A Graphical and Numerical Method for Selection of Variables in Linear Models

Munawar Iqbal
Assistant Professors
Institute of Statistics
University of the Punjab
Q. A. Campus, Lahore

Asghar Ali
Professor of Statistics
Department of Computer Science
Bahauddin Zakariya University (B.Z.U)
Multan

Abstract

A model is usually only an approximation of underlying reality. To access this reality in an adequate way, research all over the world, in different dimensions, is in progress. Most of the diagnostic methods that are being used for the selection of variables to retain in the final model are either based on theoretical methods or they are graphical, that is why model assessing becomes difficult. As a result, the regressors in a model may get very large or very small in their number. The researcher, therefore, has to look at variety of options, and has to fit a lot of models and then is found muddled with the choice to which to select and which to reject. This work is based upon introducing a diagnostic procedure for subset selection due to which one may be successful in reducing the number of possible models to be fitted. This strategy consists of graphical as well as numerical measures; this combination helps much in reducing the number of regressors in the model as well as the number of models. We have also introduced some new approaches and thus a considerable reduction in the regressors by this method does not prohibit the researcher to include regressors of his own interest.

Key words: Subset, Modelling, regressors

1. Introduction

In this article, we propose a strategy for the selection of independent variables in any model. Sometimes it was assumed that the variables which constitute the equation are chosen in advance i.e. independents in the model be fixed a priori. Examining the equation to see whether the function specification and the assumptions about the residuals, fulfill the requirements, cover the whole of the analytical process. In many applications of regression analysis however, the set of independent variables that constitute the model is not pre assumed. In these situations, previous experience in connection with underlying theoretical considerations can help the researcher/ analyst to specify the set of independent variables. Methods and criterion functions for subset selection are critically reviewed by Hocking (1976), Computational algorithms for subset selection are very well discussed by Miller (1984). Use of log linear polynomials very well explained by Ali A (1986). Stepwise Directed search which is a combination of forward selection and the stepwise backward elimination strategy described by Broerson (1984) but still the problem is there. Usually the problem consists of selecting an appropriate set of independent variables from a set that quite likely include all the important variables but we can say, with some extent, that these all are not necessary to adequately model the response y. As Montgomery (2003), “if the objective is to obtain a good description of a given process or to model a complex system, a search for regression equations with small residual sum of squares is indicated”. We have used this fact while formulating our method. Stepwise regression is used to customize the computational efforts. This search method develops a sequence of regression models, at each step adding or deleting an x variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, or F statistic being $F = \left[\dfrac{b_p}{s(b_p)}\right]^2$ the variable with the largest F-value considered to be the candidate for addition in the next stage. We have also used the same idea in combination with the ratio of coefficient of determination and mean square of residuals. We multiplied $\dfrac{\sum_{i=1}^{n}|B_i|}{S.E(B_i)}\dfrac{1}{p}$ with the aforesaid ratio and thus formulated a new criteria. Xavier de Luna and Kostas Skouras (2003) have used the graphical tools on recursive prediction errors in combination with Schwarz’s (BIC) and Akaike’s information criteria (AIC) and proposed “k” potential strategies. It seems to be useful but we are concentrating ourselves to the initial selection of variables. We are not discussing AIC, BIC and many other popular criteria because almost all of these have an extensive theoretical backgrounds. In comparison with all such methods, our strategy doesn’t require any tuff theoretical backgrounds; however, we have made comparisons with very popular Cp criterion because many authors proved it as a better criterion than AIC and BIC. Miller (1990), Fahrmeir, L & Tulz. Gerhard (1994), Mc. Cullagh et al (1989) and almost all statistical scientists unanimously describe that, the number of regressors must be as small as possible and R^2 should be large, relatively. We have considered all of these in our analysis.

While building a model, consideration should also be given to the function specification in variable selection because they both are linked together thus selection of variables or their form, are two problems which should be solved simultaneously, however for simplicity they should be treated sequentially. At the moment we confine ourselves to the selection of the variables not to the specification which is left for further research.

An important situation arises when the investigator have some prior justification for using certain variables (justification may depend upon several factors including exploratory data analysis). Thus a model driven and exploratory driven analysis both be incorporated. So we are interested in screening the potential variables to obtain the model that contain the best subset among them via exploratory analysis. In short, in most of the problems there is no single regression model that is best in terms of various evaluation criteria that have been proposed. A great deal of judgment and experience with the system being modeled is usually necessary to select an appropriate set of independent variables for a regression equation.

2. Methods

2.1 Variable Selection Strategy

Our strategy is very simple and concentrates on the strength of correlation of independent variables(x's) with dependent variable(y) and upon the Multicollinearity of different independent variables.

1. We just include those independents which have significant correlation (at 5% or 1% level) with the dependent variable (they are treated as primary variables) and exclude the independents which don't have significant correlation with dependent variable but have significant correlation with those independents which already have been declared as primary .these rejected variables are the main cause of reducing the total number of models to be fitted.
2. If two primary variables are correlated, then we treat them independently as primary variable but both of them can not appear together in any model.
3. If any pair of variables is significantly correlated and these don't include any of the primary variables then both are included one by one in combination with primary variables, but not both at a time, because of the collinearity between them. In this way, they form two different sets of models i.e. they can combine with other variables which are not mulicollinear with them. If they are "m" pairs they form "m" groups with the same conditions.
4. We include all those variables in the potential models which don't have any correlation with dependent or other independent variables but these included variables are not considered to be the primary part of the model however they are necessary to combine with the primary variables. That is, they should not constitute the model independently without the primary variables but in combination with the primary variables.

In the above paragraphs when we say multicollinearity or the correlation, we mean significant correlation between the two variables.

As for example in the Hald's data out of four independents (x_1, x_2, x_3 and x_4) there should be sixteen possible models and many authors like (Montgomery (2003)) have fitted all the sixteen models and then searched by different criteria the most suitable set of independents in the final model. By our strategy we find that out of these x_1, x_2 and x_4 are significantly correlated with dependent variable but x_2 and x_4 are correlated so in our model x_1 is confirmed and from x_2 and x_4 only one can appear hence we run two separate models

1. y on $x_1 \text{ and } x_2$
2. y on $x_1 \text{ and } x_4$

The above two models are our target models. So we have reduced sixteen models to only two models.

3. y on x_1
4. y on x_2 and
5. y on x_4

Hence the above five models, in total, can be fitted by our strategy because in other combination x_3 may be present are there may be ($x_2 \text{ and } x_4$) all of such combinations have already be rejected by our strategy. We have also applied full model for relative comparisons only.

We have introduced some other criteria (these are explained in Explanation of the terms and methods)

1. C1, Criterion
2. D1, Criterion

These are because for model fitting R^2 should be large, MSE should be small, number of variables should be less and average gain by the independents should be large.

So we have calculated the average gain by the independents as $\frac{\sum_{i=1}^{n}|B_i|}{S.E(B_i)}\frac{1}{p}$ for $1 \le i \le k$ where k represent the total number of independents in any model, and multiplied by C_1 in this way more precise model in the shape of D1, can be attained however, C_1 only can also provide best model.

We have compared our scheme with the other standard procedures like forward selection, backward elimination and stepwise regression. Also we have compared the results give by Neter et al (1987), Montgomery et al (2003) and Anderson & Bancroft (1952) and found that our strategy is simpler and give at least the same results as by other well known schemes. We have used NewR^2 which was first introduced by M.J.R. Healy (1994) in our calculations but it does not help in any improvement.

In order to explain the selection criteria and strategy for inclusion of independent variables, in any model we define the following terms.

2.2 Explanation of the Terms and methods

P= Number of parameters.
MSE= Mean Square of residuals**.**
R^2= coefficient of determination.

C1= $\frac{R^2}{MSE \times P}$ **D1=** $(\text{C1}) \times \frac{\sum_{i=1}^{n} |B_i|}{S.E(B_i)} \frac{1}{P}$

Where $|B_i|$ is the modulus value of ith regression coefficient and $S.E(B_i)$ represent its relevant standard deviation.

PARAM: variables.

The example of Hald's data Definition of Variables:

y: Calories per gram of cement
x_1: Tricalcium aluminate
x_2: Tricalciam silicate
x_3: Tetracalcium aluminoferrite
x_4: Dicalcium silicate

Significant correlation* chart

	y	x_1	x_2	x_3	x_4
y	1	.731**	.816**		-.821**
x_1		1		-.824**	
x_2			1		-.973**
x_3				1	
x_4					1

* Correlation here and afterward mean Pearson's correlation
** Correlation is significant at the 0.01 level (2-tailed)

According to our strategy, x_1, x_2 and x_4 be the primary variables, initially. the possible set of models exclude x_3 because it is correlated with primary variable x_1 and hence potential variables of the model be x_1, x_2 and x_4 however x_4 have strong correlation with x_2 this mean x_1 is compulsory in the model and there is choice between x_2 and x_4. But x_2 and x_4 both should not be included in the model because they are correlation is significant the possible set of models might be only two.

1. y on x_1, x_2.
2. y on x_1 and x_4.

However we include the final set of independent variables for further analysis as

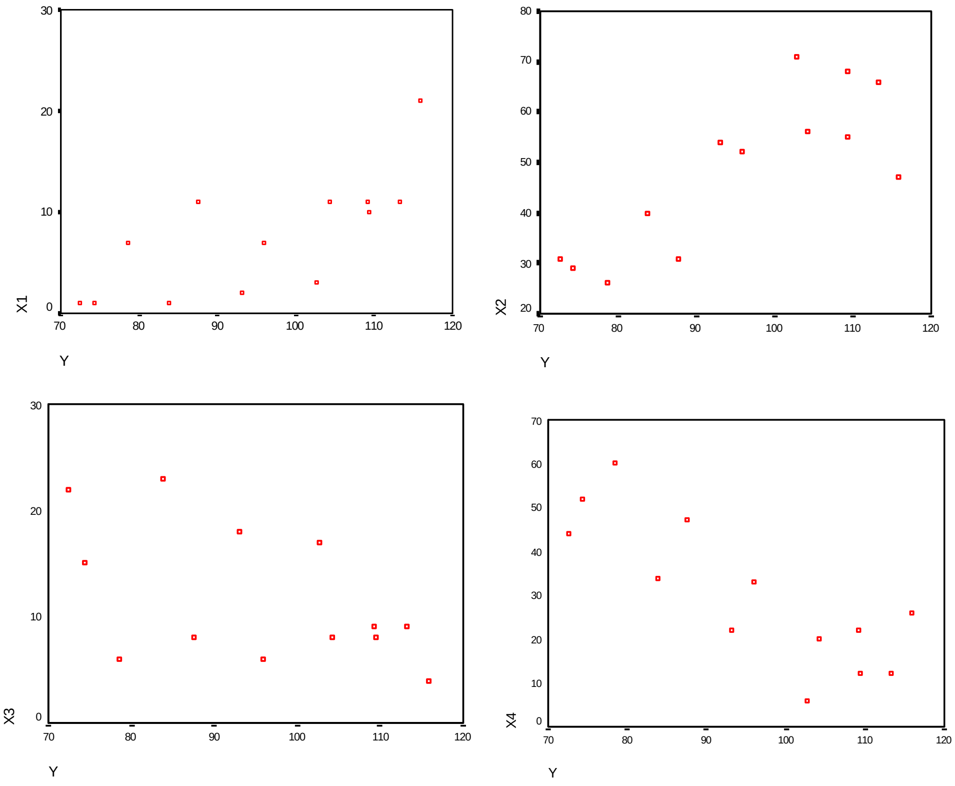
x_1
x_2
x_4
x_1 & x_2
x_1 & x_4

P	MSE	PARAM	R^2	C1	D1
1	115.6	x_1	0.534	0.005	0.016
1	82.39	x_2	0.666	0.008	0.038
1	80.35	x_4	0.675	0.008	0.04
2	5.79	x_1, x_2*	0.979	0.085	1.21
2	7.476	x_1, x_4	0.972	0.065	0.747

* the model, selected.

Montgomery D. C. (2003) have fitted 16 models for the same set of data, he used various methods including BIC, AIC and Cp criteria, and found by fitting 16 models, that the final model consist x_1 and x_2, We have also selected the same by fitting only five models. Montgomery D. C. (2003) have used well known Cp criteria while our's strategy is more simple and easy as compared to Cp criterion.

Scatter diagrams of Hald's Data



While examining the scatter diagrams we see that linear trend is available only in x_1, x_2 and x_4. Scatter diagrams reject the inclusion of x_3 in potential models. So these can be used in initial selection of the variables in a potential model.

NETER's DATA Definition of variables:

y: Survival time, ly;-Log to the base 10 of y
x_1: Blood clotting score
x_2: Prognostic Index
x_3: Enzyme Function test
x_4: Liver Function Test

Significant correlation chart

	ly	x_1	x_2	x_3	x_4
ly	1			.370**	
x_1		1			.502**
x_2			1		.369**
x_3				1	.416**
x_4					1

By our method, most favorite is x_3 and be treated as primary variable. Now the candidates are x_1, x_2 and x_4 which may combine with x_3. Here, x_4 is correlated with x_3 so it is out from the model, now we include x_1 & x_2 with x_3 because neither they are correlated with the Primary variable x_3 nor with one an other. Our proposed model consists of maximum 4 models. They are as under

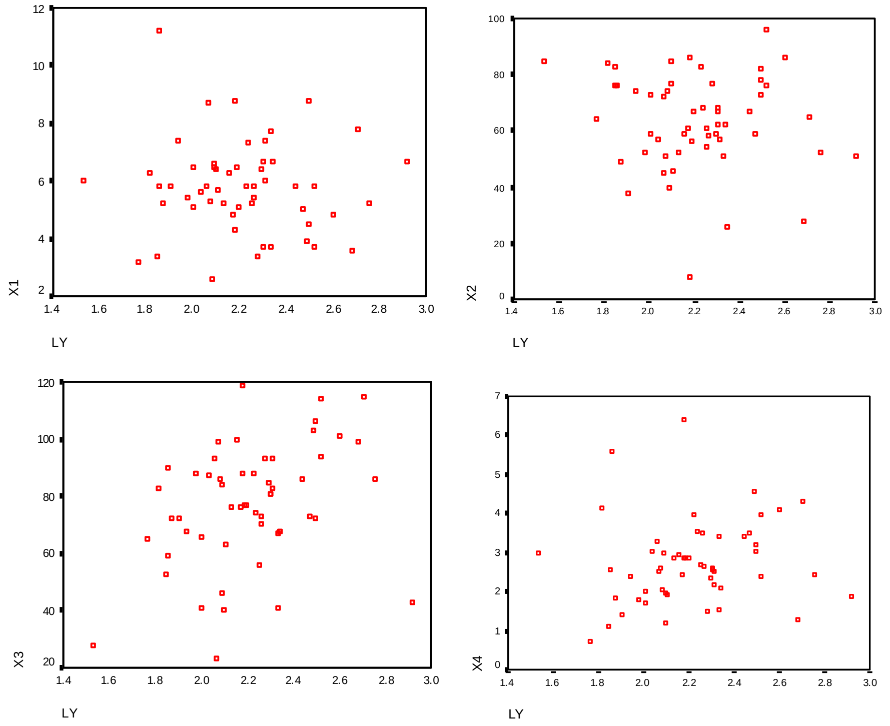
x_3
x_1 & x_3
x_2 & x_3
x_1, x_2 & x_3

P	MSE	PARAM	R^2	C1	D1
1	0.064	x_3*	0.137	2.141	5.352
2	0.066	x_1, x_3	0.137	1.038	1.321
2	0.065	x_2, x_3	0.146	1.123	1.965
3	0.066	x_1, x_2, x_3	0.146	0.737	0.860

* the best model

Neter et el (1987) selects the model x_1, x_2 and x_3 by Cp criterion but in our analysis it is rejected by all our criteria and also by MSE, because MSE from our selected model is less than the Neter's model.

Scatter diagrams of Neter's data



Yes, scatter diagram help like the earlier and we can say that linear trend is available only in x_3.

If we combine the inference from histograms and scatter diagrams we can say that only x_3 can be the member of our final selection.

Anderson and Bancoft's data Definition of variables:

y: Rate of cigarette burn in inches per 1000 seconds
x_1: Percentage of nitrogen
x_2: Percentage of chlorine
x_3: Percentage of potassium
x_4: Percentage of phosphorus
x_5: Percentage of calcium
x_6: Percentage of magnesium

Significant correlation Chart

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1		-.623**	.487*			
x_1		1				-.627**	.604**
x_2			1				
x_3				1		-.588**	-.611**
x_4					1		
x_5						1	.764**
x_6							1

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

By our strategy we can fit only 12 rather than 64 models and our most favorite model must include x_2 and x_3, so these are treated as primary variables. Other possibilities are x_1, x_4, x_5 and x_6 to combine with x_2 and x_3. Since x_5 & x_6 both are correlated with x_3 which is one of the primary variables, so x_5 and x_6 are excluded from the model. And x_1 don't have any correlation with x_4 so it is included in the model. Now we look at x_4 since it is not correlated with any other independent variable so it can be a candidate in possible models. Up to this moment there are only 4 variables in the model named x_1, x_2, x_3 and x_4. Now the required possibilities are only 12 because the models with out the combination with x_3 are also excluded.

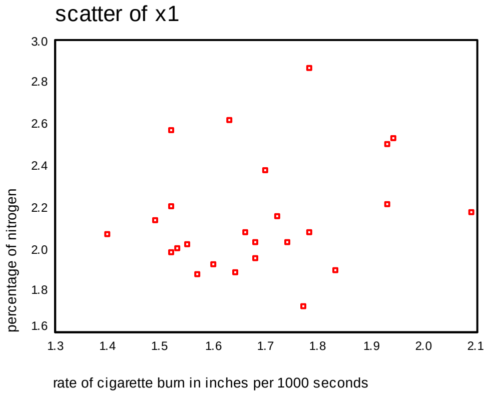
x_2
x_3
x_1 & x_2
x_1 & x_3
x_2 & x_3
x_2 & x_4
x_3 & x_4
x_1, x_2 & x_3

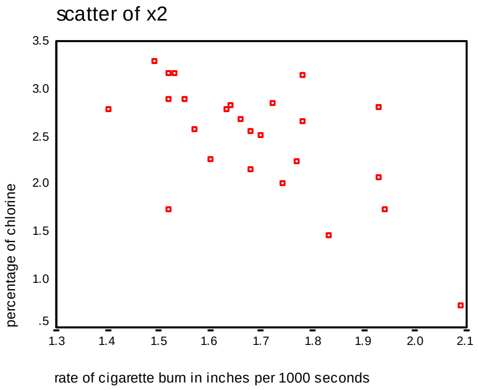
x_1, x_2, & x_4
x_1, x_3 & x_4
x_2, x_3 & x_4
x_1, x_2, x_3 & x_4

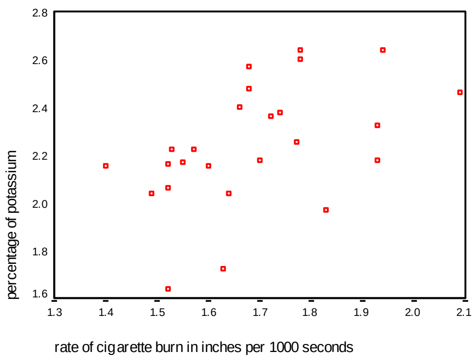
p	MSE	PARAM	R^2	C1	D1
1	0.018	x_2	0.389	21.611	82.598
1	0.022	x_3	0.237	10.773	28.94
2	0.018	x_1, x_2	0.418	11.611	27.773
2	0.022	x_1, x_3	0.29	6.5909	13.160
2	0.013	x_2, x_3*	0.574	22.077	79.655
2	0.016	x_2, x_4	0.464	14.5	40.769
2	0.022	x_3, x_4	0.288	6.5455	11.898
3	0.013	x_1, x_2, x_3	0.606	15.538	44.290
3	0.016	x_1, x_2, x_4	0.485	10.104	21.378
3	0.021	x_1, x_3, x_4	0.33	5.2381	8.211
3	0.012	x_2, x_3, x_4	0.611	16.972	47.467
4	0.012	x_1, x_2, x_3, x_4	0.636	13.25	31.158

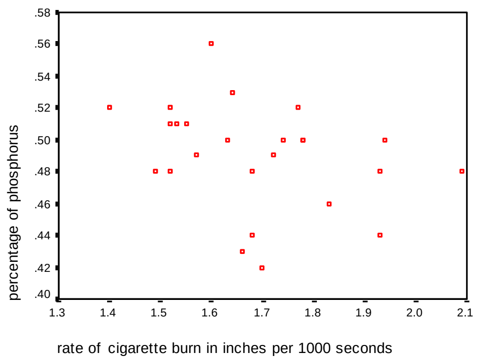
In the table above, MSE is minimum for the set of regressors (x_1, x_2, x_3, x_4) and (x_2, x_3, x_4) but on the behalf of MSE we can not say that the model which possesses only the minimum MSE is considered the best because in the traditional methods also, these sets of independent variables are not considered the best. Method of forward selection which is very well known, also rejects these sets of independent variables, and hence this method supports our strategy which is very simple in the form of C1 and D1. The Cp criterion was used on Anderson and Bancroft's data by Ali A & Al Subaihi (2001) along with some other methods, they selected x_1, x_2 and x_6 as the best set of variables, with no other details.

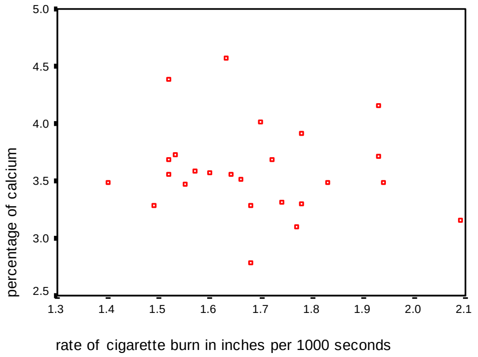
Scatter Diagrams Anderson and Bancoft's data

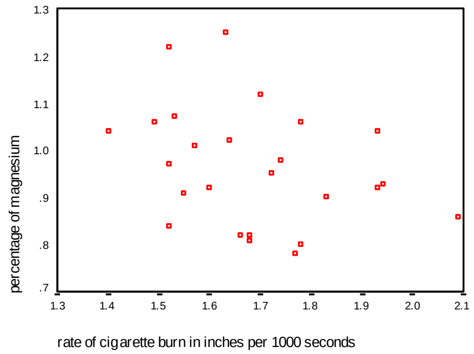





scatter of x3


scatter of x4


scatter of x5


scatter of x6

While examining the scatter diagrams we can say clearly that x_2, x_3 and x_4 have linear trend.

If we combine both the histograms and scatter diagrams, we may fairly say that model include x_1, x_2, x_3 and x_4 and thus in total 2^4 models required to be fitted rather than 2^6

Relative Comparisons by:

HALD's data

P	MSE	METHOD	PARAM	R2	C1	D1
2	5.79	OURS	x_1, x_2*	0.979	0.085	1.121
2	7.476	Forward	x_1, x_4	0.972	0.065	0.746
2	5.79	Backward	x_1, x_2	0.979	0.085	1.121
2	7.476	Stepwise	x_1, x_4	0.972	0.065	0.746

Neter's data

P	MSE	METHOD	PARAM	R^2	C1	D1
1	0.064	OURS	x_3*	0.137	2.141	5.352
1	0.064	Forward	x_3	0.137	2.141	5.352
1	0.064	Backward	x_3	0.137	2.141	5.352
1	0.064	Stepwise	x_3	0.137	2.141	5.352

Anderson and Bancroft's data

P	MSE	METHOD	PARAM	R^2	C1	D1
2	0.013	OURS	x_2, x_3*	0.574	22.08	79.65
2	0.013	Forward	x_2, x_3	0.574	22.08	79.65
3	0.011	Backward	x_2, x_3, x_5	0.645	19.55	47.01
2	0.013	Stepwise	x_2, x_3	0.574	22.08	79.65

*model selected as most suitable, by all our criteria

While comparing all three tables above, we can say easily that our strategy is simpler in application as well as in understanding and give the best possible results while selecting the variables in any model. Although with larger number of regressors it is difficult to decide whether to retain any regressors in the model or to drop it out, but it is applicable and as a result possible number of models reduce dramatically.

We have also proposed the graphical method which is also applicable. Although it is not new strategy because most of the statisticians have suggested it as primary tool but it is presented here as an alternative to some well sophisticated techniques like forward selection, backward elimination and stepwise regression.

Our graphical strategy is not so powerful but the numerical one is quite comparable to the well sophisticated techniques as mentioned earlier.

We can also compare our strategy with well known Cp criterion on Hald's data ,as discussed by Montgomery (2003) and find that our strategy is better than Cp, as in Cp criterion we have to fit 16 models and then to select x_1 & x_2 as regressors but by our strategy, the same is achieved by fitting only 5 models.

We can also make the same comparison on Neter's data and find our strategy, even more suitable, because Neter selects a model consisting x_1, x_2 and x_3 with MSE, equal to 0.066 with sixteen possible models but the model selected by our strategy consists x_1 & x_2 only with MSE equal to 0.064 with total four possible models.

It is thus recommended that Cp criterion may produce better results if applied by using our strategy.

Further research

Although a verity of variables selection methods is in practice today, there is still a plenty of work to be done viewing up the fact we are also on the track of improvement, our strategy may be improved by considering the followings

i) Detection of outliers and their removal, prior to applying our technique will be made.

ii) Use mean of present values in place of missing values if they happen to be in variables.

iii) Adjusted R^2 may be used rather than R^2.

References

1. Akaike, H. (1973), “Information theory and an extension of the maximum Likelihood Principal”, *In B.N. Petrov and F. Csaki ed., 2nd International Symposium on information theory, pp 267-281,Akademia Kiado, Budapest*.
2. Ali, A, Clark G.M and Trustrum K (1986) “Log-linear response functions and their use to model data from plant nutrition experiments” J. Sci. Food Agri., 37, 1165.
3. Ali. A and Al-Subaihi (2001) “Variable Selection in Multivariable Regression Using SAS/IML” *www.jststsoft.org/v07/i12/mv.pdf.*
4. Anderson, R.L. and Bancroft, T.A, (1952),” Statistical theory in research”, *McGraw-Hill book company, Inc., New York, NY. p205*
5. Broerson, P.M.T,(1984) “Stepwise backward elimination with Cp as selection criterion”, *Internal report ST-SV Dept. of Applied Physics, Delft., 84-100*.
6. Fahrmeir, L. and Tulz Gerhard, (1994) “Multivariate Statistical Modelling Based on Generalized Linear Models” Heidelberg*: Springer.*
7. Hald, A (1952), “Statistical theory with engineering applications”, *Wiley New York.*
8. Healy M.J.R. (1994), “ Letter to editor” *vol. 6 .Singapore journal of Statistics pp 147-148*
9. Hocking, R. R. (1976) “The analysis and selection of variables in linear regression” *Biometrics, 32, 1-49*
10. Mallows, C. L., (1973), “Some comments on Cp”, *Technometrics, 15,661-675*.
11. Mc.Cullagh, P. and Nelder, J. A. (1989), “Generalized Linear Models”, *2nd edition, Chapman & Hall, New York*
12. Miller, A. J. (1984) “Selection of subsets of regression variables (with Discussion)”, *J.R. Statist Soc. A, 147, 389-425*
13. Miller, A. J. (1990) “Subset selection in regression”, *London: Chapman and Hall.*
14. Montgomery D. C., Peck, Elizabeth A. and Vining, G. Geoffrey.(2003) “Introduction to Linear Regression Analysis”*3rd ed Wiley & sons(Asia) Pte. Ltd.*
15. Neter J. William Wasserman and Kutner Michal H. (1987), “Applied Linear Statistical Models” *2nd Edition Richard D. Irwin Tokyo Japan.*
16. Schwarz, G. (1978), “Estimating the Dimension of a Model”, Annalas *of Statistics, 6, 461-464.*
17. Xavier de Luna and Kostas Skouras (2003) “Choosing a Model Selection Strategy” *Scand. J. Statist. 30, 113-128*.