Pak. J. Agri. Sci.
25(4) 358, 1988

A STUDY ON SELECTION OF THE IMPORTANT VARIABLES SUBSET IN BARLEY BREEDING TRIALS

Aftab-i-Islam, Muhammad Ashfaq and Nazir Ahmad
epartment of Mathematics and Statistics, University of Agriculture, Faisalabad.

Six yield prediction models based on different combinations of plant characters in Himalyan primitive barleys were compared by means of (i) The R^2-adequacy test and (ii) The Residual Mean Square Ratio (RMSR) test. Percentage Relative Efficiency Estimate (PRE) was derived for each of the 15 possible pairs of regression models. The study indicated the superiority of regression model involving two independent variables namely area of flag leaf and total number of grains per plant.

INTRODUCTION

A major problem in model building studies is the choice of the independent regressors that are of real value. The reliability, of course, increases by increasing the number of independent variables but this causes much more increase in the volume of work, time and cost. To avoid complexity and minimize the effort, it is desired to have fewer regressors in the model that can serve the purpose of prediction. Stepwise procedures and all possible regression methods, based on repeated significance tests, as discussed by Draper and Smith (1981) are commonly used for the purpose of selecting such variables. A functional model having a minimal subset of regressors with a minimum mean square error estimate or high predictability for deriving a suitable optimal is considered to be the best one.

The sufficiency of a model for prediction purposes has been investigated firstly by deriving for it the (i) R^2-adequacy limits, (ii) Residual Mean Square Ratio (RMSR) of the differences in the residual variances of any two regression models with p and q regressors ($p < q$) and the Residual Mean Sum of Squares (RMSS) of the regression with q regressors, which follows the standard F-distribution. Then Percentage Relative Efficiency (PRE) of a regression model with p variables over another regression model with q variables ($p < q$) has been used (Sankar, 1968). The regression models under study belong to the class of general linear regression models

$$Y = X_1B_1 + X_2B_2 + e \text{ -------- } (1)$$

where Y (nx1) is a random vector of observed values, B_1(qx1) and B_2(px1) are vectors of unknown regression constants; X_1(nxq) and X_2(nxp) are data matrices, with full rank, on regressors; e is a (nx1) vector of residuals which are independelty and identically distributed with mean zero and variance σ^2.

THE R^2-CRITERION OF A REGRESSION MODEL

The estimates of B_2 based on model (1) viz $\tilde{B}_2$ can be compared with $\tilde{B}_2$ based on a sub-model,

$$Y = X_2B_2 + e \text{ -------- } (2)$$

where X_2 is a (nxp) matrix of fixed values and B_2 is (px1) vector of unknown regression constants; and e is as defined in (1).

The null hypothesis of equality of B_2 in model (1) and B_2 in model (2) is not rejected at a chosen level of significance, if the adequacy limit holds good for a pair of regression with p and q regressors ($p<q$) viz,

$$\frac{(R^2q - R^2p)}{(1-R^2q)/(n-q-1)} \text{ is less than } qF$$

where F is the critical value of F statistical with (q, n-q-1) degrees of freedom.

The R^2-adequacy limits can be derived for all possible pairs of subset regressions to describe the minimal adequate sets of independent variates. The subset of regressors X_2 in (2) will be infered as R^2 adequate, if R^2p is greater than R^2a.

where $R^2a = 1 - (1 - R^2q)(1 + d)$

where $= qF/(n-q-1)$

here F is at a chosen level of significance with q, n-q-1) degrees of freedom.

THE RESIDUAL MEAN SQUARE RATIO CRITERION

The sufficiency of a regression model with p variables when compared with a regression model with q variables ($P < q$), can be tested by means of an F-Statistic; and is given as

$$F = \frac{RSS(p) - RSS(q)}{(q - p)\ RMSS(q)}$$

with (q-p) (n-q-1) degrees of freedom at a chosen level of significance. Where RSS is the residual sum of squares and RMSS is the residual mean sum of squares of a regression function. The p-variate regression model is preferred to the q-variate regression model if the calculated F-value is less than the critical value of F, at a given level of significance with the necessary degrees of freedom. The q-variate regression is preferred if otherwise.

THE PERCENTAGE RELATIVE EFFICIENCY OF A REGRESSION MODEL

The Percentage Relative Efficiency (PRE) of a regression model A with p-variables over another regression model B with q-variables ($p < q$) can be derived as

$$PRE(A) = \frac{\sigma^2 B (n+q+1)/n}{\sigma^2 A (n+p+1)/n} \times 100$$

where σ_A^2 and σ_B^2 are the estimates of RMSS of regression A and B respectively.

The decision about the preference of one regression model over another model of subset regressors can be derived by comparing the estimates of Percentage Relative Efficiency (PRE) value of regression model with those of the other subsets of regression models. The model A will be preferred over the model B if the PRE is more than 100, if PRE is equal to 100, the choice remains with the experimenter to choose one of the two models.

The above criteria have been applied to the data taken from a field trial, on 75 primitive barley accessions from various altitudes in Indian Himalyan Regions, conducted in the experimental area of the Department of Plant Breeding and Genetics, University of Agriculture, Faisalabad during the year 1980-81. The data were recorded on 17 characteristics, given below, by selecting 10 plants from the middle row out of the 3 rows for each accession.

X_1 = Length of flag leaf

X_2 = Breadth of flag leaf

X_3 = Area of flag leaf

X_4 = Height of plant

X_5 = Height of stem

X_6 = Number of fertile tillers per plant

X_7 = Length of top internode

X_8 = Length of main ear

X_9 = Number of spikelets per main Ear

X_{10} = Length of apical awn

X_{11} = Length of middle awn

X_{12} = Length of basal awn

X_{13} = Average length of awn

X_{14} = Weight of main ear

X_{15} = Total number of grains per plant

X_{16} = Total grain weight per plant

X_{17} = 1000 grain weight per plant

Total grain weight X_{16} was taken as dependent variable. Six yield prediction models, given below, with various combinations of regressors were investigated.

Model A, contains all the sixteen regressors.

Model B, contains X_3 X_4, X_8, X_{13}, X_{14}, X_{15}, X_{17}

Model C, contains X_3, X_4, X_{13}, X_{15}, X_{17}

Model D, contains X_3, X_8, X_{13}, X_{15}

Model E, contains X_3, X_8, X_{15}

Model F, contains X_3, X_{15}

The estimates of the regression coefficients and estimates of experimental error (σ) under each model are given in table 1. Based on the 't' test made, the contribution of 4 out of 16 regressors in model (A), 5 out of 7 in model (B), 2 out of 5 in model (C), 4 out of 4 in model (D), 2 out of 3 in model (E) and 2 out of 2 in model (F) were found statistically significant.

The values of co-efficient of determination (R^2) are quite high for almost all the models and gradually decreased from model (A) to model (F). The value for model (A) is 0.9458 and that for model (F) is 0.7920.

The fifteen possible combinations of subset regressors were compared for the R^2-adequacy and when compared with A, models B and C only, were found to be R^2 adequate. Model D, E and F were not R^2 adequate, therefore these models are out of the race for final selection. Different pairs of models and their R^2-adequacy limits are given in the table No. 2.

Table No. 3 shows the results of residual mean square ratio criterion. We see that the model (A) is significantly different from model D, E, and F. Model B is singnificantly different

Table 1. *The estimates of regression coefficients under different models*

Variable	A	B	C	D	E	F
X_1	0.0256					
X_2	5.2681					
X_3	-0.3524	-.1223*	-0.0918	0.1582*	0.1838*	0.1949*
X_4	6.76531E-04	-.0063	-0.0111			
X_5	-0.0545					
X_6	-0.0141					
X_7	0.0898					
X_8	-0.8583*	-.9019*		-1.3882*	-0.8872	
X_9	0.1031					
X_{10}	0.3821					
X_{11}	-0.0657					
X_{12}	-0.1619					
X_{13}	-0.0222	.2052	-0.1148	0.9103*		
X_{14}	2.1102*	2.2375*				
X_{15}	0.0265*	.0279*	0.0280*	0.0299*	0.0292*	0.0282*
X_{17}	6.1756*	5.5371*	6.4301*			
Intercept	-15.9260	-12.2829	-14.6882	1.6207	4.7764	-1.8955
R^2	0.9458	0.9402	0.9254	0.8154	0.8009	0.7920
σ	6.2042	5.9266	7.1715	17.5012	18.6096	19.1749

Table 2. *Adequacy limits of different models*

		A	B	C	D	E	F
A	0.9458	--					
B	0.9402	0.9181	--				
C	0.9254	0.9181	0.9267	--			
D	0.8154	0.9181	0.9267	0.9126	--		
E	0.8009	0.9181	0.9267	0.9126	0.7889	--	
F	0.7920	0.9181	0.9267	0.9126	0.7889	0.7888	--

Table 3. *Residual meansquare ratio of different pairs of models*

Model pair	D.F. $(q - p)$	S.S. $RSS_p - RSS_q$	M.R.S.S. $\frac{RSS_p - RSS_q}{(q - p)}$	F
AB	9	37.2377	4.1375	0.6669
AC	11	134.9878	12.2720	1.9779
AD	12	865.2408	72.1030	11.6220*
AE	13	961.4384	73.9570	11.9204*
AF	14	1020.7451	72.9104	11.7500*
Residual of A	59	359.8443	6.2040	----
BC	2	97.7501	48.8750	8.2467*
BD	3	828.0031	276.0010	46.5699*
BE	4	924.2010	231.0500	38.9853*
BF	5	983.5074	196.7015	33.1900*
Residual of B	68	397.0800	5.9270	----
CD	1	730.2530	730.2530	101.8271*
CE	2	826.4509	413.2250	57.6210*
CF	3	885.7573	295.2524	41.1700*
Residual of C	70	494.8321	7.1710	----
DE	1	96.1979	96.1979	5.4970*
DF	2	155.5043	77.7522	4.4400*
Residual of D	71	1225.0851	17.5010	----
EF	1	59.3064	59.3064	3.190
Residual of E	72	1321.2830	18.6096	----

Table 4. *Relative efficiency matrix of different models*

Models	A	B	C	D	E	F
A	--	102.10	131.30	324.40	349.30	38.17
B	105.90	--	124.00	306.40	329.90	39.60
C	98.30	84.70	--	247.10	266.10	94.97
D	40.80	35.10	41.50	--	107.70	99.56
E	38.80	33.50	39.50	95.20	--	98.31
F	364.48	284.47	113.76	105.66	104.34	--

from the models C, D, R, and F. Model C is significantly different from the models C, D, R, and F. Model C is significantly different from models D, E and F. Model D is significantly different from models E and F.

The estimate of the percentage relative efficiency of regresion model when compared with each of the other regression model are given in the table No. 4. The estimates suggest that model A can be preferred over models C, D and E. Model B can be preferred over C, D, and E. Model C can be preferred over D and E. Model D can be preferred over E. Model F can be preferred over A, B, C, D and E. Considering R^2 adequacy, Residual Mean Square Ratio criterion and Percentage Relative Efficiency estimate, model B can be preferred over others.

REFERENCES

Draper, N.R. and Smith, H. 1981. Applied Regression Analysis. John Willey and Sons, Inc. New York.

Maruthi Sankar, G.R. 1968. On Screening of Regression methods for Selection of optimal variables Subsets. Jour. Ind. Soc. Ag. Statistics Vol. XXXVIII, No. 2, pp.161-168.