Regression Project

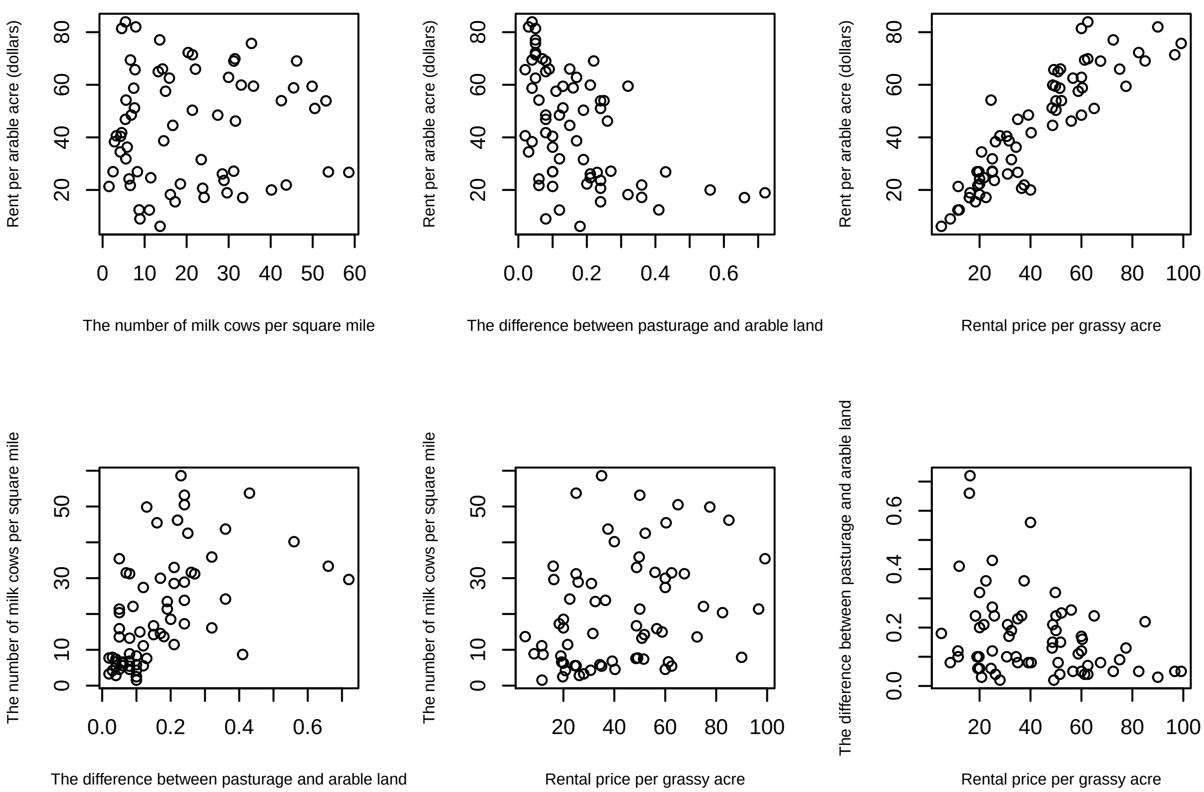
Akshay, Rohan, Soumyabrata

Description of the data : First few rows of the dataset are shown below.

Rent per arable acre (dollars)	The number of milk cows per square mile	The difference between pasturage and arable land	Rental price per grassy acre for this variety of grass
15.50	17.25	0.24	18.38
22.29	18.51	0.20	20.00
12.36	11.13	0.12	11.50
31.84	5.54	0.12	25.00
83.90	5.44	0.04	62.50

For the analysis, we consider the first column as the response variable and the other columns as the predictors. The response is denoted as y, and the predictors are denoted as x_1, x_2 and x_3 respectively.

Preliminary Analysis : We consider the plots of all pairs of columns. The plot is given below.



Comment :

1. In the 1st graph, there seems to be an polynomial relationship between y and x_1.
2. From the 2nd graph, it seems that there is a linear relationship between y and x_2 with a negative slope.
3. From the 3rd graph, it seems that there is a linear relationship between y and x_3 with a positive slope.
4. From the 4th we can say that, there is an approximate linear relationship between x_1 and x_2 with a positive slope.
5. From the 5th graph it seems that all the points are scattered randomly, resulting in a very low co-linearity between $x1$ and $x3$.
6. From the 6th graph it seems that, there is a linear relation between x_2 and x_3 with a negative slope.

Correlation: The correlation among the columns are shown below.

```
cor(data)
```

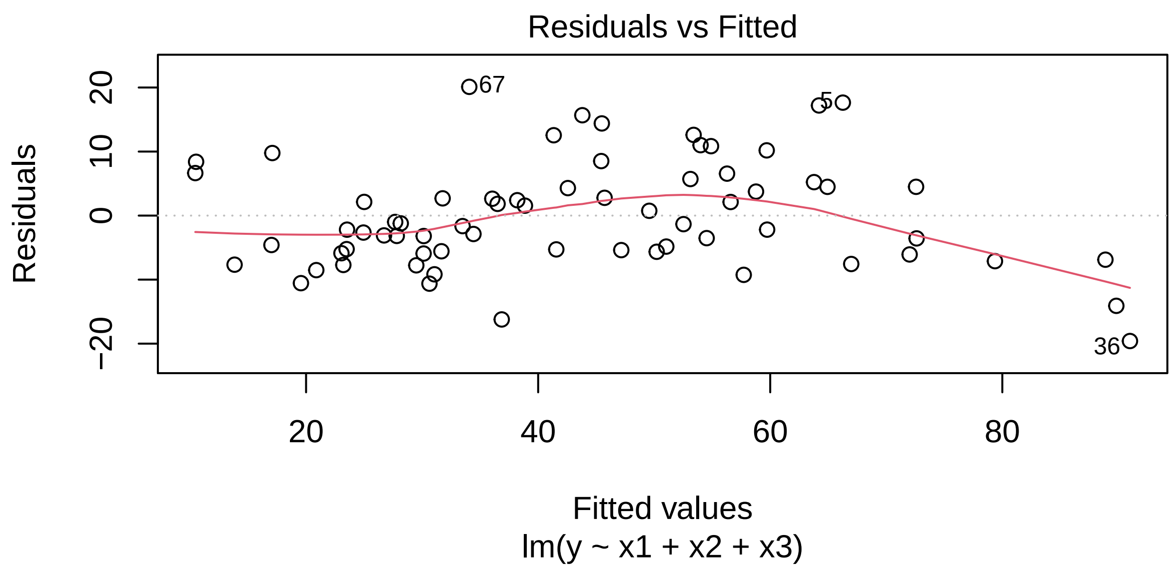
```
##               y         x1         x2         x3
## y   1.00000000 0.04550419 -0.4978930  0.8850823
## x1  0.04550419 1.00000000  0.5225979  0.3033919
## x2 -0.49789295 0.52259791  1.0000000 -0.3301773
## x3  0.88508229 0.30339186 -0.3301773  1.0000000
```

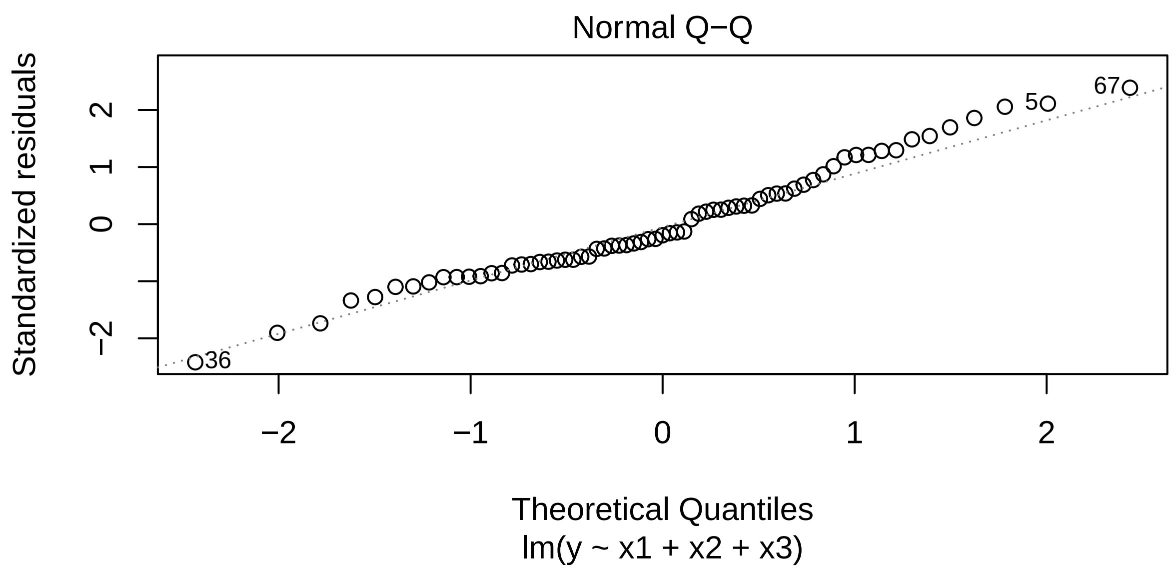
Model Fitting : First we fit the full model.

```
model.1 <- lm(y ~ x1 + x2 + x3, data)
summary(model.1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.593  -5.765  -1.646   4.858  20.106
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.84304    3.08776   5.131 2.98e-06 ***
## x1           -0.23266    0.10036  -2.318   0.0237 *
## x2          -16.72418   10.75285  -1.555   0.1249
## x3            0.83757    0.06092  13.749  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 63 degrees of freedom
## Multiple R-squared:  0.8441, Adjusted R-squared:  0.8367
## F-statistic: 113.7 on 3 and 63 DF,  p-value: < 2.2e-16
```

The residual plots are given below.



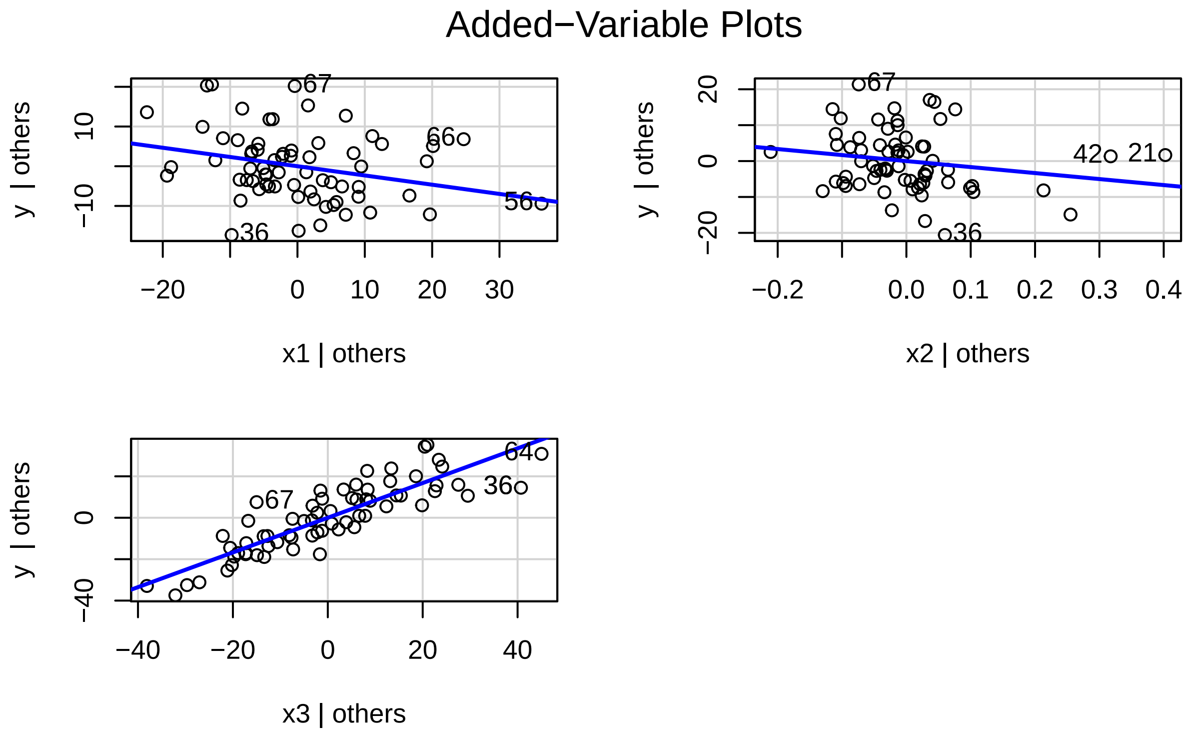


Comment: From the QQ-Plot, it seems that the residuals are almost normal.

Added Variable Plot: From the previous scatterplots of the response(y) vs.the predictors(x_1, x_2, x_3), we see a polynomial relation between y and x_1 and y and x_2. we can also consider x_1^2 and x_2^2 as predictors in the model.

We construct the *Added Variable Plot* to better serve the purpose. The plots are given below.

```
suppressMessages(library(car))
avPlots(model.1)
```



Comment: From the *Added Variable Plots*, it is clear that the predictor x_3 is almost linearly related to the response y. We can consider x_1^2 and x_2^2 as predictors in the model, but the correlation between x_1 and x_1^2 is very close to 1(exact value is 0.9618495). Similarly, the correlation between x_2 and x_2^2 is also very high(0.936776). So, considering x_1^2 and x_2^2 in the model won't give us any extra info.

<u>**Outliers and Influential Points**</u>: We calculate the measures for detecting outliers and influential points corresponding to each residuals, viz, DFFITS, Covariance Ratio, Cook's D and hat matrix diagonals.

```
influence.measures(model.1)
```

They are tabulated below.

Index	DFFITS	Cov-Ratio	Cook's D	h
1	-0.1658832	1.0444959	0.0068980	0.0321299
2	-0.0564033	1.0942107	0.0008069	0.0317672
3	-0.2520088	1.0580608	0.0158664	0.0574211
4	-0.0359064	1.1002238	0.0003273	0.0333533
5	0.5233315	0.8410725	0.0646514	0.0548422
6	-0.2292185	1.0902964	0.0131915	0.0670139
7	0.0508181	1.1052318	0.0006554	0.0393274
8	0.0334074	1.0997345	0.0002834	0.0326547
9	-0.2040705	1.1805364	0.0105249	0.1153176
10	0.1201390	1.1002074	0.0036499	0.0485060
11	-0.1801219	1.0145887	0.0080860	0.0264466
12	0.1323156	1.0877886	0.0044202	0.0437632
13	-0.1156053	1.0692970	0.0033717	0.0301633
14	0.1264924	1.0540048	0.0040261	0.0262920
15	0.1283210	1.1070660	0.0041638	0.0545413
16	0.1959373	1.1181495	0.0096788	0.0750950
17	0.3829408	0.8851801	0.0352070	0.0391216

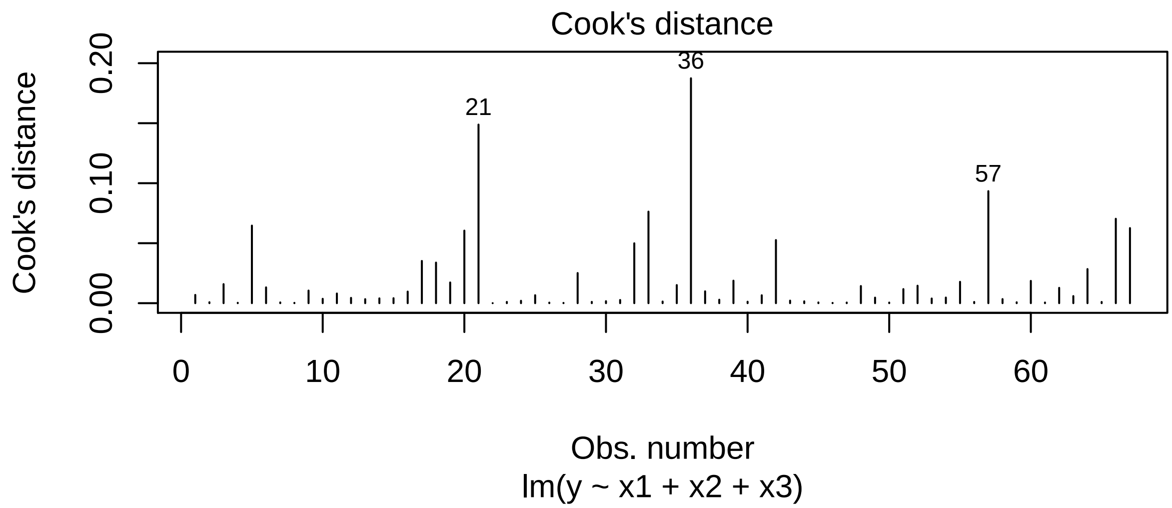
Index	DFFITS	Cov-Ratio	Cook's D	h
18	-0.3697834	1.0393466	0.0338360	0.0765434
19	-0.2686210	0.8569300	0.0172746	0.0186886
20	0.5051801	0.8535150	0.0604779	0.0540966
21	0.7739525	1.4008313	0.1488578	0.3030098
22	0.0121662	1.0864599	0.0000376	0.0192233
23	-0.0651903	1.1271901	0.0010784	0.0583738
24	0.0888811	1.0814460	0.0019986	0.0301784
25	-0.1624769	1.0833226	0.0066503	0.0482688
26	0.0431807	1.0806157	0.0004729	0.0194111
27	-0.0307114	1.1051217	0.0002395	0.0368483
28	-0.3166694	1.1262638	0.0251251	0.1042600
29	-0.0659954	1.0897705	0.0011039	0.0304346
30	-0.0812175	1.1052021	0.0016718	0.0442098
31	-0.1030839	1.0783179	0.0026854	0.0317434
32	0.4485882	1.1022306	0.0499286	0.1197979
33	-0.5560519	1.1119314	0.0763130	0.1456569
34	0.0738137	1.0827214	0.0013798	0.0275017
35	0.2481568	0.9495735	0.0150954	0.0266163
36	-0.9019691	0.8137972	0.1874469	0.1134502
37	0.1999564	0.9796401	0.0098866	0.0230451
38	-0.1071627	1.0710149	0.0028993	0.0289513
39	0.2781957	0.9074044	0.0187627	0.0254288
40	-0.0696837	1.0956977	0.0012309	0.0353032
41	-0.1618435	1.0870888	0.0066011	0.0501094
42	0.4578417	1.2960421	0.0526063	0.2164835
43	-0.0919973	1.0566419	0.0021350	0.0190021
44	0.0786382	1.1230099	0.0015683	0.0569300
45	-0.0494065	1.0812914	0.0006190	0.0211936
46	-0.0303330	1.1121276	0.0002337	0.0426682
47	0.0464710	1.0987180	0.0005480	0.0335746
48	0.2404022	0.9922949	0.0142978	0.0335801
49	-0.1351269	1.1566904	0.0046246	0.0893123
50	-0.0451217	1.0942424	0.0005166	0.0298696
51	-0.2163883	1.0639636	0.0117322	0.0516749
52	-0.2410526	1.0788864	0.0145617	0.0641229
53	-0.1238732	1.1437482	0.0038867	0.0786294
54	-0.1364769	1.0901320	0.0047023	0.0460311
55	0.2679270	1.0169015	0.0178098	0.0461829
56	-0.0636977	1.3265737	0.0010304	0.1972111
57	-0.6213744	0.9841515	0.0933802	0.1100394
58	-0.1160729	1.0735636	0.0034005	0.0323571
59	0.0547335	1.0889107	0.0007597	0.0275938
60	-0.2730610	1.0472379	0.0185778	0.0579109
61	0.0469327	1.1130817	0.0005591	0.0450912
62	0.2262492	1.0476740	0.0127910	0.0473346
63	-0.1531156	1.0831803	0.0059093	0.0460313
64	-0.3365007	1.1730356	0.0284267	0.1331396
65	0.0634382	1.1134784	0.0010210	0.0475218
66	0.5364725	1.0215956	0.0703456	0.1056051
67	0.5205079	0.7609968	0.0625866	0.0420018

The cutoff values for the measures are given below.

DFFITS > 0.4886778
Covariance Ratio $\in (-\infty, 0.8208955) \cup (1.179104, \infty)$
Cook's D > 1
Hat-Matrix Diagonals > 0.119403

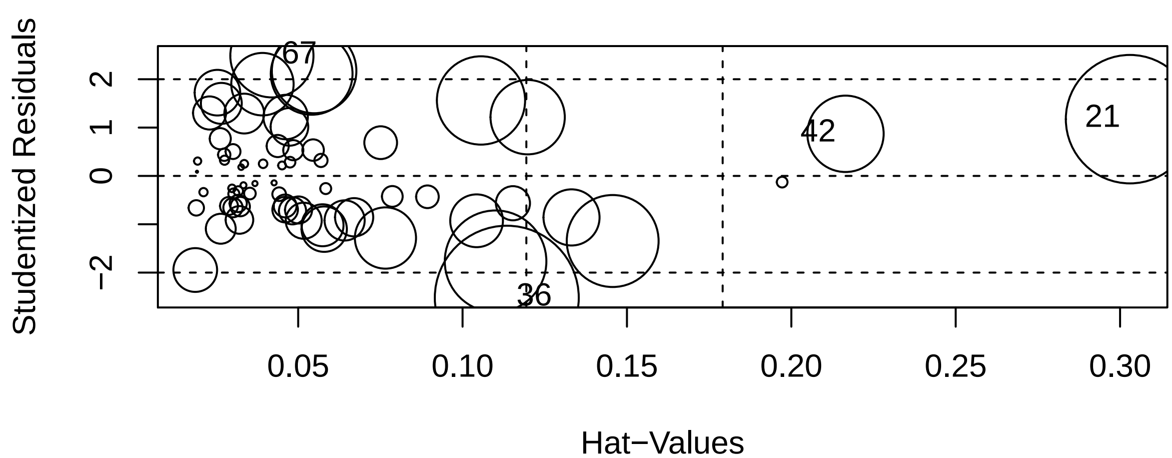
The Cook's Distances are plotted below.

```
plot(model.1, which = 4, cook.levels = 1)
```



The plot of standardized residuals and hat-matrix diagonals are given below.

```
influencePlot(model.1)
```




```
##        StudRes        Hat      CookD
## 21  1.1738139 0.30300983 0.14885783
## 36 -2.5213940 0.11345015 0.18744687
```

```
## 42  0.8710177 0.21648354 0.05260627
## 67  2.4858552 0.04200178 0.06258662
```

The suspicious points are given below.

```
influ.model.1 <- influence.measures(model.1)
summary(influ.model.1)
```

```
## Potentially influential observations of
##   lm(formula = y ~ x1 + x2 + x3, data = data) :
##
##    dfb.1_ dfb.x1 dfb.x2 dfb.x3 dffit   cov.r   cook.d hat
## 21 -0.31  -0.37   0.71   0.23   0.77_*  1.40_*  0.15   0.30_*
## 36  0.45   0.31  -0.20  -0.77  -0.90_*  0.81    0.19   0.11
## 42 -0.16  -0.16   0.39   0.09   0.46    1.30_*  0.05   0.22_*
## 56 -0.02  -0.06   0.04   0.04  -0.06    1.33_*  0.00   0.20_*
## 67  0.48  -0.01  -0.24  -0.27   0.52    0.76_*  0.06   0.04
```

For these points we apply test for outliers using outlier-shift model.

```
outlierTest(model.1)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 36 -2.521394           0.014274      0.95634
```

The points with hat-values more than 0.119403 are given by,

```
which(as.vector(hatvalues(model.1)) > 0.119403)
```

```
## [1] 21 32 33 42 56 64
```

Comment: From the test for outliers, we see that there are no outliers. But 21^{st}, 32^{nd}, 33^{rd}, 42^{nd}, 56^{th}, 64^{th} points are leverage points.

Subset Selection: Here we select the best subset of predictors, best in sense of certain criterion such as AIC, BIC, Mallow's C_p, etc., by using all possible regression and stepwise regression.

All Possible Regression: At first we select the best subset of predictors using AIC as the selection criterion.

```
suppressMessages(library(olsrr))
all_subset <- ols_step_all_possible(model.1)
```

The AIC values for the model subsets are given below.

Index	Number of Predictors	Predictors	AIC
1	1	x_3	502.3522
2	1	x_2	585.7460
3	1	x_1	604.6943
4	2	x_1, x_3	484.8196
5	2	x_2, x_3	487.7797
6	2	x_1, x_2	575.1858
7	3	x_1, x_2, x_3	484.2951

The model with all the 3 predictors is chosen to be the best w.r.t. AIC. The estimates of the parameters for the best model is given below.

term	estimate	std.error	statistic	p.value
(Intercept)	15.8430423	3.0877581	5.130921	0.0000030
x1	-0.2326559	0.1003603	-2.318207	0.0236975
x2	-16.7241770	10.7528479	-1.555325	0.1248782
x3	0.8375671	0.0609186	13.748962	0.0000000

We repeat the same process using *BIC* as the selection criterion. The *BIC* values for the model subsets are given below.

Index	Number of Predictors	Predictors	BIC
1	1	x_3	508.9662
2	1	x_2	592.3601
3	1	x_1	611.3084
4	2	x_1, x_3	493.6383
5	2	x_2, x_3	496.5984
6	2	x_1, x_2	584.0046
7	3	x_1, x_2, x_3	495.3186

The model with the predictors x_1 and x_3 is chosen to be the best model w.r.t. *BIC*. The estimates of the parameters for the best model is given below.

term	estimate	std.error	statistic	p.value
(Intercept)	12.8068372	2.4187437	5.294830	1.60e-06
x1	-0.3407358	0.0732089	-4.654292	1.68e-05
x3	0.8945668	0.0491985	18.182804	0.00e+00

We repeat the same process using Mallow's C_p as the selection criterion. The C_p values for the model subsets are given below.

Index	Number of Predictors	Predictors	C_p
1	1	x_3	24.561778
2	1	x_2	241.000550
3	1	x_1	340.363948
4	2	x_1, x_3	4.419037
5	2	x_2, x_3	7.374082
6	2	x_1, x_2	191.033950
7	3	x_1, x_2, x_3	4.000000

The model with all the 3 predictors is chosen to be the best(for which, C_p is close to p) w.r.t. C_p. The estimates of the parameters for the best model is given below.

term	estimate	std.error	statistic	p.value
(Intercept)	15.8430423	3.0877581	5.130921	0.0000030
x1	-0.2326559	0.1003603	-2.318207	0.0236975
x2	-16.7241770	10.7528479	-1.555325	0.1248782
x3	0.8375671	0.0609186	13.748962	0.0000000

Comment: The criterion AIC and C_p prefers the full model and the criterion BIC prefers the model with x_1 and x_3 as predictors. But the decrease of BIC by switching from full model to the model with x_1 and x_3 as predictors, is not very much. Thus we choose the model with all the 3 predictors.

Step-wise Regression: At first we select the best subset by step-wise regression, using the value of F-statistic as the selection criterion.

```
ols_step_both_p(model.1, prem = 0.05, pent = 0.05, details = TRUE)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. x1
## 2. x2
## 3. x3
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - x3 added
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.885       RMSE                9.978
## R-Squared               0.783       Coef. Var          22.785
## Adj. R-Squared          0.780       MSE                99.551
## Pred R-Squared          0.769       MAE                 7.790
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ----------------------------------------------------------------------
##                  Sum of
##                 Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression    23399.615         1      23399.615    235.052    0.0000
## Residual       6470.811        65         99.551
## Total         29870.426        66
## ----------------------------------------------------------------------
##
##                                  Parameter Estimates
## ---------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
## ---------------------------------------------------------------------------------------
## (Intercept)     8.750         2.590                  3.378    0.001    3.577    13.923
##          x3     0.825         0.054        0.885    15.331    0.000    0.718     0.933
## ---------------------------------------------------------------------------------------
##
##
```

```
##
## Stepwise Selection: Step 2
##
## - x1 added
##
##                         Model Summary
## ---------------------------------------------------------------
## R                       0.916       RMSE                8.691
## R-Squared               0.838       Coef. Var          19.848
## Adj. R-Squared          0.833       MSE                75.538
## Pred R-Squared          0.820       MAE                 6.884
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## -----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## -----------------------------------------------------------------------
## Regression    25035.963        2      12517.981    165.717    0.0000
## Residual       4834.464       64         75.538
## Total         29870.426       66
## -----------------------------------------------------------------------
##
##                                   Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    12.807         2.419                  5.295    0.000     7.975    17.639
##          x3     0.895         0.049        0.960    18.183    0.000     0.796     0.993
##          x1    -0.341         0.073       -0.246    -4.654    0.000    -0.487    -0.194
## ----------------------------------------------------------------------------------------
##
##
##
##                         Model Summary
## ---------------------------------------------------------------
## R                       0.916       RMSE                8.691
## R-Squared               0.838       Coef. Var          19.848
## Adj. R-Squared          0.833       MSE                75.538
## Pred R-Squared          0.820       MAE                 6.884
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## -----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## -----------------------------------------------------------------------
## Regression    25035.963        2      12517.981    165.717    0.0000
```

```
## Residual       4834.464        64         75.538
## Total         29870.426        66
## ----------------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    12.807         2.419                  5.295    0.000     7.975    17.639
##          x3     0.895         0.049        0.960    18.183    0.000     0.796     0.993
##          x1    -0.341         0.073       -0.246    -4.654    0.000    -0.487    -0.194
## ----------------------------------------------------------------------------------------
##
##
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                       Model Summary
## --------------------------------------------------------------
## R                       0.916       RMSE                8.691
## R-Squared               0.838       Coef. Var          19.848
## Adj. R-Squared          0.833       MSE                75.538
## Pred R-Squared          0.820       MAE                 6.884
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression    25035.963        2      12517.981    165.717    0.0000
## Residual       4834.464       64         75.538
## Total         29870.426       66
## ----------------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    12.807         2.419                  5.295    0.000     7.975    17.639
##          x3     0.895         0.049        0.960    18.183    0.000     0.796     0.993
##          x1    -0.341         0.073       -0.246    -4.654    0.000    -0.487    -0.194
## ----------------------------------------------------------------------------------------
##
##                             Stepwise Selection Summary
## -------------------------------------------------------------------------------------
```

```
##                         Added/                    Adj.
## Step    Variable    Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ---------------------------------------------------------------------------------------
##    1        x3       addition       0.783       0.780    24.5620    502.3522    9.9775
##    2        x1       addition       0.838       0.833     4.4190    484.8196    8.6913
## ---------------------------------------------------------------------------------------
```

Comment: The model with predictors x_1 and x_3 is the final chosen model using F-statistic as the selection criterion.

We repeat the same process using AIC as the selection criterion.

```
ols_step_both_aic(model.1, details = TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1 . x1
## 2 . x2
## 3 . x3
##
##  Step 0: AIC = 602.8332
##  y ~ 1
##
##
## Variables Entered/Removed:
##
##                              Enter New Variables
## -------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## x3            1    502.352    23399.615    6470.811    0.783        0.780
## x2            1    585.746     7404.801   22465.625    0.248        0.236
## x1            1    604.694       61.851   29808.576    0.002       -0.013
## -------------------------------------------------------------------------
##
## - x3 added
##
##
##  Step 1 : AIC = 502.3522
##  y ~ x3
##
##                              Enter New Variables
## ------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## x1            1    484.820    25035.963    4834.464    0.838       0.833
## x2            1    487.780    24817.585    5052.841    0.831       0.826
## ------------------------------------------------------------------------
##
## - x1 added
##
##
##  Step 2 : AIC = 484.8196
```

```
##  y ~ x3 + x1
##
##                         Remove Existing Variables
## --------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq         RSS       R-Sq     Adj. R-Sq
## --------------------------------------------------------------------------
## x1            1    502.352    23399.615     6470.811    0.783         0.780
## x3            1    604.694       61.851    29808.576    0.002        -0.013
## --------------------------------------------------------------------------
##
##                             Enter New Variables
## -------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq        RSS       R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## x2            1    484.295    25214.729    4655.697    0.844         0.837
## -------------------------------------------------------------------------
##
## - x2 added
##
##
##  Step 3 : AIC = 484.2951
##  y ~ x3 + x1 + x2
##
##                         Remove Existing Variables
## --------------------------------------------------------------------------
## Variable     DF      AIC       Sum Sq         RSS       R-Sq     Adj. R-Sq
## --------------------------------------------------------------------------
## x2            1    484.820    25035.963     4834.464    0.838         0.833
## x1            1    487.780    24817.585     5052.841    0.831         0.826
## x3            1    575.186    11245.130    18625.296    0.376         0.357
## --------------------------------------------------------------------------
##
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.919       RMSE                8.597
## R-Squared               0.844       Coef. Var          19.632
## Adj. R-Squared          0.837       MSE                73.900
## Pred R-Squared          0.820       MAE                 6.833
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ----------------------------------------------------------------------
##                 Sum of
##                Squares        DF    Mean Square       F         Sig.
## ----------------------------------------------------------------------
## Regression    25214.729        3       8404.910    113.734    0.0000
```

```
## Residual       4655.697        63         73.900
## Total         29870.426        66
## ----------------------------------------------------------------------
##
##                                   Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model        Beta    Std. Error    Std. Beta       t        Sig        lower     upper
## ------------------------------------------------------------------------------------------
## (Intercept)     15.843         3.088                  5.131    0.000        9.673    22.013
##          x3      0.838         0.061        0.898    13.749    0.000        0.716     0.959
##          x1     -0.233         0.100       -0.168    -2.318    0.024       -0.433    -0.032
##          x2    -16.724        10.753       -0.114    -1.555    0.125      -38.212     4.764
## ------------------------------------------------------------------------------------------
```

```
##
##
##                                 Stepwise Summary
## --------------------------------------------------------------------------------
## Variable     Method       AIC        RSS        Sum Sq       R-Sq      Adj. R-Sq
## --------------------------------------------------------------------------------
## x3          addition    502.352    6470.811    23399.615    0.78337      0.78004
## x1          addition    484.820    4834.464    25035.963    0.83815      0.83309
## x2          addition    484.295    4655.697    25214.729    0.84414      0.83671
## --------------------------------------------------------------------------------
```

Comment: The model with all 3 predictors is the final chosen model using *AIC* as the selection criterion.

Considering the 2 selection criterion, we select the model with all the predictors as the final model.

Test for Normality: We apply *Shapiro-Wilks* test, for this purpose. Some other tests are also applied.

```
final_model <- lm(y ~ ., data)
ols_test_normality(final_model)
```

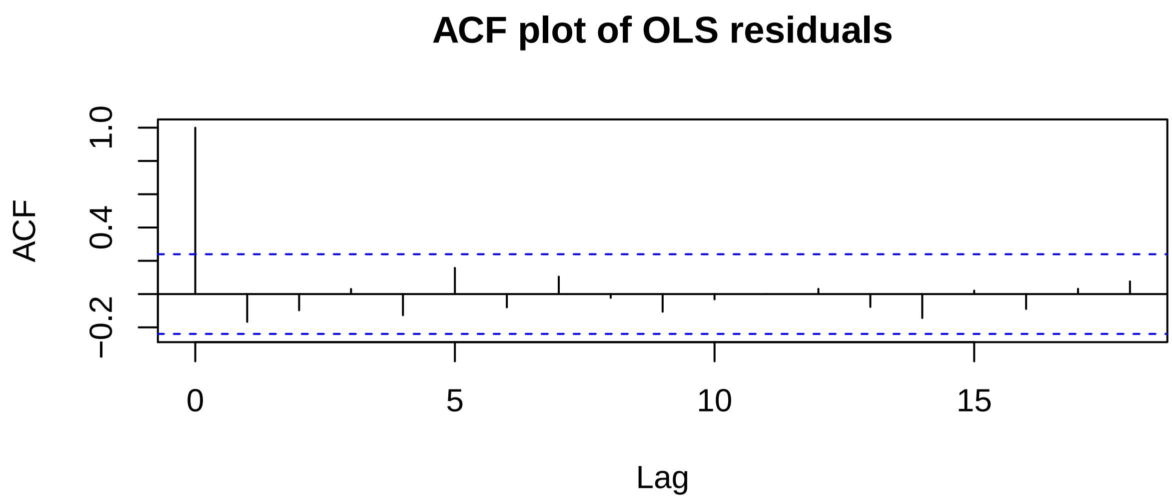
The results are tabulated below.

Test	Statistic	pvalue
Shapiro-Wilk	0.9777	0.2714
Kolmogorov-Smirnov	0.0995	0.4897
Anderson-Darling	0.6312	0.0959

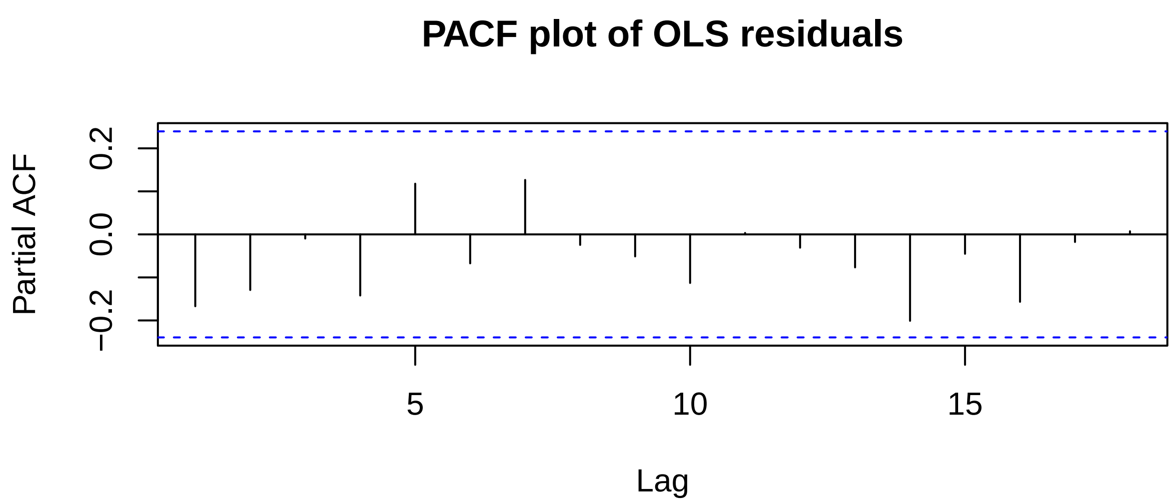
Comment: From the tests, we can say that the normality assumption is satisfied.

Autocorrelation: We first plot the *ACF* and *PACF* of the residuals. The plots are given below.

```
final_model.resi <- residuals(final_model)
acf(final_model.resi, main = "ACF plot of OLS residuals")
```



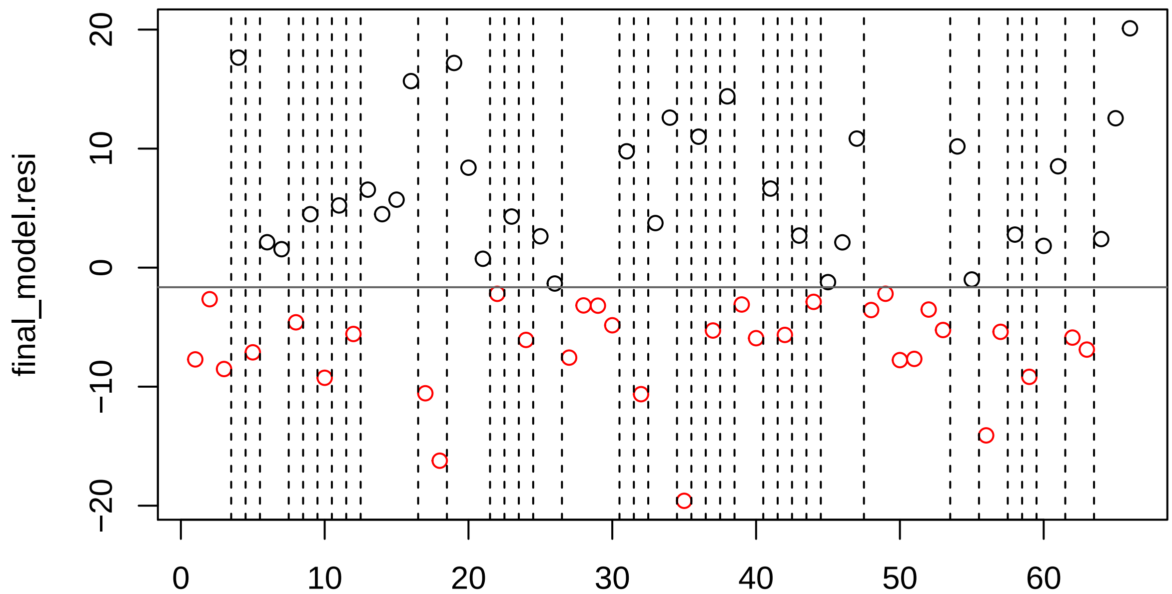

```
pacf(final_model.resi, main = "PACF plot of OLS residuals")
```



Comment: From the diagrams it seems that the residual series has no autocorrelation.

Runs Test: We can apply *Runs Test* to see if the residual series has auto correlation.

```
library(randtests)
runs.test(final_model.resi, plot = TRUE)
```




```
##
##  Runs Test
##
## data:  final_model.resi
## statistic = 0.9924, runs = 38, n1 = 33, n2 = 33, n = 66, p-value =
## 0.321
## alternative hypothesis: nonrandomness
```

Comment: From the *Runs Test*, we can say that the residuals are not auto-correlated.

Durbin-Watson Test: We also apply *Durbin-Watson* test to see if the errors(ϵ_i) follow a first-order autoregressive(*AR(1)*) model.

```
suppressMessages(library(lmtest))
dwtest(final_model, alternative = "two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  final_model
## DW = 2.2344, p-value = 0.3357
## alternative hypothesis: true autocorrelation is not 0
```

Comment: From the *Durbin-Watson* test, we can say the error(ϵ_i) doesn't follow an *AR(1)* model.

Heteroscedasticity: For this purpose, we first apply *F-test*, considering the square OLS residuals($\hat{\epsilon}_i^2$) as the response and the other variables as the predictors.

```r
summary(lm(I(final_model.resi^2) ~ x1 + x2 + x3, data))
```

```
##
## Call:
## lm(formula = I(final_model.resi^2) ~ x1 + x2 + x3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -102.38  -50.08  -34.55   26.59  357.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.7738    32.9293   0.388   0.6994
## x1           -0.5868     1.0703  -0.548   0.5855
## x2           69.0908   114.6735   0.603   0.5490
## x3            1.3435     0.6497   2.068   0.0427 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91.68 on 63 degrees of freedom
## Multiple R-squared:  0.07572,    Adjusted R-squared:  0.03171
## F-statistic:  1.72 on 3 and 63 DF,  p-value: 0.1718
```

Comment: From the test, we can see that the regression is not significant. So, the error variance is not related to linear combination of independent variables.

Breusch-Pagan Test: We also apply *BP* test to conclude on the linear dependence of error variance on the independent variables.

```r
library(skedastic)
final_model.breusch <- breusch_pagan(mainlm = final_model,
                                     koenker = FALSE,
                                     statonly = FALSE)
```

statistic	p.value	parameter	method	alternative
4.491971	0.2130076	3	Breusch-Pagan (non-studentised)	greater

Comment: From the above test, it is confirmed that the error variance is not related to linear combination of independent variables.

White's Test: Here we try find the relationship of the error variance, with all independent variables, the squares of independent variables and all the cross products.

```r
final_model.white <- white_lm(mainlm = final_model,
                              interactions = TRUE,
                              statonly = FALSE)
```

statistic	p.value	parameter	method	alternative
8.268502	0.5073307	9	White's Test	greater

Comment: From the above test, it is confirmed that the that the error variance doesn't depend on 2^{nd} degree polynomials of the independent variables.

Glejser Test: Here we use *Glejser* test to test for heteroscedasticity in the model.

```r
final_model.glejser <- glejser(mainlm = final_model,
                               statonly = FALSE)
```

statistic	p.value	parameter	alternative
5.54939	0.1357142	3	greater

Comment: From the above test, we can say that heteroscedasticity is not present in the model.

Goldfeld-Quandt Test: Here we apply *GQ* test to test for heteroscedasticity in the model.

```r
final_model.goldfeld <- goldfeld_quandt(mainlm = final_model,
                                        statonly = FALSE,
                                        prop_central = 1/3,
                                        group1prop = 1/2)
```

statistic	p.value	parameter	method	alternative
0.9106025	0.5770228	18	Goldfeld-Quandt F Test	greater

Comment: From the above test, we confirm that heteroscedasticity is not present in the model.

Collinearity: For this, we calculate the *Variance Inflation Factor* corresponding to each estimator of coefficient of the predictors. They are given below.

```r
vif(final_model)
```

Predictors	VIF
x1	2.115710
x2	2.156007
x3	1.726061

Comment: None of the *VIF*'s are more than 10.

Now, we calculate the condition number corresponding to the selected model.

```r
kappa(final_model)
```

```
## [1] 618.4017
```

The condition number of the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$ is given below.

```r
kappa(lm(y ~ x1 + x3, data))
```

```
## [1] 97.11836
```

Comment: The condition number of the regression matrix is very high, compared to the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$. We should obtain *Ridge Estimates* in this case. Otherwise, we can also work with the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$.

Ridge Regression: As, we saw earlier, our regression matrix has collinearity, so the usual *LS* estimates won't be efficient. Thus we opt for *Ridge Regression*.

```r
library(ridge)
ridge_model <- linearRidge(y ~ ., data)
summary(ridge_model)
```

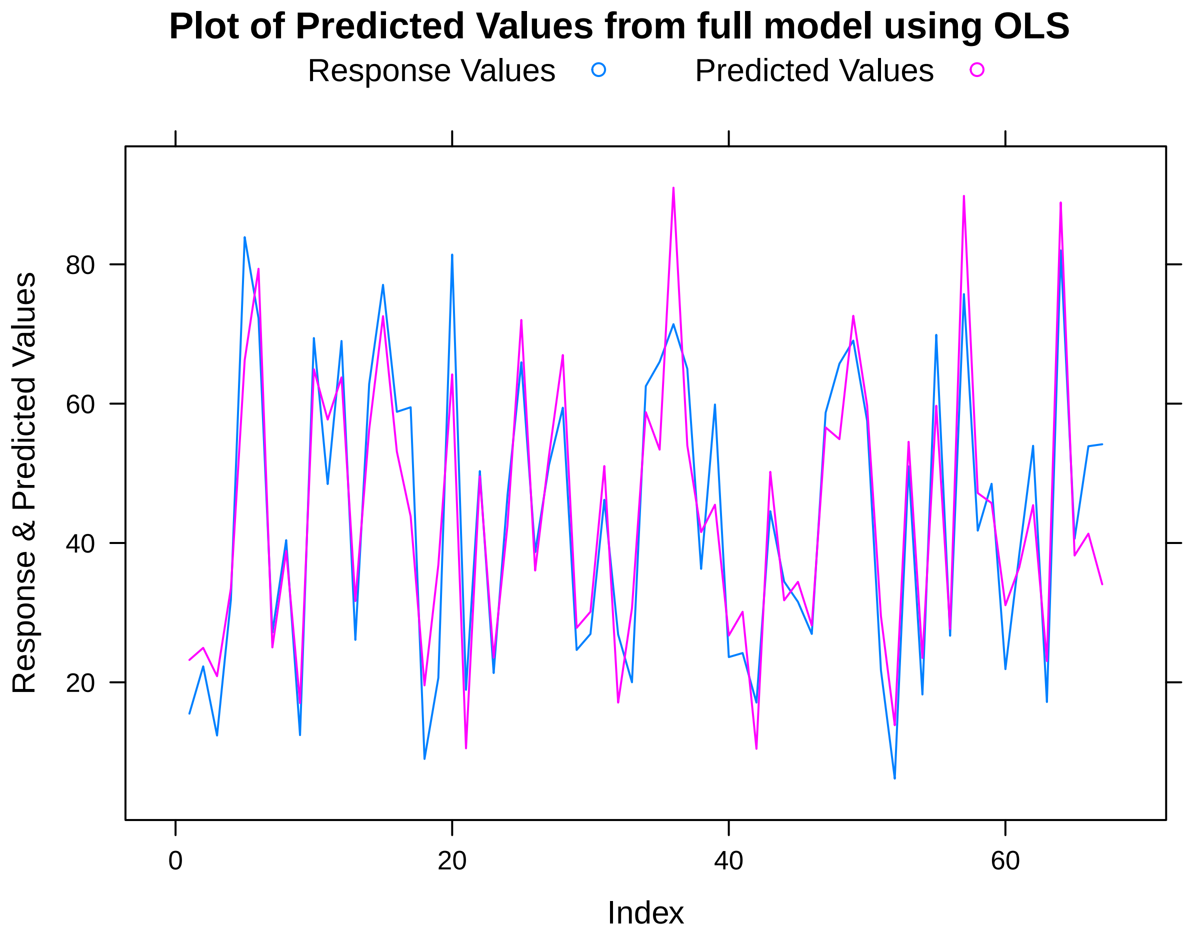
```
##
## Call:
## linearRidge(formula = y ~ ., data = data)
##
##
## Coefficients:
##              Estimate Scaled estimate Std. Error (scaled) t value (scaled)
## (Intercept)  23.24493              NA                  NA               NA
## x1           -0.06094        -7.59288             8.01409            0.947
## x2          -30.17868       -35.42614             8.03180            4.411
## x3            0.63388       117.51969             7.84075           14.988
##             Pr(>|t|)
## (Intercept)       NA
## x1             0.343
## x2          1.03e-05 ***
## x3           < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge parameter: 0.221719, chosen automatically, computed using 1 PCs
##
## Degrees of freedom: model 2.221 , variance 1.733 , residual 2.709
```

Comment: In the above section we see the final model, with the *ridge* estimates. We can also work with the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$ as the condition number for this model is 97.11836, which is very small with respect to that of the full model. Also, the decrease in AIC by including x_2 in the set of predictors, is not very significant. In the next section, we plot the fitted values from the two models, along with the response values.

Plots of Predicted Values: The plot of the predicted values from the full model along with the response values is given below.

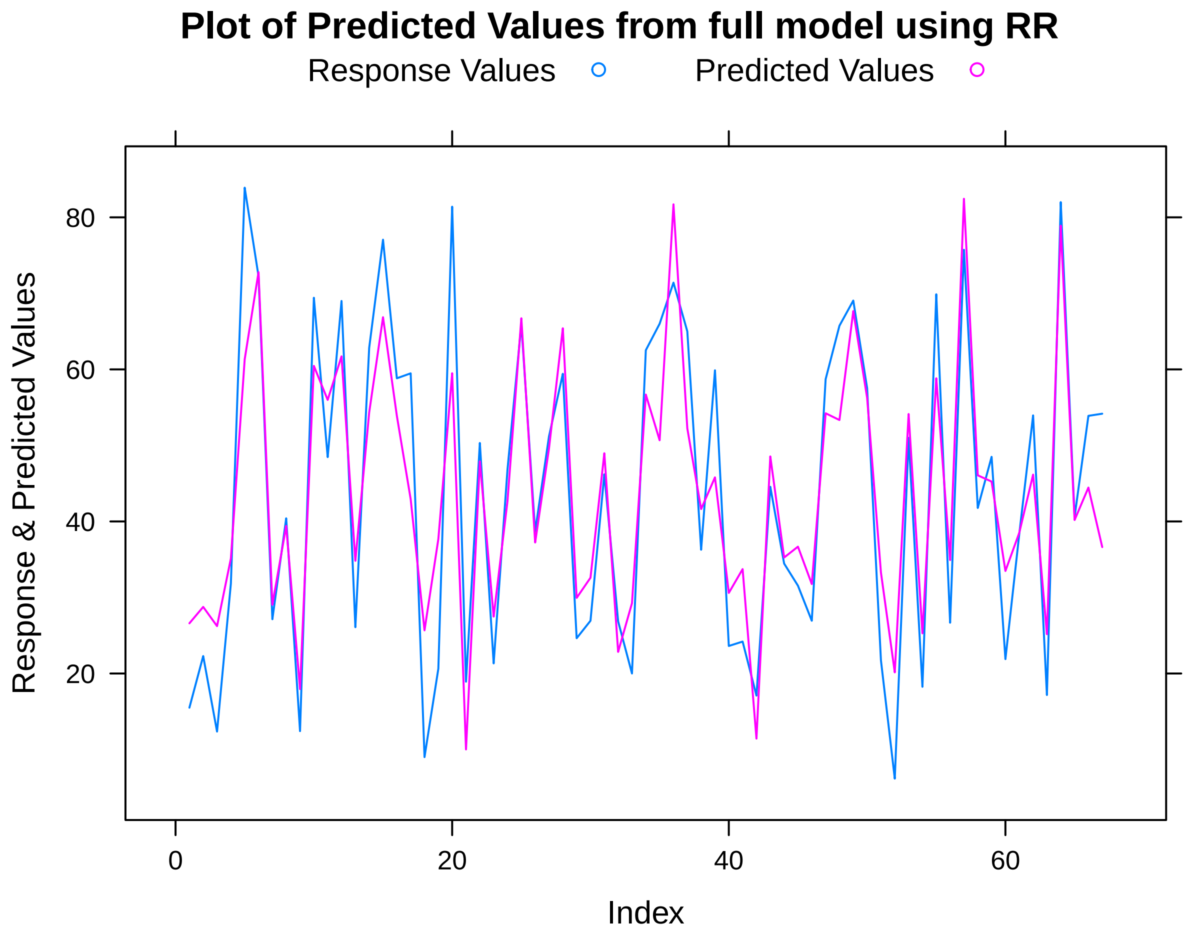
```
library(lattice)
library(latex2exp)
model.1.pred <- predict(final_model)
model.1.ridge.pred <- predict(ridge_model)
model.2.pred <- predict(lm(y ~ x1 + x3, data))

xyplot(y + model.1.pred ~ 1:67,
       data = data,
       type = c("l"),
       ylab = "Response & Predicted Values",
       xlab = "Index",
       main = list(
         "Plot of Predicted Values from full model using OLS",
         cex = 1.2
       ),
       auto.key = list(
         space = "top",
         columns = 2,
         text = c(
           "Response Values",
           "Predicted Values"
         )
       ))
```



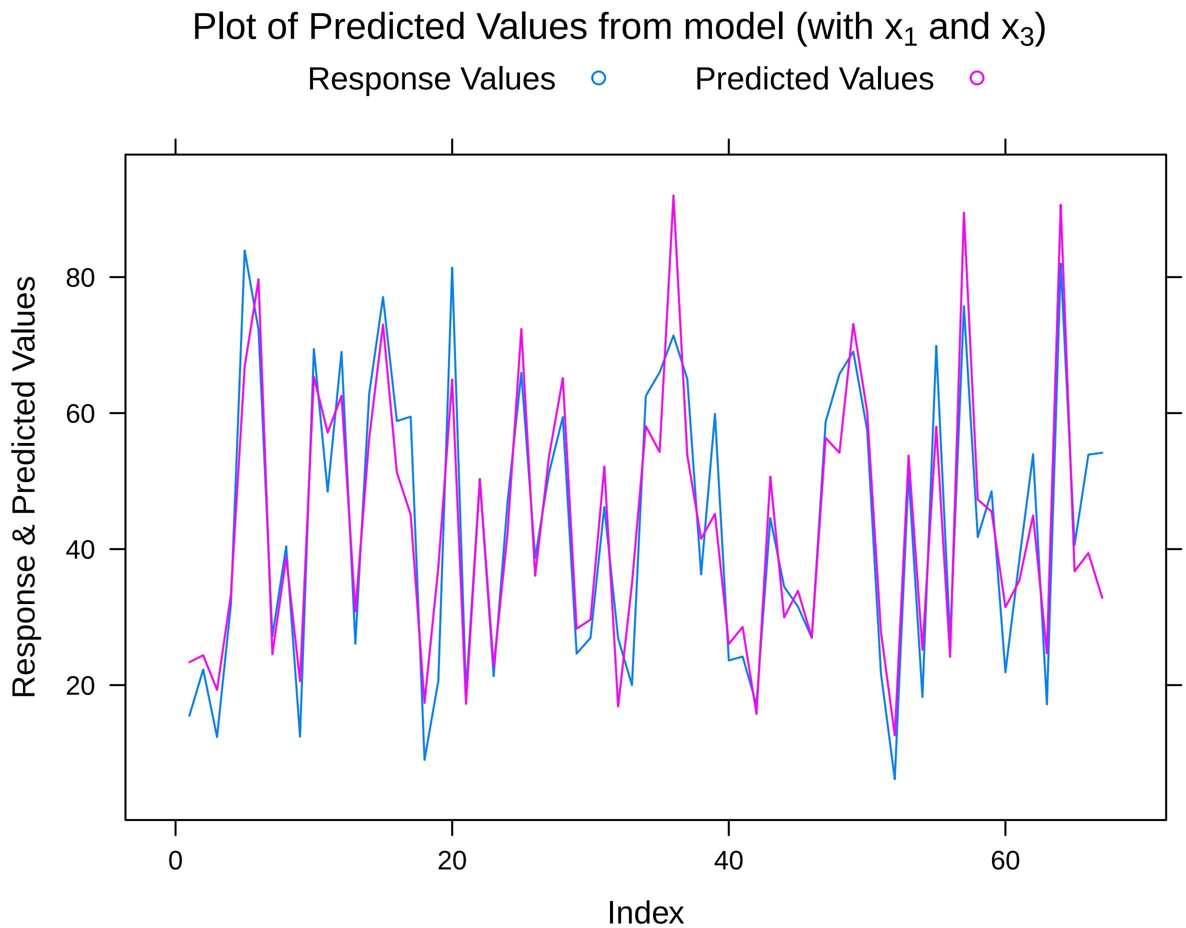
The plot of the predicted values from the full model using *Ridge* estimation, along with the response values is given below.

```
xyplot(y + model.1.ridge.pred ~ 1:67,
       data = data,
       type = c("l"),
       ylab = "Response & Predicted Values",
       xlab = "Index",
       main = list(
         "Plot of Predicted Values from full model using RR",
         cex = 1.2
       ),
       auto.key = list(
         space = "top",
         columns = 2,
         text = c(
           "Response Values",
           "Predicted Values"
         )
       ))
```



The plot of the predicted values from the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$, along with the response values is given below.

```
xyplot(y + model.2.pred ~ 1:67,
       data = data,
       type = c("l"),
       ylab = "Response & Predicted Values",
       xlab = "Index",
       main = list(
         TeX(
           "Plot of Predicted Values from model (with $x_1$ and $x_3$)"
         ),
         cex = 1.2
       ),
       auto.key = list(
         space = "top",
         columns = 2,
         text = c(
           "Response Values",
           "Predicted Values"
         )
       ))
```



Comment: We can't see any significant difference between 1st graph and 3rd graph.

The estimates of the model parameters for the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$ is given below.

term	estimate	std.error	statistic	p.value
(Intercept)	12.8068372	2.4187437	5.294830	1.60e-06
x1	-0.3407358	0.0732089	-4.654292	1.68e-05
x3	0.8945668	0.0491985	18.182804	0.00e+00

The value of R^2 and $Adj.\ R^2$, for the 2 models is tabulated below.

Predictors	R^2	$Adj.\ R^2$
x_1, x_2, x_3	0.8441369	0.8367149
x_1, x_3	0.8381522	0.8330944

Comment: From the above table we can see that nothing has been gained by switching from the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$ to the full model. In the next section we analyze the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$ for outliers, normality, heteroscedasticity and autocorrelation and collinearity.

Analysis of the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$

Outliers and Influential Observations: We calculate the measures for detecting outliers and influential points corresponding to each residuals, viz, DFFITS, Covariance Ratio, Cook's D and hat matrix diagonals.

```r
model.2 <- lm(y ~ x1 + x3, data)
influence.measures(model.2)
```

They are tabulated below.

Index	DFFITS	Cov-Ratio	Cook's D	h
1	-0.1671197	1.0405716	0.0093322	0.0319851
2	-0.0428975	1.0778233	0.0006225	0.0300631
3	-0.1737361	1.0621607	0.0101146	0.0435141
4	-0.0310206	1.0828445	0.0003257	0.0331226
5	0.4873836	0.9094046	0.0753396	0.0527792
6	-0.2353748	1.0824913	0.0185315	0.0665075
7	0.0602740	1.0851620	0.0012284	0.0381033
8	0.0347187	1.0818870	0.0004079	0.0326196
9	-0.2064142	1.0501199	0.0142198	0.0442042
10	0.1068603	1.0888542	0.0038528	0.0475774
11	-0.1607919	1.0240902	0.0086148	0.0246246
12	0.1441860	1.0579649	0.0069770	0.0353535
13	-0.0901804	1.0613052	0.0027407	0.0262405
14	0.1253613	1.0473877	0.0052729	0.0262910
15	0.1123692	1.0957141	0.0042606	0.0533386
16	0.2178052	1.0698267	0.0158637	0.0562621
17	0.3006959	0.9437274	0.0292636	0.0300822
18	-0.2271754	1.0537836	0.0172086	0.0500905
19	-0.2636764	0.8897108	0.0221599	0.0173743
20	0.4617496	0.9203043	0.0679317	0.0511213
21	0.0446799	1.1011306	0.0006756	0.0496376
22	0.0009418	1.0660898	0.0000003	0.0166154
23	-0.0374338	1.1071194	0.0004743	0.0541597
24	0.0930899	1.0660499	0.0029213	0.0297109
25	-0.1688101	1.0713550	0.0095631	0.0475623
26	0.0421665	1.0645639	0.0006012	0.0194049
27	-0.0495609	1.0770982	0.0008307	0.0302982
28	-0.2097404	1.1210206	0.0147859	0.0854560
29	-0.0733722	1.0707404	0.0018178	0.0291860
30	-0.0663897	1.0901018	0.0014902	0.0427771
31	-0.1104837	1.0515326	0.0041026	0.0251389
32	0.4534087	1.1087692	0.0679799	0.1195098
33	-0.3816324	0.9459429	0.0469502	0.0438232
34	0.0819824	1.0617211	0.0022664	0.0248600
35	0.2070834	0.9814593	0.0140995	0.0222419
36	-0.9113099	0.8614560	0.2535719	0.1078507
37	0.2005964	0.9890897	0.0132619	0.0227441
38	-0.1051864	1.0607644	0.0037247	0.0289427
39	0.2780118	0.9338905	0.0249721	0.0248879
40	-0.0517877	1.0798106	0.0009070	0.0327603
41	-0.0984508	1.0746104	0.0032688	0.0362757
42	0.0391064	1.1125006	0.0005177	0.0587363
43	-0.0951902	1.0426360	0.0030443	0.0179110
44	0.1052788	1.0764496	0.0037369	0.0387259
45	-0.0380009	1.0655988	0.0004884	0.0195457
46	-0.0009355	1.0863297	0.0000003	0.0349360

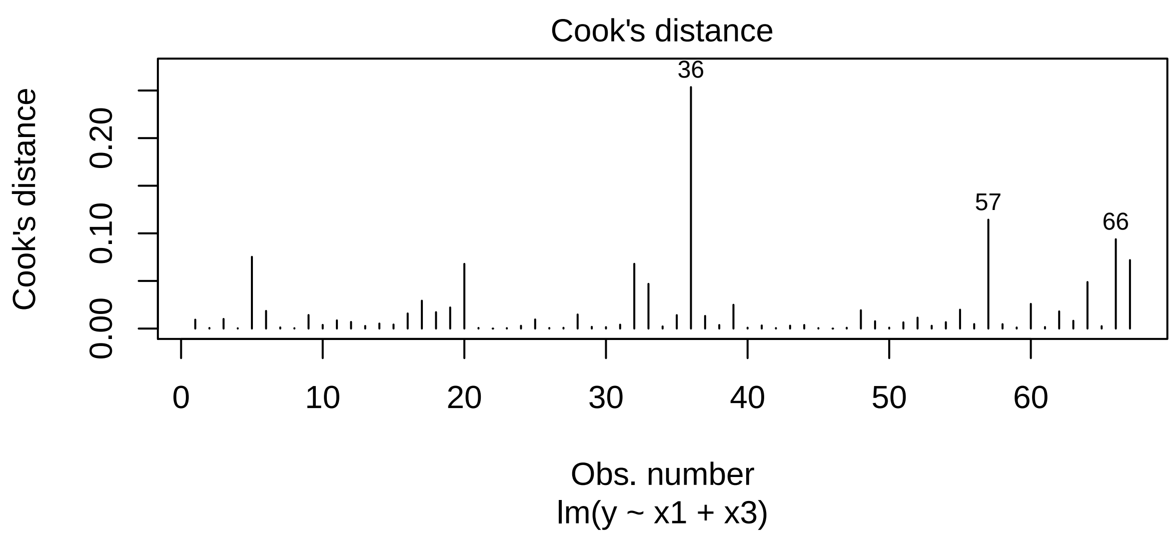
Index	DFFITS	Cov-Ratio	Cook's D	h
47	0.0507447	1.0806099	0.0008709	0.0332747
48	0.2418233	0.9913361	0.0192362	0.0305800
49	-0.1509596	1.1366092	0.0076880	0.0879308
50	-0.0539200	1.0737688	0.0009829	0.0282762
51	-0.1394311	1.0620976	0.0065302	0.0365907
52	-0.1850171	1.0802735	0.0114860	0.0558053
53	-0.0930952	1.1279790	0.0029298	0.0752656
54	-0.1415334	1.0472896	0.0067133	0.0296113
55	0.2457068	0.9859060	0.0198272	0.0299140
56	0.1183553	1.1962256	0.0047362	0.1275798
57	-0.5942371	1.0292887	0.1143483	0.1092460
58	-0.1169643	1.0624884	0.0046027	0.0322938
59	0.0578488	1.0717915	0.0011311	0.0273640
60	-0.2795539	1.0460651	0.0259317	0.0570208
61	0.0689302	1.0827208	0.0016058	0.0373437
62	0.2330236	1.0421404	0.0180647	0.0460409
63	-0.1563684	1.0427656	0.0081795	0.0306950
64	-0.3832022	1.1252095	0.0488583	0.1161501
65	0.0863508	1.0767058	0.0025169	0.0356117
66	0.5392811	0.9907092	0.0938197	0.0850280
67	0.4851187	0.7973222	0.0719225	0.0334247

The cutoff values for the measures are given below.

DFFITS > 0.4232074
Covariance Ratio $\in (-\infty, 0.8656716) \cup (1.134328, \infty)$
Cook's D > 1
Hat-Matrix Diagonals > 0.08955224

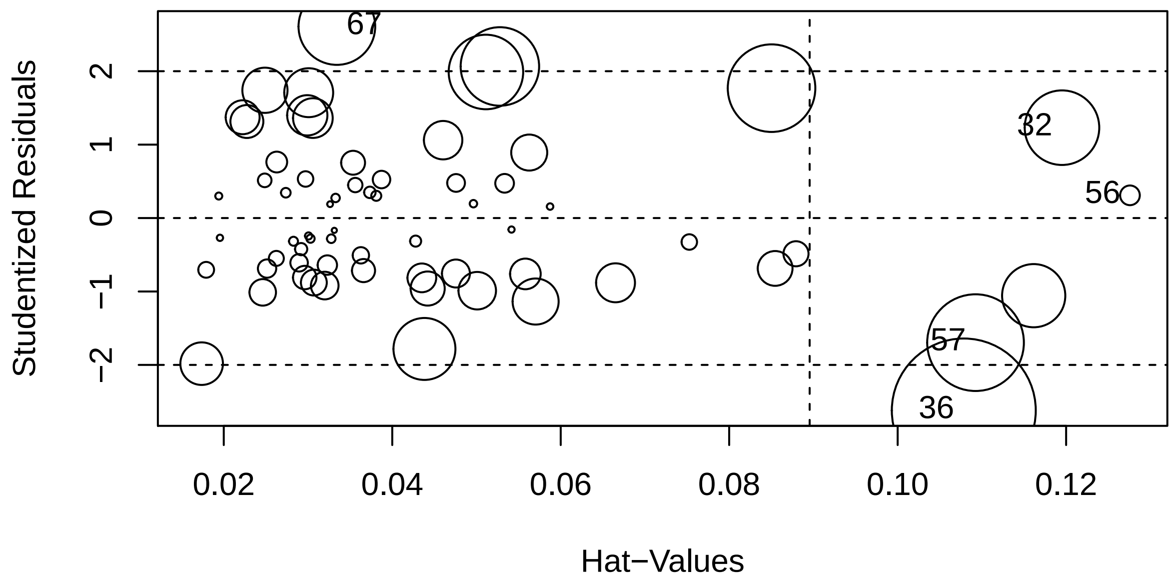
Cooks distances are plotted below.

```
plot(model.2, which = 4, cook.levels = 1)
```



The plot of standardized residuals and hat-matrix diagonals are given below.

```
influencePlot(model.2)
```




```
##        StudRes        Hat      CookD
## 32  1.2306950 0.11950976 0.06797987
## 36 -2.6210384 0.10785069 0.25357190
## 56  0.3094992 0.12757982 0.00473624
## 57 -1.6968205 0.10924599 0.11434835
## 67  2.6087497 0.03342465 0.07192248
```

The suspicious points are given below.

```
influ.model.2 <- influence.measures(model.1)
summary(influ.model.2)
```

```
## Potentially influential observations of
##   lm(formula = y ~ x1 + x2 + x3, data = data) :
##
##    dfb.1_ dfb.x1 dfb.x2 dfb.x3 dffit   cov.r   cook.d hat
## 21 -0.31  -0.37   0.71   0.23   0.77_*  1.40_*  0.15   0.30_*
## 36  0.45   0.31  -0.20  -0.77  -0.90_*  0.81    0.19   0.11
## 42 -0.16  -0.16   0.39   0.09   0.46    1.30_*  0.05   0.22_*
## 56 -0.02  -0.06   0.04   0.04  -0.06    1.33_*  0.00   0.20_*
## 67  0.48  -0.01  -0.24  -0.27   0.52    0.76_*  0.06   0.04
```

For these points we apply test for outliers using outlier-shift model.

```
outlierTest(model.2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
```

```
## 36 -2.621038             0.010975       0.73532
```

The points with hat-values more than 0.08955224 are given by,

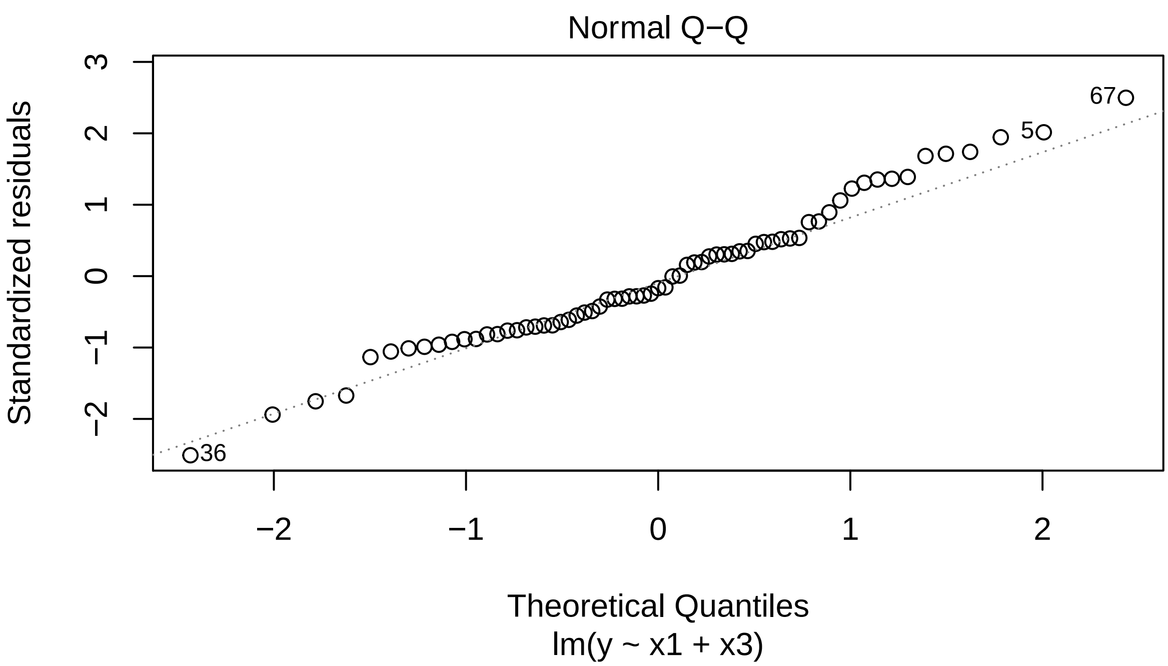
```
which(as.vector(hatvalues(model.1)) > 0.08955224)
```

```
##  [1]  9 21 28 32 33 36 42 56 57 64 66
```

Comment: From the test for outliers, we see that there are no outliers. We also see there are many leverage points. So, we can try with *Robust Regression*, which results in residuals that better identify the outlier.

Test for Normality: The *QQ-Plot* of the residuals from the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$ is given below.

```
plot(model.2, which = 2)
```



Comment: From the QQ-Plot, it seems that the residuals are almost normal. We apply *Shapiro-Wilks* test, for testing normality of the error distribution. Some other tests are also applied.

```
ols_test_normality(model.2)
```

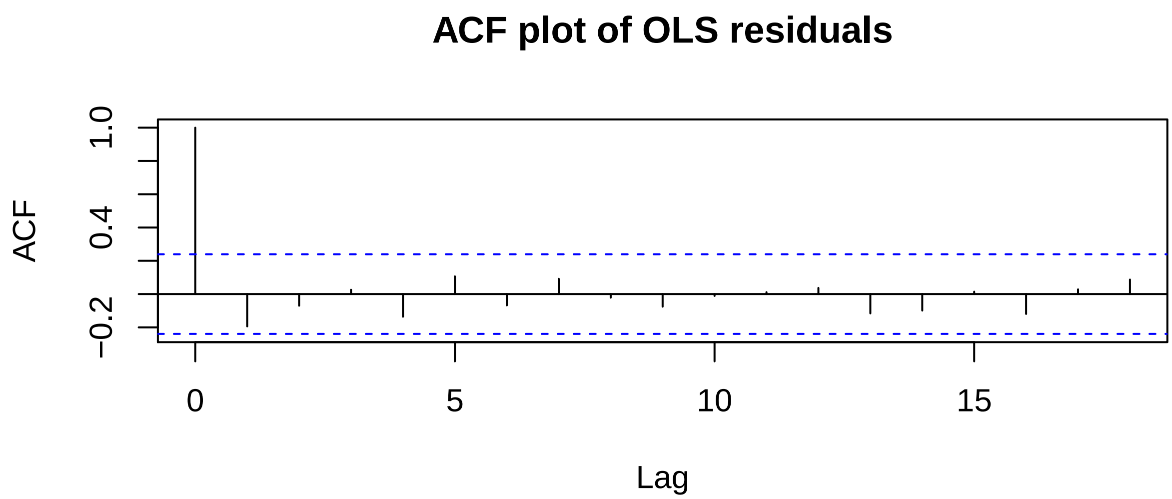
The results are tabulated below.

Test	Statistic	pvalue
Shapiro-Wilk	0.9792	0.3227
Kolmogorov-Smirnov	0.0895	0.6240
Anderson-Darling	0.6105	0.1080

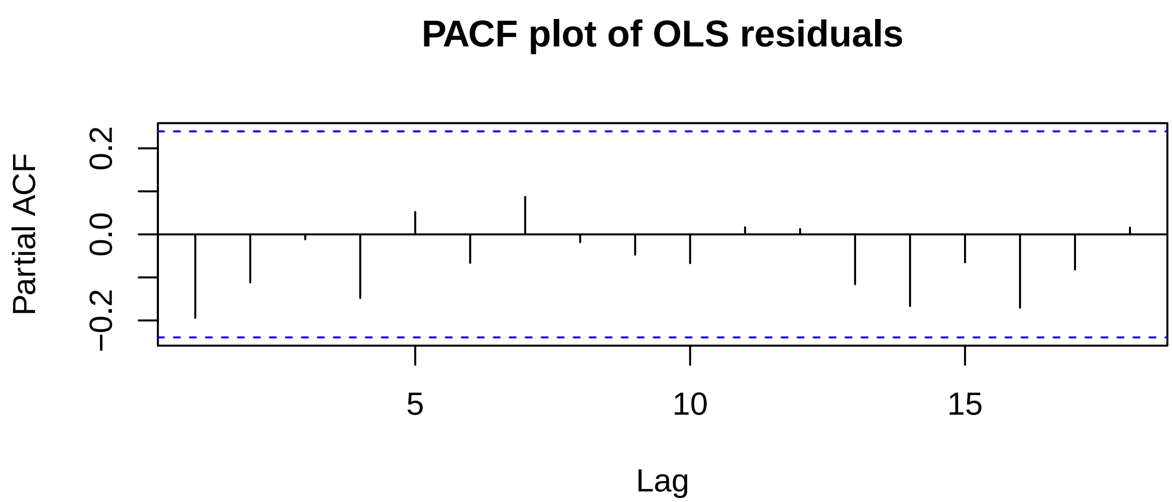
Comment: From the tests, we can say that the normality assumption is satisfied.

Autocorrelation: We first plot the *ACF* and *PACF* of the residuals. The plots are given below.

```
model.2.resi <- residuals(model.2)
acf(model.2.resi, main = "ACF plot of OLS residuals")
```



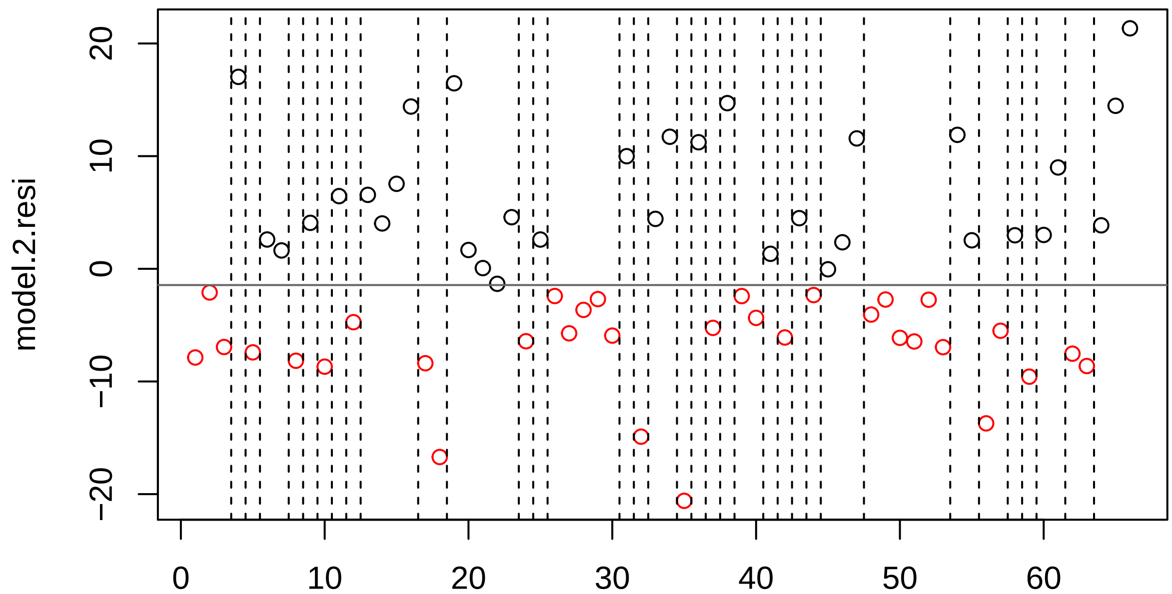

```
pacf(model.2.resi, main = "PACF plot of OLS residuals")
```



Comment: From the diagrams it seems that the residual series has no autocorrelation.

Runs Test: We can apply *Runs Test* to see if the residual series has auto correlation.

```
runs.test(model.2.resi, plot = TRUE)
```




```
##
##  Runs Test
##
## data:  model.2.resi
## statistic = 0.4962, runs = 36, n1 = 33, n2 = 33, n = 66, p-value =
## 0.6198
## alternative hypothesis: nonrandomness
```

Comment: From the *Runs Test*, we can say that the residuals are not auto-correlated.

Durbin-Watson Test: We also apply *Durbin-Watson* test to see if the errors follow a first-order autoregressive($AR(1)$) model.

```
dwtest(model.2, alternative = "two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  model.2
## DW = 2.2813, p-value = 0.2451
## alternative hypothesis: true autocorrelation is not 0
```

Comment: From the *Durbin-Watson* test, we can say the error doesn't follow an $AR(1)$ model.

Heteroscedasticity: For this purpose, we first apply *F-test*, considering the square OLS residuals as the response and the other variables as the predictors.

```
summary(lm(I(model.2.resi^2) ~ x1 + x3, data))
```

```
##
```

```
## Call:
## lm(formula = I(model.2.resi^2) ~ x1 + x3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -108.72  -54.54  -36.63   29.19  405.79
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.156488  26.781007   0.715   0.4770
## x1          -0.008448   0.810590  -0.010   0.9917
## x3           1.252133   0.544740   2.299   0.0248 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.23 on 64 degrees of freedom
## Multiple R-squared:  0.08314,    Adjusted R-squared:  0.05449
## F-statistic: 2.902 on 2 and 64 DF,  p-value: 0.06219
```

Comment: From the test, we can see that the regression is not significant. So, the error variance is not related to linear combination of independent variables.

Breusch-Pagan Test: We also apply *BP* test to conclude on the linear dependence of error variance on the independent variables.

```
library(skedastic)
model.2.breusch <- breusch_pagan(mainlm = model.2,
                                 koenker = FALSE,
                                 statonly = FALSE)
```

statistic	p.value	parameter	method	alternative
5.16104	0.0757346	2	Breusch-Pagan (non-studentised)	greater

Comment: From the above test, it is confirmed that the error variance is not related to linear combination of independent variables.

White's Test: Here we try find the relationship of the error variance, with the independent variables, the squares of independent variables and the cross products.

```
model.2.white <- white_lm(mainlm = model.2,
                          interactions = TRUE,
                          statonly = FALSE)
```

statistic	p.value	parameter	method	alternative
6.989448	0.2214263	5	White's Test	greater

Comment: From the above test, it is confirmed that the that the error variance doesn't depend on 2^{nd} degree polynomials of the independent variables.

Goldfeld-Quandt Test: Here we apply *GQ* test to test for heteroscedasticity in the model.

```
model.2.goldfeld <- goldfeld_quandt(mainlm = model.2,
                                    statonly = FALSE,
                                    prop_central = 1/3,
                                    group1prop = 1/2)
```

statistic	p.value	parameter	method	alternative
1.205794	0.3402705	19	Goldfeld-Quandt F Test	greater

Comment: From the above test, we confirm that heteroscedasticity is not present in the model.

Collinearity: For this, we calculate the *Variance Inflation Factor* corresponding to each estimator of coefficient of the predictors. They are given below.

```
vif(model.2)
```

Predictors	VIF
x1	1.101378
x3	1.101378

Comment: None of the *VIF*'s are more than 10.

The condition number for the model is 97.11836(*already calculated above*).

Comment: From the *VIF* values and the *Condition Number*, it seems that collinearity is not present in the model.

Conclusion: The final model is $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$. The estimates of the model parameters are given below.

term	estimate	std.error	statistic	p.value
(Intercept)	12.8068372	2.4187437	5.294830	1.60e-06
x1	-0.3407358	0.0732089	-4.654292	1.68e-05
x3	0.8945668	0.0491985	18.182804	0.00e+00