Sindh Univ. Res. Jour. (Sci.Ser.), Vol 33 (2): 1 – 8 (2001)

SELECTION OF THE BEST SUBSET OF YIELD COMPONENTS IN REGRESSION ANALYSIS OF WHEAT DATA

N.M. Shaikh, N.H.Leghari * and Manzoor A.Memon

Institute of Mathematics & Computer Science, University of Sindh, Jamshoro, Pakistan

Abstract

Sixty-four regression models are fitted for six different factors affecting yield of wheat. These models are compared using R^2_p, MSE_p, and C_p. It is concluded that the set comprising five factors, namely, plant height, number of fertile spikes, spike length, number of spikeletes and seed index, is the best.

Key Words: Regression Analysis

Introduction

The relationship between "Y" and X_i's can be expressed as:

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + e_i$$

Most of the criteria currently used to identify the best subset are increasing functions of the residual sum of squares (RSS). For example, the residual mean square, the C_p-statistic due to Mallows [5,6] and the statistic (n + p) RSS / (n-p), where "p" is the number of variables retained, are frequently considered. If such criteria are used, the problem of deciding on "best" subset is reduced to identifying a subset of size p for which RSS is small.

There are two basic problems. First, a criterion is required for deciding if one subset is better than the other. The second problem is computational, since the number of possible subsets that must be considered can be enormous.Two types of techniques can be used to select subsets. The best method is to search through all possible models or some proportion of them, to find those models that optimize or nearly optimize a criterion. The other method is called "stepwise regression" [10]. These techniques essentially choose a path through the possible models, looking first at a subset of one size and then by adding or deleting variables.

The objects of this study are to select the subset of independent variables to use as a model for predicting the yield of wheat in Sindh.

Materials and Methods

The data was obtained from the department of Plant Breeding and Genetics, SAU, Tando Jam, and is given in Table 2.1

* Department of Statistics, Sindh Agri. University, Tando Jam.

Table 2.1 Data on Wheat yield and Its components:

No.	X_1	X_2	X_3	X_4	X_5	X_6	Y	$\hat{Y}$	e_i
1	83	9	13.0	19	69	4.48	13.54	15.68	-2.14
2	84	8	13.0	19	59	4.16	15.26	14.41	0.85
3	88	6	13.2	19	57	4.10	13.74	11.25	2.49
4	83	12	13.7	21	64	4.88	19.58	19.75	-0.17
5	100	6	12.0	19	72	4.04	9.15	11.39	-2.24
6	95	4	11.5	19	69	4.35	9.63	8.55	1.08
7	91	5	11.5	19	57	5.10	11.40	9.84	1.56
8	90	6	13.8	21	76	4.63	14.05	13.09	0.96
9	82	4	13.6	19	62	4.48	9.32	10.62	-1.30
10	87	5	12.5	19	79	4.67	11.86	10.71	1.15
11	88	6	12.2	19	66	4.99	9.47	11.68	-2.21
12	97	6	11.0	19	63	5.05	11.21	10.57	0.64
13	93	4	14.9	19	62	4.18	8.38	9.77	-1.39
14	91	8	11.6	21	89	5.08	17.03	16.97	0.06
15	90	4	13.2	19	69	4.14	8.90	8.31	0.59
16	90	5	13.0	21	80	4.37	11.69	11.09	0.60
17	92	4	11.5	21	63	4.59	9.86	9.87	-0.01
18	94	8	14.4	19	72	4.13	18.97	14.14	4.83
19	90	9	13.0	23	79	4.19	16.42	16.63	0.21
20	82	5	12.8	21	73	4.57	12.78	10.66	2.12
21	92	5	12.8	19	71	5.01	12.02	11.41	0.61
22	97	5	12.8	21	77	4.40	9.54	10.86	-1.32
23	92	8	12.0	21	76	4.43	11.35	13.81	-2.46
24	90	5	12.2	19	69	4.16	7.44	9.98	-2.54
25	86	6	12.0	19	68	4.58	9.38	11.17	-1.79
Mean	89.99	6.12	12.65	19.80	69.40	4.51	12.08	12.09	-0.01
S.D.	4.72	1.97	0.94	1.13	7.365	0.34	3.23	2.76	1.76

The data consists of six independent variables (X_i's) and Y, where

X_1=Plant heiht (cms), X_2=Number of fertile spikes
X_3=Spike length (cms), X_4=Number of spikeletes
X_5=No. of Grains spike, X_6=Seed index (1000 grams) gms.

Y= seed yield (gms) (Independent variable).

$\hat{Y}$ is obtained by regression using the best subset.

Possible subset models are $2^k = 2^6 = 64$ (including the full model and containing only the intercepts).

Selection Criterion

In order to select the best subset some of the available criteria are summarized as:

(i) R^2_p Criterion :
R^2p = regression SS / Total SS
Large value of R^2_p will always result in a good model.

(ii) MSE Criterion :
$MSE_p = (RSS)_p / (Residual\ D.F.)_p$

As MSE_p is not a decreasing function, both parts of above ratio depend upon "p". Since all the models have the same "p" value, they have the same d.f. for RSS.

(iii) C_p Criterion :

$$C_p = \{ (RSS)_p / \sigma^2 \} - (n - 2p),$$

where σ^2 is the estimate of S^2 from the full model.

A model is considered to be appropriate if the value of C_p is very close to p, because for the full model $C_p = p$

Results and discussion

Sixty four different models have been compared using R^2_p, MSE_p and C_p.

All Possible Regressions

Table (3.1) Summary of Regression Co-efficiencets for 64 models

Reg. No (Variables)	Constant (β_0)	Co-eff: (β_1)	Co-eff: (β_2)	Reg. No. (Variables)	Constant (β_0)	Co-eff: (β_1)	Co-eff: (β_2)
1	12.063	-		5 X_4	-10.32	1.130	
2 X_1	28.69	-0.185		6 X_5	05.40	0.095	
3 X_2	3.97	1.322		7 X_6	07.00	1.105	
4 X_3	-13.36	2.005		15 X_2, X_5	00.24	1.297	0.056
8 X_1, X_2	-07.40	-0.037	1.298	16 X_2, X_6	01.60	1.316	0.534
9 X_1, X_3	-07.85	-0.048	1.913	17 X_3, X_4	-14.58	1.902	0.128
10 X_1, X_4	6.06	-0.178	1.110	18 X_3, X_5	-12.81	2.145	-0.033
11 X_1, X_5	23.48	-0.218	1.118	19 X_3, X_6	-16.78	1.989	0.803
12 X_1, X_6	24.18	-0.179	0.891	20 X_4, X_5	-10.32	1.089	0.012
13 X_2, X_3	-7.52	1.130	1.000	21 X_4, X_6	-14.57	1.121	0.982
14 X_2, X_4	-2.60	1.248	0.366	22 X_5, X_6	01.19	0.092	0.987
—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—

	Constant β_0	β_1	β_2	β_3	β_4	β_5	β_6
58	-08.999	0.0316	1.1542	1.2070	-0.2125	0.0009	-
59	-11.409	0.0345	1.1498	1.2109	-0.2129	-	0.4862
60	-12.392	0.0252	1.1314	1.0650	-	-0.0066	0.4904
61	00.210	-0.0553	1.2165	-	0.2103	0.0473	0.4272
62	-14.453	-0.0371	-	1.8469	0.2611	-0.0334	0.7942
63	-07.890	-	1.1388	1.0917	-0.1734	0.0075	0.4538
64	-11.488	0.0352	1.1494	1.2059	-0.2114	-0.0011	0.4876

Note: The results for Reg. No. 23 to 57 are available with authors.

Note: The results for Reg. No. 23 to 57 are available with authors.

Selection of Best Subset

Table (3.2) : Computation of R^2_p , MSE_p and C_p

Model	R^2_p	MSE_p	C_p	Model	R^2_p	MSE_p	C_p
X_o	0.000	11.390	40.720	X_2, X_3, X_4	0.713	3.580	1.270
X_1	0.073	10.560	38.040	X_2, X_3, X_5	0.711	3.600	1.380
X_2	0.645	4.040	1.580	X_2, X_3, X_6	0.714	3.570	1.230
X_3	0.320	7.740	22.280	X_2, X_4, X_5	0.666	4.160	4.250
X_4	0.156	9.610	32.730	X_2, X_4, X_6	0.663	4.210	4.480
X_5	0.050	10.810	39.480	X_2, X_5, X_6	0.665	4.180	4.330
X_6	0.013	11.240	41.840	X_2, X_4, X_5	0.328	8.380	25.780
X_1, X_2	0.648	4.190	3.410	X_3, X_4, X_6	0.328	8.380	24.770
X_1, X_3	0.324	8.040	24.010	X_3, X_5, X_6	0.332	8.330	25.520
X_1, X_4	0.223	9.250	30.460	X_3, X_5, X_6	0.167	10.390	51.370
X_1, X_5	0.148	10.140	35.220	X_1, X_2, X_3, X_4	0.715	3.730	3.160
X_1, X_6	0.081	10.940	39.510	X_1, X_2, X_3, X_5	0.712	3.770	3.330
X_2, X_3	0.711	3.440	0.060	X_1, X_2, X_3, X_6	0.714	3.740	3.180
X_2, X_4	0.659	4.060	2.670	X_1, X_2, X_4, X_5	0.673	4.290	5.850
X_2, X_5	0.663	4.020	2.480	X_1, X_2, X_4, X_6	0.666	4.380	6.280
X_2, X_6	0.644	4.190	3.390	X_1, X_2, X_5, X_6	0.671	4.310	5.950
X_3, X_4	0.321	8.080	24.210	X_1, X_3, X_4, X_5	0.331	8.760	27.610
X_3, X_5	0.325	8.040	23.980	X_1, X_3, X_4, X_6	0.334	8.720	27.420
X_3, X_6	0.327	8.010	23.840	X_1, X_3, X_5, X_6	0.333	8.730	27.460
X_4, X_5	0.157	10.030	34.080	X_1, X_4, X_5, X_6	0.236	10.010	33.660
X_4, X_6	0.166	9.920	34.070	X_1, X_3, X_4, X_6	0.713	3.750	3.250
X_5, X_6	0.061	11.180	40.820	X_2, X_3, X_4, X_6	0.715	3.730	3.120
X_1, X_2, X_3	0.712	3.590	1.340	X_2, X_3, X_5, X_6	0.714	3.750	3.230
X_1, X_2, X_4	0.663	4.200	4.440	X_2, X_4, X_5, X_6	0.669	4.340	6.090
X_1, X_2, X_5	0.669	4.120	4.060	X_3, X_4, X_5, X_6	0.336	8.700	27.310
X_1, X_2, X_6	0.650	4.360	5.250	X_1, X_2, X_3, X_4, X_5	0.715	9.930	5.160
X_1, X_3, X_4	0.328	8.380	25.800	X_1, X_2, X_3, X_4, X_6	0.717	3.897	5.003
X_1, X_3, X_5	0.326	8.400	25.890	X_1, X_2, X_3, X_5, X_6	0.715	3.930	5.170
X_1, X_3, X_6	0.331	8.350	25.610	X_1, X_2, X_4, X_5, X_6	0.674	4.490	7.720
X_1, X_4, X_5	0.320	9.600	32.020	X_1, X_3, X_4, X_5, X_6	0.337	9.130	29.190
X_1, X_4, X_6	0.229	9.610	32.050	X_2, X_3, X_4, X_5, X_6	0.716	3.920	5.110
X_1, X_5, X_6	0.154	10.550	36.890	$X_1, X_2, X_3, X_4, X_5, X_6$	0.717	4.110	7.000

These variables are ordered within each set on the basis of R^2_p, MSE_p and C_p criteria. In order to see if there is any consistent pattern of variables, we examine the variables in the leading equations, given in Table (3.3).

Table (3.3) C_p , R^2_p and MSE $_p$., in leading equations

Set	Variables in Leading Equation	R^2_p	MSE_p	$C_p \cong p$
1*	$Y = f(X_2)$	0.6453	4.038	1.583
2*	$Y = f(X_2, X_3)$	0.7113	3.435	-0.062
3*	$Y = f(X_2, X_3, X_6)$	0.7136	3.570	1.230
4*	$Y = f(X_2, X_3, X_4, X_6)$	0.7153	3.727	3.124
5*	$Y = f(X_1, X_2, X_3, X_4, X_6)$	0.7172	3.897	5.003
6*	$Y = f(X_1, X_2, X_3, X_4, X_5, X_6)$	0.7172	4.113	7.002

R^2_p Criterion

Tables (3.3) indicates that about 71.72% of the total variation is accounted for by model 5*, involving the set { X_1, X_2, X_3, X_4, X_6 }. The amount of percent variation explained by the other models is comparatively small. The addition of X_5 does not bring any improvement in R^2_p. Thus R^2_p criterion suggests that model involving X_1, X_2, X_3, X_4, X_6 is the best.

MSE_P Criterion

Table (3.3) shows that smallest MSE_p value (3.435) is for { X_2, X_3 }, i.e., the best subset. Thereafter it shows an increasing trend as more variables are introduced. For set 5*, the value is 3.897 which is minimum among such six subsets. Hence, set 5* seems to be the appropriate model.

C_p Criterion

Table (3.3) shows value of C_p for set 5* is 5.003 whereas the C_p value is 7.002 for the full set. As C_p value for set 5* is close to p it is the best.

Let us compare the above models by using the additional sum of squares criterion, since model 5* involving X_1, X_2, X_3, X_4, X_6 is a subset of the full model.

Additional Sum of Squares criterion

The additional SS contribution of X_5 is given below:

Table (3.4) : Analysis of Variance {F.05(1,18)=4.41}

Source	DF	SS	MS	F
Due to X_1, X_2, X_3, X_4, X_6	5	186.717	-	-
Extra due to X_5	1	0.005	0.005	-
Regression	6	186.722	-	0.012
Residual	18	75.170	4.176	-
Total	24	261.892	-	-

As the computed value of F < 1.0, so the extra SS due to X_5 does not contribute significantly in the presence of 5*. Hence X_5 is no more a candidate for inclusion in the model.

Thus it is concluded that the model with X_1, X_2, X_3, X_4, X_6 is the best with the prediction equation:

$$Y = -11.409 + \underset{(3.56)}{0.035}\, X_1 + \underset{(4.68)}{1.15}\, X_2 + \underset{(5.32)}{1.211}\, X_3 - \underset{(4.86)}{0.213}\, X_4 + \underset{(7.34)}{0.486}\, X_6$$

Aptness of a Model

The aptness of a model can be examined by studying the residuals, i.e., the plot of residuals against predicted values and independent variables.

Examination of Residuals

If the model is appropriate the residuals ($e_i = y_i - \hat{y}_i$) should exhibit the characteristics which should confirm the assumptions that the errors in the response values are independent with zero mean and constant variances.

Plot of Residuals Against Predicted Values

The residuals plotted against X_i's, indicate no unusual behaviour. The points follow a horizontal band. Moreover, no out-lier is suspected as the value does not exceed 3x (S.D).

Plot of Residuals Against Independent Values

The plot of residuals against X_i's also does not show any serious abnormalities and almost follow a horizontal band of points (These plots are available with the authors).

Normal Probability plot of Residuals

The plot does not show any serious abnormalities and follow an approximate linear trend.

Summary and Conclusions

In order to select the best set of variables for predicting the yield all subsets were tried.

The data was analyzed using IRRISTAT and STATG computer packages. The yield components such as plant height in cms, number of fertile spikes, spike length in cms, number of spikeletes, number of grain perspike, seed index (1000 gr:wt☺ in grms) were taken as indepèndent variables and seed yield as dependent variable.

On the basis of R^2p, MSEp and Cp criterion, the model involving the variables X_1 , X_2, X_3, X_4, X_6 seems to be the best.

The additional sum of squares criterion also confirms the results.

The examination of residuals suggest that there are no abnormalities in the model.
The prediction equation for the model is given as under:

$$Y = -11.409 + 0.035\ X_1 + 1.15\ X_2 + 1.211\ X_3 - 0.213\ X_4 + 0.486\ X_6$$

References:

1. Arvesen, J.N., and G.P. ZMcCabe, 1975. Subset selection problems for variances; applications to Reg. Analysis, JASA 70, 166-169.

2. Garside, M.J., 1965. The best subset in multiple regression analysis. Applied Statistics 14,186-200.

3. Garside, M.J., 1970. Some computational procedures for the best subset problems. App, Statistics 20,8-15.

4. Homsmer, D.W., B.Jovsnovic, snf D. Zlmrdhoe, 1989. Best subsets Logistic Regression, 45,1265-1270.

5. Mallows, C.L., 1964. Choosing variables in Linear Regression: Graphical Aid, Presented at the Regional meeting of the IMS, manhatntan.

6. Mallows, C.L, 1966. Choosing a subset Regression: Presented at the joint Statistical Meeting, Los Angeles.

7. McHenrey, C.E., 1978. Computation of a best subset in multivariate analysis, App. Statistics 27(3), 291-296.

8. Spojovoll., E., 1977. Alternatives to plotting C_p in multiple Regression; Biometrika 64, 1-8.

9. Sajjad, M.N. and N.M. Shaikh 1996. Econometric Models for the prices of Wheat Flour, Rice, Mongs, and Pulses in Baluchistan (1973-1992), Research Journal, Islamia University, Bahawalpur.

10. Weiberg, S., 1980. Applied Linear Regression; John Wiley & Sons, N.Y.