INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Multiple Regression of Students' Performance Using forward Selection Procedure, Backward Elimination and Stepwise Procedure

Oti, Eric Uchenna
Lecturer, Department of Statistics, Federal Polytechnic Ekowe, Bayelsa State, Nigeria
Awogbemi, Clement Adeyeye
Research Fellow, National Mathematical Centre, Abuja, Nigeria
Slink, Ruth Abiobaragha
Lecturer, Department of Statistics, Federal Polytechnic Ekowe, Bayelsa State, Nigeria

Abstract:
The building of required models aids in appropriate model prediction. The students GPA($\boldsymbol{Y}$) which is the response (dependent) variable was regressed on four predictor (independent) variables. Methodologies of the backward elimination, forward selection procedure and stepwise procedure are discussed with numerical data to find the model which gives the prediction $\boldsymbol{Y}$, given $\boldsymbol{X_1, \dots, X_4}$.

1. Introduction

The study of regression analysis which is a statistical tool for evaluating the relationship between a response and one or more predictive variables $\boldsymbol{X_1, \dots, X_K}$ has a variety of purposesand also for estimation of parameter using sample data.
Most practical application of regression analysis utilize models that are more complex than the simple linear model. In such cases, the regression analysis is called multiple linear regression. Notably, Bowerman and O'connell (1997) added that we can more accurately describe, predict and control a dependent variable by using regression model that employs more than one in dependent variable.
The linear regression model relating to $\boldsymbol{Y}$ to $\boldsymbol{X_1, \dots, X_K}$ is given as

$$\boldsymbol{Y = B_0 + B_1X_1 + \cdots + B_KX_K + \varepsilon} \quad (1.1)$$

Here

a. $\boldsymbol{\mu_{Y/X_1, \dots, X_K} = B_0 + B_1X_1 + \cdots + B_KX_K}$ (1.2)
is the mean value of the dependent Variable $\boldsymbol{Y}$ when the values of the independent variables are$\boldsymbol{X_1, \dots, X_K}$.

b. $\boldsymbol{B_0, B_1, \dots, B_K}$Are unknown regression parameters relating the mean value of $\boldsymbol{Y}$ to $\boldsymbol{X_1, \dots, X_K}$.
c. $\boldsymbol{\varepsilon}$Is an error term that describes the effects on $\boldsymbol{Y}$of all factors other than the values of the independent variables$\boldsymbol{X_1, \dots, X_K}$.

The model in Equation (1.1) is a linear regression model because the expression
$\boldsymbol{B_0 + B_1X_1 + \cdots + B_KX_K}$ expresses the mean value of $\boldsymbol{Y}$ as a linear function of the parameters$\boldsymbol{B_0, B_1, \dots, B_K}$ (Franklin and Hariharan,1994)
Based on suitable criteria, we will select more of these subset prediction functions for further consideration.
The subset prediction function considered are linear in the predictor variables as well as the parameters, when there are *K* predictor variable in all, the number of possible subset prediction function could take the form

$$\boldsymbol{Y = B_0 + B_1X_1 + \cdots + B_PX_P} \quad (1.3)$$

Which is$\boldsymbol{2^k-1}$, can be quite large for example, if ***K=20***, then the number of possible subset model is $\boldsymbol{2^{20}-1= 1,048,575}$which might amount to high computational task of evaluating all possible regressions.
Objective of Equation(1.2) is to determine how well $\boldsymbol{Y}$ may be predicted by each of the subset prediction function in equation (1.3) where $\{\boldsymbol{X_1, \dots, X_P}\}$ is a subset of $\{\boldsymbol{X_1, \dots, X_K}\}$.

Various methods of variable selection have been proposed to focus on whether a single variable should be added to a model (a forward selection method) or whether a single variable should be deleted from a model (a backward elimination method) because of the non-trivial computational problem of the all possible regression procedure.
The combination of the forward and backward selection method which permits re-examination at every step is called stepwise regression procedure (efroymson, 1966; Drapper &smith, 1996; and franklin & Hariharan).

2. Methodologies of the Variable Selection Procedures

The methodologies to be discussed are forward selection procedure, backward elimination procedure, and stepwise selection procedure.

2.1. Forward Selection Procedure

This method starts with the simplest function, namely, $\boldsymbol{B_0}$and successively one variable is been added at a time to the model in such a way that at each step a variable is added. To illustrate the idea, we describe the forward selection procedure using ***K=4***, that is, the total number of predictors under consideration is four which is denoted as $\boldsymbol{X_1, \ldots, X_4}$.
We will describe the forward selection procedure (algorithm) as follows.
At each step of the procedure, we will have a current model, and we will choose a predictor variable that is not already included in the current model as the best candidate variable for adding to that model depending on any model selection criterion measure like $\boldsymbol{R^2}$, ***Cp, adj*** $\boldsymbol{R^2}$,***Cp*** etc. can be used, and each measure will select the same best candidate variable. Whether or not this candidate variable added is actually added to the current model depends on whether a computed quantity, which is denoted by $\boldsymbol{F_C}$ exceeds a criterion value, which we denote by ***F*- in**. The researcher chooses this criterion value to somewhat correspond to a tabled ***F*-Value** with 1 degree of freedom in the numerator and ***n-(p+1)*** degree of freedom in the denominator, where (***P+1*** stands for the number of $\boldsymbol{B_s}$ in the model under consideration). This criterion value ***F-in*** is differently denoted in different statistical software packages. **MINITAB** uses the name ***F*** to enter to refer to the criterion value ***F-in***, SAS use ***P***-values in place of *F*-table values and so on.

The algorithm (process) proceeds as follows.
We begin with the model *Bo* as the current model. We calculate *SSY*, the sum of squared errors when using $\boldsymbol{\bar{Y}}$to predict ***Y.***
Calculation (or a set of calculations) is defined as a step if a predictor variable is added to the current model.
First step, for each predictor variable$\boldsymbol{X_i}$, ($\boldsymbol{i}$ = 1,…, 4) Fit the model

$$\boldsymbol{B_0 + B_iX_i} \quad (2.1)$$

by least squares and obtain ***SSE*** $\boldsymbol{(X_i)}$, the sum of squared errors when using $\boldsymbol{\hat{B}_0 + \hat{B}_iX_i}$ to predict Y.
Choose the variable $\boldsymbol{X_i}$ that will result in the smallest value for ***SSE***$\boldsymbol{(X_i)}$ as the best candidate variable to be added to the current model.
It is observed that the variable $\boldsymbol{X_i}$ with the smallest value of ***SSE*** $\boldsymbol{(X_i)}$ at this step is the same variable $\boldsymbol{X_i}$with the largest value of $\boldsymbol{R^2}$ $\boldsymbol{(X_i)}$ **or**$\boldsymbol{adjR^2}$ $\boldsymbol{(X_i)}$, or the smallest value of ***Cp***$\boldsymbol{(X_i)}$. Thus, any of the criteria can be used in place of ***SSE*** to identify the best candidate variable at each step.
Their results are the same. However, we will use the smallest ***SSE*** to simplify discussion. Let assume the variable to be$\boldsymbol{X_1}$.

$$\boldsymbol{F_C = \frac{SSY - SSE\,(X1)}{MSE(X1)}} \quad (2.2)$$

Where ***MSE*** $\boldsymbol{(X_i)}$ **=** ***SSE*** $\boldsymbol{(X1)/(n-2)}$ and ***SSY*** is the total(corrected) sum of squares of the dependent variable ***SST***. If $\boldsymbol{F_c \leq F\text{-}in}$, then the algorithm stops and the original model $\boldsymbol{B_0}$ is the final model. If $\boldsymbol{F_c > F\text{-}in}$, then add X_1 to the current model which makes it.

$$\boldsymbol{B_0 + B_1X_1} \quad (2.3)$$

We proceed to the second step.
Second step, the current model is$\boldsymbol{B_0 + B_1X_1}$. The predictor variables that is not in this step are$\boldsymbol{X_2, X_3, X_4}$,for ***i***= 2,3,4. At this point,we fit the model

$$\boldsymbol{B_0 + B_1X_1 + B_iX_i} \quad (2.4)$$

and obtain ***SSE*** $\boldsymbol{(X_1, X_i)}$**.** Choose the variable $\boldsymbol{X_i}$ that result in the smallest value for ***SSE***$\boldsymbol{(X_1, X_i)}$ as the best variable to be added to the current model. Assume that the variable is $\boldsymbol{X_2}$. Then we calculate

$$\boldsymbol{F_c = \frac{SSE(X_1) - SSE(X_1, X_2)}{MSE\,(X_1, X_2)}} \quad (2.5)$$

Where,
MSE $\boldsymbol{(X_1, X_2)}$ **=** ***SSE*** $\boldsymbol{(X_1, X_2)}$ ***/ (n-3)***

Like before ,if$Fc \leq F\text{- }in$,the algorithm stops and we choose the model in Equation(2.3) as the final model, but if $Fc > F\text{-}in$, then add X_2 to the current model which makes it

$$B_0 + B_1X_1 + B_2X_2 \quad (2.6)$$

This process continues until all the predictor variables are already included in the current model, which implies that there is no need to proceed further and the algorithm stops.
In this case at the fourth step the final model becomes

$$B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 \quad (2.7)$$

2.2backward Elimination Procedure

This method begins with the present of a constant model B_0 with a model that includes all of the available predictor variables, namely

$$B_0 + B_1X_1 + \cdots + B_KX_K$$

Which proceed by successively eliminating one variable at a time from the model, such that in every step, the variable removed is the variable contributing the least to the prediction of *Y* at that step. The algorithm is illustrated also using $K=4$, that is, $X_1,\ldots,X_4$.

At each step of the algorithm, we will have a current model and will also label a predictor variable included in the current model as the best variable for deletion from the model.
Whether or not this variable is deleted from the current model depends on whether quantity computed which is denoted by F_C is smaller than a criterion value that we call ***F-out***. The model begins with

$$B_0 + B_1X_1 + \cdots + B_4X_4 \quad (2.8)$$

as the current model. The model is fitted using the least square method and calculate $SSE\ (X_1, \ldots, X_4)$,the sum of squared error using $\widehat{B}_0 + \widehat{B}_1X_1 + \cdots + \widehat{\widehat{B}}_4X_4$to predict Y.

First step, X_1, X_2, X_3 **and** X_4 are variables in the model of this step. For each predictor variable X_i,i= 1,,4. Fit the model obtained by deleting this prediction variable from the current model and calculate the corresponding ***SSE***. Leading us to consider the following from models because $K=4$.

$B_0 + B_1X_1 + B_2X_2 + B_3X_3$	Entails that X_4isomitted
$B_0 + B_1X_1 + B_2X_2 + B_4X_4$	Entails that $X_3$is omitted
$B_0 + B_1X_1 + B_3X_3 + B_4X_4$	Entails that $X_2$is omitted
$B_0 + B_2X_2 + B_3X_3 + B_4X_4$	Entails that $X_1$is omitted

And the corresponding ***SSE*** are $\mathbf{SSE}(X_1, X_2, X_3)$, $SSE(X_1, X_2, X_4)$, $SSE(X_1, X_3, X_4)$,$SSE(X_2, X_3, X_4)$ respectively, suppose the first ***SSE*** is the smallest amongst them i.e. $SSE\ (X_1, X_2, X_3)$. Which implies that if we want to delete one of the predictor variable in the current model, the best choice variable to delete will be X_4 because the three remaining predictors X_1, X_2 and X_3 are the best predictor subset models of the current model.
The computed quantity becomes

$$Fc = \frac{SSE(X_1,X_2,X_3) - SSE(X_1,X_2,X_3,X_4)}{MSE(X_1,X_2,X_3,X_4)} \quad (2.9)$$

Where $MSE\ (X_1,\ldots,X_4) = SSE\ (X_1,\ldots,X_4)/(n\text{-}5)$.
If $F_c > F\text{-}out$, then the algorithm stops and the model in equation (2.8) is chosen as the final model, which means that no variables are deleted in the first step and the variable in the model are$X_1, \ldots, X_4$. But if $F_c \leq F\text{-}out$, then $X_4$is deleted from the current model. When $X_4$is deleted from the first step, then

$$B_0 + B_1X_1 + \ldots + B_3X_3 \quad (2.10)$$

is the remaining model containing variables X_1, X_2, X_3.

Second step, since Equation (2.10) is the current model for each predictor X_i, ($i = 1, 2, 3$)
Fit the model obtained by deleting the predictor variable from the current model and calculate the corresponding ***SSE*** that leads us to consider the following three models

$B_0 + B_1X_1 + B_2X_2$ Entails that X_3 Is omitted
$B_0 + B_1X_1 + B_3X_3$ Entails that X_2 Is omitted
$B_0 + B_2X_2 + B_3X_3$ Entails that X_1 Is omitted

Which corresponding ***SSE*** are $SSE(X_1.X_2)$,$SSE(X_1.X_3)$ and $SSE\ (X_2.X_3)$ Respectively. Suppose the smallest amongst these ***SSE*** is $SSE\ (X_1.X_2)$**.** We calculate

$$F_c = \frac{SSE\ (X_1,X_2) - SSE(X_1,X_2,X_3)}{MSE(X_1,X_2,X_3)} \quad (2.11)$$

Where MSE $(X_1, X_2, X_3) = SSE\ (X_1, X_2, X_3) / (n-4)$
If $Fc > F\text{-}out$, then the algorithm stops and Equation (2.10) is chosen as the final model. Otherwise, if $Fc \leq F\text{-}out$, then delete $X_3$from the current model. When $X_3$is deleted in the second step then

$$B_0 + B_1X_1 + B_2X_2 \quad (2.12)$$

is the remaining model containing variables $X_1$and X_2. We proceed till the fourth step where X_1 is deleted and the procedure terminates.

2.3. Stepwise Procedure

The stepwise procedure is the combination of the forward selection and the backward elimination procedure which allows re-examination at every step. Although, there are many versions of stepwise procedures but we will only discuss one in detail.
The predictor variable are $K=4$, then we start with an initial (current) model with no predictors B_0or the full model $B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4$ or any other subset model. The algorithm will proceed in two stages
First stage, we start with the current model and perform the backward selection procedure as many times as is necessary until no more variables can be deleted. If the current model is B_0, omit this stage and go to the second stage.
Second stage, we begin with the final model of the first stage and perform the forward selection procedure once. If a predictor variable is added to the current model, then go back to the first stage.
If no predictor variable is added to the current model at this stage, then the procedure terminates because no variable can be added to the current model and no variable can be removed from the current model. In this case, that current model is selected as the final model.

3. Model Selection Criteria

Many selection criteria for choosing the best model has been proposed. These criteria are based on the principle of parsimony which suggests selecting a model with small residual sum of squares with as few parameters as possible.
Hocking(1976)reviewed eight criteria while Bendel and Afifi (1977) compared also eight criteria but not all the same as hockings.
A selection criterion is an index that can be computed for each candidate model and used to compare models (Kleinbqum et al, 1987). We shall consider four criteria:R^2,MSE, $R^2\ adj$, and Cp**.**

3.1. Multiple Coefficient of Determination

The Multiple coefficient of determination R^2 is the proportion of the total (corrected) sum of squares of the dependent variables explained by the independent variables in the model:

$$R^2 = \frac{SSR}{SSY} = \frac{SSY - SSE}{SSY} = 1 - \frac{SSE}{SSY} \quad (3.1)$$

The objective is to select a model that accounts for as much of the variation in Y. Observe that in the above Equation (3.1) , SSY is the same as SST. The use of the R^2 Criterion for models building requires a judgment as to whether the increase in R^2 from additional variables justifies the increased complexity of the model (Rawlings et al., 1998).

3.2. Mean Square Error

The mean square error (residual) is an estimate of σ^2if the model contains all relevant independent variables. If relevant independent variables have been omitted, then the mean square error is biased. Including an unimportant independent variable will have little impact on the mean square error, as variables are added to the model, is for it to decrease towards σ^2 and fluctuate around σ^2 once all relevant variables have been included.

3.3. Adjusted Multiple Coefficient of Determination

The adjusted R^2, denoted as $adj\ R^2$, is a rescaling of $R^2$by degree of freedom so that it involves a ratio of mean square rather than sum of squares:

$$adj\ R^2 = 1 - \frac{MSE}{MSY} = 1 - \frac{(1-R^2)(n-1)}{(n-p)} = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)} \quad (3.2)$$

3.4. MALLOWS'Cp Criterion

The ***Cp*** criterion was proposed by mallows (1973), and

$$Cp = \frac{SSE_P}{S^2} + 2p - n = \frac{SSE_p}{S^2} + 2(K+1) - n \quad (3.3)$$

Where $S^2$is an estimate of σ^2,n is the number of observation, SSE_p is the sum of squares error(residual)from the ***P*** variable subset model.

4. Numerical Data and Result Analysis

The data in the research work is a secondary data collected from the department of statistics , University of port Harcourt 1998/1999 Academics Session, where the predictive variables are the joint Admission and Matriculation Board (JAMB) subjects offered by the respective students of the department that were given provisional Admission into the university , while the response variableis the grade point aggregate(GPA) of the respective students at the end of the first year that comprises of first and second semester examination.

4.1. Regression Analysis: GPA versus English, Maths, Chemistry, Physics

```
The regression equation is
GPA = 3.51 + 0.00072 ENGLISH - 0.0254 MATHS - 0.0159 CHEMISTRY + 0.0312 PHYSI
CS


Predictor       Coef   SE Coef      T      P
Constant      3.5137    0.7203   4.88  0.000
ENGLISH     0.000718  0.006843   0.10  0.917
MATHS      -0.025421  0.007288  -3.49  0.001
CHEMISTRY  -0.015873  0.006042  -2.63  0.012
PHYSICS     0.031188  0.006444   4.84  0.000


S = 0.345456   R-Sq = 43.3%   R-Sq(adj) = 38.3%


Analysis of Variance

Source          DF      SS      MS     F      P
Regression       4  4.1060  1.0265  8.60  0.000
Residual Error  45  5.3703  0.1193
Total           49  9.4763
```

4.2. Stepwise Regression: GPA versus English, Maths, Chemistry, Physics

```
Forward selection.  F-to-Enter: 4


Response is GPA on 4 predictors, with N = 50


Step             1        2        3
Constant     1.521    2.503    3.550

PHYSICS     0.0267   0.0306   0.0313
T-Value       3.77     4.54     4.94
P-Value      0.000    0.000    0.000

MATHS               -0.0217  -0.0254
T-Value               -2.89    -3.53
P-Value               0.006    0.001

CHEMISTRY                    -0.0159
T-Value                        -2.68
P-Value                        0.010

S            0.390    0.363    0.342
R-Sq         22.87    34.48    43.32
R-Sq(adj)    21.26    31.69    39.62
Mallows C-p   15.2      8.0      3.0
```

4.3. Stepwise Regression: GPA versus English, Maths, Chemistry, Physics

```
Backward elimination.  F-to-Remove: 4


Response is GPA on 4 predictors, with N = 50


Step              1        2
Constant      3.514    3.550

ENGLISH      0.0007
T-Value        0.10
P-Value       0.917

MATHS       -0.0254  -0.0254
T-Value       -3.49    -3.53
P-Value       0.001    0.001

CHEMISTRY   -0.0159  -0.0159
T-Value       -2.63    -2.68
P-Value       0.012    0.010

PHYSICS      0.0312   0.0313
T-Value        4.84     4.94
P-Value       0.000    0.000

S             0.345    0.342
R-Sq          43.33    43.32
R-Sq(adj)     38.29    39.62
Mallows C-p     5.0      3.0
```

4.4. Stepwise Regression: GPA versus English, Maths, Chemistry, Physics

```
  F-to-Enter: 4  F-to-Remove: 3


Response is GPA on 4 predictors, with N = 50


Step              1        2
Constant      3.514    3.550

ENGLISH      0.0007
T-Value        0.10
P-Value       0.917

MATHS       -0.0254  -0.0254
T-Value       -3.49    -3.53
P-Value       0.001    0.001

CHEMISTRY   -0.0159  -0.0159
T-Value       -2.63    -2.68
P-Value       0.012    0.010

PHYSICS      0.0312   0.0313
T-Value        4.84     4.94
P-Value       0.000    0.000

S             0.345    0.342
R-Sq          43.33    43.32
R-Sq(adj)     38.29    39.62
Mallows C-p     5.0      3.0
```

4.5. Stepwise Regression: GPA versus English, Maths, Chemistry, Physics

```
  F-to-Enter: 4  F-to-Remove: 4


Response is GPA on 4 predictors, with N = 50


Step             1        2
Constant     3.514    3.550

ENGLISH     0.0007
T-Value       0.10
P-Value      0.917

MATHS      -0.0254  -0.0254
T-Value      -3.49    -3.53
P-Value      0.001    0.001

CHEMISTRY  -0.0159  -0.0159
T-Value      -2.63    -2.68
P-Value      0.012    0.010

PHYSICS     0.0312   0.0313
T-Value       4.84     4.94
P-Value      0.000    0.000

S            0.345    0.342
R-Sq         43.33    43.32
R-Sq(adj)    38.29    39.62
Mallows C-p    5.0      3.0
```

Using MINITAB statistical software in analyzing the above mentioned procedures, the same subsets of independent (predictor) variables were selected for forward selection procedure, backward elimination procedure and stepwise procedure which includes Maths ***(X_2)***, Chemistry ***(X_3),*** and Physics ***(X_4).***

Their regression equation is given as

- *GPA = 3.55 – 0.0254MATHS – 0.0159CHEMISTRY + 0.0313PHYSICS*

Their root mean squared error (***S***)is 0.342, multiple coefficient of determination (R^2) is 43.32, adjusted multiple coefficient of determination (***adj*R^2**) is 39.62 and Mallows'***C_P*** is 3. Irrespective of the procedure used, ***R^2,*** (***adj*R^2**), S, CP yielded the same result.

5. Conclusion

The forward selection procedure is a version of stepwise procedure since it gives the same result as the stepwise method. The backward elimination procedure performed exactly the same as stepwise procedure in terms of selecting variables. In stepwise procedure, the full model in Equation (2.8) was chosen as the initial model. None of the three methods examines all possible subsets of the procedures. It is often used because it is less computationally expensive compared to the all-subset regression procedure, especially when the number of explanatory variable is so large, that is, for ***k≥20***.

The R^2 method is actually reasonable for the purpose of selection and it gives a clear idea about increase in variation explained by regression equation in terms of adding new variable in the model.

6. References

i. Bendel, R.B. and Afifi, A. A (1977) Comparison of stopping rules in forward "stepwise" regression. Journal of the American Statistical Association, 72: 46-53.
ii. Bowerman, L. Bruce and Richard T. O'Connell (1997) Applied Statistics: Improving Business Process.
iii. Draper, N. R. and Smith, H. (1966) Applied Regression Analysis. Wiley and Sons. New York.
iv. Efroymson, M. A. (1966) Stepwise regression – a backward and forward look. Presented at the Eastern Regional Meetings of the institution of mathematical statistics, Florham park, New Jersey.
v. Graybill, A. Franklin and Haraharan, k. lyer (1994) Regression Analysis: Concepts and Applications.
vi. Hocking, R. R. (1976) The Analysis and Selection of Variables in Linear Regression. Biometrics, Vol. 32.No. 1. Pp. 1-49.
vii. Kleinbaum, G. D., Kupper, L. L., and Muller, E. K. (1988) Applied Regression Analysis and Other Multivariable Methods
viii. Mallow, C. L. (1973) Some Comments on C_P, Technometrics 15, 661-75.
ix. Rawlings, O. John, Pantula G. S., and Dickey, A. D.(1998) Applied Regression Analysis: A Research tool, Second Edition.