On Identification of Relevant Variables to Predict Output of Oilfield Using Support Vector Regression

Desai S. S.
Department of Statistics,
Gopal Krishna Gokhale College, Kolhapur.
Maharashtra State, India.
Pin code : 416012

Abstract:

The production of oil has great significance as a world energy source. Broadly speaking, factors affecting output of oilfield can be classified into two groups namely human factors and geological factors. Each group consists of number of factors affecting output in oilfield. Identifying a prediction model with relevant factors (predictors) is a difficult task in the absence of prior knowledge. This could be done by using subset selection techniques in regression. Mostly, such techniques are based on least squares method (LS). Regression model is fitted under certain assumptions like, independence of predictors; error variable follows normal distribution with constant variance etc. Oilfield output data may not satisfy some of these assumptions and model selection techniques based on LS fail to select parsimonious model. As an alternative we use support vector regression. In this article, we study performance of different model selection techniques for oilfield data when predictors are linearly related.

Keywords: *Oilfield output prediction; Least Squares Method; Multiple linear regression; Support Vector Regression; Subset selection; Mallow's C_p.; Prediction risk.*

1.0 Introduction:

Now-a-days petroleum products have become necessary commodities in the day to day works of life. Oilfield is mother of petroleum products. Production in the oilfield plays a significant role in the economy of a nation. An oilfield is an area under the sedimentary rock with abundance of petroleum or crude oil. Typically an oilfield extends over a large area encompassing hundreds of kilometers with a large number of oil wells. Therefore, prediction of oilfield output based on factors affecting it is essential for the oil industry. Moreover, identifying the relevant factors for the accurate prediction is a serious problem. Regression analysis is a widely used tool for this purpose. Usually, multiple linear regression is employed in such cases. A multiple linear regression model is defined as

$$Y = X\beta + e \qquad (1.1)$$

where Y is known as response variable and is a vector of n observations, $\beta = (\beta_0, \beta_1, \ldots, \beta_{k-1})'$ is a vector of unknown regression coefficients, X is a matrix of order $(n \times k)$ of observations on $k-1$ predictors (regressors) $X_1, X_2, \ldots, X_{k-1}$ with 1's in the first column and e is a vector of errors with following assumptions.

Assumptions:

i. Observations on response variable are independently distributed.
ii. E(e) = 0 and V(e) = $\sigma^2 I_n$, where, I_n is an identity matrix of order n.
iii. e ~ N(**0**, $\sigma^2 I_n$)

In regression, the least squares estimation method is mostly used for parameter estimation. The least squares estimator (LS) of β (Montgomery et al 2006) is given by

$$\hat{\beta} = (X'X)^{-1}X'Y. \qquad (1.2)$$

This method performs well under above assumptions on errors. If these assumptions are violated, estimator in (1.2) may not perform well. Moreover, if the assumptions are satisfied but some of the predictors are linearly related, then data may exhibit problem of multicollinearity. In such situation, estimator given in (1.2) will have large standard error and inference based on it will be misleading. Generally oilfield output depends on previous and current information of some variables. Consequently these variables may be highly correlated to each other.

Oilfield Data :

Let us consider the oilfield output data analyzed by Mustafar et al. (2011). The data contains oilfield output (Y) as response variable and eight different predictor variables as follows.

X_1 : the total number of wells, X_5 : the oil moisture content of previous year,

X_2 : the startup number of wells, X_6 : the oil production rate of previous year,

X_3 : the number of new adding wells, X_7 : the recovery percent of previous year,

X_4 : the injected water volume last year, X_8 : the oil output of previous year.

The multiple linear regression equation fitted to the above data by LS method is

$$\hat{Y} = 2019687 + 178\,X_1 + 218\,X_2 + 194\,X_3 + 0.0768\,X_4 - 54502\,X_5 - 983461\,X_6 + 271927\,X_7 + 0.026\,X_8 \quad (1.3)$$

The predictor variable X_8 : the oil output of previous year may have some linear relationship with the other predictors. Also, it seems that the predictors X_1 and X_2 may be related. To investigate these relationships, we obtained the correlation matrix for predictors.

Correlation Matrix:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1							
X_2	0.9844	1						
X_3	0.9286	0.9345	1					
X_4	0.9887	0.9552	0.9204	1				
X_5	0.8391	0.7578	0.7170	0.8831	1			
X_6	– 0.8646	– 0.7995	– 0.7171	– 0.8989	– 0.9724	1		
X_7	0.9014	0.8357	0.7674	0.9221	0.9613	– 0.9422	1	
X_8	0.9946	0.9714	0.9253	0.9916	0.8481	– 0.8689	0.9169	1

The correlation matrix reveals that there exists strong linear relationship between any two predictors used. This may not confirm the presence of multicollinearity. So we obtained condition indices and variance inflation factor (VIF) for each predictor. The VIFs for X_1, X_2, …., X_8 are 386.1, 112.0, 20.6, 245.0, 114.6, 41.4, 85.7, 391.6 respectively. The condition indices are 1, 4.1321, 7230.6, $6.6945 \times 10^{+05}$, $1.9345 \times 10^{+06}$, $3.3209 \times 10^{+07}$, $4.2189 \times 10^{+012}$, $8.9033 \times 10^{+015}$ and condition number is $8.9033 \times 10^{+015}$.

We observe that severe multicollinearity is present in the data as indicated by high VIF's (e.g. 391.6, 386.1, 245) and condition indices (e.g. $8.9033x10^{+015}$, $4.2189x10^{+012}$). The performance of LS estimator is 'poor' in the presence of multicollinearity in the data. This is pointed out by many researchers. The effects of the presence of multicollinearity on LS estimator are discussed in the standard texts like Montgomery et al. (2006) and Draper and Smith (2003).

In the literature, many techniques are available for dealing with the problem caused by multicollinearity. Ridge regression (Hoerl and Kennard, 1970) and Principal component regression (Marquardt, 1970) are suggested for the estimation purpose. Among these Ridge estimator is widely used for estimating parameters in the presence of multicollinearity. Support vector regression method can also be independently used.

Rest of the paper is organized as; Section 2 gives meaning of subset selection and also describes some methods for subset selection. The performance evaluation of these methods using oilfield data is done in Section 3. Section 4 gives discussion.

2.0 Variable selection in regression:

One of the main objectives of regression analysis is to predict the future value of the response variable using the given values of X_1, X_2, …, X_{k-1} regressors. In practice, the data contains large number of variables for instance, rainfall data, oilfield data, micro array data, socio economic data, etc. A model based on a smaller subset of variables gives more accurate prediction than a model based on a large set (Miller, 2002). A large number of variables are introduced in the earlier stage of analysis and to enhance the predictive ability of the model, some variables are deleted by using some variable selection techniques. Hence, variable selection plays a vital role in regression analysis.

The problem of subset selection is that of searching for the 'best' subset of size 'p' from the all possible subsets such that the selected subset gives an accurate prediction. The literature of variable selection techniques in regression is very rich. An appropriate technique should be used for better results. When the data satisfies all the assumptions mentioned in Section 1, it is said to be clean data. Mallows' C_p (Mallows, 1973), R^2, Adjusted R^2, Sequential procedures (stepwise selection, forward selection and backward elimination), etc. are some of the methods used for variable

selection in clean data. These methods are based on LS estimation procedure. In the literature few methods are available for variable selection in presence of multicollinearity based on ridge estimator such as Rp (Dorugade and Kashid, 2010a) and RG_p (Dorugade and Kashid, 2010b).

In this study, we consider model (1.1) as full model. The fitted equation (in vector notation) is

$$\hat{Y}_k = X\hat{\beta} \qquad (2.1)$$

and the residual sum of squares is defined as,

$$RSS_k = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{ik}\right)^2 \qquad (2.\ 2)$$

We can write model (1.1) as

$$Y\ = X_1\beta_1 + X_2\beta_2 + e.$$

where X_1 is an n × p matrix of the observations on p (≤ k) predictors and β_1 is a p × 1 vector of the regression coefficients.

Here, we consider the subset model as

$$Y = X_1\beta_1 + e\ . \qquad (2.3)$$

The fitted equation (vector notation) for subset model is

$$\hat{Y}_p = X\hat{\beta}_1 \qquad (2.\ 4)$$

and the residual sum of squares for subset model is defined as

$$RSS_p = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{ip}\right)^2 \qquad (2.5)$$

In this article, we consider following variable selection techniques which are used in different scenarios. First two methods are used for clean data and remaining methods are used in the presence of multicollinearity. Below we discuss these methods in brief.

2.1 Mallows' C_p criterion:

Mallow's C_p (1973) is one of the most popular variable selection methods, it is defined as

$$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p) \qquad (2.6)$$

where, RSS_p is the residual sum of squares of subset model, σ^2 is error variance replaced by its suitable estimate (RSS_k / n–k), n is the number of observations and p is the number of parameters in subset model. This method is based on LS estimation

method and as stated earlier, its performance is poor in case of collinear data. However, for clean data its performance is good.

2.2 Method based on significance of regression coefficients:

Another approach to variable selection is to select the variables to be included in the model on the basis of p-values of test for significance of individual coefficients. This approach is suitable in case of clean data. In presence of multicollinearity, p-values may signal in opposite direction.

Variable selection with collinear data

Presence of multicollinearity in the data introduces serious distortions in the analysis. Thus, the data with multicollinearity should be handled carefully. There are two approaches for variable selection in such case.

2.3 Variable selection after removing multicollinearity:

This method is explained in Chatterjee and Hadi (2006). In this method, we delete judiciously the set of variables responsible for multicollinearity in the data, so that the resultant set is free from multicollinearity. Based on the values of variance inflation factor (VIF), variables responsible for multicollinearity are decided and deleted. VIF for predictor X_j is defined as reciprocal of $(1 - R_j^2)$, where, R_j^2 is the multiple correlation coefficient obtained by regressing X_j on all the remaining predictors.

The other approach is to use ridge regression (Hoerl and Kennard, 1970) based method for variable selection in the presence of multicollinearity, which is discussed below.

2.4 R_p criterion:

The problem of multicollinearity has attracted several researchers. Some of them have developed alternative estimators when the multicollinearity is severe. Hoerl and Kennard (1970) proposed the ridge estimator which is widely used because of its optimality properties (see Vinod and Ullah, 1981). It is defined as,

$$\hat{\beta}_R = (X'X + rI)^{-1} X'Y, \qquad (2.7)$$

where, r is the ridge constant or ridge parameter. Hoerl, Kennard and Baldwin (1975) have recommended,

$$r = (k-1)\hat{\sigma}^2 / \hat{\beta}'\hat{\beta} \quad (2.8)$$

where $\hat{\beta}$ is the LS estimator of the β and

$$\hat{\sigma}^2 = (Y'Y - \hat{\beta}'X'Y)/n-k \quad (2.9)$$

Recently, Dorugade and Kashid (2010a) proposed R_p statistic for subset selection based on ridge estimator of β in the presence of multicollinearity. It is defined as

$$R_p = \frac{\sum_{i=1}^{n}(\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\sigma_R^2} - tr(H'H) + tr(H_1'H_1) + p \quad (2.10)$$

where p is the number of parameters in the subset model, σ_R^2 is error variance and is replaced by its suitable estimate $[(Y'Y - \hat{\beta}_R'X'Y)/n-k]$, $= X(X'X + rI)^{-1}X'$, $H_1 = X_A(X_A'X_A + r_A I)^{-1}X_A'$, r_A is ridge constant or ridge parameter for subset model. Note that the matrix H and H_1 are equivalent to hat matrix when LS estimator is used.

2.5 Support Vector Regression and Sp –criterion :

An alternative to above methods is to use a data dependent method such as Support Vector Machine (SVM). The SVM methodology is fast growing area in machine learning. SVM has been introduced by Boser et al. (1992) in COLT. The basic task of SVM is to explore data (input-output pairs) and provide optimally accurate predictions for unseen data. A version of SVM for regression has been proposed in 1997 by Vapnik, Golowich and Smola. This method is called Support Vector Regression (SVR).

In SVR, the goal is to estimate an unknown function based on data ($\mathbf{x}_i$, y_i), i = 1, 2, ….., n of input vectors $\mathbf{x}_i \in R^{k-1}$ and associated targets $y_i \in R$, of the form,

$$y_i = f(\mathbf{x}_i) + e_i, \quad (2.11)$$

where, $f(\mathbf{x}_i)$ is unknown regression function and e_i is error term.

In case of linear regression, the function $f(\mathbf{x}_i)$ is described as follow,

$$f(\mathbf{x}_i) = b + \mathbf{x}_i\mathbf{w} \ , \quad (2.12)$$

where, $\mathbf{w} = (w_1, w_2, \ldots, w_{k-1})' \in R^{k-1}$, $b \in R$ is bias and $\mathbf{x}_i = (x_1, x_2, \ldots, x_{k-1})$.

Therefore, Equation (2.11) becomes,

$$y_i = b + \mathbf{x}_i\mathbf{w} + e_i \ , \qquad i = 1, 2, \ldots, n. \quad (2.13)$$

In matrix notation, we write

$$Y = X\beta + e$$

where, $\beta = (b, w_1, w_2, \ldots, w_{k-1})'$, Y, X and e are the same as defined in Section 1. This equation is equivalent to Equation (1.1).

In SVR, for formulation of optimization problem we use the following 'ε -insensitive loss function' proposed by Vapnik (1995)

$$L_\varepsilon(y_i, f(\mathbf{x_i})) = \text{Max}\{ |f(\mathbf{x_i}) - y_i| - \varepsilon , 0 \} \quad (2.14)$$

where, $\varepsilon > 0$ is a pre-defined constant which controls the noise tolerance. The goal of SVR is to find a function $f(\mathbf{x})$ that has at most ε deviations from the actually obtained targets y_i for all training data at the same time as flat as possible.

Using the ε -insensitive loss function, the regression problem can be written in the form of convex optimization problem (Smola and Schölkopf, 2004) as follows:

$$\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \quad (2.15)$$

$$\text{Subject to:} \quad y_i - (\mathbf{x}_i \mathbf{w} + b) \le \varepsilon , \ i = 1, 2,\ldots,n. \quad (2.16)$$

$$(\mathbf{x}_i \mathbf{w} + b) - y_i \le \varepsilon , \ i = 1, 2,\ldots,n \quad (2.17)$$

The above optimization problem is feasible in case function *f* actually exists and approximates all pairs ($\mathbf{x}_i$, y_i) without error with ε precision,

To cope with infeasible constraints of above problem, we introduce non negative slack variables ξ_i and ξ_i^*, which measure the deviations of training samples outside ε - insensitive zone. The above optimization problem becomes (Vapnik, 1995),

$$\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (2.18)$$

$$\text{Subject to :} \quad y_i - (\mathbf{x}_i \mathbf{w} + b) \le \varepsilon + \xi_i , \ i = 1, 2,\ldots,n. \quad (2.19)$$

$$(\mathbf{x}_i \mathbf{w} + b) - y_i \le \varepsilon + \xi_i^* , \ i = 1, 2,\ldots,n \quad (2.20)$$

$$\text{and} \quad \xi_i , \xi_i^* \ge 0 , \ i = 1, 2,\ldots,n$$

The constant $C > 0$ determines the trade-off between the flatness of *f* and the amount up to which deviations larger than ε are tolerated.

Using Lagrange's multipliers method and exploiting the optimum constraints, the weight vector is given by (Vapnik, 1997, Gunn, 1998),

$$\mathbf{w}' = \sum_{i=1}^{n_{nsv}} (\alpha_i - \alpha_i^*)\mathbf{x}_i \qquad (2.21)$$

and the regression function is given by

$$f(\mathbf{x}) = \sum_{i=1}^{n_{nsv}} (\alpha_i - \alpha_i^*)\mathbf{x}_i\mathbf{x}' + b \qquad (2.22)$$

where, α_i , α_i^* for i = 1, 2,….,n are Lagrange's multipliers and n_{nsv} – number of support vectors. The value of bias b is given by (Gunn, 1998),

$$b = -\frac{1}{2}(\mathbf{x}_r + \mathbf{x}_s)\mathbf{w} \qquad (2.23)$$

where, $\mathbf{x}_r$ and $\mathbf{x}_s$ are the support vectors (i.e. any input vector which has nonzero value of either α_i or α_i^* respectively).

The role of meta parameters C and ε:

The performance of SVR (estimation accuracy) strongly depends on proper setting of regularization parameter (C) and width of insensitive zone (ε). Such parameters are called as meta parameters. Parameter ε controls the width of the ε - insensitive zone used to fit the training data. The ε decides the level of accuracy of the regression function through number of support vectors. To achieve certain accuracy, we need to choose a smaller value of ε to have maximum number of support vectors. Parameter C determines the tradeoff between the model complexity (flatness) and the degree to which deviations larger than ε are tolerated in optimization formulation. Existing methods for selection of meta parameter C are

A priori knowledge and/or user expertise,

C = Range (Mattera and Haykin,1999),

C = $Max(|\bar{y} - 3\sigma_y|, |\bar{y} + 3\sigma_y|)$ (Cherkassky and Ma,2004),

C = PR (Percentile Range) = ($P_{(100+\gamma)/2} - P_{(100-\gamma)/2}$) (Desai and Kashid, 2013)

and C= Max (|Me – 3Q.D.|, |Me + 3Q.D.|) (Desai and Kashid, 2013)

Sp –criterion:

Kashid and Kulkarni (2002) proposed the more general Sp - criterion based on M-estimator (Montgomery et al., 2006, chap. 11) for outlier data. It is defined as,

$$Sp = \sum_{i=1}^{n} (\hat{y}_{ik} - \hat{y}_{ip})^2/\hat{\sigma}^2 - (k - 2p) \qquad (2.24)$$

Where, $\hat{y}_{ik}$ and $\hat{y}_{ip}$ are predicted values based on full model and subset model respectively. Also k and p are the parameters of the full and subset model respectively. Further, note that σ^2 is usually unknown and so it has to be replaced by its suitable estimate.

3.0 Comparison of subset selection methods:

To compare the performance of various subset selection methods, we obtain the mean absolute percentage error (MAPE) defined as,

$$\text{MAPE} = \sum_{i=1}^{n}[(\left|y_i - \hat{y_i}\right|/y_i) * 100]/n \qquad (3.1)$$

In this section, we demonstrate numerically how variable selection methods give misleading results if they are applied without considering the nature of data. We reconsider the oilfield data discussed in section 1 and analyze it by using the methods mentioned in above section.

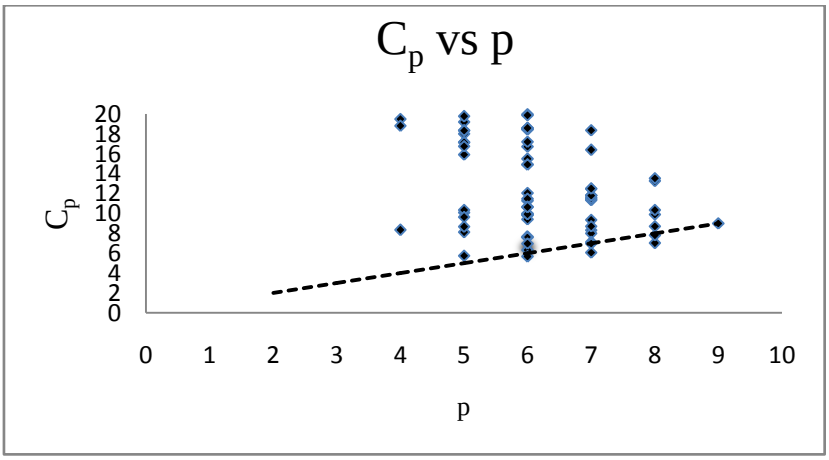
3.1 Variable selection using Mallows' C_p:

We obtain the values of C_p statistic for all possible (2^k –1= 255) subsets. In the following table, we list two values of C_p statistic which are minimum in same size of subsets.

Table 1: Values of C_p statistic.

Predictors in the Model	C_p	p
X_1	60.4	2
X_8	65.3	
X_1X_7	27.7	3
X_1X_8	33.4	
$X_2X_7X_8$	8.3	4
$X_1X_5X_7$	18.8	
$X_2X_4X_5X_7$	**5.7**	5
$X_1X_2X_7X_8$	8.1	
$X_2X_4X_5X_6X_7$	5.7	6
$X_1X_2X_4X_5X_7$	6.3	
$X_1X_2X_4X_5X_6X_7$	6.1	7
$X_2X_4X_5X_6X_7X_8$	7.0	
$X_1X_2X_3X_4X_5X_6X_7$	7.0	8
$X_1X_2X_4X_5X_6X_7X_8$	7.8	
$X_1X_2X_3X_4X_5X_6X_7X_8$	9.0	9

Fig. 1 Plot of C_p vs p



Value of C_p corresponding to predictors $\{X_2, X_4, X_5, X_7\}$ is 5.7 which is close to 5. Hence, according to this method $\{X_2, X_4, X_5, X_7\}$ is proper subset. This fact is also demonstrated through graphically. In Fig. 1, the dotted line represents $C_p = p$ and point denotes value of C_p. Naturally, if C_p is close to p, the corresponding points will be close to the line $C_p = p$. From Fig. 1, it is clear that the subsets for which C_p is close to p are proper subsets. Among these, $\{X_2, X_4, X_5, X_7\}$ is of the smallest size.

3.2 Variable selection using method based on p-value:

In this method, we remove predictors one by one corresponding to larger p value (of test for significance of individual predictor). The same method is used by Mustafar et al. (2011). Here we fix the significance indicator 0.05 and apply this method to oilfield data. The largest p-value is 0.901, which corresponds to X_8, so we remove X_8 from the model and regress Y on remaining predictors. In the same way, X_3 (p value = 0.307), X_1 (p value = 0.210), X_6 (p value = 0.613) are removed in subsequent stages. Finally, the predictors X_2, X_4, X_5 and X_7 remain in the regression model whose p values are significant. So, this method selects the set $\{X_2, X_4, X_5, X_7\}$ as proper subset. The same subset is selected for significance indicator 0.1 and 0.01. The regression coefficients, p – values and VIF values for the subset model are computed and presented in the following table.

Table 2 : p and VIF values for significant predictors.

Predictor	Coeff.	p	VIF
Constant	–259910	0.313	--
X_2	352.21	0	19.5
X_4	0.12302	0	35.7
X_5	–36606	0.001	19
X_7	277024	0	20.5

It is clear that p values in the above table indicate the significance of individual predictors, but VIF values indicate that still the multicollinearity is present in the predictors selected by this method. Mallows' C_p and the p – value based method agree on the importance of same subset because both are based on LS estimator.

3.3 Variable selection after removing multicollinearity:

We obtained VIF values corresponding to all predictors. The VIF value corresponding to predictor X_8 is 319.6, which is larger; we remove X_8 from the model. We obtained VIF values corresponding to remaining predictors X_i, i= 1,2, 3, …, 7 and maximum VIF is 337.2 which corresponds to X_1. So, we remove X_1 in second stage. On the same way we remove X_5 (VIF = 57.7) in third stage, X_4 (VIF = 47.6) in fourth stage and regressed Y on remaining variables X_2, X_3, X_6, X_7. The fitted regression equation is

$$\hat{Y} = -1430935 + 535\,X_2 + 398\,X_3 - 206908\,X_6 + 259257\,X_7 \qquad (3.2)$$

The VIF values for remaining variables X_2, X_3, X_6, X_7 are 11.1, 8.1, 9.1, 10.8 respectively. This indicates that the data doesn't contain severe multicollinearity. Here, we consider model in (3.2) as fitted full model and apply Mallows C_p for variable selection. Following table presents values of C_p statistic for all possible subset models when full model contains the predictors X_2, X_3, X_6, X_7.

Table 3 : The values of C_p for non collinear data.

Predictors in the Model	C_p	p
X_2	115.5	2
X_3	471.4	
X_6	703.8	
X_7	453.9	
X_2X_3	116.9	**3**
X_2X_6	37	
X_2X_7	**3.4**	
X_3X_6	161.4	
X_3X_7	112.3	
X_6X_7	453.4	
$X_2X_3X_6$	35	4
$X_2X_3X_7$	3.2	
$X_2X_6X_7$	5.3	
$X_3X_6X_7$	109.9	
$X_2X_3X_6X_7$	5	5

From above table it is clear that C_p statistic selects the set of predictors $\{X_2, X_7\}$ as proper subset. This method selects the subset model with smaller size as compared to the methods given in 3.1 and 3.2.

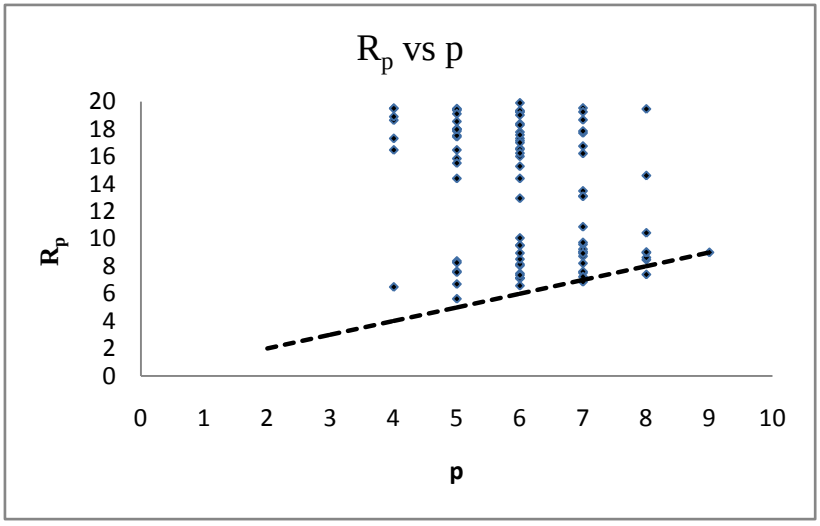
3.4 Variable selection using R_p criterion:

We apply method based on R_p statistic to oilfield data and obtain the values of Rp statistics for all possible (2^k –1= 255) subsets. Two values of R_p statistic, which are minimum in each size of subsets are presented in Table 4.

Table 4 :The values of R_p statistic

Predictors in the Model	R_p	p
X_1	56.23302	2
X_8	58.52541	
X_1X_7	25.60112	3
X_1X_8	28.60801	
$X_2X_7X_8$	6.491387	4
$X_2X_5X_8$	16.46874	
$X_1X_2X_7X_8$	**5.62053**	**5**
$X_2X_4X_7X_8$	6.702801	
$X_2X_4X_5X_7X_8$	6.58359	6
$X_1X_2X_5X_7X_8$	7.125915	
$X_1X_2X_4X_5X_7X_8$	6.848504	7
$X_1X_2X_3X_4X_6X_7$	7.227894	
$X_1X_2X_4X_5X_6X_7X_8$	7.388682	8
$X_1X_2X_3X_4X_5X_6X_7$	8.458443	
$X_1X_2X_3X_4X_5X_6X_7X_8$	9.0	9

Fig. 2 Plot of R_p vs p



The Value of R_p corresponding to predictors $\{X_1, X_2, X_7, X_8\}$ is 5.62053, which is close to 5. Hence $\{X_1, X_2, X_7, X_8\}$ is proper subset. Fig. 2 demonstrates this fact. The VIF values for the selected subset of variables are 229.9, 62.0, 10.6 and 120.5. This indicates that multicollinearity is present in the selected subset. Since, ridge regression is used, it does not affects the prediction ability of the model.

3.5 Variable selection Using Support Vector Regression and Sp-Statistic:

In this method, we use support vector regression to estimate regression coefficients for oilfield data. To perform SVR, we have used meta parameter

C = Max (|Me – 3Q.D.|, |Me + 3Q.D.|) suggested by Desai and Kashid (2013) and

$\varepsilon = C \times 10^{-6}$ (see Gunn,1998). For subset selection we have used more general Sp criterion (Kashid and Kulkarni, 2002). Obtained the values of Sp statistics for all possible subsets using SVR. Two values of Sp statistic which are minimum in each group of equal number of predictors are presented in Table No.5.

Table 5: Values of Sp statistic.

Predictors in the Model	Sp	p
X_1	58.72623939	2
X_8	61.10392123	
X_2X_8	26.50206833	3
X_1X_8	46.70830141	
$X_2X_4X_8$	22.19079591	4
$X_1X_2X_8$	22.43180781	
$X_2X_5X_6X_8$	11.54141709	5
$X_2X_3X_5X_8$	11.76939773	
$X_2X_3X_4X_5X_7$	7.886057118	6
$X_2X_4X_5X_6X_8$	13.65096426	
$X_2X_3X_4X_5X_6X_7$	**7.485941498**	**7**
$X_2X_3X_4X_5X_7X_8$	9.534545572	
$X_2X_3X_4X_5X_6X_7X_8$	7.077579478	8
$X_1X_2X_3X_4X_5X_6X_7$	9.950798094	
$X_1X_2X_3X_4X_5X_6X_7X_8$	9	9

Value of Sp corresponding to predictors $\{X_2, X_3, X_4, X_5, X_6, X_7\}$ is 7.485941498, which is closer to 7 in small subsets. Hence $\{X_2, X_3, X_4, X_5, X_6, X_7\}$ is proper subset.

Interestingly, different methods selected different subsets. In order to compare these subsets and consequently the methods which select them, we assess the mean absolute percentage error. We generated 10000 bootstrap samples each for sample size 5, 10, 15, 20 and 24 from the oilfield data. The MAPE's corresponding to each selected subset are reported in the Table No. 6.

Table 6 : Mean Absolute Percentage Error.

Method	Proper Subset	Bootstrap Sample Size				
		n = 5	n = 10	n = 15	n = 20	n = 24
P-value	$\{ X_2, X_4, X_5, X_7\}$	33.08442	33.07843	33.10258	33.09600	33.12569
Cp	$\{ X_2, X_4, X_5, X_7\}$	33.08442	33.07843	33.10258	33.09600	33.12569
Rp	$\{ X_1, X_2, X_7, X_8\}$	3.493844	3.484232	3.476870	3.484011	3.465856
Chat.	$\{ X_2, X_7\}$	4.815408	4.793653	4.764784	4.774153	4.757813
SVR	$\{ X_2, X_3, X_4, X_5, X_6, X_7\}$	2.777376	2.762925	2.770807	2.775243	2.750963

The subsets $\{ X_1, X_2, X_7, X_8\}$ and $\{ X_2, X_3, X_4, X_5, X_6, X_7\}$ give least MAPE among those considered. R_p criterion and Sp criterion selected the corresponding subsets. Thus, R_p criterion and Sp criterion using SVR estimates perform better than other criteria in the presence of multicollinearity.

4.0 Discussion:

In this article, we discussed the use of subset selection methods for the purpose of building a model less complex in nature but giving higher prediction accuracy in the contest of oilfield data. As discussed earlier, there are many subset selection methods available in the literature. The user of statistics may find difficult to choose one of them. Naturally, if one uses a subset selection method without knowing the nature of the data, then the results may be misleading. It is important to understand the nature of the data and problems associated with it. Based on the problems in the data, an appropriate method of subset selection should be used.

References:

1. Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992): A Training Algorithm for Optimal Margin Classifiers, *ACM COLT 92, Pittsburgh, PA, pp. 144–152.*
2. Chatterjee S. and Hadi A. S. (2006): Regression Analysis by Example, *Forth Edition, John Wiley and Sons Inc, New York.*
3. Cherkassky, V. and Yunqian, Ma (2004): Practical Selection of SVM Parameters and Noise Estimation for SVM Regression, *Neural Networks Vol-17, n-1, 113-126.*
4. Desai S. S. and Kashid D. N. (2013): Estimation of Regression Parameters Using SVM with New Methods for Meta Parameter. (Accepted for publication in), *International Journal of Data Mining, Modeling and Management*
5. Dorugade A. V., Kashid D. N. (2010a): Variable Selection in Linear Regression based on Ridge Estimator. *Journal of Statistical Computation and Simulation. 80 (11), 1211-1224.*
6. Dorugade A. V., Kashid D. N. (2010b): Subset Selection in Linear Regression Using Generalized Ridge Estimator, *Journal of Statistical Theory and Practice, 4, 375-389.*
7. Draper N. R. and Smith H. (2003): Applied Regression Analysis. *Third edition - John Wiley and Sons Inc, New York.*
8. Gunn S. R. (1998): Support Vector Machines for Classification and Regression . *Technical Report. School of Electronics and Computer Science, University of Suthoampton.*
9. Hoerl A. E. and Kennard R. W. (1970): Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics, 12, 55-67.*
10. Hoerl, Kennardand Baldwin (1975): Ridge Regression: Some Simulation, *Computation in Statistics 4, 105-123.*
11. Kashid, D. N. and Kulkarni, S. R. (2002): A more General Criteria for Subset Selection in Multiple Linear Regressions, *Communication in Statistics Theory and Method. 31(5), 795-811.*
12. Mallow's C. L. (1973): Some Comments on C_p, *Technometrics, 15, 661-675.*
13. Marquardt D. W. (1970): Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation, *Technometrics, 12, 591-612.*

14. Mattera, D. and Haykin, S. (1999): Support Vector Machines for Dynamic Reconstruction of a Chaotic System, in : B. Schölkopf, J. Burges, A. Smola, Eds., Advances in Kernel Methods: *Support Vector Learning (pp 211–242), Cambridge, MA, MIT Press.*
15. Miller A. J. (2002): Subset Selection in Regression, *Chapman and Hall.*
16. Montgomery D. C., Peck E. A. and Vining G. G. (2006): Introduction to Linear Regression Analysis. *Third edition - John Wiley and Sons Inc.*
17. Mustafar I. B., Razali R. (2011): A Study on Prediction of Output in oilfield Using Linear Regression. *International Journal of Applied Science and Technology, Vol. 1, No. 4, 107-113.*
18. Smola, A. J. and Schölkopf, B. (2004): A Tutorial on Support Vector Regression. *Statistics and Computing - 14, pp. 199 - 222.*
19. Vapnik, V., Golowich, S. and Smola, A. (1997): Support Vector Method for Function Approximation, Regression Estimation and Signal Processing, *In Mozer M., Jordan M. and Petshe T. editors, NIPS, Vol. 9, pp. 281-287, Cambridge, MA, MIT Press.*
20. Vapnik V. (1998): The Nature of Statistical Learning Theory. *Second Edition, Springer, New York.*
21. Vinod H. D. and Ullah A. (1981): Recent Advances in Regression Methods, *Marcel Dekker, New York.*