▼ Journal of Research (Science), Bahauddin Zakariya University, Multan, Pakistan.
Vol.13, No.2, December 2002, pp. 119-127 ISSN 1021-1012

SELECTION OF VARIABLES IN MULTIPLE REGRESSION USING STEPWISE REGRESSION

G. R. Pasha
Chairman, Department of Statistics, Dean, Faculty of Science and Agriculture, Bahauddin Zakariya University, Multan, Pakistan.

Abstract: This paper focuses on the selection of variables using stepwise regression. Methodologies of backward elimination and forward selection are discussed. Detailed explanations of stepwise regression procedures are presented. Above methodologies in various terms with empirical data are explained in detail.

Keywords: Backward elimination, forward selection, multiple and stepwise regression, swapping.

INTRODUCTION

In using regression models for prediction, too many regressors cause a higher prediction variance whereas too few regressor variables give a biased prediction. This requires a compromise and is the reason for subset selection. The problem of determining the best subset of variables has long been of interest to applied statisticians and, primarily because of the current availability of high-speed computation; this problem has received considerable attention in the recent statistical feature.

Methods and criterion functions for subset selection are critically reviewed by Hocking [1976]. Miller [1984] discussed computational algorithms for subset selection. He divided search strategies into those, which guarantee to find the best fitting subsets of some or all sizes, and the cheap methods, which sometime find the best fitting subsets. Thompson [1978] has also reviewed subset selection in regression, and Copas [1983] has reviewed developments in regression as a whole over the period 1959-82.

Efroymson [1960] described the stepwise regression, the basis for this procedure is just the Jordan reduction method for solving linear equations with a specific criterion for determining the order in which variables are entered or removed. However for specified stopping rules, stepwise regression also implies the selection of a particular subset of variables.

DEFINITION AND CONCEPTS OF F-TO-ENTER AND F-TO-REMOVE

Variables are entered or removed utilizing a test statistic which namely, the t test for testing the hypothesis that the co-efficient of the variable is zero or equivalently testing the hypothesis that the partial correlation co-efficient is zero. However, the square of this test statistic which is distributed as F and is called either the F-to-enter or F-to-remove. The distribution of the maximum F-to-enter is of course not even remotely like as F-distribution. This was pointed out by Draper and Smith [1981] and by

Pope and Webster [1972]. The true distribution of the maximum F-to-enter is a function of the values of the predictor variables.

In symbols then suppose that a subset C containing P variables has been entered in to the equation P=0,1,2……..K-1, then the F-to-enter rule tests whether a variable X (not in C) significantly improves the prediction of Y over that of the variables in C i.e. testing

$$H_0: \rho_{XY.C} = 0$$

with $$F_{YX.C} = \frac{r^2_{YX.C}\,(n-P-2)}{1-r^2_{YX.C}} \quad F(1, n-P-2)\ d.f.$$

where $\rho_{YX.C}$ = Population partial correlation coefficient

$r^2_{YX.C}$ = Sample partial correlation coefficient.

In stepwise regression if each step is denoted by *q* then F-to-enter for each variable at each step not in the equation is calculated as:

$$F = \frac{SSR_q - SSR_{q-1}}{(SST - SSR_q)\,/\,(n-q-1)}$$

e.g. at fist step the F-to-enter for each variable is

$$F = \frac{SSR_1}{(SST - SSR_1)\,/\,(n-2)}$$

where SSR_1 = sum of square of regression for each variable and at 2nd step.

The F-to-enter for each variable is

$$\frac{SSR_2 - SSR_1}{(SST - SSR_2)\,/\,(n-3)}$$

where SSR_2 = Sum of square of regression for each variable at 2nd step and SSR_1 = Sum of square of regression for each variable at 1st step.

$$F_{YX.C'} = \frac{r^2_{YX.C'}\,(n-P'-2)}{1-r^2_{YX.C'}} \quad F(1, n-P'-2)\ \ d.f$$

This can be interpreted as at each q^{th} step

$$F = \frac{\left[SSR_{(K-q+1)} - SSR_{K-q}\right](n-k+q-2)}{\left[SST - SSR_{(K-q+1)}\right]}$$

This follow F-distribution with (1,*n*-*k*+*q*-2) *d.f.*

MINIMUM-F-TO-ENTER AND MINIMUM F-TO-REMOVE

The number of variables which enter the regression equation in the sequential procedures of selecting variables are controlled through the parameter called the minimum F-to-enter or minimum F-to-remove, on the other hand at each step in selection procedures the calculated value of F-to-enter or F-to-remove is compared with the theoretical pre-assumed value with priori considered significance level, which is called the minimum F-to-enter or minimum F-to-remove. In symbols, min. F-to-enter is equal to $F_{\alpha}(1,V)$ with V = mean squared error degree of freedom with different values of α , mostly α is set equal to 0.05.

METHODOLOGY OF BACKWARD ELIMINATION

The steps in the procedure are outlined as:

i) A regression equation containing all the variables is computed, e.g. for K=3 regression equation

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \in$$

ii) Find partial correlation co-efficient for each variable in the model and using these partials correlation coefficient Find F-to-remove for each variable, e.g. for X_1

$$F_{YX_1X_2X_3} = \frac{r^2_{YX1.X2X3}(n-2-2)}{1-r^2_{YX_1.X_2X_3}} \quad \text{with} \quad (1, n-4)\ d.f.$$

$$F = \frac{SSR_{123} - SSR_{23}}{(SST - SSR_{123})/n-4} \sim F(1, n-4)\ d.f.$$

iii) Compare these calculated F-values with minimum F-to-remove defined in preceding section and variable having smallest value say F_2 than minimum F-to-remove is dropped from the model.

iv) Now repeat the step (ii) with F-to-remove for each remaining variable and find smallest F-to-remove F_L and draw conclusion in such a way.

   a) if $F_L <$ MIN-F-to-remove then drop the variable

   b) if $F_L >$ MIN-F-to-remove then terminate the process and adopt the resulting regression equation and estimate regression coefficient and Anova to test the goodness of fit.

METHODOLOGY OF FORWARD SELECTION

This method start by finding the partial correlation coefficient of the *X*'s with *Y*.

i) Note in first step the partial correlation coefficient is the simple correlation. Using these correlation co-efficient find F-to-enter for each *X*. Then compare the largest F-to-enter with MIN-F-to-enter defined above. It is equivalent to testing the null hypothesis that

$$H_0 : \rho_{YXi} = 0$$

If the largest F-to-enter is less than Min-F-to-enter then the regression is meaningless, and we should seek other method of analyzing the data. If largest F-to-enter is greater than MIN-F-to-enter then include the corresponding variable is regression equation

ii) Now calculate the F-to-enter for remaining variables using partial correlation Coefficient with Y given that X_1 has been included and compare the largest F-to-enter with Min-F-to-enter and draw conclusion such as: If largest F-to-enter is greater than Min-F-to-enter then include the corresponding variable in regression equation and repeat the step (ii) otherwise, if largest F-to-enter is less than Min-F-to-enter then terminate the procedure and adopt the regression equation.

BRIEF ACCOUNT OF DISCUSSION ON STEPWISE REGRESSION

Stepwise regression method was developed to economize on computational efforts, as compared with the all-possible regression approach, while arriving at a reasonably good "best" set of independent variables.

Stepwise regression is an extension of the forward selection procedure, where at each stage provision is made for inclusion and deletion of variables. The method proceeds same as for forward selection with the addition that at each stage, before the inclusion of a new variable, the partial F-criterion for each variable in the regression equation is evaluated and compared with a pre-selected percentage point of the appropriate F-distribution This provide a judgment on the contribution made by each variable as though it had been the most recent variable entered, irrespective of its actual point of entry into the model. Any variable that provides a non-significant contribution due to many reasons such as multi-collinearity among explanatory variables, is removed from the model. This process is continued until no more variables will be admitted to the equation and no more are repeated.

CONCEPT OF SWAPPING

An alternative to stepwise regression, which often finds better fitting subset, is that of replacing predictors rather than deleting them. Suppose that we have 26 explanatory variables denoted by *A* to *Z* and that we start with the subset *ABCD* consider first replacing predictor *A* with that one from the remaining 22 which gives the smallest residual sum of squares in a subset with *BC* and *D* if no reduction can be obtained then *A* is not replaced. They try replacing *B*, then *C* then *D*, and then back to the new first predictor, continuing until no further reduction can be found. This

procedure must converge, and usually converges rapidly, as the residual sum of squares decreases each time that a predictor is replaced.

Many variations on the basic replacement method are possible. As described above the method could converge upon a different final subset if started from subset *DBAC* instead of *ABCD* that is if we carry out the replacement in a different order. A variation on the method is to find the best replacement for *A* but not to make the replacement similarly the best of the four replacements is implemented the process is repeated until no further improvement can be found.

METHODOLOGY OF STEPWISE REGRESSION

In stepwise regression the independent variables $X_1, X_2, \ldots\ldots X_K$ are entered one by one into the equation according to some pre-established criterion. Once a variable is in the equation, however it may be swapped with a variable not in the equation or it may be removed from the equation altogether.

The set of criterion, which helps us to determine how a variable is entered, swapped, or removed is called stepwise criterion, which is as follows:

i) Stepwise procedure with F-test
ii) Swapping with F-test
iii) Stepwise procedure with multiple correlation co-efficient R^2.
iv) Swapping with R^2.

Now we discuss in detail the above procedures.

STEPWISE PROCEDURE WITH F-TEST

First of all we decide about the MIN-F-to-enter and MIN-F-to-remove as discussed above in definition section. Some package computer programs assume a default value for MIN-F-to-enter is 4.0 and for MIN-F-to-remove 3.9. Let we have a set of *K* explanatory variables from which we have to choose best subset.

Step-1. Find the matrix of simple correlation coefficient between *Y* and X_i, i=1,2,3,……,*K* and find F-to-enter for each variable as:

$$F_{YX\,i} = \frac{r_{YXi}^2\ (n-P-2)}{1-r_{YX_i}^2} \quad i = 1,2,\ldots\ldots,K$$

P = Number of variables that has been included in the model, in first step *P*=0.

$$F_i = \frac{SSR_i}{(SST - SSR_i)\big/(n-2)}$$

SSR_i = sum of square of regression while regression *Y* on X_i compare the largest F-to-enter (F_L) with MIN-F-to-enter if F_L < MIN-F-to-enter then the regression is meaningless then we should look any other method to

analyze data. If F_L > MIN-F-to-enter then include the corresponding variable.

Step-2. Find the partial correlation coefficient between Y and remaining variables i.e. $r_{YX_i.X_1}$ and find F-to-enter for each variable rather then X_1

$$F_{YX_i.X_1} = \frac{r^2_{YX_i.X_1}(n-P-2)}{1-r^2_{YX_i.X_1}} \qquad \text{Where} \quad P=1, \; i=1,2,....K \quad i \neq X_1$$

$$F_i = \frac{SSR_{i1} - SSR_1}{(SST - SSR_{i1})/n-3} \qquad \text{with } (1, n-3)\ d.f.$$

SSR_{i1} = Regression SS due to regressing Y on i^{th} and X_1 variables, $SSR_{i1} - SSR_1$ = additional SS due to including ith variable. If F_L is less than MIN-F-to-enter then stop process and adopt equation with one variable if F_L is greater than MIN-F-to-enter then include corresponding variable in equation say (X_2). After including variable X_2 also find F-to-remove for each variable in the equation as:

$$F_{YX_1.X_2} = \frac{r^2_{YX_1.X_2}(n-1-2)}{1-r^2_{YX_1.X_2}} \cdot F_{YX_2.X_1} = \frac{r^2_{YX_2.X_1}(n-1-2)}{1-r^2_{YX_2.X_1}}$$

$$F_1 = \frac{SSR_{12} - SSR_2}{(SST - SSR_{12})/n-3} \cdot F_2 = \frac{SSR_{12} - SSR_1}{(SST - SSR_{12})/n-3}$$

If smallest F-to-remove (F_S) is less than MIN-F-to-remove then delete the corresponding variable, and if F_S is greater than MIN-F-to-remove then go to next step.

Step-3. Repeat step-2 and so on.

Swapping with F-Test

This procedure utilizes the same rules for removal and entering variables as discussed in stepwise methodology, except that at each step a possible exchange is made between a variable in the equation with one that not in the equation.

Stepwise with Multiple Correlation Co-efficient R^2

This procedure uses the F-to-enter for entering variables, but uses a different rule for the removal of variables. In this method at any step a variable is removed if its removal results in a larger R^2 (squared multiple correlation coefficient). It is possible to obtain an increase in R^2. As we know that R^2 is a function of residual mean square and P, the number of variables in the equation. It is therefore possible that the MSE and P, changes in such a way that a small set of variables will indeed give a larger R^2. This procedure then considers in order:

i) removal of any variable using R^2.
ii) entering of a variable using the F-to-enter.

Swapping with R^2

This procedure is the same as defined in (iii) except that the swapping rule is also utilized. The order of the rules is

i) remove a variable using the R^2
ii) swap or exchange two variable to increase R^2
iii) enter a new variable using the F-to-enter.

EXAMPLE

We will apply method of selection of independent variables on data about employed persons in Pakistan with 5 dependencies; using this data we will show an important advantage of stepwise regression on *FS* and *BE* methods.

Year	Y	X_1	X_2	X_3	X_4	X_5
1963-64	16.24	18.33	0.0420	3179	50.310	17.5
64-65	16.47	18.72	0.0479	3212	51.760	18.0
65-66	16.79	19.24	0.0253	3252	53.260	20.0
66-67	16.93	19.26	0.0857	3508	54.790	20.5
67-68	17.17	19.43	0.0358	3528	56.370	20.6
68-69	17.40	19.29	0.0159	3554	58.000	20.8
69-70	17.75	19.23	0.0412	3587	59.700	21.0
70-71	18.37	19.21	0.0572	3549	61.490	21.2
71-72	18.55	19.09	0.0472	3497	63.340	21.5
72-73	19.24	1912	0.0969	3415	65.300	21.7
73-74	19.76	19.38	0.2997	3329	66.879	21.9
74-75	20.30	19.55	0.2671	3289	68.924	22.2
75-76	20.08	19.82	0.1166	3248	71.033	22.5
76-77	21.89	19.76	0.1178	3373	73.205	22.8
77-78	22.73	20.10	0.0779	3676	75.444	23.2
78-79	23.62	19.98	0.0663	3715	77.516	23.4
79-80	24.15	20.23	0.1072	3750	80.130	23.7
80-81	24.70	20.30	0.1237	3815	82.580	24.0
81-82	25.27	20.42	0.1000	3882	84.254	26.2
82-83	25.85	20.31	0.0448	3931	87.758	26.5
83-84	26.40	20.33	0.0836	4047	90.480	26.9
84-85	26.96	20.61	0.0746	4423	93.286	27.2
85-86	27.93	20.67	0.0483	4349	96.180	27.5
86-87	28.70	20.92	0.0387	4544	99.162	27.9
87-88	28.99	20.66	0.3884	4873	102.238	28.0
88-89	29.99	20.73	0.3087	4595	105.409	28.1
89-90	30.82	20.73	0.3854	4543	108.678	28.3
1990-91	31.78	20.77	0.3868	4568	111.938	28.6

Where Y = No. of persons employed in millions, X_1 = Land cultivated in hectors, X_2 = Inflation rate, X_3 = No. of establishments, X_4 = Population in millions, X_5 = Literacy rate.
SOURCE: Economic Survey of Pakistan for 1991-92.

***FS* AND *BE* METHODS**

By using *FS* and *BE* methods we select same subset of independent variables which are X_2 X_4 X_5. Their regression equation and Anova is given below:

$$Y = 4.28 - 2.44X_2 + 0.310X_4 - 0.215X_5$$

Predictor	Coefficients	Standard Deviation	t-Ratio
Constant	4.279	1.273	3.36
X_2	-2.4392	0.8941	-2.73
X_4	0.30990	0.02363	13.12
X_5	-0.2153	0.1246	-1.73
S = 0.3954		R^2 - Sq = 99.4%	

ANOVA

SOV	DF	SS	MS
Regression	3	657.79	219.26
Error	24	3.75	0.16
Total	27	661.54	

STEPWISE REGRESSION

Stepwise regression selects the final model with only two variables X_2 and X_4 it excludes X_5, their regression equation and Anova is given below:

$$Y = 2.16 - 1.73X_2 + 0.270X_4$$

Predictor	Coefficients	Standard Deviation	t-Ratio
Constant	2.1626	0.36	6.01
X_2	- 1.7291	0.825	-2.10
X_4	0.270	0.0052	31.39

CONCLUSIONS AND DISCUSSION

We have seen that both *FS* and *BE* methods allow to include X_5 in final model, but it is not appropriated because the t-Ratio of X_5 which is –1.73 is not significant, on the other hand stepwise regression automatically remove X_5 and select X_2 and X_4, it show an advantage of stepwise on *FS* and *BE* methods. It is clear that there is no unique way of searching for a best set of independent variables, and subjective approach also play an important role in the search process.

R^2 method is reasonable for selection purpose; it gives a clear idea about the increase in variation explained by regression equation due to adding a new variable in the model. A limitation of this method is that there is computational problem, if there is a large number of explanatory variables say 20, then 2^{20} –1 regression equations are needed to estimate which is very difficult however with the access of high speed computer it become easy.

FS method is a version of stepwise method it gives same results obtained by stepwise regression. *BE* method selects same variables as obtained by stepwise regression. Stepwise regression seems suitable method for data considered; it reached the final model with less correlated variables. It screen out less informative variables which can be observed in example.

References

Bendel, R.B. and Afifi, A.A. (**1977**) "Comparison of stopping rules in forward stepwise regression", *J. American Statist. Assoc.*, 72, 46-53.

Brocrsen, P.M.T. (**1984**) "Stepwise backward elimination with C_p as selection criterion", International report ST-SV 84-01, Department of Applied Physics, Delft.

Broersen, P.M.T. (**1986**) “Subset regression with stepwise directed search”, *Appl. Statist.*, 35(2), 168-177.

Chirstenson, P.D. (**1982**) "Variable selection in multiple regression”, Ph.D. Dissertation No. 8307741, Iowa State University, Ann. Arbor and London.

Copas, J.B. (**1983**) "Regression, prediction and shrinkage with discussion", *J. R. Statist, Soc*. B, 45, 311-354.

Draper, N. and Smith, H. (**1981**) "Applied Regression Analysis”, 2nd ed. Wiley, New York.

Efroymson, M.A. (**1960**) "Multiple Regression Analysis", In: A. Ralston and H.S. Wilf (Eds.), *Mathematical Methods for Digital Computers*, John Wiley, New York.

Elden, L. (**1972**) "Stepwise regression analysis with orthogonal transformations", Unpublished report, Mathematics Department, Linkoping University, Sweden.

Feiveson, A.H. (**1973**) "Selecting variables in regression and classification by thresholding, Ph.D. Dissertation, Texas A and M University, College Station Texas.

Forsythe, A.B., Engelman, L., Jennrich, R. and May, P.R.A. (**1973**) "A stopping rule for variable selection in multiple regression", *Amer. J. Statist. Assoc.*, 68, 75-77.

Gorman, J.W. and Toman, R.J. (**1966**) "Selection of variables for fitting equations to data", *Technometries*, 8, 27-51.

Hocking, R.R. (**1976**) “The analysis and selection of variables in linear regression”, *Biometrics*, 32, 1-49.

Lindley, D.V. (**1968**) "The choice of variables in multiple regression", *J. Royal, Statist. Soc.* (London), B 30, 31-53.

Madhakrishnan, S. (**1974**) "Selection of variables in multiple regression”, Ph.D. Dissertation, University of Houston, Texas.

Mallows, C.L. (**1973**) "Some comments on C_p", *Technometrics*, 15, 661-675.

Mantel, N. (**1970**) "Use of stepdown procedure in variable selection", *Technometrics*, 12, 621-625.

Miller, A.J. (**1984**) “Selection of subsets of regression variables”, *J. R. Statist.* A, 147, Part-3, 389-425.

Pope, P.T. and Webster, J.T. (**1972**) “The use of an F-statistics in stepwise regression procedure”, *Technometrics*, 14, 327-340.

Thompson, M.L. (**1978**) “Selection of variables in multiple regression: Part I. A review and evaluation”, *International Statistical Review*, 46, 1-49.

Thompson, M.L. (**1978**) “Selection of variables in multiple regression: Part II. Chosen procedure computation and examples“, *International Statistical Review*, 46, 129-146.