Regression Analysis Of Population Drinking Data

Instructor : Dr. Swagata Nandi

Stat-Math Unit, Indian Statistical Institute Delhi.

Group VIII

Karnajit Bhowmick

Rajatsubhra Mistry

Spandan Ghoshal

Contents

About The Data

- Here we have 46 observations from different places of the following quantiles :-
- Here our response variable is *Cirrhosis Death Rate* and others are all covariates.
- To get an overview of the data, we first perform the exploratory data analysis.

About The Data

- Here we have 46 observations from different places of the following quantiles :-

  Urban Population

  Late Births

  Wine Consumption Per Capita

  Liquor Consumption Per Capita

  Cirrhosis Death Rate

- Here our response variable is *Cirrhosis Death Rate* and others are all covariates.
- To get an overview of the data, we first perform the exploratory data analysis.

About The Data

- Here we have 46 observations from different places of the following quantiles :-

  Urban Population

  Late Births

  Wine Consumption Per Capita

  Liquor Consumption Per Capita

  Cirrhosis Death Rate

- Here our response variable is *Cirrhosis Death Rate* and others are all covariates.
- To get an overview of the data, we first perform the exploratory data analysis.

About The Data

- Here we have 46 observations from different places of the following quantiles :-

  Urban Population

  Late Births

  Wine Consumption Per Capita

  Liquor Consumption Per Capita

  Cirrhosis Death Rate

- Here our response variable is *Cirrhosis Death Rate* and others are all covariates.
- To get an overview of the data, we first perform the exploratory data analysis.

Loading the Dataset

- To get an overview of the data, we first load it in R and print first few values:-

	Ind	Ind_1	Urban population	Late births	Wine consumption per capita
1	1	1	44	33.2	5
2	2	1	43	33.8	4
3	3	1	48	40.6	3
4	4	1	52	39.2	7
5	5	1	71	45.5	11
6	6	1	44	37.5	9

	Liquor consumption per capita	Cirrhosis death rate
1	30	41.2
2	41	31.7
3	38	39.4
4	48	57.5
5	53	74.8
6	65	59.8

- Here we have “Cirrhosis death rate” as the response and “Urban population”, “Late births”, “Wine consumption per capita”, “Liquor consumption per capita” as the covariates.

Loading the Dataset

- To get an overview of the data, we first load it in R and print first few values:-

	Ind	Ind_1	Urban population	Late births	Wine consumption per capita
1	1	1	44	33.2	5
2	2	1	43	33.8	4
3	3	1	48	40.6	3
4	4	1	52	39.2	7
5	5	1	71	45.5	11
6	6	1	44	37.5	9

	Liquor consumption per capita	Cirrhosis death rate
1	30	41.2
2	41	31.7
3	38	39.4
4	48	57.5
5	53	74.8
6	65	59.8

- Here we have “Cirrhosis death rate” as the response and “Urban population”, “Late births”, “Wine consumption per capita”, “Liquor consumption per capita” as the covariates.

Type of the covariates

- To know type of each covariates, we use the **str()** function :-

```
'data.frame': 46 obs. of  7 variables:
 $ Ind                          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Ind_1                        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Urban population             : int  44 43 48 52 71 44 57 34 70 54 ...
 $ Late births                  : num  33.2 33.8 40.6 39.2 45.5 37.5 44.2 31.
 $ Wine consumption per capita  : int  5 4 3 7 11 9 6 3 12 7 ...
 $ Liquor consumption per capita: int  30 41 38 48 53 65 73 32 56 57 ...
 $ Cirrhosis death rate         : num  41.2 31.7 39.4 57.5 74.8 59.8 54.3 47.
```

- So the dataset contains no factor covariate hence we can perform multiple linear regression here. For ease of indexing, we name the columns as "I", "1", "A1", "A2", "A3", "A4", "Y".

Type of the covariates

- To know type of each covariates, we use the **str()** function :-

```
'data.frame': 46 obs. of  7 variables:
 $ Ind                          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Ind_1                        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Urban population             : int  44 43 48 52 71 44 57 34 70 54 ...
 $ Late births                  : num  33.2 33.8 40.6 39.2 45.5 37.5 44.2 31.
 $ Wine consumption per capita  : int  5 4 3 7 11 9 6 3 12 7 ...
 $ Liquor consumption per capita: int  30 41 38 48 53 65 73 32 56 57 ...
 $ Cirrhosis death rate         : num  41.2 31.7 39.4 57.5 74.8 59.8 54.3 47.
```

- So the dataset contains no factor covariate hence we can perform multiple linear regression here. For ease of indexing, we name the columns as “I”, “1”, “A1”, “A2”, “A3”, “A4”, “Y”.

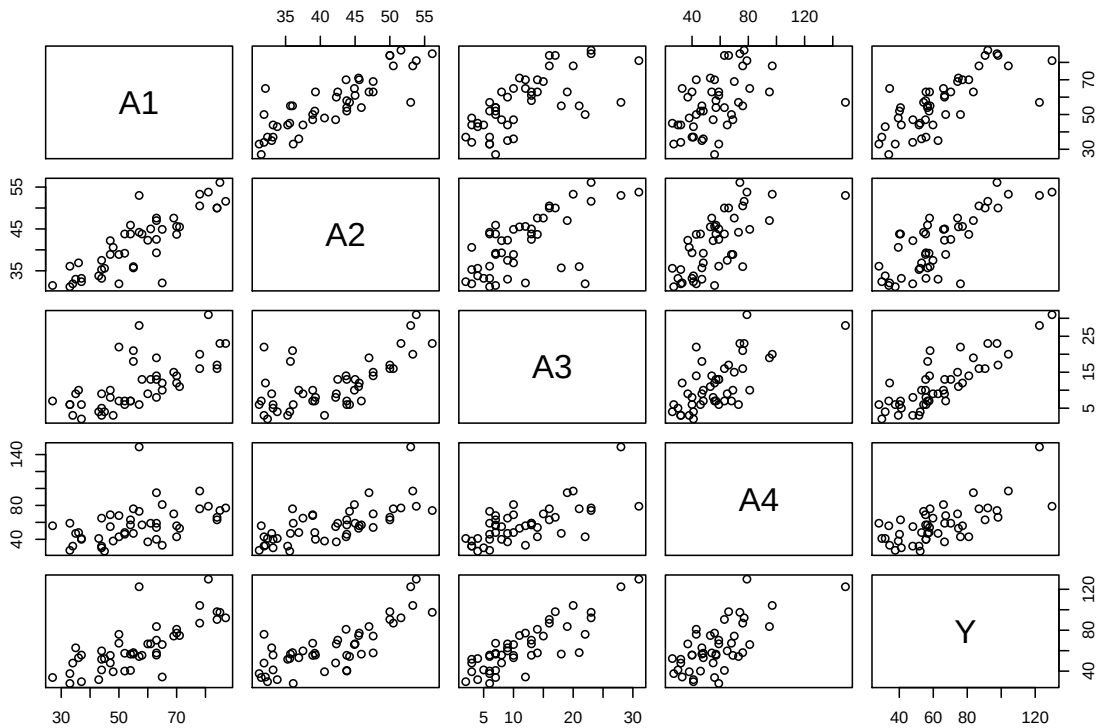
5-number Summary of Covariates

- To get an idea of the values of each covariate we calculate the 5-number summary for each of them :-

```
       A1              A2              A3              A4
 Min.   :27.00   Min.   :31.20   Min.   : 2.00   Min.   : 26.00
 1st Qu.:44.25   1st Qu.:35.62   1st Qu.: 6.25   1st Qu.: 41.50
 Median :55.00   Median :42.25   Median :10.00   Median : 56.00
 Mean   :56.26   Mean   :41.48   Mean   :11.59   Mean   : 57.50
 3rd Qu.:65.00   3rd Qu.:45.83   3rd Qu.:15.75   3rd Qu.: 68.75
 Max.   :87.00   Max.   :56.10   Max.   :31.00   Max.   :149.00
       Y
 Min.   : 28.00
 1st Qu.: 48.90
 Median : 57.65
 Mean   : 63.49
 3rd Qu.: 75.70
 Max.   :129.90
```

Pairwise Scatterplots

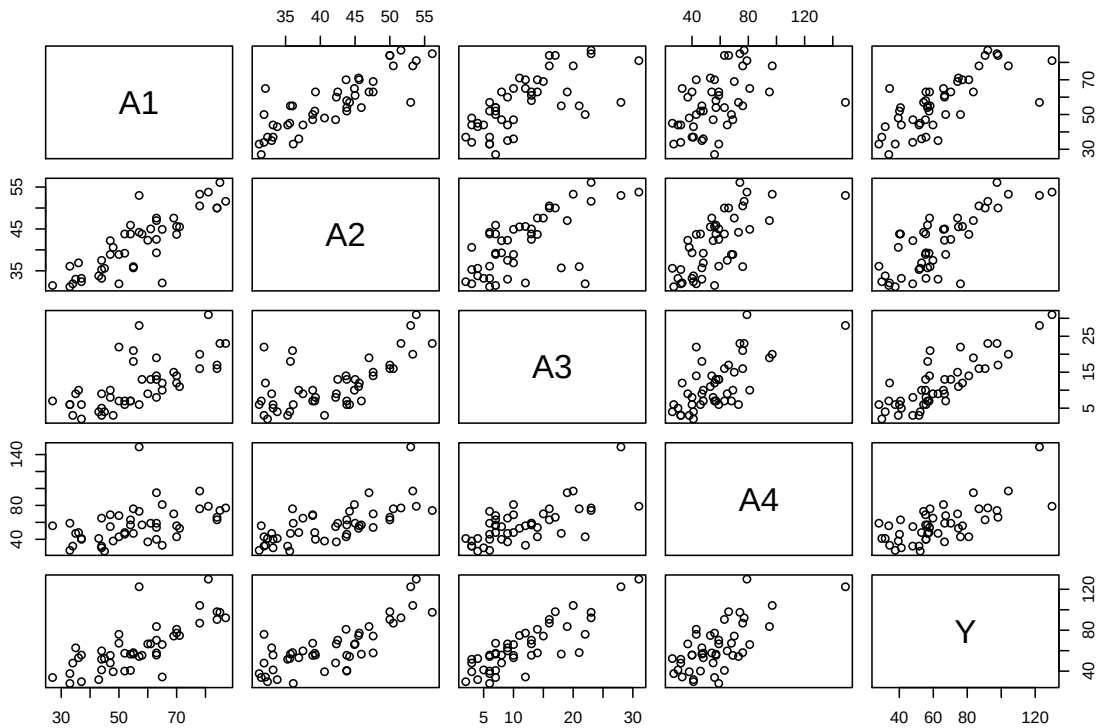
- To get an idea of the relationship between covariates and response, we make pairwise scatterplots using **pairs()** function :-



- This plot clearly indicates linear relationship between the covariates and response also. This might lead to the problem of multicollinearity which we will formally diagnose.

Pairwise Scatterplots

- To get an idea of the relationship between covariates and response, we make pairwise scatterplots using **pairs()** function :-



- This plot clearly indicates linear relationship between the covariates and response also. This might lead to the problem of multicollinearity which we will formally diagnose.

Correlation Between Covariates

- We calculate the correlation between the covariates and response to get even better idea of linear dependence between them :-

```
           A1        A2        A3        A4
A1 1.0000000 0.8432812 0.6786230 0.4402957
A2 0.8432812 1.0000000 0.6398407 0.6863643
A3 0.6786230 0.6398407 1.0000000 0.6759206
A4 0.4402957 0.6863643 0.6759206 1.0000000
```

- We can see that the correlations are high between many predictors which can lead to problem of multicollinearity.

Correlation Between Covariates

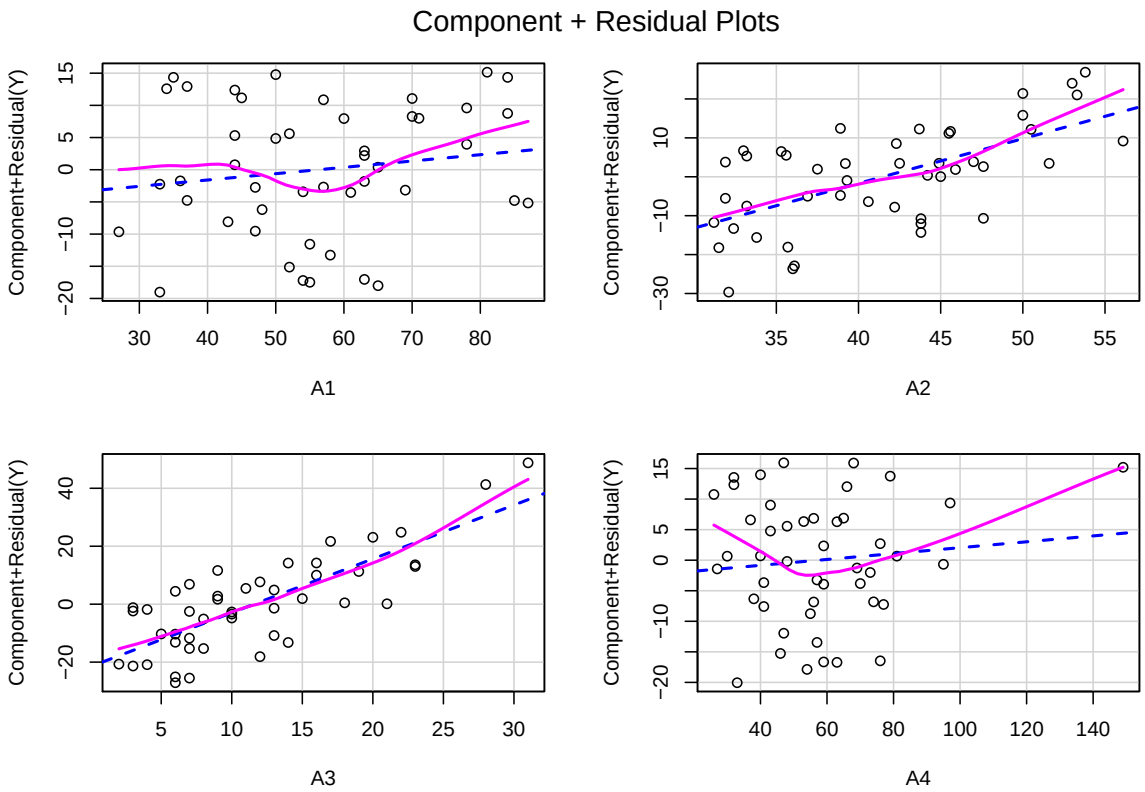
- We calculate the correlation between the covariates and response to get even better idea of linear dependence between them :-

```
           A1        A2        A3        A4
A1 1.0000000 0.8432812 0.6786230 0.4402957
A2 0.8432812 1.0000000 0.6398407 0.6863643
A3 0.6786230 0.6398407 1.0000000 0.6759206
A4 0.4402957 0.6863643 0.6759206 1.0000000
```

- We can see that the correlations are high between many predictors which can lead to problem of multicollinearity.

Partial Residual Plots

- To get an idea of the nature of relationship between the covariates and response, we make the partial residual plot for all the covariates :-



Conclusion

- The plot indicates the linear relationship between the covariates and response.
- Hence, we will fit the usual multiple linear regression model with no higher order polynomial terms.
- Later we will see if other models with interaction terms are better or not.

Conclusion

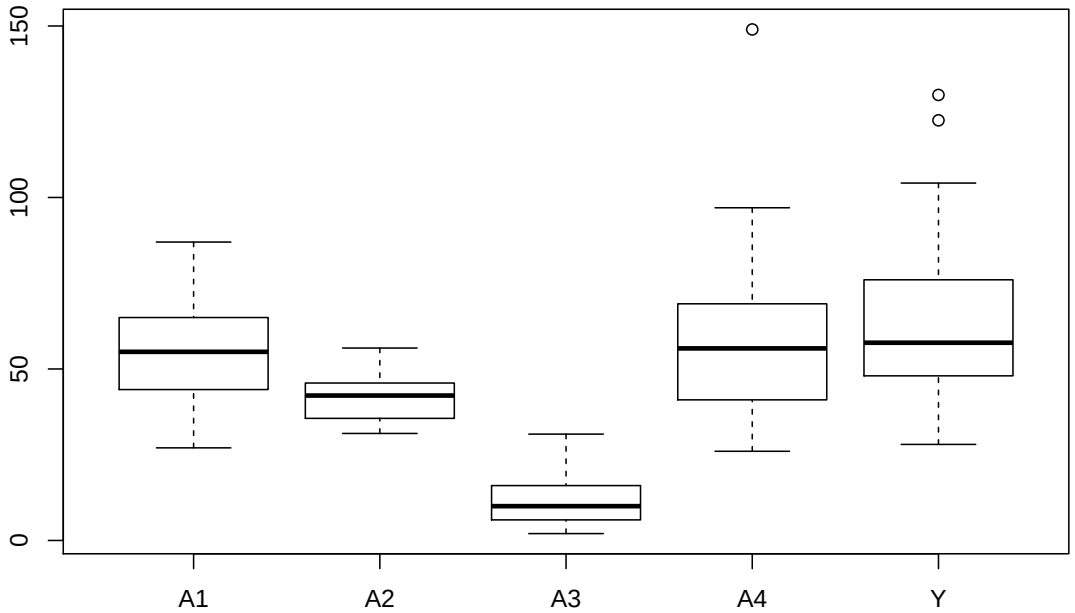
- The plot indicates the linear relationship between the covariates and response.
- Hence, we will fit the usual multiple linear regression model with no higher order polynomial terms.
- Later we will see if other models with interaction terms are better or not.

Conclusion

- The plot indicates the linear relationship between the covariates and response.
- Hence, we will fit the usual multiple linear regression model with no higher order polynomial terms.
- Later we will see if other models with interaction terms are better or not.

Boxplots of Covariates and Response

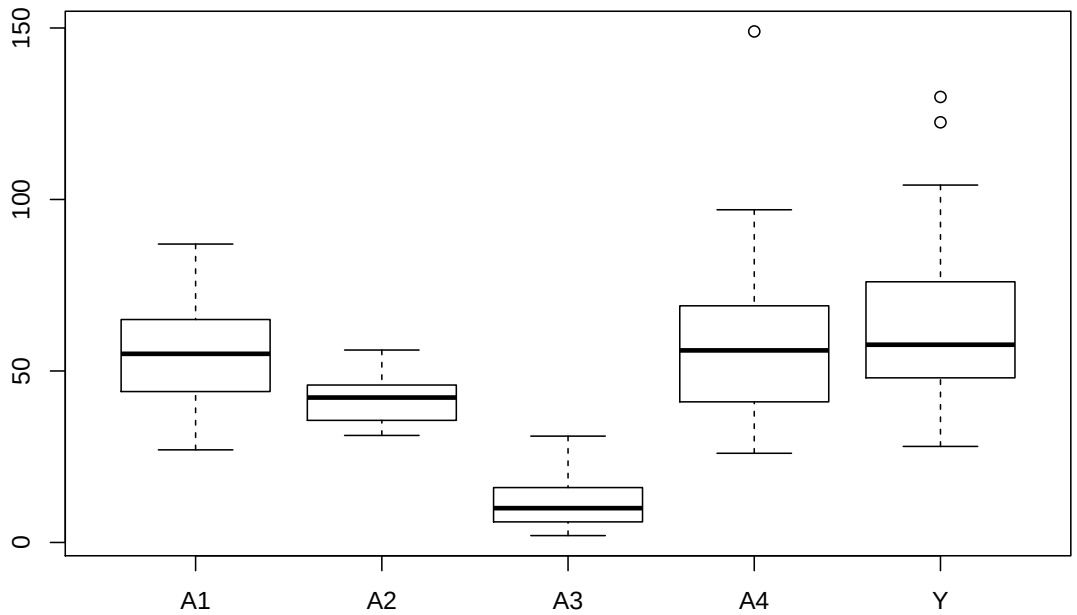
- We draw the boxplots for differnet covariates to get idea of presence of outlier / high leavarage points :-



- Here also we get some indication of possible presence of those points.

Boxplots of Covariates and Response

- We draw the boxplots for differnet covariates to get idea of presence of outlier / high leavarage points :-



- Here also we get some indication of possible presence of those points.

Fitting a Linear Model to the Dataset

- We fit a linear model of the form :-

$$\boldsymbol{Y}^{n\times 1} = \boldsymbol{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \boldsymbol{\epsilon}^{n\times 1}$$

  where $n = 46$ (total number of observed responses) and $p = 5$ where columns of $\boldsymbol{X} = \begin{bmatrix}\mathbf{1}_n & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4\end{bmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \cdots & \beta_4\end{pmatrix}$ each corresponding to the 4 different covariates.

- We fit a linear model based on the given dataset in R and then verify the different assumptions of it.

Fitting a Linear Model to the Dataset

- We fit a linear model of the form :-

$$\boldsymbol{Y}^{n\times 1} = \boldsymbol{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \boldsymbol{\epsilon}^{n\times 1}$$

  where $n = 46$ (total number of observed responses) and $p = 5$ where columns of $\boldsymbol{X} = \begin{bmatrix}\mathbf{1}_n & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4\end{bmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \cdots & \beta_4\end{pmatrix}$ each corresponding to the 4 different covariates.
- We fit a linear model based on the given dataset in R and then verify the different assumptions of it.

Features of the fitted model

- We fit the linear model specified before in the dataset using **lm()** function and to get an idea about the estimates we use the **summary()** function :-

```
Call:
lm(formula = Y ~ A1 + A2 + A3 + A4, data = X)

Residuals:
     Min       1Q   Median       3Q      Max
-18.8723  -6.7803   0.1507   7.3252  16.4419

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.96310   11.40035  -1.225   0.2276
A1            0.09829    0.24407   0.403   0.6893
A2            1.14838    0.58300   1.970   0.0556 .
A3            1.85786    0.40096   4.634 3.61e-05 ***
A4            0.04817    0.13336   0.361   0.7198
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.61 on 41 degrees of freedom
Multiple R-squared:  0.8136,Adjusted R-squared:  0.7954
F-statistic: 44.75 on 4 and 41 DF,  p-value: 1.951e-14
```

Explanation of the fitted model

- As we can see only the coefficients β_2, β_3 for covariates "A2","A3" are statistically significant.
- This does not imply that other covariates are insignificant since there maybe many problems that are hidden in the model.
- So, before concluding anything we verify all the assumptions of a linear model.

Explanation of the fitted model

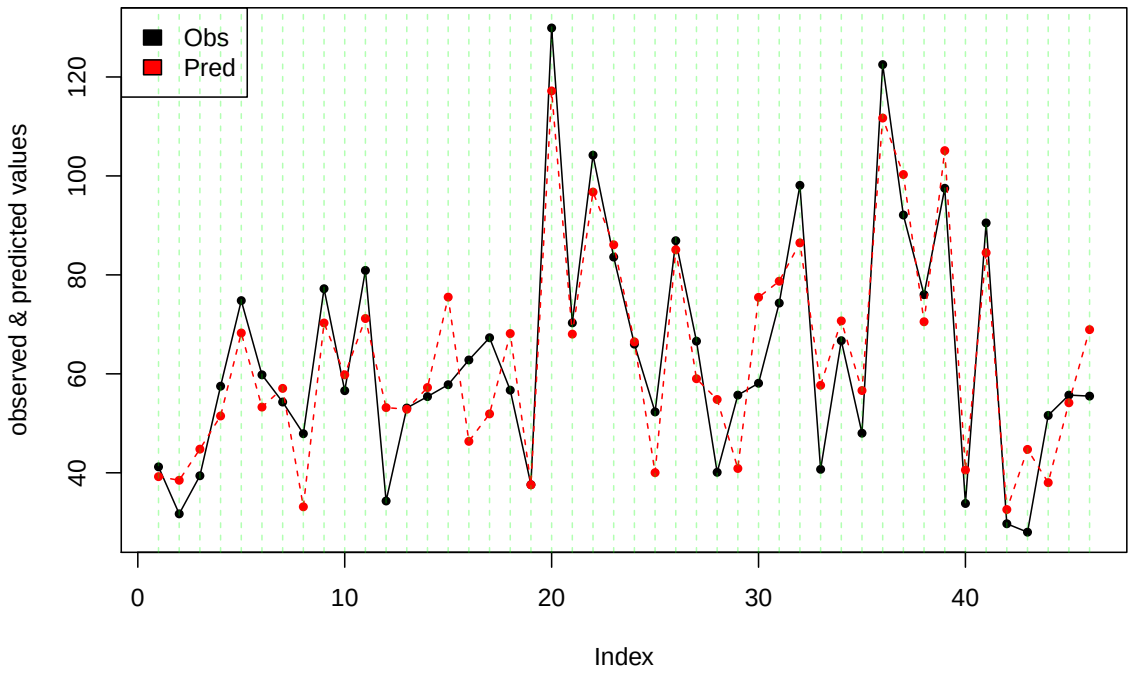
- As we can see only the coefficients β_2, β_3 for covariates "A2","A3" are statistically significant.
- This does not imply that other covariates are insignificant since there maybe many problems that are hidden in the model.
- So, before concluding anything we verify all the assumptions of a linear model.

Explanation of the fitted model

- As we can see only the coefficients β_2, β_3 for covariates "A2","A3" are statistically significant.
- This does not imply that other covariates are insignificant since there maybe many problems that are hidden in the model.
- So, before concluding anything we verify all the assumptions of a linear model.

Obs vs Fitted Values

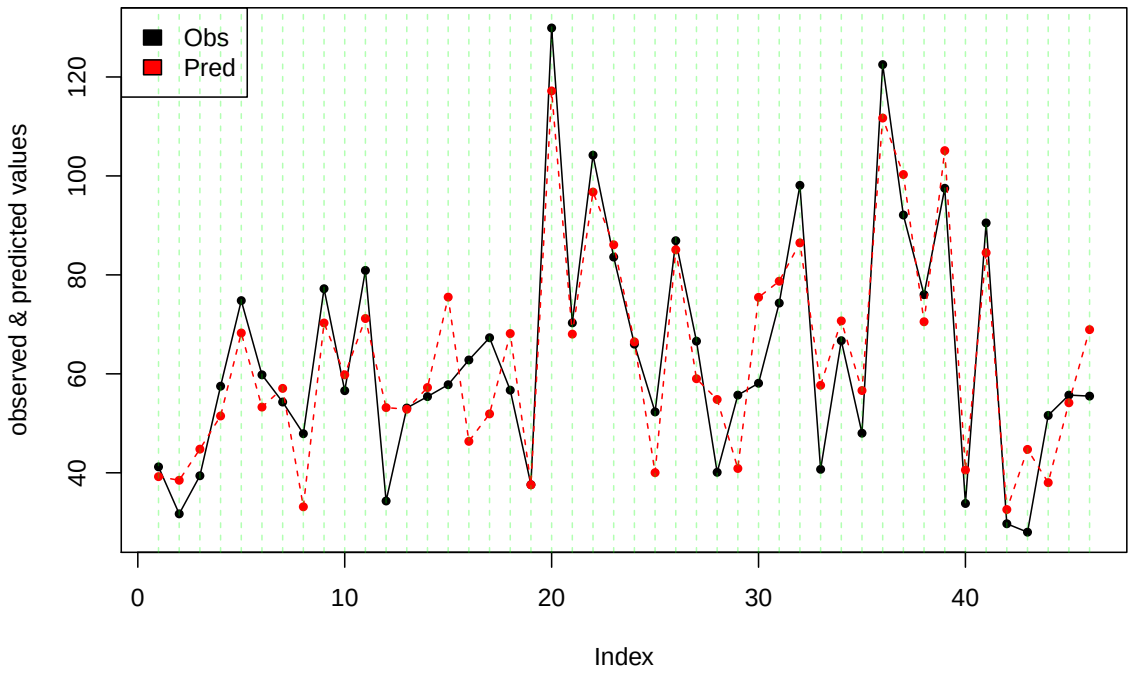
- We plot the observed vs fitted values to get some idea about prediction :-



- The fit is good except a few observations.
- There may be many reasons for this which we will eventually look into.

Obs vs Fitted Values

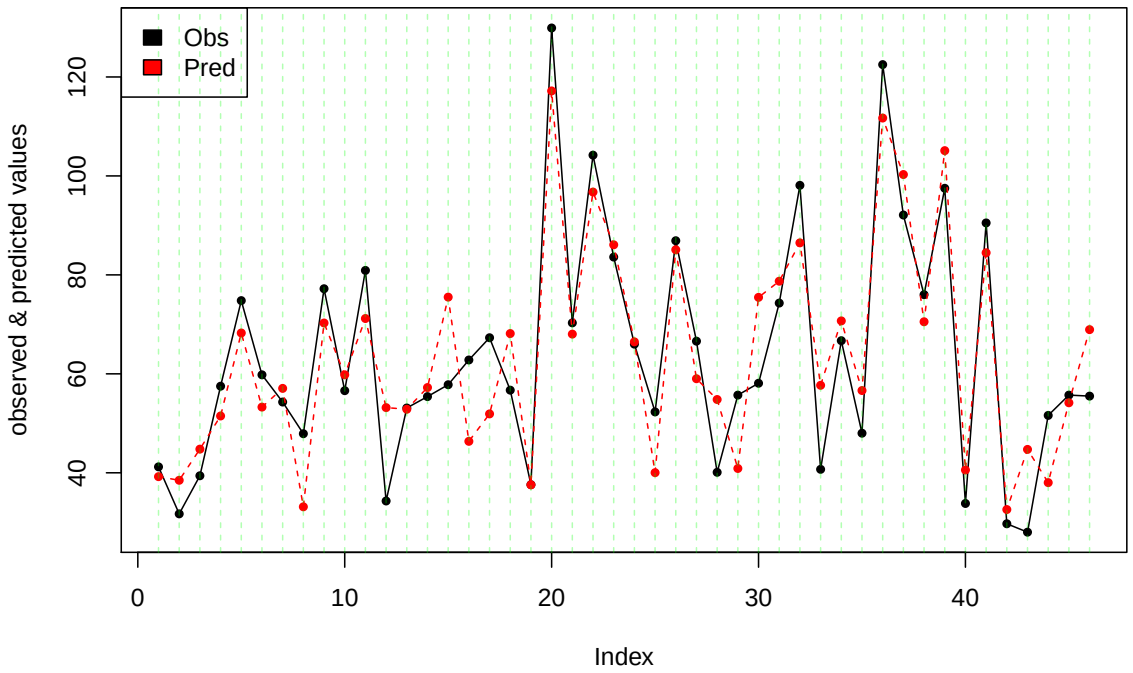
- We plot the observed vs fitted values to get some idea about prediction :-



- The fit is good except a few observations.
- There may be many reasons for this which we will eventually look into.

Obs vs Fitted Values

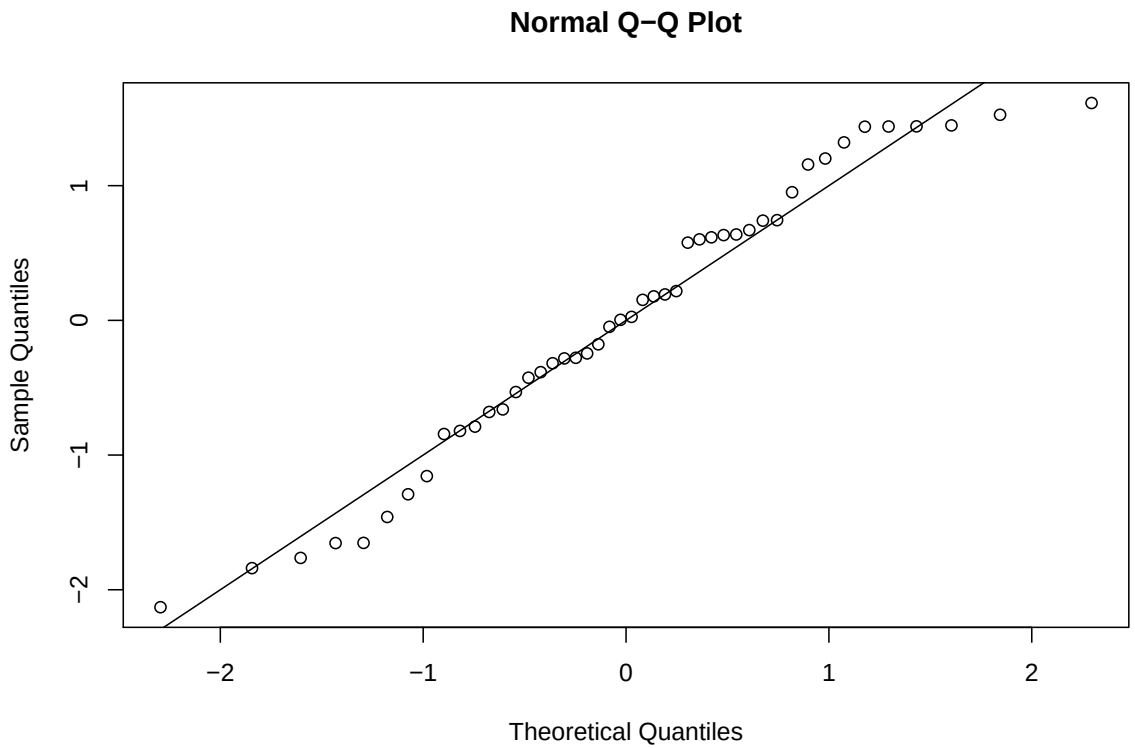
- We plot the observed vs fitted values to get some idea about prediction :-



- The fit is good except a few observations.
- There may be many reasons for this which we will eventually look into.

QQ-plot of residuals

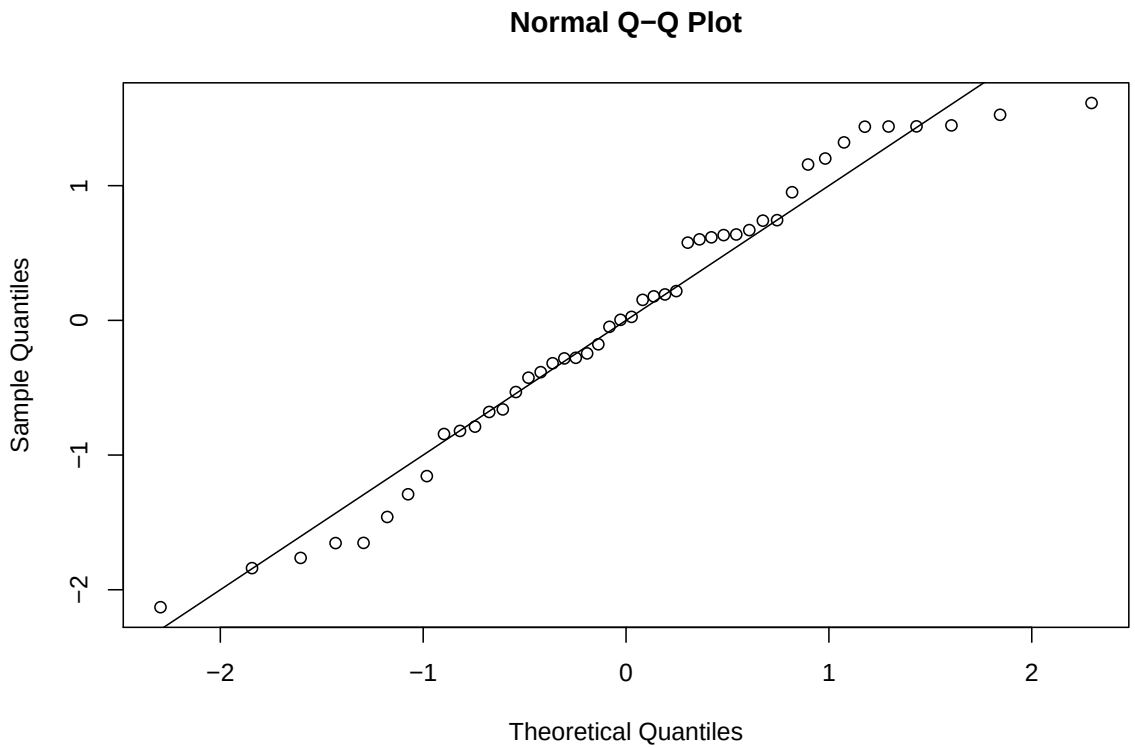
- We now plot the sorted residuals (quantiles) against the population quantiles of a normal distribution :-



- We can see the qq-plot indicates ligh tailed residuals with possible deviation from normality.
- There maybe some outlier points present which we will verify later.

QQ-plot of residuals

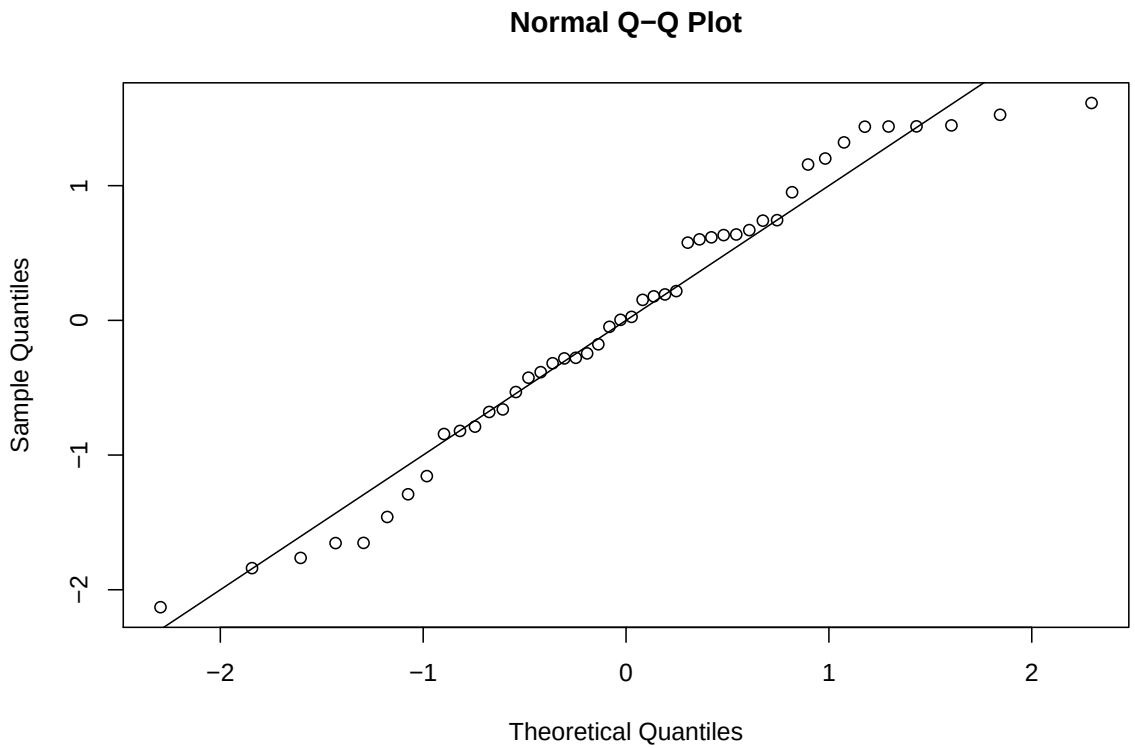
- We now plot the sorted residuals (quantiles) against the population quantiles of a normal distribution :-



- We can see the qq-plot indicates ligh tailed residuals with possible deviation from normality.
- There maybe some outlier points present which we will verify later.

QQ-plot of residuals

- We now plot the sorted residuals (quantiles) against the population quantiles of a normal distribution :-



- We can see the qq-plot indicates ligh tailed residuals with possible deviation from normality.
- There maybe some outlier points present which we will verify later.

Shapiro-Wilk Test

- We test the following hypothesis
  H_0 : residuals are normally distributed against $H_1 : H_0$ is false using Shapiro-Wilk test in R as :-

```
Shapiro-Wilk normality test

data:  resi
W = 0.95987, p-value = 0.1133
```

- Though the p-value is more than 0.1 but this doesn't give strong evidence in favour of H_0 so we will further check for presence of correlation between the errors and other issues also.

Shapiro-Wilk Test

- We test the following hypothesis
  H_0 : residuals are normally distributed against $H_1 : H_0$ is false using Shapiro-Wilk test in R as :-

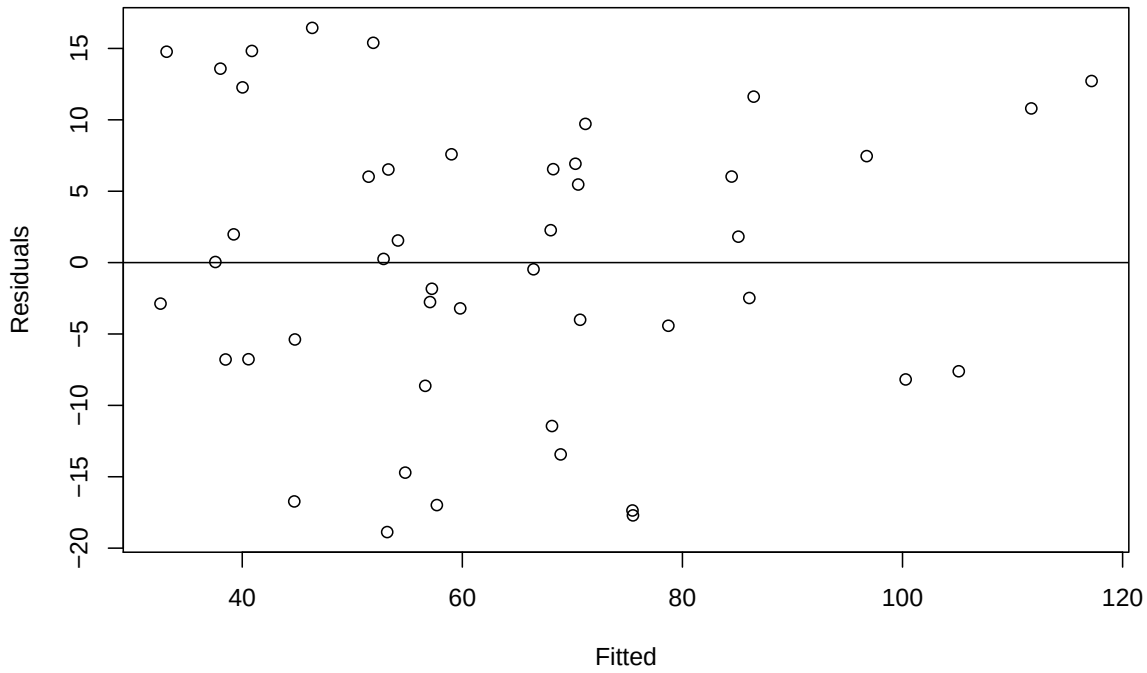
```
Shapiro-Wilk normality test

data:  resi
W = 0.95987, p-value = 0.1133
```

- Though the p-value is more than 0.1 but this doesn't give strong evidence in favour of H_0 so we will further check for presence of correlation between the errors and other issues also.

Checking homoscedasticity Assumptions

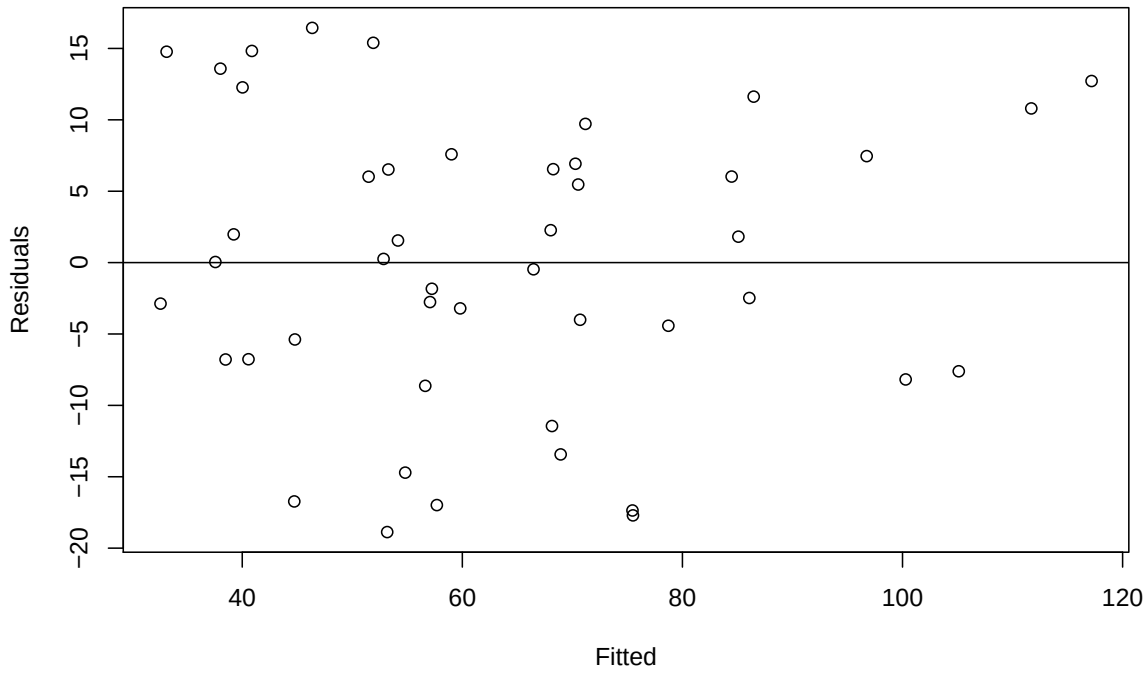
- First to check homoscedasticity assumption, we make the residuals $(\widehat{\varepsilon})$ vs fitted $(\widehat{y})$ plots :-



- We can see the plot doesn't give any clear indication of presence of heteroscedasticity. So, we can't conclude anything hence we will perform confirmatory tests.

Checking homoscedasticity Assumptions

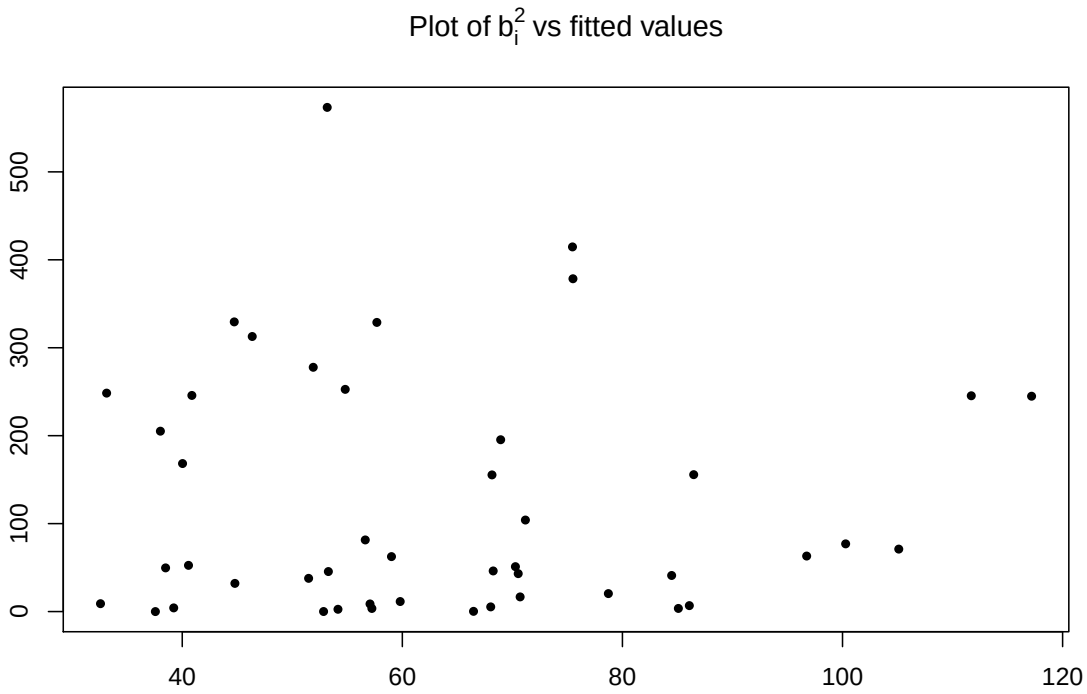
- First to check homoscedasticity assumption, we make the residuals $(\widehat{\varepsilon})$ vs fitted $(\widehat{y})$ plots :-



- We can see the plot doesn't give any clear indication of presence of heteroscedasticity. So, we can't conclude anything hence we will perform confirmatory tests.

b_i vs $\widehat{y}_i$ plot

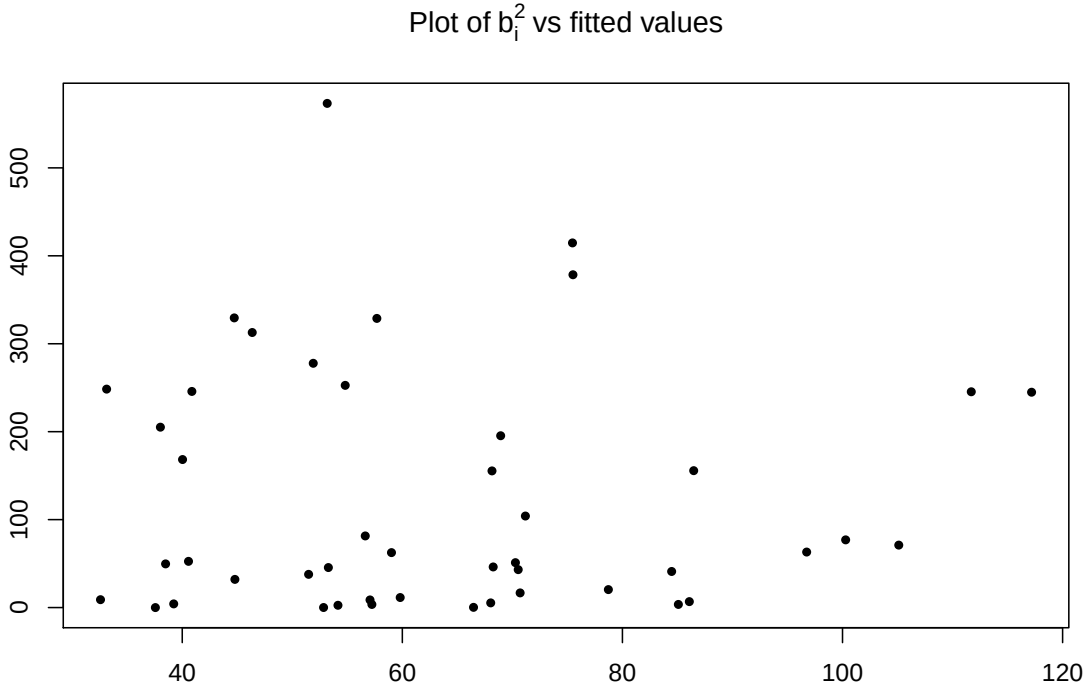
- A standard technique to detect presence of heteroscedasticity is to plot the quantities $b_i = \frac{e_i^2}{1-h_i}$ against the fitted values $\widehat{y}_i$.
- We make the plot using R :-



- This plot gives no indication of any heteroscedasticity present in the residuals.

b_i vs $\widehat{y}_i$ plot

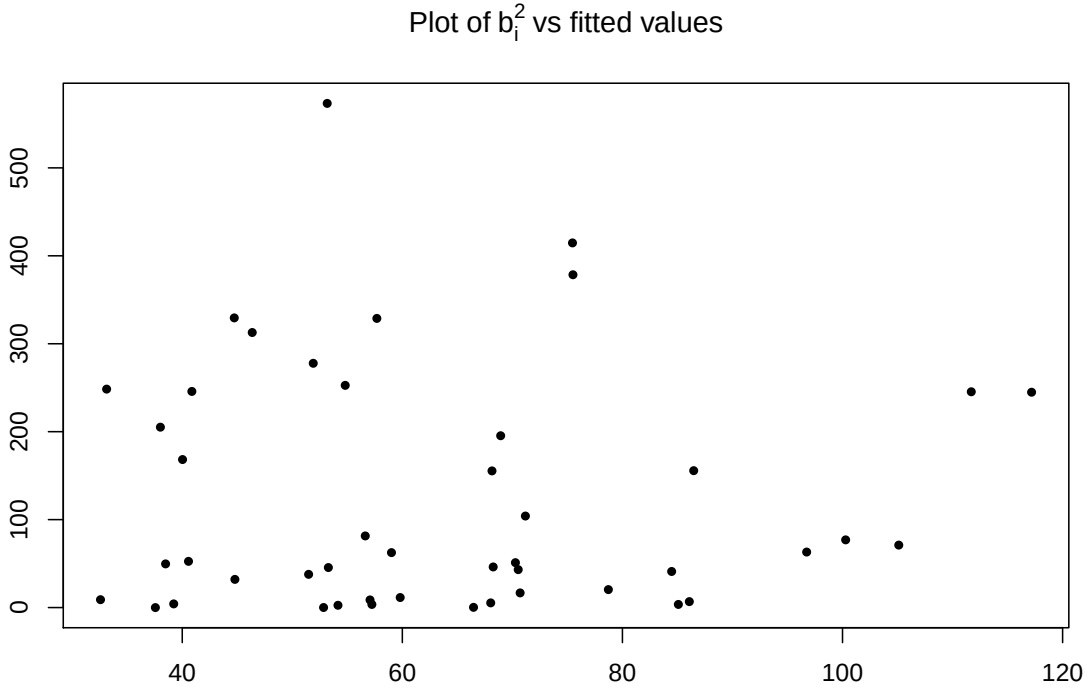
- A standard technique to detect presence of heteroscedasticity is to plot the quantities $b_i = \frac{e_i^2}{1-h_i}$ against the fitted values $\widehat{y}_i$.
- We make the plot using R :-



- This plot gives no indication of any heteroscedasticity present in the residuals.

b_i vs $\widehat{y}_i$ plot

- A standard technique to detect presence of heteroscedasticity is to plot the quantities $b_i = \frac{e_i^2}{1-h_i}$ against the fitted values $\widehat{y}_i$.
- We make the plot using R :-



- This plot gives no indication of any heteroscedasticity present in the residuals.

Breusch-Pagan Test

- We perform the Breusch–Pagan test for testing the homoscedasticity assumptions using R :-

```
studentized Breusch-Pagan test

data:  reg
BP = 2.6929, df = 4, p-value = 0.6105
```

- We can see that the p-value of the outcome is satisfactorily high so we can safely assume the error variances to be equal.

Breusch-Pagan Test

- We perform the Breusch–Pagan test for testing the homoscedasticity assumptions using R :-

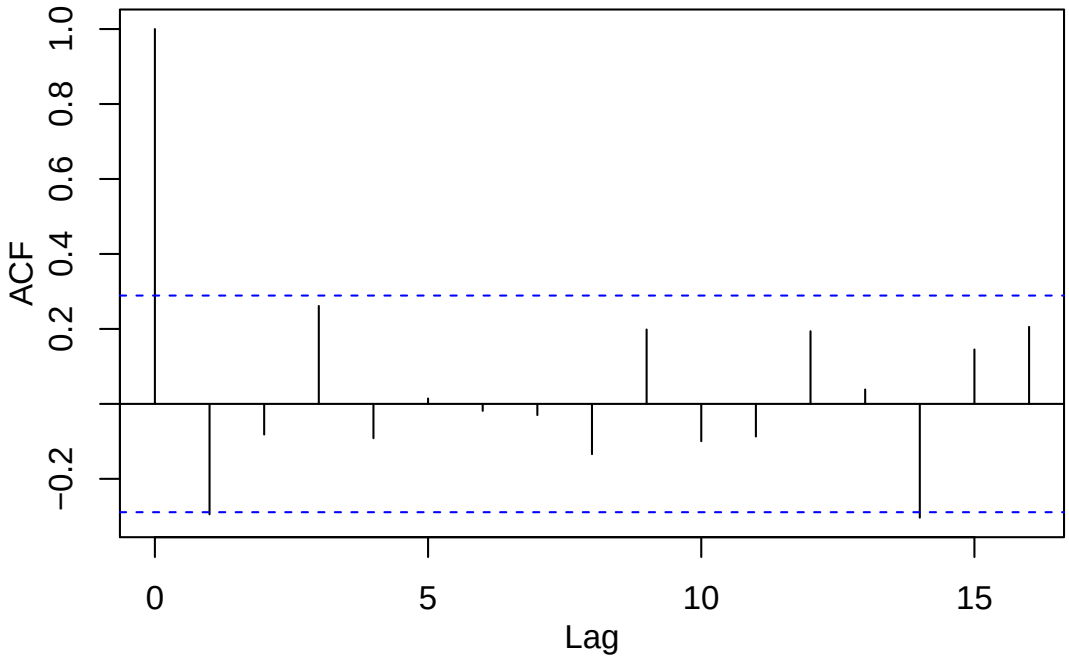
```
studentized Breusch-Pagan test

data:  reg
BP = 2.6929, df = 4, p-value = 0.6105
```

- We can see that the p-value of the outcome is satisfactorily high so we can safely assume the error variances to be equal.

ACF plot

- If the errors in the model are truely independent, then we will expect the sample autocorrelation coefficients for different lags k to be insignificant.

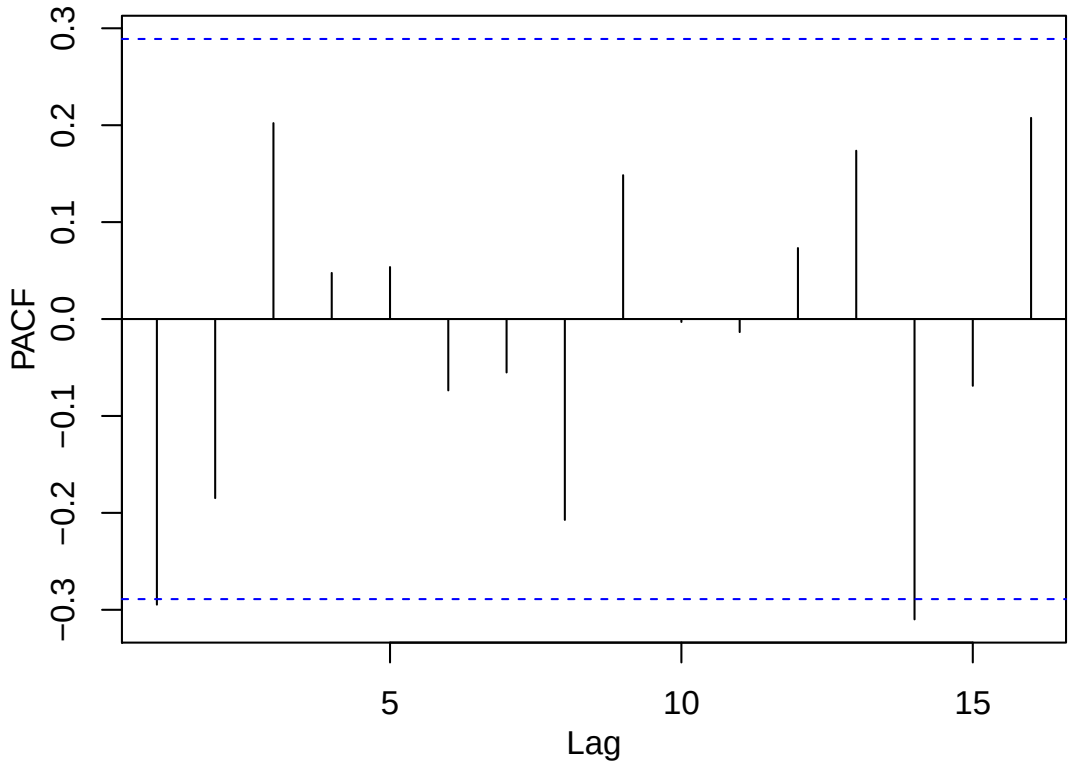


ACF plot

- This plot clearly gives indication of no presence of any type of correlation between the residuals.

PACF plot

- Similarly, we plot the sample partial autocorrelation coefficients for different lags and got the same kind of observations indicating no presence of correlations.



Durbin-Watson Test

- To test the null hypothesis H_0 : errors are uncorrelated against H_1 : errors are correlated, we perform Durbin-Watson test which gives the following results :-
- We get the observed p-value of the Durbin-Watson statistic to be equal to 0.9734 favouring the uncorrelated assumption.

Durbin-Watson Test

- To test the null hypothesis H_0 : errors are uncorrelated against H_1 : errors are correlated, we perform Durbin-Watson test which gives the following results :-
- We get the observed p-value of the Durbin-Watson statistic to be equal to 0.9734 favouring the uncorrelated assumption.

Breusch–Godfrey test

- To check whether residuals are uncorrelated for higher orders, we perform the Breusch–Godfrey test upto order 20.

```
Breusch-Godfrey test for serial correlation of order up to 20

data:  lm(Y ~ A1 + A2 + A3 + A4, data = X)
LM test = 24.951, df = 20, p-value = 0.2033
```

- Since the test gives p-value more than 0.05, we can accept H_0 hence the assumption of uncorrelated residuals can be assumed to be satishfied.

Breusch–Godfrey test

- To check whether residuals are uncorrelated for higher orders, we perform the Breusch–Godfrey test upto order 20.

```
Breusch-Godfrey test for serial correlation of order up to 20

data:  lm(Y ~ A1 + A2 + A3 + A4, data = X)
LM test = 24.951, df = 20, p-value = 0.2033
```

- Since the test gives p-value more than 0.05, we can accept H_0 hence the assumption of uncorrelated residuals can be assumed to be satishfied.

Hat Matrix Diagonals

- To detect high leverage points, we compute the hat matrix diagonals h_i of the matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$:-

```
        36         20         12         38         30         39
0.4994692 0.3042241 0.3015417 0.2998329 0.2078749 0.1721364
```

- We find out if there is any diagonal element with value $> \frac{2p}{n}$ as they should be looked at more closely.

```
        12         20         36         38
0.3015417 0.3042241 0.4994692 0.2998329
```

- Hence we will apply other procedures also to confirm whether these points are influential or not.

Hat Matrix Diagonals

- To detect high leverage points, we compute the hat matrix diagonals h_i of the matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$:-

```
        36         20         12         38         30         39
0.4994692 0.3042241 0.3015417 0.2998329 0.2078749 0.1721364
```

- We find out if there is any diagonal element with value $> \frac{2p}{n}$ as they should be looked at more closely.

```
        12         20         36         38
0.3015417 0.3042241 0.4994692 0.2998329
```

- Hence we will apply other procedures also to confirm whether these points are influential or not.

Hat Matrix Diagonals

- To detect high leverage points, we compute the hat matrix diagonals h_i of the matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$:-

```
       36        20        12        38        30        39
0.4994692 0.3042241 0.3015417 0.2998329 0.2078749 0.1721364
```

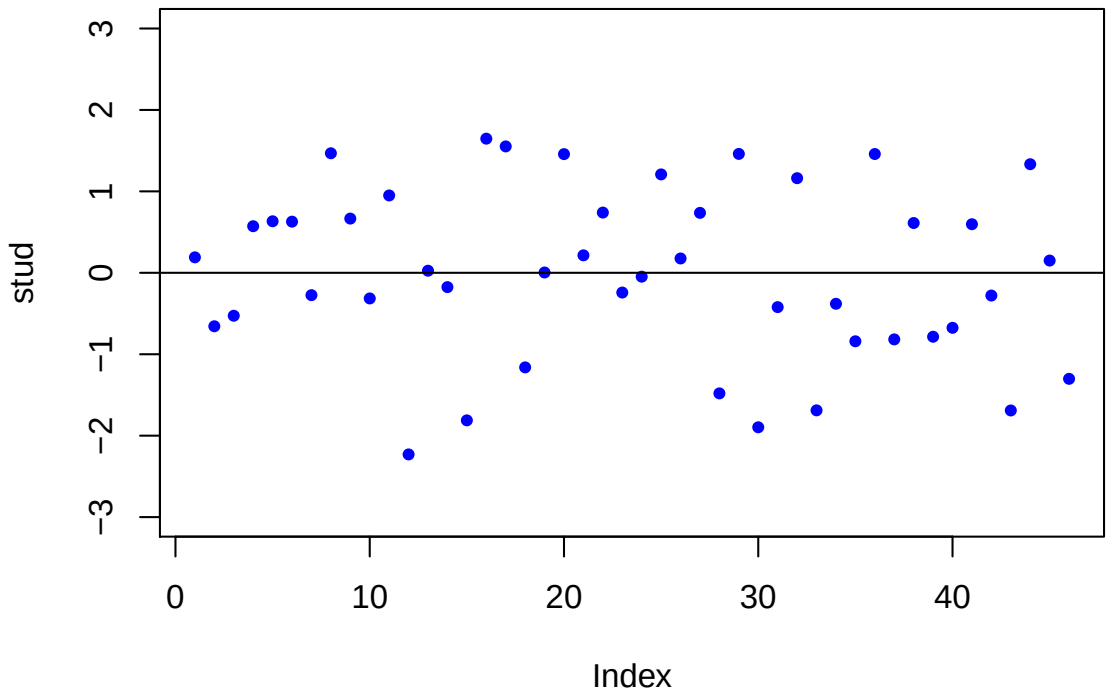
- We find out if there is any diagonal element with value $> \frac{2p}{n}$ as they should be looked at more closely.

```
       12        20        36        38
0.3015417 0.3042241 0.4994692 0.2998329
```

- Hence we will apply other procedures also to confirm whether these points are influential or not.

Externally Studentized Residuals

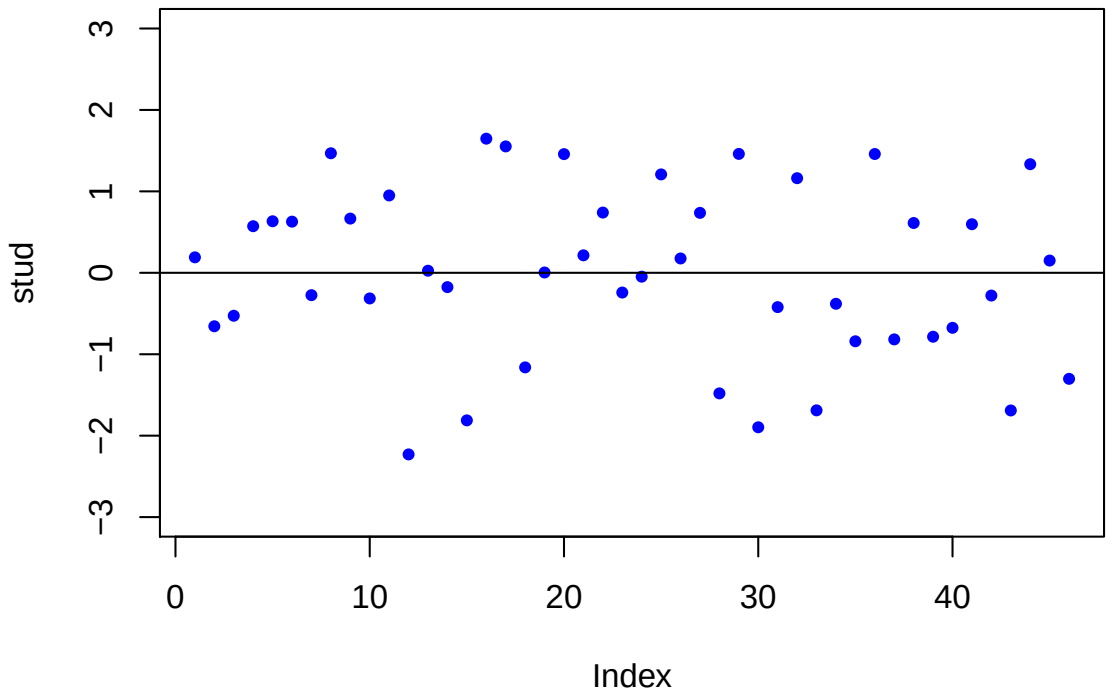
- We plot the externally studentized residuals using the formula $t_i^2 = r_i^2 \left(\frac{n-p-1}{n-p-r_i^2} \right)$:-



- If the assumptions are correct i.e. $\epsilon \sim N_n\left(\mathbf{0}, \sigma^2 I\right)$ then we should get that $t_i \sim t_{n-p-1}$.

Externally Studentized Residuals

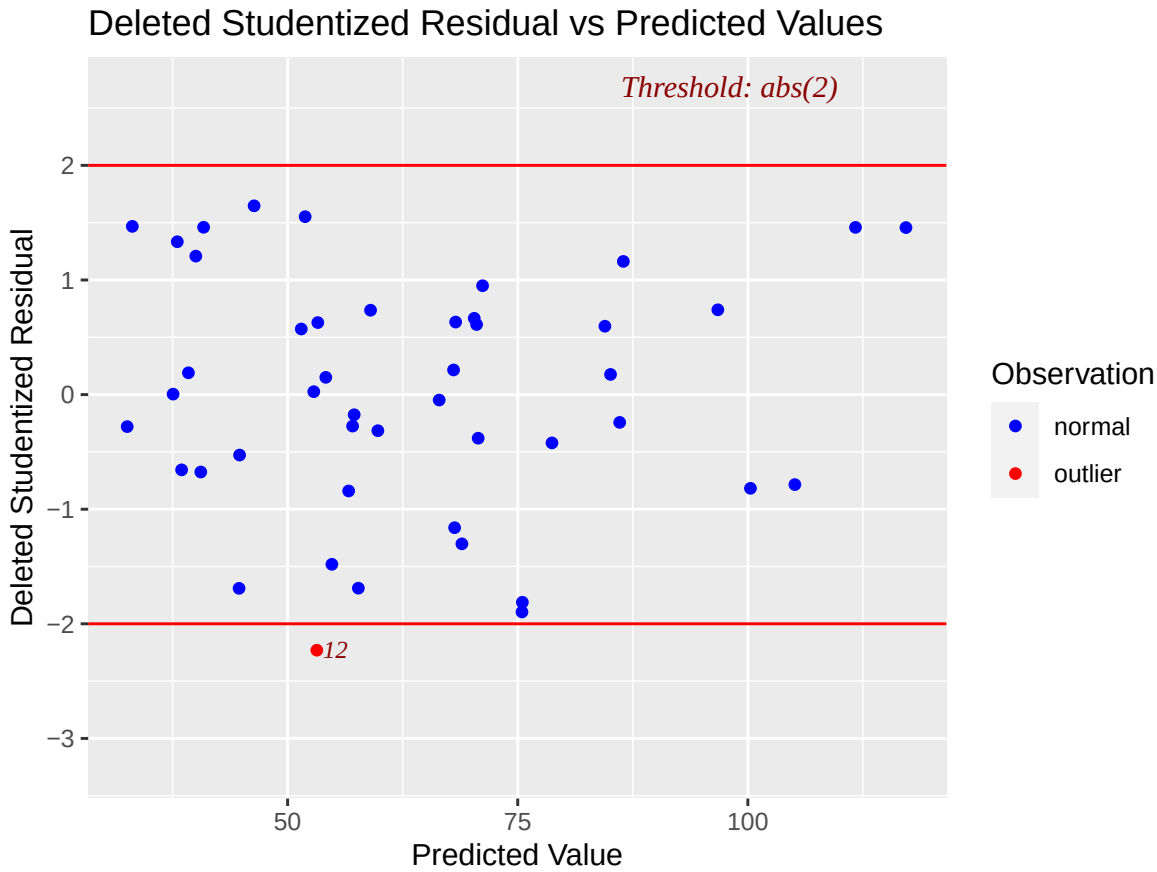
- We plot the externally studentized residuals using the formula $t_i^2 = r_i^2 \left(\frac{n-p-1}{n-p-r_i^2} \right)$:-



- If the assumptions are correct i.e. $\boldsymbol{\epsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$ then we should get that $t_i \sim t_{n-p-1}$.

Studentized Residuals

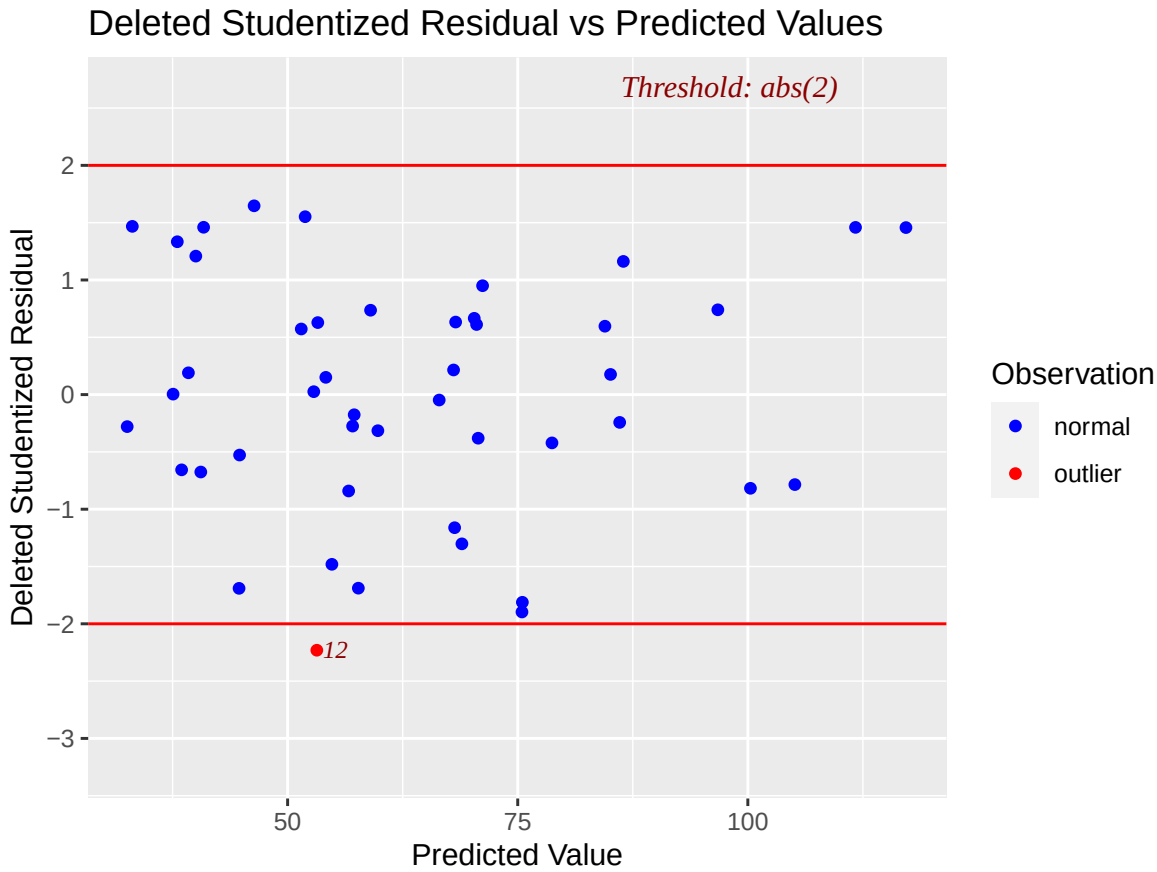
- Hence the significant externally studentized residuals will have values $|t_i| > t_{n-p-1;\frac{\alpha}{2}} \iff t_i^2 > F_{n-p-1;\alpha}$:-



- We can see from the plot that one residual is significant hence we treat that as an outlier.

Studentized Residuals

- Hence the significant externally studentized residuals will have values $|t_i| > t_{n-p-1;\frac{\alpha}{2}} \iff t_i^2 > F_{n-p-1;\alpha}$:-



- We can see from the plot that one residual is significant hence we treat that as an outlier.

$DFBETAS$

- After outliers , we check for presence of high leavarage points, which can be detected using DFBETAS measure for different parameters $DFBETAS_{ij} = \frac{\widehat{\beta}_j - \widehat{\beta}(i)_j}{S(i)\sqrt{\sum_i c^2_{j+1,i}}}$ where

$$\boldsymbol{C} = ((c_{ij})) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X} \text{ :-}$$

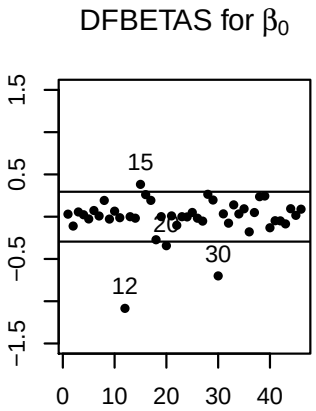
- We will consider the points for which $|DFBETAS_{ij}| > \frac{2}{\sqrt{n}}$. In the next slide we plot the values for all the 5 coefficients $\beta_i, i = 0, ..., 4$.

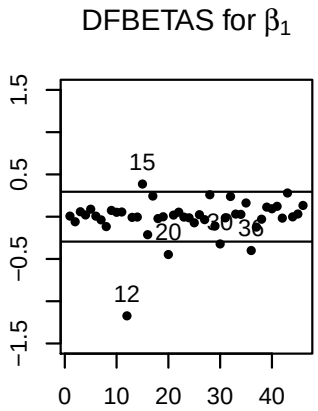
$DFBETAS$

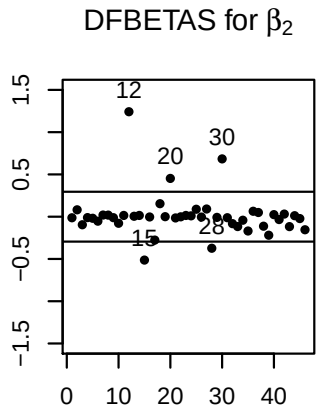
- After outliers , we check for presence of high leavarage points, which can be detected using DFBETAS measure for different parameters $DFBETAS_{ij} = \frac{\widehat{\beta}_j - \widehat{\beta}(i)_j}{S(i)\sqrt{\sum_i c^2_{j+1,i}}}$ where

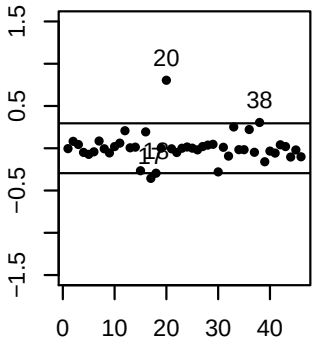
  $\boldsymbol{C} = ((c_{ij})) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}$:-
- We will consider the points for which $|DFBETAS_{ij}| > \frac{2}{\sqrt{n}}$. In the next slide we plot the values for all the 5 coefficients $\beta_i, i = 0, ..., 4$.

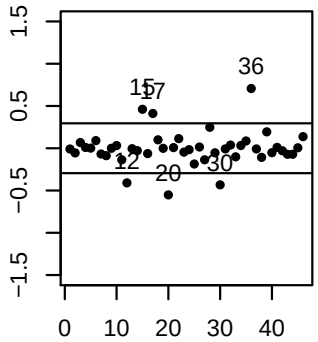
DFBETAS


DFBETAS for β_0
1.5
0.5
−0.5
−1.5
0
10
20
30
40
15
20
30
12


DFBETAS for β_1
1.5
0.5
−0.5
−1.5
0
10
20
30
40
15
20
30
36
12


DFBETAS for β_2
1.5
0.5
−0.5
−1.5
0
10
20
30
40
12
30
20
15
28

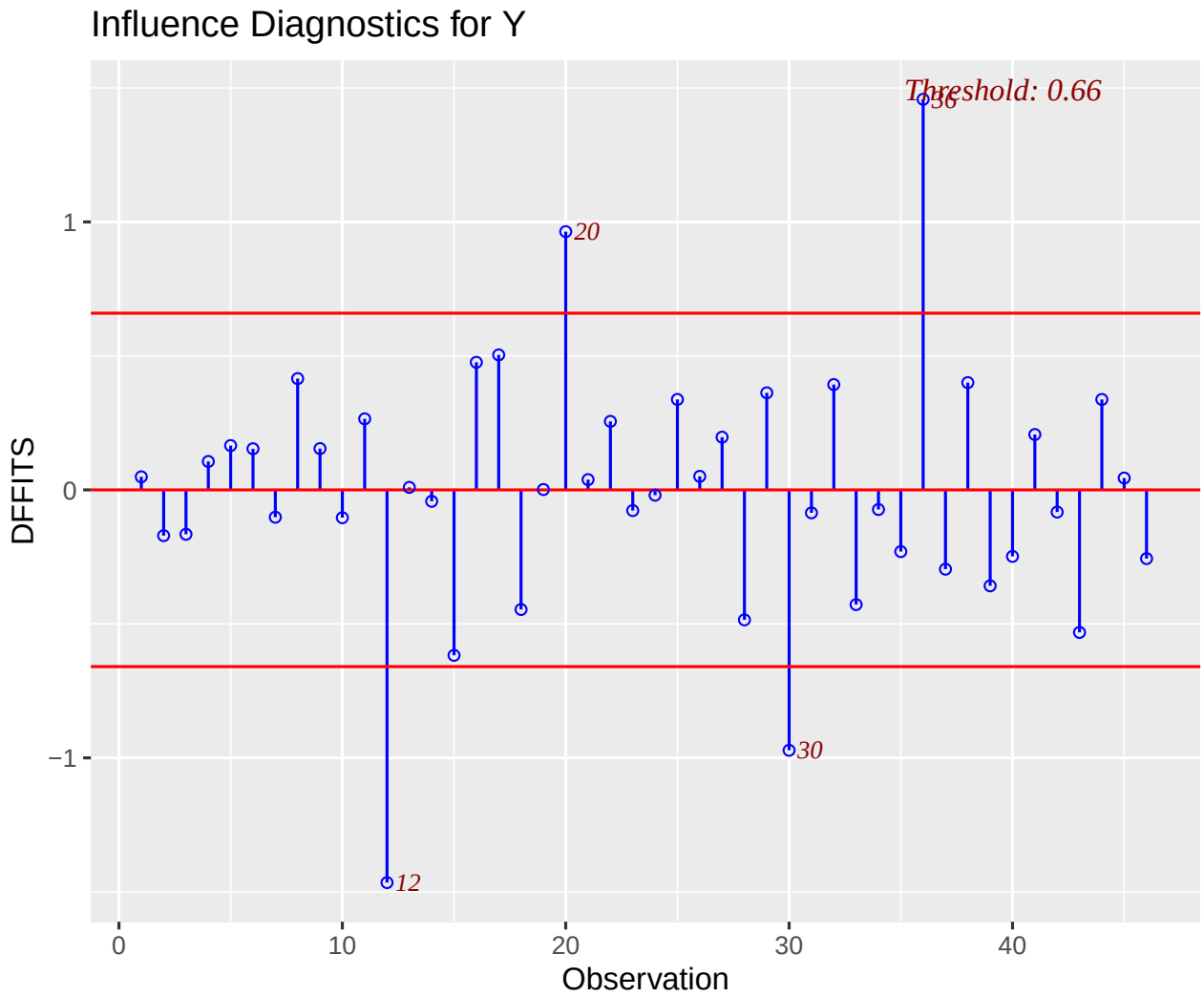

DFBETAS for β_3
1.5
0.5
−0.5
−1.5
0
10
20
30
40
20
38


DFBETAS for β_4
1.5
0.5
−0.5
−1.5
0
10
20
30
40
36
20
30

$DFFITS$

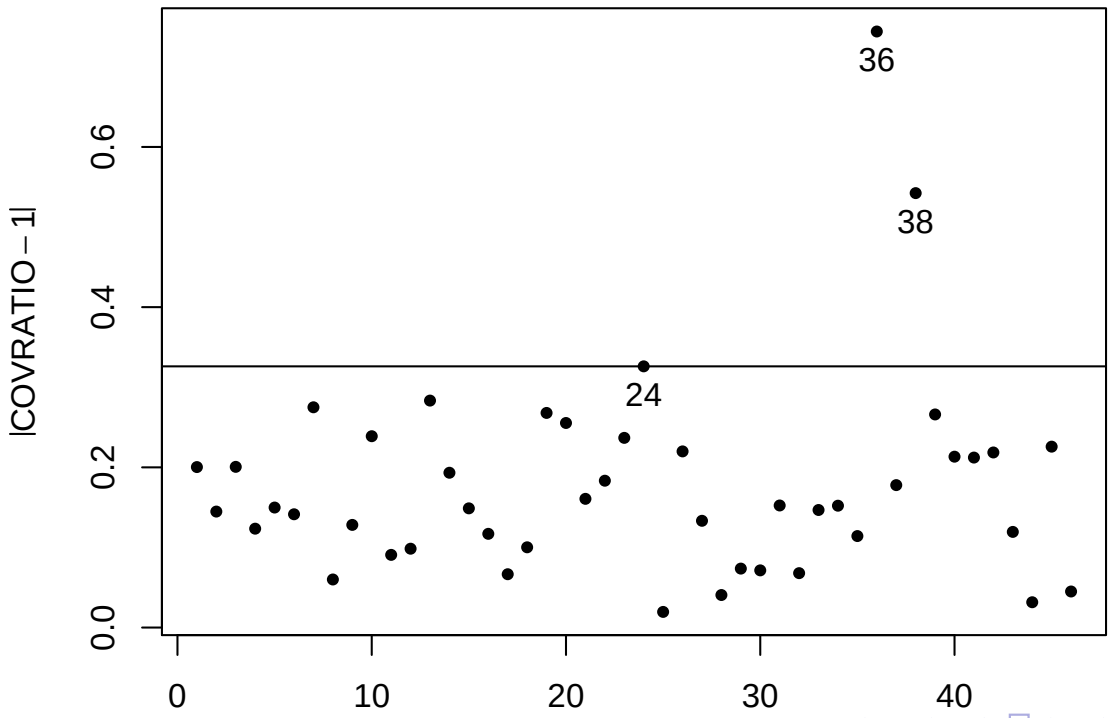
- To notice the change in fitted values, we plot the DFFITS values for all the points where $DFFITS_i = t_i \left(\frac{h_i}{1-h_i}\right)^{1/2}$ and we will check for the points for which $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$.

DFFITS


Influence Diagnostics for Y
Threshold: 0.66
36
20
12
30
DFFITS
1
0
−1
0
10
20
30
40
Observation

$COVRATIO$

- We also plot the COVRATIO values which are defined as $COVRATIO_i = \left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p}\right)^{-p} (1-h_i)^{-1}$ and we consider the points to have high fluence for which $|COVRATIO - 1| > \frac{3p}{n}$.



Cook's D

- Lastly, we calculate the Cook's Distance $D_i = \frac{\left(\widehat{\beta}(i)-\widehat{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X} \left(\widehat{\beta}(i)-\widehat{\beta}\right)}{pS^2} = r_i^2 \frac{h_i}{p(1-h_i)}$ for all the n points.
- We flag the points as suspicious for which $D_i > \frac{4}{n}$ here $n = 46, p = 5$. Whose value equals to 0.087.
- We plot the values and see if such suspicious points exists or not.

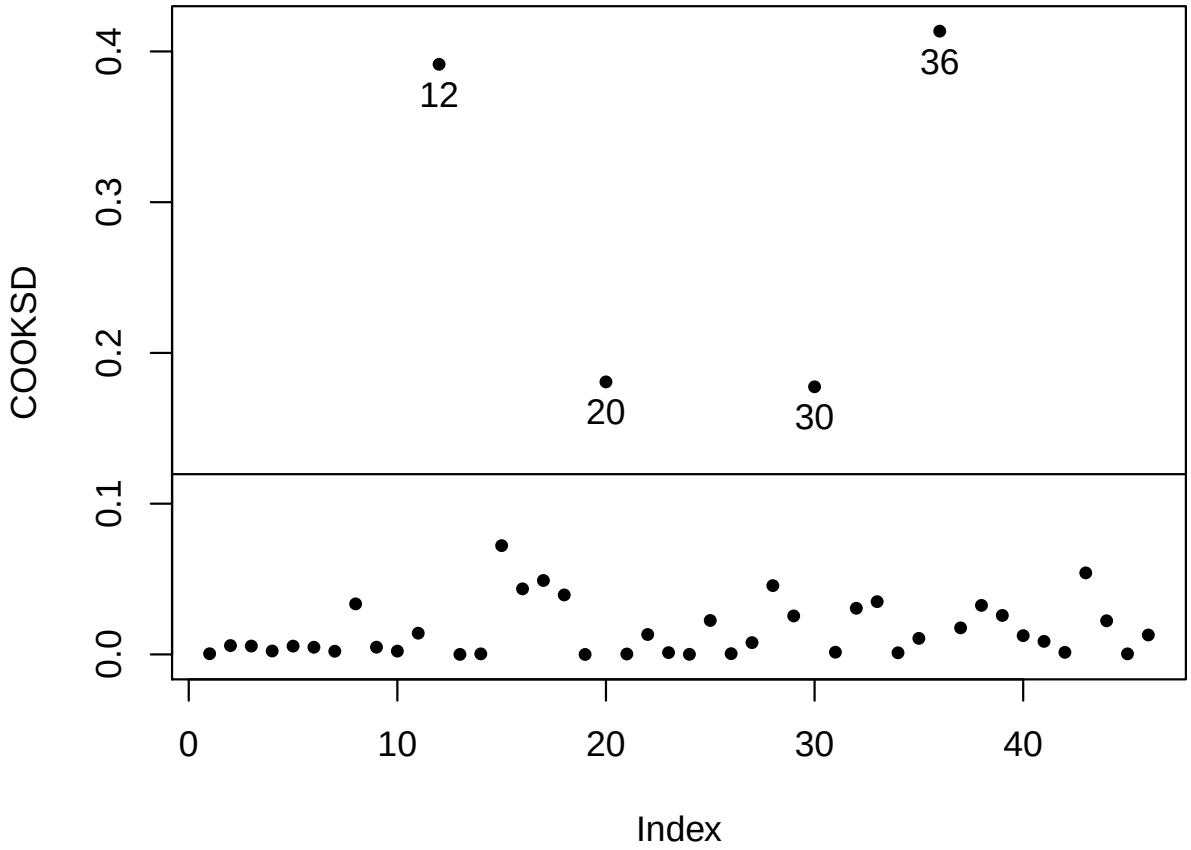
Cook's D

- Lastly, we calculate the Cook's Distance $D_i = \frac{\left(\widehat{\beta}(i)-\widehat{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X} \left(\widehat{\beta}(i)-\widehat{\beta}\right)}{pS^2} = r_i^2 \frac{h_i}{p(1-h_i)}$ for all the n points.
- We flag the points as suspicious for which $D_i > \frac{4}{n}$ here $n = 46, p = 5$. Whose value equals to 0.087.
- We plot the values and see if such suspicious points exists or not.

Cook's D

- Lastly, we calculate the Cook's Distance $D_i = \frac{\left(\widehat{\beta}(i)-\widehat{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X} \left(\widehat{\beta}(i)-\widehat{\beta}\right)}{pS^2} = r_i^2 \frac{h_i}{p(1-h_i)}$ for all the n points.
- We flag the points as suspicious for which $D_i > \frac{4}{n}$ here $n = 46, p = 5$. Whose value equals to 0.087.
- We plot the values and see if such suspicious points exists or not.

Cook's D



- We can clearly notice that the points 12,20,30,36 have significant values of D_i. So we will investigate them further.

Conclusion

- From all the diagnostics performed for finding influential observations, we can make the following table of our findings :-

- Hence, from the table, we conclude the points $12, 20, 30, 36$ to be influential points and we we will later remove them from the model and see the changes occuring in all aspects of the fitted linear models.

Conclusion

- From all the diagnostics performed for finding influential observations, we can make the following table of our findings :-

Diagnostic Measures	Points Detected
h_i	$12, 20, 36, 38$
t_i	12
$DFBETAS$	$12, 15, 17, 18, 20, 28, 30, 36$
$DFFITS$	$12, 20, 30, 36$
$COVRATIO$	$24, 36, 38$
$Cook's\ D$	$12, 20, 30, 36$

- Hence, from the table, we conclude the points $12, 20, 30, 36$ to be influential points and we we will later remove them from the model and see the changes occuring in all aspects of the fitted linear models.

Conclusion

- From all the diagnostics performed for finding influential observations, we can make the following table of our findings :-

Diagnostic Measures	Points Detected
h_i	$12, 20, 36, 38$
t_i	12
$DFBETAS$	$12, 15, 17, 18, 20, 28, 30, 36$
$DFFITS$	$12, 20, 30, 36$
$COVRATIO$	$24, 36, 38$
$Cook's\ D$	$12, 20, 30, 36$

- Hence, from the table, we conclude the points $12, 20, 30, 36$ to be influential points and we we will later remove them from the model and see the changes occuring in all aspects of the fitted linear models.

Removing Influential Points

- We remove the influential points and then again fit a linear model with all the covariates and write the summary output of the fitted model here :-

```
Call:
lm(formula = Y ~ A1 + A2 + A3 + A4, data = X[-c(12, 20, 30, 36),
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-16.6468  -5.0683  -0.3998   6.1885  16.7016

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.2021    11.7358   1.721  0.09353 .
A1            0.8783     0.2582   3.401  0.00162 **
A2           -0.7104     0.6205  -1.145  0.25960
A3            1.2890     0.3960   3.255  0.00243 **
A4            0.1489     0.1312   1.134  0.26392
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.672 on 37 degrees of freedom
Multiple R-squared:  0.8236,Adjusted R-squared:  0.8045
F-statistic: 43.18 on 4 and 37 DF,  p-value: 1.867e-13
```

Removing Influential Points

- This model has increased value of $R^2_{adj} = 0.8045$.

Improvements in the fitted model

- From the output,we can see that the R^2_{adj} value has increased from that of the full model which was $= 0.7954$.
- Also we can see that the model indicates the estimates of the intercept term $\left(\widehat{\beta}_0\right)$, variables A1 & A3 $\left(\widehat{\beta}_1, \widehat{\beta}_3\right)$ to be signifcant.
- Whether covariates A2 & A4 are signifcant or not, will be verified later.

Improvements in the fitted model

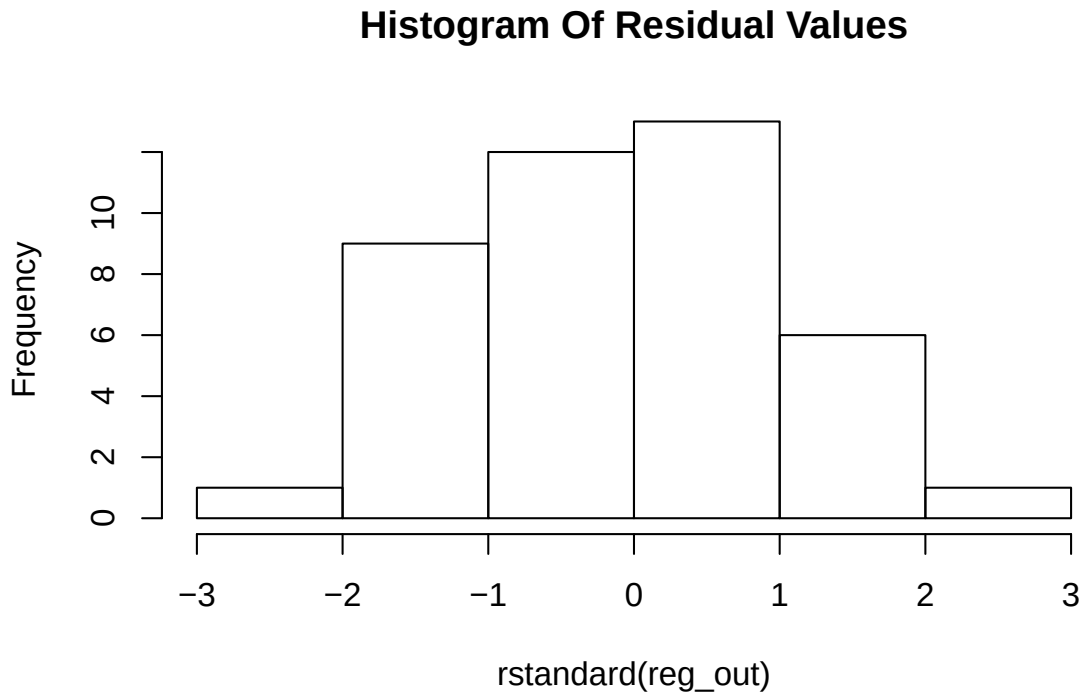
- From the output,we can see that the R^2_{adj} value has increased from that of the full model which was $= 0.7954$.
- Also we can see that the model indicates the estimates of the intercept term $\left(\widehat{\beta}_0\right)$, variables A1 & A3 $\left(\widehat{\beta}_1, \widehat{\beta}_3\right)$to be signifcant.
- Whether covariates A2 & A4 are signifcant or not, will be verified later.

Improvements in the fitted model

- From the output,we can see that the R^2_{adj} value has increased from that of the full model which was $= 0.7954$.
- Also we can see that the model indicates the estimates of the intercept term $\left(\widehat{\beta}_0\right)$, variables A1 & A3 $\left(\widehat{\beta}_1, \widehat{\beta}_3\right)$to be signifcant.
- Whether covariates A2 & A4 are signifcant or not, will be verified later.

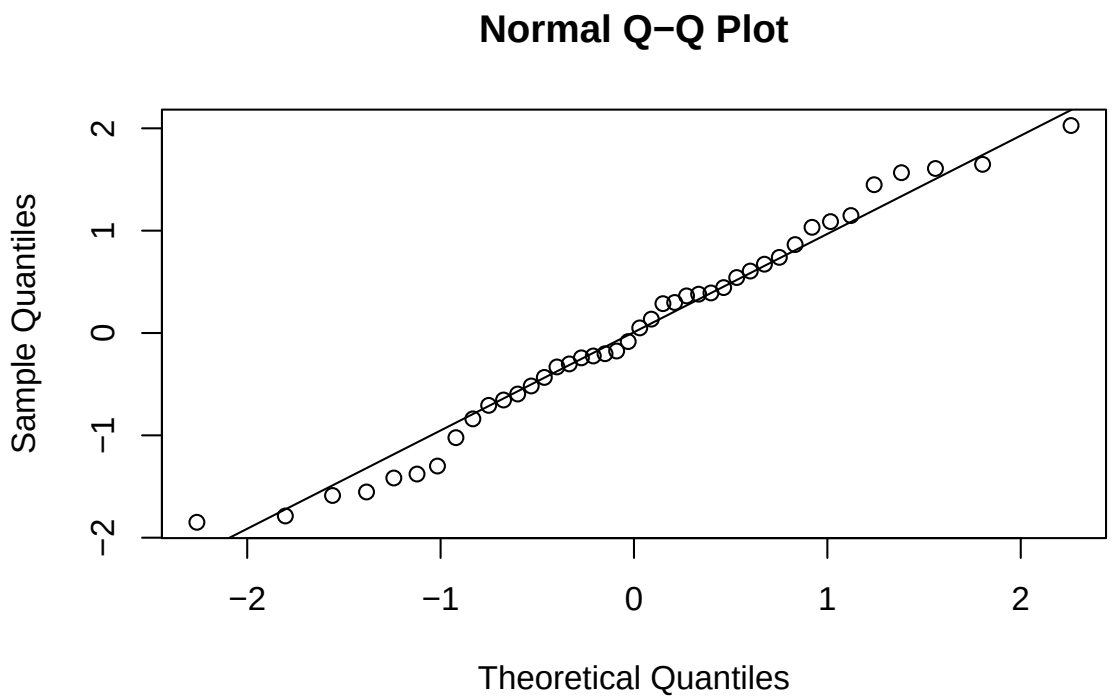
Histogram of Residuals

- We also plot the histogram of the residual values and can notice it's symmetric about 0 and seems to be normally distributed :-



QQ-plot

- For checking the assumptions for the residuals we again make the quantile-quantile plot of the standardized residuals :-



Checking Other Assumptions

- We again perform both the Shapiro-Wilk test and Durbin-Watson test for checking normality and presence of correlation between the residuals, respectively. We write down the observations in the following table :-

- Hence we can see considerable improvement in the normality assumptions of the residuals whereas the uncorrelated & homoscedasticity assumptions are more or less remains equally acceptable.
- Hence, the model can be considered to be better than the previous model as a result of removing the influential points.

Checking Other Assumptions

- We again perform both the Shapiro-Wilk test and Durbin-Watson test for checking normality and presence of correlation between the residuals, respectively. We write down the observations in the following table :-

Tests	Model with influential points	Model without influential points
Shapiro-Wilk	0.1133	0.8429
Durbin-Watson	0.9734	0.6943
Breusch-Pagan	0.6105	0.5742
Breusch-Godfrey	0.2033	0.7973

- Hence we can see considerable improvement in the normality assumptions of the residuals whereas the uncorrelated & homoscedasticity assumptions are more or less remains equally acceptable.
- Hence, the model can be considered to be better than the previous model as a result of removing the influential points.

Checking Other Assumptions

- We again perform both the Shapiro-Wilk test and Durbin-Watson test for checking normality and presence of correlation between the residuals, respectively. We write down the observations in the following table :-

Tests	Model with influential points	Model without influential points
Shapiro-Wilk	0.1133	0.8429
Durbin-Watson	0.9734	0.6943
Breusch-Pagan	0.6105	0.5742
Breusch-Godfrey	0.2033	0.7973

- Hence we can see considerable improvement in the normality assumptions of the residuals whereas the uncorrelated & homoscedasticity assumptions are more or less remains equally acceptable.
- Hence, the model can be considered to be better than the previous model as a result of removing the influential points.

Checking Other Assumptions

- We again perform both the Shapiro-Wilk test and Durbin-Watson test for checking normality and presence of correlation between the residuals, respectively. We write down the observations in the following table :-

Tests	Model with influential points	Model without influential points
Shapiro-Wilk	0.1133	0.8429
Durbin-Watson	0.9734	0.6943
Breusch-Pagan	0.6105	0.5742
Breusch-Godfrey	0.2033	0.7973

- Hence we can see considerable improvement in the normality assumptions of the residuals whereas the uncorrelated & homoscedasticity assumptions are more or less remains equally acceptable.
- Hence, the model can be considered to be better than the previous model as a result of removing the influential points.

Multicollinearity

- Next we consider the problem of multicollinearity that may be present in our dataset as suspected from the pairwise scatterplots.
- We calculate the condition number for the scaled and centred model matrix $\boldsymbol{X}^*$ which is $\kappa\left(\boldsymbol{X}^*\right) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ where λ_i's are the eigenvalues of $\boldsymbol{X}^{*T}\boldsymbol{X}^* = \boldsymbol{R}_{xx}$. We calculate $\kappa\left(\boldsymbol{X}^*\right)$ using R :-

```
[1] 8.624355
```

- Hence, the square of condition number $\kappa^2\left(\boldsymbol{X}^*\right) \approx 74.379$ is an upper bound for the VIFs which is quite large !

Multicollinearity

- Next we consider the problem of multicollinearity that may be present in our dataset as suspected from the pairwise scatterplots.
- We calculate the condition number for the scaled and centred model matrix $\boldsymbol{X}^*$ which is $\kappa\left(\boldsymbol{X}^*\right) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ where λ_i's are the eigenvalues of $\boldsymbol{X}^{*T}\boldsymbol{X}^* = \boldsymbol{R}_{xx}$. We calculate $\kappa\left(\boldsymbol{X}^*\right)$ using R :-

```
[1] 8.624355
```

- Hence, the square of condition number $\kappa^2\left(\boldsymbol{X}^*\right) \approx 74.379$ is an upper bound for the VIFs which is quite large !

Multicollinearity

- Next we consider the problem of multicollinearity that may be present in our dataset as suspected from the pairwise scatterplots.
- We calculate the condition number for the scaled and centred model matrix $\boldsymbol{X}^*$ which is $\kappa\left(\boldsymbol{X}^*\right) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ where λ_i's are the eigenvalues of $\boldsymbol{X}^{*T}\boldsymbol{X}^* = \boldsymbol{R}_{xx}$. We calculate $\kappa\left(\boldsymbol{X}^*\right)$ using R :-

```
[1] 8.624355
```

- Hence, the square of condition number $\kappa^2\left(\boldsymbol{X}^*\right) \approx 74.379$ is an upper bound for the VIFs which is quite large !

VIF

- Now, to determine whether some covariate with corresponding column $\boldsymbol{x}_j$, can be predicted accurately using other covariates or not, we compute the variance inflation factors $\text{VIF}_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination of the regression of $\boldsymbol{x}^{*(j)}$ on the columns of $\boldsymbol{X}^{*(j)}$.
- We calculate the VIF_j values for $j = 1, 2, 3, 4$ in R :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- For the variables A1 and A2, we can see that the VIF values are greater than 5 and even close to 10! So we can interpret this as "the standard error of $\widehat{\beta_1}$ and $\widehat{\beta_2}$ would be $\sqrt{9.26} \approx 3.043$ and $\sqrt{9.361} \approx 3.059$ times more (respectively) than it would have been without the presence of collinearity".

VIF

- Now, to determine whether some covariate with corresponding column $\boldsymbol{x}_j$, can be predicted accurately using other covariates or not, we compute the variance inflation factors $\text{VIF}_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination of the regression of $\boldsymbol{x}^{*(j)}$ on the columns of $\boldsymbol{X}^{*(j)}$.
- We calculate the VIF_j values for $j = 1, 2, 3, 4$ in R :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- For the variables A1 and A2, we can see that the VIF values are greater than 5 and even close to 10! So we can interpret this as "the standard error of $\widehat{\beta_1}$ and $\widehat{\beta_2}$ would be $\sqrt{9.26} \approx 3.043$ and $\sqrt{9.361} \approx 3.059$ times more (respectively) than it would have been without the presence of collinearity".

VIF

- Now, to determine whether some covariate with corresponding column $\boldsymbol{x}_j$, can be predicted accurately using other covariates or not, we compute the variance inflation factors $\text{VIF}_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination of the regression of $\boldsymbol{x}^{*(j)}$ on the columns of $\boldsymbol{X}^{*(j)}$.
- We calculate the VIF_j values for $j = 1, 2, 3, 4$ in R :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- For the variables A1 and A2, we can see that the VIF values are greater than 5 and even close to 10! So we can interpret this as "the standard error of $\widehat{\beta_1}$ and $\widehat{\beta_2}$ would be $\sqrt{9.26} \approx 3.043$ and $\sqrt{9.361} \approx 3.059$ times more (respectively) than it would have been without the presence of collinearity".

Effect of Influential Points on Collinearity

- This is a very interesting observation that we have made in the dataset.
- If we remove the influential points and then calculate the VIF values, we get :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- But if we do the same without removing those points, we get :-

```
      A1       A2       A3       A4
5.910333 6.748416 3.080737 3.488172
```

- So, as we can see the VIF values increased after removal of the influential points.
- This is intuitive from the fact that actual linear dependence between the covariates was being slightly nullified by the presence of such influential points.

Effect of Influential Points on Collinearity

- This is a very interesting observation that we have made in the dataset.
- If we remove the influential points and then calculate the VIF values, we get :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- But if we do the same without removing those points, we get :-

```
      A1       A2       A3       A4
5.910333 6.748416 3.080737 3.488172
```

- So, as we can see the VIF values increased after removal of the influential points.
- This is intuitive from the fact that actual linear dependence between the covariates was being slightly nullified by the presence of such influential points.

Effect of Influential Points on Collinearity

- This is a very interesting observation that we have made in the dataset.
- If we remove the influential points and then calculate the VIF values, we get :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- But if we do the same without removing those points, we get :-

```
      A1       A2       A3       A4
5.910333 6.748416 3.080737 3.488172
```

- So, as we can see the VIF values increased after removal of the influential points.
- This is intuitive from the fact that actual linear dependence between the covariates was being slightly nullified by the presence of such influential points.

Effect of Influential Points on Collinearity

- This is a very interesting observation that we have made in the dataset.
- If we remove the influential points and then calculate the VIF values, we get :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- But if we do the same without removing those points, we get :-

```
      A1       A2       A3       A4
5.910333 6.748416 3.080737 3.488172
```

- So, as we can see the VIF values increased after removal of the influential points.
- This is intuitive from the fact that actual linear dependence between the covariates was being slightly nullified by the presence of such influential points.

Effect of Influential Points on Collinearity

- This is a very interesting observation that we have made in the dataset.
- If we remove the influential points and then calculate the VIF values, we get :-

```
      A1       A2       A3       A4
9.261059 9.360894 2.857795 2.751179
```

- But if we do the same without removing those points, we get :-

```
      A1       A2       A3       A4
5.910333 6.748416 3.080737 3.488172
```

- So, as we can see the VIF values increased after removal of the influential points.
- This is intuitive from the fact that actual linear dependence between the covariates was being slightly nullified by the presence of such influential points.

Demonstrating Effect of Collinearity

- To demonstrate how collinearity can affect the estimates badly, we delibarately introduce some random noise in the response observations ($\boldsymbol{\delta} \sim N(0, 1)$) and then fit a linear model and see the changes in the estimate.

```
 (Intercept)           A1           A2           A3           A4
-13.96310010   0.09828590   1.14837707   1.85786103   0.04817018
(Intercept)           A1           A2           A3           A4
   1.304038    2.353402   -2.501280   -4.056390    1.437409
```

- Hence, we can clearly see the how the estimates change a lot for introducing random noise in the response.

Demonstrating Effect of Collinearity

- To demonstrate how collinearity can affect the estimates badly, we delibarately introduce some random noise in the response observations ($\boldsymbol{\delta} \sim N(0, 1)$) and then fit a linear model and see the changes in the estimate.

```
 (Intercept)           A1           A2           A3           A4
-13.96310010   0.09828590   1.14837707   1.85786103   0.04817018
(Intercept)           A1           A2           A3           A4
   1.304038    2.353402   -2.501280   -4.056390    1.437409
```

- Hence, we can clearly see the how the estimates change a lot for introducing random noise in the response.

Dealing with Collinearity

- To deal with the collinearity present in the dataset, we first try to remove one of the correlated covariates “A1” or “A2” and see what improvements are observed in the variation inflation factors :-

- We also check for other assumptions between the models :-

Dealing with Collinearity

- To deal with the collinearity present in the dataset, we first try to remove one of the correlated covariates "A1" or "A2" and see what improvements are observed in the variation inflation factors :-

Model	Condition Number (κ)	R^2_{adj}
$Y = \beta_0 + \beta_1 \mathsf{A1} + \beta_2 \mathsf{A2} + \beta_3 \mathsf{A3} + \beta_4 \mathsf{A4}$	8.512	0.795
$Y = \beta_0 + \beta_2 \mathsf{A2} + \beta_3 \mathsf{A3} + \beta_4 \mathsf{A4}$	5.625	0.755
$Y = \beta_0 + \beta_1 \mathsf{A1} + \beta_3 \mathsf{A3} + \beta_4 \mathsf{A4}$	3.951	0.803

- We also check for other assumptions between the models :-

Dealing with Collinearity

- To deal with the collinearity present in the dataset, we first try to remove one of the correlated covariates "A1" or "A2" and see what improvements are observed in the variation inflation factors :-

Model	Condition Number (κ)	R^2_{adj}
$Y = \beta_0 + \beta_1 \mathsf{A1} + \beta_2 \mathsf{A2} + \beta_3 \mathsf{A3} + \beta_4 \mathsf{A4}$	8.512	0.795
$Y = \beta_0 + \beta_2 \mathsf{A2} + \beta_3 \mathsf{A3} + \beta_4 \mathsf{A4}$	5.625	0.755
$Y = \beta_0 + \beta_1 \mathsf{A1} + \beta_3 \mathsf{A3} + \beta_4 \mathsf{A4}$	3.951	0.803

- We also check for other assumptions between the models :-

Dealing with Collinearity

- To deal with the collinearity present in the dataset, we first try to remove one of the correlated covariates "A1" or "A2" and see what improvements are observed in the variation inflation factors :-

Model	Condition Number (κ)	R^2_{adj}
$Y = \beta_0 + \beta_1 \text{A1} + \beta_2 \text{A2} + \beta_3 \text{A3} + \beta_4 \text{A4}$	8.512	0.795
$Y = \beta_0 + \beta_2 \text{A2} + \beta_3 \text{A3} + \beta_4 \text{A4}$	5.625	0.755
$Y = \beta_0 + \beta_1 \text{A1} + \beta_3 \text{A3} + \beta_4 \text{A4}$	3.951	0.803

- We also check for other assumptions between the models :-

Tests	Full Model	Without "A1"	Without "A2"
Shapiro-Wilk	0.1133	0.563	0.2461
Durbin-Watson	0.9734	0.7376	0.7638
Breusch-Pagan	0.6105	0.3742	0.9423

Dealing with Collinearity

- Hence, the model with covariates "A1", "A3", "A4" seems to be a much better model in terms of both prediction and accuracy of the estimates of β. Also if we calculate the VIF values for the last model, we get them to be considerably small :-

```
      A1       A3       A4
2.360578 2.491333 1.677148
```

- This is also relatable from the fact that initially covariate "A2" had the maximum VIF value and the condition number also decreased significantly due to its removal.

Dealing with Collinearity

- Hence, the model with covariates "A1", "A3", "A4" seems to be a much better model in terms of both prediction and accuracy of the estimates of β. Also if we calculate the VIF values for the last model, we get them to be considerably small :-

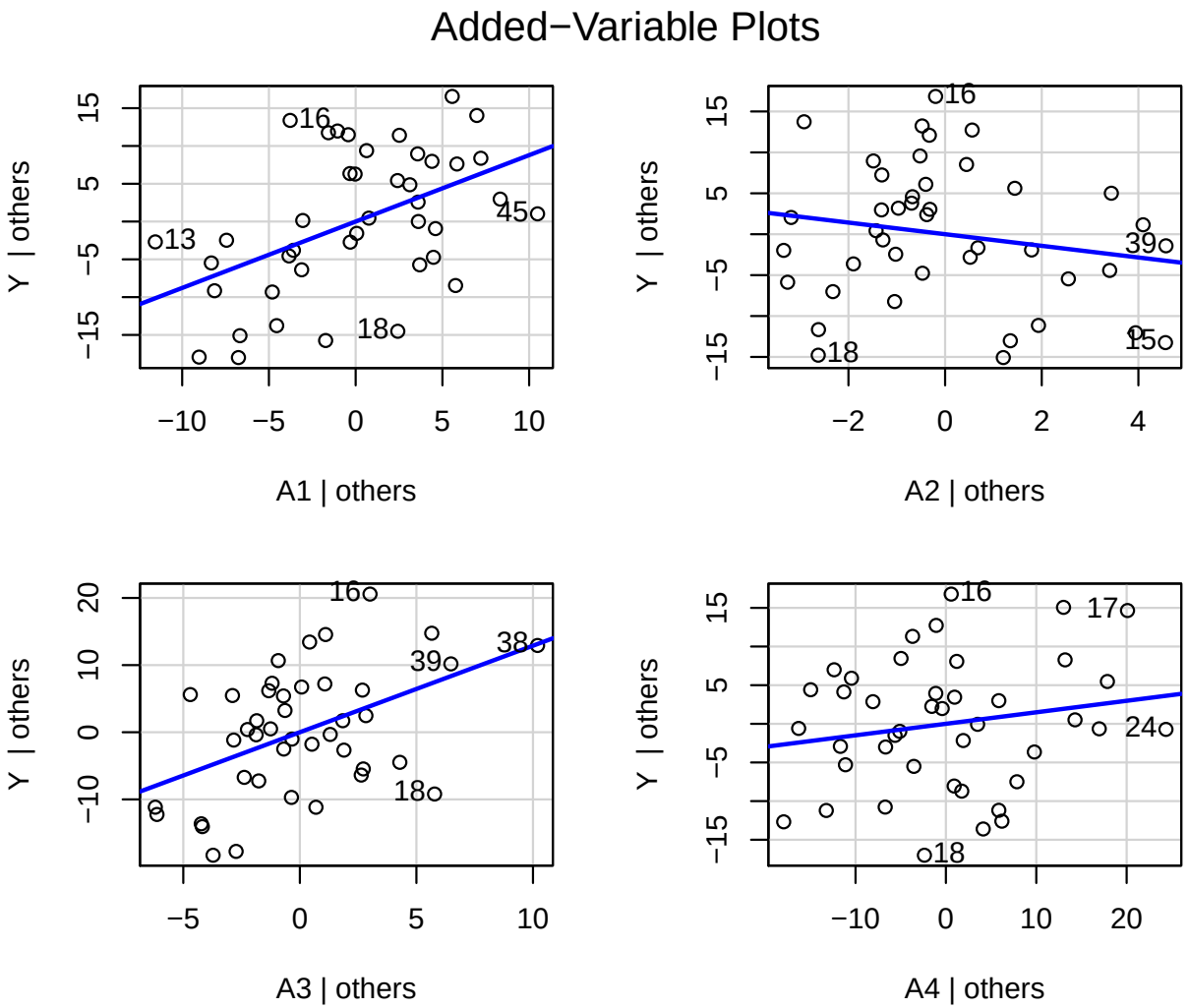
```
      A1       A3       A4
2.360578 2.491333 1.677148
```

- This is also relatable from the fact that initially covariate "A2" had the maximum VIF value and the condition number also decreased significantly due to its removal.

Added Variable Plot

- For better understanding the contribution of a covariate in the regression model, we make a scatter plot of $\boldsymbol{e}^{(i)} = (\boldsymbol{I} - \boldsymbol{P}_i)\,\boldsymbol{Y}$ against $(\boldsymbol{I} - \boldsymbol{P}_i)\,\boldsymbol{x}^{(i)}$ where $\boldsymbol{e}^{(i)}$ are the residuals of the model with variable Ai excluded and $\boldsymbol{x}^{(i)}$ is the column of observations of Ai. This is also called the *added variable plot*.

Added Variable Plot



Conclusion

- From the 4 plots, we can see that slopes of the fitted lines for the added variable plots of A1 & A3 are much more signifcant than other two plots for A2 & A4.
- We can make some important conclusions from here.
- This indicates that once predictor A1, A3, A4 is included, A2 can be excluded from the model for the high collinearity present between them.
- Similar can be said for A4.
- Now, for much better conclusions, we perform further model selection procedures based on several criterias.

Conclusion

- From the 4 plots, we can see that slopes of the fitted lines for the added variable plots of A1 & A3 are much more signifcant than other two plots for A2 & A4.
- We can make some important conclusions from here.
- This indicates that once predictor A1, A3, A4 is included, A2 can be excluded from the model for the high collinearity present between them.
- Similar can be said for A4.
- Now, for much better conclusions, we perform further model selection procedures based on several criterias.

Conclusion

- From the 4 plots, we can see that slopes of the fitted lines for the added variable plots of A1 & A3 are much more signifcant than other two plots for A2 & A4.
- We can make some important conclusions from here.
- This indicates that once predictor A1, A3, A4 is included, A2 can be excluded from the model for the high collinearity present between them.
- Similar can be said for A4.
- Now, for much better conclusions, we perform further model selection procedures based on several criterias.

Conclusion

- From the 4 plots, we can see that slopes of the fitted lines for the added variable plots of A1 & A3 are much more signifcant than other two plots for A2 & A4.
- We can make some important conclusions from here.
- This indicates that once predictor A1, A3, A4 is included, A2 can be excluded from the model for the high collinearity present between them.
- Similar can be said for A4.
- Now, for much better conclusions, we perform further model selection procedures based on several criterias.

Conclusion

- From the 4 plots, we can see that slopes of the fitted lines for the added variable plots of A1 & A3 are much more signifcant than other two plots for A2 & A4.
- We can make some important conclusions from here.
- This indicates that once predictor A1, A3, A4 is included, A2 can be excluded from the model for the high collinearity present between them.
- Similar can be said for A4.
- Now, for much better conclusions, we perform further model selection procedures based on several criterias.

Stepwise Selection

- We perform the stepwise selection algorithm which performs a forward selection (FS) followed by a backward elimination (BE) using the AIC criterion and get to an optimum model.
- We get the following sequence of models in the selection procedure :-

Stepwise Selection

- We perform the stepwise selection algorithm which performs a forward selection (FS) followed by a backward elimination (BE) using the AIC criterion and get to an optimum model.
- We get the following sequence of models in the selection procedure :-

Stepwise Selection

- We perform the stepwise selection algorithm which performs a forward selection (FS) followed by a backward elimination (BE) using the AIC criterion and get to an optimum model.
- We get the following sequence of models in the selection procedure :-

Model	AIC Value
$Y = \beta_0$	372.1803
$Y = \beta_0 + \beta_1 \text{A1}$	319.7608
$Y = \beta_0 + \beta_1 \text{A1} + \beta_3 \text{A3}$	305.0896

Stepwise Selection

- We give the final model as an output we get in R :-

```
                              Stepwise Summary
-------------------------------------------------------------------------------
Variable     Method       AIC        RSS        Sum Sq       R-Sq      Adj. R-Sq
-------------------------------------------------------------------------------
A1          addition    319.761    4316.877    11455.113    0.72629      0.71945
A3          addition    305.090    2902.575    12869.415    0.81597      0.80653
-------------------------------------------------------------------------------
```

- Hence, this method gives the model containing covariates “A1”, “A3” as the optimum one.

Stepwise Selection

- We give the final model as an output we get in R :-

```
                              Stepwise Summary
-------------------------------------------------------------------------------
Variable     Method       AIC        RSS        Sum Sq       R-Sq      Adj. R-Sq
-------------------------------------------------------------------------------
A1          addition    319.761    4316.877    11455.113    0.72629      0.71945
A3          addition    305.090    2902.575    12869.415    0.81597      0.80653
-------------------------------------------------------------------------------
```

- Hence, this method gives the model containing covariates “A1”, “A3” as the optimum one.

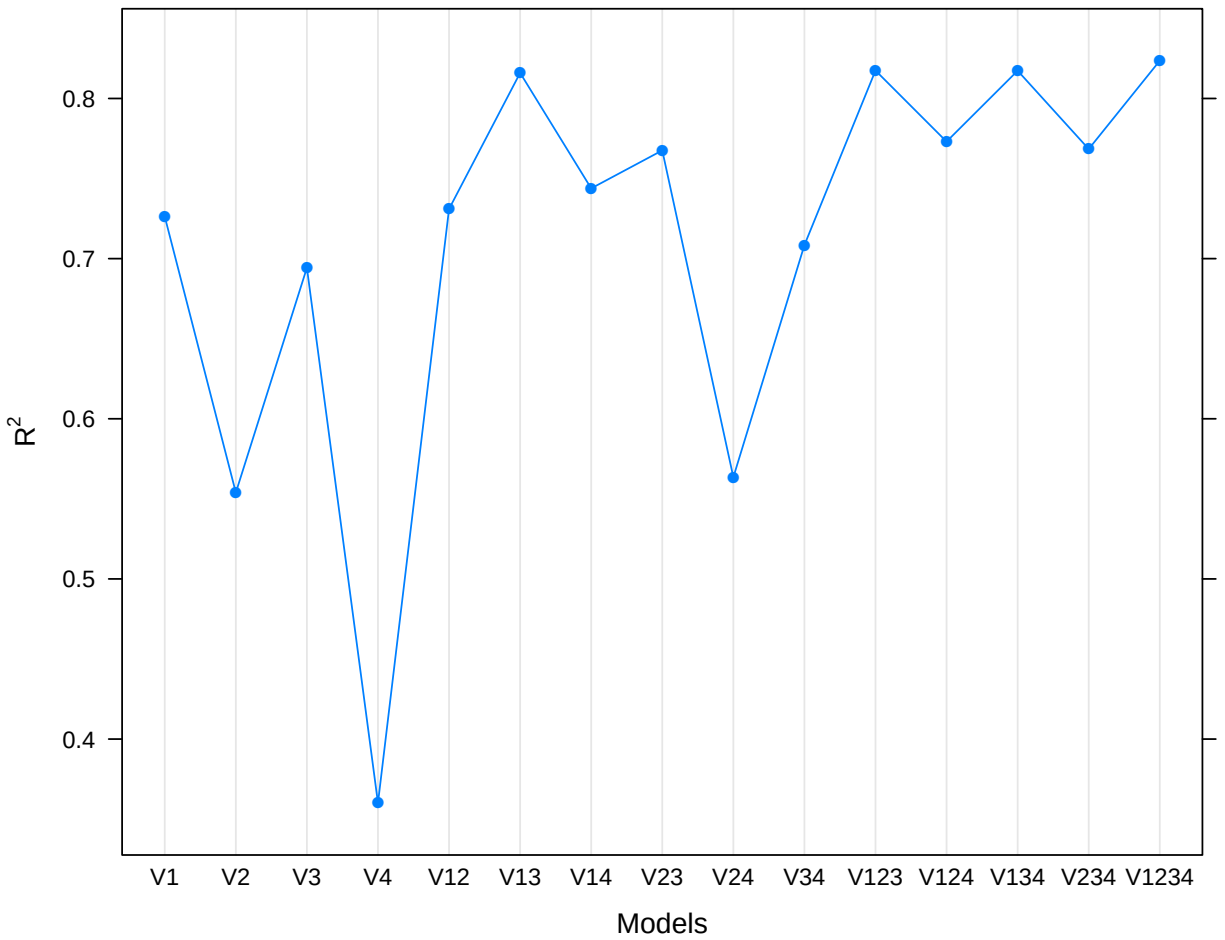
Best Subset Selection

- We use different criterions for chosing optimal model among all the $2^4 - 1 = 15$ possible linear models and plot the diagrams for all of them one by one for comparison :-
- We name each of the models as "Vi" , "Vij" , "Vijk" etc where i, j, k denotes the variables included in the model.

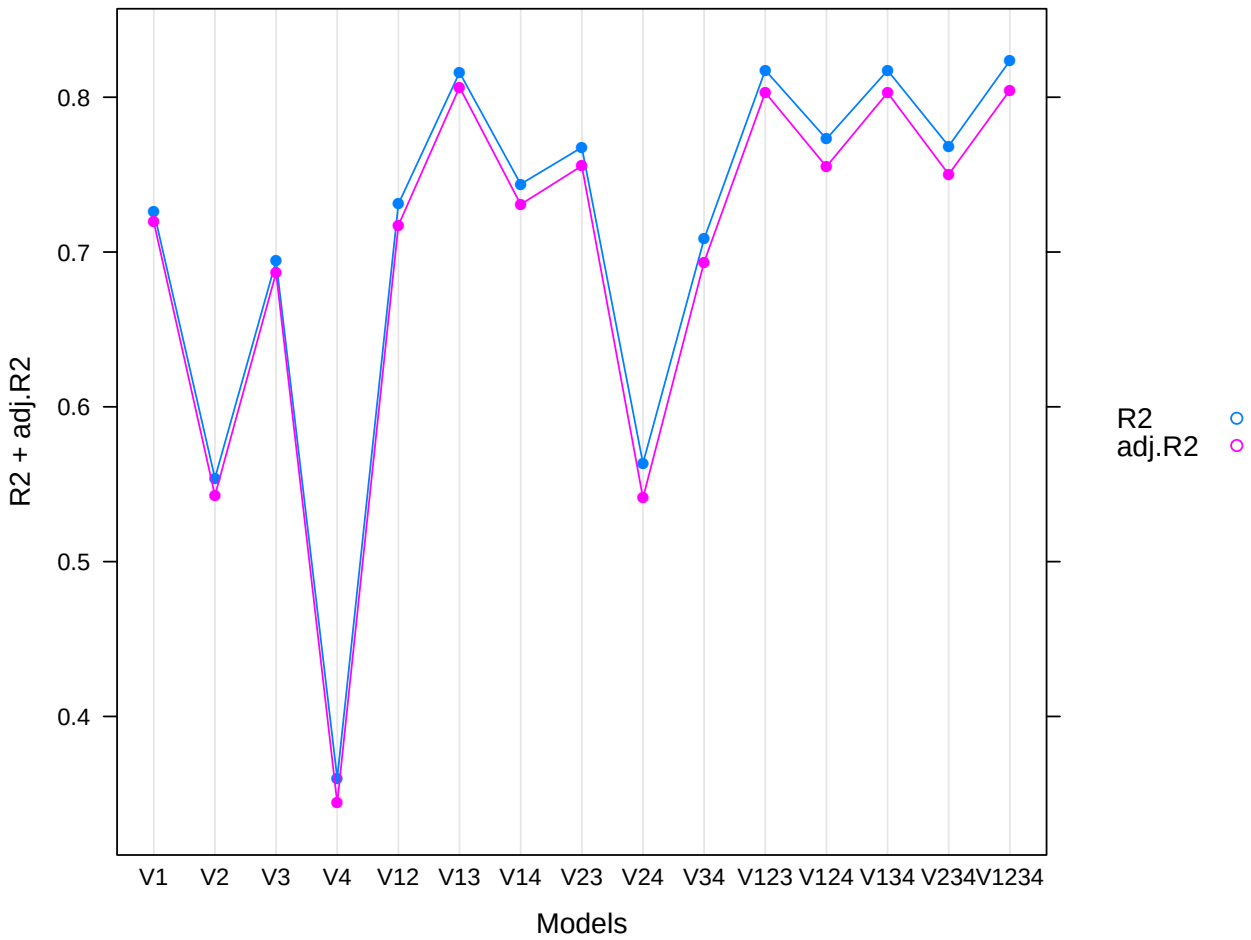
Best Subset Selection

- We use different criterions for chosing optimal model among all the $2^4 - 1 = 15$ possible linear models and plot the diagrams for all of them one by one for comparison :-
- We name each of the models as “Vi” , “Vij” , “Vijk” etc where i, j, k denotes the variables included in the model.

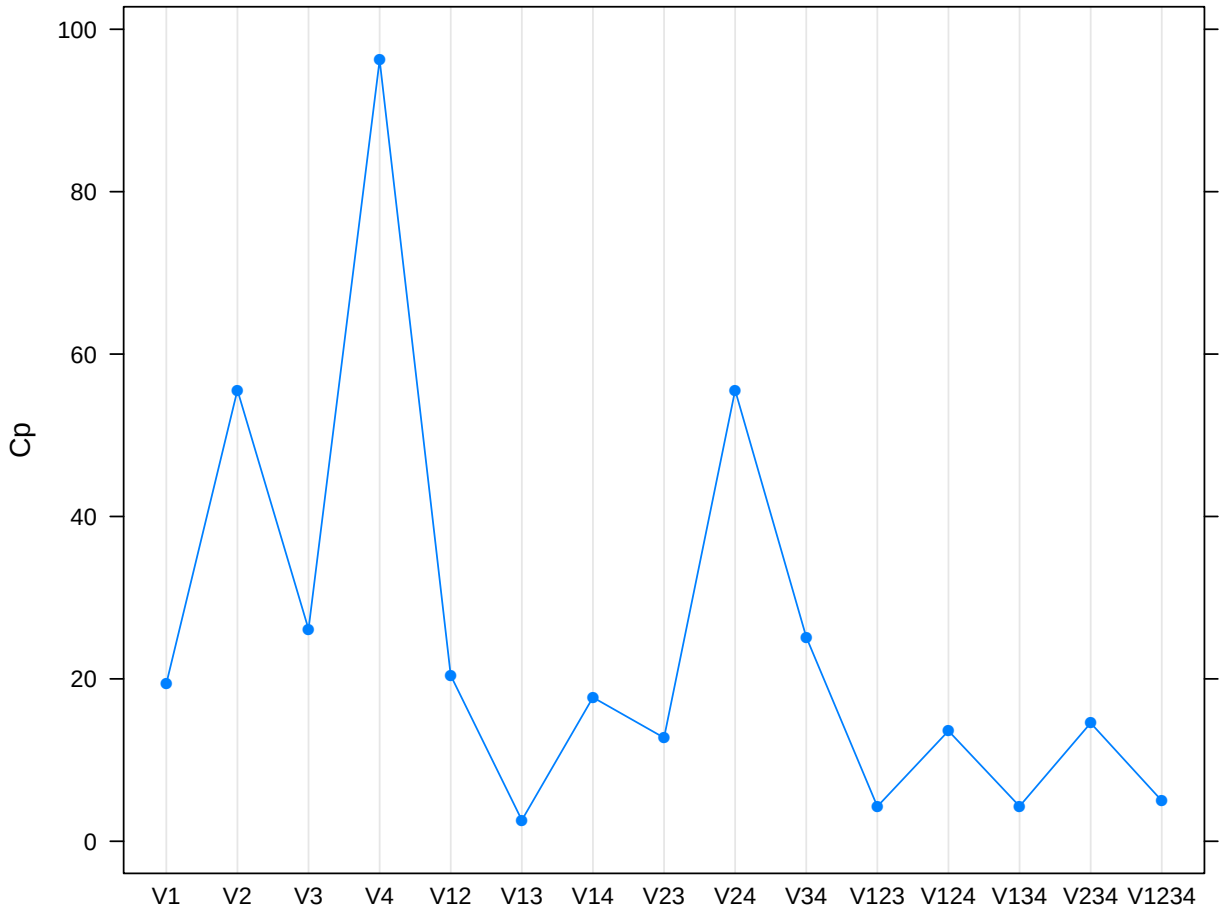
Coefficient of determination (R^2)



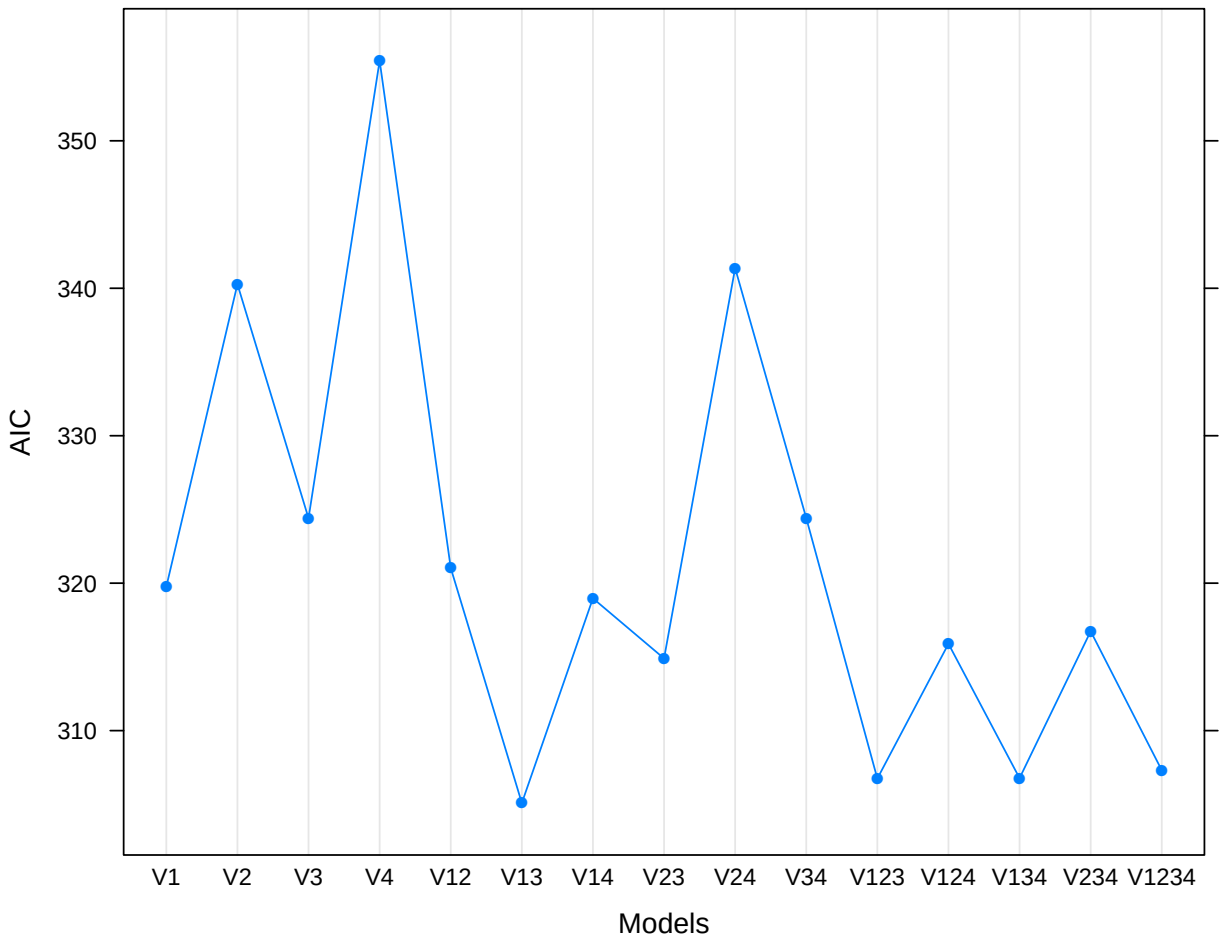
R^2 & R^2_{adj}



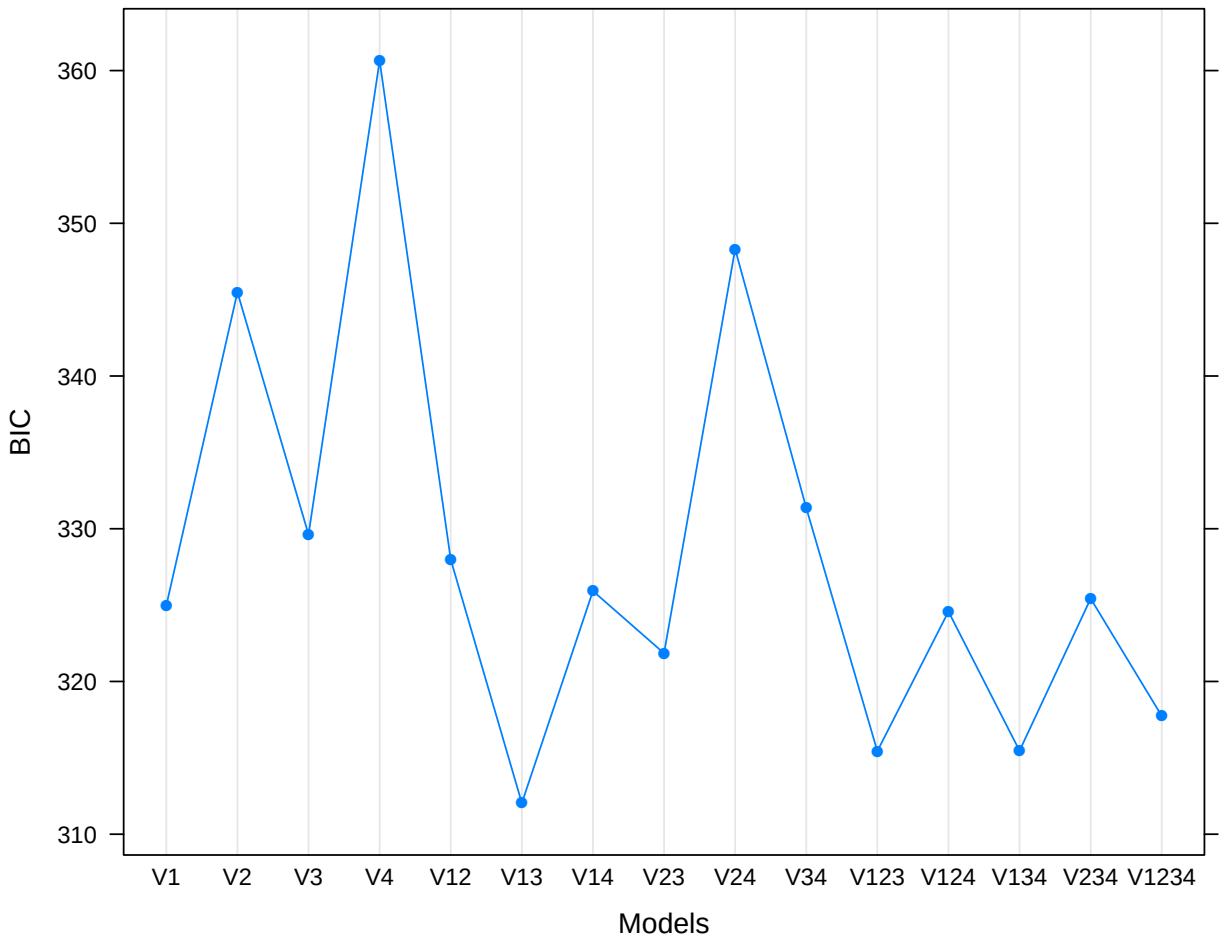
Mallow's C_p



Akaike information criterion (AIC)



Bayesian information criterion (BIC)



Leave-One-Out CV

- We calculate the leave one out CV values for estimating prediction error.
- Instead of fitting a model each time and then calculating $CV(1) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}(i)\right]^2$, we use the more formula that is $CV(1) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\right]^2}{[1-h_i]^2}$ where h_i are the i^{th} diagonal entries of the hat matrix.
- We calculate this value for each of the model and then make the line diagram to get an idea of optimum model.

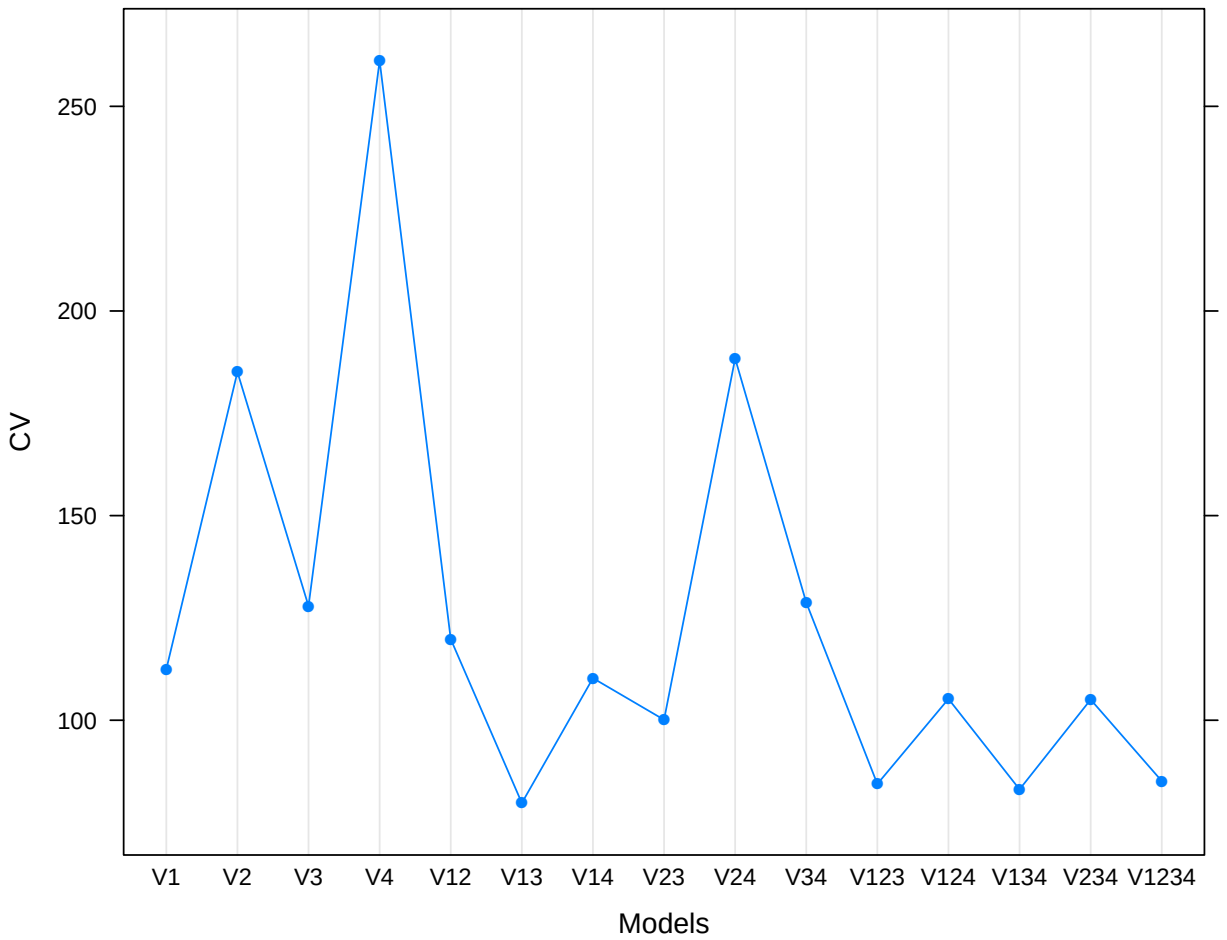
Leave-One-Out CV

- We calculate the leave one out CV values for estimating prediction error.
- Instead of fitting a model each time and then calculating $CV\left(1\right)=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\left(i\right)\right]^2$, we use the more formula that is $CV\left(1\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\right]^2}{\left[1-h_i\right]^2}$ where h_i are the i^{th} diagonal entries of the hat matrix.
- We calculate this value for each of the model and then make the line diagram to get an idea of optimum model.

Leave-One-Out CV

- We calculate the leave one out CV values for estimating prediction error.
- Instead of fitting a model each time and then calculating $CV\left(1\right)=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\left(i\right)\right]^2$, we use the more formula that is $CV\left(1\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\right]^2}{\left[1-h_i\right]^2}$ where h_i are the i^{th} diagonal entries of the hat matrix.
- We calculate this value for each of the model and then make the line diagram to get an idea of optimum model.

Leave-One-Out CV



Values of different measures for all the models

- We list down the values of R^2, R^2_{adj}, Mallow's $C_p, AIC, BIC, CV(1)$ values in one table for all the 15 models for better comparison :-

	R2	adj.R2	Cp	AIC	BIC	CV
V1	0.7262947	0.7194521	19.406493	319.7608	324.9738	112.46225
V2	0.5539177	0.5427657	55.560548	340.2757	345.4888	185.08170
V3	0.6942925	0.6866498	26.118573	324.4050	329.6180	127.88888
V4	0.3600579	0.3440593	96.220397	355.4325	360.6455	261.15078
V12	0.7309995	0.7172046	20.419718	321.0326	327.9832	119.82317
V13	0.8159665	0.8065289	2.598883	305.0896	312.0402	79.76883
V14	0.7438759	0.7307413	17.719047	318.9724	325.9231	110.28385
V23	0.7675782	0.7556591	12.747757	314.8939	321.8445	100.10949
V24	0.5635089	0.5411247	55.548917	341.3629	348.3135	188.38242
V34	0.7084529	0.6935017	25.148610	324.4131	331.3638	128.75523
V123	0.8174541	0.8030425	4.286879	306.7487	315.4370	84.42043
V124	0.7730732	0.7551579	13.595257	315.8890	324.5773	105.20886
V134	0.8173399	0.8029194	4.310820	306.7749	315.4633	83.10493
V234	0.7684411	0.7501601	14.566777	316.7376	325.4260	105.04886
V1234	0.8235897	0.8045183	5.000000	307.3127	317.7387	85.14023

Conclusion

Now, we write down the optimals models we get from different model selection criterions with corresponding values :-

Criterions	Optimum Model	Value
R^2_{adj}	$Y = \beta_0 + \beta_1 \mathrm{A1} + \beta_3 \mathrm{A3}$	0.8065
Mallow's C_p	$Y = \beta_0 + \beta_1 \mathrm{A1} + \beta_3 \mathrm{A3}$	2.598
	$Y = \beta_0 + \beta_1 \mathrm{A1} + \beta_3 \mathrm{A3} + \beta_4 \mathrm{A4}$	4.286879
	$Y = \beta_0 + \beta_1 \mathrm{A1} + \beta_2 \mathrm{A2} + \beta_3 \mathrm{A3}$	4.310820
AIC	$Y = \beta_0 + \beta_1 \mathrm{A1} + \beta_3 \mathrm{A3}$	305.089
BIC	$Y = \beta_0 + \beta_1 \mathrm{A1} + \beta_3 \mathrm{A3}$	312.0402
$CV\ (1)$	$Y = \beta_0 + \beta_1 \mathrm{A1} + \beta_3 \mathrm{A3}$	79.76883

Conclusion

- Hence, clearly this indicates among all the linear models, $Y = \beta_0 + \beta_1 \text{A1} + \beta_3 \text{A3}$ is optimum based on several criterions.
- This is also intuitive from the fact that here we are removing the covariates which had linear dependence.
- In terms of terminology of the given dataset, the optimum predictors of $Y =$ Cirrhosis death rate are
  A1 = Urban population & A3 = Wine consumption per capita.
- So the optimum fitted model can be written as :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Also we observed that all the covariates in the model are signifcant.

Conclusion

- Hence, clearly this indicates among all the linear models, $Y = \beta_0 + \beta_1 \text{A1} + \beta_3 \text{A3}$ is optimum based on several criterions.
- This is also intuitive from the fact that here we are removing the covariates which had linear dependence.
- In terms of terminology of the given dataset, the optimum predictors of $Y =$ Cirrhosis death rate are
  A1 = Urban population & A3 = Wine consumption per capita.
- So the optimum fitted model can be written as :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Also we observed that all the covariates in the model are signifcant.

Conclusion

- Hence, clearly this indicates among all the linear models, $Y = \beta_0 + \beta_1 \text{A1} + \beta_3 \text{A3}$ is optimum based on several criterions.
- This is also intuitive from the fact that here we are removing the covariates which had linear dependence.
- In terms of terminology of the given dataset, the optimum predictors of $Y =$ Cirrhosis death rate are A1 = Urban population & A3 = Wine consumption per capita.
- So the optimum fitted model can be written as :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Also we observed that all the covariates in the model are signifcant.

Conclusion

- Hence, clearly this indicates among all the linear models, $Y = \beta_0 + \beta_1 \text{A1} + \beta_3 \text{A3}$ is optimum based on several criterions.
- This is also intuitive from the fact that here we are removing the covariates which had linear dependence.
- In terms of terminology of the given dataset, the optimum predictors of $Y =$ Cirrhosis death rate are
  A1 = Urban population & A3 = Wine consumption per capita.
- So the optimum fitted model can be written as :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Also we observed that all the covariates in the model are signifcant.

Conclusion

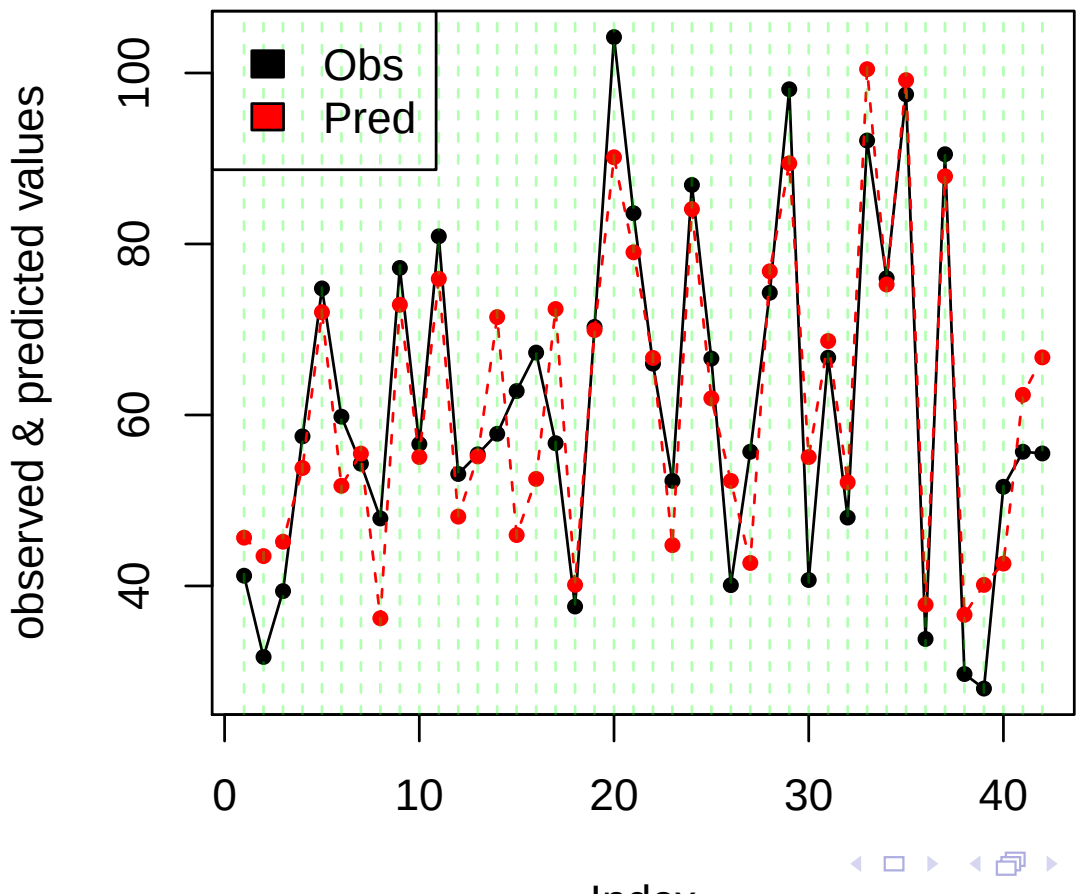
- Hence, clearly this indicates among all the linear models, $Y = \beta_0 + \beta_1 \text{A1} + \beta_3 \text{A3}$ is optimum based on several criterions.
- This is also intuitive from the fact that here we are removing the covariates which had linear dependence.
- In terms of terminology of the given dataset, the optimum predictors of $Y =$ Cirrhosis death rate are
  A1 = Urban population & A3 = Wine consumption per capita.
- So the optimum fitted model can be written as :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Also we observed that all the covariates in the model are signifcant.

Obs vs Fitted Values

- We plot the observed vs fitted values obtained using this model :-



Conclusion

- Again we perform all the diagonistic tests for this final model and find the following :-

- All the assumptions seem to be satishfied here. Hence we can really consider this to be a good model.

Conclusion

- Again we perform all the diagonistic tests for this final model and find the following :-

Tests	p-values
Shapiro-Wilk	0.6282
Durbin-Watson	0.645
Breusch-Pagan	0.2723
Breusch-Godfrey	0.9345

- All the assumptions seem to be satishfied here. Hence we can really consider this to be a good model.

Conclusion

- Again we perform all the diagonistic tests for this final model and find the following :-

Tests	p-values
Shapiro-Wilk	0.6282
Durbin-Watson	0.645
Breusch-Pagan	0.2723
Breusch-Godfrey	0.9345

- All the assumptions seem to be satishfied here. Hence we can really consider this to be a good model.

Models with Interaction Terms

- One class of models that we have not considered yet are those with interaction terms (upto second order). Since there can be too many of them, we will not perform best subset selection here. Rather we again perform stepwise regression for choosing an optimal one among them.
- In this class, we get the following model :-

```
Call:
lm(formula = Y ~ A1 + A3 + A4 + A1:A4, data = X[-c(12, 20, 30,
    36), ])

Coefficients:
(Intercept)           A1           A3           A4        A1:A4
  36.182148     0.121091     1.366505    -0.448029     0.008963
```

Models with Interaction Terms

- One class of models that we have not considered yet are those with interaction terms (upto second order). Since there can be too many of them, we will not perform best subset selection here. Rather we again perform stepwise regression for choosing an optimal one among them.
- In this class, we get the following model :-

```
Call:
lm(formula = Y ~ A1 + A3 + A4 + A1:A4, data = X[-c(12, 20, 30,
    36), ])

Coefficients:
(Intercept)           A1           A3           A4        A1:A4
  36.182148     0.121091     1.366505    -0.448029     0.008963
```

Models with Interaction Terms

- But if we calculate the VIF values, we get :-

```
       A1        A3        A4     A1:A4
16.892829  2.545754 18.460498 52.872700
```

- This model has an adjusted R^2 value equal to 0.8299. But since, the main problem is this model has very high vif values and many of the predictors are not signifcant. So we don't consider these type of models.

Conclusion

- Hence we conclude our final multiple linear regression model is :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Obviously this model also has some drawback and there is no such "best" model that we can have but this performs more or less better than most of the models hence, it is a good one.
- Next we use other types of regression models with different interpretations.

Conclusion

- Hence we conclude our final multiple linear regression model is :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Obviously this model also has some drawback and there is no such "best" model that we can have but this performs more or less better than most of the models hence, it is a good one.
- Next we use other types of regression models with different interpretations.

Conclusion

- Hence we conclude our final multiple linear regression model is :-

$$Y = 9.9241 + 0.6397\text{A1} + 1.5159\text{A3}$$

- Obviously this model also has some drawback and there is no such "best" model that we can have but this performs more or less better than most of the models hence, it is a good one.
- Next we use other types of regression models with different interpretations.

Ridge Regression

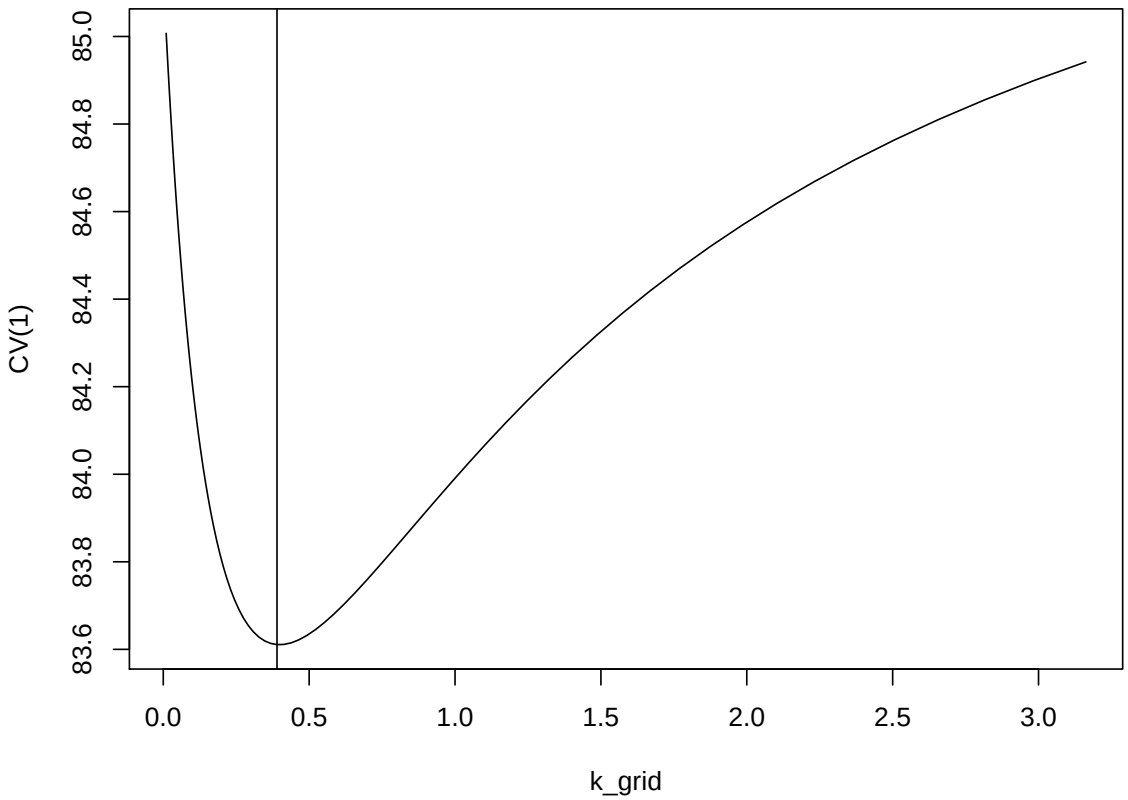
- An alternate approach to deal with collinearity is fitting a ridge regression model as it can improve the accuracy of the predictions. The ridge estimate of the model parameters is $\widehat{\boldsymbol{\beta}}(k) = \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ where k is the ridge parameter.
- For an optimal choice of k, we calculate estimates of prediction errors of the ridge predictors for different choices of k over a set of trial values. This can be expressed as $CV_k(1) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}(k)\right]^2}{\left[1-a_{ii}(k)\right]^2}$ and choose the k_{opt} for which this quantity is minimum.

Ridge Regression

- An alternate approach to deal with collinearity is fitting a ridge regression model as it can improve the accuracy of the predictions. The ridge estimate of the model parameters is $\widehat{\boldsymbol{\beta}}(k) = \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ where k is the ridge parameter.
- For an optimal choice of k, we calculate estimates of prediction errors of the ridge predictors for different choices of k over a set of trial values. This can be expressed as $CV_k(1) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}(k)\right]^2}{\left[1 - a_{ii}(k)\right]^2}$ and choose the k_{opt} for which this quantity is minimum.

Ridge Regression

- We plot the $CV_k\,(1)$ values for different choices of k :-



Ridge Regression

- We find that the $CV_k\,(1)$ is minimum for $k \approx 0.3898$ hence, we calculate the corresponding ridge estimates.
- The fitted model then becomes :-

$$Y = 11.777 + 0.767\text{A1} + -0.322\text{A2} + 1.324\text{A3} + 0.114\text{A4}$$

- This model has estimated prediction error ≈ 83.612.
- This is close to the optimum OLS model that we have fitted.

Ridge Regression

- We find that the $CV_k(1)$ is minimum for $k \approx 0.3898$ hence, we calculate the corresponding ridge estimates.
- The fitted model then becomes :-

$$Y = 11.777 + 0.767\text{A1} + -0.322\text{A2} + 1.324\text{A3} + 0.114\text{A4}$$

- This model has estimated prediction error ≈ 83.612.
- This is close to the optimum OLS model that we have fitted.

Ridge Regression

- We find that the $CV_k\,(1)$ is minimum for $k \approx 0.3898$ hence, we calculate the corresponding ridge estimates.
- The fitted model then becomes :-

$$Y = 11.777 + 0.767\text{A1} + -0.322\text{A2} + 1.324\text{A3} + 0.114\text{A4}$$

- This model has estimated prediction error ≈ 83.612.
- This is close to the optimum OLS model that we have fitted.

Ridge Regression

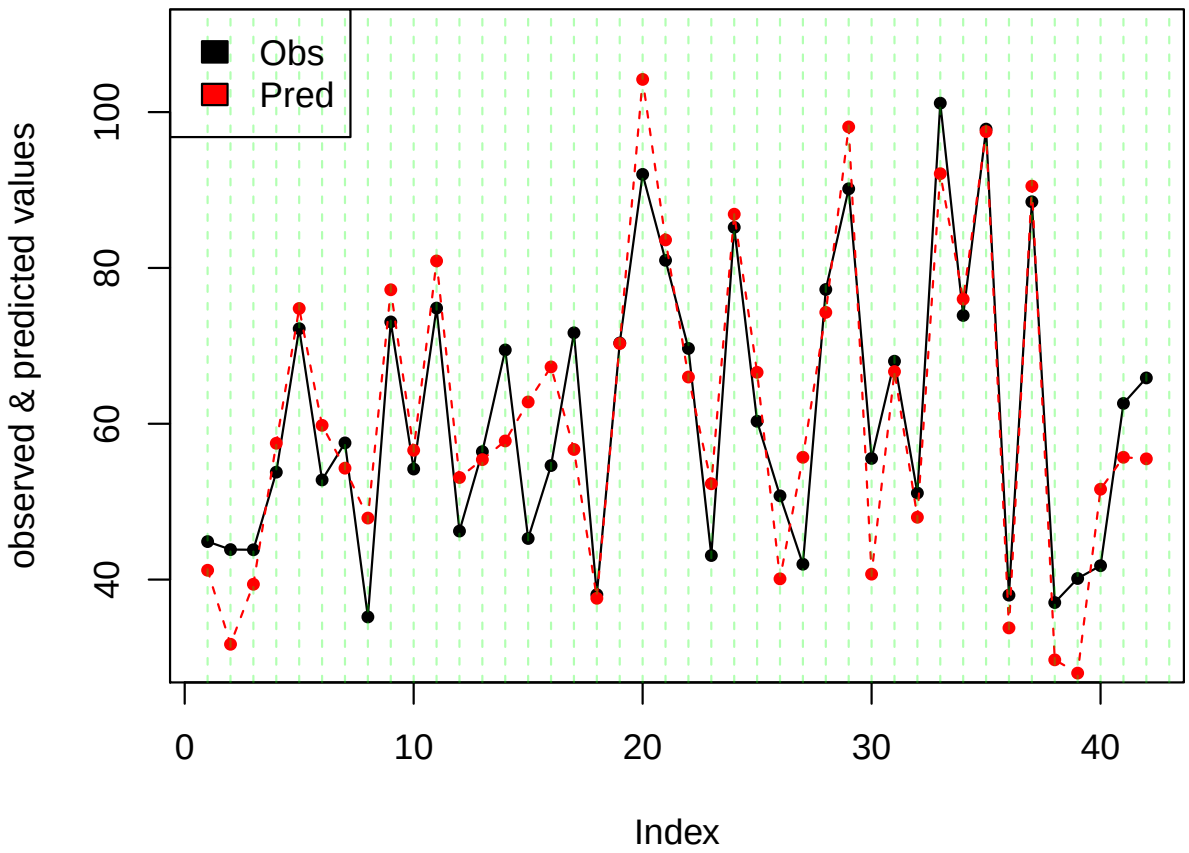
- We find that the $CV_k\,(1)$ is minimum for $k \approx 0.3898$ hence, we calculate the corresponding ridge estimates.
- The fitted model then becomes :-

$$Y = 11.777 + 0.767\text{A1} + -0.322\text{A2} + 1.324\text{A3} + 0.114\text{A4}$$

- This model has estimated prediction error ≈ 83.612.
- This is close to the optimum OLS model that we have fitted.

Ridge Regression

- We plot the observed & fitted values with the same index in the x-axis and get the following output :-



Lasso Regression

- Another efficient way of model selection is using the Lasso Regression method. Here we minimize the sum of squares $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2$ subject to the constraint $\sum_j |\beta_j| \leq \lambda$ for some $\lambda > 0$.
- Using R, we find the Lasso Estimates of β where the value of λ is chosen using k-fold cross-validation criteria.
- The optimum value of λ chosen by the criteria approximately equals ≈ 0.501 and the model is :-

```
$s0
[1] 25.2888

[[2]]
1 x 4 sparse Matrix of class "dgCMatrix"
          A1 A2       A3 A4
s0 0.4579335  . 1.013119  .
```

Lasso Regression

- Another efficient way of model selection is using the Lasso Regression method. Here we minimize the sum of squares $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2$ subject to the constraint $\sum_j |\beta_j| \leq \lambda$ for some $\lambda > 0$.
- Using R, we find the Lasso Estimates of $\boldsymbol{\beta}$ where the value of λ is chosen using k-fold cross-validation criteria.
- The optimum value of λ chosen by the criteria approximately equals ≈ 0.501 and the model is :-

```
$s0
[1] 25.2888

[[2]]
1 x 4 sparse Matrix of class "dgCMatrix"
          A1 A2       A3 A4
s0 0.4579335  . 1.013119  .
```

Lasso Regression

- Another efficient way of model selection is using the Lasso Regression method. Here we minimize the sum of squares $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2$ subject to the constraint $\sum_j |\beta_j| \leq \lambda$ for some $\lambda > 0$.
- Using R, we find the Lasso Estimates of $\boldsymbol{\beta}$ where the value of λ is chosen using k-fold cross-validation criteria.
- The optimum value of λ chosen by the criteria approximately equals ≈ 0.501 and the model is :-

```
$s0
[1] 25.2888

[[2]]
1 x 4 sparse Matrix of class "dgCMatrix"
          A1 A2       A3 A4
s0 0.4579335  . 1.013119  .
```

Lasso Regression

- Hence the Lasso Estimates of the parameters

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 25.28 \\ 0.457 \\ 0 \\ 1.013 \\ 0 \end{pmatrix}.$$

- As we can see from the output, here also, the variables "A2", "A4" has been dropped and this also gives strong evidence in favour of the optimum linear model.

Lasso Regression

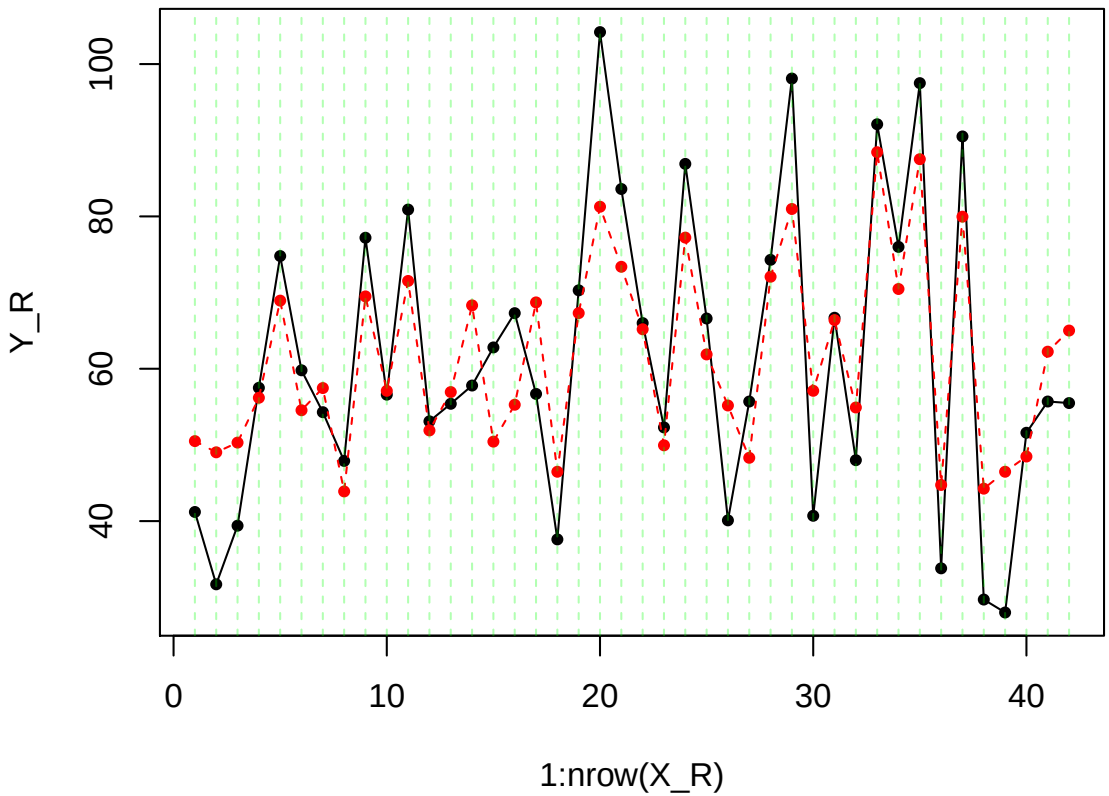
- Hence the Lasso Estimates of the parameters

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 25.28 \\ 0.457 \\ 0 \\ 1.013 \\ 0 \end{pmatrix}.$$

- As we can see from the output, here also, the variables "A2", "A4" has been dropped and this also gives strong evidence in favour of the optimum linear model.

Obs & Fitted Values

- Here also make the observed and fitted plot for different index values :-



Robust Regression Methods

- We have detected influential points in our dataset, and also removed them to get better models.
- Now, we demonstrate using different robust regression methods how they can be used even if we have outliers in our dataset.
- So we perform the rest of the methods using the full dataset, without removing any observation.

Robust Regression Methods

- We have detected influential points in our dataset, and also removed them to get better models.
- Now, we demonstrate using different robust regression methods how they can be used even if we have outliers in our dataset.
- So we perform the rest of the methods using the full dataset, without removing any observation.

Robust Regression Methods

- We have detected influential points in our dataset, and also removed them to get better models.
- Now, we demonstrate using different robust regression methods how they can be used even if we have outliers in our dataset.
- So we perform the rest of the methods using the full dataset, without removing any observation.

Least Absolute Deviation

- Here we minimize the quantity $\sum_i |e_i(\boldsymbol{b})|$ i.e.

$$\widehat{\boldsymbol{\beta}}_{LAD} = \underset{\boldsymbol{b}}{\text{argmin}} \sum_i |e_i(\boldsymbol{b})| \text{where } e_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}.$$

- The estimated values of $\widehat{\beta}_{LAD}$ equals :-

```
(Intercept)          A1          A2          A3          A4
-7.19977780  0.42727435  0.61749765  1.96154792 -0.02363183
```

Least Absolute Deviation

- Here we minimize the quantity $\sum_i |e_i(\boldsymbol{b})|$ i.e.

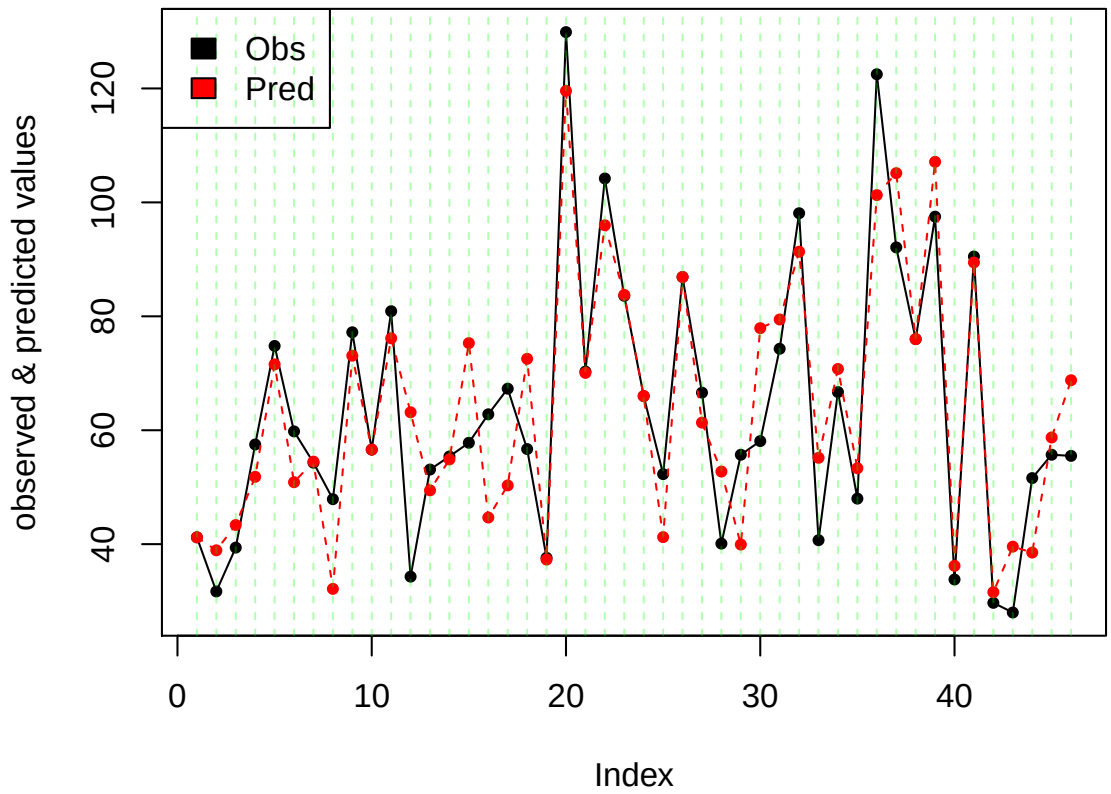
$$\widehat{\boldsymbol{\beta}}_{LAD} = \underset{\boldsymbol{b}}{\text{argmin}} \sum_i |e_i(\boldsymbol{b})| \text{where } e_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}.$$

- The estimated values of $\widehat{\boldsymbol{\beta}}_{LAD}$ equals :-

```
(Intercept)          A1          A2          A3          A4
-7.19977780  0.42727435  0.61749765  1.96154792 -0.02363183
```

Obs vs Fitted Values

- We plot the observed vs fitted values obtained using this model :-



Obs vs Fitted Values

- The plot shows here the predicted values are more or less accurate for all the observations.

Least Median Square

- Here we minimize the median of the squared residuals $\widehat{\boldsymbol{\beta}}_{LMS} = \underset{\boldsymbol{b}}{\text{argmin}} \underset{i}{\text{med}}\, e_i^2(\boldsymbol{b})$.
- Using R, we get the estimated value of $\widehat{\boldsymbol{\beta}}_{LMS}$ as :-

```
(Intercept)          A1          A2          A3          A4
100.67646579  1.43011898 -3.48160404  3.15085397 -0.04186281
```

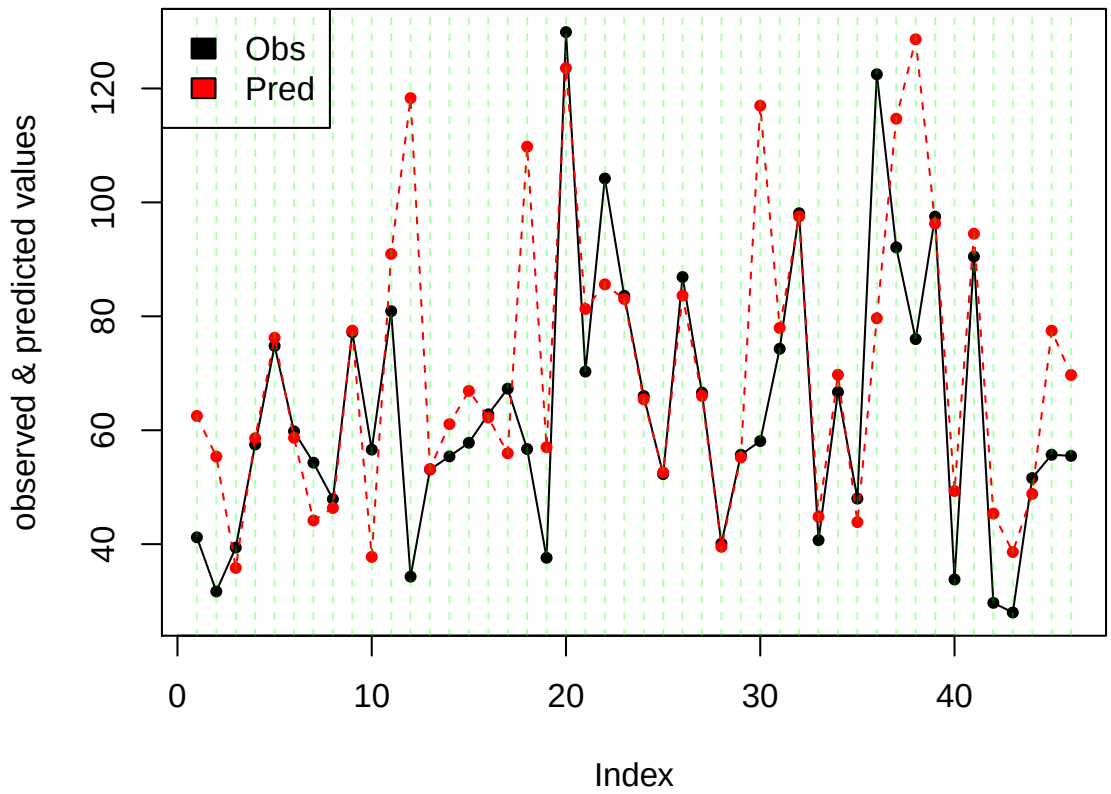
Least Median Square

- Here we minimize the median of the squared residuals $\widehat{\boldsymbol{\beta}}_{LMS} = \underset{\boldsymbol{b}}{\text{argmin}} \underset{i}{\text{med}}\, e_i^2(\boldsymbol{b})$.
- Using R, we get the estimated value of $\widehat{\boldsymbol{\beta}}_{LMS}$ as :-

```
 (Intercept)           A1           A2           A3           A4
100.67646579   1.43011898  -3.48160404   3.15085397  -0.04186281
```

Obs vs Fitted Values

- We plot the observed vs fitted values obtained using this model :-



Obs vs Fitted Values

- We can see the prediction is quite accurate at some places where as it is bad at some others possibly due to the presence of outliers.

Least Trimmed Squares Estimate

- Lastly, we compute the LTS estimates of $\boldsymbol{\beta}$ where we minimize the trimmed mean of the squared residuals

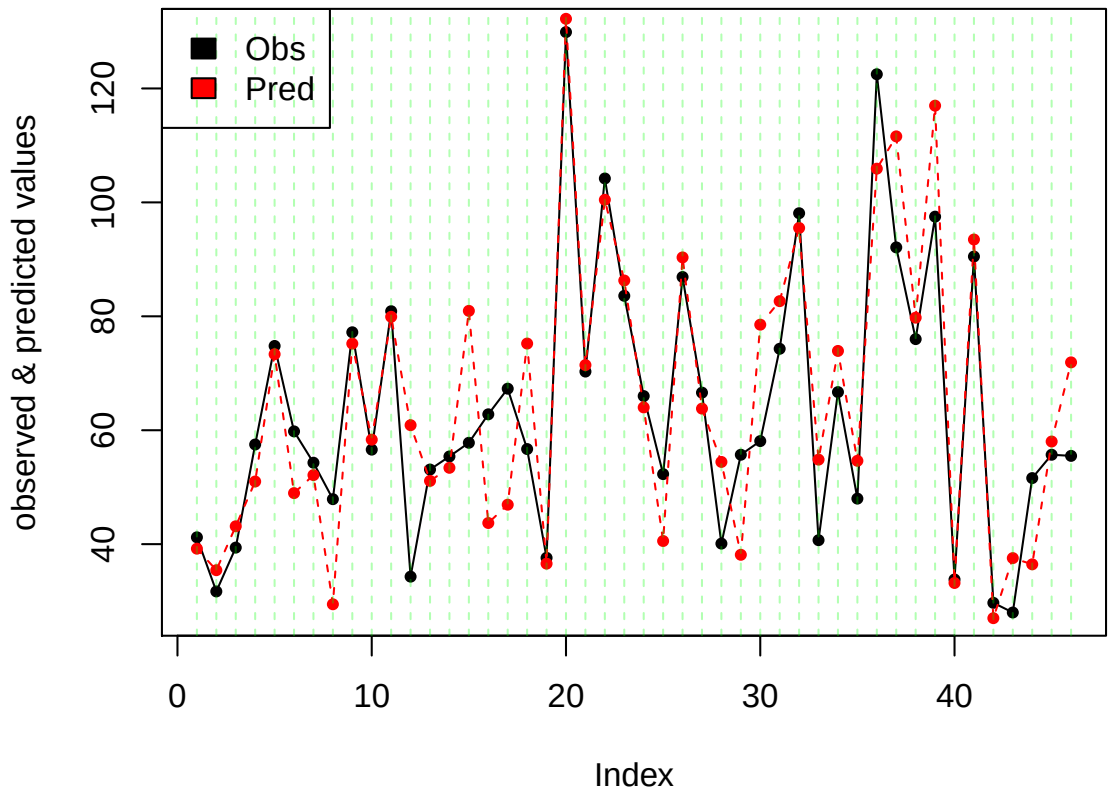
  $\widehat{\boldsymbol{\beta}}_{LMS} = \underset{\boldsymbol{b}}{\text{argmin}} \frac{1}{h} \sum_{i=1}^{h} e_i^2(\boldsymbol{b})$ for some appropriate choice of h.

  (Here we choose $h = [n/2] + 1$)

```
(Intercept)          A1          A2          A3          A4
-21.7304500   0.2861684   1.2215938   2.4980685  -0.1567162
```

Obs vs Fitted Values

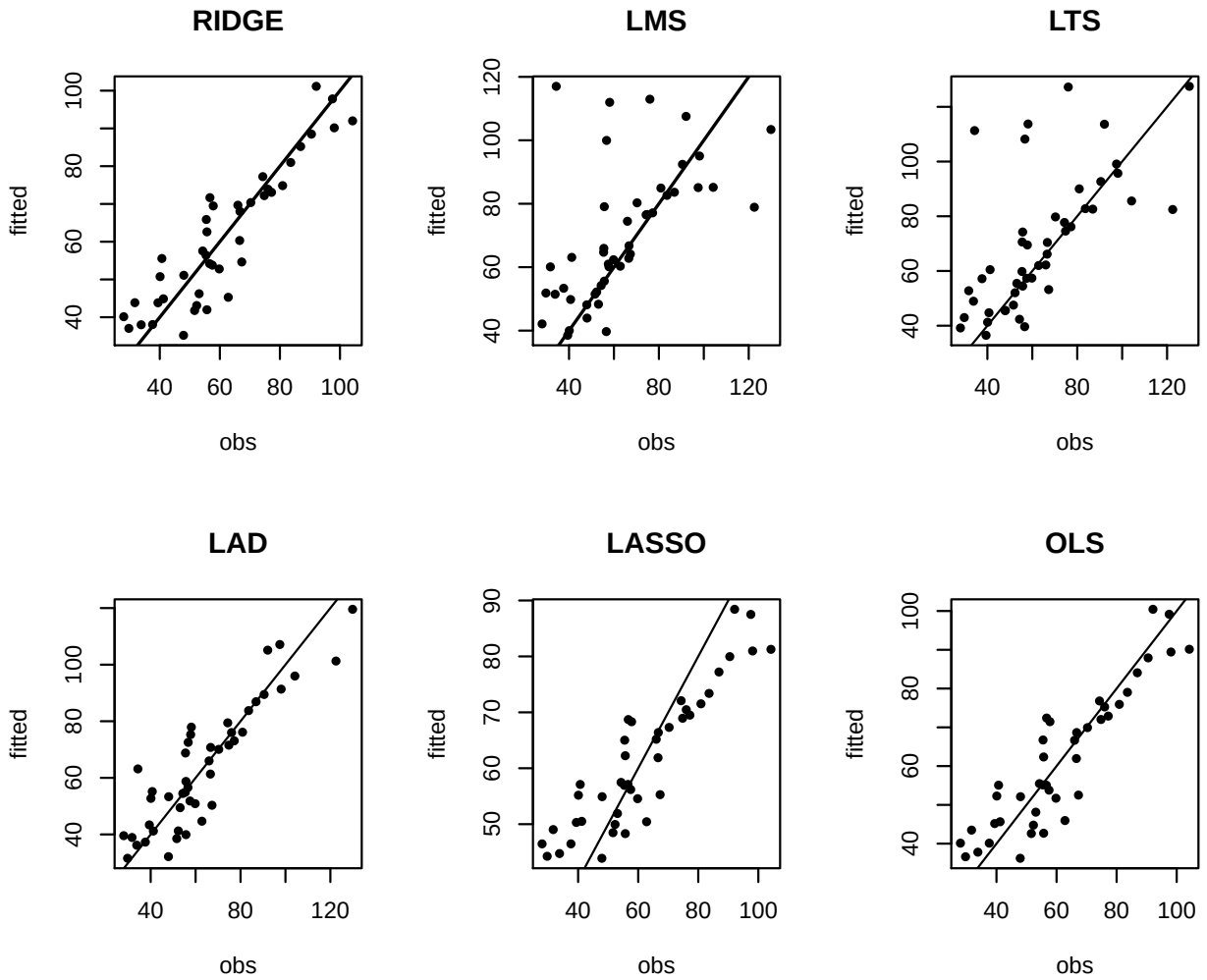
- We plot the observed vs fitted values obtained using this model :-



Comparative Study Of All Models

- We make the observed vs fitted value plots for all the "good" models that we have considered so far :-

Comparative Study Of All Models



Comparative Study Of All Models

- Finally, as a measure of comparison of different models, we use the root mean square error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}$ where $\widehat{y}_i$ denotes the fitted values using different regression models. We calculate this measure for all the "good" models we have found so far :-

- Hence, we can conclude the Ridge and the OLS model with influential points removed with covariates "A1" & "A3" performs more or less better than the others in terms of prediction accuracy.

Comparative Study Of All Models

- Finally, as a measure of comparison of different models, we use the root mean square error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}$ where $\widehat{y}_i$ denotes the fitted values using different regression models. We calculate this measure for all the "good" models we have found so far :-

Methods	$RMSE$	CV
OLS model with "A1" & "A3"	8.313178	79.769
Ridge Model	8.195655	83.612
Lasso Model	9.889319	79.984
LAD Model	10.59494	-
LMS Model	11.6656	-
LTS Model	11.0904	-

- Hence, we can conclude the Ridge and the OLS model with influential points removed with covariates "A1" & "A3" performs more or less better than the others in terms of prediction accuracy.

Comparative Study Of All Models

- Finally, as a measure of comparison of different models, we use the root mean square error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}$ where $\widehat{y}_i$ denotes the fitted values using different regression models. We calculate this measure for all the "good" models we have found so far :-

Methods	$RMSE$	CV
OLS model with "A1" & "A3"	8.313178	79.769
Ridge Model	8.195655	83.612
Lasso Model	9.889319	79.984
LAD Model	10.59494	-
LMS Model	11.6656	-
LTS Model	11.0904	-

- Hence, we can conclude the Ridge and the OLS model with influential points removed with covariates "A1" & "A3" performs more or less better than the others in terms of prediction accuracy.

Acknowledgement

We would like to express our **special thanks of gratitude** to our respected **Prof.Swagata Nandi** ma'am for helping us throught the presentation work and also for giving us this wonderful opportunity as we learned many things during the making of this presentation.