**International Journal in Management and Social Science**
Volume 09 Issue 11, November 2021 ISSN: 2321-1784 Impact Factor: 7.088
Journal Homepage: http://ijmr.net.in, Email: irjmss@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal

# A STUDY OF REGULARIZATION, NON-PARAMETRIC REGRESSIONS AND COMPARISONS OF REGRESSION MODELS

**HARIPAL SINGH**
RESEARCH SCHOLAR, SUNRISE UNIVERSITY, ALWAR
**DR. SATENDRA KUMAR**
ASSOCIATE PROFESSOR, SUNRISE UNIVERSITY, ALWAR

## ABSTRACT

If the datasets contain large number of variables, OLS assumption of independence of the variables may get violated. Some of the variables may depend on each other. This problem is called as problem of multicollinearity. In case if multicollinearity, we cannot calculate the coefficients of regression using the method of ordinary least square as matrix (W‟W)-1 is found to be singular with correlated predicted variables, hence unbiased estimates do not exists. The regression estimates calculated in this case have very large values than expected which may lead to incorrect results. Incorrect values of correlation coefficients will produce large values of predictions resulting in the large residual errors. In such cases regularized regression methods are used to deal with multicollinearity and help to manage regression coefficients in order to reduce variance and sample error. Regularization is useful when high multi-collinearity is present among repressors, in case of no. of variables are large than the numbers of observations are less than number of variables. In presence of multicollinearity, high variability in coefficients terms is observed and the coefficients of the correlated variables become over-inflated. So it is expected to identify and remove strongly correlated variables using some tools.

**KEYWORDS:**Regularization, Non-Parametric Regressions, Regression Models, OLS assumption

## INTRODUCTION

When the observation numbers are less than the no. of variables (n<p) then the inverse of the matrix (W'W) does not exists. That means, there is no unique solution set for least square

estimates. This is called as problem of insufficient solutions. In this case, some variables should be removed until we get p<n to fit regression model using OLS method.

In case of large number of variables, we wish to identify smaller subset of variables which exhibits the strongest relation to be included in the model.

In such cases we regularized regression is used to control the estimates of parameter.

Regularization is a regression model with constraint on the magnitude of regression coefficients. Constraint added in the model will help to reduce magnitude and fluctuations of the regression coefficients and decrease variance of model. Regularization method solves multi-collinarity problem by introducing to objective function a penalty term that controls complexity of model.

Regularization is done to shrink the coefficients towards zero to avoid the risk of over fitting.

The objective function with penalty parameter (P) is given by

Minimize [SSE + P]

Following are some models used when collinearity is observed amongst the variables.

**RIDGE REGRESSION**

Mathematically speaking, estimates of regression coefficients using the ordinary least squares

(OLS) method are given by the formula,

$\hat{Y}$ = ( W'W )-1 W' Y

Here W is data matrix of predictor variables and Y is vector of response variable. When the predictors are highly correlated, the matrix (W'W)-1 is near singular hence unbiased estimates does not exists. Ridge regression introduces a ridge parameter Δ and modifies the OLS estimator to ^

$\hat{Y}_\Delta$ = (W'W +Δ I)-1 W' Y

The value of the ridge parameter is very small just enough to remove the singularity of the matrix. If this value is large then the effects of other variables will get inflated. It is obvious that new regression coefficient estimates are no more remained unbiased.

Ridge model retains all variables yet reduce the noise by less influential variables and minimize multicollinearity.

**Mathematical form of Ridge Regression-**

In case of multicollinearity, the huge value and the variability is observed in the regression coefficients, which needs to be stabilized to control the error in the prediction. So the objective is to minimize the errors and the values of estimated regression coefficients. Some of the regression coefficients are positive or some may be negative, they may cancel the effects of each other while adding up the coefficients. That‟s why the squares of the values of the coefficients are taken. Ridge regression uses $\lambda\Sigma\gamma j2$ as a penalty parameter. This parameter is referred as L2 regularization norm. L2 signifies second-order penalty on the coefficients. The objective function becomes,

$$\text{Minimize } [SSE + \lambda\Sigma\gamma j2]$$

With $\lambda = 0$, the objective function is same as normal OLS regression objective function. As value of $\lambda$ increases, the penalty becomes large and forcing coefficients to zero. $\Sigma\gamma j2 < c$ for some positive constant c the last condition is called regularization condition and its effect is restricting regression coefficients within hypersphare with radius c. It is called L2 regularization norm.

The ridge Regression is given by

$$y = \gamma 0 + \gamma 1\ w1 + \gamma 2\ w2 + \ldots + \gamma pwp + \varepsilon$$

Subject to the constraint $\Sigma\gamma i2 < c$ for $c > 0$

**Alternative way to treat multi co-linearity**

In this case of multi co-linearity, we propose to use the recursive partitioning method same as used in decision trees. In this procedure most dominating variable is chosen as partitioning variable, and the data set is partitioned into two mutually exclusive sub groups and the procedure is repeated on these sub groups to finally get homogeneous, mutually exclusive partitioned sets. As we divide the data set according to the most dominating variable, its influence on other variables may get reduce and the new sub sets may not be having multi co-linearity property. For such data sets we can use our usual regression model.

So it is proposed that while processing decision tree algorithm, we use the stopping rule that at every node we calculate (W'W), if it is singular then we stop the procedure otherwise we continue partitioning.

**Important points of Ridge regression.**

1. Ridge regression makes the same assumptions that linear regression makes except for the assumption of normality.
2. Ridge regression shrinks the values of regression Coefficient but does not reduce them to zero. As a result Ridge regression does not lead to variable selection that can avoid multicollinearity.
3. Reason regression is called a regularization method and it uses L2 regularization because it controls the L2 norm of the regression coefficients
4. Depending on the form of Ridge regression model a parameter $\Delta$ or C is known as shrinkage parameter it is also called the bias in parameter due to the fact that it causes the estimates of regression Coefficient to be biased.

**LASSO REGRESSION**

Lasso regression uses the absolute values of regression coefficients instead of squares of regression coefficients as regularization. Lasso Regression is similar to ridge regression except

for the fact that lasso regression results in selection of variables as a result of regularization. The name Lasso is the short form (that is„acronym‟) of the descriptive name "Least Absolute

Shrinkage and Selection Operator". It is obtained by subjecting the regression coefficients to the linear constraint.

$$\text{Minimize } [SSE + \lambda \Sigma| \gamma j |]$$

The constraint is also called as L1 penalty (L1 norm). With this penalty, lasso does variable selection and shrinkage that is; it shrinks the coefficients and set others to zero. This overcomes the limitation of Ridge regression. As a result of using L1 norm, Lasso regression reduces some regression coefficients to zero, leading to the removal of the corresponding variables from the model. Predictor variables that have non-zero coefficients in Lasso regression are the variables that are selected for inclusion in the model. It can be noticed that, larger the penalty applied, closer the parameter estimates get to zero. The model of Lasso Regression is given by

$$y = \gamma 0 + \gamma 1\, w1 + \gamma 2\, w2 + \ldots.. + \gamma p wp + \varepsilon$$

Subject to the constraint $\Sigma\, |\gamma i| < C$ for some $C > 0$

**Important points in the context of lasso regression.**

1. Lasso regression has some assumptions as linear regression except for the assumption of normality.
2. Lasso regression shrinks some of the regression coefficients to zero thus affecting selection of predictor variables for inclusion in the model.
3. Lasso regression is a regularization method that uses the L1 norm for regularization fore if there is a group of highly correlated predictor variables Lhasa regression retains only one of these variables in the model and shrinks coefficient of others to zero.

## ELASTIC NET REGRESSION

Elastic Net Regression combines the regularization conditions (or the constraints imposed) of both methods viz. ridge regression and lasso regression. More precisely Elastic Net minimizes total squared error subject to the following two constraints

Constraint 1 - for $\Sigma |\gamma i| < C$ for some constant $C > 0$

Constraint 2 - for $\Sigma \gamma i2 < D$ for some constant $D > 0$

The objective function can be written as

$$\text{Minimize } [SSE + \lambda 1 \Sigma \gamma j2 + \lambda 2 \Sigma |\gamma j|]$$

Elastic net regression is supposed to inherit benefits if both ridge regression and lasso regression, while avoiding disadvantages or limitations of anyone of the two. Elastic net regression appears to be more of a theoretical interest than of much practical use. The only mentionable application of elastic net regression in the literature is in Support Vector Machines.

## QUANTILE REGRESSION

Quantile regression is linear regression‟s extension used in presence of the outliers, high degree of skewness and heteroscedasticity the objective of quantile regression is to predict the specified quantile of the response variable instead of predicting its arithmetic mean. In particular, median regression is a specific form of quantile regression model. It is very useful in describing the distribution of target variable when it is known to be non-normal and hence cannot be described by only mean and variance. Quantile regression can be useful in estimating the average income of low income group since it is known that income distribution is not normal.

While fitting quantile regression, first, values of regressor variable (W) and that of output (Y) are ordered. Percentiles of W and Y are calculated denoted by w(i) and y(i) are calculated.

Regression equation is then fitted to these percentile values.

In this method, regression equation is fitted to (w (i), y (i) ) and for this we have to break the original pair of w and y.

**PRINCIPAL COMPONENT REGRESSION (PCR)**

PCR is used in presence of multicollinearity or large the number of predictor variables. It is expected to derive low-dimensional data that contains maximum possible information. In case, where no. of variables p is exceeds no. of observations (p > n), n observations are contained in p dimensional space (p principal components) and PCR is linear combination of these p components. Thus one can say that PCA reduces the dimensionality while explaining the most of the variability.

A set of new variables are obtained from original variables such that the new variables (called principal components) are uncorrelated.

i Since principle components are independent of one another, PCR has no problem of multicollinearity.

ii Also, since principal components decrease dimensionality of data set, PCR also achieves reduction in dimensionality of data.

The first principal component (PC) has maximum variance. Second PC is calculated such a way that it is not correlated to first PC and has second largest variance. In same way p number of PCs is derived from p number of variables.

**Construction of principle components-**

Let Correlation Matrix is R

Largest Eigen value is $\lambda 1$

Corresponding Eigen vector is $\eta 1 = (\eta 11 + \eta 12 + \cdots + \eta 1p)$

Then $PC1 = \eta 11\, W1 + \eta 12\, W2 + \cdots + \eta 1p\, Wp$

And Regression of Y on PC1is

$Y = \gamma 0 + \gamma 1\ PC1$

Let $R1 = R - \lambda 1\ \eta 1\ \eta 1'$

Largest Eigen value of R1 is the second largest value of R $\lambda 2$ say.

Corresponding Eigen vector is $\eta 2$

Then $PC2 = \eta 21\ W1 + \eta 22\ W2 + \cdots + \eta 2pWp$

Regression of Y on PC1 and PC2 is

$Y = \gamma 0 + \gamma 1 PC1 + \gamma 2\ PC2$

Computing in the same way, we get

For Eigen value $\lambda p$ and Eigen vector $\eta p$

$PCp = \eta p1\ W1 + \eta p2\ W2 + \cdots + \eta ppWp$

Regression of Y on PC1, PC2, …PCp is

$Y = \gamma 0 + \gamma 1\ PC1 + \gamma 2\ PC2 + \cdots + \gamma pPCp$

Since all the principal components are independent, we can fit the regression line using usual OLS method.

**Illustrative example of Principal Component Regression**

For illustration, the dataset „Boston" is taken from the library „MASS" of R software. The structure or data frame of the Boston dataset is described below.

The response variable is „medv"

There are 506 observations on 14 variables:

1. „crim‟ : crime rate per capita : numeric 0.02731, 0.02729, 0.03237, 0.06905...
2. „indus‟: non-retail business proportion : numeric 2.31, 7.07, 7.0,7, ...
3. „zn‟ : residual land proportion : numeric 18, 0, 0, 12.5, 12.5, 12.5, 12.5...
4. „chas‟ : dummy variable : integer 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ...
5. „age‟ : owner occupied unit proportion : numeric 65.2, 78.9, 61.1,...
6. „nox‟ : concentration of nitric oxide : numeric 0.538, 0.469, 0.469, 0.458..
7. „rm‟ : room number average : numeric 6.58, 6.42, 7.18, 7, 7.15...
8. „dis‟ : distance weighted : numeric 4.09, 4.97, 4.97, 6.06, 6.06...
9. „rad‟ : accessibility index : integer 1, 2, 2, 3, 3, 3, 5, 5, 5, 5...
10. „tax‟ : property tax rate : numeric 296, 242, 242, 222, 222,...
11. „ptratio‟: per town pupil-teacher ratio :numeric 15.3, 17.8, 17.8,...
12. „black‟: proportion of blacks : numeric 397, 397, 393, 395, 397...
13. „lstat‟: lower status percentage: numeric 4.98, 9.14, 4.03, 2.94, 5.33...
14. „medv‟ : owner occupied homes median : numeric 24, 21.6, 34.7, 33.4, …

We will check for the NA and missing observations. Note that the dataset does not contain any missing or NA observations or missing observations, so it can be taken directly for the analysis.

The variables are scaled using the „centre‟ and „scale‟ values and then principal components are computed. The outputs of PC analysis calculations are listed as,

$names

[1] "sdev" "rotation" "center" "scale" "w"

$class

[1] "prcomp"

$sdev

[1] 2.4470390 1.2318221 1.0661141 0.9007497 0.8023388 0.7230142 0.6335543

[8] 0.5055041 0.4627149 0.4368680 0.4269747 0.3687497

**LINEAR DISCRIMINANT ANALYSIS (LDA)**

Also known as Fisher Discriminant Analysis (FDA). It is classification and dimensionality reduction technique. These techniques are highly used in machine learning since high dimensional data sets exist in these days. Linear discriminant analysis is a generalization of Fisher‟s discriminant analysis.

LDA assumes that the predictor variables are normally distributed (Gaussian distribution) and classes have specific means and equal variances/covariance. LDA uses the concept of ratio maximization of between class variance and within class variance to get maximum severability. In other classification techniques, the classes are predetermined and the probability of the observations to get included in that class is calculated. In case of linear discriminant analysis, classes are not predetermined. In the analysis the class boundaries are calculated using the principle of maximizing the ratio of between class variances to the within class variances. And the classification is done using that principle for future.

For simplicity, let‟s considers there be two classes in the data whose boundaries are to be determined. Consider there be two classes with means $\mu_1$ and $\mu_2$ with constant variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

TSS. = $\Sigma\Sigma(w_{ij} - \bar{\mu})^2 = \Sigma\Sigma(w_{ij} - \mu_i + \mu_i - \mu)^2$

$= \Sigma\Sigma(w_{ij} - \mu_i)^2 + \Sigma\Sigma(\mu_i - \mu)^2$ …… (Using assumptions)

= within S.S. + between S.S.

To have both the classes separable, error sum of square within the classes should be minimum and the error sum of squares between the classes should be maximum so as to get distinguish classes. Thus the criterion for better classification will be either

**International Journal in Management and Social Science**
Volume 09 Issue 11, November 2021 ISSN: 2321-1784 Impact Factor: 7.088
Journal Homepage: http://ijmr.net.in, Email: irjmss@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal

(i) Maximizing $\frac{\text{between S.S.}}{\text{within S.S.}}$ or (ii) minimizing $\frac{\text{within S.S.}}{\text{Between S.S.}}$

In second case, if both the classes are very much near and not clearly separable the criterion is not possible as between S.S. is nearly equal to zero. So the classification criteria is to maximize the proportion of between error total of squares to within total squares of errors.

Principle of classification can be generalized for more than two classes.

We can write the classification rule as follows.

Let distance of W from first population mean is small than second population

$$(\underline{W}-\underline{\mu}_1')\,\Sigma^{-1}\left(\underline{W}-\underline{\mu}_1\right) < \left(\underline{W}-\underline{\mu}_2'\right)\Sigma^{-1}(\underline{W}-\underline{\mu}_2')$$

$$\underline{W}'\Sigma^{-1}\underline{W} - 2\underline{W}'\Sigma^{-1}\underline{\mu}_1 + \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 < \underline{W}'\Sigma^{-1}\underline{W} - 2\underline{W}'\Sigma^{-1}\underline{\mu}_2 + \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2$$

$$2\underline{W}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) > \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2$$

$$2\underline{W}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) > \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_2 + \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_2 - \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2$$

$$2\underline{W}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) > \underline{\mu}_1'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) + \underline{\mu}_2'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

$$\underline{W}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) > \frac{1}{2}\ (\underline{\mu}_1'+\underline{\mu}_2')\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

$$\underline{W}'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \frac{1}{2}\ (\underline{\mu}_1'+\underline{\mu}_2')\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) > 0$$

Let $D_{ij}(\underline{W}) = \underline{W}'\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j) - \frac{1}{2}\ (\underline{\mu}_i'+\underline{\mu}_j')\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j)$

If $D_{ij}(\underline{W}) > 0$ for all $j \neq i$, then $\underline{W}$ is assign to population j.

$$D_p^2 = (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

Define $\delta_i^2 = \frac{(\mu_{1i}-\mu_{2i})^2}{\sigma_{ii}}$, i= 1, 2, …, p

Rearrange the variables in the vector W so that the following condition is satisfied.

$\delta_1^2 \geq \delta_2^2 \geq \cdots \geq \delta_p^2$.

Note that $D_p^2$ can be decomposed as follows.

$$D_p^2 = \delta_1^2 + (\underline{\mu}_{1(2)} - \underline{\mu}_{2(2)})' \Sigma_{22.1}^{-1} (\underline{\mu}_{1(2)} - \underline{\mu}_{2(2)}),$$

Where

$$\underline{\mu}_1 = \begin{bmatrix} \mu_{11} \\ \underline{\mu}_{1(2)} \end{bmatrix}_{p-1}^{1}, \qquad \underline{\mu}_2 = \begin{bmatrix} \mu_{21} \\ \underline{\mu}_{2(2)} \end{bmatrix}_{p-1}^{1}$$

$\Sigma = [\quad \bar{} \quad]$ and

$$\Sigma_{22.1} = \Sigma_{22} - \frac{\sigma\,\sigma'}{\sigma_{11}}.$$

$D_p^2 = \delta_1^2 + D_{p-1-1}^2$, so that $= (\_ \quad - \_ \quad)" \quad (\_ \quad - \_ \quad)$

Repeating the same step, we obtain

$$D_{p-1.1}^2 = \delta_{2.1}^2 + D_{p-2.2}^2$$

Where $\delta_{2.1}^2 = \dfrac{(\mu_{12} - \mu_{22})^2}{\sigma_{22.1}}$

And $D_{p-2.2}^2$ is defined in the same way as $D_{p-1.1}^2$ above with p-1 written in the place of p.

The discriminant analysis is useful only when $\mu_1 \neq \mu_2$. It can be shown that

$\underline{\mu}_1 \neq \underline{\mu}_2$ if and only if $D_p^2 > 0$.

When $\mu_1 \neq \mu_2$, it is not necessary that $\mu_{1j} \neq \mu_{2j}$ for all j=1, 2,….,p.

Rearranging the components of W so that $\delta_1^2 \geq \delta_2^2 \geq \cdots\cdots \geq \delta_p^2$ implies that if some of the components of $\mu_1$ and $\mu_2$ are equal, these components will be last in the rearranged vector $\underline{X}$. In other words, if a positive integer k can be found between 1 and p-1 so that $\underline{\mu}_{1(2)} = \underline{\mu}_{2(2)}$, where

$$\underline{\mu}_1 = \begin{bmatrix} \underline{\mu}_{1(1)} \\ \underline{\mu}_{1(2)} \end{bmatrix} \begin{matrix} k \\ p-k \end{matrix}, \quad \underline{\mu}_2 = \begin{bmatrix} \underline{\mu}_{2(1)} \\ \underline{\mu}_{2(2)} \end{bmatrix}$$

Then, it is easy to show that

$$D^2_{p-k\,.k} = (\underline{\mu}_{1(2)} - \underline{\mu}_{2(2)})' \, \Sigma^{-1}_{22.1} \, (\underline{\mu}_{1(2)} - \underline{\mu}_{2(2)}) = 0, \text{ where}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{matrix} k \\ p-k \end{matrix} \text{and} \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\, \Sigma_{11}^{-1}\, \Sigma_{12}$$

If $\underline{\quad} = \underline{\quad}$ then it is easy to find the last p-k components of the random vector $\underline{W}$ do not contribute to the discriminant function. This is same as the statement that the first k components $W_1, W_2 \ldots W_k$ of the random vector $\underline{W}$ are sufficient for constructing the discriminant function. Consequently, only these k variables will be selected for discriminant analysis. This is how feature selection works in the linear discriminant analysis.

## CONCLUSI1ON

The research reported in the present thesis aims at unifying and generalizing the linear regression model as the most general predictive statistical model. As the result of the research carried out for the present thesis leads to some conclusions. Some recommendations can be made for future research on this topic on the basis of these conclusions. Some of the most important recommendations are listed below for the benefit of the readers and future researchers who wish to work on predictive statistical modelling. The study is regarding the choice of the mathematical model. The choice of the mathematical model obviously depends on the nature of the relationship between the independent variable(s) and the dependent variable. Therefore, the choice of the mathematical model can be made only after carrying out some graphical and numerical analysis in order to understand the nature of the relationship between the independent variable(s) and the dependent variable. The graphical analysis includes drawing scatter plots of the dependent or response variable against each and every predictor or independent variable to determine the nature of the relationship between the two. If this relationship is reasonably linear, then the corresponding predictor or independent variable can occur with a constant coefficient in the predictive model. If this relationship does not appear to be reasonably linear, then an appropriate transformation may be necessary on the concerned predictor or independent variable, so that the resulting (transformed) variable has a linear relationship with the dependent of response variable. The numerical analysis will involve calculations of some measures of association between the independent variables and the dependent variable. If all the variables are measured on an interval scale and scatter diagrams show a reasonably linear relationship, then the most common measure of the relationship is given by the coefficient of correlation, popularly known as Pearson's correlation coefficient. If the relationship does not appear to be reasonably linear, then the correlation ratio can be useful. The correlation ratio is a coefficient of non-linear association.

## REFERENCES

1. Feipeng Zhang and Qunhua Li (2016). Robust bent line regression. arXiv: 1606.02234v1 [stat.ME],(2016).

2. Feng Li, Shoumei Li, Nana Tang and Thierry Denoeux (2017). Constrained IntervalValued Linear Regression Model.

3. Francis L. Huang (2014). Analysing Group Level Effects with clustered data using Taylor Series Linearization. Practical Assessment, Research and Evaluation (2014), Vol. 19, No. 13.

4. Francisco Cribari-Neto, AchimZeileis (2009). Beta regression in R. Journal of Statistical Software, Vol. 34, Issue 2.

5. G.Sun, S.J.Hoff, B.C. Zelle, M.A. Nelson (2008). Development and Comparison of Back Propagation and Generalized Regression Neural Network. Agriculture and Bio-systems Engineering Publications, Vol.2, No.4, 685-694.

6. Gary King and Michael Tomz (2000). Making the Most of Statistical Analysis. American Journal of Political science, Vol. 44, No. 2, Pp 341-355.

7. Gonzalo Mateos and Georgios B. Ginnakis (2010). Distributed Sparse Linear Regression. IEEE transactions on signal processing, vol.58, No.10, 5261-5276.

8. GuangliNie, Wei Rowe, Lingling Zhang and YingjieTian (2011). Credit Card Churn Forecasting by Logistic Regression and Decision Tree. Expert systems with Applications, Vol. 38, pp 15273-15285.

9. HaifengZheng, Liding Chen, Xiaozeng and Yan Ma (2009). CART for analysis of soybean yield. Agriculture, Ecosystems and Environment, Vol. 132, pp. 98-105.

10. Halbert White (1980). A Heteroscedasticity- Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. Econometrica, Vol. 48, No. 4, pp. 817-838.

11. Hans- Geprg, Muller &Alrich, Stadt Muller (2005). Generalized Functional Linear Models. The Annals of Statistics, Vol. 33, No. 2, 774 - 805.

12. Hansheng Wang (2007), Regression Coefficient and Autoregressive Order Shrinkage and Selection via Lasso. Journal of the Royal Statistical Society Series B, Vol. 69, issue 1, pp 63-78.

13. HarvinderChauhan, AnuChauhan (2013). Implementation of decision tree algorithm C4.5. International Journal of Scientific and Research Publications, Vol.3, Issue 10.

14. Helmut Kuchenhoff (1996). An exact algorithm for estimating break points in segmented generalized linear models. Computational statistics, Vol. 12, No. 2.

15. Henriette Koch, PoulJennum and Julie A.E. Christensen (2018). Automatic sleep classification using adaptive segmentation reveals an increased number of rapid eye movement sleep transitions. Journal of Sleep Research.