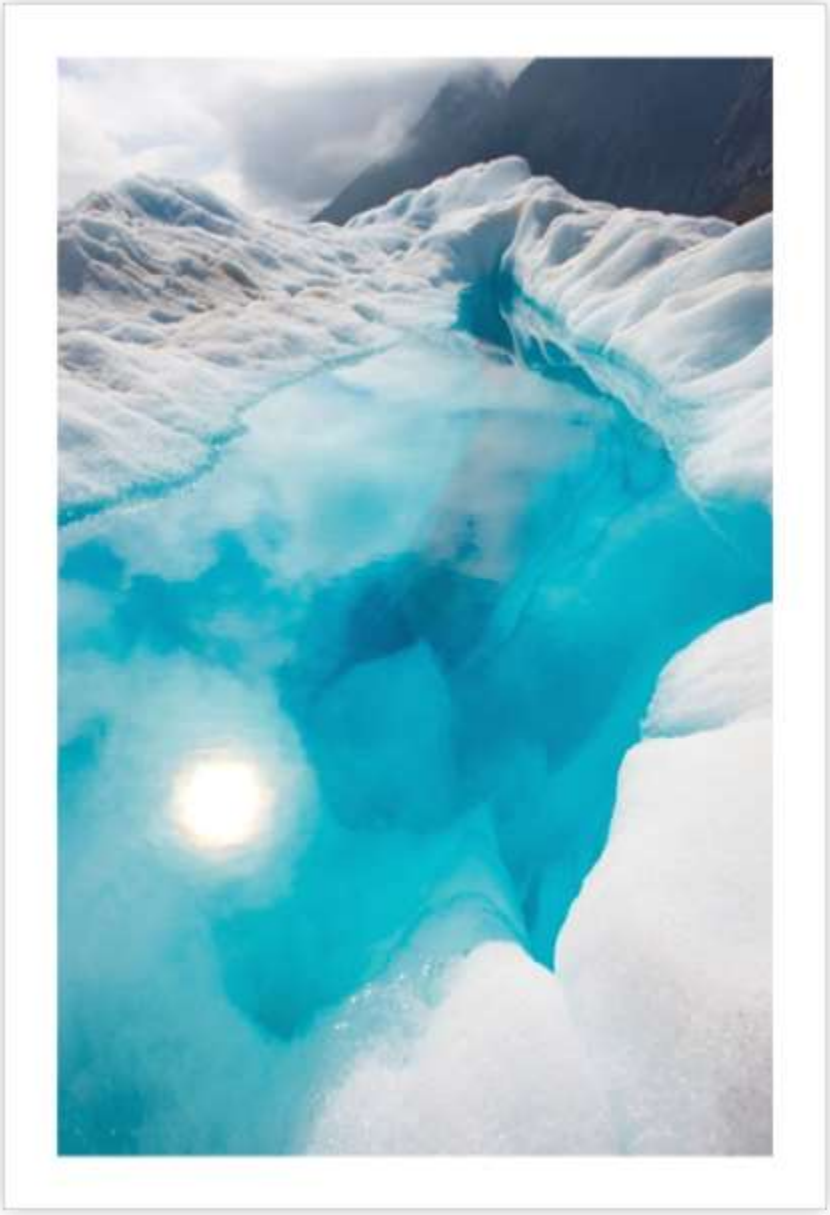


STUDY ON DEPENDENCE OF COMPRESSIVE STRENGTH ON VARIOUS INPUT FACTORS OF CONCRETE FORMATION

Abhradipta Ghosh | MD2101
Arijit Naskar | MD2102
Piyush Swarnkar | MD2115
Regression Techniques | December 27, 2021

1 Data Description:

Description of the Covariates

We have 7 input variables and 3 output variables in our data which have been described as follows:

Input Variable:

Cement: Cement is a binder, a substance used for construction that sets, hardens, and adheres to other materials to bind them together. It mixed with other matreials produces mortar, concrete, etc.

Slag Cement: Slag cement is most widely used in concrete, either as a separate cementitious component or as part of a blended cement.It works synergistically with portland cement to increase strength, reduce permeability,improve resistance to chemical attack and inhibit rebar corrosion.

Water

Fly ash: Fly ash use in concrete improves the workability of plastic concrete, and the strength and durability of hardened concrete.

SP: Superplasticizers, also known as high range water reducers, are additives used in making high strength concrete. Plasticizers are chemical compounds that enable the production of concrete with approximately 15% less water content. Superplasticizers allow reduction in water content by 30% or more.

Course aggr: Coarse aggregates are defined as any material greater than 4.75 mm. Aggregates make up 60-80% of the volume of concrete and 70-85% of the mass of concrete. Aggregate is also very important for strength, thermal and elastic properties of concrete, dimensional stability and volume stability.

Fine Aggr: Fine aggregates are usually sand or crushed stone that are less than 9.55mm in diameter. functionality same as course aggr.

Output Variables:

Slump: The 'slump' of concrete refers to the consistency of fresh concrete before it sets – the higher the slump, the more fluid the concrete is.

Flow: The percentage increase in the average diameter of the spreading concrete over the base diameter of the mould is called the flow of concrete. It also measures how fluid concrete is.

Compressive Strength: Test pieces are used to measure the compressive strength of concrete. The test pieces generally consist of cylinders 16 cm in diameter and 32 cm in height that are manufactured in cardboard moulds.The test pieces are subjected to a crushing force (generally 28 days after manufacture) and the force required to break the specimen is measured.The ratio between the crushing force and the cross-sectional area of the test piece gives the compressive strength of the concrete. it is measured to check whether the strength of the concrete has reached the requirement of specified strength.

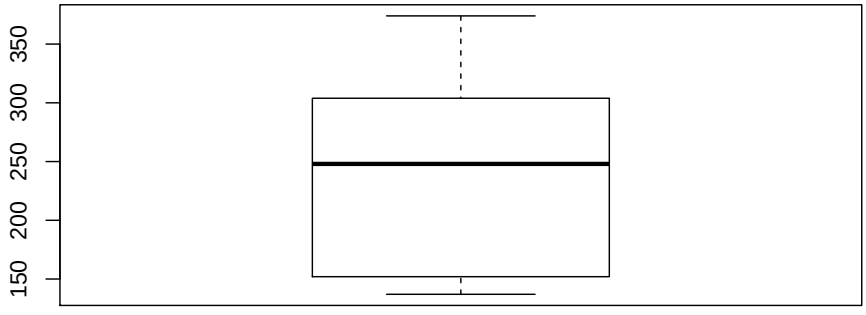
Slump test and flow test measures the same attribute , hence highly correlated.

Descriptive Statistics

Cement:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   137.0   152.0   248.0   229.9   303.9   374.0
```

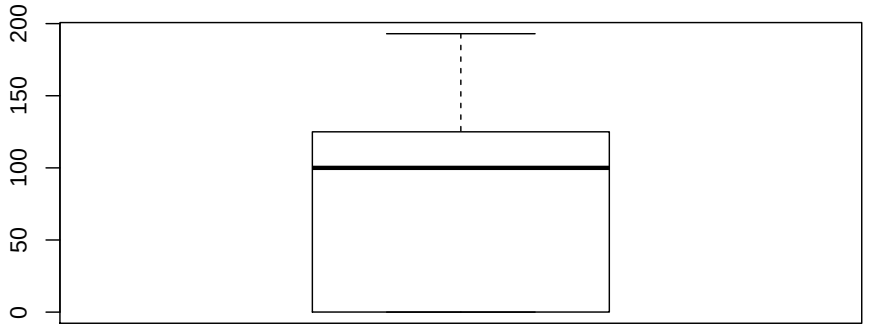
The box plot of cement is given below,



Slag Cement:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.05  100.00   77.97  125.00  193.00
```

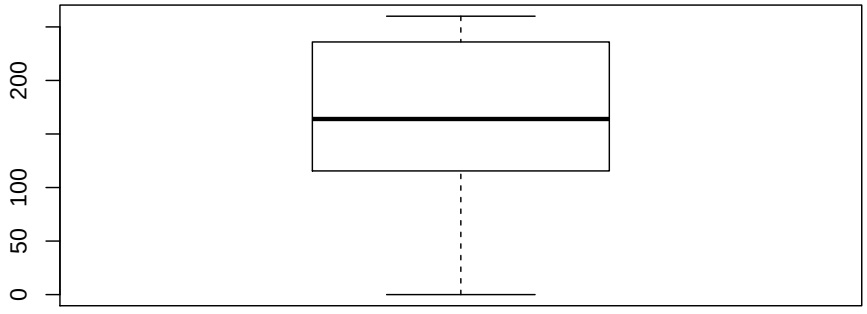
The box plot of slag cement is given below,



Water:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   115.5   164.0   149.0   235.9   260.0
```

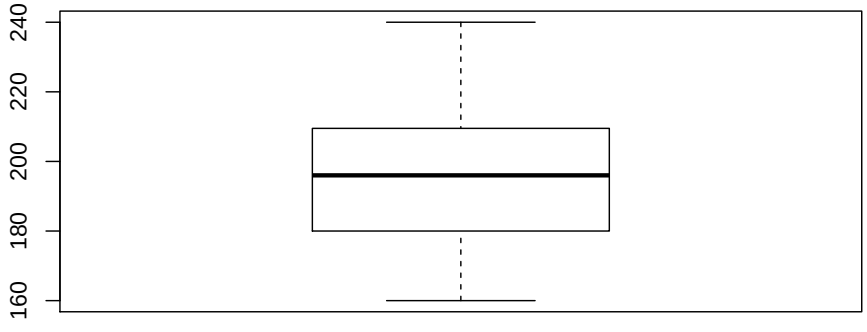
The box plot of water is given below,



Fly Ash:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   160.0   180.0   196.0   197.2   209.5   240.0
```

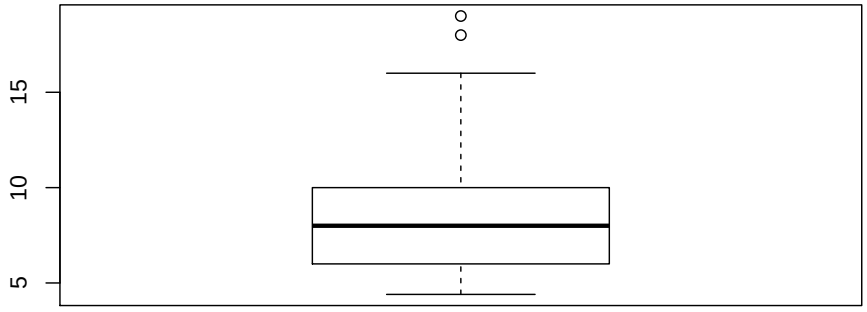
The box plot of fly ash is given below,



SP:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.40    6.00    8.00    8.54   10.00   19.00
```

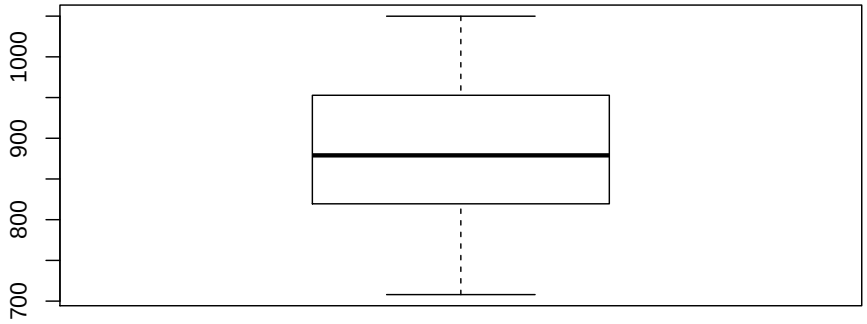
The box plot of sp is given below,



CA:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   708.0   819.5   879.0   884.0   952.8  1049.9
```

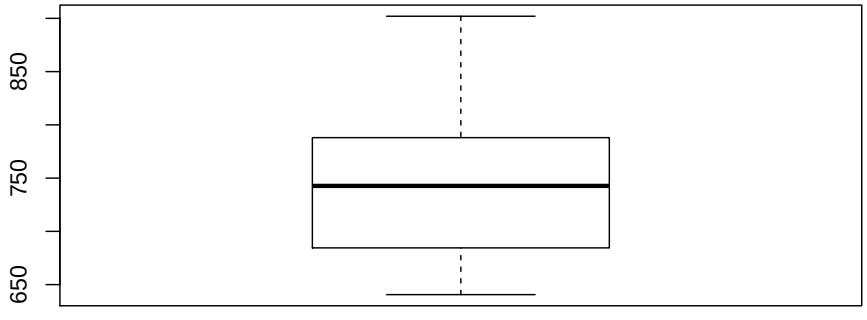
The box plot of ca is given below,



FA:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   640.6   684.5   742.7   739.6   788.0   902.0
```

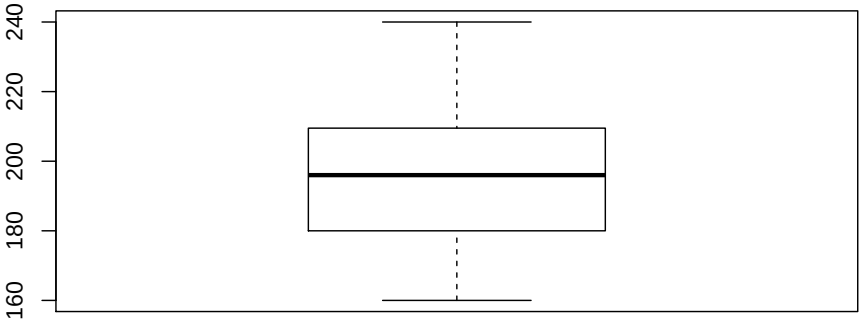
The box plot of FA is given below,



Slump:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   14.50   21.50   18.05   24.00   29.00
```

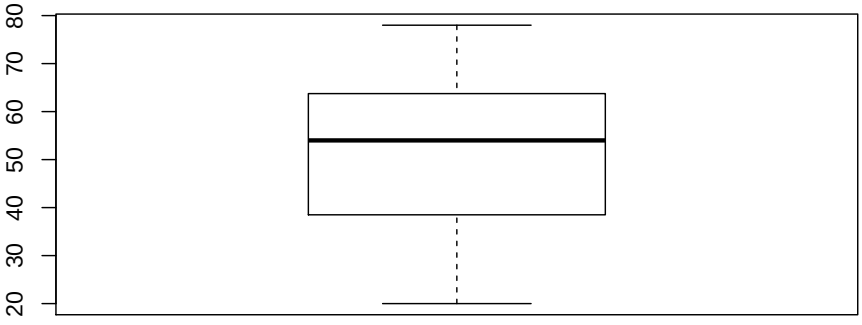
The box plot of slump is given below,



Flow:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   38.50   54.00   49.61   63.75   78.00
```

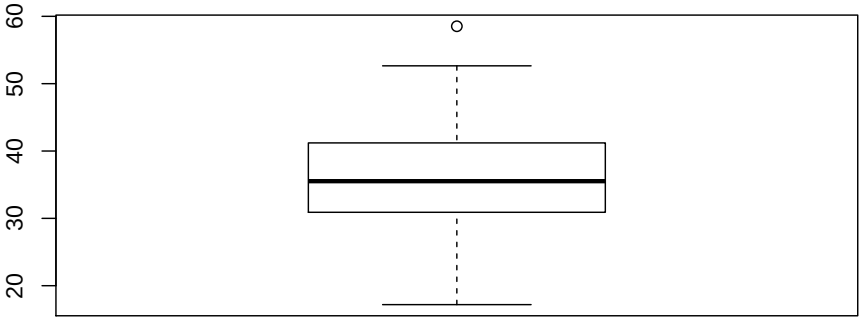
The box plot of flow is given below,



Comprehensive Strength:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.19   30.90   35.52   36.04   41.20   58.53
```

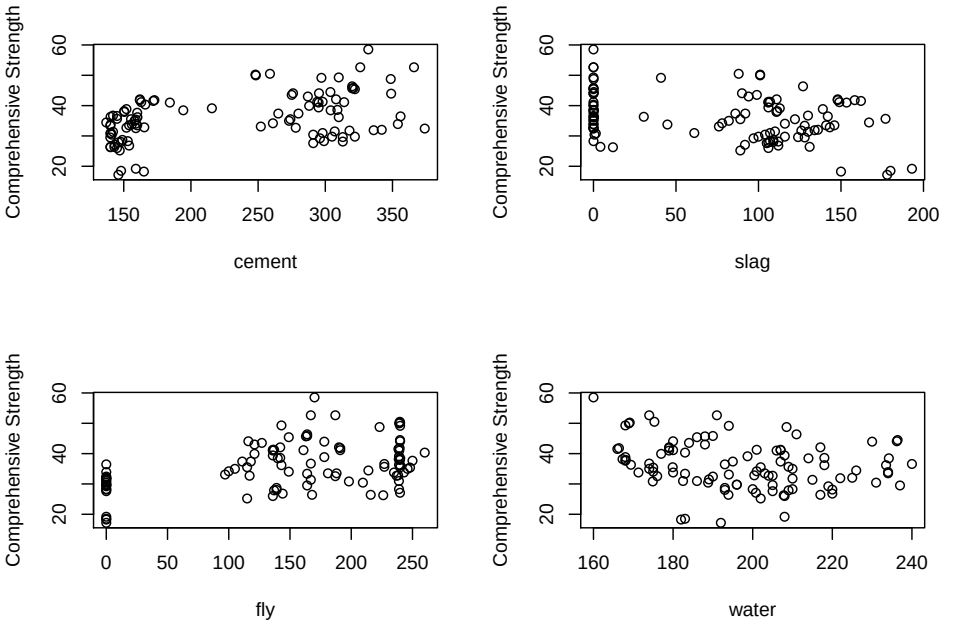
The box plot of comprehensive strength is given below,

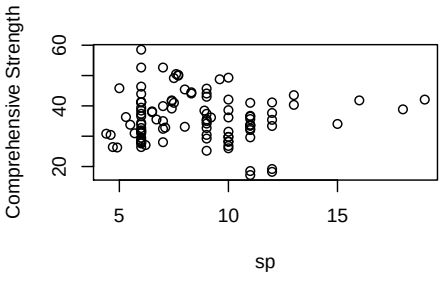


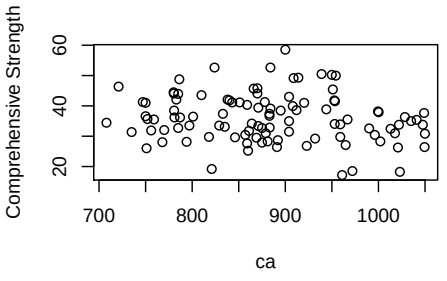
Relationship between response and predictor

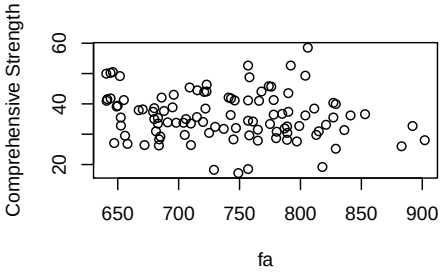
Scatter plot

The scatter plot of response and each predictor variables is given below





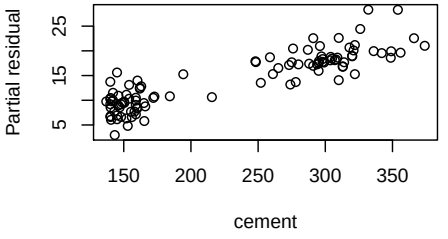


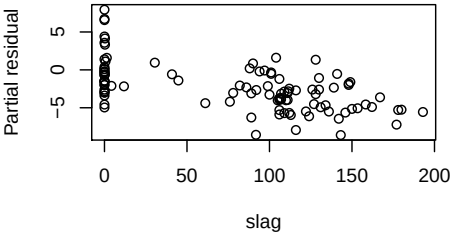


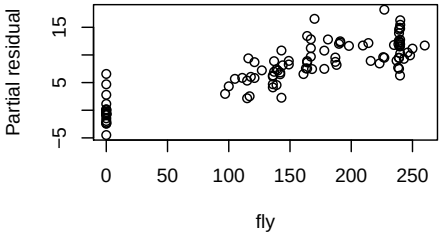
Comment: We can't observe linear relation between these response and the predictors using scatter plot.

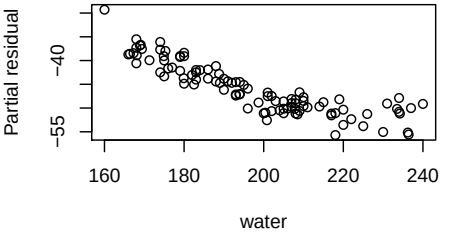
Partial residual plot

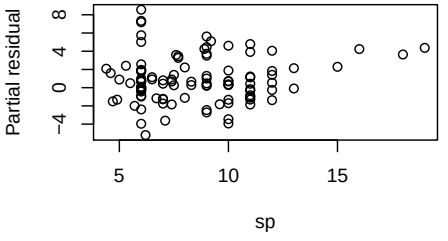
The partial residual plot of response and each predictor variables is given below

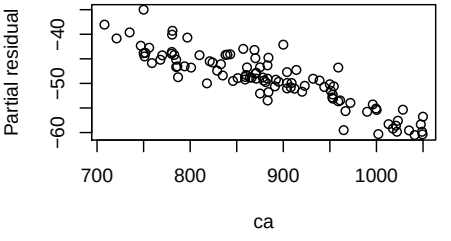

Partial residual
5
15
25
150
200
250
300
350
cement

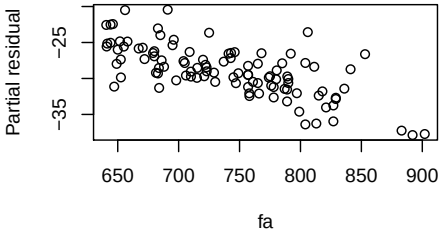

Partial residual
−5
0
5
0
50
100
150
200
slag


Partial residual
−5
5
15
0
50
100
150
200
250
fly


Partial residual
−55
−40
160
180
200
220
240
water







Comment: We can observe that the response variable is linearly related with each variable.

Model and Assumptions:

Following the given data, we are interested to find how Compressive Strength and Slump or Flow of concrete depends on several components of concrete formation. To deal with this problem we consider the following model : $Y = X\beta + \epsilon$, where Y denote the observations of response variable , X denote the design matrix with obervations on the i^{th}explanatory variable along the ithcolumn. β is the unknown parameter vector and ϵ denote the vector containing the corresponding error terms.

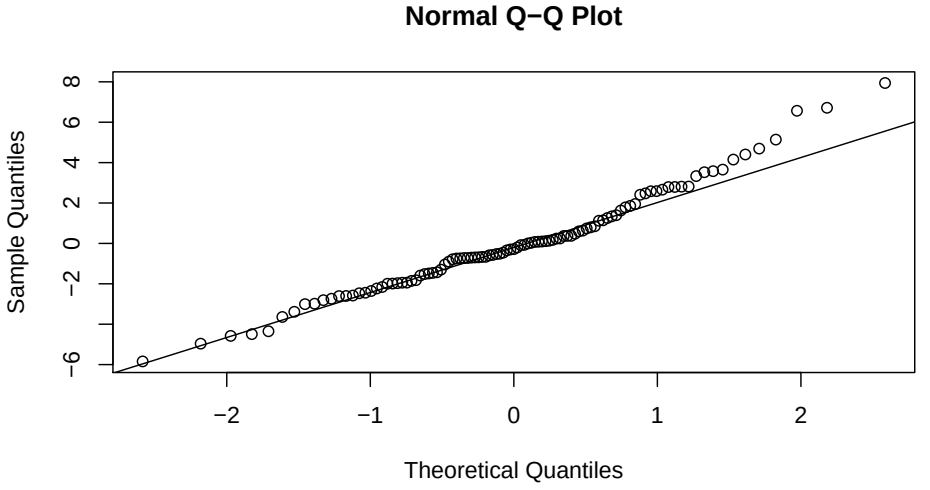
The model assumptions are as follows

- $\epsilon_i's$ are Normally Distributed.
- $\epsilon_i's$ are homoscedastic with variance σ^2.
- $\epsilon_i's$ are uncorrelated.
- X matrix is non-stochastic and of full rank.

Normality in the Data :

Residual Plot

According to our assumption,The error ϵ_i's are independently distributed NORMAL random variables. $e_i's$ can be considered as a reasonable estimate of the errors in the model when the other assumptions hold true . We inspect the Normal Probability Plot based on the residuals to judge over the validity of the assumption.



Comment : From the above Q-Q plot, we can observe that quantiles of the residual resemble very closely to the quantiles of Normal Distribution. Hence there is no visual evidence of non-Normality in the data.

Shapiro-Wilk Test

To test , H_o : $e_i's$ has come from a Normally distributed population vs. H_1 : not H_0

Test Statistic :

$$W = \frac{(\sum_{i=1}^{n} a_i e_{(i)})^2}{(\sum_{i=1}^{n} e_i)^2},$$

$e_{(i)}$ denote the ith residuals in the ordered sample.

The coefficients $a_i's$ are given by :

$$(a_1, a_2, ...a_n)' = \frac{m' V^{-1}}{||m' V^{-1}||}$$

where, the vector m is the expected values of ordered sample from Standard Normal Distribution . V is the covariance matrix of those Normal ordered statistics.

We compute p-value of the test using R , as given below –

```
Attaching package:  'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


Shapiro-Wilk normality test

data:  e
W = 0.97486, p-value = 0.0467
```

Comment : From the output of Shapiro-Wilk Test, we see that the p-value is 0.0467. The Null Hypothesis is accepted at level 0.01.Hence we can conclude that the Normality assumptions of the errors is valid.

Heteroscedasticity :

From our assumption ,the error $\epsilon_i's$ are homoscedastic, i.e, each of the error terms has equal variance denoted by σ^2.we need to check whether this assumption is plausible in the considered model based on the data. When all the assumptions hold true, residuals can be treated as estimates of the error terms . Thus the residual plot may reflect the true state of the error terms.

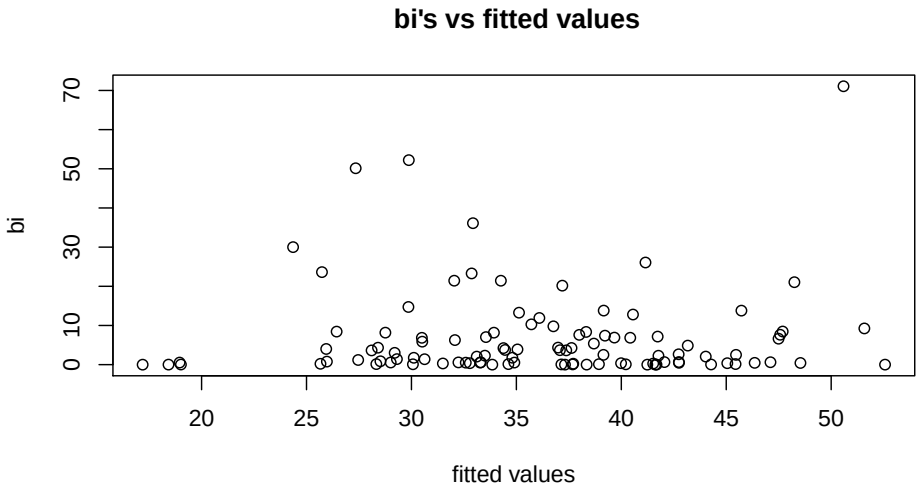
Several Plots :

- Plot of $b_i's$ vs. fitted values

Let us consider the quantity $b_i = \frac{e_i^2}{1-h_i}$, i=1(1)n

when the variances are all equal to σ^2,E$(b_i) = \sigma^2$, i =1(1)n . Thus, plotting the $b_i's$ against the fitted values should result in a wedge-shaped display, when the variances increase with the mean.

By plotting $b_i's$ against the fitted values we get the plot as following :



Comment : From the plot, we do not see any significant wedge shape, hence the variances puportedly do not increase with the increase in the fitted values.

Breusch-Pagan Test

The Breusch–Pagan test is used to test for heteroskedasticity in a linear regression models. It tests the null hypothesis that the variance of error terms is independent of the explanatory variables against dependence. It involves calculating the residuals by Ordinary Least Squares and then regressing the square of these residuals, divided by the MLE of error variance, against the explanatory variables of the linear regression model to obtain the auxiliary regression equation.

Assume that the error variances are of the linear form of independent variables as–

$$\sigma_i^2 = \delta_0 + \delta_1 x_{i1} + ... + \delta_{k-1} x_{i,k-1}$$

Then the Lagrange Multiplier test statistic for testing the null is given by:

$$LM = 1/2(TSS - RSS) \sim \chi^2_{p-1}$$

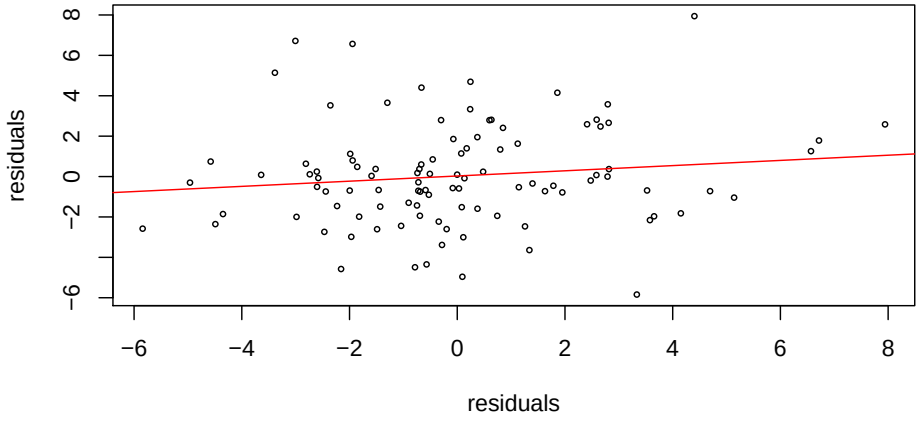
Where TSS is the sum of squared deviations of the response and RSS is the sum of squared residuals, both obtained from the auxiliary equation. A robust variant of this test involve using only the squared residuals as response in auxiliary equation and then the statistic nR^2 from this equation has asymptotically same distribution as LM.

The p value of BP test is 0.6584. So we can conclude that the residuals are homoscedastic.

Autocorrelation :

Residual plot

According to our assumption the error $\epsilon_i's$ are uncorrelated among themselves. To check the validity of this assumption we assume that the other assumptions hold true. Since the residuals are expected to reflect the nature of the errors, we observe the plot of residuals vs residuals with lag 1 to check for the possible presence of autocorrelation , in the data.



Comment: From the above plot, no significant correlation is visible among the residuals (considering lag1). Also the autocorrelation coefficient obtained from these residuals is 0.13 which is not very significant.

Durbin Watson Test

To draw more credible conclusion regarding autocorrelation we conduct Durbin Watson Test on the data.

Suppose that the error ϵ_i follow a first order autoregressive process , $\epsilon_i = \rho\epsilon_{i-1} + \delta_i$, where $\delta_i's$ are independently and identically distributed as N(0,σ^2).

Since autocorrelation from the residuals is positive we test for $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. The test statistic is given by –

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

We conduct the test from the residuals, using R . We get the output as–

```
## Loading required package:  carData

## Warning:  package 'carData' was built under R version 3.6.3

##
## Attaching package:  'car'

## The following object is masked from 'package:dplyr':
##
##    recode


##  lag Autocorrelation D-W Statistic p-value
##    1       0.1284928      1.730584     0.1
##  Alternative hypothesis: rho != 0
```

Comment : in the test, the p-value is 0.108. Hence the null hypothesis gets accepted at level α= 0.01. We get no significant evidence in favour of the autocorrelation from the data.

PRESENCE OF MULTICOLLINEARITY

In the formulation of regression technique one crucial assumption is that the design matrix X is of full rank.Otherwise the matrix X'X will become singular. We know

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}.$$

So, near collinearity will have considerable effect on the precision with which β can be estimated . To diagnose the presence of collinearity or linear relationship between the explanatory variables , we consider the following steps.

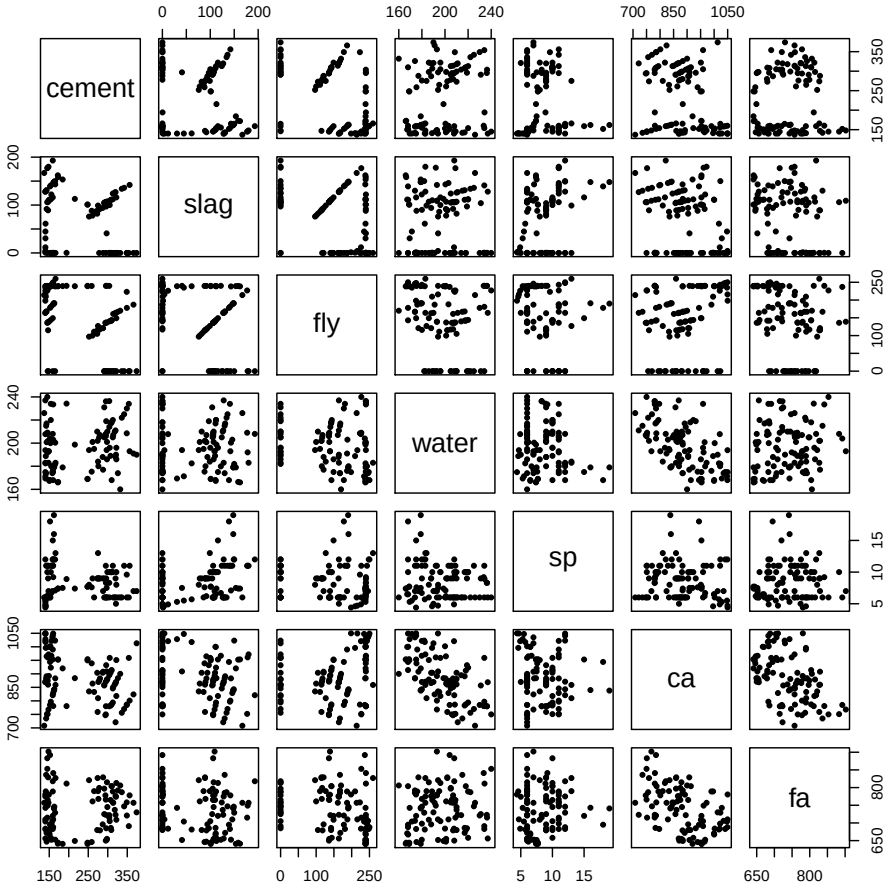
Correlation Matrix & Graphical plot

To inspect the pairwise correlation, we obtain the correlation matrix and check whether pairwise correlations are significantly large.

```
       cement  slag   fly water    sp    ca    fa
cement   1.00 -0.24 -0.49  0.22 -0.11 -0.31  0.06
slag    -0.24  1.00 -0.32 -0.03  0.31 -0.22 -0.18
fly     -0.49 -0.32  1.00 -0.24 -0.14  0.17 -0.28
water    0.22 -0.03 -0.24  1.00 -0.16 -0.60  0.11
sp      -0.11  0.31 -0.14 -0.16  1.00 -0.10  0.06
ca      -0.31 -0.22  0.17 -0.60 -0.10  1.00 -0.49
fa       0.06 -0.18 -0.28  0.11  0.06 -0.49  1.00
```

Comment : from this matrix , the off-diagonal element are not very high , hence we can say the pairwise correlations are not very significant.

We can also observe the graphical plot as given below which also depicts the pairwise correlation.



VIF & CONDITION NUMBER

A numerical evidence of the presence or absence of Multicollinearity is the Variance Inflation Factors and the Condition number (obtained from correlation matrix) . Very high VIF's and condition number indicate the presence of collinearity in the data and require further remedial measures to be taken.

We obtain the values as follow –

```
[1] "The VIF's are : "
   cement      slag       fly     water        sp        ca        fa
48.570807 55.276977 58.649500 31.431899  2.139998 88.171895 49.961057
```

```
Attaching package:  'pracma'

The following object is masked from 'package:car':

    logit

[1] "Condition Number :  728.697067865526"
```

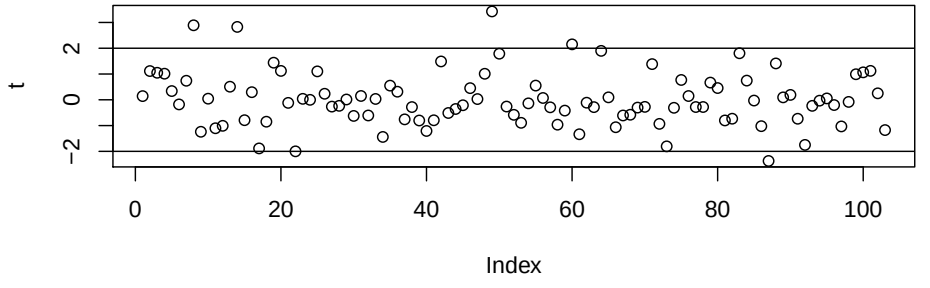
Observation : We observe that in the data VIF for Coarse Aggr. is extremely large. We can recalculate the measures of multicollinearity by deleting this covariate. The obtained values of VIF's and Condition Number are as follows.

```
[1] "The VIF's are : "
  cement     slag      fly    water       sp       fa
1.886804 1.832936 2.318386 1.117704 1.158483 1.323080
[1] "Condition Number :  5.68689368800087"
```

Comment : We observe that after eliminating the covariate Coarse Aggr. all the VIF's and Condition Number decrease considerably.

Outlier detection

The external studentised residual $t_i(= \frac{e_i}{S(i)})$ shows the impact of the i th observation on the regression analysis .If $|t_i| > 2$,then we suspect the i th observation to be an outlier.



Comment: 8,14,49,60 and 87 th observation is suspected to be outlier

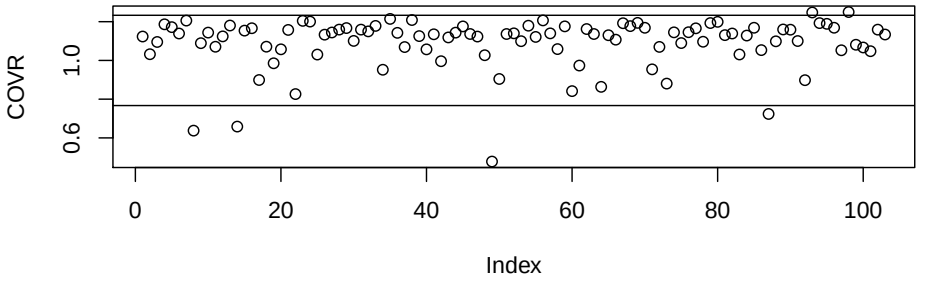
We can use various robust measures for outlier detection.

Covratio

It is the ratio measuring the scalar change in estimated error covariance matrix compared to its leave-one out counterpart:

$$COVRATIO = \frac{det\{S(i)^2[X(i)'X(i)]^{-1}\}}{det\{S^2[X'X]^{-1}\}} = \left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p}\right)^{-p}(1-h_i)^{-1}$$

Observation having $|COVRATIO - 1| > 3p/n$ are considered to be high influencial points.

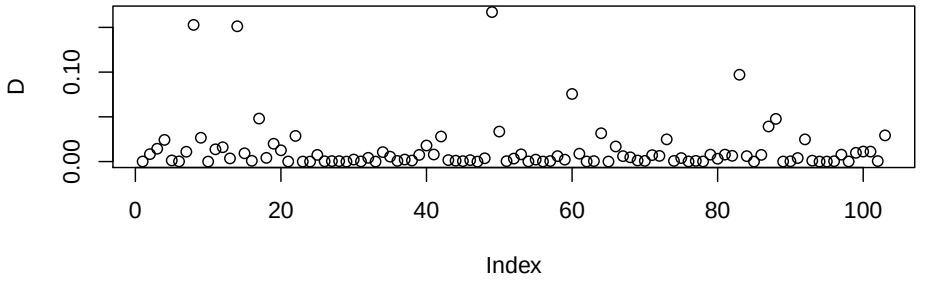


|COVRATIO| of the 8,14,49,87,93 and 98 th observation is larger than the cutoff.

Cooks Distance

The cooks distance measures the distance between $\hat{\beta}$ and $\hat{\beta}(i)$ which is denoted by $D_i = \frac{(\hat{\beta}(i)-\hat{\beta})^T X^T X(\hat{\beta}(i)-\hat{\beta})}{pS^2}$

If $D_i > F_{p,n-p}^{0.10}$ then we can suspect that the i th observation is an outlier.

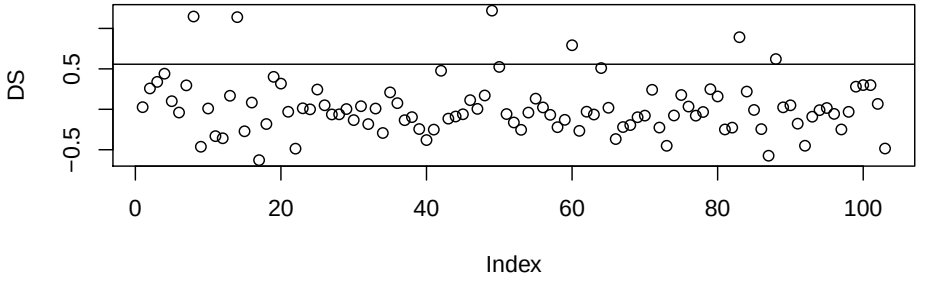


The cook's distance of the observations are significantly low.So, we can't find any outlier by this measure.

DFFITSi

DFFITSi measures the change in the i th fitted value which is denoted by $DFFITS_i = t_i(\frac{h_i}{1-h_i})^{\frac{1}{2}}$

The cut off for this measure is $2\sqrt{p/n}$



DFFITSi of the 8,14,17,49,60,83,87 and 88 th observation is larger than the cutoff.

Hypothesis testing

For testing the hypothesis whether the i th observation is an outlier,we will use outlier shifted model,

$y = X\beta + Z\gamma + \epsilon$ where Z is a diagonal matrix with ith diagonal element as 1 and other elements as zero.

We have to test $H_0 : \gamma = 0$ vs $\gamma \neq 0$

The test statistic is $F = t_i^2 \sim F_{1,n-p-1}$ under H_0

We reject H_0 if observed $t_i^2 > F_{1,n-p-1}^{0.05}$

```
## [1]  8 14 22 49 60 87
```

Conclusion:In the light of the given data, we can reject the null hypothesis and conclude that 8,14,22,49,60 and 87 th observation are the outliers.

Masking and Swamping

To find the effect of masking and swamping we have to use leave many out technique.

We have to estimate the regression estimates after removing two or more observations from the data to find the impact of the removed observations.

Cook's suggested the measure for estimating the distance of $\hat{\beta}$ and $\hat{\beta}(\underset{\sim}{i})$ to evaluate the impact of $\underset{\sim}{i}$ th observation which is denoted by $D_{\underset{\sim}{i}} = \frac{(\hat{\beta}(\underset{\sim}{i})-\hat{\beta})^T X^T X(\hat{\beta}(\underset{\sim}{i})-\hat{\beta})}{pS^2}$

The cut off value is $F_{p,n-p}^{0.10}$

Using leave-two-out technique,all the values $D_{\underset{\sim}{i}}$are found to smaller than the cut off value.

Using leave-three-out technique,we found the same result.

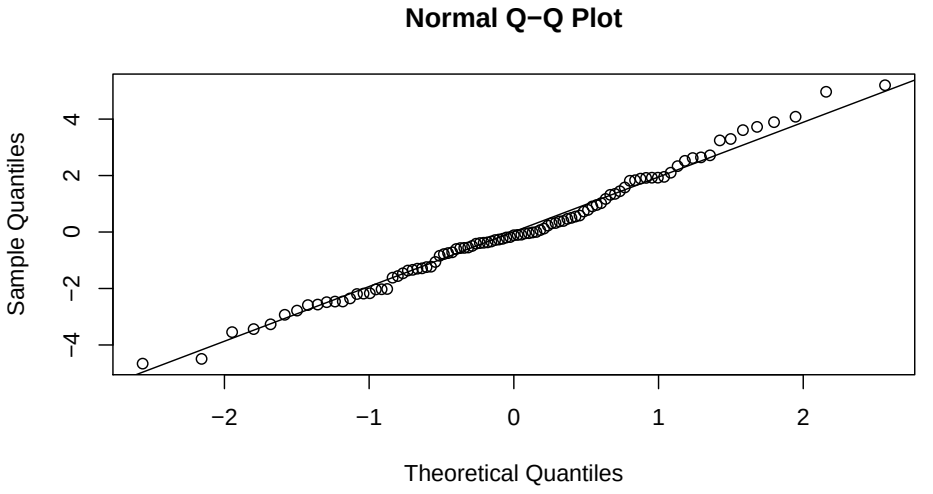
Then we can't suspect any pair or tuple of the observations to be outlier.

—- We remove the detected influencial points from the data and check for the assumptions once again. We truncate 8,14,22,49,60 and 87 th observations and now further deal with the data with 97 rows.

Normality in the Data:

Residual Plot

Truncating the possible influential points we again inspect the Q-Q plot.



Comment : From the above Q-Q plot, we can observe that quantiles of the residual resemble very closely to the quantiles of Normal Distribution. Hence there is no visual evidence of non-Normality in the data.

Shapiro-Wilk Test

To test , $H_o : e_i's$ has come from a Normally distributed population vs. H_1 : not H_0

Test Statistic :

$$W = \frac{(\sum_{i=1}^{n} a_i e_{(i)})^2}{(\sum_{i=1}^{n} e_i)^2},$$

$e_{(i)}$ denote the ith residuals in the ordered sample.

The coefficients $a_i's$ are given by :

$$(a_1, a_2, ...a_n)' = \frac{m' V^{-1}}{||m' V^{-1}||}$$

where, the vector m is the expected values of ordered sample from Standard Normal Distribution . V is the covariance matrix of those Normal ordered statistics.

After eliminating the influencial points, we conduct the test in similar way.

We compute p-value of the test using R , as given below –

```
Shapiro-Wilk normality test

data:  e
W = 0.98908, p-value = 0.6121
```

Comment : From the output of Shapiro-Wilk Test, we see that the p-value is 0.6121. The Null Hypothesis is accepted at level 0.01.Hence we can conclude that the Normality assumptions of the errors is valid.

Heteroscedasticity:

To check whether elimination of the influential points induce any sign of heteroscedasticity in our diagnosis , we conduct the similar tests and observations with the shortened dataset.

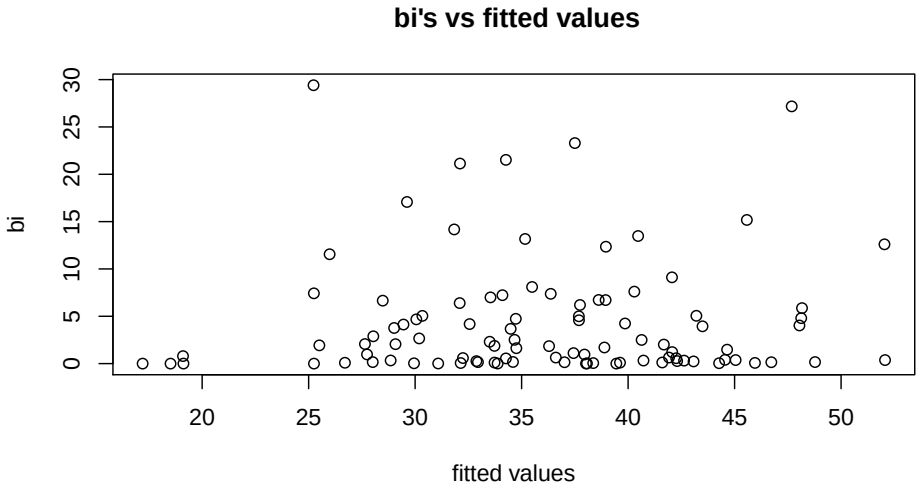
Several Plots :

- Plot of $b_i's$ vs. fitted values

Let us consider the quantity $b_i = \frac{e_i^2}{1-h_i}$, i=1(1)n

when the variances are all equal to σ^2,E$(b_i) = \sigma^2$, i =1(1)n . Thus, plotting the $b_i's$ against the fitted values should result in a wedge-shaped display, when the variances increase with the mean.

By plotting $b_i's$ against the fitted values we get the plot as following :



Comment : From the plot, we do not see any significant wedge shape, hence the variances puportedly do not increase with the increase in the fitted values.

Breusch-Pagan Test

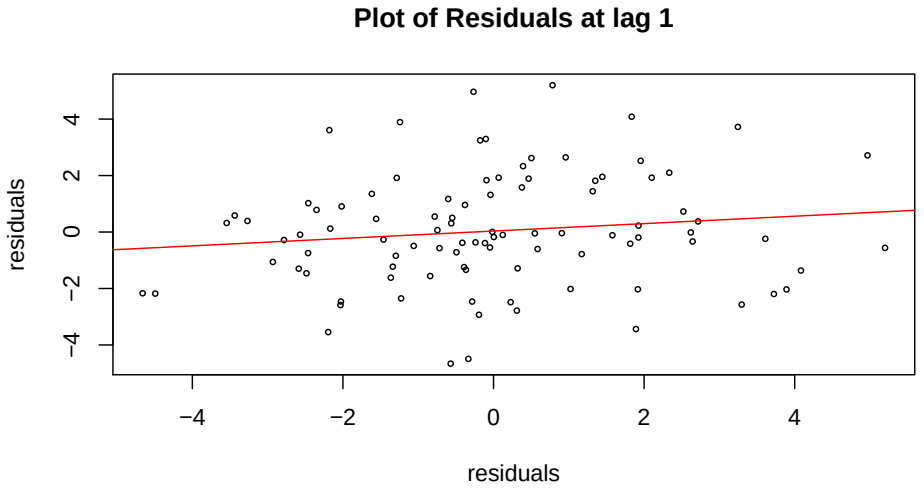
After the removal of high influential points, we again test for heteroscedasticity of the residuals.

The p value of BP test is 0.268. So we can conclude that the residuals are homoscedastic.

Autocorrelation :

Residual plot

According to our assumption the error $\epsilon_i's$ are uncorrelated among themselves. To check the validity of this assumption we assume that the other assumptions hold true. Since the residuals are expected to reflect the nature of the errors, we observe the plot of residuals vs residuals with lag 1 to check for the possible presence of autocorrelation , in the data.



Comment: From the above plot, no significant correlation is visible among the residuals (considering lag1). Also the autocorrelation coefficient obtained from these residuals is 0.13 which is not very significant.

Durbin Watson Test

Eliminating the outliers we again conduct the Durbin Watson Test.

Suppose that the error ϵ_i follow a first order autoregressive process , $\epsilon_i = \rho\epsilon_{i-1} + \delta_i$, where $\delta_i's$ are independently and identically distributed as N(0,σ^2).

Since autocorrelation from the residuals is positive we test for $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. The test statistic is given by –

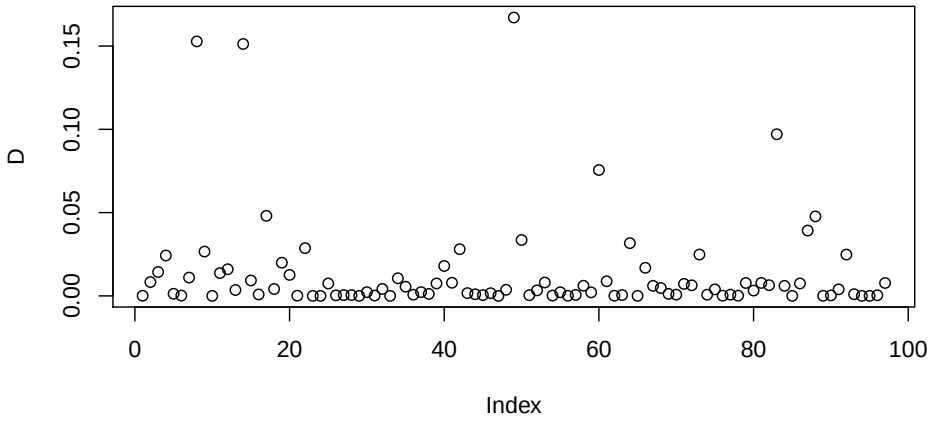
$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

We conduct the test from the residuals, using R . We get the output as,

Comment : in the test, the p-value is 0.076. Hence the null hypothesis gets accepted at level α= 0.01. We get no significant evidence in favour of the autocorrelation from the data.

Outlier Detection

We have evaluated the cooks distance.



Comment: No observation can be suspected to be outlier.

PRESENCE OF MULTICOLLINEARITY

In the formulation of regression technique one crucial assumption is that the design matrix X is of full rank.Otherwise the matrix X'X will become singular. We know

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

So, near collinearity will have considerable effect on the precision with which β can be estimated . To diagnose the presence of collinearity or linear relationship between the explanatory variables , we consider the following steps.

Correlation Matrix & Graphical plot

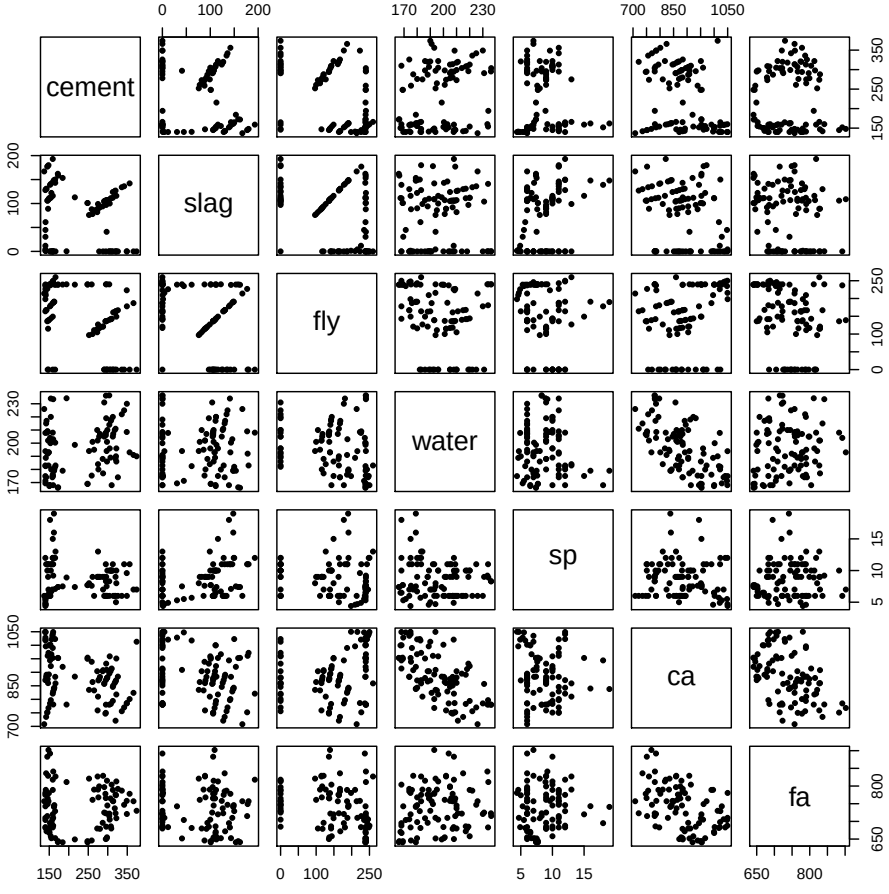
To inspect the pairwise correlation, we obtain the correlation matrix and check whether pairwise correlations are significantly large.

```
       cement  slag   fly water    sp    ca    fa
cement   1.00 -0.21 -0.47  0.29 -0.12 -0.35  0.04
```

```
slag    -0.21  1.00 -0.36  0.00  0.30 -0.26 -0.14
fly     -0.47 -0.36  1.00 -0.26 -0.14  0.20 -0.32
water    0.29  0.00 -0.26  1.00 -0.14 -0.64  0.15
sp      -0.12  0.30 -0.14 -0.14  1.00 -0.10  0.05
ca      -0.35 -0.26  0.20 -0.64 -0.10  1.00 -0.47
fa       0.04 -0.14 -0.32  0.15  0.05 -0.47  1.00
```

Comment : from this matrix , the off-diagonal element are not very high , hence we can say the pairwise correlations are not very significant.

We can also observe the graphical plot as given below which also depicts the pairwise correlation.



VIF & CONDITION NUMBER

A numerical evidence of the presence or absence of Multicollinearity is the Variance Inflation Factors and the Condition number (obtained from correlation matrix) . Very high VIF's and condition number indicate the presence of collinearity in the data and require further remedial measures to be taken.

We obtain the values as follow –

```
[1] "The VIF's are : "
   cement      slag       fly     water        sp        ca        fa
46.414528 53.338878 57.891303 27.175000  2.154775 86.929121 46.811982
```

```
## [1] "Condition Number :  729.098753365719"
```

Observation : We observe that in the data VIF for Coarse Aggr. is extremely large. We can recalculate the measures of multicollinearity by deleting this covariate. The obtained values of VIF's and Condition Number are as follows.

```
[1] "The VIF's are : "
  cement     slag      fly    water       sp       fa
1.895517 1.822619 2.391863 1.151795 1.147166 1.375894
```

```
## [1] "Condition Number :  6.20316995213431"
```

Comment : We observe that after eliminating the covariate Coarse Aggr. all the VIF's and Condition Number decrease considerably.

Since this is not enough evidence to delete the covariate from our model , we would try to find a suitable linear model for our data by some statistically efficient methods such as Stepwise Regression , and Lasso Regression . We would compare the methods to select a single suitable model.

MODEL 1 : STEPWISE REGRESSION

Since evaluating all possible regressions and then comparing model adequacy measure can be computationally burdensome, various methods are used for evaluating only a

small number of subset regression models by either adding or deleting regressors one at a time using Partial F statistic for these regressors.

$$F = \frac{(RSS_p - RSS_{k+1}).(n-k-1)}{RSS_{k+1}.(k-p+1)}$$

And when K=p,

$$F = \frac{(RSS_p - RSS_{p+1}).(n-p-1)}{RSS_{p+1}}$$

Stepwise regression is a method of selection of explanatory variable , in which at each step all regressors previously entered into the model are reassessed via their partial F statistics. It requires 2 cut-offs, one for entering the regressors and other for removing them; F_{IN} and F_{OUT}. We start with the model containing only the intercept term. Then at each step Forward Selection is followed by Backward Elimination, i.e. we add the regressor with greatest partial F value as well as exceeding F_{IN} and then remove the regressor with smallest partial F value which is lower than F_{OUT} as well. This procedure is repeated until Forward Selection does not add a new variable to the model.

Using R we conduct this method ,

```
Error in eval(expr, envir, enclos):  object 'stepwise' not found

Start:  AIC=395.98
cs ~ 1

         Df Sum of Sq    RSS    AIC
+ fly     1   1309.35 4323.5 372.32
+ cement  1   1056.58 4576.2 377.83
+ slag    1    561.37 5071.4 387.80
+ fa      1    319.29 5313.5 392.32
+ water   1    246.29 5386.5 393.64
+ ca      1    155.19 5477.6 395.27
<none>                5632.8 395.98
+ sp      1      6.19 5626.6 397.87

Step:  AIC=372.32
cs ~ fly

         Df Sum of Sq    RSS    AIC
+ cement  1   3158.04 1165.4 247.15
+ ca      1    404.66 3918.8 364.79
+ slag    1    133.33 4190.1 371.28
```

```
<none>                4323.5 372.32
+ fa      1     44.16 4279.3 373.32
+ water   1     43.68 4279.8 373.33
+ sp      1      7.69 4315.8 374.15
- fly     1   1309.35 5632.8 395.98

Step:  AIC=247.15
cs ~ fly + cement

         Df Sum of Sq    RSS    AIC
+ water   1     322.4  843.0 217.74
+ slag    1     266.3  899.1 223.99
+ sp      1     237.9  927.5 227.00
<none>                1165.4 247.15
+ ca      1      15.4 1150.1 247.87
+ fa      1       1.0 1164.4 249.07
- cement  1    3158.0 4323.5 372.32
- fly     1    3410.8 4576.2 377.83

Step:  AIC=217.74
cs ~ fly + cement + water

         Df Sum of Sq    RSS    AIC
+ ca      1     336.7  506.3 170.29
+ slag    1     257.9  585.1 184.32
+ sp      1     168.5  674.5 198.11
<none>                843.0 217.74
+ fa      1       7.9  835.1 218.83
- water   1     322.4 1165.4 247.15
- fly     1    3048.6 3891.6 364.11
- cement  1    3436.8 4279.8 373.33

Step:  AIC=170.29
cs ~ fly + cement + water + ca

         Df Sum of Sq    RSS    AIC
+ fa      1     79.61  426.7 155.70
+ sp      1     62.31  444.0 159.55
+ slag    1     59.48  446.8 160.17
<none>                506.3 170.29
- ca      1    336.68  843.0 217.74
- water   1    643.72 1150.1 247.87
- cement  1   2822.28 3328.6 350.95
```

```
- fly     1   2936.19 3442.5 354.22

Step:  AIC=155.7
cs ~ fly + cement + water + ca + fa

         Df Sum of Sq     RSS    AIC
+ sp      1     36.80  389.92 148.95
+ slag    1     13.39  413.33 154.61
<none>                 426.72 155.70
- fa      1     79.61  506.33 170.29
- ca      1    408.40  835.13 218.83
- water   1    722.07 1148.79 249.76
- fly     1   2083.73 2510.46 325.59
- cement  1   2271.15 2697.87 332.57

Step:  AIC=148.95
cs ~ fly + cement + water + ca + fa + sp

         Df Sum of Sq     RSS    AIC
<none>                 389.92 148.95
+ slag    1      0.10  389.82 150.93
- sp      1     36.80  426.72 155.70
- fa      1     54.10  444.02 159.55
- ca      1    275.89  665.81 198.85
- water   1    534.46  924.39 230.68
- fly     1   2024.23 2414.16 323.80
- cement  1   2223.42 2613.34 331.49

Call:
lm(formula = cs ~ fly + cement + water + ca + fa + sp, data = dt1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6270 -1.3448 -0.1063  1.2794  5.2016

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.088501   9.434765   8.595 2.40e-13 ***
fly          0.071697   0.003317  21.615  < 2e-16 ***
cement       0.081433   0.003595  22.654  < 2e-16 ***
water       -0.181730   0.016362 -11.107  < 2e-16 ***
ca          -0.032905   0.004123  -7.980 4.48e-12 ***
fa          -0.015986   0.004524  -3.534  0.00065 ***
```

```
sp           0.245782   0.084334   2.914  0.00450 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.081 on 90 degrees of freedom
Multiple R-squared:  0.9308,Adjusted R-squared:  0.9262
F-statistic: 201.7 on 6 and 90 DF,  p-value: < 2.2e-16
```

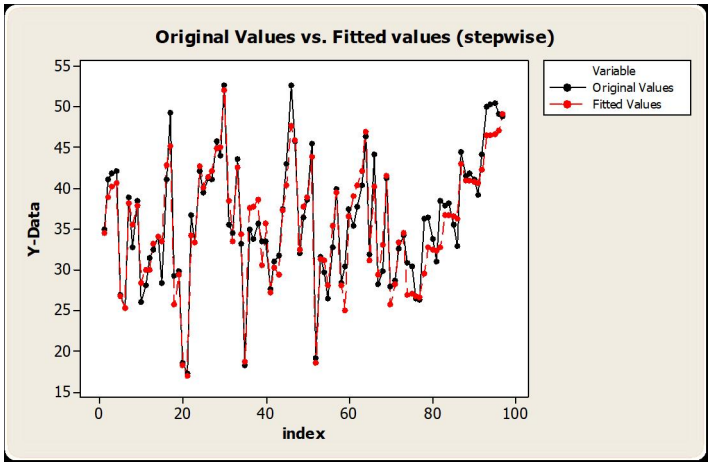
COMMENT : By stepwise regression, we obtain the fitted regression line as —

$$Y_i = 81.1 + 0.07x_{i3} + 0.08x_{i1} - 0.18x_{i4} - 0.03x_{i6} - 0.02x_{i7} + 0.25x_{i5}, i = 1(1)97$$

$\overline{R}^2 = 0.9262$ and Cross validation error of the model is 7.52 and the mean absolute deviation is 1.71

The original vs fitted response is given below,

The plot of fitted vs original values is given below,



MODEL 2 : LASSO REGRESSION

Lasso Estimate is one of the very useful shrienkage estimates which gives specifically good model in presence of multicollinearity.The SSE is minimised subject to the constraint $\sum_{i=1}^{n} |\beta_i| < s$.

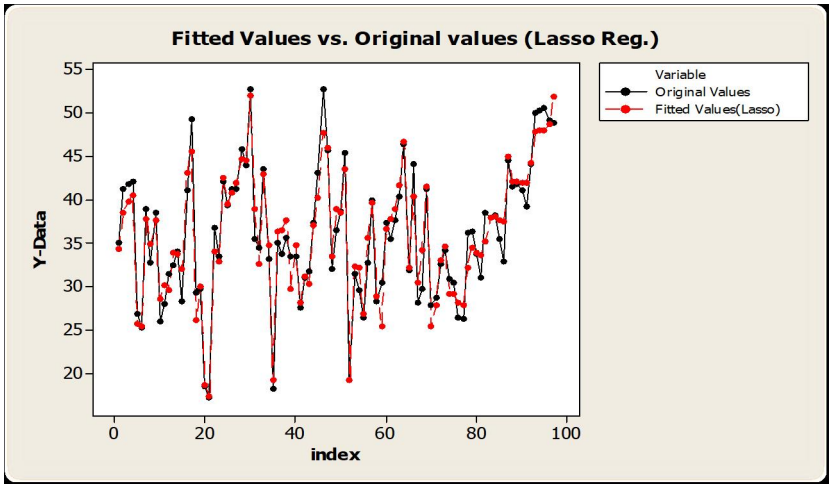
We perform the Lasso Regression using R and choose the tuning parameter by minimising CV (by in-built algorithm in R) .

The fitted regression is given by,

$$Y_i = 78.79 + 0.08x_{i1} + 0.07x_{i3} - 0.178x_{i4} + 0.24x_{i5} - 0.032x_{i6} - 0.016x_{i7}, i = 1(1)97$$

The cross validation error of the model is 4.726 and the mean absolute deviation is 1.56

The plot of fitted vs original values is given below,



MODEL 3 : ROBUST REGRESSION

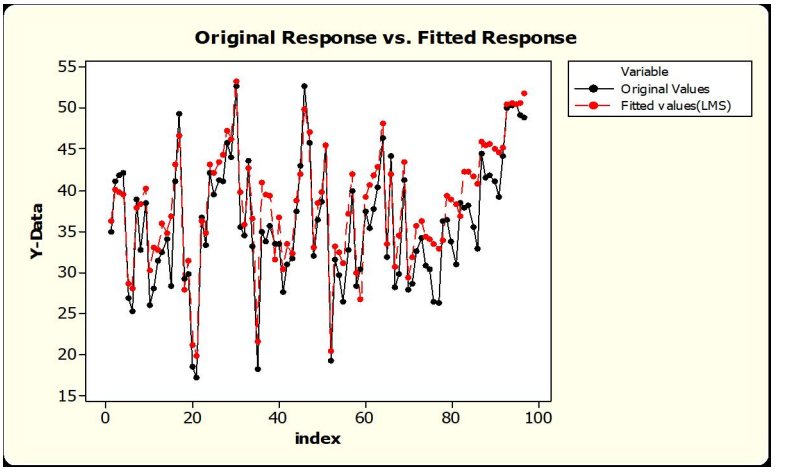
We have seen that in our data there exists a number of potential outliers which might have a considerable effect on the Least square regression . Robust regression can often give a good model based on the data. Since deletion of data reduces amount of information available for model fitting and also may delete some crucial members, we fit a linear regression model using Least median of Squares Estimates (LMS). This minimises

$$\underset{i=1(1)n}{Mediane_i^2}(\boldsymbol{\beta})$$

Using an in-built algorithm in R, we obtain the LMS estimates. The fitted regression line is given by

$$Y_i = 116.85 + 0.06x_{i1} - 0.016_{i2} + 0.06x_{i3} - 0.21x_{i4} - 0.15x_{i5} - 0.04x_{i6} - 0.03x_{i7}, i = 1(1)n$$

The original vs fitted value by LMS model is given below,



The model by LMS technique follows the general pattern of the response and it is not affected by outliers.But the cross validation error of this model is very high (3781.84) as the error got affected by influential points.

CONCLUSION:

Based on the Cross Validation Error for the 3 models described above, we conclude that Model 2 (LASSO) is the best fit considering its lowest CV among the 3. The final model for the given data is:

$$Y_i = 78.79 + 0.08x_{i1} + 0.07x_{i3} - 0.178x_{i4} + 0.24x_{i5} - 0.032x_{i6} - 0.016x_{i7}, i = 1(1)n$$