Regression Analysis

Sub. Code.: MTH416A
Semester II

Project

A Study of Psychochemical Properties of Protein Tertiary Structure

Sayan Bhowmik - 201409
Suryasis Jana - 201447
Arghyamalya Biswas - 201275
Soumyadip Sarkar - 201431
Shubha Sankar Banerjee - 201416

Under Supervision of
Dr. Sharmishtha Mitra

Department of Statistics
Indian Institute of Technology Kanpur

Acknowledgement

Real learning comes from a practical work. We are very grateful to our instructor Dr. Sharmistha Mitra (Regression Analysis- MTH416A) for providing all of us such an opportunity to be engaged in a practical project work based on our own area of interests under the context of Regression Analysis and its applications.
We are sincerely thankful to our instructor and project guide Dr. Sharmistha Mitra, ma'am for her continuous guidance, monitoring and constant support throughout the course.

Contents

1 Introduction

Proteins are one of the most important molecules in the living organisms so they play vital structural role in the cells of living organisms. They are constructed of several polypeptide chains of amino acids, which fold into complex tertiary structure. The knowledge of the protein function is directly dependent on its three dimensional (tertiary) structure.
In this project we focus on explaining the dependence of Root Mean Square Deviation which is used as a quantitative measure of similarity between two or more protein structures on different Psychochemical attributes related to a Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment.
The main objective of the project is to properly explain protein's tertiary structure which we have taken as RMSD (Root Mean Square Deviation) by a linear model of 9 other components given as F1, F2, F3, F4, F5, F6, F7, F8, F9 using different tools of regression analysis. We aim to model the relationship between response and the regressors using an MLR model. Along with an empirical evaluation, we will be performing a detailed mathematical analysis of different aspects of the model to establish its validity. This includes Residual analysis to check if our model assumptions hold true, tests for checking Multi-collinearity to check if there exists any near-linear relationship among any subsets of the regressors, tests of significance to guarantee that the derived model parameters do make sense, and also variable selection to see if we can get the same result as our full model using only lesser number of regressors. To summarize, we are aiming to build a model of RMSD by identifying the most important factors (or regressors) through the regression analysis and model diagnostics.To have make analysis as exhaustive as our knowledge on regression can permit, we are motivated to apply certain more tools of analysis which we will be covering over time in the Regression Analysis (MTH416A) course.

2 Data Description

This is a dataset of Physicochemical Properties of Protein Tertiary Structure. Data set is taken from CASP 5-9. There are a total of 45730 decoys and size varying from 0 to 21 angstrom.
The variable of concern under the setup are listed as follows:

- RMSD- Size of the residue
- F1- Total Surface area
- F2- Non polar exposed area
- F3- Fractional area of exposed non polar residue
- F4- Fractional area of exposed non polar part of the residue
- F5- Molecular mass of weighted exposed area
- F6- Average deviation from standard exposed area of residue.
- F7- Euclidean distance
- F8- Secondary structure penalty
- F9- Spatial Distribution constraints (N, K- value)

The data is of available in the following form:

	RMSD	F1	F2	F3	F4	F5	F6	F7	F8	F9
1	17.284	13558.30	4305.35	0.31754	162.1730	1872790.5	215.3590	4287.87	102	27.0302
2	6.021	6191.96	1623.16	0.26213	53.3894	803446.7	87.2024	3328.91	39	38.5468
3	9.275	7725.98	1726.28	0.22343	67.2887	1075647.6	81.7913	2981.04	29	38.8119
4	15.851	8424.58	2368.25	0.28111	67.8325	1210471.6	109.4390	3248.22	70	39.0651
5	7.962	7460.84	1736.94	0.23280	52.4123	1021019.7	94.5234	2814.42	41	39.9147
6	1.700	5117.30	1120.99	0.21905	51.6732	672722.7	79.5911	3234.21	15	41.2382
7	9.314	5924.16	1625.27	0.27434	70.2103	828514.5	76.8064	2821.40	70	39.4964
8	1.985	6882.15	1791.22	0.26027	77.2501	916516.5	96.6785	3490.88	74	37.4203
9	1.915	12090.00	4190.74	0.34662	129.0020	1687508.0	186.3090	4262.78	39	30.3916
10	1.495	7400.24	1881.95	0.25430	82.9320	1023845.6	104.6970	3852.40	26	35.4140
11	12.118	6556.77	1612.77	0.24597	71.6315	891544.3	93.5329	3161.33	76	38.0433
12	0.884	8828.21	2658.63	0.30115	90.8578	1233384.3	123.6860	3194.30	22	37.2413

Figure 1: Glimpse of the dataset

2.1 Exploratory Analysis

Exploratory Data Analysis is an approach to summarize the main characteristics of the data set, often with visual methods. The primary objective of this is to see what the data can tell us other than formal modelling or hypothesis testing task.

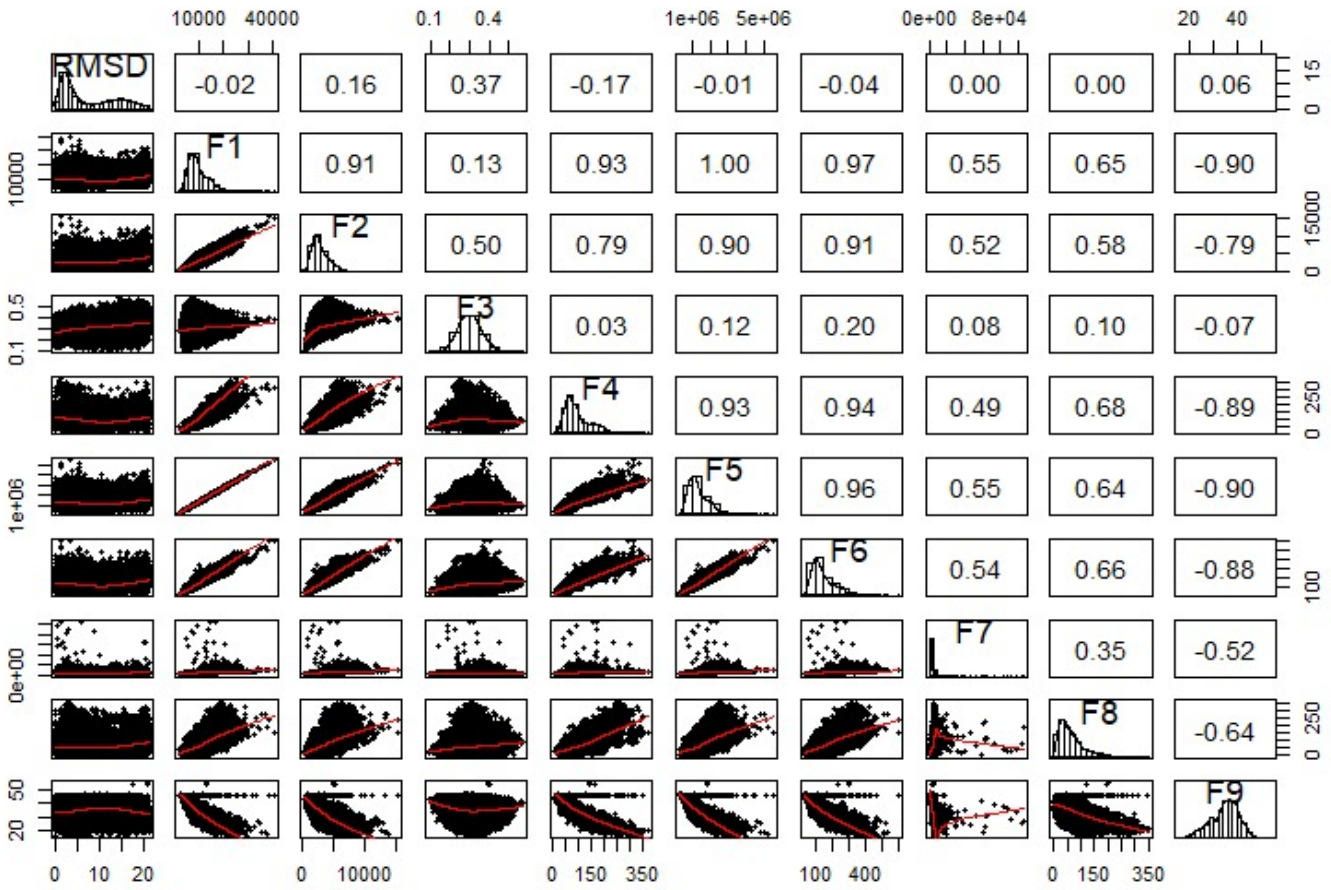

Figure 2: The pairs plot for all the variables under consideration

We observe the following:

- The response variable(RMSD) is positively skewed which indicates the presence of some outliers.
- Among the regressors, except for F3 and F9, all the variables are positively skewed where as F3 and F9 are almost symmetrically distributed.
- None of the individual regressor is found to be highly correlated with the response variable.
- However some of the pairs of regressors are supposed to be highly correlated viz
  - F1 has high positive correlation with F2,F4,F5,F6,F8 and high negative correlation with F9. Moreover F1 is perfectly correlated with F5 in a positive way.
  - F2 is positively correlated with F4,F5,F6 and negatively correlated with F9.
  - F4 has positive correlation with F5,F6,F8 and negative correlation with F9.
  - (F5,F6),(F5,F8) pairs have positive correlation and (F5,F9) has negative correlation.
  - F6 is positively correlated with F8 and negatively correlated with F9.
  - And finally F8 and F9 share a negative correlation.

3 Multiple Linear Regression

We assume that we are provided with n sets on observations on $RMSD, F_1, F_2, ..., F_9$. Our multiple linear model involves $p = 9$ regressors $x_1 = F_1, ..., x_9 = F_9$ and an intercept term. The response variable being y=RMSD. The **MLR** model is:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + \epsilon_i, \ \forall \ i = 1, ..., n$$

Where, ϵ is the error term in the model, with the following assumptions:

$$\begin{aligned} E[\epsilon_i] &= 0, \ \forall i = 1(1)n \\ Var[\epsilon_i] &= \sigma^2, \ \forall i = 1(1)n \\ Cov[\epsilon_i, \epsilon_j] &= 0, \ \forall i \neq j \end{aligned}$$

3.1 Normal Equations

We can write the above stated MLR equations in the matrix form as follows:

$$Y = X\beta + \epsilon$$

Where,

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \ X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We want to find the estimate of β from the given data. We will apply the least squares technique to obtain the estimates. The technique involves minimizing the Sum of Squares of errors with respect to β i.e. to minimize the following function:

$$S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$$

Differentiating the above equations with respect to β, we get the **Least Squares Normal Equations** of out MLR model, given as:

$$X'X\beta = X'Y$$

Thus the least squares estimates of out MLR model is given by:

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

provided $(X'X)^{-1}$ exists.

3.2 Least Squares Estimates

The estimated values of the parameters obtained are as follows:

```
Call:
lm(formula = RMSD ~ F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 + F9)

Coefficients:
(Intercept)           F1           F2           F3           F4           F5           F6
  6.037e+00    1.572e-03    1.429e-03    1.803e+01   -1.082e-01   -4.075e-06   -2.388e-02
         F7           F8           F9
 -1.387e-04    1.485e-02   -1.101e-01
```

Figure 3: Least Squares estimates of Model parameters

Thus the estimated model stands:

$$Y = 6.037 + 1.572 \times 10^{-3}X_1 + 1.429 \times 10^{-3}X_2 + 1.803 \times 10^{1}X_3 - 1.082 \times 10^{-1}X_4 \\ - 4.075 \times 10^{-6}X_5 - 2.388 \times 10^{-2}X_6 - 1.387 \times 10^{-4}X_7 + 1.485 \times 10^{-2}X_8 - 1.101 \times 10^{-1}X_9$$

3.3 Tests for significance of Regression

This test is performed to check whether atleast one of the regressors have linear relationship with the response or not. The testing hypothesis is:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \; vs \; H_A : \; \beta_i \neq 0 \; for \; some \; i \in \{1, 2, ..., p\}$$

We will be using the Analysis of Variance method to make a conclusion on the test. We make use of the following results:

- $SS_{Total} = SS_{reg} + SS_{res}$
- $SS_{reg}/\sigma^2 \sim \chi^2_p$ where p is the number of regressors
- $SS_{res}/\sigma^2 \sim \chi^2_{n-p-1}$
- SS_{reg} and SS_{res} are linearly independent
- $F_{stat} = \frac{SS_{reg}/p}{SS_{res}/(n-p-1)} = \frac{MS_{reg}}{MS_{res}} \sim F_{p;\, n-p-1}$

The summary of the model is given as follows:

```
Call:
lm(formula = RMSD ~ F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 + F9)

Residuals:
     Min       1Q   Median       3Q      Max
-24.1639  -4.0934  -0.9466   3.8620  20.0977

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.037e+00  5.876e-01  10.274  < 2e-16 ***
F1           1.572e-03  1.231e-04  12.764  < 2e-16 ***
F2           1.429e-03  1.130e-04  12.653  < 2e-16 ***
F3           1.803e+01  1.065e+00  16.927  < 2e-16 ***
F4          -1.082e-01  1.645e-03 -65.773  < 2e-16 ***
F5          -4.075e-06  7.550e-07  -5.397 6.79e-08 ***
F6          -2.388e-02  1.752e-03 -13.628  < 2e-16 ***
F7          -1.387e-04  1.481e-05  -9.361  < 2e-16 ***
F8           1.485e-02  5.933e-04  25.035  < 2e-16 ***
F9          -1.101e-01  1.000e-02 -11.013  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.184 on 45720 degrees of freedom
Multiple R-squared:  0.2823,    Adjusted R-squared:  0.2822
F-statistic:  1998 on 9 and 45720 DF,  p-value: < 2.2e-16
```

Figure 4: Summary of the Estimated MLR Model

We can clearly see that the p-values of all the parameter estimates i.e. $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_9$ are less than 0.05. Thus at 95% of significance we can conclude that all the paramater estimates are statistically significant.

The ANOVA table is thus as follows:

Source of Variation	df	SS	MS	F_{stat}
Regression	9	483256	53695.11	1998.2
Residual	45720	1228552	26.8712	
Total	45729	1711807		

We see that $F_{stat} > F_{0.05;\ p,n-p-1}$. Thus we reject the null hypothesis in favour of the alternate hypothesis and conclude that out linear regression line is significant.
R^2 and Adjusted R^2 are used to explain the overall adequacy of the model, where,

$$\begin{aligned} R^2 &= 1 - \frac{SS_{Res}}{SS_{Tot}} \\ R^2_{Adj} &= 1 - \frac{SS_{res}/(n-p-1)}{SS_{Tot}/(n-1)} \end{aligned}$$

We also see that the value of Coefficient of Determination i.e. R^2 turns out to be 0.2823. It implies that about 28.23% of the variation in observed response variable i.e. RMSD is explained by the assumed Multiple Linear Regression Model. The Model is thus not very efficient in explaining the observed responses.

4 Model Diagnostics

Goodness of Fit indicators like R^2, the χ^2, t or F statistics are useful in testing model adequacy. However, these measure do not shed light on whether the normal error assumptions are violated or not.

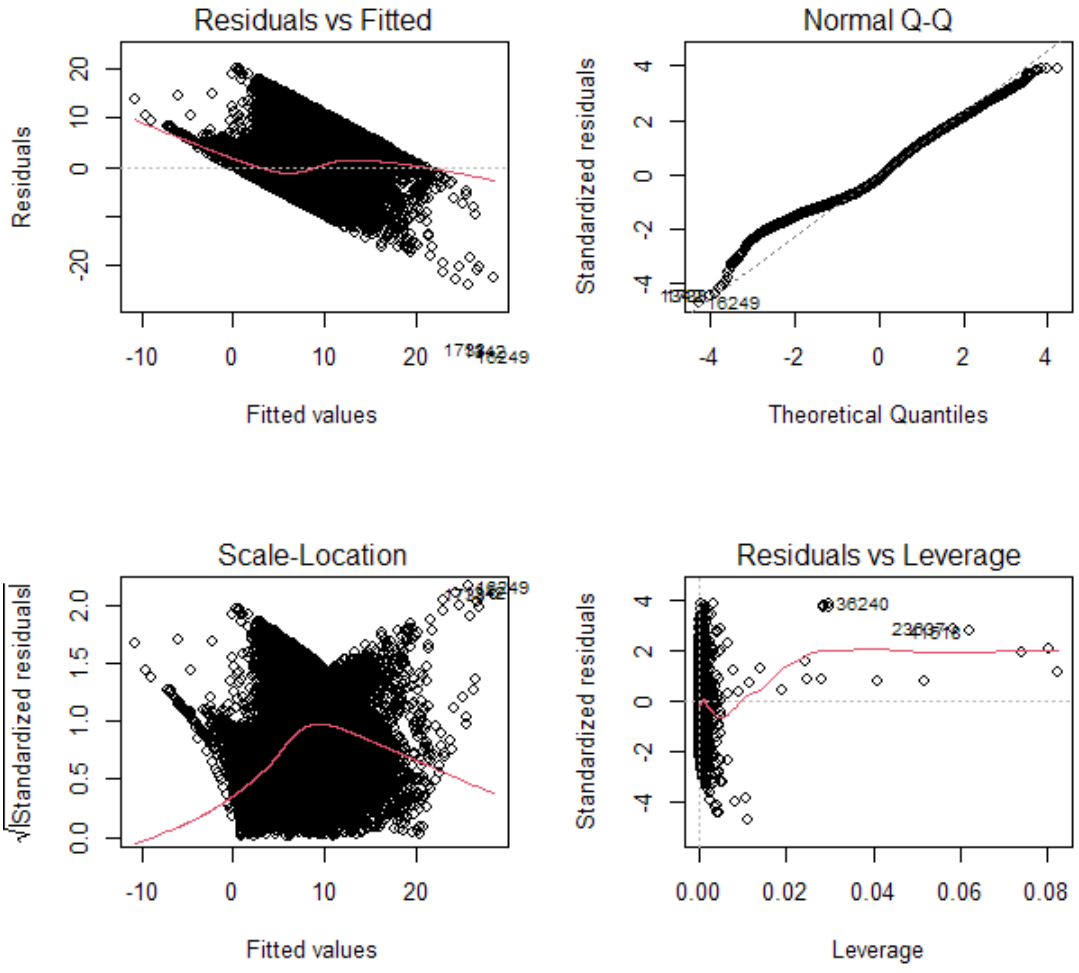

Figure 5: Plots on which we will base our analysis on

4.1 Residual Analysis

We already have our estimated model as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 + \hat{\beta}_7 X_7 + \hat{\beta}_8 X_8 + \hat{\beta}_9 X_9$$

which gives our estimated response values $\hat{Y}_i,\ \forall i = 1(1)n$.
Thus the residual is calculated by:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i,\ \forall i = 1(1)n$$

We will prepare a scatterplot of residuals $(\hat{\epsilon}_i)$ vs the fitted values $(\hat{Y}_i)$. If a pattern is obvious, we can conclude that error variance is non-constant indicating that the model is heteroscedastic.

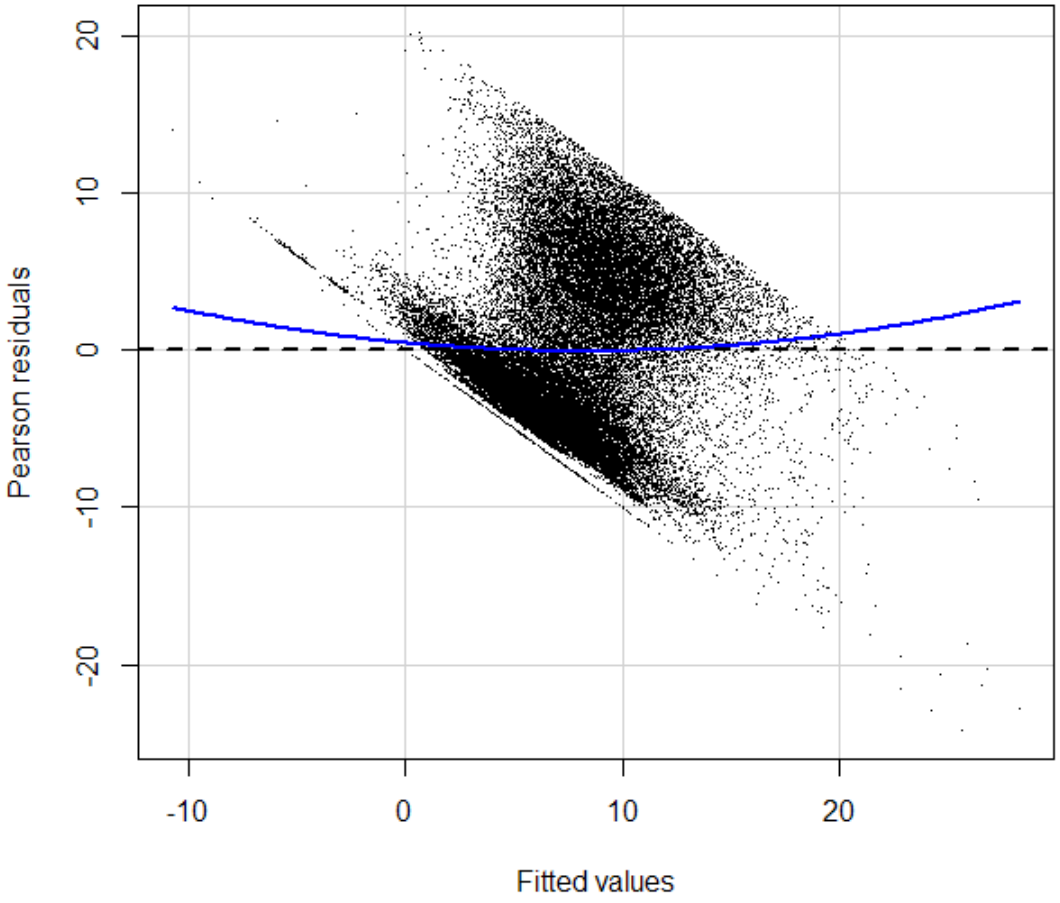

Figure 6: Residual Plot for MLR model

We can clearly see that a distinct pattern is apparent. It indicates the presence of heteroscedasticity, i.e., the errors vary with one or more regressors.

Next we plot the residuals against each regressor to check if we can get an idea of the nature of heteroscedasticity.

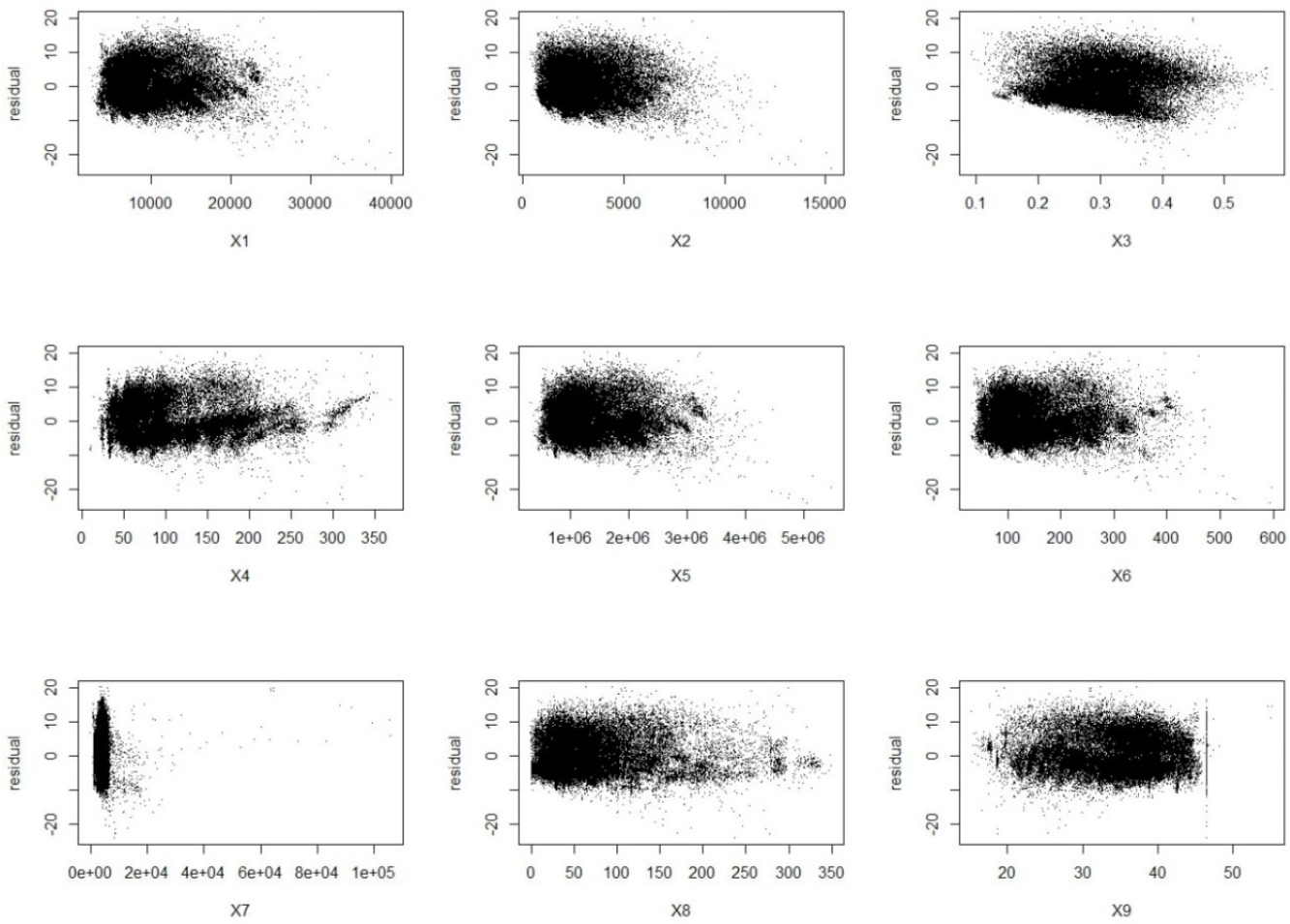

Figure 7: Plot for residual vs each regressor

We can see that there seems to be some form of dependence of residuals on the regressors. However the nature of dependence on any of the regressors is not clear.

4.2 Testing Normality of Error Assumption

1. **Graphical Method:** We first check the Q-Q plot for the standardized residuals.

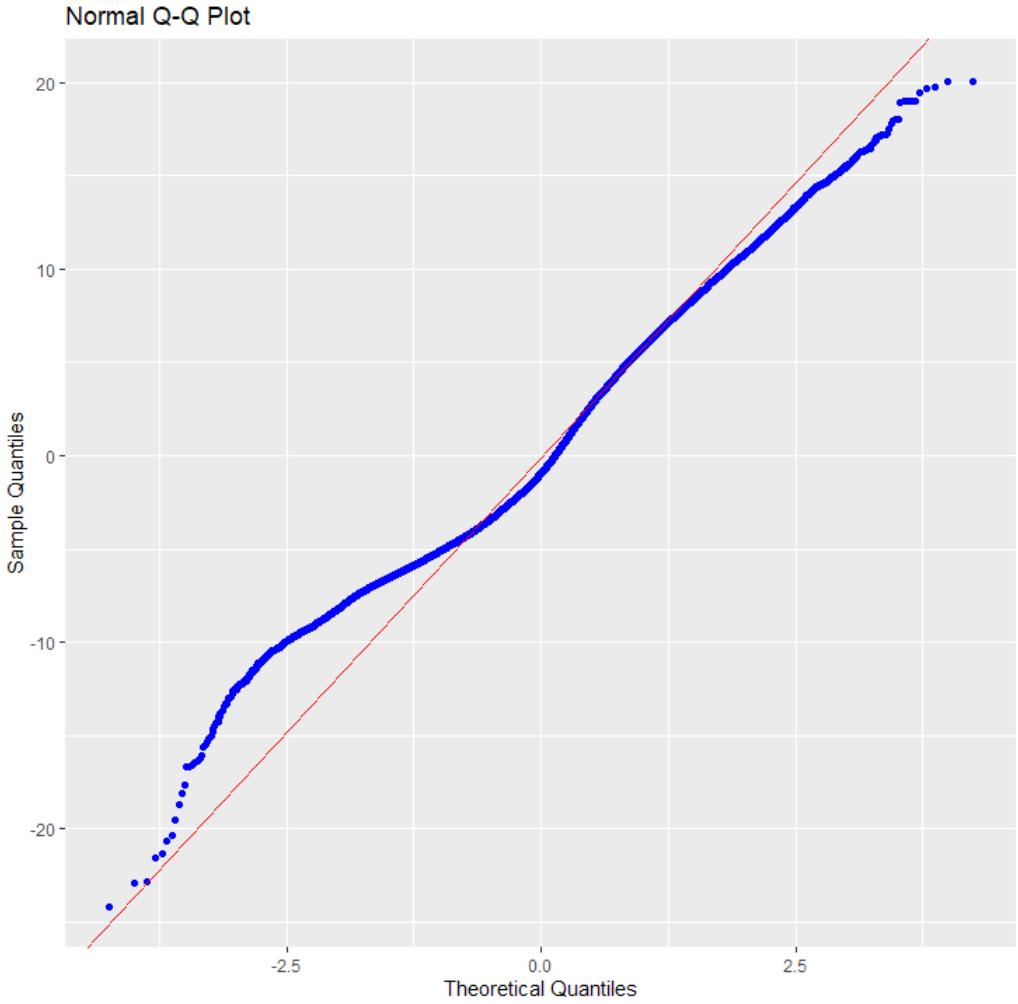

Figure 8: Normal Q-Q plot for the MLR model

We can see that the data points do not lie entirely on the 45° diagonal line. Thus we can suspect that the errors are actually not normally distributed.

2. **Kolmogorov-Smirnov Test for Normality**: We test the following hypothesis:

$$H_0 : \text{Errors are Normally Distributed} \quad ag. \quad H_A : H_0 \text{ is not true}$$

```
        One-sample Kolmogorov-Smirnov test

data:  uni_std_res
D = 0.075978, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 9: Summary of Kolmogorov-Smirnov Test

We get the value of the test statistic $D = 0.075978$ and the p-value as 2.2×10^{-16} which is smaller than 0.05 (testing at 95% level of significance). Thus we reject the null hypothesis in favour of the alternate hypothesis and conclude that the errors are not normally distributed.

4.3 Detection of Outliers

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or due to experimental error.
We will try to identify outliers using standardized and studentized residuals.
Standardized residuals are given by:

$$d_i = \frac{\hat{\epsilon}_i}{\sqrt{MSRes}}, \ \forall i = 1(1)n$$

where, $\hat{\epsilon}_i = Y_i - \hat{Y}_i, \ \forall i$ and $MSRes = (n-p-1)^{-1} \sum_i \hat{\epsilon}_i^2$. A large value of d_i, say, $|d_i| > 3$ indicates possible outlier.

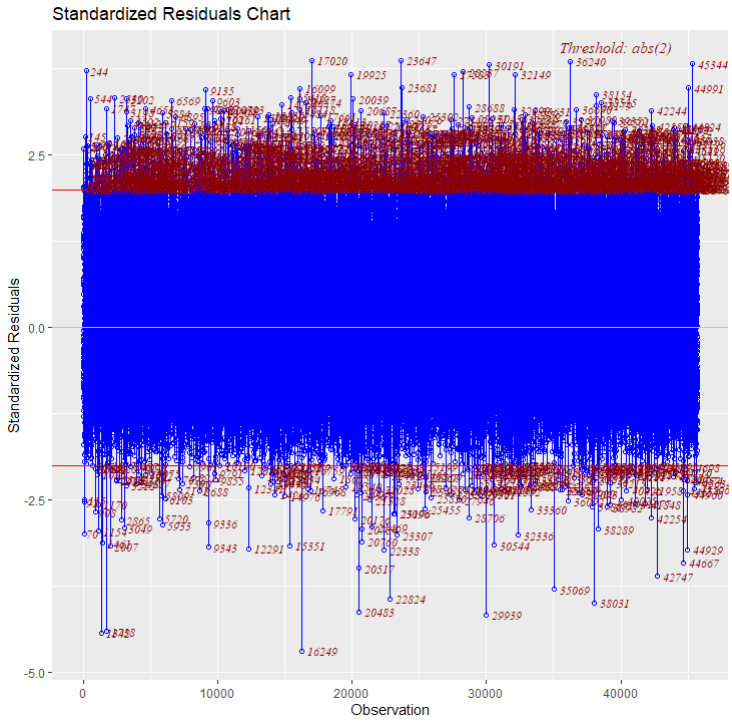

Figure 10: Plot for Standardized residual indicating points outside a limit

We can see that a substantial number of data points lie in regions which can indicate presence of outliers.

Studentized residuals are given by:

$$r_i = \frac{\hat{\epsilon}_i}{\sqrt{MSRes}\sqrt{1-h_{ii}}},\ \forall i = 1(1)n$$

where, $h_{ii} = (i,i)^{th}$ element of $X(X'X)^{-1}X'$. A large value of r_i, say, $|r_i| > 3$ indicates possible outlier.

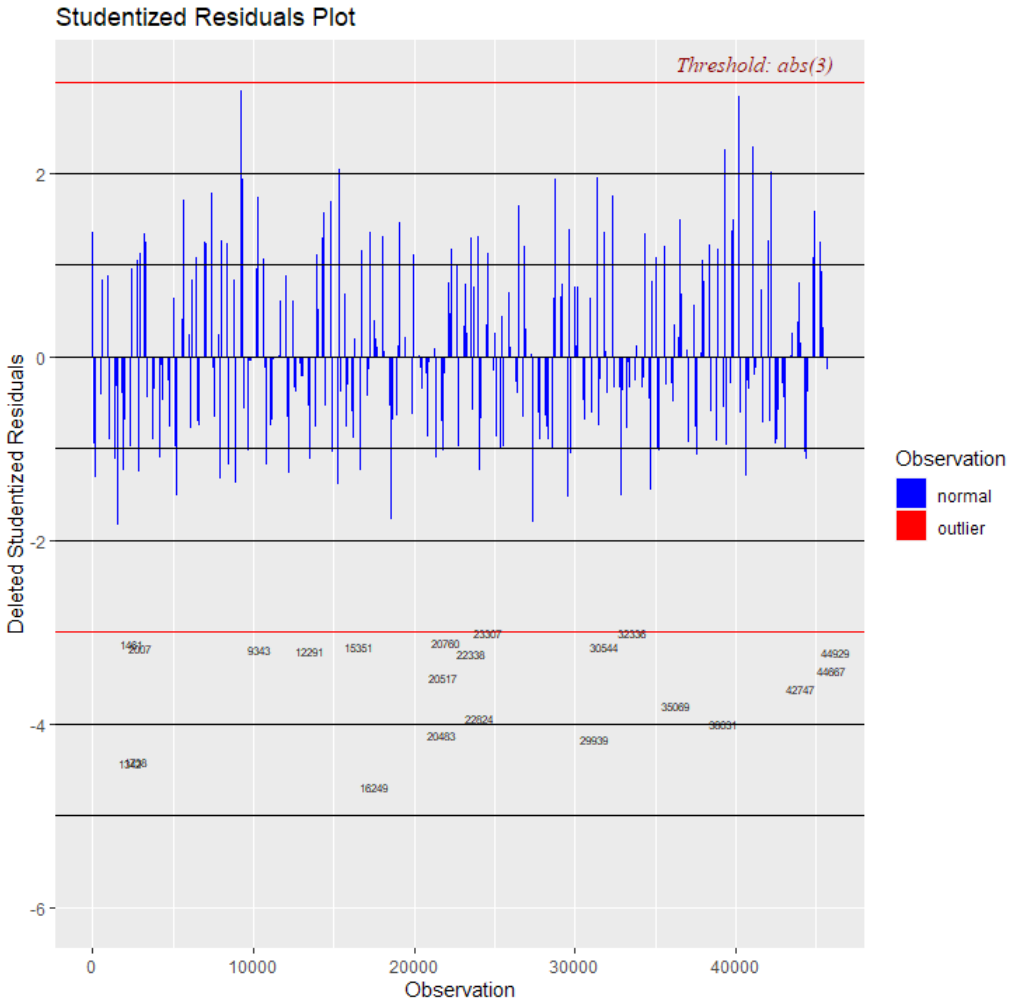

Figure 11: Plot for Studentized residual indicating points outside a limit

We can see that a few points can be believed to be possible outliers.

4.4 Detection of Leverage Points

A leverage point is determined on the basis of location of the point on the x-space and hence remote points impart more effect on the parameters of the model.
A point is considered to be a leverage point if:

$$h_{ii} > 2p/n$$

$h_{ii} = (i,i)^{th}$ element of $X(X'X)^{-1}X'$, p is the total number of variables in our model (both response and regressors).

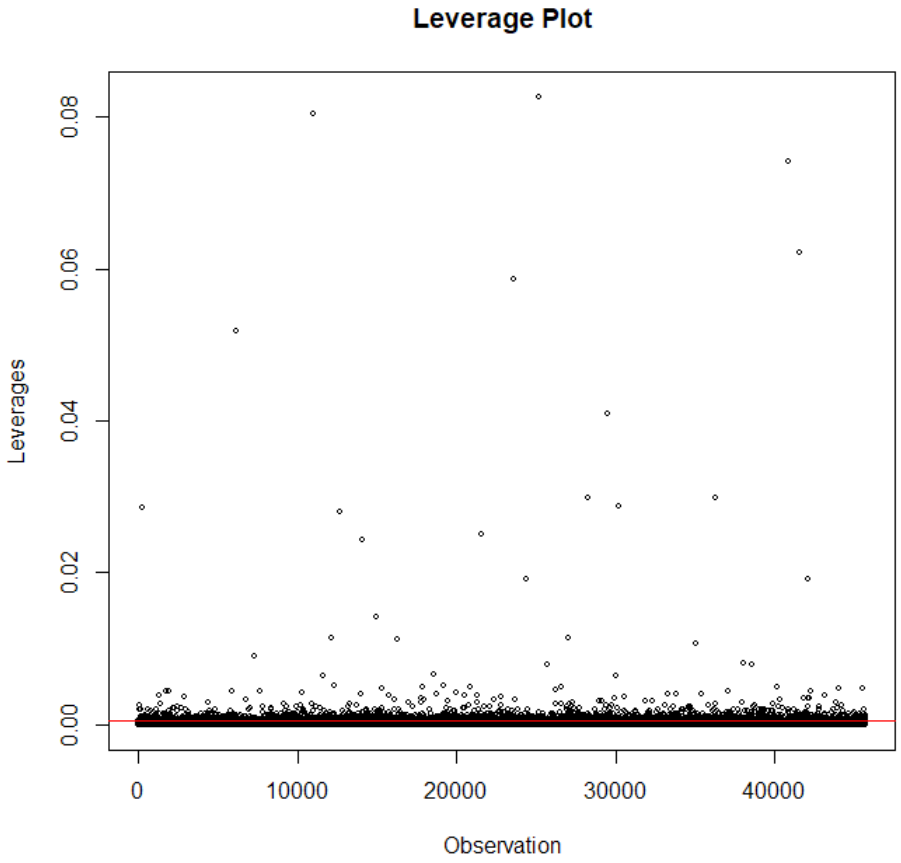

Figure 12: Plot for checking leverages

As we can see, quite a many points apparent have large leverages.

4.5 Detection of Influential points

We use Cook's Distance or Cook's D to estimate the influence of a data point.
Technically, Cook's D is calculated by removing the i^{th} data point from the model and recalculating the regression. It summarizes the extent to which all the values in the regression model change when the i^{th} observation is removed. The formula for Cook's distance is:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)s^2}$$

where, $\hat{Y}_{j(i)}$ is the fitted response value obtained when excluding the i^{th} regressor, and s^2 is the MSE.
Any point having $D_i > 4/n$ can be investigated to be an influential point and in this case we will be considering them as such.

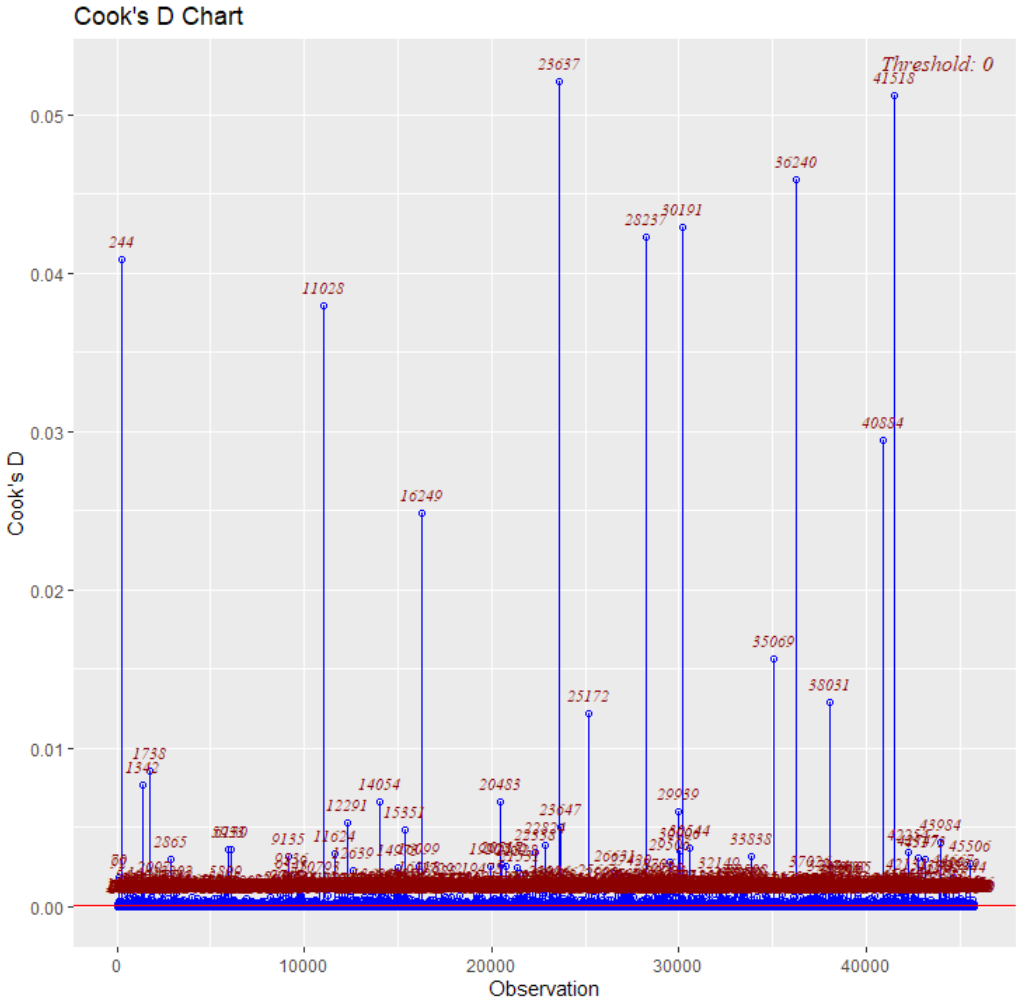

Figure 13: Plot for Cook's Distance

We see that there are substantial number of points which can be regarded as influential points.

Since we are interested in finding the true nature of the dependence of response on the regressors, we will drop these points.
We fit the MLR model again on this reduced set of data points. The residuals are noted.
We notice the histogram of the residuals:

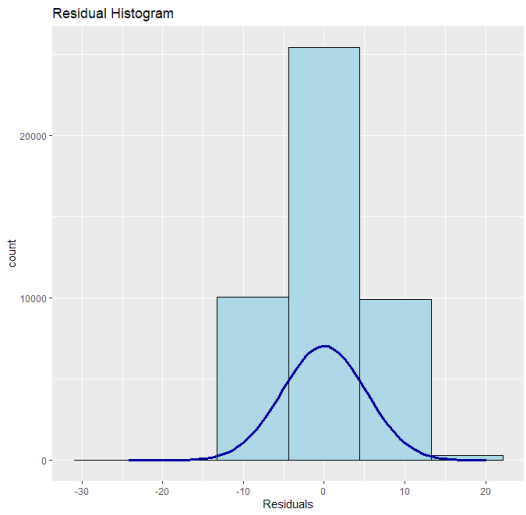

Figure 14: Residual Histogram for MLR model on Original Dataset

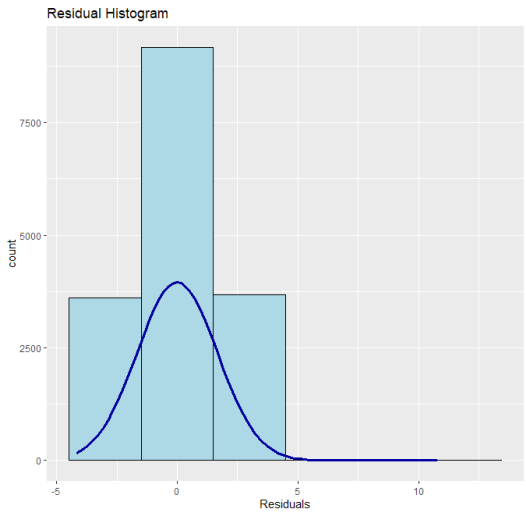

Figure 15: Residual Histogram for MLR model on Cleaned Dataset

We can see that data points with extreme influence has been removed, as is indicated by the two histograms.
We will make our next analyses on this cleaned data set.

5 Multicollinearity

Multicollinearity problem arises when the variance of the estimated parameters is too high. This can lead to a huge change in the estimated parameter on addition or deletion of a data point.
To detect multicollinearity we first reframe our model.
Out MLR model is given by:

$$Y_i = \beta_0 + \sum_{i=1}^{n} \beta_j x_{ij} + \epsilon_i, \ \forall i = 1(1)n \tag{1}$$

Where, x_{ij} is the value of j^{th} standardized regressor F_j, i.e.,

$$x_{ij} = \frac{F_{ij} - \bar{F}_j}{sd(F_j)}, \ \forall j = 1(1)p$$

The design matrix is $X = ((x_{ij}))$ of order $n \times p$, n is the total number of observations in the deleted model.

5.1 Prima Facie Detection

1. We can simply check the correlation matrix (related to the standardized regressors) to detect whether there is some potential in our model to have multicollinearity.

	F1	F2	F3	F4	F5	F6	F7	F8	F9
F1	1.000	0.889	0.053	0.941	0.998	0.974	0.781	0.661	-0.921
F2	0.889	1.000	0.472	0.770	0.885	0.893	0.701	0.572	-0.777
F3	0.053	0.472	1.000	-0.067	0.051	0.116	0.031	0.047	0.024
F4	0.941	0.770	-0.067	1.000	0.936	0.944	0.707	0.682	-0.917
F5	0.998	0.885	0.051	0.936	1.000	0.969	0.781	0.652	-0.920
F6	0.974	0.893	0.116	0.944	0.969	1.000	0.757	0.660	-0.908
F7	0.781	0.701	0.031	0.707	0.781	0.757	1.000	0.506	-0.757
F8	0.661	0.572	0.047	0.682	0.652	0.660	0.506	1.000	-0.640
F9	-0.921	-0.777	0.024	-0.917	-0.920	-0.908	-0.757	-0.640	1.000

Figure 16: Correlation Matrix
Yellow marked entries in the abive matrix indicate high values of pairwise correlation coefficient

We observe clealy that there are several entries in the matrix indicating very high value of pairwise correlation coefficient and therefore we may have enough reason to believe that our model contains multicollinearity.

2. We also check whether the matrix $(X'X)^{-1}$ is unstable or ill-conditioned or not.
We compute the eigen values of $X'X$. The are given below:

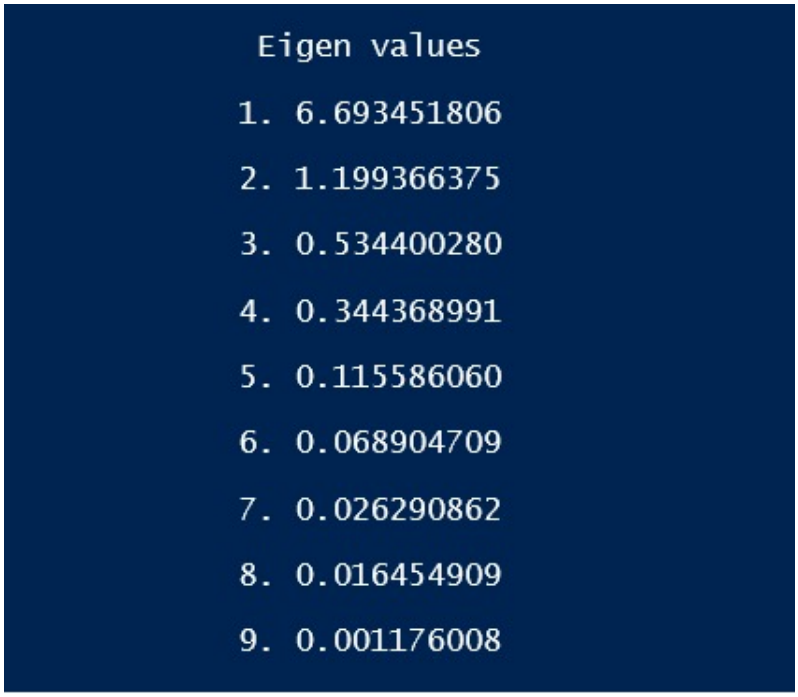

Figure 17: Eigen values of matrix $X'X$

It is noted that the last two eigen values (0.01645 and 0.001176) are very close to 0 and they might be responsible for the instability in the matrix $(X'X)^{-1}$.

3. We also note the determinant of the matrix $X'X$ given by:

$$|X'X| = \prod_{i=1}^{9} \{i^{th} \ eigen \ value \ of \ X'X\} = 5.986286 \times 10^{-9}$$

The determinant value is very close to 0 and hence is a near singular matrix.

4. The condition number of the matrix $X'X$ is given by:

$$\text{Condition Number } k = \sqrt{\frac{max \ eigenvalue}{min \ eigenvalue}} = \sqrt{\frac{6.6934}{0.001176}} = 75.4432$$

We can clearly see that the condition number is also very high, hinting towards potential near-linear relationship among some subsets of regressors.

Conclusion from Prime Facie Multicollinearity Detection- We have shown that, the matrix $X'X$ is near singular and ill-conditioned with a large condition number and also the pairwise correlation matrix has some large entries.
Thus, we can observe that there are some reasons to believe prima facie that our model may be affected by multicollinearity. So, we would perform further analysis on it to detect the regressors which are mainly responsible for multicollinearity.

5.2 Methodological Detection and Diagostics

5.2.1 Variance Inflation factor (VIF) and Variance Decompostion Method

VIF is calculated for parameters corresponding to each regresor. It provides us with an index that measures the inflation in the variance of an estimated regression coefficient due to collinearity. VIF is calculated as:

$$VIF_j = \frac{1}{1 - R_j^2}, \ \forall j = 1, 2, .., p$$

where R_j^2=Coefficient of Determination when x_j is regressed with other regresors.
We calculate the VIF of all the 9 regressors and then we tend to drop the regressor having highest VIF among those having VIF>5.

```
 VIF Multicollinearity Diagnostics

          VIF detection
F1.s 506.0250         1
F2.s  40.6471         1
F3.s   7.7823         0
F4.s  17.6277         1
F5.s 362.4306         1
F6.s  30.5046         1
F7.s   2.7855         0
F8.s   1.9744         0
F9.s   8.2883         0

Multicollinearity may be due to F1.s F2.s F4.s F5.s F6.s regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

==================================
```

Figure 18: Yellow marked VIF's are less than 5

VIF method indicates that there may exist multicollinearity in the model due to all regressors except F7 and F8.
As VIF does not provide us the information that which subset of regressors are basically responsible for multicollinearity we now go for **Variance Decomposition Method**.
We used Singular Value Decomposition of the input matrix X to calculate variance decomposition of the estimated parameters along various singular values. The variance decomposition matrix π is calculated as:

$$\begin{aligned} X &= UDV^T \\ ((\pi))_{kj} &= \frac{\frac{1}{\lambda_k} v_{kj}^2}{\sum_{k=1}^{p} \frac{1}{\lambda_k} v_{kj}^2}, \ \forall k, j = 1(1)p \end{aligned}$$

The Variance Decomposition of the original model (with all 9 regressors) is given as follows:

```
Call:
eigprop(mod = model1)

   Eigenvalues      CI (Intercept)   F1.s   F2.s   F3.s   F4.s   F5.s   F6.s   F7.s   F8.s   F9.s
1       6.6935  1.0000           0 0.0000 0.0004 0.0000 0.0011 0.0001 0.0007 0.0054 0.0059 0.0024
2       1.1994  2.3624           0 0.0000 0.0027 0.0871 0.0013 0.0000 0.0000 0.0008 0.0015 0.0017
3       1.0000  2.5872           1 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
4       0.5344  3.5391           0 0.0000 0.0004 0.0011 0.0001 0.0001 0.0002 0.0849 0.7875 0.0013
5       0.3444  4.4087           0 0.0001 0.0006 0.0004 0.0122 0.0002 0.0035 0.7780 0.1171 0.0037
6       0.1156  7.6098           0 0.0011 0.0159 0.0582 0.0068 0.0016 0.0056 0.0001 0.0093 0.7365
7       0.0689  9.8560           0 0.0013 0.0317 0.0644 0.4435 0.0027 0.0357 0.1251 0.0369 0.2152
8       0.0263 15.9559           0 0.0029 0.0008 0.0378 0.3689 0.0077 0.8766 0.0055 0.0077 0.0237
9       0.0165 20.1687           0 0.0108 0.8350 0.6553 0.1350 0.0353 0.0326 0.0000 0.0014 0.0127
10      0.0012 75.4432           0 0.9838 0.1123 0.0957 0.0310 0.9525 0.0450 0.0002 0.0327 0.0028

==============================
Row 10==> F1.s, proportion 0.983757 >= 0.50
Row 9==> F2.s, proportion 0.835000 >= 0.50
Row 9==> F3.s, proportion 0.655336 >= 0.50
Row 10==> F5.s, proportion 0.952483 >= 0.50
Row 8==> F6.s, proportion 0.876577 >= 0.50
Row 5==> F7.s, proportion 0.777998 >= 0.50
Row 4==> F8.s, proportion 0.787522 >= 0.50
Row 6==> F9.s, proportion 0.736504 >= 0.50
```

Figure 19: Yellow marked entries indicate π_{kj}'s which are high

Observations and Steps

1. We observe from the above variance decomposition table that for rownumber 10, condition index >15(i.e. for the row of condition number 75.4432>15) and corresponding that row number 10 there are two regressors F1 and F5 which are forming a subset as they have only larger (>0.5) variance decomposition proportions among all other regressors in that row. Thus row 10 indicates that regressors F1 and F5 (corresponding to the standardized regressors X1 and X5) may be involved in multicollinearity.
   Similarly, we check that for F9, condition index=20.1687>15. And we observe that regressors F2 and F3 might be involved in multicollinearity.

2. Now we got to subsets of regressors, (F1,F5) and (F1,F2). We observe,

   - $VIF_{F1} = 506.02 > VIF_{F5} = 362.4306$, Hence we drop F5 from our model.
   - $VIF_{F2} = 40.64 > VIF_{F3} = 7.7823$, Hence we drop F2 from our model.

From now on we would simultaneously look at VIF method and Variance Decomposition method for tuning our model in order to eradicate multicollinearity from our model.

```
         VIF detection
F3.s  1.4912         0
F4.s 16.0369         1
F5.s 21.2501         1
F6.s 26.4760         1
F7.s  2.7588         0
F8.s  1.9033         0
F9.s  8.2414         0

Multicollinearity may be due to F4.s F5.s F6.s regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Figure 20: VIF for model with 7 regressors (F3,F4,F5,F6,F7,F8,F9)
Yellow marked entries indicate VIF<5

We observe VIF>5 for the regressors F4, F5, F6 and F9. Therefore, by VIF method we may say this model still has multicollinearity due to these regressors. For the detection of subset we again check for Variance decomposition method.

```
  Eigenvalues      CI (Intercept)   F3.s   F4.s   F5.s   F6.s   F7.s   F8.s   F9.s
1      4.9548  1.0000           0 0.0000 0.0023 0.0018 0.0014 0.0101 0.0120 0.0045
2      1.0214  2.2025           0 0.6407 0.0007 0.0000 0.0002 0.0000 0.0007 0.0005
3      1.0000  2.2259           1 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
4      0.5215  3.0823           0 0.0008 0.0000 0.0014 0.0006 0.1329 0.7785 0.0027
5      0.3304  3.8725           0 0.0037 0.0204 0.0049 0.0073 0.7122 0.1747 0.0120
6      0.0984  7.0973           0 0.0339 0.0238 0.0345 0.0480 0.0118 0.0005 0.9337
7      0.0495 10.0010           0 0.1298 0.7095 0.3655 0.0119 0.1329 0.0299 0.0133
8      0.0240 14.3706           0 0.1910 0.2432 0.5918 0.9305 0.0001 0.0036 0.0333

==============================
Row 2==> F3.s, proportion 0.640727 >= 0.50
Row 7==> F4.s, proportion 0.709480 >= 0.50
Row 8==> F5.s, proportion 0.591810 >= 0.50
Row 8==> F6.s, proportion 0.930539 >= 0.50
Row 5==> F7.s, proportion 0.712177 >= 0.50
Row 4==> F8.s, proportion 0.778549 >= 0.50
Row 6==> F9.s, proportion 0.933746 >= 0.50
```

Figure 21: Variance Decomposition of model with 7 regressors (F3,F4,F5,F6,F7,F8,F9)

Observation and Steps

1. We observe from the above variance decomposition table for 7 regressors that for row number 8, condition index= condition number=14.3706 <15 but corresponding to that row number 8 there are two regressors F5 and F6 which are forming a subset as they have only larger (>0.5) variance decomposition proportions among all other regressors in that row. Thus row 8 indicates that regressors F5 and F6 (corresponding to the standardized regressors X1 and X5) may be involved in multicollinearity. We consider them as potential candidates for multicollinearity as VIF values for both of them >15.
   It is to be noted that there are no such other rows with subset formation among regressors in the above case of variance decomposition.

2. Now we got a group of regressors, namely(F5,F6)which are suspected to be involved in multicollinearity. We observe:

- $VIF_{F6} = 26.47 > VIF_{F5} = 21.25$, Hence drop F6 from our model.

Now, after dropping F6, our model becomes,

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \epsilon$$

Now, at this stage, we could have stopped our procedure for multicollinearity diagnostics as the variance decomposition method shows us that condition number of (X'X) (maximum condition index) is less than 15.

```
  Eigenvalues     CI (Intercept)   F3.s   F4.s   F5.s   F7.s   F8.s   F9.s
1      4.0290 1.0000           0 0.0000 0.0048 0.0044 0.0158 0.0189 0.0068
2      1.0131 1.9942           0 0.8485 0.0006 0.0001 0.0004 0.0024 0.0002
3      1.0000 2.0072           1 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
4      0.5171 2.7913           0 0.0029 0.0000 0.0028 0.1691 0.7362 0.0037
5      0.3065 3.6255           0 0.0218 0.0434 0.0161 0.6736 0.2106 0.0334
6      0.0855 6.8646           0 0.0020 0.1589 0.1496 0.0084 0.0081 0.9547
7      0.0488 9.0867           0 0.1247 0.7922 0.8269 0.1327 0.0238 0.0013

==============================
Row 2==> F3.s, proportion 0.848544 >= 0.50
Row 7==> F4.s, proportion 0.792197 >= 0.50
Row 7==> F5.s, proportion 0.826924 >= 0.50
Row 5==> F7.s, proportion 0.673621 >= 0.50
Row 4==> F8.s, proportion 0.736236 >= 0.50
Row 6==> F9.s, proportion 0.954654 >= 0.50
```

Figure 22: Variance Decomposition for model with regressors (F3,F4,F5,F7,F8,F9)

But we observe carefully that variance decomposition proportions corresponding to the two regressors F4 and F5 are very high so they are forming a group and are suspected to be involved in multicollinearity. Not only that, VIF values are also very high corresponding to these two regressors under this model.

```
 VIF Multicollinearity Diagnostics

         VIF detection
F3.s  1.1440         0
F4.s 11.6954         1
F5.s 12.8991         1
F7.s  2.7396         0
F8.s  1.9026         0
F9.s  8.2405         0

Multicollinearity may be due to F4.s F5.s regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Figure 23: VIF for model with regressors (F3,F4,F5,F7,F8,F9)

Also, the pairwise correlation coefficient between standardized regressors X4 and X5 (corresponding to F4 and F5 respectively) is very high **0.936**.
We observe that F5 can be well explained by F4 and hence our decision of dropping F5 is quite justifiable.

```
Call:
lm(formula = F5.s ~ F4.s)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03174 -0.21472 -0.06699  0.14097  2.76565

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.500e-16  2.743e-03     0.0        1
F4.s        9.361e-01  2.743e-03   341.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3518 on 16453 degrees of freedom
Multiple R-squared:  0.8762,    Adjusted R-squared:  0.8762
F-statistic: 1.165e+05 on 1 and 16453 DF,  p-value: < 2.2e-16

> cor(F5.s,F4.s)
[1] 0.936077
```

Figure 24: Summary table for simple linear regression $F5 \sim F4$

Hence, we won't ignore this issue; although the condition number=9.0867<15 (by theoretical thumb rule we should have ignored the group formation and could have claimed that this underlying model is the final model free from multicollinearity).
With the similar proceedings and the same philosophy, we drop F9 in the next step. We are then finally left with the model with 4 regressors F3, F4, F7, F8.

5.2.2 Reduced Model

Therefore, our reduced model is:

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \beta_7 X_7 + \beta_8 X_8 + \epsilon$$

We calculate the VIF.

```
 VIF Multicollinearity Diagnostics

       VIF detection
F3.s 1.0329         0
F4.s 2.8685         0
F7.s 2.0269         0
F8.s 1.9007         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

Figure 25: VIF for each of 4 regressors

We see that VIF for each of 4 regressors is very low and in fact all of them are < 5 under the above model.
We obtain the variance decomposition of the model:

```
  Eigenvalues     CI (Intercept)   F3.s   F4.s   F7.s   F8.s
1      2.2672 1.0000           0 0.0000 0.0577 0.0691 0.0714
2      1.0102 1.4981           0 0.9457 0.0028 0.0005 0.0022
3      1.0000 1.5057           1 0.0000 0.0000 0.0000 0.0000
4      0.4940 2.1423           0 0.0017 0.0007 0.4657 0.5663
5      0.2286 3.1490           0 0.0526 0.9388 0.4647 0.3601

================================
Row 2==> F3.s, proportion 0.945660 >= 0.50
Row 5==> F4.s, proportion 0.938784 >= 0.50
Row 4==> F8.s, proportion 0.566294 >= 0.50
```

Figure 26: Variance Decomposition for the model

So, we can observe that under this reduced model there does not exist any subset of regressors causing multicollinearity in the model.

For the reduced model minimum eigen value of $X'X$ becomes 0.2286 which is quite larger than 0 where as for the original model the minimum eigen value of $X'X$ was 0.001176 which might be responsible for the instability of $X'X$ under the original model.

We also see that,

- Condition no. of $X'X$ under original model=75.4432 (ill-conditioned)
- Condition no. of $X'X$ under reduced moel=3.1490

Correlation Matrix under the reduced model

	F3	F4	F7	F8
F3	1.00000000	-0.06737759	0.03140161	0.0469082
F4	-0.06737759	1.00000000	0.70697939	0.6816355
F7	0.03140161	0.70697939	1.00000000	0.5062295
F8	0.04690820	0.68163546	0.50622954	1.0000000

Figure 27: Correlation Matrix for reduced model

We see that any pair of regressors are not highly correlated.

6 Ridge Regression: A remedy of Multicollinearity

We will consider the MLR model on standardized regressors that we have already stated as equation (1) (Under Multicollinearity).
Multicollinearity leads to coefficients with large magnitudes. Ridge regression shrinks the coefficients by imposing a penalty on their size. The technique yields biased estimates of the model parameters but with lower variances.
We obtain the ridge parameter estimates by minimizing the $(Y-X\beta)'(Y-X\beta)+\lambda\beta'\beta$ constrained on $\beta'\beta \leq d^2$ (say), where λ is the shrinkage/complexity parameter that controls the amount of shrinkage.
The Ridge estimate is given by:

$$\hat{\beta}_R = (X'X + \lambda I)^{-1} X'y$$

Thus λ is a positive constant added to the diagonal elements of $X'X$, so that $(X'X + \lambda I)$ is not near singular and $(X'X + \lambda I)^{-1}$ is no longer unstable.

6.1 Ridge Trace

We have considered the Ridge traceplot of different elements (or regressors) of $\hat{\beta}_R$ against λ for different values of $\lambda = 0(0.001)0.5$ and observe that after $\lambda = 0.1335939$ all elements of the ridge parameter vector $\hat{\beta}_R$ are becoming stable.

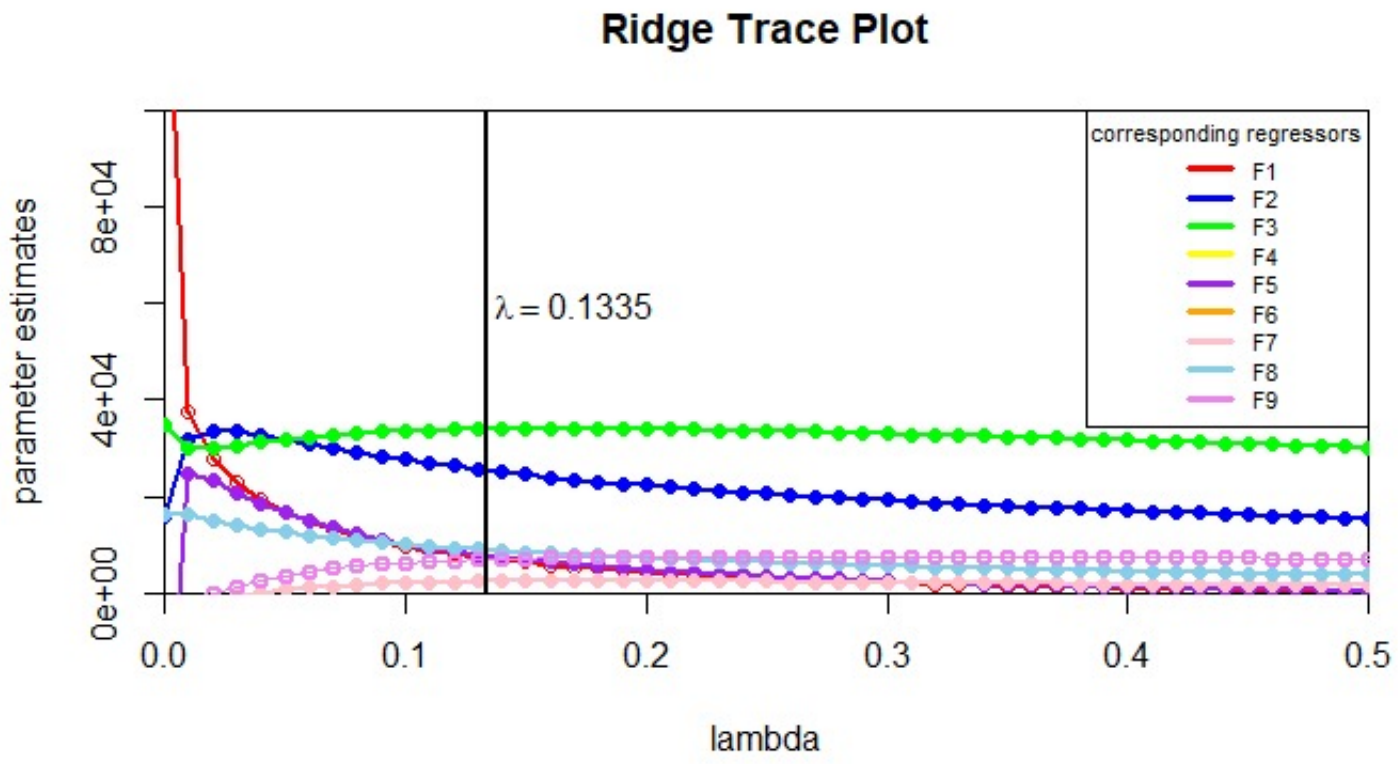

Figure 28: Ridge Trace plot

Then by taking $\lambda = 0.13359$ in our Ridge estimator for the model coefficients, we get the

estimates as follows:

Parameter	Estimate
$\hat{\beta}_0$	6.465791
$\hat{\beta}_1$	8.529023
$\hat{\beta}_2$	0.9720707
$\hat{\beta}_3$	2.126493
$\hat{\beta}_4$	-7.399937
$\hat{\beta}_5$	-2.687172
$\hat{\beta}_6$	-1.421968
$\hat{\beta}_7$	-0.3081103
$\hat{\beta}_8$	0.9998739
$\hat{\beta}_9$	-0.3554566

Thus our estimated model is of the form:

$$\widehat{RMSD} = \hat{\beta}_0 + \sum_{j=1}^{9} \hat{\beta}_j \left(\frac{F_j - \bar{F}_j}{s.d.(F_j)} \right)$$

Thus our final model is:

$$\begin{aligned} \widehat{RMSD} = -0.7103491 &+ 0.001975163F_1 + 0.0006459997F_2 + 31.00019F_3 \\ &- 0.1173224F_4 - 4.479266 \times 10^{-6}F_5 - 0.0189902F_6 \\ &- 0.0002242405F_7 + 0.01718049F_8 - 0.05516927F_9 \end{aligned}$$

6.2 Comparison of Parameter Estimates

Although $\hat{\beta}_R = (X'X + \lambda I)^{-1}X'y$ is a biased estimate of parameter vector β under the ridge regression, we may compare the variances of each of the individual parameter estimates roughly with that of the usual parameter estimates(usual parameter estimate $\hat{\beta}$ is an unbiased estimate) obtained by least squares technique.
Theoretically, we have the result,

$$Var(\hat{\beta}_{R_j}) \leq Var(\hat{\beta}_j)$$

On the basis of our model and data, here we get the same result; in fact in our case, variance of every j^{th} parameter estimate under ridge regression is strictly lesser than that of usual parameter estimate.Following table shows the fact.

	Var of usual parameter estimate	var of Ridge estimate
1	1394.402436	1.113740
2	112.007127	2.284601
3	21.444977	2.151903
4	48.575000	3.543990
5	998.713694	1.532398
6	84.058611	2.697886
7	7.675660	3.674262
8	5.440684	3.145826
9	22.839148	4.351826

Figure 29: Variances of parameter estiates obtained by OLS and Ridge regression

But we actually should have compared the MSE's of the parameter estimates.
We observe that:

$$24.56 = Total\ MSE(\hat{\beta}_R) < Total\ MSE(\hat{\beta}) = 2695.17$$

Hence, we may say that our fitted ridge regression is worthwhile in pulling up those very small eigen values of original matrix (X'X) efficiently. *The above results also signify that our ridge regression model is quite able to provide us better parameter estimates in terms of Mean Squared Errors (MSE) without dropping a single regressor.*

7 Variable Selection

Here our aim is to choose a subset of 'best' regressors (in some sense) from a pool of 9 possible regressors. There are several ways of choosing this subset of 'best' regressors. In our analysis we shall use 2 model selection criteria viz Akaike Information Criterion and Mallow's Cp Statistic.

Theoretic criteria: Penalized likelihood function

Our MLR model is

$$Y = X\beta + \epsilon$$

where we assume $\epsilon \sim N_n(0, \sigma^2 I_n)$ such that $Y \sim N_n(X\beta, \sigma^2 I_n)$
The likelihood function given by,

$$L(\beta, \sigma^2|y) = (2n)^{-n/2}(\sigma^2)^{-n/2}exp\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\}$$

So, the general form of the penalized likelihood function,

$$-2ln\ L^* + penalty\ term$$

. Where,

$$L^* = \max_{\beta,\sigma^2} L(\beta, \sigma^2|y) = L(\hat{\beta_{mle}}, \hat{\sigma^2_{mle}})$$

By further calculation we get the **penalized likelihood function** as,

$-2ln\ L^*$+ penalty term $= nln(SS_{Res})$+ penalty term

7.1 Akaike Information Criterion

AIC is a single number score that can be used to determine which of multiple models is most likely to be the best model for a given dataset. It estimates models relatively, meaning that AIC scores are only useful in comparison with other AIC scores for the same dataset. A lower AIC score is better.
Here the penalty term is $2p$, p is the number of model parameters.

We calculate the AIC(p) using the following formula:

$$AIC(p) = -2ln\ L^* + 2p = nln\ SS_{Res(p)} + 2p$$

So, we compute AIC for all possible models and choose the model for which AIC is minimum.

7.2 Mallow's Cp Statistic

Let the true Model has K regressors and intercept term and the Subset Model has p-1 regressors and intercept. So that $r = k - (p - 1)$ regressors are detected.
Define, Γ_p=Scaled sum of MSE's of $\hat{Y_i}$ for some subset model of order p, as

$$\Gamma_p = p + \frac{SSB_{(p)}}{\sigma^2}$$

Where, $SSB_{(p)} = \beta'_{(r)} X'_{(r)} \left(I_p - P_{X_{(r)}}\right) X_{(r)} \beta_{(r)}$
Now, consider the following statistic,

$$\widehat{\Gamma}_p = C_p = \frac{SS_{Res(p)}}{\widehat{\sigma^2}} + 2p - n$$

This is called the Mallow's Cp statistic. Furthermore, $E(SS_{Res(p)}) = (n-p)\sigma^2 + SSB_{(p)}$ (from estimation from σ^2). Thus,

$\Gamma_p = \frac{E(SS_{Res(p)})}{\sigma^2} + 2p - n = p + \frac{SSB_{(p)}}{\sigma^2} = p$ iff the sum of squrare Bias, $SSB_{(p)} = 0$.

Now, with $E\left(\widehat{\Gamma}_p\right) = E\left(C_p\right) \approx \Gamma_p$, we consider the model for which, $C_p \approx p$ for the first time, i.e. we choose the smallest value of p for which $C_p \approx p$.

Best Subsets Regression

Model Index	Predictors
1	F3
2	F2 F4
3	F1 F3 F4
4	F1 F3 F4 F8
5	F1 F3 F4 F6 F8
6	F1 F3 F4 F5 F6 F8
7	F1 F2 F3 F4 F6 F7 F8
8	F1 F2 F3 F4 F5 F6 F7 F8
9	F1 F2 F3 F4 F5 F6 F7 F8 F9

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP
1	0.5302	0.5302	0.5301	40008.4677	83676.6820	36975.5766	83699.8072	155662.7763
2	0.7903	0.7902	0.7902	8759.9839	70410.6443	23711.0524	70441.4778	69507.1900
3	0.8291	0.8291	0.829	4089.8696	67039.7206	20340.7844	67078.2625	56628.4971
4	0.8565	0.8564	0.8564	804.0214	64170.8764	17473.1447	64217.1267	47565.4548
5	0.8589	0.8588	0.8588	516.7655	63894.5959	17196.9698	63948.5546	46770.6536
6	0.8602	0.8602	0.8601	359.7947	63741.6290	17044.0729	63803.2961	46335.0702
7	0.8615	0.8615	0.8614	201.0886	63585.4730	16888.0277	63654.8485	45894.6480
8	0.8624	0.8623	0.8622	99.2285	63484.4463	16787.0907	63561.5301	45610.9670
9	0.8632	0.8631	0.863	10.0000	63395.4146	16698.1596	63480.2068	45362.0950

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality

Figure 30: AIC and Mallow's Cp for different subsets of regressors

Conclusion

Based on the values seen, the value for Mallow's Cp is indeed equal to 10 (the number of parameters) for Model 9. Further on, we notice that model 9, involving all the regressors from the full model, is the one associated with the lowest value of AIC.
Hence by Variable Selection Technique, we end up with our full model itself.

8 Model Adequacy

Here we will consider the model derived by eliminating Multicolliearity.
Thus in this case, our model is:

$$Y_i = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \beta_7 X_7 + \beta_8 X_8 + \epsilon$$

The least square estimates of the model parameters are:

Parameter	β_0	β_3	β_4	β_7	β_8
Least Square Estimate	-4.970	4.338×10^{1}	-4.915×10^{-2}	7.439×10^{-4}	1.862×10^{-2}

The summary of the model is presented below:

```
Call:
lm(formula = tt$RMSD ~ tt$F3 + tt$F4 + tt$F7 + tt$F8)

Residuals:
     Min       1Q   Median       3Q      Max
-22.6140  -1.6548  -0.3382   1.4880  14.0301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.970e+00  9.809e-02  -50.67   <2e-16 ***
tt$F3        4.338e+01  2.747e-01  157.94   <2e-16 ***
tt$F4       -4.915e-02  4.978e-04  -98.73   <2e-16 ***
tt$F7        7.439e-04  1.921e-05   38.72   <2e-16 ***
tt$F8        1.862e-02  4.392e-04   42.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.378 on 16450 degrees of freedom
Multiple R-squared:  0.7192,    Adjusted R-squared:  0.7192
F-statistic: 1.053e+04 on 4 and 16450 DF,  p-value: < 2.2e-16
```

Figure 31: Summary Chart for the model

R^2 and Adjusted R^2 are used to explain the overall adequacy of the model, where,

$$\begin{aligned} R^2 &= 1 - \frac{SS_{Res}}{SS_{Tot}} \\ R^2_{Adj} &= 1 - \frac{SS_{res}/(n-p-1)}{SS_{Tot}/(n-1)} \end{aligned}$$

We get the values of R^2 and Adjusted R^2 to be 0.7192 and 0.7192 respectively, implying that about 71.92% of the variation in observed responses is explained by our assumed linear model. Thus we can conclude that this model is quite efficient in explaining the dependence of the response (RMSD) on Fractional area of exposed non-polar residue (F3), Fractional area of exposed non-polar part of the residue (F4), Euclidean Distance (F7) and Secondary structure penalty (F8).

R^2 for Prediction

The Prediction Error Sum of Squares (or PRESS statistic) is defined as the sum of squared PRESS residuals. We will use it as a measure of model quality.
The press statistic is given by:

$$PRESS = \sum_{i=1}^{n} \left(\frac{\hat{\epsilon}_i}{1 - h_{ii}} \right)^2$$

PRESS is generally regarded as a measure of how well a regression model will perform in predicting new data. A model with small value of PRESS is desired.

R^2 for prediction based on PRESS is given by:

$$R^2_{PRESS} = 1 - \frac{PRESS}{SS_{Tot}}$$

We get the value of PRESS as 93273.93. Correspondingly, we get, $R^2_{PRESS} = 0.7184843$. Therefore we would expect the model to explain about 71.85% of the variability in predicting new observations.

Graphical Overview of Model

We plot the observed and estimated response values against a random sample of 100 observations.

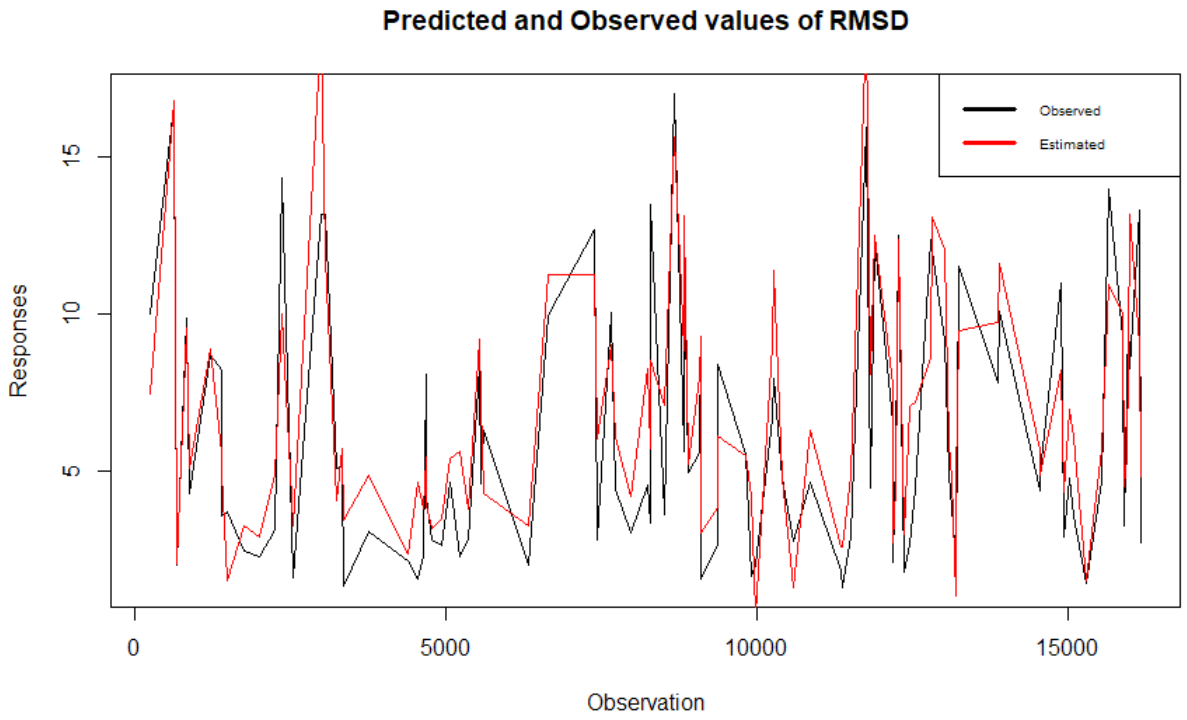

Figure 32: Plot for Observed and Estimated RMSD

We can see from this plot that our model is quite efficient in estimating the RMSD.

We take a look at the Q-Q plot for residuals.

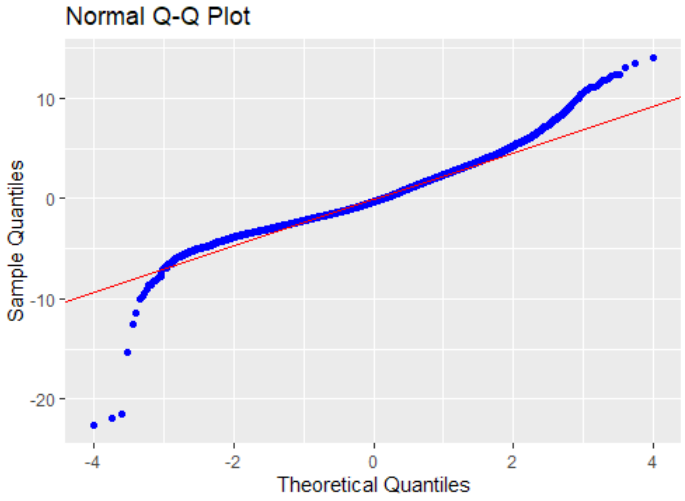

Figure 33: Q-Q plot for residuals

We can see that there is an indication that the errors may have heavier tails implying that the errors may not be normally distributed.
Further,we take a look at the scatter plot of residuals vs. fitted response values.

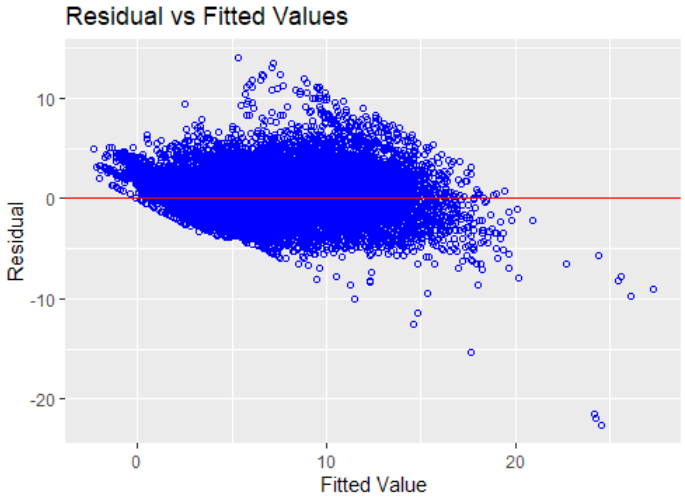

Figure 34: Residual Plot for MLR Model

We can see that the nature of the plot is quite random. We do get a hint of association, but the nature of dependence is not clear.

9 Conclusion

We see that the dependence of protein's tertiary structure is more efficiently explained by a subset of the 9 available regressors (i.e. F3, F4, F7 and F8) (71.92% efficient in explaining the observed response variance) than by all the full model involving 9 regressors (28.23% efficient in explaining the observed response variance) as we have seen that some of the regressors have a near-linear relationship amongst them.

We have also tried to keep the full model by allowing a trade-off between the biasness of the parameter estimates and their mean square errors. We have seen that although in doing so, we get biased estimates of the true model parameters, we however get estimates with lesser mean square errors that normal estimates.

We have made attempts to select a subset of regressors which explain the response most effiecient by looking at the AIC and Mallow's Cp of different subset models. We however have found that in doing so, we get back our full model itself.

We consider the final model as the one including regressors which do not have near-linear relationship amongst themselves.

10 Bibliography

Below are a list of references we used for the completion of this project.

- Lecture Notes of Regression Analysis (MTH416A), instructed by Dr. Sharmishtha Mitra
- Introduction to Linear Regression Analysis- Montgomery, Peck, Vining.
- Wikipedia
- Practical Regression and Anova using R- Julian J. Faraway (July, 2002)

A First Appendix

Data Source

The data has been obtained from UCI Machine Learning Repository. The link is attached:
https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure

B Second Appendix

R Codes used:

```
# Loading data
t <- read.csv("D:/CASP.csv")

# Library's used
library(plyr)
library(dplyr)
attach(t)
library(MASS)
library(olsrr)

# Pairs Plot for variables
pairs(t,density=TRUE)

# MLR model on whole subset
p=10
n=length(RMSD)
model1 <- lm(RMSD~F1+F2+F3+F4+F5+F6+F7+F8+F9)
summary(model1)

# Detection of Outliers, Leverages, Influential points
ols_plot_cooksd_chart(model1)   # Cooks distance plot
ols_plot_resid_stand(model1)    # Standardized residual plot
ols_plot_resid_stud(model1)     # Studentised residual plot

# Leverage Plot
h_ii=lm.influence(model1)$hat
plot(seq(1,n,1),h_ii,pch=,xlab="Observation",ylab="Leverages",
main="Leverage Plot")

# After deletion, plot of residuals for model 1 and model 2
ols_plot_resid_qq(model1)
ols_plot_resid_qq(model2)

library(mctest)
imcdiag(model,method='VIF')
eigprop(model)

# Q-Q Plots
ols_plot_resid_qq(model1)

# Two residual histrohrams for original and cleaned data model
```

```
ols_plot_resid_hist(model1)
ols_plot_resid_hist(model2)

# Multicollinearity Checking

n=length(RMSD)
f=(n-1)/n

F1.s=(F1-mean(F1))/sqrt(var(F1)*f)
F2.s=(F2-mean(F2))/sqrt(var(F2)*f)
F3.s=(F3-mean(F3))/sqrt(var(F3)*f)
F4.s=(F4-mean(F4))/sqrt(var(F4)*f)
F5.s=(F5-mean(F5))/sqrt(var(F5)*f)
F6.s=(F6-mean(F6))/sqrt(var(F6)*f)
F7.s=(F7-mean(F7))/sqrt(var(F7)*f)
F8.s=(F8-mean(F8))/sqrt(var(F8)*f)
F9.s=(F9-mean(F9))/sqrt(var(F9)*f)

model1=(lm(RMSD~F1.s+F2.s+F3.s+F4.s+F5.s+F6.s+F7.s+F8.s+F9.s))

imcdiag(model1,method="VIF")

eigprop(model1)

#(F1,F5) and (F2,F3) are the subsets
# F1 and F2 is removed

model2=(lm(RMSD~F3.s+F4.s+F5.s+F6.s+F7.s+F8.s+F9.s))
imcdiag(model2,method="VIF")
eigprop(model2)


#(F5,F6) subset forms and F6 is removed


model3=(lm(RMSD~F3.s+F4.s+F5.s+F7.s+F8.s+F9.s))
imcdiag(model3,method="VIF")
eigprop(model3)

#(F4,F5) subset and F5 is removed though cond no is<5

model4=(lm(RMSD~F3.s+F4.s+F7.s+F8.s+F9.s))
imcdiag(model4,method="VIF")
eigprop(model4)

#(F4,F9) and F9 is removed

model5=(lm(RMSD~F3.s+F4.s+F7.s+F8.s))
imcdiag(model5,method="VIF")
eigprop(model5)
```

```
#no subset forming also all VIFs <15

#eigen values checking~original model

X=matrix(c(F1.s,F2.s,F3.s,F4.s,F5.s,F6.s,F7.s,F8.s,F9.s),ncol=9)
X1=(t(X)%*%X)/n
round(X1,3)
eval1=eigen(X1)$values

#Reduced model

model5=(lm(RMSD~F3.s+F4.s+F7.s+F8.s))
X.red=matrix(c(F3.s,F4.s,F7.s,F8.s),ncol=4)
X1.red=((t(X.red)%*%X.red))/n

rownames(X1.red)=c("F3","F4","F7","F8")
colnames(X1.red)=c("F3","F4","F7","F8")
View(X1.red)
eval2=eigen(x1.red)$values
eval2
summary(model1)

# Ridge Regression
f=(n-1)/n
F1s <- (tt$F1-mean(tt$F1))/sqrt(var(tt$F1)*f)
F2s <- (tt$F2-mean(tt$F2))/sqrt(var(tt$F2)*f)
F3s <- (tt$F3-mean(tt$F3))/sqrt(var(tt$F3)*f)
F4s <- (tt$F4-mean(tt$F4))/sqrt(var(tt$F4)*f)
F5s <- (tt$F5-mean(tt$F5))/sqrt(var(tt$F5)*f)
F6s <- (tt$F6-mean(tt$F6))/sqrt(var(tt$F6)*f)
F7s <- (tt$F7-mean(tt$F7))/sqrt(var(tt$F7)*f)
F8s <- (tt$F8-mean(tt$F8))/sqrt(var(tt$F8)*f)
F9s <- (tt$F9-mean(tt$F9))/sqrt(var(tt$F9)*f)
ttt<- data.frame(c(tt$RMSD,F1s,F2s,F3s,F4s,F5s,F6s,F7s,F8s,F9s))
X <- matrix(c(F1s,F2s,F3s,F4s,F5s,F6s,F7s,F8s,F9s),ncol=9)
w <- t(X)%*%X
n1 <- length(tt$RMSD)
w=w/n1

library(genridge)
par(mfrow=c(1,1))
model4 <- ridge(tt$RMSD,X/sqrt(n1),lambda=seq(0,.5,0.001))
model4$kHKB
# 0.1335939
model_ridge<- ridge(tt$RMSD,X/sqrt(n1),lambda=0.1335939)
coef(model_ridge)
fl=function(x)
{
m1=solve(X1+(x*diag(9)))%%t(X)%%RMSD
m1
}
```

```
z=seq(0,3,0.01)
length(z)
Mat=matrix(0,nrow=9,ncol=length(z))
for(i in 1:length(z))
{
Mat[,i]=fl(z[i])
}


lam_fin=0.1335939

#trace plot

par(mfrow=c(1,1))
plot(z,Mat[1,],col='red',xlim=c(0,0.5), ylim=c(0,100000),xaxs="i",yaxs="i",
main="Ridge Trace Plot",col.main="Black"
,xlab='lambda',ylab='parameter estimates',text(0.17, 6e+04,expression(lambda==0.1335)))
lines(z,Mat[1,],col='red',lwd=2)
points(z,Mat[2,],col='blue',pch=19)
lines(z,Mat[2,],col='blue',lwd=2)

points(z,Mat[3,],col='green',pch=19)
lines(z,Mat[3,],col='green',lwd=2)

points(z,Mat[4,],col='yellow',pch=19)
lines(z,Mat[4,],col='yellow',lwd=8)

points(z,Mat[5,],col='purple',pch=19)
lines(z,Mat[5,],col='purple',lwd=2)

points(z,Mat[6,],col='orange',pch=19)
lines(z,Mat[6,],col='orange',lwd=6)

points(z,Mat[7,],col='pink',pch=19)
lines(z,Mat[7,],col='pink',lwd=2)

points(z,Mat[8,],col='skyblue',pch=19)
lines(z,Mat[8,],col='skyblue',lwd=2)

lines(z,Mat[9,],col='violet',pch=19)
points(z,Mat[9,],col='violet',lwd=2)
abline(v=lam_fin,col='black',lwd=2)
legend("topright",legend=c("F1","F2","F3","F4","F5","F6","F7","F8","F9"),
title="corresponding regressors",col=c("red","blue","green",
    "yellow","purple","orange","pink","skyblue","violet"),
        lty=1,cex=0.7,lwd=3)

#comparison Ridge regression


betar.h=solve((n*X1)+(lam_fin*diag(9)))%%t(X)%%RMSD
eval1
```

```
evect1=eigen(X1)$vectors
summary(model)
sig.hat=1.66^2
library(Matrix)
# install.packages("psych")
library(psych)

tr(solve(t(X)%*%X))
m1=matrix(0,nrow=9,ncol=9)
m2=matrix(0,nrow=9,ncol=9)
for(i in 1:9)
{
m1=m1+((evect1[,i]%*%t(evect1[,i]))/eval1[i])
m2=m2+(eval1[i](evect1[,i]%%t(evect1[,i]))/(eval1[i]+lam_fin)^2)
}
cov.betah=sig.hat*(m1)
cov.bridge=sig.hat*(m2)

# sum of variances under usual estimate
sig.hat*tr(m1)

#sum of variances under ridge estimate
sig.hat*tr(m2)

var.usual_e=diag(cov.betah)
var.ridge_e=diag(cov.bridge)
var.dfm=matrix(c(var.usual_e,var.ridge_e),nrow=9)
rownames(var.dfm)=1:9
colnames(var.dfm)=c("Var of usual parameter estimate", "var of Ridge estimate")
View(var.dfm)

cov.betah=sig.hat*(m1)
cov.bridge=sig.hat*(m2)

cov.bridge
cov.betah
tr(m2*sig.hat)+(lam_fin^2)t(betar.h)%%solve((n*X1)+
    (lam_fin*diag(9)))%*%betar.h
mse.ridge=tr(m2*sig.hat)+(lam_fin^2)t(betar.h)%%solve((n*X1)+
    (lam_fin*diag(9)))%*%betar.h
cov.betah
mse.usual=tr(cov.betah)
mse.usual
mse.ridge

# Work on Reduced Model
model_req <- lm(tt$RMSD~tt$F3+tt$F4+tt$F7+tt$F8)

plot(sort(rs),response[sort(rs)],type="l",main="Predicted
    and Observed values of RMSD",xlab="Observation",ylab="Responses")
lines(sort(rs),predict(model_req)[sort(rs)],col="red")
legend("topleft",legend=c("Observed","Estimated"),lwd=3,lty=c(1,1),col=c("black","red"))
```