Jour. Ind. Soc. Ag. Statistics
Vol. XXXVIII, No. 2 (1986), pp. 161-168

ON SCREENING OF REGRESSION MODELS FOR SELECTION OF OPTIMAL VARIABLE SUBSETS

G. R. MARUTHI SANKAR
All India Co-ordinated Soil Test-Crop Response Correlation Project, ICAR, Hyderabad

(Received : June, 1983)

SUMMARY

The yield prediction models involving six different combinations of soil-fertility regressor variables were compared by means of (i) the R^2-adequacy test and (ii) the Residual Mean Square ratio test. The estimates of the percentage relative efficiency (PRE) of each regression model for the preference over each of the other regression models were derived for the 15 possible pairs of regression models. The results indicated the superiority of a functional model involving the linear a:d quadratic trends of soil and fertiliser parameters along with their interactions as independent regressors, when compared with each of the other regression models.

Keywords : Regression models; R^2-adequacy; RMSR Test; Efficiency-of models.

Introduction

In any model building study, the experimenter faces the problem of prediction of the response surface with an optimal subset of independent regressors which are of real value. Stepwise procedures viz., Forward selection and Backward elimination are in common use to predict the variable subsets based on repeated significance tests as discussed by Draper and Smith [2].

The usefulness of a regression function differs from model to model with varying predictors. The derived optima would differ from model to model depending upon the magnitudes of the non-orthogonal regression coefficients. The best functional model would be the one having a minimal subset of predictors with a minimum mean square error estimate and having high predictability and feasibility for deriving a suitable optimal, The regression models investigated in this paper involved different combinations of soil and fertiliser nutrients of N, P and K as predictor variables. The R^2-adequacy limits of each regression model are derived to infer on its sufficiency for prediction purposes (Aitkin, [1]).

The sufficiency of a regression has also been derived by means of testing the residuals viz., by deriving the 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.

2. Adequacy Criterion for Regressions

The regression models investigated belong to the class of general regression models,

$$Y = X_1 \beta_1 + X_2 \beta_2 + E \tag{1}$$

where $Y (n \times 1)$ is a random vector of observed values β_1 and β_2 are $(q \times 1)$ and $(p \times 1)$ vectors of unknown regression constants; X_1 and X_2 are $(n \times q)$ and $(n \times p)$ matrices of fixed values with full rank; and E is a $(n \times 1)$ vector of residuals which are independent and identically distributed with mean zero and variance σ^2.

2.1 *The R^2 Criterion of a Regression Model*

The estimates of β_2 based on model (1), viz $\hat{\beta}_2$ can be compared with $\tilde{\beta}_2$ based on a sub-model,

$$Y = X_2 \tilde{\beta}_2 + E \tag{2}$$

where X_2 is a $(n \times p)$ matrix of fixed values and $\tilde{\beta}_2$ is $(p \times 1)$ vector of unknown regression constant; and E is as defined in (1).

The null hypothesis $H_0 : \hat{\beta}_2 = \tilde{\beta}_2$ 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.,

$$\frac{(R_q^2 - R_p^2)}{(1 - R_q^2)/(n - q - 1)} < qF \tag{3}$$

where F is the critical value of F-statistic 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 inferred as R^2-adequate, if

$$R_p^2 > R_a^2 \tag{4}$$

where $R_a^2 = 1 - (1 - R_q^2)(1 + d)$

and $d = qF/(n - q - 1)$

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

2.2 *The Residual Mean Square Ratio (RMSR) Criterion of a Regression Model*

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-Statistics; and is given as

$$F = \frac{\text{RSS}(p) - \text{RSS}(q)}{(q - p)\,\text{RMSS}(q)} \tag{5}$$

with $(q - p)$, $(n - q - 1)$ degrees of freedom at a chosen level of significance. In (5), 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.

2.3. *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

$$\text{PRE}(A) = \frac{\sigma_B^2\,(n + q + 1)/n}{\sigma_A^2\,(n + p + 1)/n} \times 100 \tag{6}$$

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

The inferences on the frequency preferences of a regression model over other subset of regressions can be derived by comparing the estimates of PRE values of a regression model with those of the other subsets of regression models. For example, the model A will be preferred to the model F, if PRE is more than 100, rejected if PRE is less than 100. If PRE is equal to 100 the choice remains with the experimenter to choose one of the two models.

2.4. *Calibration of Optima under Different Models*

Three types of optimal nutrient calibrations viz., (i) general optima, (ii) economic optima and (iii) soil test based optima can be derived for the nutrients which behaved with the law of diminishing returns under the different models. The calibration for the *j*th nutrient under the *i*th model can be given as :

(i) Calibration for General Optima $(G_{ij}) = \dfrac{L_j}{2Q_j}$

(ii) Calibration for economic optima $(E_{ij}) = \dfrac{L_j - R_j}{2Q_j}$

(iii) Calibration for soil test based Optima $(S_{ij}) = \dfrac{L_j - R_j - I_jS_j}{2Q_j}$

where L_j and Q_j are the linear and quadratic coefficients of *j*th fertiliser nutrient; I_j is the soil and fertiliser interaction coefficient of the *j*th nutrient; R_j is the ratio of the per unit cost of the fertiliser nutrient and the value of the crop; and S_j is the soil test value of the *j*th nutrient respectively.

3. Results and Discussion

The different models discussed in this paper were calibrated for the yield data of a wheat (variety: Kalyansona) experiment conducted in the medium black soils of Madhya Pradesh under the All India Co-ordinated Soil Test Crop Response Correlation Project. The experiment involved 15 fertilised treatments out of a total of 125 treatments based on 5 × 5 × 5 levels of Nitrogen (N), Phosphorus (P) and Potassium (K) nutrients respectively. There were four replications each having 15 fertilised and 4 unfertilised treatments. Thus there were 76 experimental plots and the estimates of the initial soil available N, P and K nutrients were derived for each of the experimental plots. The means and the standard deviations (S.D.) of the soil and fertiliser N, P and K nutrients are given below in Table 1.

TABLE 1—DISTRIBUTION OF SOIL AND FERTILISER NUTRIENTS (kg/ha)

	Soil N	*Soil P*	*Soil K*	*Fertiliser Nutrients*		
	(O.C%)	*(Olsen)*	*(Am. Ac.)*	N	P_2O_5	K_2O
Mean	0.5	17.5	415	105	84	40
S.D.	0.05	8.6	41	69	55	28

The six different combinations of a set of 18 regressors generated from soil and fertiliser N, P & K nutrients are included as independent variates in the yield prediction models. The 18 regressors generated for the model-building investigation are linear (6), quadratic (6) of soil and fertiliser nutrients; soil and fertiliser interactions (3); and fertiliser nutrient interactions (3). However, the fertiliser interaction term of P and K, i.e. PK got eliminated from the model building due to the existence of a significant intercorrelation between P and PK and K and PK regressor variables. The different curvi-linear multiple regressions calibrated for the data are :

(A) a model with all the 17 regressors;

(B) a model with 15 regressors comprising of linear (6), quadratic (6), and interactions (3) of the soil and fertiliser N, P and K nutrients.

(C) a model with 11 regressors comprising of linear (3), quadratic (3), and interactions (2) of fertiliser N, P and K nutrients and interactions (3) of the soil and fertiliser N, P and K nutrients;

(D) a stepwise regression model based on the Gauss Forward selection procedure;

(E) a model with the significant regressors as included in and derived from the model (A); and,

(F) a model with 8 regressors comprising of linear (3), quadratic (3), and interactions (2) of the fertiliser N, P and K nutrients.

The estimates of the regression coefficients and the estimates of experimental error (σ) under each model are given in Table 2. Based on the t-test made, the magnitudes of 9 out of 17 regressors in the model (A), 6 out of 15 in the model (B), 7 out of 11 in the model (C), 10 out of 17 in the model (D) 8 out of 9 in the model (E) and 7 out of 8 in the model (F) were found to be significantly contributing to the estimated Y. The Coefficients of determination (R^2) were highly significant and were found to vary from 0 9 to 0.93 for the 6 models calibrated for the experimental data.

The 15 possible combinations of subset regressions were compared for the R^2-adequacy and it was found that the R^2 values of subset regressions were greater than the adequacy limits and were adequate enough for all prediction purposes. The different pairs of models and their R^2-adequacy limits are given in Table 3. Although the R^2 values of different regressions were adequate and on par with each other, on testing the residual with RMSR criterion, it was found that the model A was significantly different from the models C and F, the model B was significantly different from the models C and F; the model C was significantly different from the model D, the model D was significantly different from

TABLE 2—THE ESTIMATES OF REGRESSION COEFFICIENTS UNDER DIFFERENT MODELS

Sl. No.	*Variable*	*A*	*B*	*C*	*D*	*E*	*F*
1.	*FN*	20.34**	28.68**	43.17**	31.11**	32.55**	32.87**
2.	*FP*	6.86*	18.10*	2.00	11.27*	14.69*	5.61
3.	*FK*	10.12	−49.33	14.85			41.70*
4.	*SN*	—18602	—13153		—19950*	—3283*	
5.	*SP*	83.19*	78.25*		29.20*	29.59*	
6.	*SK*	18.87*	4.50			2.46	
7.	*FN* × *FN*	—0.12**	—0.07**	—0.12**	—0.11**	—0.11**	—0.13**
8.	*FP* × *FP*	—0.11*	—0.05*	—0.10*	—0.09*	—0.09*	—0.09*
9.	*FK* × *FK*	—0.16	0.06	—0.16*			—0.11*
10.	*SN* × *SN*	12103	9010		15405*		
11.	*SP* × *SP*	—1.00	—0.89				
12.	*SK* × *SK*	—0 02	—0.01				
13.	*FN* × *SN*	17. 6	7.80	—24.81*			
14.	*FP* × *SP*	—0.29**	—0.26*	—0.02	—0.22*	—0.21*	
15.	*FK* × *SK*	0.04	0.09	0.06*	0.01*		
16.	*FN* × *FP*	0.16*		0.17**	0.10**	0.08**	—0.19**
17.	*FN* × *FK*	—0.07		—0.15			—0.18*
	Intercept	3439	4750	2222	7876	2373	2222
	R^2	0.93**	0.91**	0.91**	0.92**	0.92**	0.90**
	σ	528	521	573	535	542	589
	C. V.	10.9	10.8	11.8	11.1	11.2	12.2

TABLE 3—ADEQUACY LIMITS OF DIFFERENT MODELS

		A	*B*	*C*	*D*	*E*	*F*
A	(0.93)	—					
B	(0.91)	0.8933	—				
C	(0.91)	0.8933	0.8687	—			
D	(0.92)	0.8933	0.8687	0.8799	—		
E	(0.92)	0.8933	0.8687	0.8799	0.8956	—	
F	(0.90)	0.8933	0.8687	0.8799	0.8956	0.8978	—

the model *F* and the model *E* was significantly different from the model *F*. The RMSR and the *F*-statistic values of different models are given in Table 4.

TABLE 4—RESIDUAL MEAN SQUARE RATIOS OF DIFFERENT PAIRS OF MODELS

Source (*Model-pair*)	*D. f.* ($q-p$)	*Sum of squares* $RSS(p)-RSS(q)$	*Mean sum of squares* (*MRSS*)	*F*
A, *B*	2	108350	54175	0.19
A, *C*	6	4801400	800233	2.87*
A, *D*	7	2429480	347069	1.24
A, *E*	8	3184000	398000	1.43
A, *F*	9	7038900	782100	2.80*
Residual of *A*	58	16193800	279203	
B, *C*	4	4693050	1173263	4.32**
B, *D*	5	2321130	464226	1.71
B, *E*	6	3075650	512608	1.89
B, *F*	7	6930550	990079	3.64**
Residual of *B*	60	16302150	271703	
C, *D*	1	2371920	2371920	7.23**
C, *E*	2	1617400	808700	2.47
C, *F*	3	2237500	745833	2.27
Residual of *C*	64	20995200	328050	
D, *E*	1	754520	754520	2.63
D, *F*	2	4609420	2304710	8.04**
Residual of *D*	65	18623280	286512	
E, *F*	1	3854900	3854900	13.13**
Residual of *E*	66	19377800	293603	

The estimates of the percentage relative efficiency of regression model when compared with each of the other regression models are given as a matrix in Table 5. The estimates suggest that the model *A* can be preferred to the models *C, D, E,* and *F,* the model *B* can be preferred to the models *A, C, D, E* and *F,* the model *C* can be preferred to the model *F,* the model *D* can be preferred to the models *C, E* and *F* and the model *E* can be preferred to the models *C* and *F.* In other words for the numerical example discussed in the paper the models *A, B, C, D, E* and *F* can be preferred in the order *B, A; D, E, C* and *F* for further interpretations.

TABLE 5—RELATIVE EFFICIENCY MATRIX OF DIFFERENT MODELS

Model	*A*	*B*	*C*	*D*	*E*	*F*
A	—	99.42	125.51	110.88	114.94	137.35
B	100.58	—	126.23	111.51	115.60	138.14
C	79.67	79.22	—	88.34	91.59	109.43
D	90.19	89.68	113.20	—	103.66	123.87
E	87.00	86.51	109.18	96.47	—	119.49
F	72.81	72.39	91.38	80.73	83.69	—

The soil test based fertiliser calibrations were found to be possible for *N* under the model *C* and for *P* under the models *A, B, C, D* and *E.* However, the general and the economic fertiliser calibrations were found to exist for *N* and *P* under all the six models and for *K* under the models *A, C* and *F* respectively.

ACKNOWLEDGEMENT

The author is grateful to Dr. M. Velayutham, Project Co-ordinator, All India Co-ordinated Soil Test Crop Response Correlation Project, Hyderabad for useful suggestions in the preparation of the paper.

REFERENCES

[1] Aitkin, A. Murray (1974) : Simultaneous inference and the Choice of variable subsets in Multiple regression, *Technometrics*, **16** (2) : 221-227.

[2] Draper, N. R. and Smith, H. (1966) : *Applied Regression Analysis*. John Wiley and Sons, Inc., New York,