment in the sample.

$\bar{x}_1$ = the mean number of pixels per segment in the sample which were classified as the crop in question.

$\bar{X}_1$ = the mean number of pixels per segment in the population which were classified as the crop in question.

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}$$ = the consistent estimate of the constant regression coefficient which minimizes the variance.

y_i = the hectarage of the particular crop in question for the i'th sample segment.

x_{1i} = the number of pixels in the i'th segment which were classified as the crop in question. (Note: the subscript 1 is used to prevent confusion when the multiple regression estimator is introduced).

This estimator is known to be biased and according to Cochran [4] the bias to the order 1/n is:

$$-\frac{1}{n}(S_{e1l}/S_{x_1}^2)$$

where

$$S_{e1l} = \frac{1}{N-1} \sum_{i=1}^{N} e_i (x_{1i} - \bar{X}_1)^2$$

$$e_i = y_i - \bar{Y} - B(x_{1i} - \bar{X}_1)$$

$$S_x^2 = \sum_{i=1}^{N} (x_{1i} - \bar{X}_1)^2 / (N - 1).$$

The approximate variance of this estimator for a large sample can be found in most survey sampling books and is:

$$V(\bar{y}_r) = \left(\frac{1 - f}{n}\right) S_y^2 (1 - \rho^2) \qquad (2)$$

where

n = the number of randomly chosen segments for the given stratum.

$f = \frac{n}{N}$ = the sampling fraction.

N = the total number of segments in the given stratum.

$S_v^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / (N - 1)$ = the variance of the y_i's.

$$\rho = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})(x_{1i} - \bar{X}_1)}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{Y})^2 \sum_{i=1}^{N} (x_{1i} - \bar{X}_1)^2}}$$ = the correlation coefficient of v_i and x_{1i}.

From equation (2) it follows that as ρ^2 increases, the closer ρ^2 is to 1, the smaller is the variance of $\bar{v}_r$.

MULTIPLE INDEPENDENT VARIABLE CASE

When considering multiple independent variables, say q variables, the regression estimator is of the form:

$$\bar{y}_{mr} = \bar{y} + \sum_{i=1}^{q} B_i(\bar{X}_i - \bar{x}_i) \qquad (3)$$

where

B_i = known constant corresponding to i'th independent variable (crop).

$\bar{X}_i$ = mean number of pixels per segment in the population which were classified as the i'th independent variable (crop).

$\bar{x}_i$ = mean number of pixels per segment in the sample which were classified as the i'th independent variable (crop).

The variance of this estimator is more easily expressed in matrix form. Therefore, using matrix notation [5/] the estimator becomes $\bar{y}_{mr} = \bar{y} + B'(\bar{X} - \bar{x})$ (4)

where

$$\bar{X}' = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_q)$$

$$\bar{x}' = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_q)$$

and $$B' = (B_1, B_2, \ldots, B_q)$$

The estimate $\bar{y}_{mr}$ can be shown to be unbiased for $\bar{Y}$ when B' is fixed and its variance is given by

$$V(\bar{y}_{mr}) = \frac{1 - f}{n} [S_y^2 + B' \Sigma_{xx} B - 2 B' \Sigma_{xy}] \qquad (5)$$

where

$$\Sigma_{xx} = \begin{pmatrix} S_{x_1}^2, S_{x_1x_2}, \ldots, S_{x_1x_q} \\ S_{x_2x_1}, S_{x_2}^2, \ldots, \cdot \\ \cdot \\ \cdot \\ S_{x_qx_1}, \ldots, S_{x_q}^2 \end{pmatrix} \qquad \Sigma_{xy} = \begin{pmatrix} S_{x_1y} \\ S_{x_2y} \\ \cdot \\ \cdot \\ S_{x_qy} \end{pmatrix}$$

and $$S_{x_ix_j} = \frac{\sum_{k=1}^{N} (x_{ik} - \bar{X}_i)(x_{jk} - \bar{X}_j)}{N - 1}, \quad i,j=1, \ldots, q$$

Since the vector B is a vector of known constants, we would like to choose the vector of constants which would minimize the variance of $\bar{y}_{mr}$. Taking the derivation of (5) with respect to B', setting the resulting equation to zero, and solving for B gives

$$\hat{B} = \Sigma_{xx}^{-1} \Sigma_{xy} \qquad (6)$$

The minimum variance of $\bar{y}_{mr}$ is then found by substituting (6) into (5). This yields

$$V(\bar{y}_{mr}) = \frac{(1 - f)}{n} (S_y^2 - \hat{B}' \Sigma_{xx} \hat{B}) \qquad (7)$$

This variance can be expressed as:

$$V(\bar{y}_{mr}) = \frac{(N - n)}{N} \frac{S_y^2}{n} \cdot [1 - R^2_{yx_1x_2 \ldots x_q}] \qquad (8)$$

where $R_{yx_1x_2 \ldots x_q}$ is the population multiple correlation coefficient as defined in Anderson [6], i.e.

$$R_{yx_1x_2 \ldots x_q} = \frac{\sqrt{\hat{B}' \Sigma_{xx} \hat{B}}}{S_y}$$

Note the similarity between equation (2) for the one independent variable case and equation (8) for the multiple independent variable case. For the multiple variable case the simple correlation coefficient between y and x is replaced by the multiple correlation between y and the linear combination of $x_1, \ldots, x_q$.

THE TWO INDEPENDENT VARIABLE CASE

In the two independent variable case the estimator becomes

$$\bar{y}_{mr} = \bar{y} + B_1 (\bar{X}_1 - \bar{x}_1) + B_2 (\bar{X}_2 - \bar{x}_2)$$

and the formulas for the constants B_1 and B_2 are explicitly given by

$$B_1 = \frac{S_{yx_2} S_{x_1x_2} - S_{yx_1} S^2_{x_2}}{S^2_{x_1x_2} - S^2_{x_1} S^2_{x_2}}$$

and

$$B_2 = \frac{S_{yx_1} S_{x_1x_2} - S_{yx_2} S^2_{x_1}}{S^2_{x_1x_2} - S^2_{x_1} S^2_{x_2}}$$

Since the variances and covariances are unknown, B_1 and B_2 are estimated from the sample by the consistent estimates

$$\hat{b}_1 = \frac{s_{yx_2} s_{x_1x_2} - s_{yx_1} s^2_{x_2}}{s^2_{x_1x_2} - s^2_{x_1} s^2_{x_2}}$$

and

$$\hat{b}_2 = \frac{s_{yx_1} s_{x_1x_2} - s_{yx_2} s^2_{x_1}}{s^2_{x_1x_2} - s^2_{x_1} s^2_{x_2}}$$

The minimum variance is then estimated by

$$(1 - f) \frac{s^2_y}{n} (1 - \hat{R}^2_{yx_1x_2})$$

where $\hat{R}^2_{yx_1x_2}$ is the sample multiple correlation coefficient squared.

The variance formula is valid only for large n. According to KONIJN,[7] the term of order 1/n in the bias of the multiple regression estimate $\bar{y}_{mr}$ is

$$(1 - f)\,\frac{1}{n}\left[\frac{1}{1 - \rho^2_{x_1x_2}}\right]\left\{(2\,S_{e12}\,\rho_{x_1x_2} / S_{x_1}\,S_{x_2})\,(S_{e11} / S^2_{x_1}) - (S_{e22} / S^2_{x_2})\right\}$$

where

$$S_{ekj} = \frac{1}{N - 1}\sum_{i=1}^{N} e_i\,(x_{ki} - \bar{X}_k)(x_{ji} - \bar{X}_j)$$

for k, j = 1, 2

and $\quad e_i = y_i - \bar{Y} - B_1(x_{1i} - \bar{X}_1) - B_2(x_{2i} - \bar{X}_2)$

Since the divisor is $[1 - \rho^2_{x_1x_2}]$ it follows that if x_1 and x_2 are highly correlated, then the bias of the estimate can be appreciable.

ILLUSTRATIONS OF THE GAINS IN PRECISION

Three examples will be presented which demonstrate varying degrees of gain in reduced variance. LANDSAT data obtained from land use stratum 11 (intensive cultivation) in the Western Pass [8/] area of Illinois is used for the examples. Variance estimates are calculated for the two cases: (a) mean hectares using the regression estimator with one independent variable and (b) mean hectares using the multiple regression estimator with two independent variables. The coefficients of variation are not calculated due to the unavailability, at present, of the data required to make the estimates. However, the relative efficiency of the multiple regression estimator over the simple regression estimator is estimated by the formula $v(\bar{y}_r)/v(\bar{y}_{mr})$. All calculations ignore the finite population correction factor.

Example 1:

In the first example, soybean hectarage was first regressed on classified soybean pixels and secondly regressed on classified soybean pixels and classified

corn pixels. The estimated variances for the simple and multiple regression estimators are 4.65 (hectares)^2 and 4.22 (hectares)^2, respectively, yielding a relative efficiency estimate of 1.10. In this case the estimate of ρ was .884 while the estimate of $R_{yx_1x_2}$ was .897.

Example 2:

In the second example alfalfa hectarage was first regressed on classified alfalfa pixels and then on classified alfalfa pixels and classified soybean pixels. The variance estimates were respectively 3.277 (hectares)^2 and 2.518 (hectares)^2 resulting in an estimated relative efficiency of 1.30. In this case $\hat{\rho}$ = .402 and $\hat{R}_{yx_1x_2}$ = .569.

Example 3:

In this example dense woods hectarage was regressed first on classified dense wood pixels and then on classified dense wood pixels and classified permanent pasture pixels. The variance estimates are 3.677 (hectares)^2 and 3.00 (hectares)^2 respectively. The relative efficiency estimate was 1.23 while $\hat{\rho}$ = .288 and $\hat{R}_{yx_1x_2}$ = .501.

Generally speaking it appears that the gain is greater when the simple correlation coefficient ρ is poor.

SELECTION OF THE INDEPENDENT VARIABLES

All three examples in the previous section exhibited a substantial reduction in the variance with the addition of a second independent variable. Actually the multiple correlation coefficient will usually increase and never decrease with the addition of new variables [9/]. Therefore, the variance will at worst stay the same with the addition of new variables. The question then arises as to how many variables and which variables to include. These questions

can be more adequately addressed after the multiple regression estimator, its variance and bias are implemented on the EDITOR software system at Bolt, Beranek, and Newman Data Processing Center as their answers will require considerable future research.

The three major factors to consider in developing a variable selection procedure are the variance and bias of the estimator and the amount of data available (i.e. the degrees of freedom). Conceivably, classical variable selection procedures and criteria such as stepwise, backward, forward or select procedures can be adapted. However, one must pay special attention to the bias of the estimator. Since each coefficient of regression has a bias the addition of each variable contributes to the overall bias. Also, the relation among the dependent variables must be considered since, as indicated previously, highly correlated variables might increase the bias to the point that it is no longer negligible.

FOOTNOTES

1/ A pixel is the smallest area resolution element of LANDSAT data and is approximately equivalent to 1.1 acres or 0.44 hectares.

2/ Economics Factors in Implementation of Remote Sensing Into Agricultural Information Systems, Huddleston H.F., Ray R.M., The Second Annual Conference on the Economics of Remote Sensing Information Systems, San Jose, California, January, 1978.

3/ The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop Acreage Experiment, Gleason C., Starbuck R., Sigman R., Hanuschak G., Craig M., Cook P., Allen R., Statistical Reporting Service, U.S. Department of Agriculture, October 1977.

4/ Sampling Techniques, Cochran W.G., John Wiley and Sons, Inc., 1963.

5/ Final Report on Design of Sample Surveys To Reduce Respondent Burden, Hocking R.R., Mississippi State University, September 30, 1977.

6/ An Introduction to Multivariate Statistical Analysis, Anderson T.W., John Wiley and Sons, Inc., 1958.

7/ Statistical Theory of Sample Survey Design and Analysis, Konijn H.S., Elsevier-North Holland Pub. Co., 1974.

8/ Same as 3/.

9/ Applied Regression Analysis, Draper N.R., Smith H., John Wiley and Sons, Inc., 1966.