DEATH RATE PREDICTION

Math 644 Regression Analysis Models
Final Project Fall 2021

Sanjana Guntumadugu

Yaksh Patel

INTRODUCTION

The total age adjusted mortality rate, our response variable in each regression equation, can be obtained for the years 1959-1961 for 201 Standard Metropolitan Statistical Areas (SMSA). The age-adjusted death rates are given for the categories male white, female white, male non-white and female non-white.

Previous workers, e.g., Glasser and Greenburg [1], Holland, et al [2], and Oechsli and Buechley [3], have found climate or weather variables account for some of the variation in disease rates. Precipitation, mean January temperature, mean July temperature and household size, schooling and population per square mile, poor families have been included in the present study. The pollution potential of three pollutants, namely HC, NO, SO, have been estimated by Benedict [4]. The pollution potential is determined as the product of the tons emitted per day per square kilometer of each pollutant and a dispersion factor which accounts for mixing height, wind speed, number of episode days and dimension of each SMSA. These factors included in the dataset account for Mortality rate.

OBJECTIVE

In our project, we would like to predict the people's death rate. We acquired the data from people.sc.fsu.edu.. This data has 60 observations and

V1, the average annual precipitation;
V2, the average January temperature;
V3, the average July temperature;
V4, the size of the population older than 65;
V5, the number of members per household;
V6, the number of years of schooling for persons over 22;
V7, the number of households with fully equipped kitchens;
V8, the population per square mile;
V9, the size of the nonwhite population;
V10, the number of office workers;
V11, the number of families with an income less than $3000;
V12, the hydrocarbon pollution index;
V13, the nitric oxide pollution index;
V14, the sulfur dioxide pollution index;
V15, the degree of atmospheric moisture.
V16, the death rate.

This method can predict the death rate based on the above factors listed. We would like to identify the main factors and their relationships to the response to the Death rate. We want to find a model which can be the best prediction of the real measurement and also be economically efficient in practice.

V16 = V1 * X1 + V2 * X2 + V3 * X3 + V4 * X4 + V5 * X5 + V6 * X6 + V7 * X7 + V8 * X8 + V9 * X9 + V10 * X10 + V11 * X11 + V12 * X12 + V13 * X13 + V14 * X14 + V15 * X15

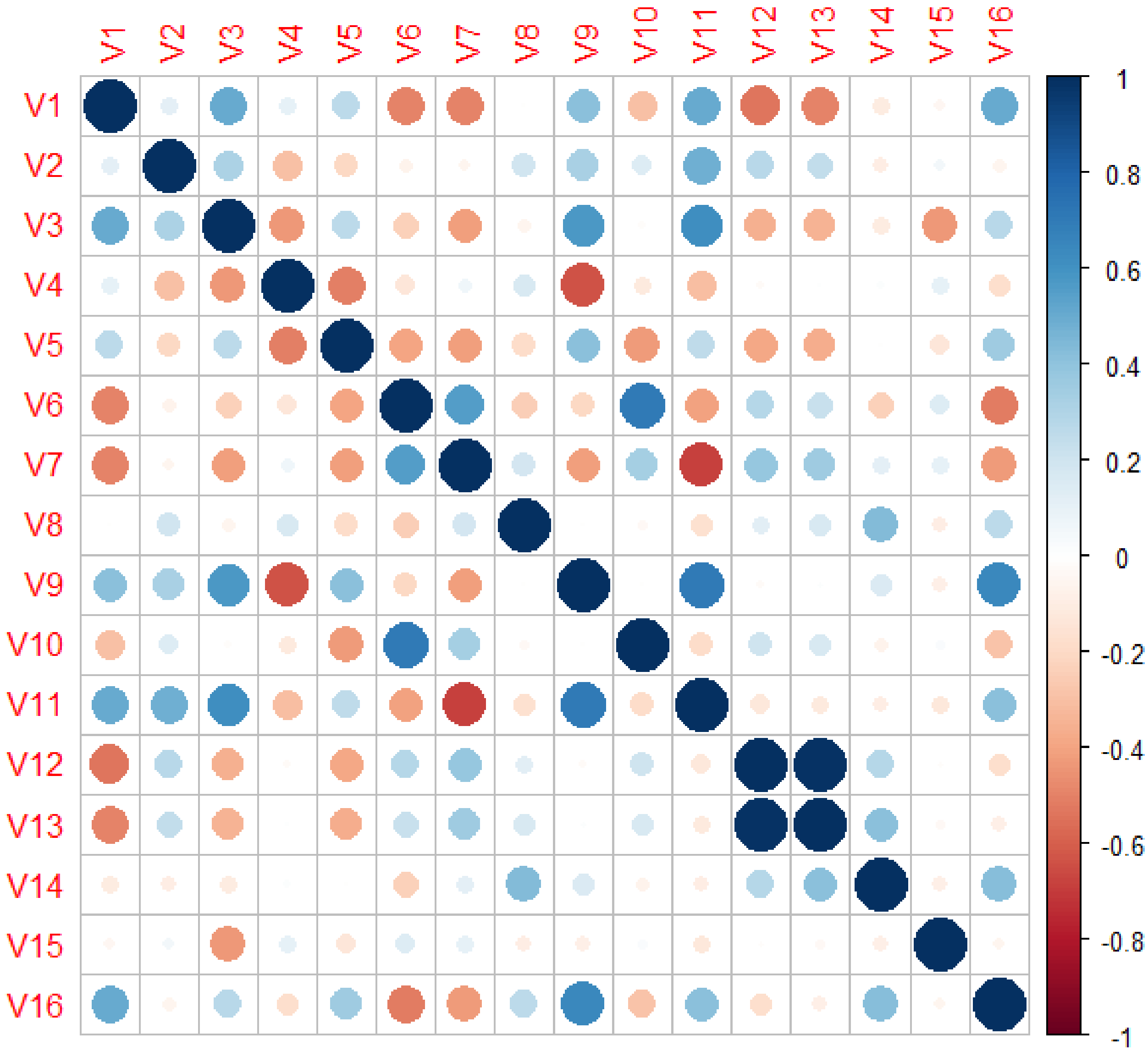
STATISTICAL ANALYSIS

- **Assumption** - Number of observations must be greater than number of Xs. This data satisfies this assumption as it has 60 observations and 15 variables.

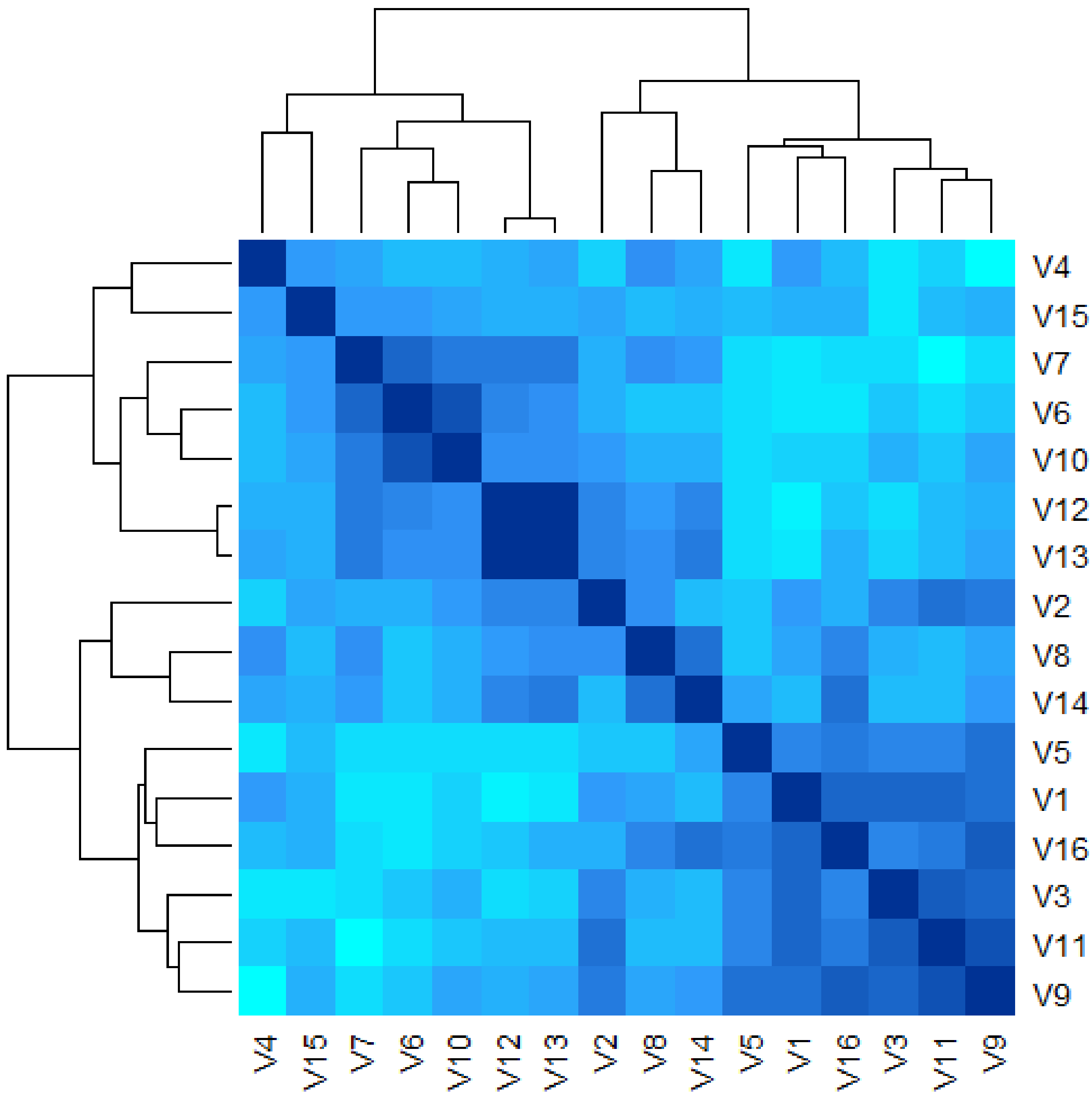
- **Description**

Variables	Mean	Median	SE. mean	Variance	Std. Deviation	Coef. Variance
V1	37.367	38.00	1.289	99.694	9.985	0.267
V2	34.817	31.500	1.546	143.406	11.975	0.344
V3	74.600	74.000	0.6153	22.7186	4.7664	0.0639
V4	8.798	9.000	0.189	2.145	1.465	0.1666
V5	3.2632	3.265	0.0175	0.0183	0.1353	0.0414
V6	10.973	11.050	0.109	0.715	0.845	0.077
V7	80.913	810150	0.663	26.433	5.141	0.0635
V8	3.88	3.57	1.88	2.11	1.45	3.75
V9	11.873	10.400	1.152	79.564	8.920	0.751
V10	46.073	45.500	0.597	21.408	4.627	0.100
V11	14.373	13.200	0.537	17.306	4.160	0.100
V12	37.85	14.50	11.87	8459.89	91.98	2.43
V13	22.52	9.00	5.98	2149.10	46.36	2.06
V14	53.77	30.00	8.18	4018.35	63.39	1.18
V15	57.533	57.00	0.704	29.8124	5.460	0.0949
V16	9.40	9.44	8.03	3.87	6.22	6.62

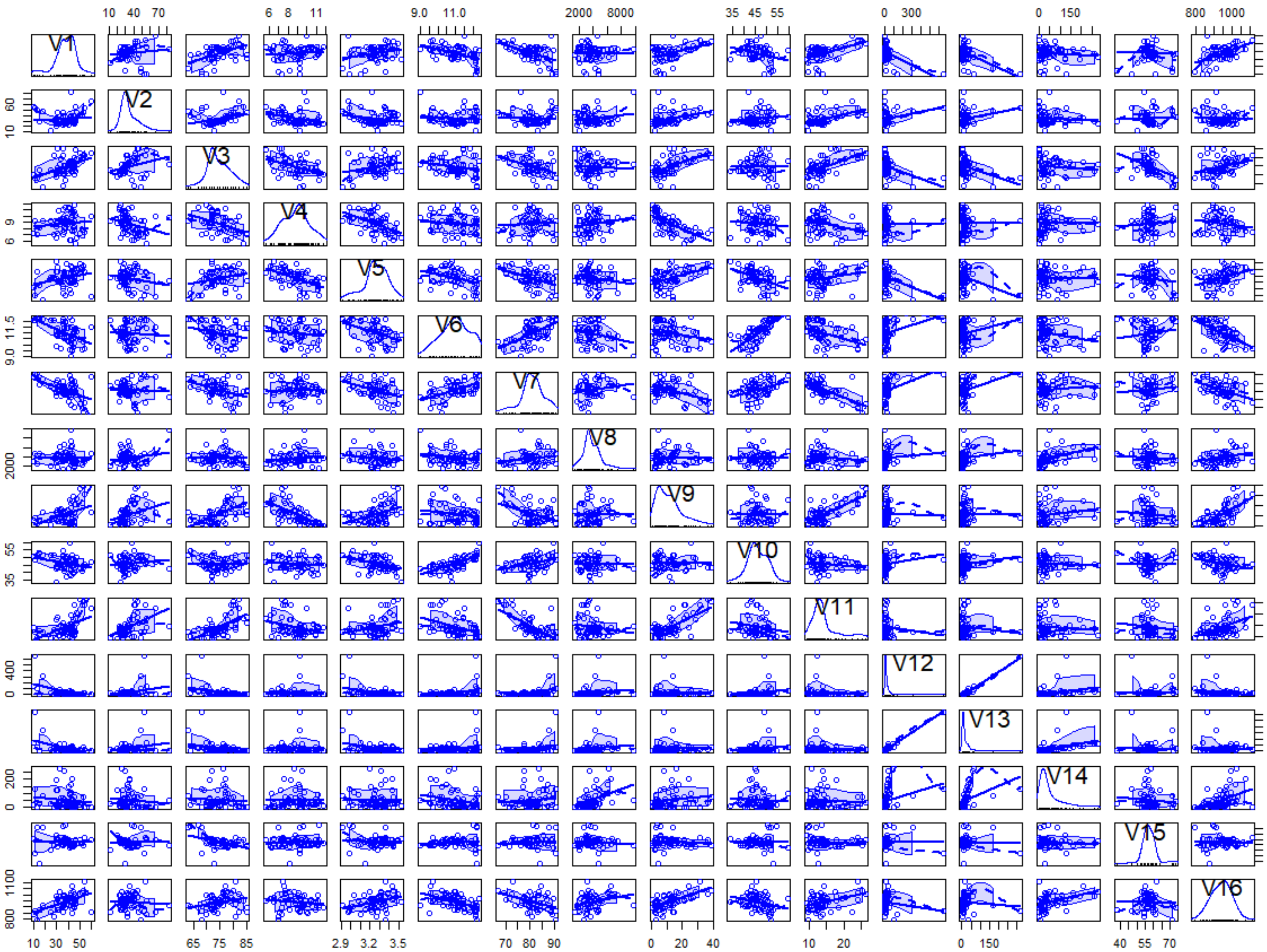
- **Correlation**



- **Heatmap**



- **Scatterplot**
  How variables change with respect to each other.



MODEL BUILDING

The dataset is split into training and testing data in the split ratio of 0.8 i.e.: 44 and 16 observations respectively. The models are based on the training dataset which will later be used for testing the accuracy of the overall model.

- **Variability in Predictor Variables Assumption**: V5 and V6 does not have significantly larger values than zero hence they are excluded for further Analysis.

V1	103	V2	160	V3	24	V4	2.4
V5	0.019	V6	0.76	V7	25	V8	2e+06
V9	65	V10	25	V11	16	V12	11246
V13	2841	V14	3249	V15	33		

1. **Full Model:** V16 = V1+V2+V3+V4+V7+V8+V9+V10+V11+V12+V13+V14+V15

- **No perfect multicollinearity assumption (VIF)**: Using Variance Inflation Factor we can eliminate predictors whose values are greater than 4; V4, V9, V11, V12, V13.

```
  V1    V2    V3    V4    V7    V8    V9   V10    V11   V12     V13     V14     V15
  3.5   3.1   3.7   5.1   2.3   1.9   7.4   1.5    5.1   128.5  138.2   3.9      2.0
```

New Model: V16 = V1+V2+V3+V7+V8+V10+V14+V15

- **Coefficients**:

```
              Estimate      Std. Err    t value Pr(>|t|)
(Intercept)  1.04e+03   2.21e+02    4.70  3.9e-05 ***
V1           1.92e+00   7.37e-01    2.61  0.01324 *
V2          -1.07e+00   5.16e-01   -2.07  0.04610 *
V3           4.92e-01   1.81e+00    0.27  0.78673
V7          -3.08e+00   1.36e+00   -2.27  0.02931 *
V8           6.30e-03   4.56e-03    1.38  0.17574
V10         -6.07e-01   1.28e+00   -0.47  0.63897
V14          4.04e-01   1.07e-01    3.79  0.00056 ***
V15          1.00e+00   1.23e+00    0.82  0.42051
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37 on 35 degrees of freedom
Multiple R-squared:  0.574,    Adjusted R-squared:  0.476
F-statistic: 5.89 on 8 and 35 DF,  p-value: 8.68e-05
```

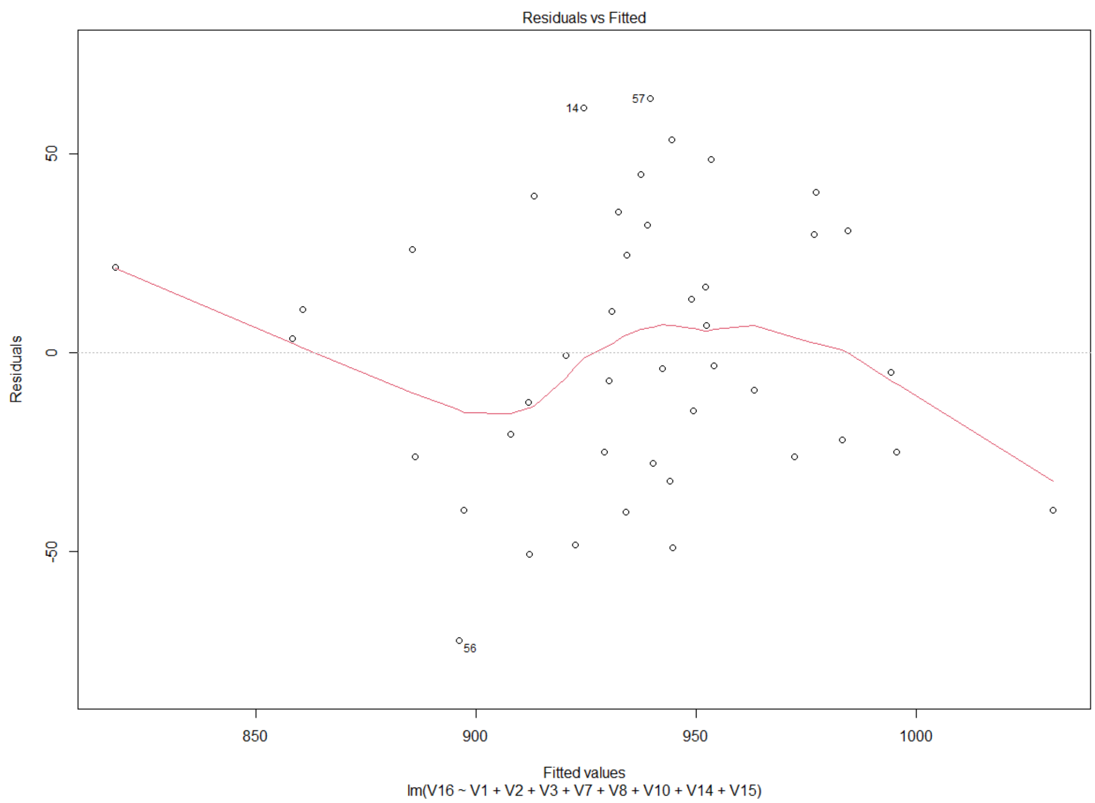
- **Analysis of Variance** Table
  Response: V16

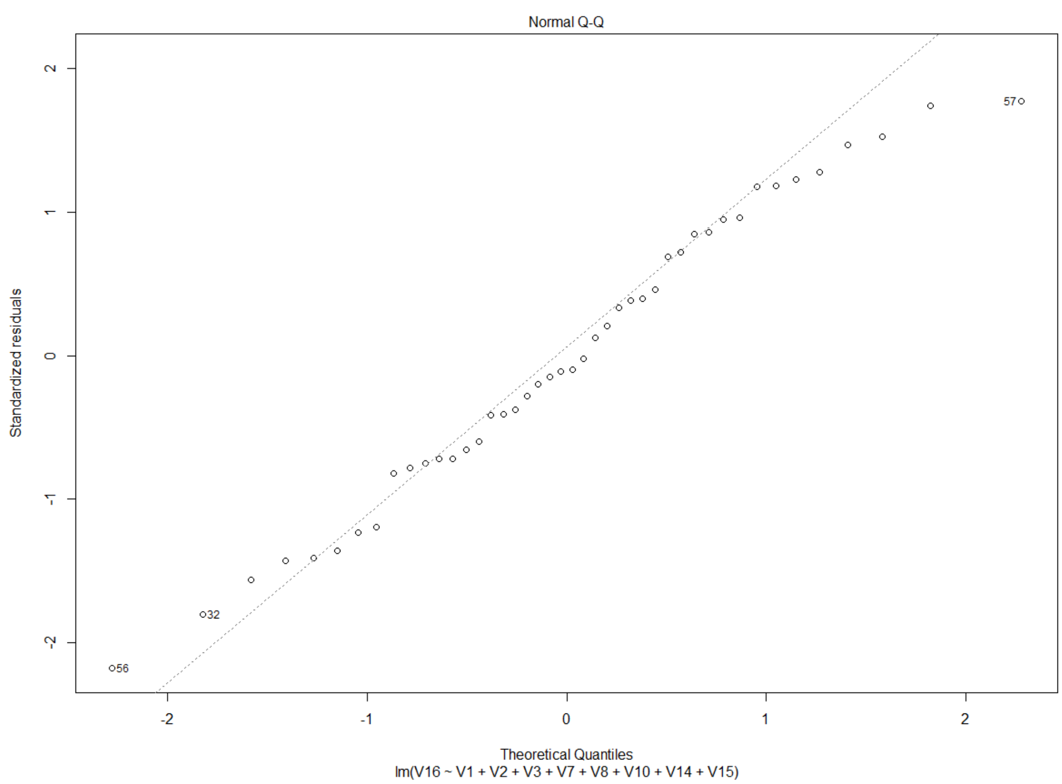
	Df	Sum Sq.	Mean Sq.	F value	Pr(>F)
V1	1	20394	20394	14.79	0.00049 ***
V2	1	5034	5034	3.65	0.06430 .
V3	1	54	54	0.04	0.84460
V7	1	10549	10549	7.65	0.00901 **
V8	1	8368	8368	6.07	0.01884 *
V10	1	402	402	0.29	0.59289
V14	1	19274	19274	13.97	0.00066 ***
V15	1	917	917	0.66	0.42051
Residuals	35	48278	1379		

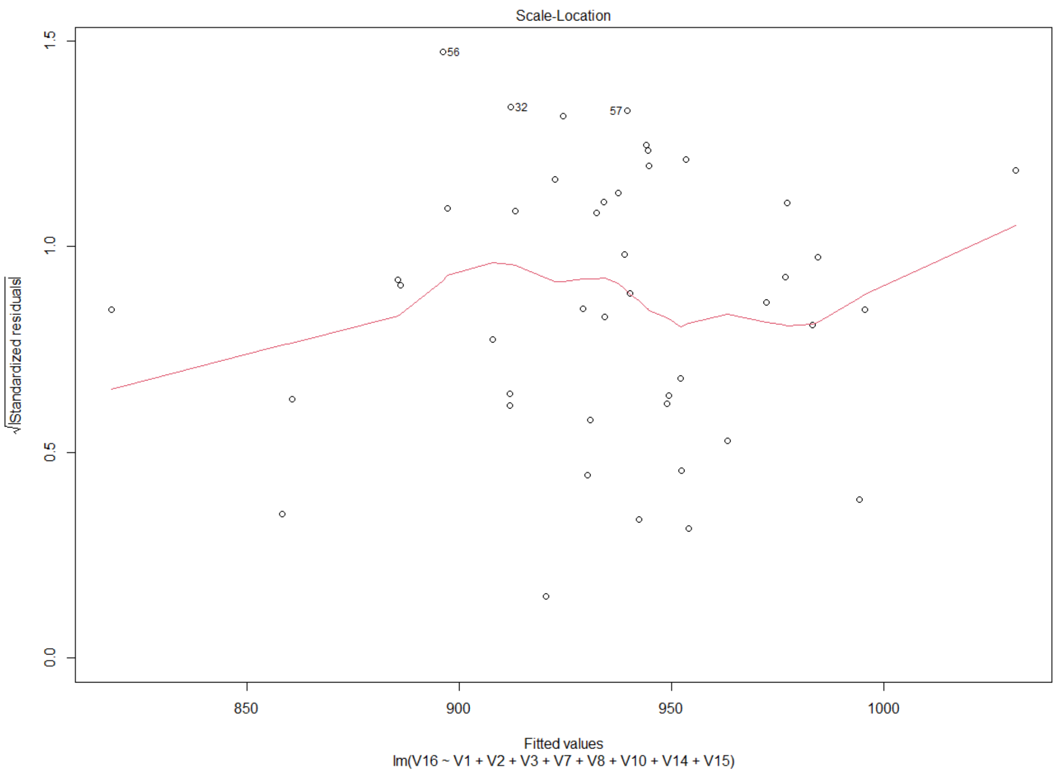
  Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- Regression Model Linearity Assumption: It is linear in parameters. Hence, assumption is satisfied.
- Mean Residual Value Assumption: Mean of the Residuals: 8.3e-16 which is approximately equal to 0. Hence, the assumption is true for this model.
- Homoscedasticity and Normality Assumption:
  Using 4 Degrees of Freedom, Level of Significance = 0.05

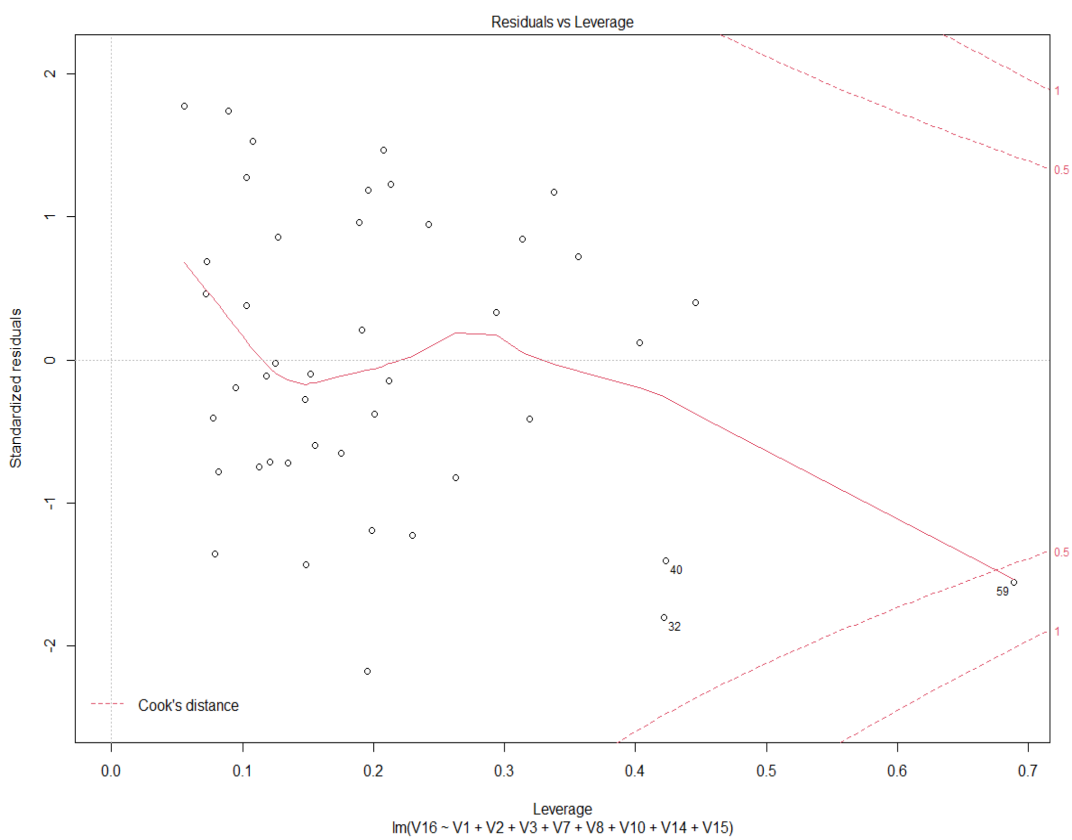
	Value	p-value	Decision
Global Stat	1.9607	0.743	Assumptions acceptable.
Skewness	0.0204	0.886	Assumptions acceptable.
Kurtosis	1.1881	0.276	Assumptions acceptable.
Link Function	0.2576	0.612	Assumptions acceptable.
Heteroscedasticity	0.4946	0.482	Assumptions acceptable.

  The points appear random and the line quite pretty flat, without increasing or decreasing trend. So, the condition of homoscedasticity can be accepted. Thus, Homoscedasticity assumption is satisfied.


Residuals vs Fitted
Residuals
Fitted values
lm(V16 ~ V1 + V2 + V3 + V7 + V8 + V10 + V14 + V15)
850
900
950
1000
50
0
-50
14
57
56


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(V16 ~ V1 + V2 + V3 + V7 + V8 + V10 + V14 + V15)
-2
-1
0
1
2
57
32
56

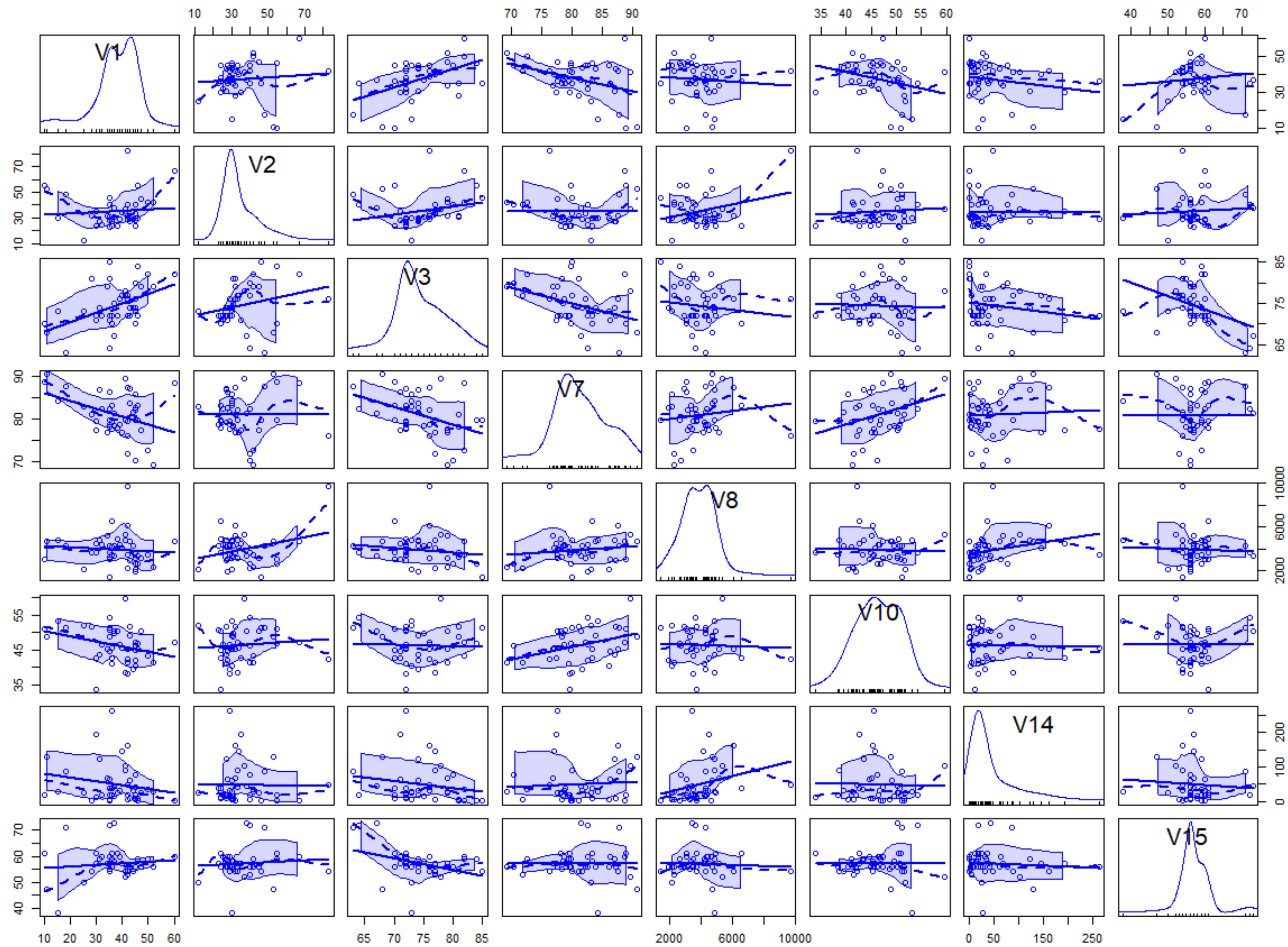

Scale-Location
√|Standardized residuals|
Fitted values
lm(V16 ~ V1 + V2 + V3 + V7 + V8 + V10 + V14 + V15)
56
32
57
0.0
0.5
1.0
1.5
850
900
950
1000


Residuals vs Leverage
Standardized residuals
Leverage
lm(V16 ~ V1 + V2 + V3 + V7 + V8 + V10 + V14 + V15)
Cook's distance
40
32
59
1
0.5
0.5
1
-2
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

- Normality assumption is satisfied.

- Accuracy

	ME	RMSE	MAE	MPE	MAPE
Test set	-158	173	158	-17	17



2. **Reduced Model using AIC (direction: both)**

```
Start:  AIC=348
V16 ~ 1

       Df         Sum of Sq    RSS AIC
+ V7    1     21182  92088 340
+ V1    1     20394  92876 341
+ V14   1     17660  95610 342
+ V10   1     11959 101311 345
<none>          113270 348
+ V2    1      3468 109802 348
+ V8    1      2327 110943 349
+ V3    1      2320 110950 349
+ V15   1       497 112773 349

Step:  AIC=340
V16 ~ V7

       Df Sum of Sq    RSS AIC
+ V14   1     20320  71768 331
+ V1    1      9001  83086 338
+ V8    1      4714  87374 340
<none>               92088 340
+ V10   1      3741  88346 341
+ V2    1      3448  88639 341
+ V15   1       574  91514 342
+ V3    1       110  91978 342
- V7    1     21182 113270 348

Step:  AIC=331
V16 ~ V7 + V14

       Df Sum of Sq   RSS AIC
+ V1    1     15121 56647 323
<none>              71768 331
+ V2    1      3142 68626 331
+ V10   1      2890 68878 332
+ V15   1      1286 70482 333
+ V8    1       899 70869 333
+ V3    1       165 71602 333
- V14   1     20320 92088 340
- V7    1     23843 95610 342
```

```
Step:  AIC=323
V16 ~ V7 + V14 + V1

       Df Sum of Sq   RSS AIC
+ V2    1      4506 52141 321
<none>              56647 323
+ V3    1      1287 55360 324
+ V10   1       874 55773 324
+ V8    1       767 55880 324
+ V15   1       507 56141 325
- V7    1      9489 66136 328
- V1    1     15121 71768 331
- V14   1     26439 83086 338

Step:  AIC=321
V16 ~ V7 + V14 + V1 + V2

       Df Sum of Sq   RSS AIC
+ V8    1      2609 49532 321
<none>              52141 321
+ V15   1       729 51413 323
+ V10   1       495 51647 323
- V2    1      4506 56647 323
+ V3    1       370 51772 323
- V7    1      9024 61166 326
- V1    1     16485 68626 331
- V14   1     26386 78527 337

Step:  AIC=321
V16 ~ V7 + V14 + V1 + V2 + V8

       Df Sum of Sq   RSS AIC
<none>              49532 321
- V8    1      2609 52141 321
+ V15   1       899 48633 322
+ V10   1       255 49277 323
+ V3    1       135 49397 323
- V2    1      6348 55880 324
- V7    1     10055 59587 327
- V1    1     16604 66136 332
- V14   1     19713 69245 334
```

V16 = V7 + V14 + V1 + V2 + V8

- Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.11e+03	1.07e+02	10.42	1.1e-12	***
V7	-3.30e+00	1.19e+00	-2.78	0.00846	**
V14	4.00e-01	1.03e-01	3.89	0.00039	***
V1	2.13e+00	5.98e-01	3.57	0.00099	***
V2	-1.01e+00	4.60e-01	-2.21	0.03344	*
V8	6.09e-03	4.30e-03	1.41	0.16525	

  Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 36 on 38 degrees of freedom
  Multiple R-squared: 0.563,   Adjusted R-squared: 0.505
  F-statistic: 9.78 on 5 and 38 DF, p-value: 4.58e-06

- Analysis of Variance Table

  Response: V16

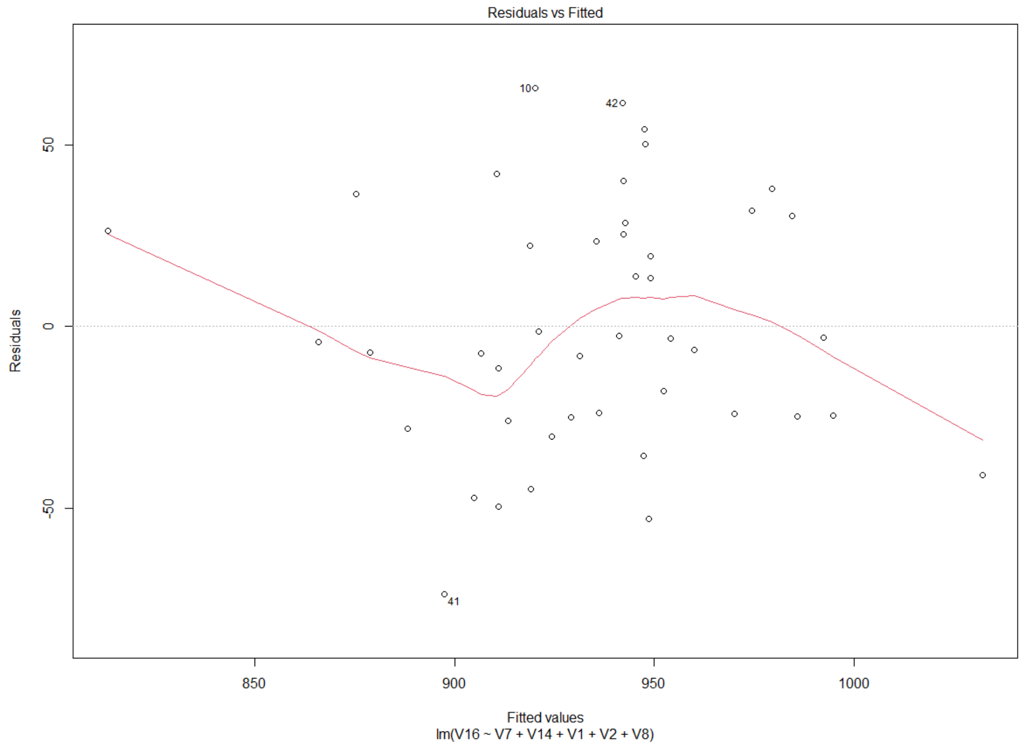
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
V7	1	21182	21182	16.25	0.00026	***
V14	1	20320	20320	15.59	0.00033	***
V1	1	15121	15121	11.60	0.00157	**
V2	1	4506	4506	3.46	0.07075	.
V8	1	2609	2609	2.00	0.16525	
Residuals	38	49532	1303			

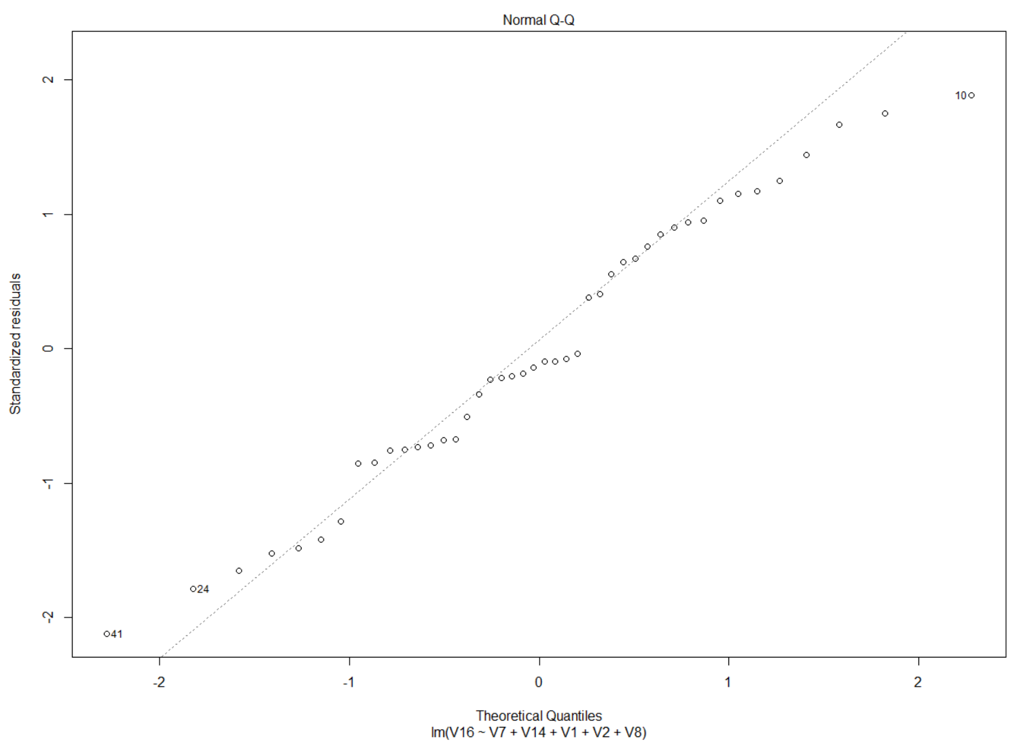
  Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

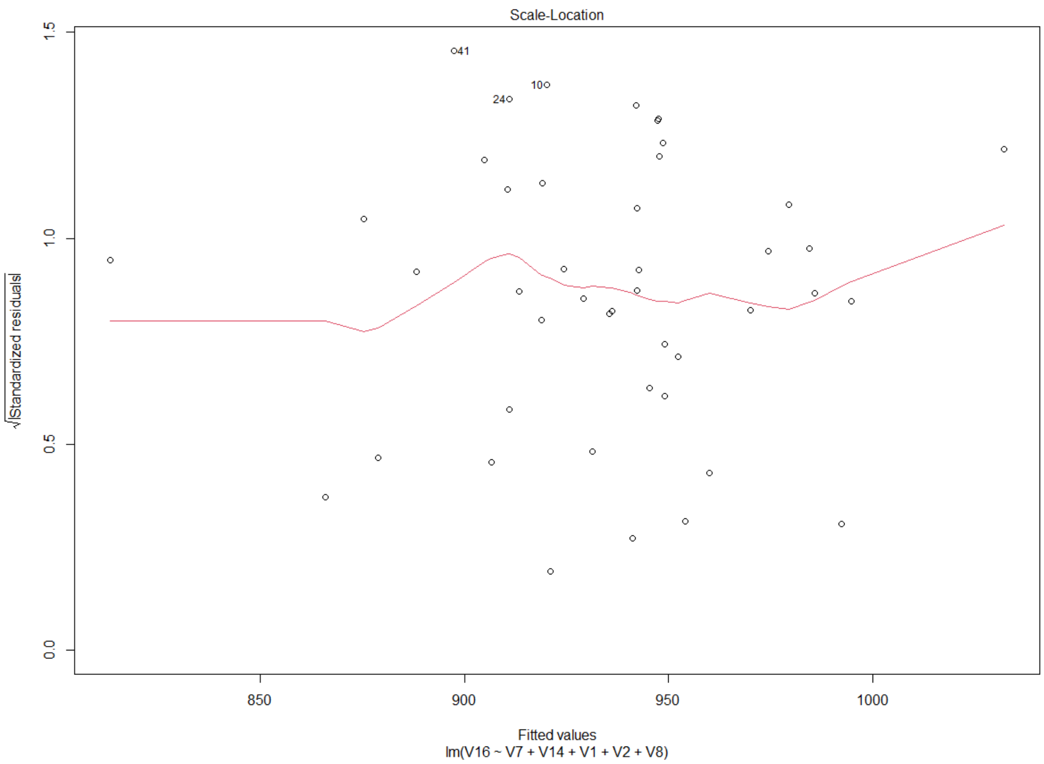
- Regression Model Linearity Assumption: It is linear in parameters. Hence, assumption is satisfied.
- Mean Residual Value Assumption: Mean of the 1.4e-15 which is approximately equal to 0. Hence, the assumption is true for this model.
- Homoscedasticity and Normality Assumption:
  Using 4 Degrees of Freedom, Level of Significance = 0.05

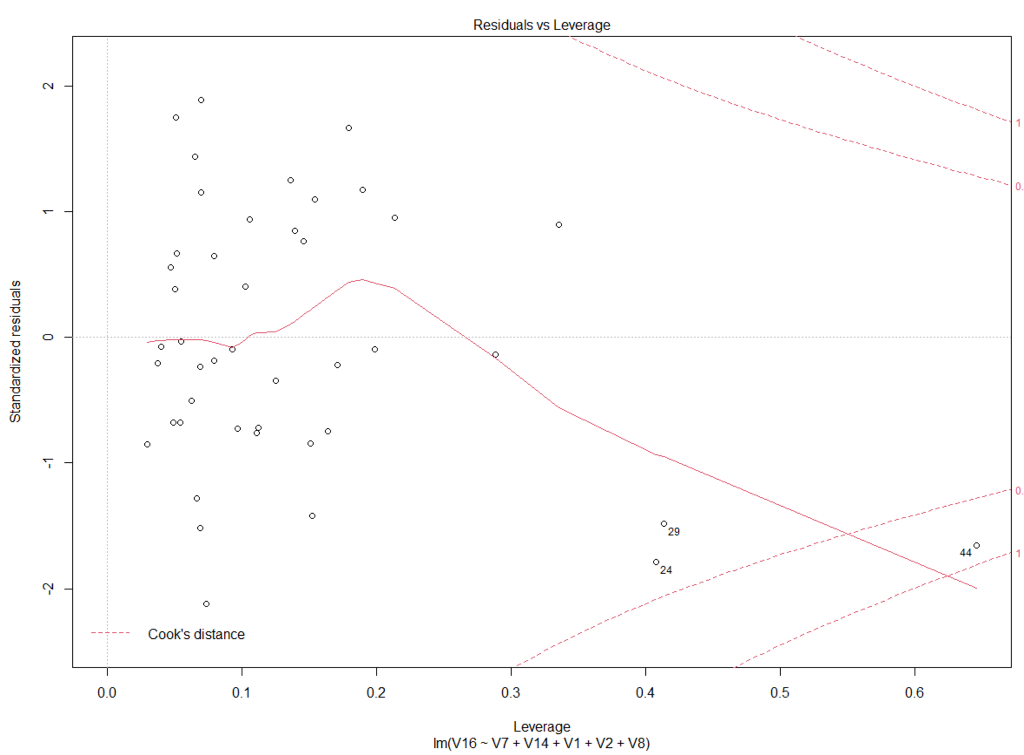
	Value	p-value	Decision
Global Stat	1.7932	0.774	Assumptions acceptable.
Skewness	0.0147	0.903	Assumptions acceptable.
Kurtosis	1.1935	0.275	Assumptions acceptable.
Link Function	0.1059	0.745	Assumptions acceptable.
Heteroscedasticity	0.4790	0.489	Assumptions acceptable.

The points appear random and the line quite pretty flat, without increasing or decreasing trend. So, the condition of homoscedasticity can be accepted. Thus, Homoscedasticity assumption is satisfied.

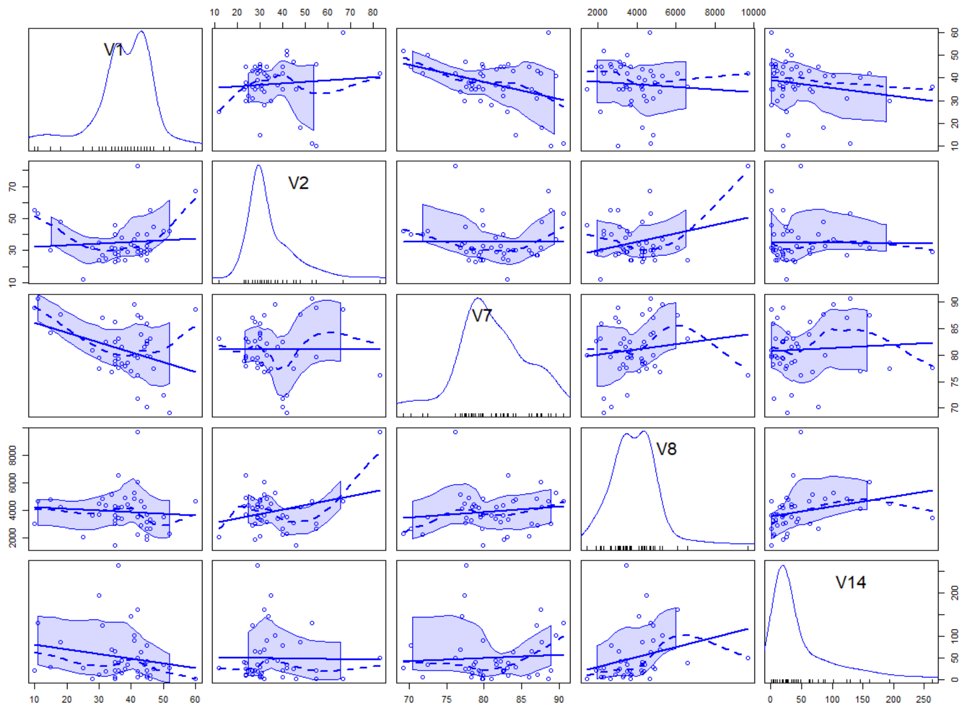





Scale-Location
√|Standardized residuals|
Fitted values
lm(V16 ~ V7 + V14 + V1 + V2 + V8)
41
10
24
0.0
0.5
1.0
1.5
850
900
950
1000


Residuals vs Leverage
Standardized residuals
Leverage
lm(V16 ~ V7 + V14 + V1 + V2 + V8)
Cook's distance
29
24
44
1
0.5
0.5
1
-2
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
0.6

- Normality assumption is satisfied.



- Accuracy

	ME	RMSE	MAE	MPE	MAPE
Test set	-201	225	201	-22	22

3. **Reduced Model using AIC (direction: forward)**

```
Start:  AIC=348
V16 ~ 1

       Df Sum of Sq    RSS AIC
+ V7    1     21182  92088 340
+ V1    1     20394  92876 341
+ V14   1     17660  95610 342
+ V10   1     11959 101311 345
<none>              113270 348
+ V2    1      3468 109802 348
+ V8    1      2327 110943 349
+ V3    1      2320 110950 349
+ V15   1       497 112773 349

Step:  AIC=340
V16 ~ V7

       Df Sum of Sq   RSS AIC
+ V14   1     20320 71768 331
+ V1    1      9001 83086 338
+ V8    1      4714 87374 340
<none>              92088 340
+ V10   1      3741 88346 341
+ V2    1      3448 88639 341
+ V15   1       574 91514 342
+ V3    1       110 91978 342

Step:  AIC=331
V16 ~ V7 + V14

       Df Sum of Sq   RSS AIC
+ V1    1     15121 56647 323
<none>              71768 331
+ V2    1      3142 68626 331
+ V10   1      2890 68878 332
+ V15   1      1286 70482 333
+ V8    1       899 70869 333
+ V3    1       165 71602 333
```

```
Step:  AIC=323
V16 ~ V7 + V14 + V1


Df Sum of Sq   RSS AIC
+ V2    1      4506 52141 321
<none>              56647 323
+ V3    1      1287 55360 324
+ V10   1       874 55773 324
+ V8    1       767 55880 324
+ V15   1       507 56141 325

Step:  AIC=321
V16 ~ V7 + V14 + V1 + V2

       Df Sum of Sq   RSS AIC
+ V8    1      2609 49532 321
<none>              52141 321
+ V15   1       729 51413 323
+ V10   1       495 51647 323
+ V3    1       370 51772 323

Step:  AIC=321
V16 ~ V7 + V14 + V1 + V2 + V8

       Df Sum of Sq   RSS AIC
<none>              49532 321
+ V15   1       899 48633 322
+ V10   1       255 49277 323
+ V3    1       135 49397 323
```

V16 = V7 + V14 + V1 + V2 + V8

- Coefficients:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.11e+03   1.07e+02   10.42  1.1e-12 ***
V7          -3.30e+00   1.19e+00   -2.78  0.00846 **
V14          4.00e-01   1.03e-01    3.89  0.00039 ***
V1           2.13e+00   5.98e-01    3.57  0.00099 ***
V2          -1.01e+00   4.60e-01   -2.21  0.03344 *
V8           6.09e-03   4.30e-03    1.41  0.16525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 36 on 38 degrees of freedom
```

Multiple R-squared: 0.563, Adjusted R-squared: 0.505
F-statistic: 9.78 on 5 and 38 DF, p-value: 4.58e-06

- Analysis of Variance Table

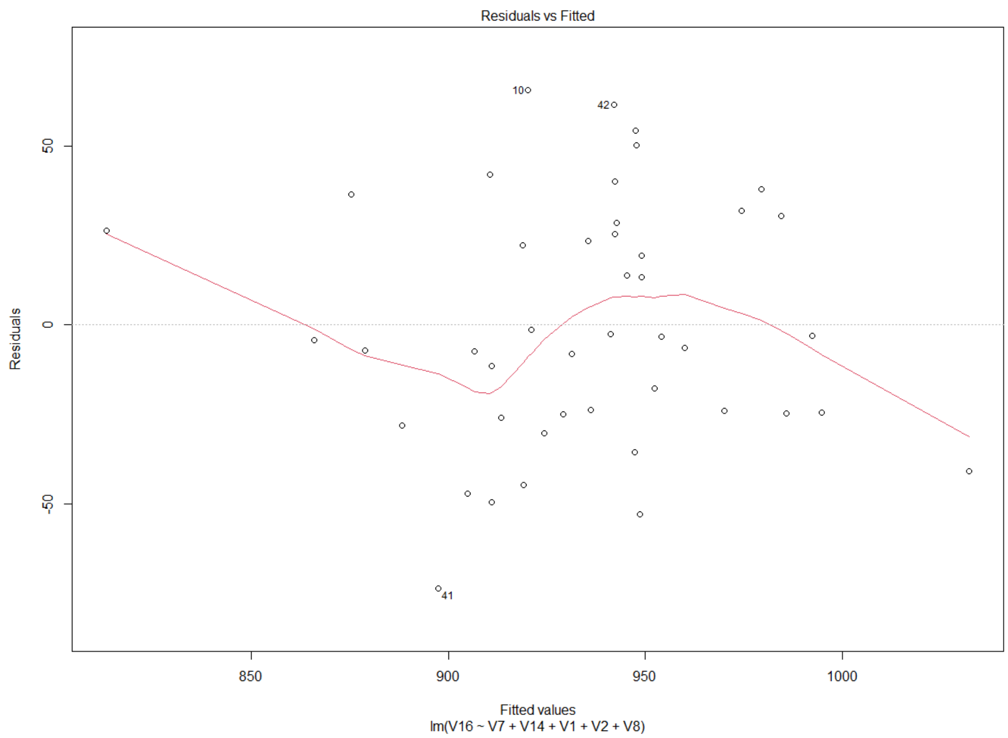
  Response: V16

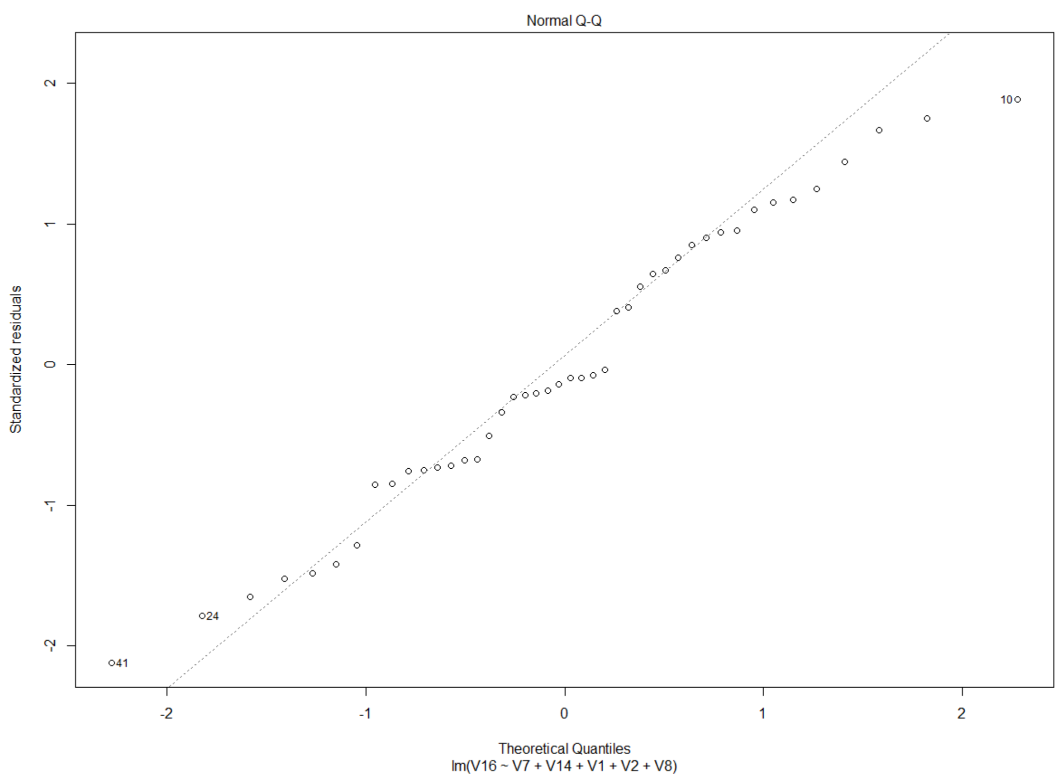
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
V7	1	21182	21182	16.25	0.00026	***
V14	1	20320	20320	15.59	0.00033	***
V1	1	15121	15121	11.60	0.00157	**
V2	1	4506	4506	3.46	0.07075	.
V8	1	2609	2609	2.00	0.16525	
Residuals	38	49532	1303			

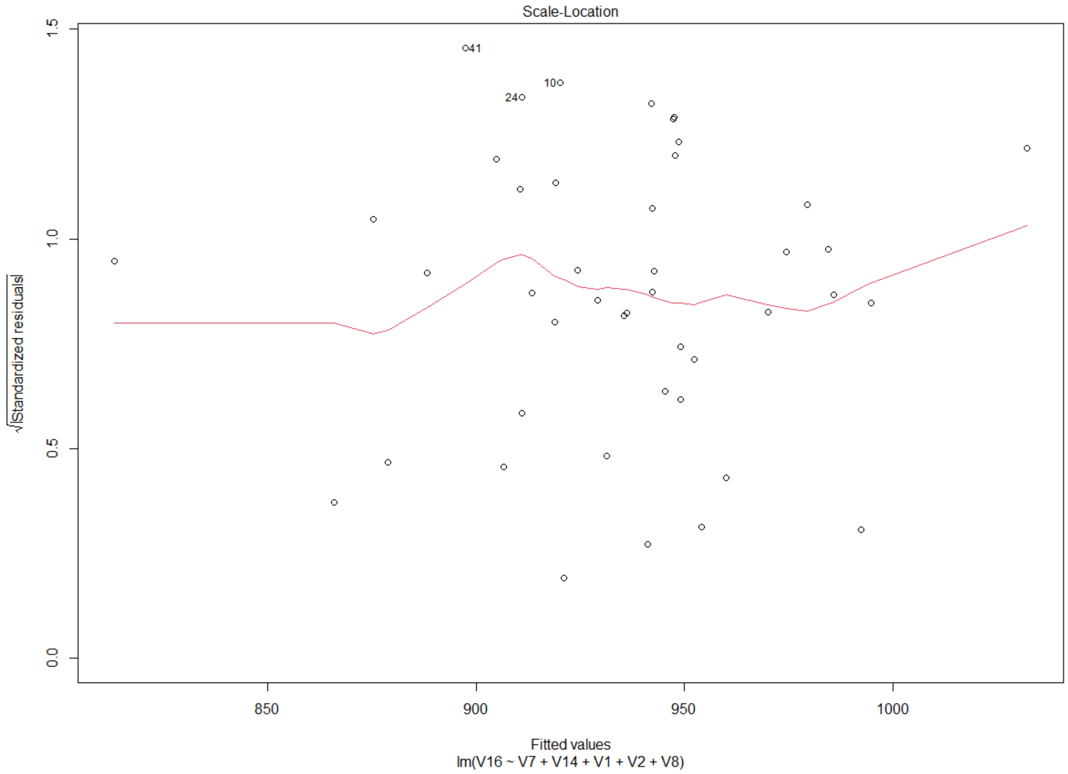
  Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
- Regression Model Linearity Assumption: It is linear in parameters. Hence, assumption is satisfied.
- Mean Residual Value Assumption: Mean of the 1.4e-15 which is approximately equal to 0. Hence, the assumption is true for this model.
- Homoscedasticity and Normality Assumption:
  Using 4 Degrees of Freedom, Level of Significance = 0.05

	Value	p-value	Decision
Global Stat	1.7932	0.774	Assumptions acceptable.
Skewness	0.0147	0.903	Assumptions acceptable.
Kurtosis	1.1935	0.275	Assumptions acceptable.
Link Function	0.1059	0.745	Assumptions acceptable.
Heteroscedasticity	0.4790	0.489	Assumptions acceptable.

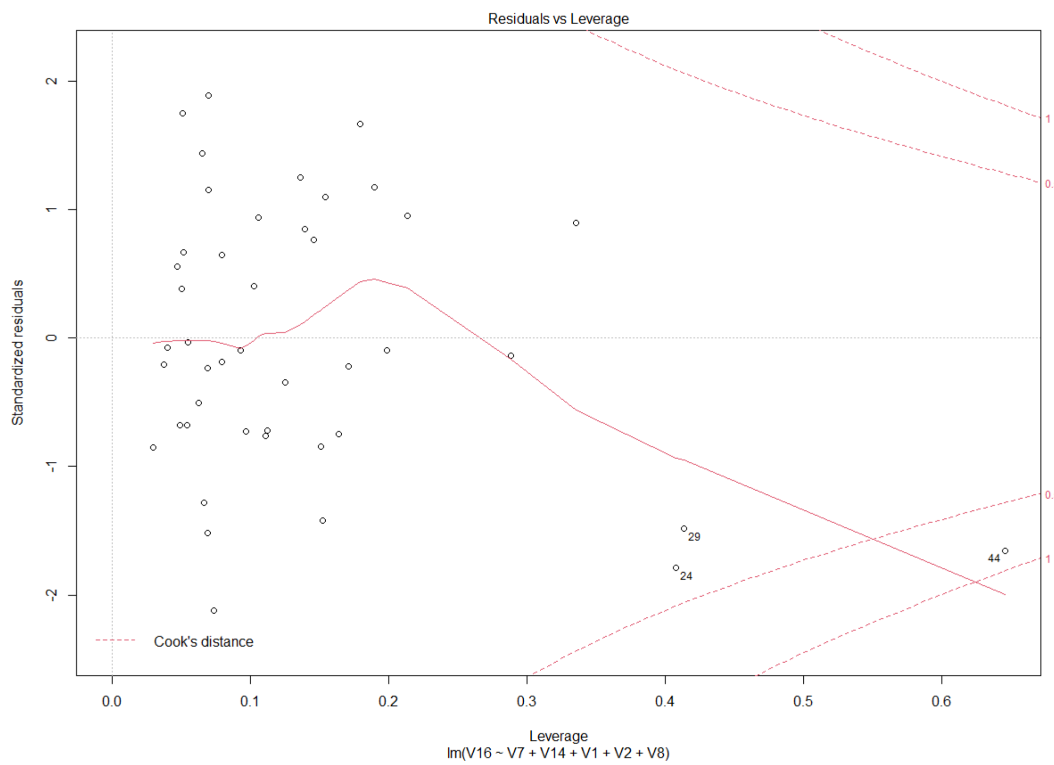
  The points appear random and the line quite pretty flat, without increasing or decreasing trend. So, the condition of homoscedasticity can be accepted. Thus, Homoscedasticity assumption is satisfied.


Residuals vs Fitted
Residuals
Fitted values
lm(V16 ~ V7 + V14 + V1 + V2 + V8)
850
900
950
1000
50
0
-50
10
42
41


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(V16 ~ V7 + V14 + V1 + V2 + V8)
-2
-1
0
1
2
10
24
41

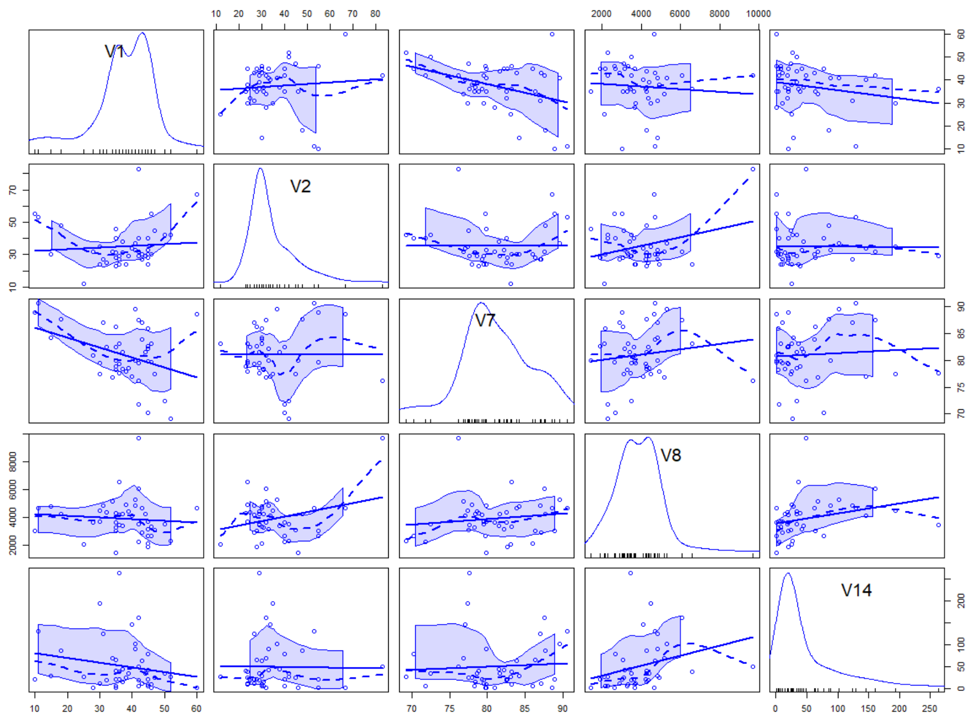


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
850
900
950
1000
41
10
24
Fitted values
lm(V16 ~ V7 + V14 + V1 + V2 + V8)



Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
0.5
0.5
1
29
24
44
Cook's distance
Leverage
lm(V16 ~ V7 + V14 + V1 + V2 + V8)

- Normality assumption is satisfied.



- Accuracy

	ME	RMSE	MAE	MPE	MAPE
Test set	-201	225	201	-22	22

4. **Reduced Model using BIC (direction: forward)**

```
Start:  AIC=348
V16 ~ 1

       Df Sum of Sq    RSS AIC
+ V7    1     21182  92088 342
+ V1    1     20394  92876 342
+ V14   1     17660  95610 344
+ V10   1     11959 101311 346
<none>               113270 348
+ V2    1      3468 109802 350
+ V8    1      2327 110943 350
+ V3    1      2320 110950 350
+ V15   1       497 112773 351

Step:  AIC=342
V16 ~ V7

       Df Sum of Sq   RSS AIC
+ V14   1     20320 71768 334
+ V1    1      9001 83086 340
<none>               92088 342
+ V8    1      4714 87374 342
+ V10   1      3741 88346 343
+ V2    1      3448 88639 343
+ V15   1       574 91514 344
+ V3    1       110 91978 345

Step:  AIC=334
V16 ~ V7 + V14

       Df Sum of Sq   RSS AIC
+ V1    1     15121 56647 326
<none>               71768 334
+ V2    1      3142 68626 335
+ V10   1      2890 68878 335
+ V15   1      1286 70482 336
+ V8    1       899 70869 336
+ V3    1       165 71602 336
```

```
Step:  AIC=326
V16 ~ V7 + V14 + V1

       Df Sum of Sq   RSS AIC
+ V2    1      4506 52141 325
<none>              56647 326
+ V3    1      1287 55360 328
+ V10   1       874 55773 328
+ V8    1       767 55880 328
+ V15   1       507 56141 329

Step:  AIC=325
V16 ~ V7 + V14 + V1 + V2

       Df Sum of Sq   RSS AIC
<none>              52141 325
+ V8    1      2609 49532 326
+ V15   1       729 51413 327
+ V10   1       495 51647 328
+ V3    1       370 51772 328
```

V16 = V7 + V14 + V1 + V2

- Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1114.1034   108.3761   10.28  1.2e-12 ***
V7            -3.1089     1.1966   -2.60   0.0132 *
V14            0.4429     0.0997    4.44  7.1e-05 ***
V1             2.1266     0.6056    3.51   0.0011 **
V2            -0.8124     0.4425   -1.84   0.0740 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37 on 39 degrees of freedom
Multiple R-squared:  0.54,      Adjusted R-squared:  0.492
F-statistic: 11.4 on 4 and 39 DF,  p-value: 3.1e-06
```

- Analysis of Variance Table

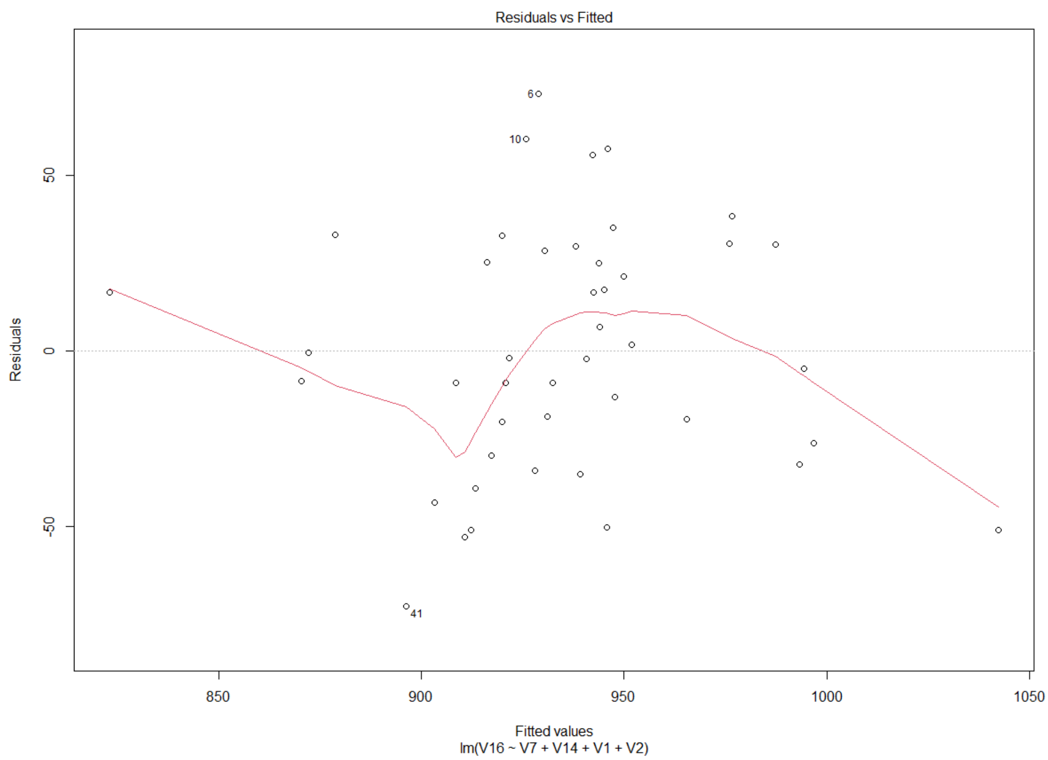
```
Response: V16
          Df Sum Sq Mean Sq F value  Pr(>F)
V7         1  21182   21182   15.84 0.00029 ***
V14        1  20320   20320   15.20 0.00037 ***
V1         1  15121   15121   11.31 0.00174 **
```

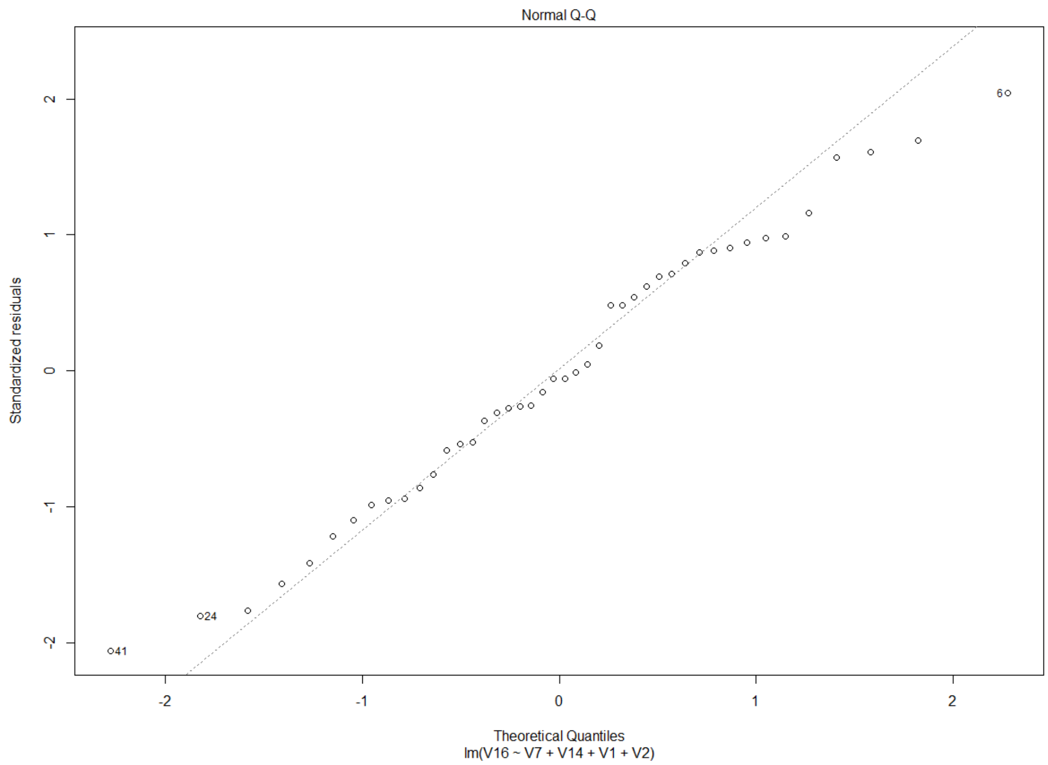
V2 1 4506 4506 3.37 0.07402 .
Residuals 39 52141 1337

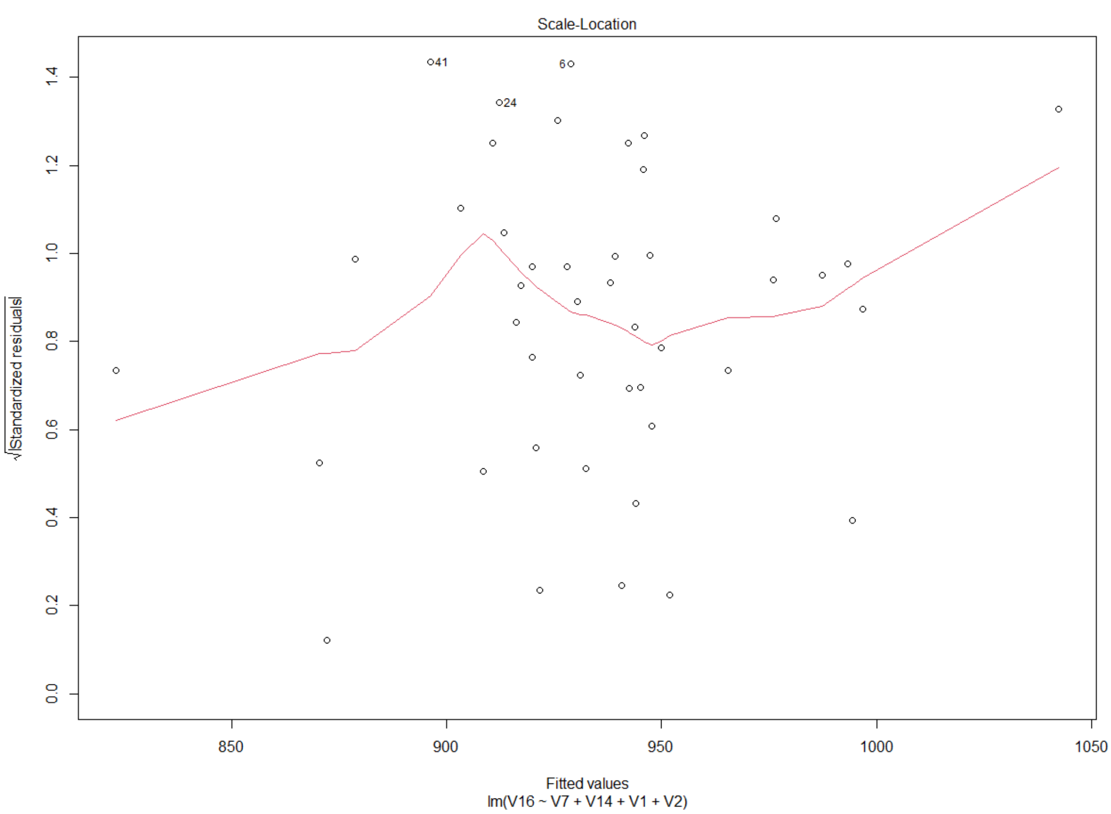
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

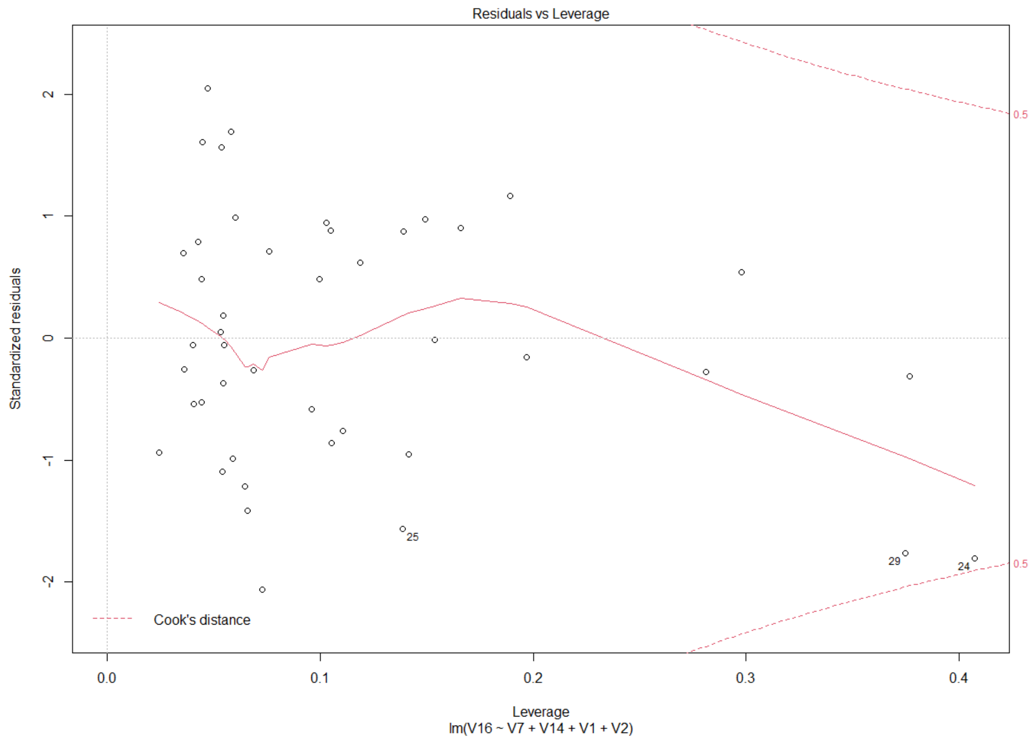
- Regression Model Linearity Assumption: It is linear in parameters. Hence, assumption is satisfied.
- Mean Residual Value Assumption: Mean of the 1.3e-15 which is approximately equal to 0. Hence, the assumption is true for this model.
- Homoscedasticity and Normality Assumption:
  Using 4 Degrees of Freedom, Level of Significance = 0.05

	Value	p-value	Decision
Global Stat	2.77618	0.596	Assumptions acceptable.
Skewness	0.00592	0.939	Assumptions acceptable.
Kurtosis	1.02728	0.311	Assumptions acceptable.
Link Function	1.72572	0.189	Assumptions acceptable.
Heteroscedasticity	0.01726	0.895	Assumptions acceptable.

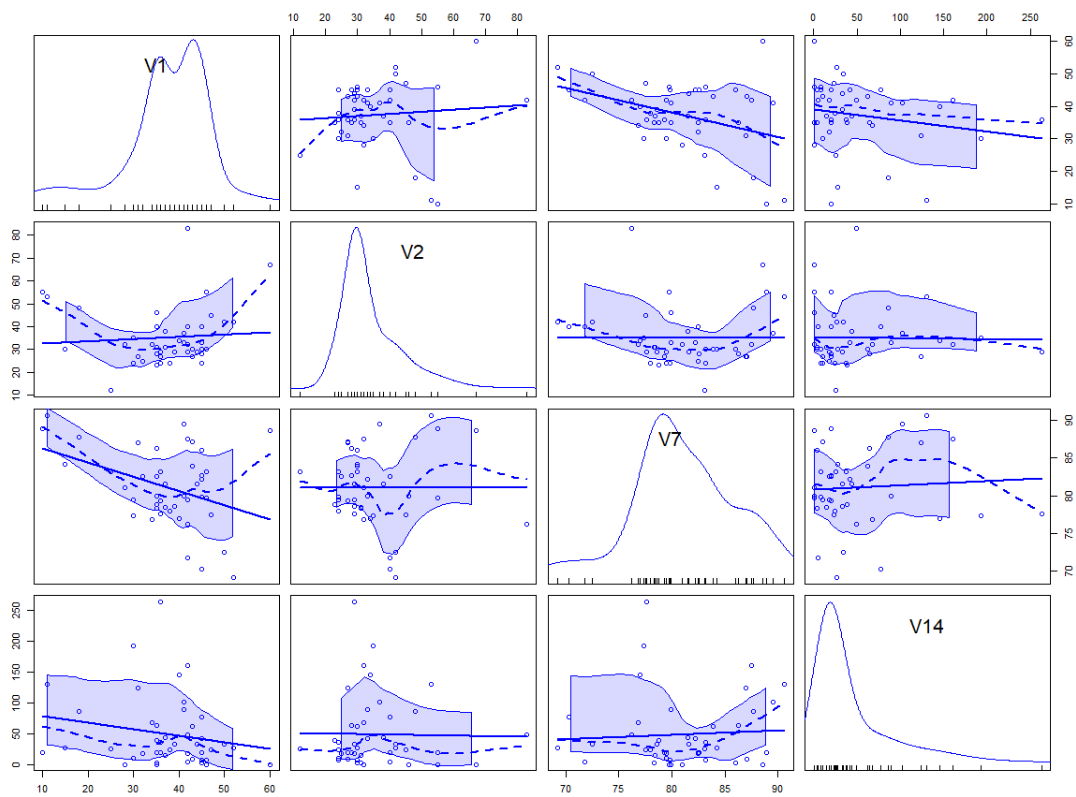



Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(V16 ~ V7 + V14 + V1 + V2)
41
24
6
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
lm(V16 ~ V7 + V14 + V1 + V2)
41
6
24
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
850
900
950
1000
1050



- Normality assumption is satisfied.



- Accuracy

	ME	RMSE	MAE	MPE	MAPE
Test set	-217	239	217	-24	24

5. **Reduced Model using BIC (direction: both)**

```
Start:  AIC=333
V16 ~ V1 + V2 + V3 + V7 + V8 + V10 + V14 + V15

       Df Sum of Sq   RSS AIC
- V3    1       103 48381 330
- V10   1       309 48587 330
- V15   1       917 49195 331
- V8    1      2634 50913 333
<none>              48278 333
- V2    1      5899 54177 335
- V7    1      7123 55402 336
- V1    1      9396 57675 338
- V14   1     19857 68135 345

Step:  AIC=330
V16 ~ V1 + V2 + V7 + V8 + V10 + V14 + V15

       Df Sum of Sq   RSS AIC
- V10   1       252 48633 328
- V15   1       897 49277 328
- V8    1      2532 50913 330
<none>              48381 330
- V2    1      6190 54571 333
+ V3    1       103 48278 333
- V7    1      8636 57017 335
- V1    1     14119 62500 339
- V14   1     19759 68140 343

Step:  AIC=328
V16 ~ V1 + V2 + V7 + V8 + V14 + V15

       Df Sum of Sq   RSS AIC
- V15   1       899 49532 326
- V8    1      2780 51413 327
<none>              48633 328
+ V10   1       252 48381 330
+ V3    1        46 48587 330
- V2    1      6696 55329 331
```

```
- V7    1     10386 59019 333
- V1    1     15640 64273 337
- V14   1     20036 68669 340

Step:  AIC=326
V16 ~ V1 + V2 + V7 + V8 + V14

       Df Sum of Sq   RSS AIC
- V8    1      2609 52141 325
<none>              49532 326
+ V15   1       899 48633 328
- V2    1      6348 55880 328
+ V10   1       255 49277 328
+ V3    1       135 49397 328
- V7    1     10055 59587 331
- V1    1     16604 66136 336
- V14   1     19713 69245 338

Step:  AIC=325
V16 ~ V1 + V2 + V7 + V14

       Df Sum of Sq   RSS AIC
<none>              52141 325
+ V8    1      2609 49532 326
- V2    1      4506 56647 326
+ V15   1       729 51413 327
+ V10   1       495 51647 328
+ V3    1       370 51772 328
- V7    1      9024 61166 330
- V1    1     16485 68626 335
- V14   1     26386 78527 341
```

- **V16 = V1 + V2 + V7 + V14**

- Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1114.1034   108.3761   10.28  1.2e-12 ***
V1             2.1266     0.6056    3.51   0.0011 **
V2            -0.8124     0.4425   -1.84   0.0740 .
V7            -3.1089     1.1966   -2.60   0.0132 *
V14            0.4429     0.0997    4.44  7.1e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

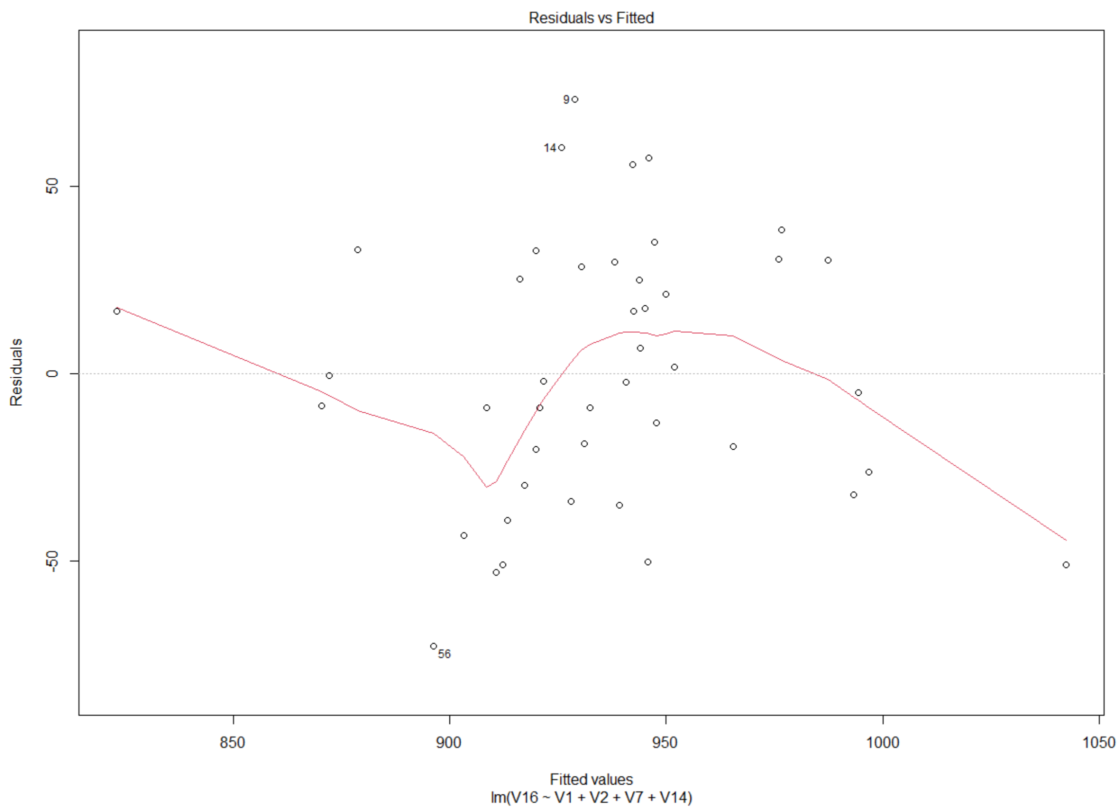
```
Residual standard error: 37 on 39 degrees of freedom
Multiple R-squared:  0.54,	Adjusted R-squared:  0.492
F-statistic: 11.4 on 4 and 39 DF,  p-value: 3.1e-06
```

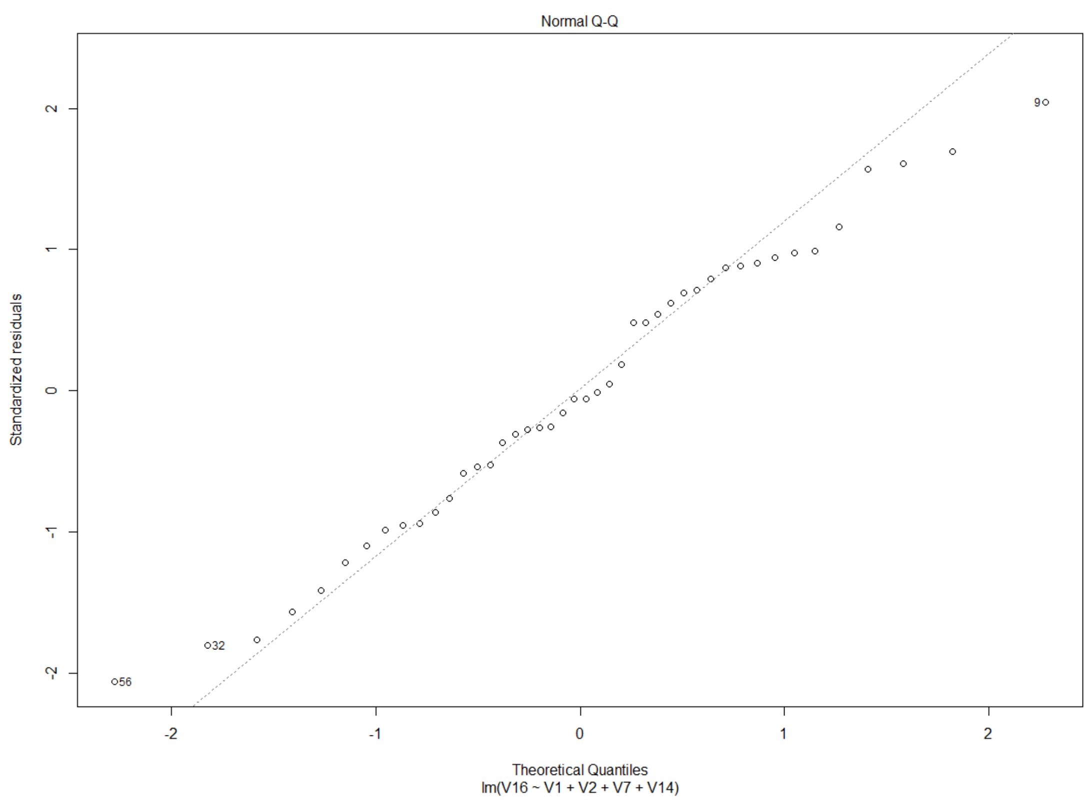
- Analysis of Variance Table

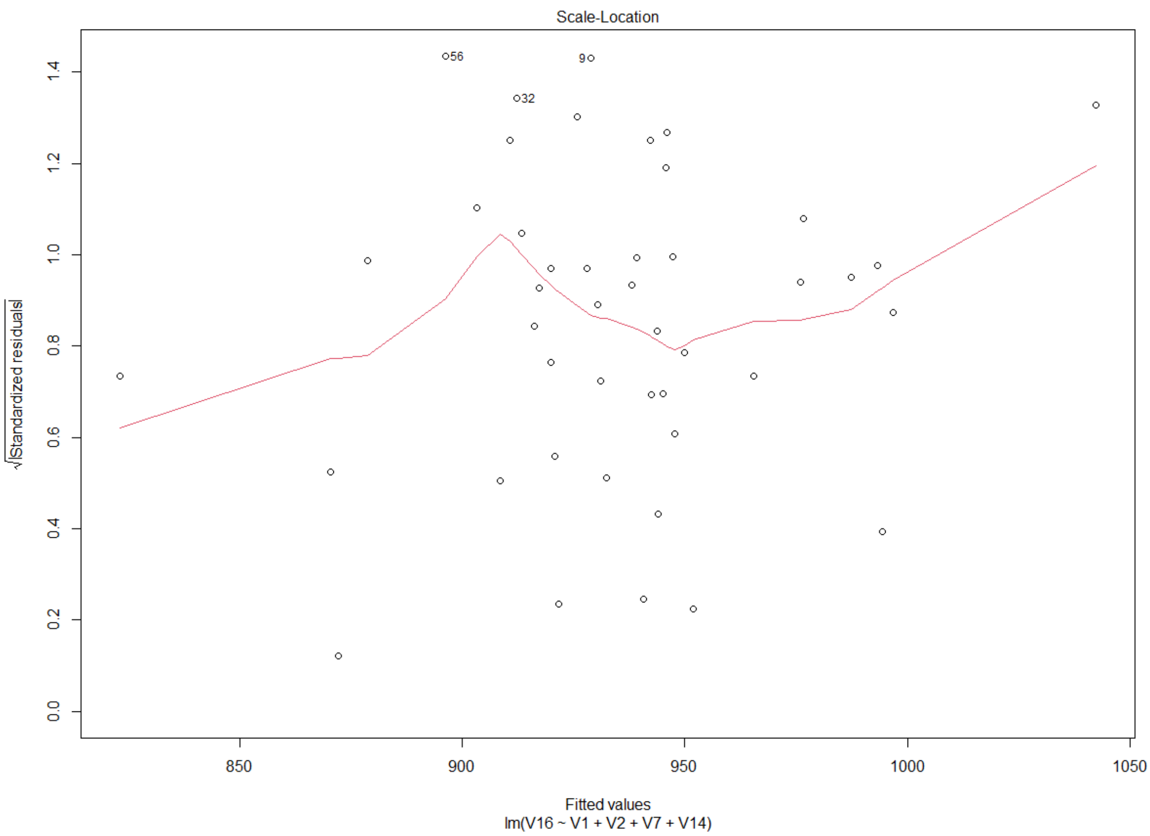
```
Response: V16
          Df Sum Sq Mean Sq F value  Pr(>F)
V1         1  20394   20394   15.25 0.00036 ***
V2         1   5034    5034    3.77 0.05957 .
V7         1   9314    9314    6.97 0.01187 *
V14        1  26386   26386   19.74 7.1e-05 ***
Residuals 39  52141    1337
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

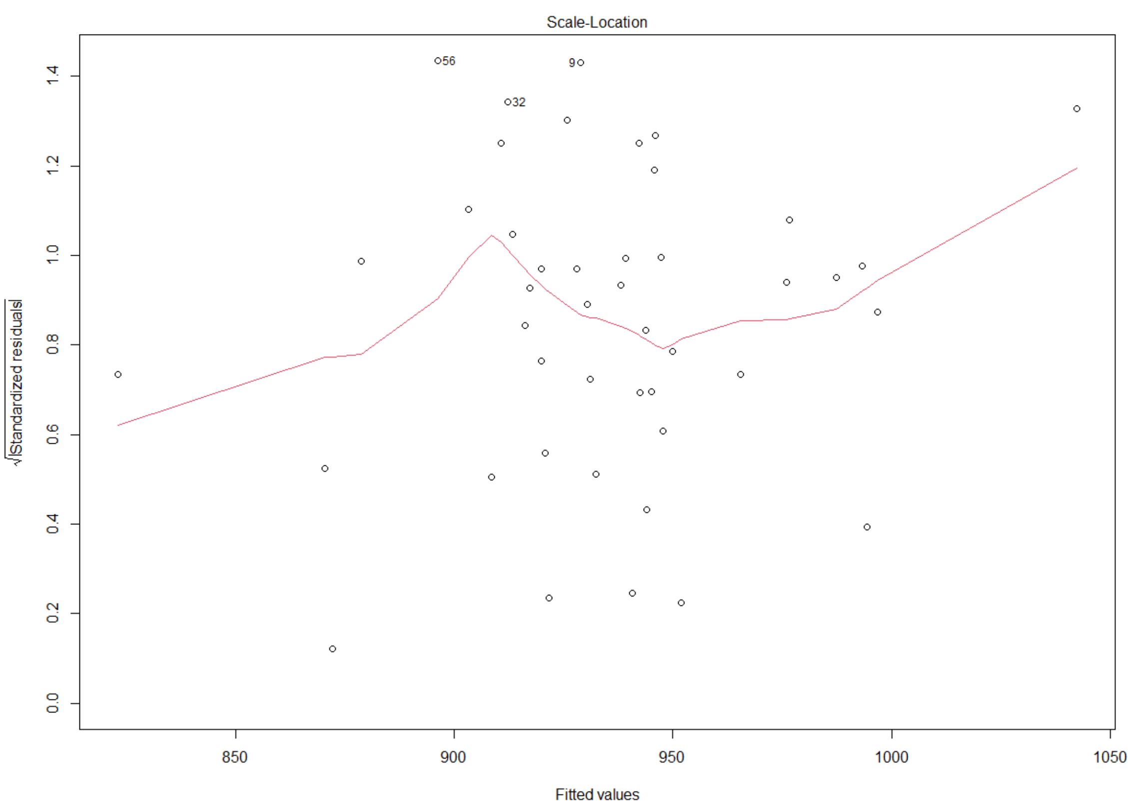
- Regression Model Linearity Assumption: It is linear in parameters. Hence, assumption is satisfied.
- Mean Residual Value Assumption: Mean of the 1.3e-15 which is approximately equal to 0. Hence, the assumption is true for this model.
- Homoscedasticity and Normality Assumption:
  Using 4 Degrees of Freedom, Level of Significance = 0.05

	Value	p-value	Decision
Global Stat	2.77618	0.596	Assumptions acceptable.
Skewness	0.00592	0.939	Assumptions acceptable.
Kurtosis	1.02728	0.311	Assumptions acceptable.
Link Function	1.72572	0.189	Assumptions acceptable.
Heteroscedasticity	0.01726	0.895	Assumptions acceptable.

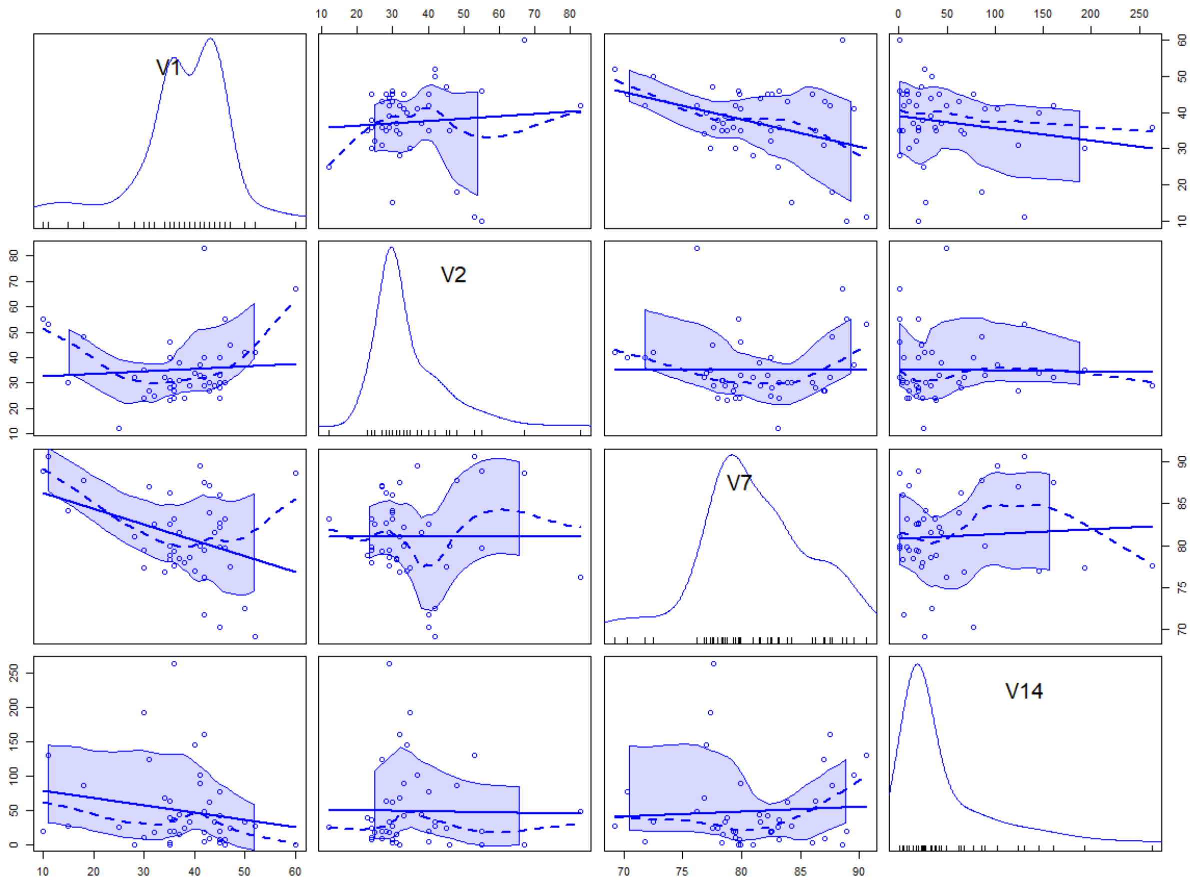

Residuals vs Fitted
Residuals
Fitted values
lm(V16 ~ V1 + V2 + V7 + V14)
9
14
56
50
0
-50
850
900
950
1000
1050


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(V16 ~ V1 + V2 + V7 + V14)
9
32
56
2
1
0
-1
-2


Scale-Location
√|Standardized residuals|
Fitted values
lm(V16 ~ V1 + V2 + V7 + V14)
56
9
32


Scale-Location
√|Standardized residuals|
Fitted values
lm(V16 ~ V1 + V2 + V7 + V14)
56
9
32

- Normality assumption is satisfied.



- Accuracy

	ME	RMSE	MAE	MPE	MAPE
Test set	22	83	67	1.6	6.9

MODEL COMPARISON: -

- ☐ **Analysis of Variance Table of all the Models**

Model 1: V16 ~ V1 + V2 + V3 + V7 + V8 + V10 + V14 + V15

(Intercept)	V1	V2	V3	V7	V8	V10	V14	V15
1.0e+03	1.9e+00	-1.1e+00	4.9e-01	-3.1e+00	6.3e-03	-6.1e-01	4.0e-01	1.0e+00

Model 2: V16 ~ V7 + V14 + V1 + V2 + V8

(Intercept)	V7	V14	V1	V2	V8
1.1e+03	-3.3e+00	4.0e-01	2.1e+00	-1.0e+00	6.1e-03

Model 3: V16 ~ V7 + V14 + V1 + V2 + V8

(Intercept)	V7	V14	V1	V2	V8
1.1e+03	-3.3e+00	4.0e-01	2.1e+00	-1.0e+00	6.1e-03

Model 4: V16 ~ V7 + V14 + V1 + V2

(Intercept)	V7	V14	V1	V2
1114.103	-3.109	0.443	2.127	-0.812

Model 5: V16 ~ V1 + V2 + V7 + V14

(Intercept)	V1	V2	V7	V14
1114.103	2.1127	-0.812	-3.109	0.443

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	35	48278				
2	38	49532	-3	-1254	0.30	0.82
3	38	49532	0	0		
4	39	52141	-1	-2609	1.89	0.18

☐ Comparing the Values

```
broom::glance(regFull)
A tibble: 1 x 12
r.squared adj.r.squared sigma statistic    p.value    df logLik   AIC   BIC deviance df.residual  nobs
    <dbl>         <dbl> <dbl>     <dbl>      <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
    0.574         0.476  37.1      5.89 0.0000868     8  -216.  453.  471.   48278.          35    44
broom::glance(regAICboth)
A tibble: 1 x 12
r.squared adj.r.squared sigma statistic     p.value    df logLik   AIC   BIC deviance df.residual  nobs
    <dbl>         <dbl> <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
    0.563         0.505  36.1      9.78 0.00000458     5  -217.  448.  461.   49532.          38    44
broom::glance(regAICfwd)
A tibble: 1 x 12
r.squared adj.r.squared sigma statistic     p.value    df logLik   AIC   BIC deviance df.residual  nobs
    <dbl>         <dbl> <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
    0.563         0.505  36.1      9.78 0.00000458     5  -217.  448.  461.   49532.          38    44
broom::glance(regBICfwd)
A tibble: 1 x 12
r.squared adj.r.squared sigma statistic     p.value    df logLik   AIC   BIC deviance df.residual  nobs
    <dbl>         <dbl> <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
    0.540         0.492  36.6      11.4 0.00000310     4  -218.  448.  459.   52141.          39    44
broom::glance(regBICboth)
A tibble: 1 x 12
r.squared adj.r.squared sigma statistic     p.value    df logLik   AIC   BIC deviance df.residual  nobs
    <dbl>         <dbl> <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
    0.540         0.492  36.6      11.4 0.00000310     4  -218.  448.  459.   52141.          39    44
```

☐ Comparing the Accuracy

	ModelNames	*Rsq*	*AdjRsq*	*MeanErr*	*MeanAbsErr*	*MeanPercentErr*	*MeanAbsPercentErr*
1	*regFull*	0.57	0.48	22	67	1.6	6.9
2	*regAICboth*	0.56	0.51	-201	201	-22.1	22.1
3	*regAICfwd*	0.56	0.51	-201	201	-22.1	22.1
4	*regBICfwd*	0.54	0.49	-217	217	-23.8	23.8
5	*regBICboth*	0.54	0.49	22	67	1.6	6.9

CONCLUSION

We have concluded that the best model is the one made using BIC, both directional;
Final Model:

$$\mathbf{Y = (1114.103) + (2.1127)V1 + (-0.812)V2 + (-3.109)V7 + (0.443)V14}$$

Y=V16, the death rate

V1, the average annual precipitation;
V2, the average January temperature;
V7, the number of households with fully equipped kitchens;
V14, the sulfur dioxide pollution index;

The Final Model does not have the best Adjusted R-squared, however it has the best BIC value (least BIC value) and the best Mean Error (lowest error).

CODE

Data Set Description and Introduction

```
XB = read.table('D:/Studies/NJiT/Sem_2/Regression_Analysis/Project/x28.txt')
XB

str(XB) # Compact structure of Data
summary(XB)

library(Hmisc)
describe(XB)

library(pastecs)
stat.desc(XB)
options(digits=3)
StatAn <- pd.DataFrame(stat.desc(XB, basic=F))
StatAn

XB.cor = cor(XB)# using Pearson

XB.cor

library(corrplot)
dev.off()
corrplot(XB.cor)

palette = colorRampPalette(c("cyan", "#3296fa", "#003294")) (20)
heatmap(x = XB.cor, col = palette, symm = TRUE)

pairs(~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13+V14+V15+V16, data = XB)

library(car)
scatterplotMatrix(~ V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13+V14+V15+V16,
data = XB)

set.seed(125)
library(caTools)

data_split = sample.split(XB, SplitRatio = 0.8)
train <- subset(XB, data_split == TRUE)
test <-subset(XB, data_split == FALSE)
summary(train)

V16 = train$V16
V1 <- train$V1
V2 = train[,2]
V3 <- train[,3]
V4 = train$V4
V5 <- train$V5
```

```
V6 = train[,6]
V7 <- train[,7]
V8 = train$V8
V9 <- train$V9
V10 = train[,10]
V11 <- train[,11]
V12 = train$V12
V13 <- train$V13
V14 = train[,14]
V15 <- train[,15]

k=1  # test Variability is larger than 0;  therefore 5th and 6th columns not considered
while (k<16){
  print(var(train[,k]))
  k = k+1
}

Y_og <- test[nrow(test)]   # Original Values
Y_og

Y_calc <- test[nrow(test)]*0 # Initialization For calculated values
Y_calc

regBase = lm(V16 ~ 1)
regBase
```

1. FULL MODEL

```
regFull = lm(V16 ~ V1+V2+V3+V4+V7+V8+V9+V10+V11+V12+V13+V14+V15, data = train)
regFull

# VIF factor: No perfect multicollinearity
vif(modeL)

regFull = lm(V16 ~ V1+V2+V3+V7+V8+V10+V14+V15, data = train)
regFull

Y_calcF1 <- Y_calc  # Initialization For calculated values
Y_calc1 <- Y_calcF1  # Initialization For calculated values
modeL <- regFull

summary(modeL)
anova(modeL)
attributes(modeL)
residuals(modeL)
sum(residuals(modeL))
mean(residuals(modeL))
```

```
# homoscedasticity
par(mfrow=c(length(test),length(test)))
gvlma::gvlma(modeL, alphalevel = 0.05)
dev.off()
plot(modeL)

# TEST DATA
coefficients(modeL)

pairs(~V1+V2+V3+V7+V8+V10+V14+V15, data = train)

library(car)
scatterplotMatrix(~ V1+V2+V3+V7+V8+V10+V14+V15, data = train)

# CODE for getting calculated values of the test model and later comparing them with the
Original values

j = 1
while (j<=nrow(test)) {
  i = 1
  Y_calc1[j,1] <- coef(modeL)[i]

  while (i<(length(coef(modeL))-1)) {
    Y_calc1[j,1] = Y_calc1[j,1] + test[j,i]*coef(modeL)[i+1]
    i=i+1
  }
  j=j+1

}
   #   Checking Accuracy
library(forecast)
Acc_Y_calc1 = accuracy(Y_calc1[,1], Y_og[,1])
CompTab = Y_og
CompTab[,2] = Y_calc1[,1]
CompTab[,3] = Y_og[,1] - Y_calc1[,1]
CompTab[,4] = CompTab[,3]*CompTab[,3]
library(dplyr)
CompTab <- rename(CompTab, Y_og = V16, Y_calcl1 = V2, error = V3, Sq.error = V4)
CompTab
Y_calcF1 <- Y_calc1
Acc_Y_calcF1 <- Acc_Y_calc1
Acc_Y_calcF1
```

2. Reduced Model: F-test-based backward selection using rms::fastbw()

```
library(rms) # rms: root mean sqaure; ols: ordinary least squares
ols.full <- ols(V16 ~ V1+V2+V3+V4+V7+V8+V9+V10+V11+V12+V13+V14+V15, data = train)
regPval = fastbw(ols.full, rule = "p", sls = 0.5)
```

```
regPval

Y_calc2 <- Y_calc  # Initialization For calculated values
Y_calc1 <- Y_calc2  # Initialization For calculated values
modeL = regPval

summary(modeL)
#anova(modeL)
attributes(modeL)
residuals(modeL)
sum(residuals(modeL))
mean(residuals(modeL))

#par(mfrow=c(length(test),length(test)))
#gvlma::gvlma(modeL, alphalevel = 0.05)
#dev.off()
#plot(modeL)

# TEST DATA
coefficients(modeL)
# Graph plotted on the basis of Coefficients
pairs(~V1+V2+V3+V7+V8+V9+V14, data = train)

library(car)
scatterplotMatrix(~ V1+V2+V3+V7+V8+V9+V14, data = train)

# VIF factor: No perfect multicollinearity
vif(modeL)

# CODE for getting calculated values of the test model and later comparing them with the
Original values

j = 1
while (j<=nrow(test)) {
  i = 1
  Y_calc1[j,1] <- coef(modeL)[i]

  while (i<(length(coef(modeL))-1)) {
    Y_calc1[j,1] = Y_calc1[j,1] + test[j,i]*coef(modeL)[i+1]
    i=i+1
  }
  j=j+1

}

library(forecast)
Acc_Y_calc1 = accuracy(Y_calc1[,1], Y_og[,1])
CompTab = Y_og
CompTab[,2] = Y_calc1[,1]
CompTab[,3] = Y_og[,1] - Y_calc1[,1]
```

```
CompTab[,4] = CompTab[,3]*CompTab[,3]
library(dplyr)
CompTab <- rename(CompTab, Y_og = V16, Y_calcl1 = V2, error = V3, Sq.error = V4)
CompTab
Y_calc2 <- Y_calc1
Acc_Y_calc2 <- Acc_Y_calc1
Acc_Y_calc2
```

3. Reduced Model: AIC both direction

```
regAICboth <- step(regFull, scope = list(upper=regFull, lower=~1), direction = "both",  k = 2,
trace = 1)
regAICboth
Y_calc3 <- Y_calc  # Initialization For calculated values
Y_calc1 <- Y_calc3  # Initialization For calculated values
modeL = regAICboth

summary(modeL)
anova(modeL)
attributes(modeL)
residuals(modeL)
sum(residuals(modeL))
mean(residuals(modeL))
par(mfrow=c(length(test),length(test)))
gvlma::gvlma(modeL, alphalevel = 0.05)
dev.off()
plot(modeL)

# TEST DATA
coefficients(modeL)

# Graph plotted on the basis of Coefficients
pairs(~V1+V2+V7+V8+V14, data = train)

library(car)
scatterplotMatrix(~ V1+V2+V7+V8+V14, data = train)

# VIF factor: No perfect multicollinearity
#vif(modeL)

# CODE for getting calculated values of the test model and later comparing them with the
Original values
j = 1
while (j<=nrow(test)) {
  i = 1
  Y_calc1[j,1] <- coef(modeL)[i]

  while (i<(length(coef(modeL))-1)) {
```

```
    Y_calc1[j,1] = Y_calc1[j,1] + test[j,i]*coef(modeL)[i+1]
    i=i+1
  }
  j=j+1

}
library(forecast)
Acc_Y_calc1 = accuracy(Y_calc1[,1], Y_og[,1])
CompTab = Y_og
CompTab[,2] = Y_calc1[,1]
CompTab[,3] = Y_og[,1] - Y_calc1[,1]
CompTab[,4] = CompTab[,3]*CompTab[,3]
library(dplyr)
CompTab <- rename(CompTab, Y_og = V16, Y_calcl1 = V2, error = V3, Sq.error = V4)
CompTab
Y_calc3 <- Y_calc1
Acc_Y_calc3 <- Acc_Y_calc1
Acc_Y_calc3
```

4. Reduced Model: # AIC Forward

```
regAICfwd = step(regBase, scope = list(upper=regFull, lower=~1), direction = "forward",  k = 
2, trace = 1)
regAICfwd

Y_calc4 <- Y_calc  # Initialization For calculated values
Y_calc1 <- Y_calc4  # Initialization For calculated values
modeL = regAICfwd

summary(modeL)
anova(modeL)
attributes(modeL)
#residuals(modeL)
mean(residuals(modeL))

par(mfrow=c(length(test),length(test)))
gvlma::gvlma(modeL, alphalevel = 0.05)
dev.off()
plot(modeL)

# TEST DATA
coefficients(modeL)

# Graph plotted on the basis of Coefficients
pairs(~V1+V2+V7+V8+V14, data = train)

library(car)
scatterplotMatrix(~ V1+V2+V7+V8+V14, data = train)
```

```
# VIF factor: No perfect multicollinearity
#vif(modeL)

# CODE for getting calculated values of the test model and later comparing them with the
Original values
j = 1
while (j<=nrow(test)) {
  i = 1
  Y_calc1[j,1] <- coef(modeL)[i]

  while (i<(length(coef(modeL))-1)) {
    Y_calc1[j,1] = Y_calc1[j,1] + test[j,i]*coef(modeL)[i+1]
    i=i+1
  }
  j=j+1

}

library(forecast)
Acc_Y_calc1 = accuracy(Y_calc1[,1], Y_og[,1])
CompTab = Y_og
CompTab[,2] = Y_calc1[,1]
CompTab[,3] = Y_og[,1] - Y_calc1[,1]
CompTab[,4] = CompTab[,3]*CompTab[,3]
library(dplyr)
CompTab <- rename(CompTab, Y_og = V16, Y_calcl1 = V2, error = V3, Sq.error = V4)
CompTab
Y_calc4 <- Y_calc1
Acc_Y_calc4 <- Acc_Y_calc1
Acc_Y_calc4
```

5. Reduced Model: BIC Forward Direction

```
regBICfwd = step(regBase, scope = list(upper=regFull, lower=~1), direction = "forward",  k =
log(length(test)),  trace = 1)
regBICfwd

Y_calc5 <- Y_calc  # Initialization For calculated values
Y_calc1 <- Y_calc5  # Initialization For calculated values
modeL = regBICfwd

summary(modeL)
anova(modeL)
attributes(modeL)
#residuals(modeL)
mean(residuals(modeL))
```

```
par(mfrow=c(length(test),length(test)))
gvlma::gvlma(modeL, alphalevel = 0.05)
dev.off()
plot(modeL)

# TEST DATA
coefficients(modeL)

# Graph plotted on the basis of Coefficients
pairs(~V1+V2+V7+V14, data = train)

library(car)
scatterplotMatrix(~ V1+V2+V7+V14, data = train)

# VIF factor: No perfect multicollinearity
#vif(modeL)

# CODE for getting calculated values of the test model and later comparing them with the
Original values
j = 1
while (j<=nrow(test)) {
  i = 1
  Y_calc1[j,1] <- coef(modeL)[i]

  while (i<(length(coef(modeL))-1)) {
    Y_calc1[j,1] = Y_calc1[j,1] + test[j,i]*coef(modeL)[i+1]
    i=i+1
  }
  j=j+1

}

library(forecast)
Acc_Y_calc1 = accuracy(Y_calc1[,1], Y_og[,1])
CompTab = Y_og
CompTab[,2] = Y_calc1[,1]
CompTab[,3] = Y_og[,1] - Y_calc1[,1]
CompTab[,4] = CompTab[,3]*CompTab[,3]
library(dplyr)
CompTab <- rename(CompTab, Y_og = V16, Y_calcl1 = V2, error = V3, Sq.error = V4)
CompTab
Y_calc5 <- Y_calc1
Acc_Y_calc5 <- Acc_Y_calc1
Acc_Y_calc5
```

6. Reduced Model: BIC Both Direction

```
regBICboth = step(regFull, scope = list(upper=regFull, lower=~1), direction = "both",  k =
log(length(test)),  trace = 1)
regBICboth

Y_calc6 <- Y_calc  # Initialization For calculated values
Y_calc1 <- Y_calc6  # Initialization For calculated values
modeL = regBICboth

summary(modeL)
anova(modeL)
attributes(modeL)
#residuals(modeL)
mean(residuals(modeL))

par(mfrow=c(length(test),length(test)))
gvlma::gvlma(modeL, alphalevel = 0.05)
dev.off()
plot(modeL)

# TEST DATA
coefficients(modeL)

# Graph plotted on the basis of Coefficients
pairs(~V1+V2+V7+V14, data = train)

library(car)
scatterplotMatrix(~ V1+V2+V7+V14, data = train)

# VIF factor: No perfect multicollinearity
#vif(modeL)

# CODE for getting calculated values of the test model and later comparing them with the
Original values
j = 1
while (j<=nrow(test)) {
  i = 1
  Y_calc1[j,1] <- coef(modeL)[i]

  while (i<(length(coef(modeL))-1)) {
    Y_calc1[j,1] = Y_calc1[j,1] + test[j,i]*coef(modeL)[i+1]
    i=i+1
  }
  j=j+1

}

library(forecast)
Acc_Y_calc1 = accuracy(Y_calc1[,1], Y_og[,1])
```

```
CompTab = Y_og
CompTab[,2] = Y_calc1[,1]
CompTab[,3] = Y_og[,1] - Y_calc1[,1]
CompTab[,4] = CompTab[,3]*CompTab[,3]
library(dplyr)
CompTab <- rename(CompTab, Y_og = V16, Y_calcl1 = V2, error = V3, Sq.error = V4)
CompTab
Y_calc6 <- Y_calc1
Acc_Y_calc6 <- Acc_Y_calc1
Acc_Y_calc6
```

6. Comparing all the Models

```
anova(regFull, regAICboth, regAICfwd, regBICfwd, regBICboth)
AIC(regFull, regAICboth, regAICfwd, regBICfwd, regBICboth)
BIC(regFull, regAICboth, regAICfwd, regBICfwd, regBICboth)
#CompModel<-as.data.frame(matrix(nrow=4,ncol=6))
#colnames(CompModel)<-c("regFull","regAICboth","regAICfwd","regBICfwd","regBICboth")
#
ModelNames <- c("regFull","regAICboth","regAICfwd","regBICfwd","regBICboth","regPval")
ModelAcc <- c(Acc_Y_calc1, Acc_Y_calc3, Acc_Y_calc4, Acc_Y_calc5, Acc_Y_calc6,
Acc_Y_calc2)
AdjRsq <- c(summary(regFull)$adj.r.squared, summary(regAICboth)$adj.r.squared,
summary(regAICfwd)$adj.r.squared, summary(regBICfwd)$adj.r.squared,
summary(regBICboth)$adj.r.squared, NA)
Rsq <- c(summary(regFull)$r.squared, summary(regAICboth)$r.squared,
summary(regAICfwd)$r.squared, summary(regBICfwd)$r.squared,
summary(regBICboth)$r.squared, NA)
MeanErr <- c(Acc_Y_calc1[,1], Acc_Y_calc3[,1], Acc_Y_calc4[,1], Acc_Y_calc5[,1],
Acc_Y_calc6[,1], Acc_Y_calc2[,1])
RootMeanSqErr <- c(Acc_Y_calc1[,2], Acc_Y_calc3[,2], Acc_Y_calc4[,2], Acc_Y_calc5[,2],
Acc_Y_calc6[,2], Acc_Y_calc2[,2])
MeanAbsErr <- c(Acc_Y_calc1[,3], Acc_Y_calc3[,3], Acc_Y_calc4[,3], Acc_Y_calc5[,3],
Acc_Y_calc6[,3], Acc_Y_calc2[,3])
MeanPercentErr <- c(Acc_Y_calc1[,4], Acc_Y_calc3[,4], Acc_Y_calc4[,4], Acc_Y_calc5[,4],
Acc_Y_calc6[,4], Acc_Y_calc2[,4])
MeanAbsPercentErr <- c(Acc_Y_calc1[,5], Acc_Y_calc3[,5], Acc_Y_calc4[,5],
Acc_Y_calc5[,5], Acc_Y_calc6[,5], Acc_Y_calc2[,5])
Coeff <- c(coefficients(regFull), coefficients(regAICboth),
coefficients(regAICfwd),coefficients(regBICfwd), coefficients(regBICboth))
Coeff
CompareModel <- data.frame(ModelNames, Rsq, AdjRsq, MeanErr, MeanAbsErr,
MeanPercentErr, MeanAbsPercentErr)
CompareModel[1:6,]

broom::glance(regFull)
broom::glance(regAICboth)
broom::glance(regAICfwd)
```

```
broom::glance(regBICfwd)
broom::glance(regBICboth)
broom::glance(regPval)
```

REFERENCES

[1] GLASSER, M., and GREENBURO, L. (1971). “Air Pollution Mortality and Weather,” Archives Environmental Health bb, 334-343.

[2] HOLLAND, W. W., SPICER, C. C., and WILSON, J. M. G. (1961). “Influence of the Weather on Respiratory and Heart Disease,” The Lancet d, 338-341.

[3] OECHSLI, F. W. and BUECHLEY, R. W. (1970). “Excess Mortality Associated with Three Los Angeles September Hot Spells,” Environmental Research S, 277-284.

[4] BENEDICT, H. M. (1971). “Plant Damage by Air Pollutants: CRC-APRAC Project No. CAPA-2-68,” presented at the Automotive Air Pollution Res. Symp., Chicago.

[5] REGRESSION - Linear Regression Datasets (fsu.edu)

[6] DANIEL, C., and WOOD, F. S. (1971). Fitting Equations to Data (Computer Analysis of Multifactor Data for Scientists and Engineers), John Wiley.

[7] HEXTER, A. C., and GOLDSMITH, J. R. (1971). “Carbon Monoxide: Association of Community Air Pollution and Mortality,” Science I?‘d, 265-268.

[8] HICKEY, R. J., BOYCE, D. E., HARNER, E. B., and CLELLAND, R. C. (1970). “Ecological Statistical Studies Concerning Environmental Pollution and Chronic Disease,” IEEE Transactions on Geoscience Electronics GE-g, 186-202.

[9] OECHSLI, F. W. and BUECHLEY, R. W. (1970). “Excess Mortality Associated with Three Los Angeles September Hot Spells,” Environmental Research S, 277-284.

[10] SHY, C. M., CREASON, J. P., PEARLMAN, M. D., MCCLAIN, K. E., BENSON, F. B., and YOUNG, M. M. (1970). “The Chattanooga School Children Study: Effects of Community Exposure, to Nitrogen Dioxide. I. Methods, Description of Pollutant Exposure, and Results of Ventilatory Function Testing,” J. Air Pollution Control Assoc. $0, 539-545.