Predictive Modelling DS432 - Regression

Saha Debanshee Gopal - U101113FCS074

4th October 2016

Question 1

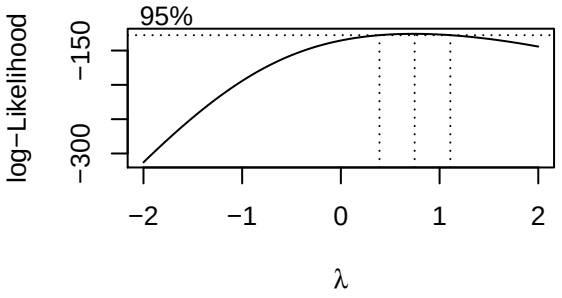
RegD1

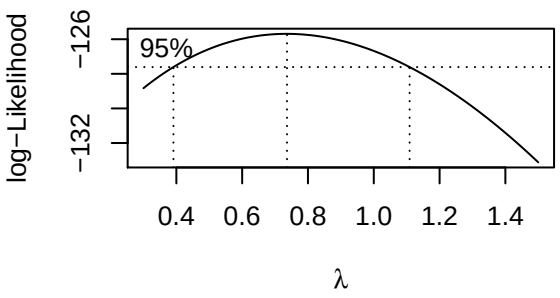
```
##
## Call:
## lm(formula = frml, data = trndata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0886 -1.5006  0.0718  1.1500  3.4174
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.60337    0.34712   7.500 5.08e-11 ***
## X1           0.19608    0.03679   5.330 7.65e-07 ***
## X2           0.01041    0.64983   0.016    0.987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.782 on 87 degrees of freedom
## Multiple R-squared:  0.2489, Adjusted R-squared:  0.2317
## F-statistic: 14.42 on 2 and 87 DF,  p-value: 3.912e-06
```

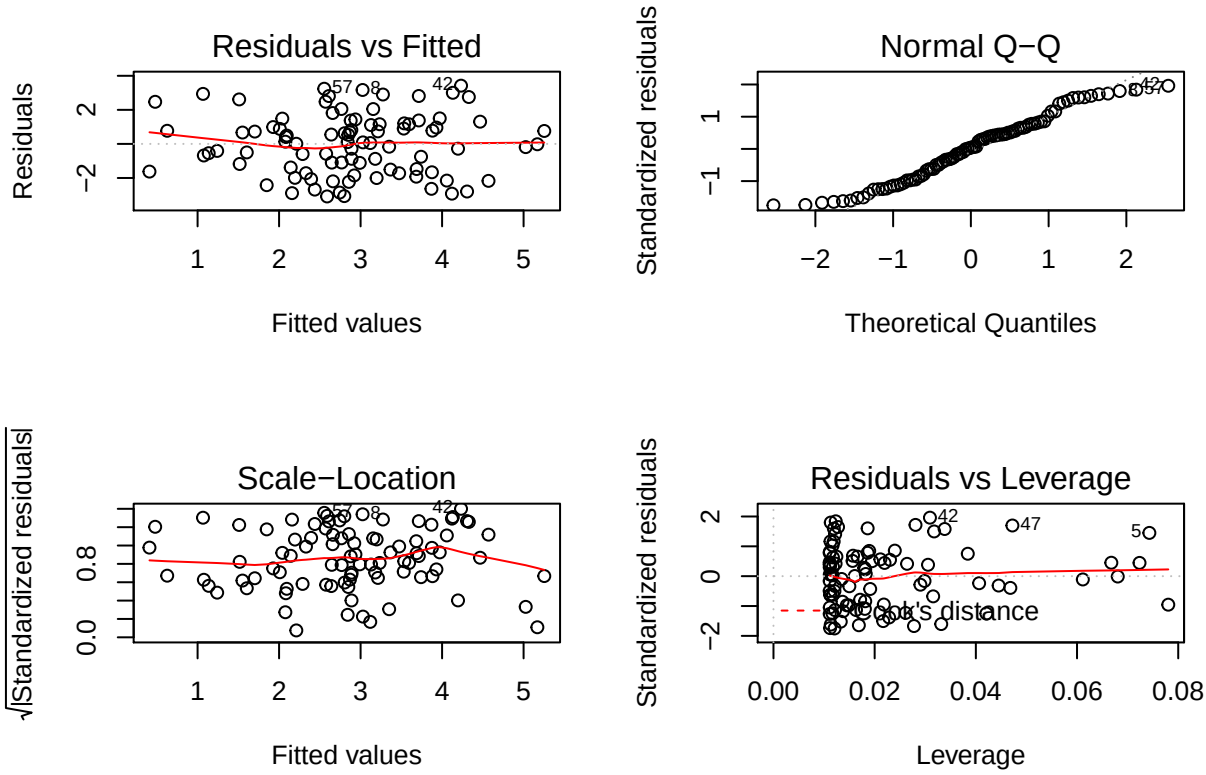
Model fit is not good. Transforming and re-evaluating.

```
##
## Call:
## lm(formula = Y1 ~ X1 + I(X2^pp$X2$lambda), data = trndata, subset = (cook <
##     max(cook)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0676 -1.4910  0.0895  1.1103  3.3458
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.4882     0.4669   5.329 7.83e-07 ***
## X1                   0.2077     0.0369   5.630 2.22e-07 ***
## I(X2^pp$X2$lambda)   0.1178     0.7296   0.161    0.872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.763 on 86 degrees of freedom
## Multiple R-squared:  0.2712, Adjusted R-squared:  0.2543
## F-statistic:    16 on 2 and 86 DF,  p-value: 1.233e-06
```

```
##
## Call:
## lm(formula = Y1 ~ X1, data = trndata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0848 -1.4992  0.0697  1.1549  3.4198
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.60797    0.19373   13.46  < 2e-16 ***
## X1           0.19615    0.03632    5.40 5.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.772 on 88 degrees of freedom
## Multiple R-squared:  0.2489, Adjusted R-squared:  0.2404
## F-statistic: 29.16 on 1 and 88 DF,  p-value: 5.587e-07
```







The variable X2 is eliminated using backward elimination

```
##     X1     X2
## 1.0143 1.0143
```

The VIF value is less than 5 implying low co-linearity

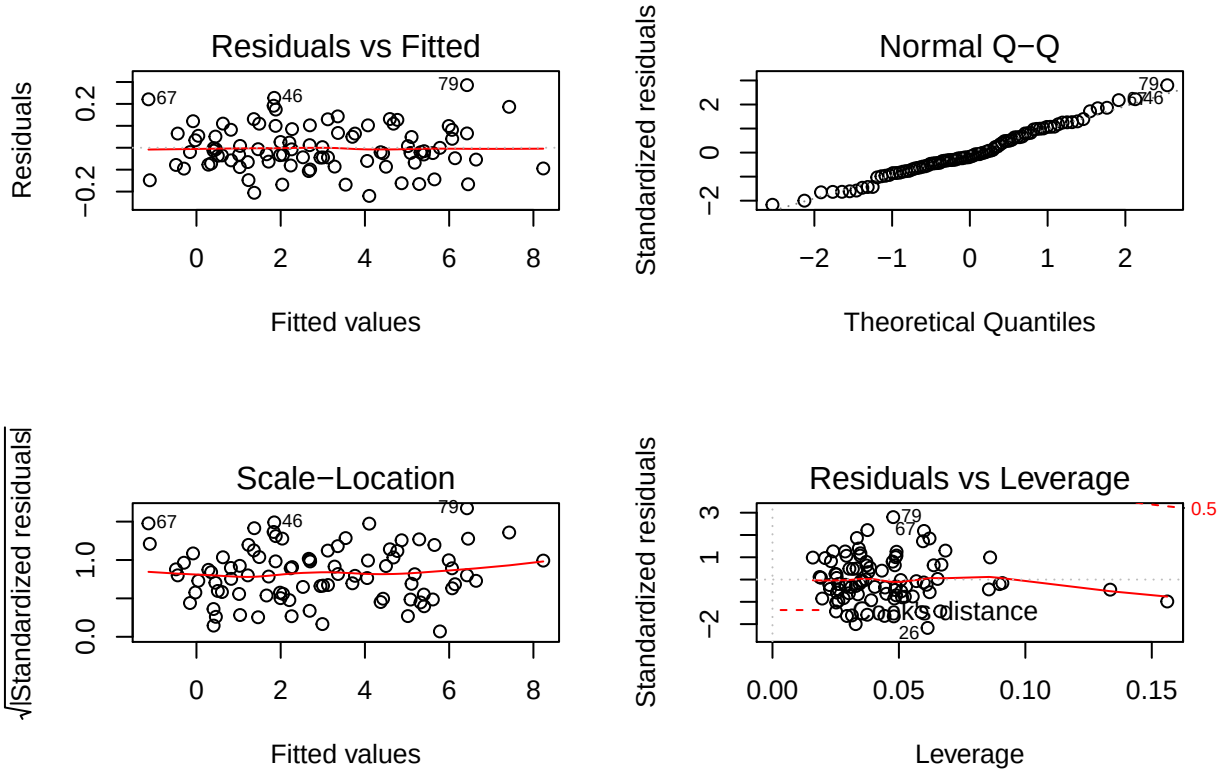
Correlation

```
## [1] 0.714581
```

RegD2

```
##
## Call:
## lm(formula = frml, data = trndata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21982 -0.06485 -0.01780  0.06692  0.28568
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.049007   0.026746  -1.832   0.0704 .
## X1           0.200967   0.002325  86.426   <2e-16 ***
## X2          -0.448742   0.037525 -11.958   <2e-16 ***
## X3           0.706049   0.003856 183.101   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1045 on 86 degrees of freedom
## Multiple R-squared:  0.9978, Adjusted R-squared:  0.9978
## F-statistic: 1.325e+04 on 3 and 86 DF,  p-value: < 2.2e-16
```




```
## Start:  AIC=-402.62
## Y1 ~ X1 + X2 + X3
##
##        Df Sum of Sq    RSS     AIC
## <none>                0.94 -402.62
## - X2    1      1.56   2.50 -316.47
## - X1    1     81.58  82.52   -1.81
## - X3    1    366.18 367.12  132.53
```

The VIF value is less than 5 implying low co-linearity

```
##     X1     X2     X3
## 1.0065 1.0098 1.0150
```

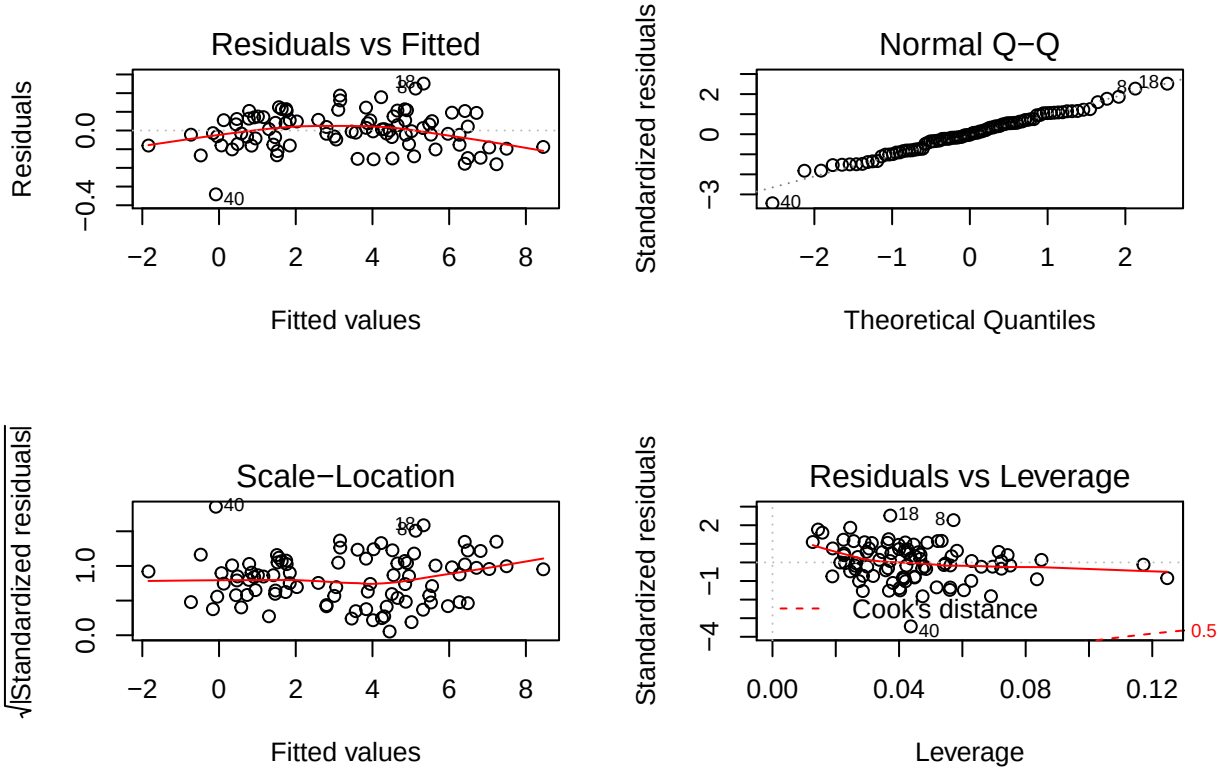
Correlation

```
## [1] 0.9983277
```

RegD3

```
##
```

```
## Call:
## lm(formula = frml, data = trndata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34173 -0.07485 -0.00193  0.06305  0.25112
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.027233   0.028608   0.952    0.344
## X1           0.202293   0.002453  82.479   <2e-16 ***
## X2          -1.004993   0.039089 -25.710   <2e-16 ***
## X4           1.162795   0.006041 192.484   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1016 on 86 degrees of freedom
## Multiple R-squared:  0.9981, Adjusted R-squared:  0.9981
## F-statistic: 1.536e+04 on 3 and 86 DF,  p-value: < 2.2e-16
```




```
## Start:  AIC=-407.7
## Y1 ~ X1 + X2 + X4
##
##        Df Sum of Sq    RSS     AIC
## <none>                 0.89 -407.70
## - X2    1      6.82   7.71 -215.15
```

```
## - X1    1     70.22  71.11  -15.21
## - X4    1    382.43 383.32  136.42
```

The VIF value is less than 5 implying low co-linearity

```
##     X1     X2     X4
## 1.0084 1.0075 1.0030
```

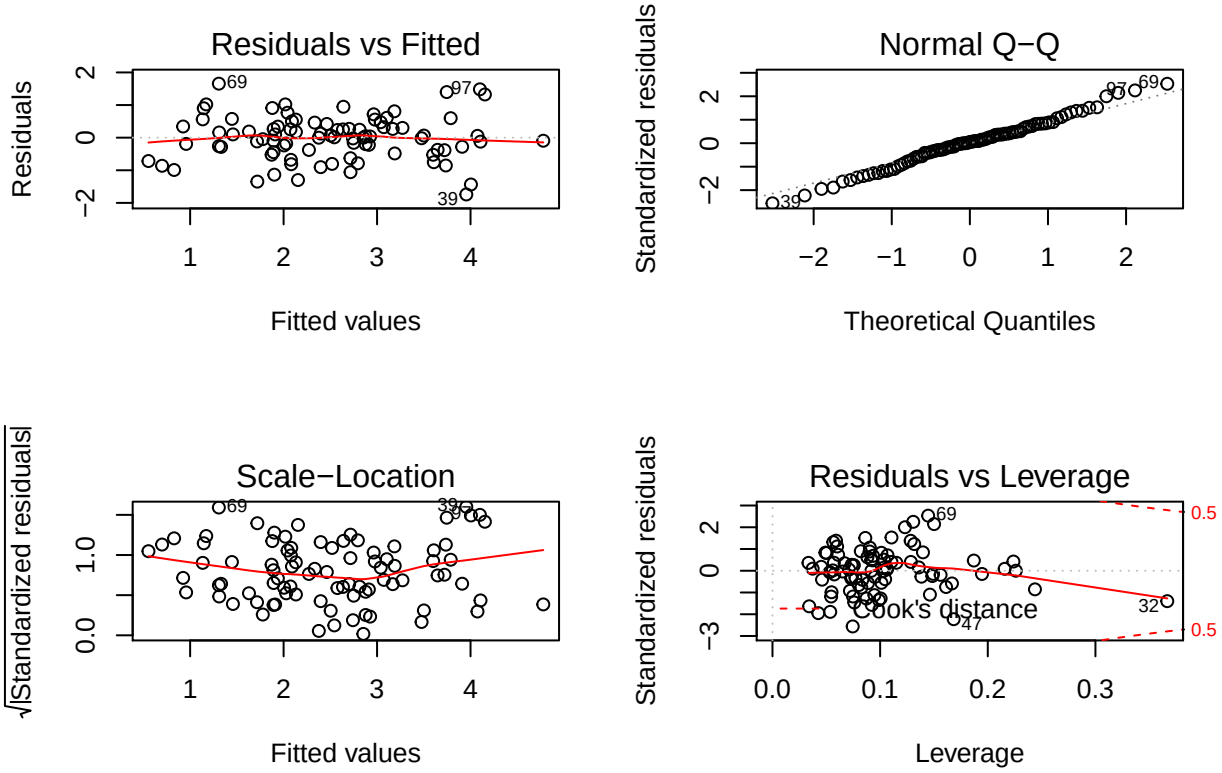
Correlation

```
## [1] 0.9972059
```

Question 2

Summary of the data :

```
##
## Call:
## lm(formula = frml, data = trndata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74127 -0.38032  0.03694  0.38065  1.65529
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.169553   1.356734   0.862 0.391310
## X1           0.539640   0.093852   5.750 1.66e-07 ***
## X2           0.416647   0.177288   2.350 0.021297 *
## X3          -0.024590   0.011482  -2.142 0.035341 *
## X4           0.110637   0.062704   1.764 0.081575 .
## X5           0.862998   0.248862   3.468 0.000857 ***
## X6          -0.081674   0.097436  -0.838 0.404465
## X7           0.046457   0.168551   0.276 0.783565
## X8           0.003688   0.004578   0.805 0.422988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7067 on 78 degrees of freedom
## Multiple R-squared:  0.6551, Adjusted R-squared:  0.6198
## F-statistic: 18.52 on 8 and 78 DF,  p-value: 3.126e-15
```




```
## Start:  AIC=-51.9
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X7    1    0.0379 38.996 -53.814
## - X8    1    0.3241 39.282 -53.178
## - X6    1    0.3509 39.309 -53.118
## <none>               38.958 -51.898
## - X4    1    1.5549 40.513 -50.494
## - X3    1    2.2909 41.249 -48.927
## - X2    1    2.7585 41.716 -47.946
## - X5    1    6.0062 44.964 -41.424
## - X1    1   16.5128 55.471 -23.155
##
## Step:  AIC=-53.81
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X6    1    0.3338 39.329 -55.072
## - X8    1    0.6444 39.640 -54.388
## <none>               38.996 -53.814
## - X4    1    1.5771 40.573 -52.365
## - X3    1    2.2532 41.249 -50.927
## - X2    1    2.7213 41.717 -49.945
## - X5    1    5.9790 44.975 -43.403
## - X1    1   17.5039 56.500 -23.556
```

```
##
## Step:  AIC=-55.07
## Y ~ X1 + X2 + X3 + X4 + X5 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X8    1    0.3524 39.682 -56.296
## <none>              39.329 -55.072
## - X4    1    1.7201 41.050 -53.348
## - X3    1    2.0543 41.384 -52.643
## - X2    1    2.6438 41.973 -51.412
## - X5    1    5.7576 45.087 -45.186
## - X1    1   19.7690 59.099 -21.643
##
## Step:  AIC=-56.3
## Y ~ X1 + X2 + X3 + X4 + X5
##
##        Df Sum of Sq    RSS     AIC
## <none>              39.682 -56.296
## - X3    1    1.7855 41.467 -54.467
## - X4    1    1.9413 41.623 -54.141
## - X2    1    2.4821 42.164 -53.018
## - X5    1    7.3300 47.012 -43.549
## - X1    1   22.5569 62.239 -19.139
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = trndata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83288 -0.42752  0.08003  0.41092  1.62176
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.45460    0.87636   1.660 0.100817
## X1           0.52600    0.07752   6.786 1.74e-09 ***
## X2           0.39094    0.17368   2.251 0.027103 *
## X3          -0.02077    0.01088  -1.909 0.059790 .
## X4           0.12215    0.06136   1.991 0.049894 *
## X5           0.83983    0.21712   3.868 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6999 on 81 degrees of freedom
## Multiple R-squared:  0.6487, Adjusted R-squared:  0.627
## F-statistic: 29.92 on 5 and 81 DF,  p-value: < 2.2e-16
```

Data nature

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
```

```
##
##     as.Date, as.Date.numeric
##
##  Breusch-Pagan test
##
## data:  regressionAIC
## BP = 4.8583, df = 5, p-value = 0.4334
```

The test shows that the data is heteroscedastic in nature. The residuals vs the fitted values graph suggests constant variance

Normality Test

```
##
##  Shapiro-Wilk normality test
##
## data:  regressionAIC$residuals
## W = 0.99192, p-value = 0.8739
```

Given sample is normalzied

Leverage points

```
##          2          3          4          6          9         12
## 0.07360412 0.13892324 0.08356536 0.09303144 0.10287667 0.09325019
##         19         32         33         38         39         47
## 0.13303370 0.36600555 0.06641019 0.09831964 0.06263047 0.09457382
##         49         55         57         58         62         63
## 0.11935763 0.09285728 0.08467047 0.08907343 0.08235008 0.09433647
##         64         69         70         72         73         74
## 0.08339922 0.14255116 0.08563351 0.07640520 0.09545987 0.08482798
##         75         76         78         79         80         83
## 0.07252479 0.07211625 0.08813566 0.06730275 0.07927584 0.08653123
##         86         88         89         90         92         93
## 0.06809809 0.06821115 0.14041047 0.09446324 0.07835639 0.07017668
##         94         95         96
## 0.19271042 0.09596020 0.06927420
```

Outliers

```
##
## Attaching package: 'car'
## The following object is masked from 'package:DAAG':
##
##     vif
## The following object is masked from 'package:usdm':
##
##     vif
##
## No Studentized residuals with Bonferonni p < 0.05
```

```
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 39 -2.818263          0.0060821      0.52914
```

```
##          Y         X1         X2         X3         X4         X5
##   5.582930  -1.347074   6.107600  41.000000   2.326302   1.000000
##         X6         X7         X8
##   2.904170   9.000000 100.000000
```

Influential Points

```
## Potentially influential observations of
##   lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = trndata) :
##
##    dfb.1_ dfb.X1 dfb.X2  dfb.X3 dfb.X4 dfb.X5 dffit   cov.r   cook.d
## 19   0.05  -0.03   0.00   -0.06   0.00   0.01   0.09    1.24_*  0.00
## 32   0.61   0.28  -1.02_*  0.10   0.21   0.05  -1.08_*  1.46_*  0.19
## 39   0.04   0.01  -0.08    0.03  -0.25  -0.49  -0.73    0.65_*  0.08
## 69  -0.71  -0.54   0.72    0.36  -0.68  -0.02   1.06_*  0.78_*  0.17
## 89   0.04   0.00  -0.05    0.00   0.04  -0.02  -0.07    1.25_*  0.00
## 94  -0.02  -0.02  -0.01    0.03   0.00  -0.01  -0.05    1.33_*  0.00
##    hat
## 19  0.13
## 32  0.37_*
## 39  0.06
## 69  0.14
## 89  0.14
## 94  0.19
```

Condition Number

```
##
## Attaching package: 'perturb'
```

```
## The following object is masked from 'package:raster':
##
##     reclassify
```

```
## Condition
## Index    Variance Decomposition Proportions
##           intercept X1    X2    X3    X4    X5
## 1   1.000 0.000     0.013 0.001 0.001 0.001 0.014
## 2   1.986 0.000     0.003 0.000 0.000 0.605 0.061
## 3   2.404 0.001     0.029 0.001 0.001 0.135 0.422
## 4   4.332 0.001     0.942 0.002 0.001 0.008 0.488
## 5  18.668 0.004     0.000 0.655 0.435 0.018 0.001
## 6  29.119 0.994     0.012 0.341 0.562 0.232 0.013
```

VIF

```
##       X1       X2       X3       X4       X5
## 1.494402 1.291066 1.195905 1.364675 1.482044
```

VIF value if less than 5, implying low col

Correlaton

```
## [1] 0.8073483
```

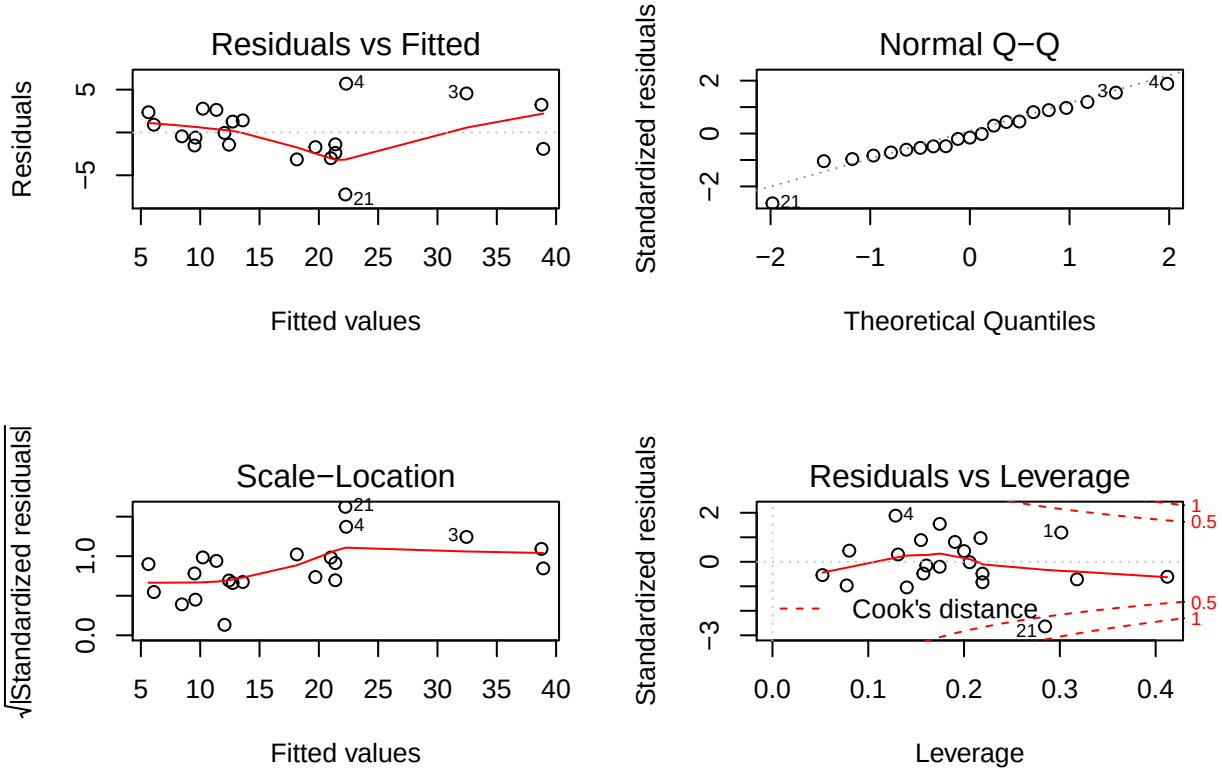
Correlation value is high

Question 3

```
##    Y1 X1 X2 X3
## 1  42 80 27 89
## 2  37 80 27 88
## 3  37 75 25 90
## 4  28 62 24 87
## 5  18 62 22 87
## 6  18 62 23 87
## 7  19 62 24 93
## 8  20 62 24 93
## 9  15 58 23 87
## 10 14 58 18 80
## 11 14 58 18 89
## 12 13 58 17 88
## 13 11 58 18 82
## 14 12 58 19 93
## 15  8 50 18 89
## 16  7 50 18 86
## 17  8 50 19 72
## 18  8 50 19 79
## 19  9 50 20 80
## 20 15 56 20 82
## 21 15 70 20 91
```

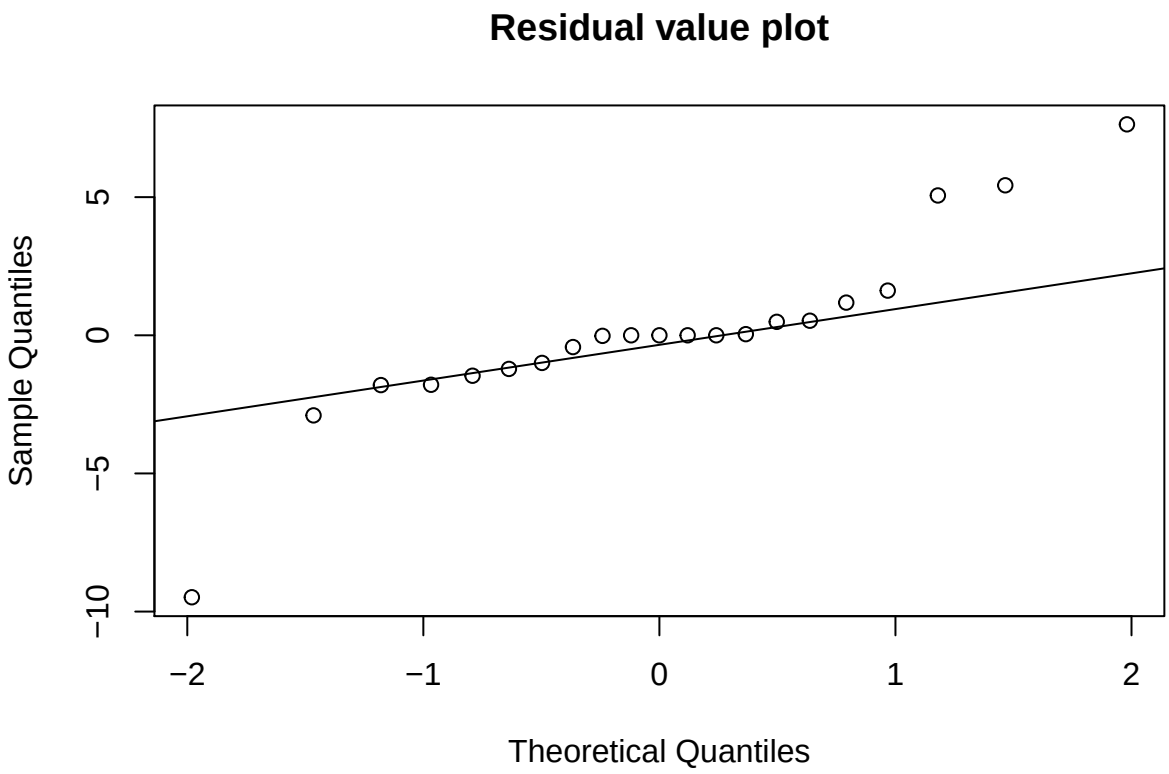
Least Squares

```
##
## Call:
## lm(formula = frml, data = reg2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2377 -1.7117 -0.4551  2.3614  5.6978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.9197    11.8960  -3.356  0.00375 **
## X1            0.7156     0.1349   5.307  5.8e-05 ***
## X2            1.2953     0.3680   3.520  0.00263 **
## X3           -0.1521     0.1563  -0.973  0.34405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 17 degrees of freedom
## Multiple R-squared:  0.9136, Adjusted R-squared:  0.8983
## F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```



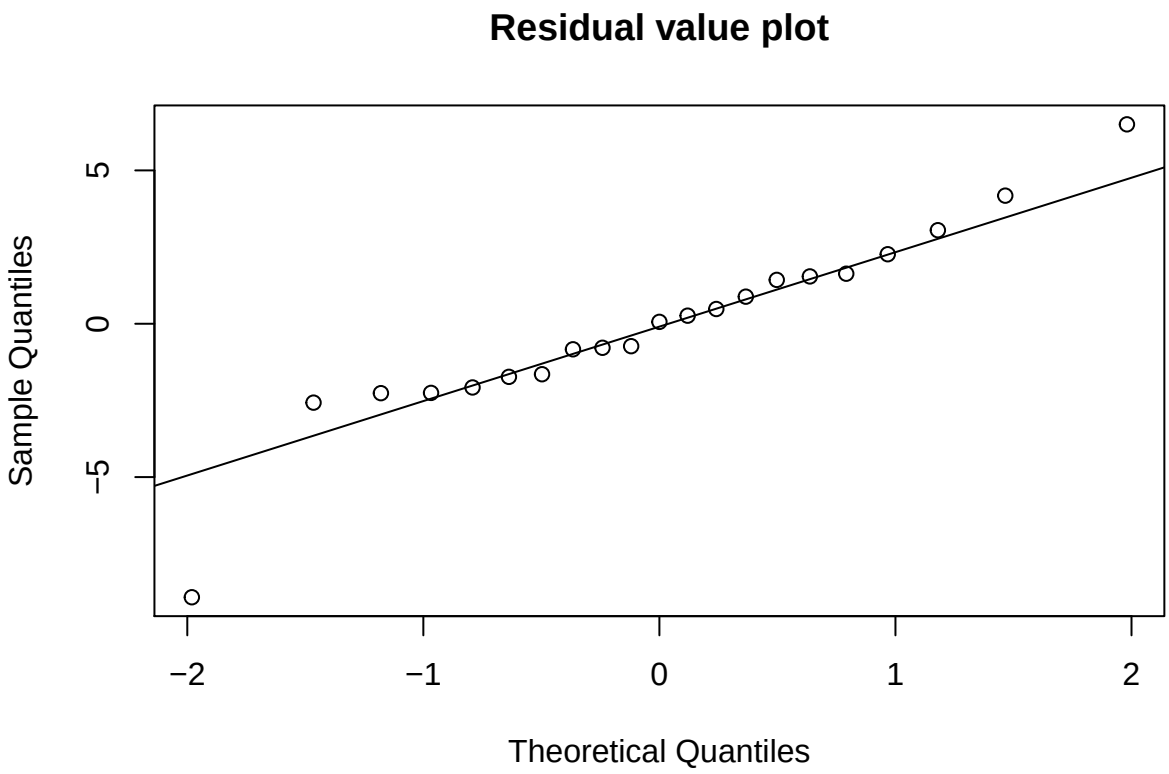
Least Abosolute Deviations

```
##                Length Class      Mode
## call            4     -none-     call
## dims            2     -none-     numeric
## coefficients    4     -none-     numeric
## scale           1     -none-     numeric
## minimum         1     -none-     numeric
## fitted.values  21     -none-     numeric
## residuals      21     -none-     numeric
## numIter         1     -none-     numeric
## control         4     -none-     numeric
## weights        21     -none-     numeric
## logLik          1     -none-     numeric
## speed           5     proc_time  numeric
## converged       1     -none-     logical
## xlevels         0     -none-     list
## terms           3     terms      call
## model           4     data.frame list
```



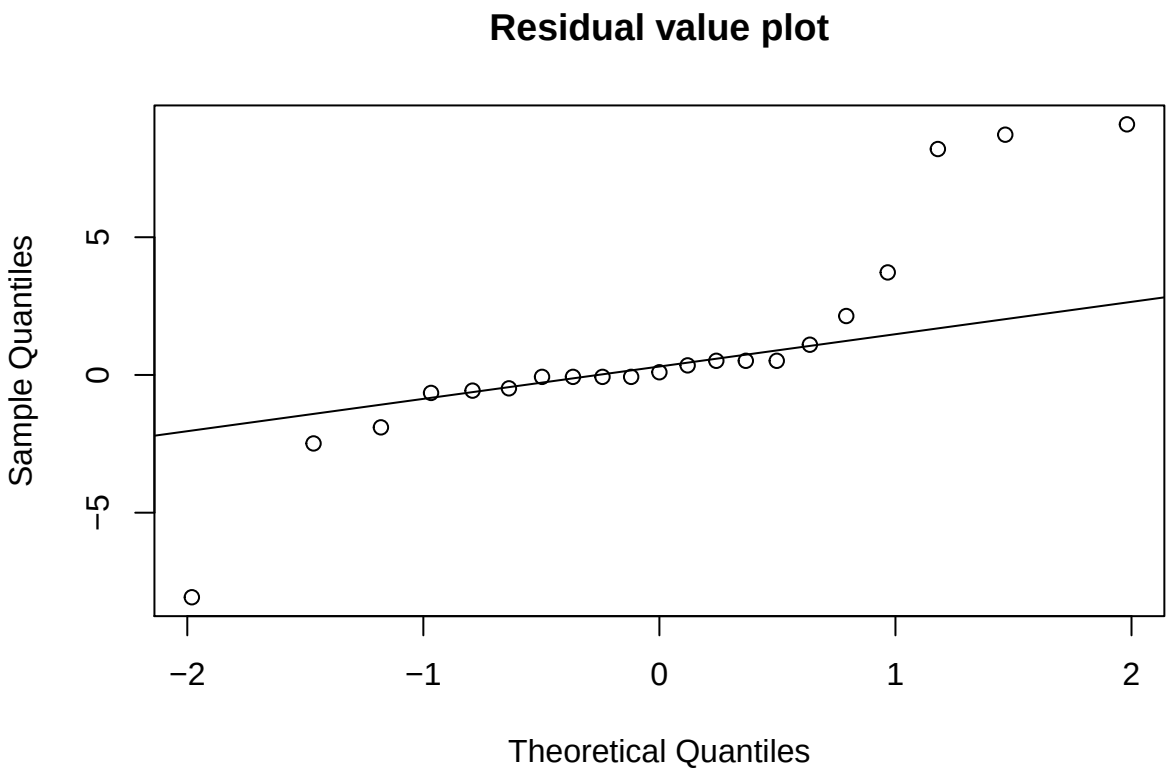
Huber Method

```
##
## Call: rlm(formula = frml, data = reg2)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.91753 -1.73127  0.06187  1.54306  6.50163
##
## Coefficients:
##             Value    Std. Error t value
## (Intercept) -41.0265   9.8073    -4.1832
## X1            0.8294   0.1112     7.4597
## X2            0.9261   0.3034     3.0524
## X3           -0.1278   0.1289    -0.9922
##
## Residual standard error: 2.441 on 17 degrees of freedom
```



Least Trimmed Squares

```
##                Length Class      Mode
## crit            1     -none-     numeric
## sing            1     -none-     character
## coefficients    4     -none-     numeric
## bestone         4     -none-     numeric
## fitted.values 21     -none-     numeric
## residuals     21     -none-     numeric
## scale           2     -none-     numeric
## terms           3     terms      call
## call            4     -none-     call
## xlevels         0     -none-     list
## model           4     data.frame list
```



The difference in the graph is shown in the graph which indicates a change in the line as well the distributed values in the plot of the residual values.

Result of Least Square after removal of outliers and influential points

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 21 -3.330493           0.004238     0.088999
```

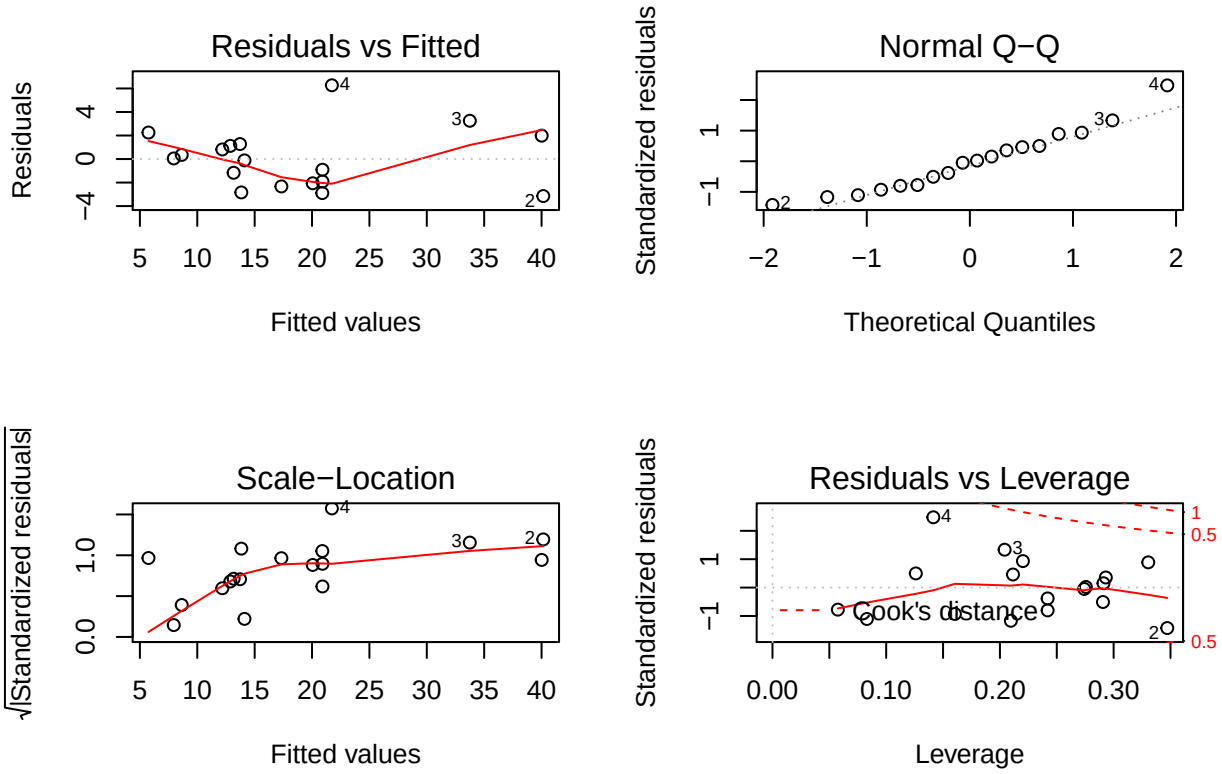
```
## Y1 X1 X2 X3
## 42 80 27 72
```

```
## Y1 X1 X2 X3
##  7 50 17 93
```

```
## Potentially influential observations of
##   lm(formula = frml, data = reg2) :
##
##    dfb.1_ dfb.X1  dfb.X2  dfb.X3 dffit   cov.r   cook.d hat
## 17 -0.46   0.02   -0.06    0.42  -0.50    1.98_*  0.07   0.41
## 21  0.40  -1.62_*  1.64_* -0.36  -2.10_*  0.22_*  0.69   0.28
```

```
##                   2.5 %      97.5 %
## (Intercept) -65.0180339 -14.8213150
## X1            0.4311143   1.0001661
## X2            0.5188228   2.0717495
```

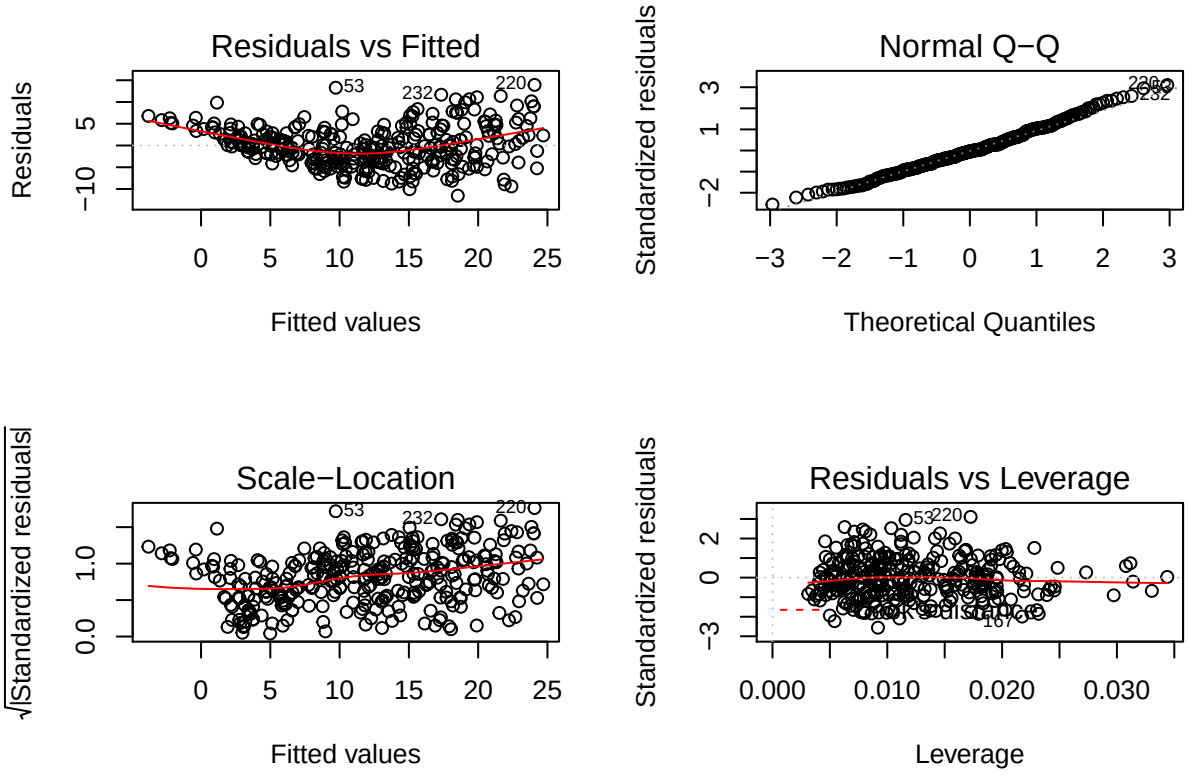
```
## X3           -0.4818741   0.1776291
##
## Call:
## lm(formula = frml, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1485 -2.0228 -0.0323  1.2354  6.2700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41.6803    12.8789  -3.236  0.00597 **
## X1            0.8918     0.1286   6.936 6.92e-06 ***
## X2            0.8346     0.3502   2.383  0.03189 *
## X3           -0.1369     0.1642  -0.833  0.41863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.732 on 14 degrees of freedom
## Multiple R-squared:  0.943,  Adjusted R-squared:  0.9308
## F-statistic: 77.22 on 3 and 14 DF,  p-value: 5.979e-09
```



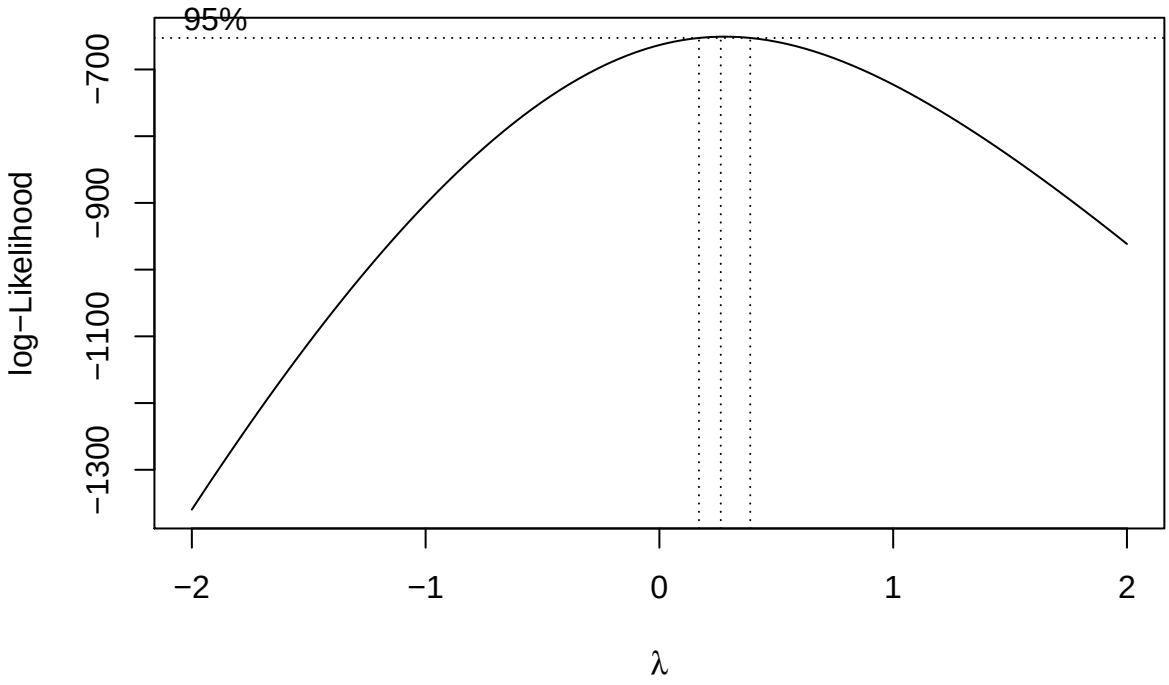
Question 4

Before transformation

```
##
## Call:
## lm(formula = frml, data = regD7)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5291  -3.0137  -0.2249   2.8239  13.9303
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.049e+01  1.616e+00  -6.492 3.16e-10 ***
## X3           7.738e-02  1.339e-02   5.777 1.77e-08 ***
## X4           3.296e-01  2.109e-02  15.626  < 2e-16 ***
## X5          -1.004e-03  1.639e-04  -6.130 2.54e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.524 on 326 degrees of freedom
## Multiple R-squared:  0.684,  Adjusted R-squared:  0.6811
## F-statistic: 235.2 on 3 and 326 DF,  p-value: < 2.2e-16
```

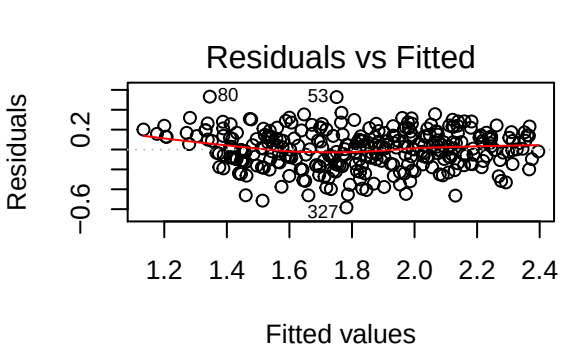


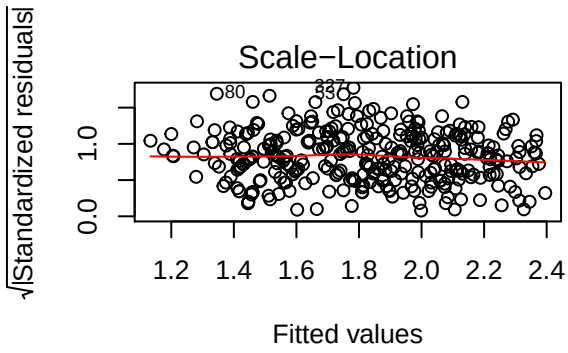
After transform

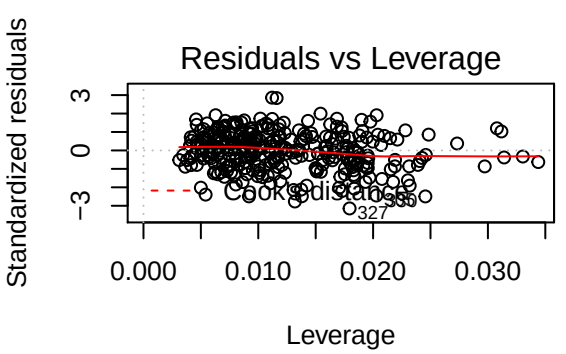


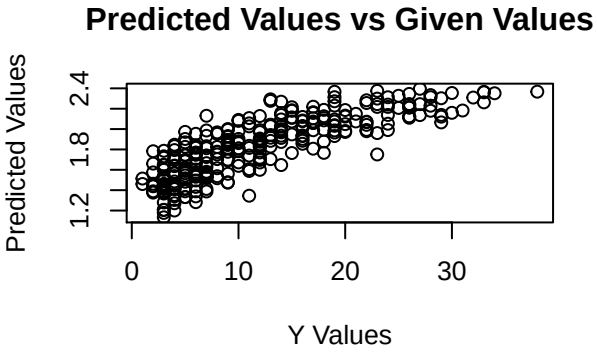

```
##
## Call:
## lm(formula = I(regD7$Y^lambda) ~ regD7$X3 + regD7$X4 + regD7$X5,
##     data = regD7)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58271 -0.11566  0.00969  0.13409  0.53169
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.963e-01  6.679e-02  13.420  < 2e-16 ***
## regD7$X3     3.227e-03  5.534e-04   5.831 1.33e-08 ***
## regD7$X4     1.413e-02  8.716e-04  16.218  < 2e-16 ***
## regD7$X5    -5.229e-05  6.770e-06  -7.723 1.41e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1869 on 326 degrees of freedom
## Multiple R-squared:  0.7157, Adjusted R-squared:  0.7131
## F-statistic: 273.5 on 3 and 326 DF,  p-value: < 2.2e-16
```

```
## [1] 0.8263281
```


Residuals vs Fitted
Residuals
Fitted values
80
53
327
−0.6
0.2
1.2
1.4
1.6
1.8
2.0
2.2
2.4


Scale−Location
√|Standardized residuals|
Fitted values
80
337
0.0
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
327
−3
0
3
0.000
0.010
0.020
0.030

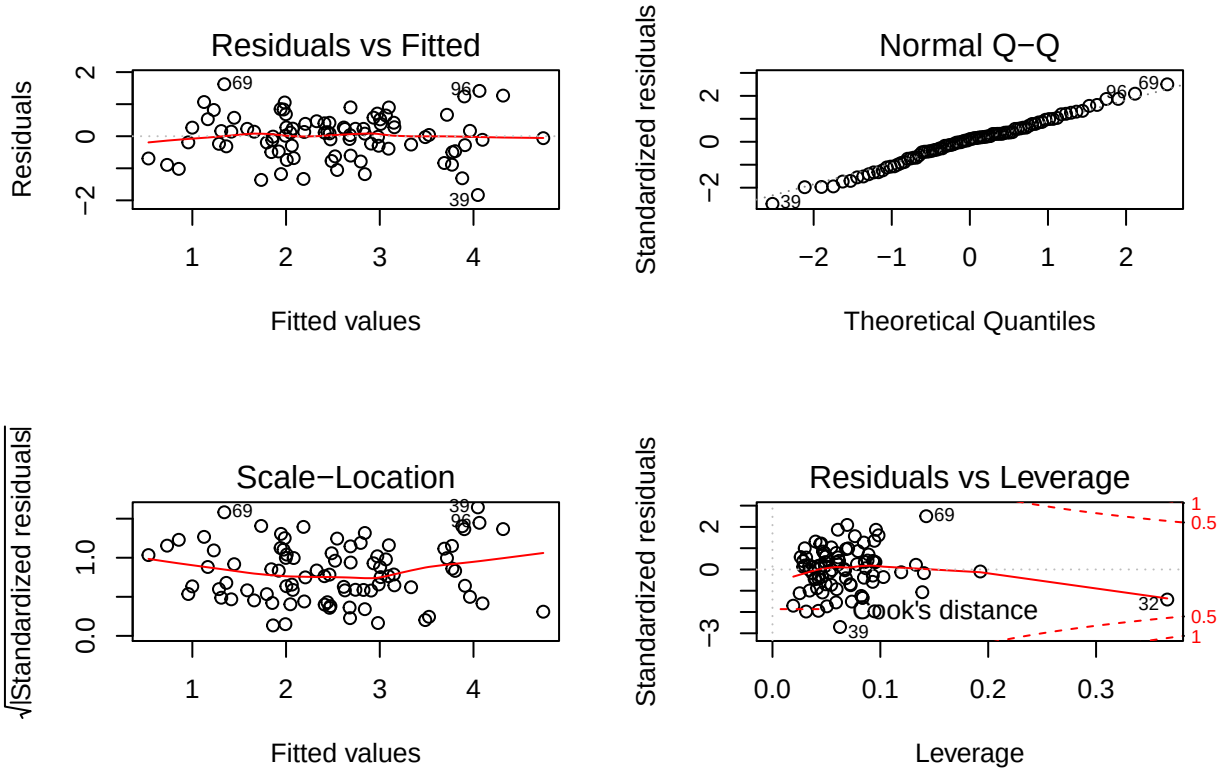

Predicted Values vs Given Values
Predicted Values
Y Values
1.2
1.8
2.4
0
10
20
30

Question 5

Backward Elimination

```
## Start:  AIC=-51.9
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X7    1    0.0379 38.996 -53.814
## - X8    1    0.3241 39.282 -53.178
## - X6    1    0.3509 39.309 -53.118
## <none>              38.958 -51.898
## - X4    1    1.5549 40.513 -50.494
## - X3    1    2.2909 41.249 -48.927
## - X2    1    2.7585 41.716 -47.946
## - X5    1    6.0062 44.964 -41.424
## - X1    1   16.5128 55.471 -23.155
##
## Step:  AIC=-53.81
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X6    1    0.3338 39.329 -55.072
## - X8    1    0.6444 39.640 -54.388
## <none>              38.996 -53.814
## - X4    1    1.5771 40.573 -52.365
## - X3    1    2.2532 41.249 -50.927
## - X2    1    2.7213 41.717 -49.945
## - X5    1    5.9790 44.975 -43.403
## - X1    1   17.5039 56.500 -23.556
##
## Step:  AIC=-55.07
## Y ~ X1 + X2 + X3 + X4 + X5 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X8    1    0.3524 39.682 -56.296
## <none>              39.329 -55.072
## - X4    1    1.7201 41.050 -53.348
## - X3    1    2.0543 41.384 -52.643
## - X2    1    2.6438 41.973 -51.412
## - X5    1    5.7576 45.087 -45.186
## - X1    1   19.7690 59.099 -21.643
##
## Step:  AIC=-56.3
## Y ~ X1 + X2 + X3 + X4 + X5
##
##        Df Sum of Sq    RSS     AIC
## <none>              39.682 -56.296
## - X3    1    1.7855 41.467 -54.467
## - X4    1    1.9413 41.623 -54.141
## - X2    1    2.4821 42.164 -53.018
## - X5    1    7.3300 47.012 -43.549
## - X1    1   22.5569 62.239 -19.139
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = trndata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83288 -0.42752  0.08003  0.41092  1.62176
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.45460    0.87636   1.660 0.100817
## X1           0.52600    0.07752   6.786 1.74e-09 ***
## X2           0.39094    0.17368   2.251 0.027103 *
## X3          -0.02077    0.01088  -1.909 0.059790 .
## X4           0.12215    0.06136   1.991 0.049894 *
## X5           0.83983    0.21712   3.868 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6999 on 81 degrees of freedom
## Multiple R-squared:  0.6487, Adjusted R-squared:  0.627
## F-statistic: 29.92 on 5 and 81 DF,  p-value: < 2.2e-16
```



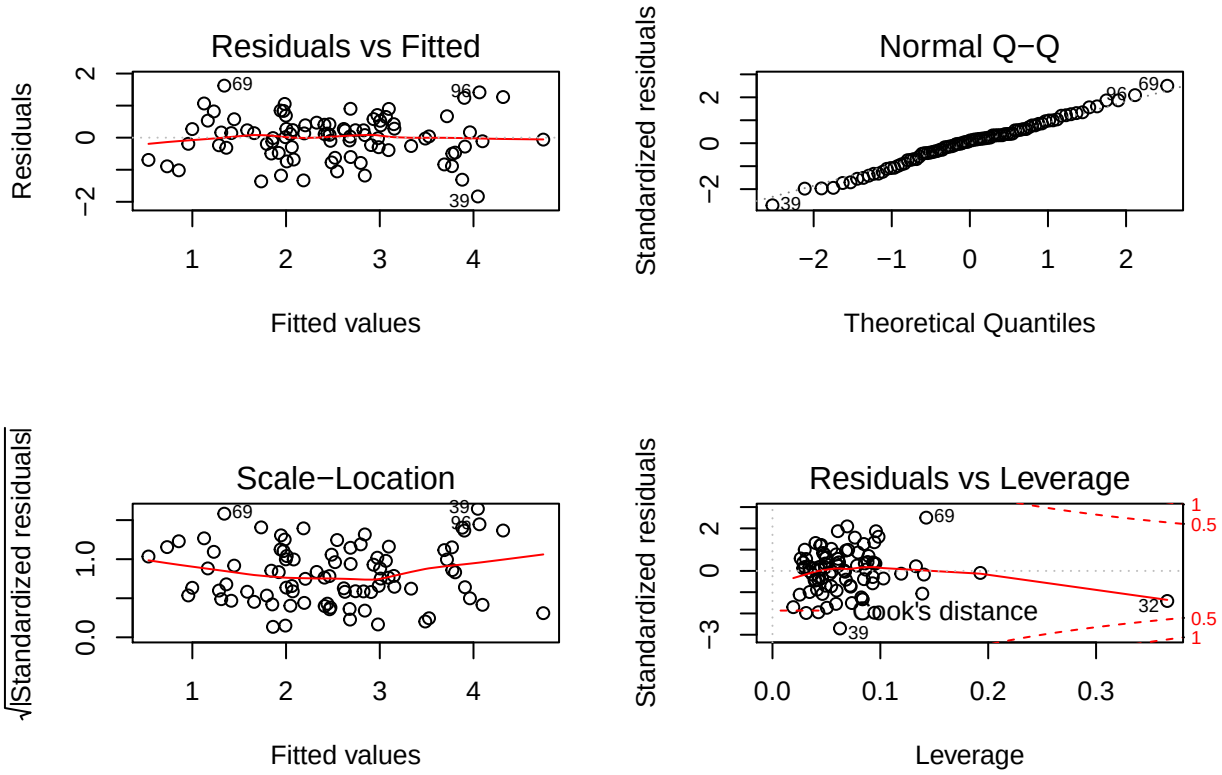
The correlation is high, but Ra^2 is low.

AIC

```
## Start:  AIC=-56.3
## Y ~ X1 + X2 + X3 + X4 + X5
##
##        Df Sum of Sq    RSS     AIC
## <none>              39.682 -56.296
## - X3    1    1.7855 41.467 -54.467
## - X4    1    1.9413 41.623 -54.141
## - X2    1    2.4821 42.164 -53.018
## - X5    1    7.3300 47.012 -43.549
## - X1    1   22.5569 62.239 -19.139
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = trndata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83288 -0.42752  0.08003  0.41092  1.62176
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.45460    0.87636   1.660 0.100817
## X1           0.52600    0.07752   6.786 1.74e-09 ***
## X2           0.39094    0.17368   2.251 0.027103 *
## X3          -0.02077    0.01088  -1.909 0.059790 .
## X4           0.12215    0.06136   1.991 0.049894 *
## X5           0.83983    0.21712   3.868 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6999 on 81 degrees of freedom
## Multiple R-squared:  0.6487, Adjusted R-squared:  0.627
## F-statistic: 29.92 on 5 and 81 DF,  p-value: < 2.2e-16
```

```
## [1] 192.5991
```



AICC

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   221.5   221.5   221.5   221.5   221.5   221.5
```

BIC

```
## [1] 244.6993
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   244.7   244.7   244.7   244.7   244.7   244.7
```

R^2 & Ra^2

```
##
## Call:
## lm(formula = frml, data = RegD8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.669337   1.296387   0.516  0.60693
## X1           0.587022   0.087920   6.677 2.11e-09 ***
## X2           0.454467   0.170012   2.673  0.00896 **
## X3          -0.019637   0.011173  -1.758  0.08229 .
## X4           0.107054   0.058449   1.832  0.07040 .
## X5           0.766157   0.244309   3.136  0.00233 **
## X6          -0.105474   0.091013  -1.159  0.24964
## X7           0.045142   0.157465   0.287  0.77503
## X8           0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
## Start:  AIC=-58.32
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X7    1    0.0412 44.204 -60.231
## - X8    1    0.5258 44.689 -59.174
## - X6    1    0.6740 44.837 -58.853
## <none>              44.163 -58.322
## - X3    1    1.5503 45.713 -56.975
## - X4    1    1.6835 45.847 -56.693
## - X2    1    3.5861 47.749 -52.749
## - X5    1    4.9355 49.099 -50.046
## - X1    1   22.3721 66.535 -20.567
##
## Step:  AIC=-60.23
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X6    1    0.6623 44.867 -60.789
## <none>              44.204 -60.231
## - X8    1    1.1920 45.396 -59.650
## - X3    1    1.5166 45.721 -58.959
## - X4    1    1.7053 45.910 -58.560
## + X7    1    0.0412 44.163 -58.322
## - X2    1    3.5462 47.750 -54.746
## - X5    1    4.8984 49.103 -52.037
## - X1    1   23.5039 67.708 -20.872
##
## Step:  AIC=-60.79
## Y ~ X1 + X2 + X3 + X4 + X5 + X8
##
##        Df Sum of Sq    RSS     AIC
## - X8    1    0.6590 45.526 -61.374
## <none>              44.867 -60.789
## + X6    1    0.6623 44.204 -60.231
## - X3    1    1.2649 46.131 -60.092
## - X4    1    1.6465 46.513 -59.293
## + X7    1    0.0296 44.837 -58.853
```

```
## - X2    1    3.5647 48.431 -55.373
## - X5    1    4.2503 49.117 -54.009
## - X1    1   25.4189 70.285 -19.248
##
## Step:  AIC=-61.37
## Y ~ X1 + X2 + X3 + X4 + X5
##
##        Df Sum of Sq    RSS     AIC
## <none>              45.526 -61.374
## - X3    1    0.9592 46.485 -61.352
## + X8    1    0.6590 44.867 -60.789
## + X7    1    0.4560 45.070 -60.351
## + X6    1    0.1293 45.396 -59.650
## - X4    1    1.8568 47.382 -59.497
## - X2    1    3.2251 48.751 -56.735
## - X5    1    5.9517 51.477 -51.456
## - X1    1   28.7665 74.292 -15.871
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = RegD8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## X1           0.56561    0.07459   7.583 2.77e-11 ***
## X2           0.42369    0.16687   2.539 0.012814 *
## X3          -0.01489    0.01075  -1.385 0.169528
## X4           0.11184    0.05805   1.927 0.057160 .
## X5           0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

R^2 = 0.65 & Ra^2 = 0.62

Mallows Cp

```
##
## Call:
## lm(formula = frml, data = RegD8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## X1           0.587022   0.087920   6.677 2.11e-09 ***
## X2           0.454467   0.170012   2.673  0.00896 **
## X3          -0.019637   0.011173  -1.758  0.08229 .
## X4           0.107054   0.058449   1.832  0.07040 .
## X5           0.766157   0.244309   3.136  0.00233 **
## X6          -0.105474   0.091013  -1.159  0.24964
## X7           0.045142   0.157465   0.287  0.77503
## X8           0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

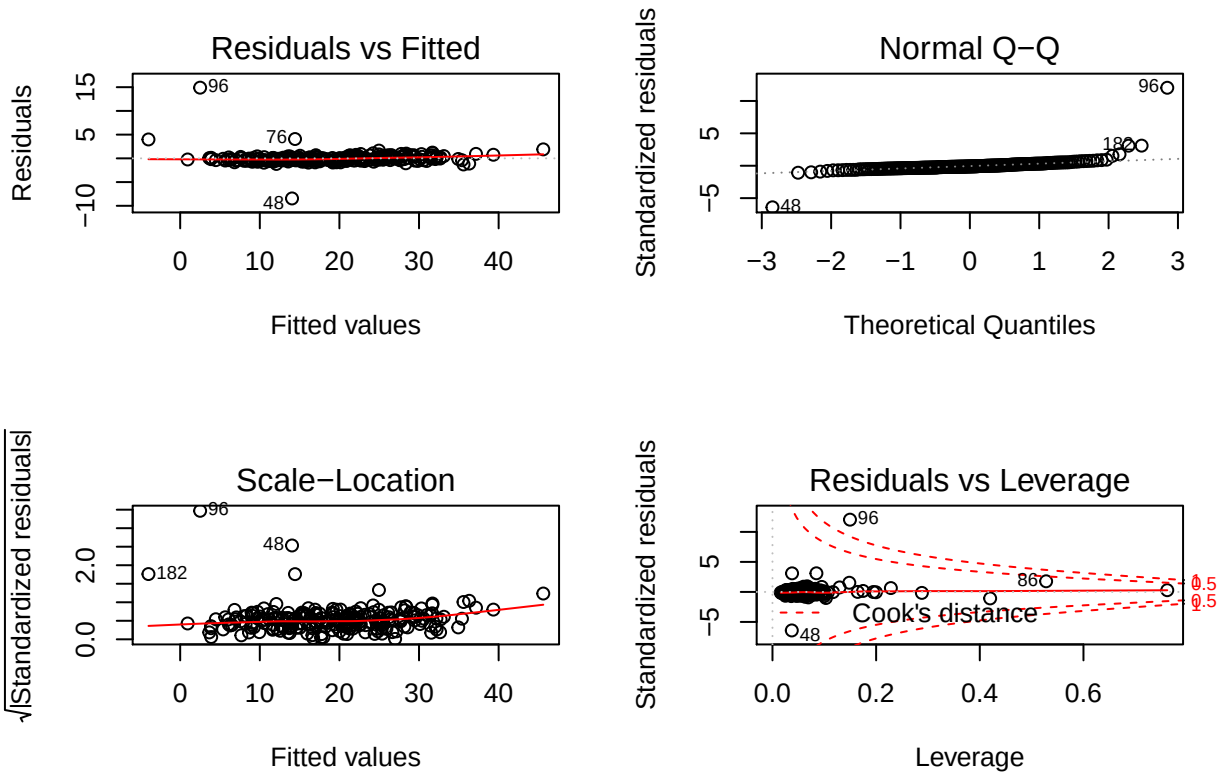
```
##          [,1]
## [1,] 1167.839
```

In conclusion, the R^2 & Ra^2 is the best model determiner.

Question 6

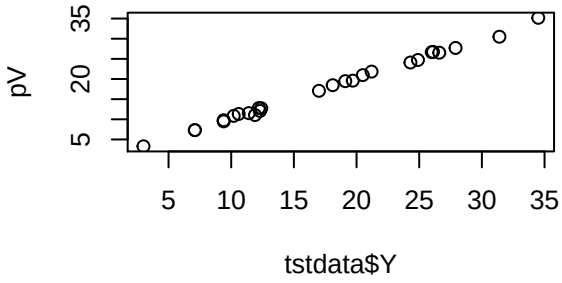
Linear Regression with all predictors

```
##
## Call:
## lm(formula = frml, data = trndata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4495 -0.3918 -0.1386  0.2502 14.8961
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.481e+02  1.226e+01  36.555   <2e-16 ***
## X1          -4.089e+02  9.332e+00 -43.818   <2e-16 ***
## X2           1.304e-02  1.048e-02   1.244    0.215
## X3           1.251e-02  1.840e-02   0.680    0.497
## X4          -9.698e-03  3.061e-02  -0.317    0.752
## X5          -4.506e-02  7.587e-02  -0.594    0.553
## X6           3.106e-02  3.211e-02   0.967    0.335
## X7           2.048e-02  3.510e-02   0.583    0.560
## X8           1.558e-02  4.927e-02   0.316    0.752
## X9          -1.390e-02  4.694e-02  -0.296    0.767
## X10         -1.076e-02  8.181e-02  -0.132    0.895
## X11         -1.151e-01  8.643e-02  -1.332    0.184
## X12         -6.810e-02  5.646e-02  -1.206    0.229
## X13          4.644e-02  6.431e-02   0.722    0.471
## X14          2.638e-02  1.860e-01   0.142    0.887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.342 on 211 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9743
## F-statistic: 609.1 on 14 and 211 DF,  p-value: < 2.2e-16
```




```
##        X1        X2        X3        X4        X5        X6        X7
##  3.944818  2.153677 36.689002  1.660995  4.330486  9.230707 18.063627
##        X8        X9       X10       X11       X12       X13       X14
## 15.353380  7.642016  4.989283  2.268680  3.713877  2.145120  3.535124
```

```
## [1] 0.9986404
```



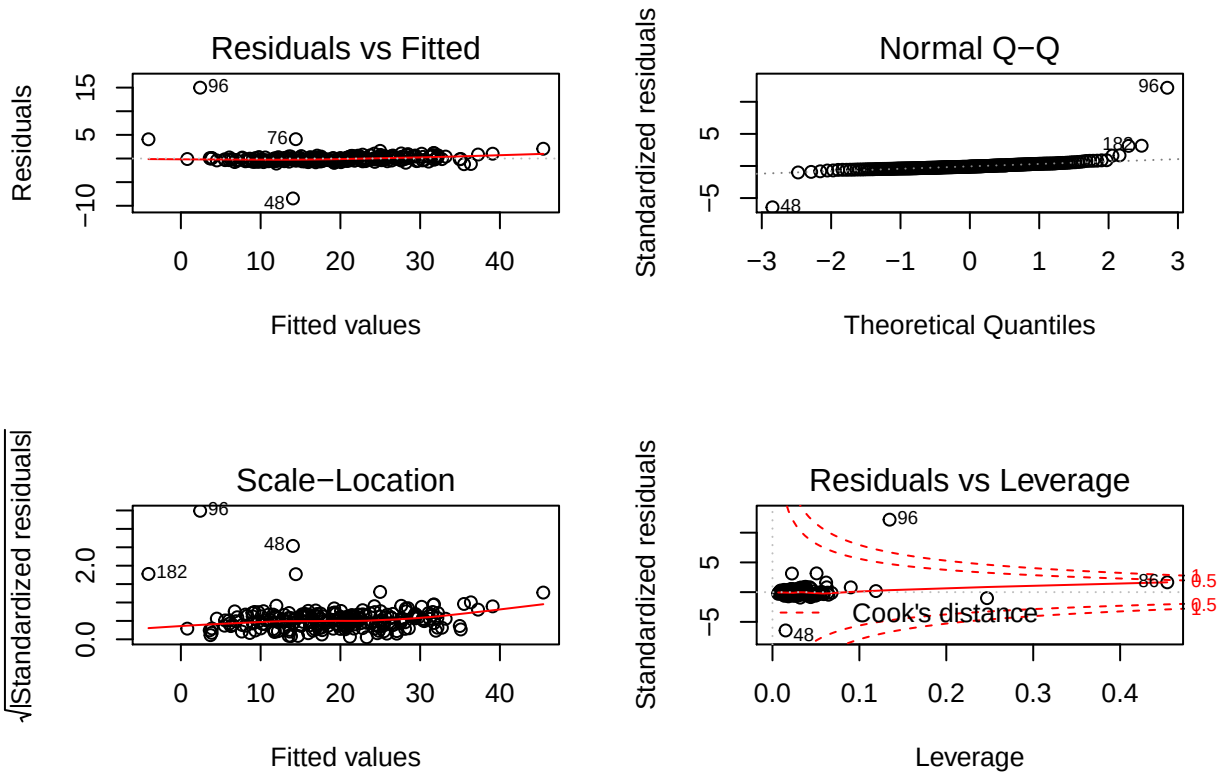
Linear Regression done using AIC

```
## Start:  AIC=147.45
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14
##
##        Df Sum of Sq    RSS    AIC
## - X10   1       0.0  380.0 145.46
## - X14   1       0.0  380.0 145.47
## - X9    1       0.2  380.2 145.54
## - X8    1       0.2  380.2 145.55
## - X4    1       0.2  380.2 145.55
## - X7    1       0.6  380.6 145.81
## - X5    1       0.6  380.6 145.82
## - X3    1       0.8  380.8 145.94
## - X13   1       0.9  381.0 146.00
## - X6    1       1.7  381.7 146.45
## - X12   1       2.6  382.6 147.00
## - X2    1       2.8  382.8 147.10
## - X11   1       3.2  383.2 147.34
## <none>              380.0 147.45
## - X1    1    3458.0 3838.0 668.07
##
## Step:  AIC=145.46
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X11 + X12 +
```

```
##     X13 + X14
##
##        Df Sum of Sq    RSS    AIC
## - X14   1       0.0  380.1 143.48
## - X8    1       0.2  380.2 143.56
## - X4    1       0.2  380.2 143.58
## - X9    1       0.2  380.3 143.59
## - X5    1       0.6  380.7 143.83
## - X7    1       0.6  380.7 143.84
## - X3    1       0.8  380.8 143.94
## - X13   1       0.9  381.0 144.01
## - X6    1       1.7  381.8 144.48
## - X12   1       2.6  382.6 145.00
## - X2    1       2.8  382.8 145.12
## <none>                380.0 145.46
## - X11   1       3.4  383.5 145.50
## - X1    1    3458.7 3838.7 666.11
##
## Step:  AIC=143.48
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X11 + X12 +
##     X13
##
##        Df Sum of Sq    RSS    AIC
## - X8    1       0.2  380.2 141.58
## - X4    1       0.2  380.3 141.59
## - X9    1       0.2  380.3 141.61
## - X5    1       0.6  380.7 141.83
## - X7    1       0.6  380.7 141.86
## - X3    1       0.9  380.9 141.99
## - X13   1       1.0  381.1 142.07
## - X6    1       1.7  381.8 142.49
## - X12   1       2.6  382.6 143.00
## <none>                380.1 143.48
## - X2    1       3.5  383.6 143.54
## - X11   1       3.6  383.6 143.59
## - X1    1    3623.5 4003.6 673.62
##
## Step:  AIC=141.58
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X9 + X11 + X12 + X13
##
##        Df Sum of Sq    RSS    AIC
## - X9    1       0.1  380.4 139.65
## - X4    1       0.3  380.5 139.76
## - X5    1       0.8  381.0 140.04
## - X7    1       0.8  381.0 140.06
## - X13   1       0.9  381.2 140.13
## - X6    1       1.6  381.8 140.51
## - X3    1       2.0  382.2 140.76
## - X12   1       2.6  382.9 141.13
## <none>                380.2 141.58
## - X2    1       3.4  383.7 141.61
## - X11   1       3.7  384.0 141.79
## - X1    1    3651.9 4032.1 673.22
##
```

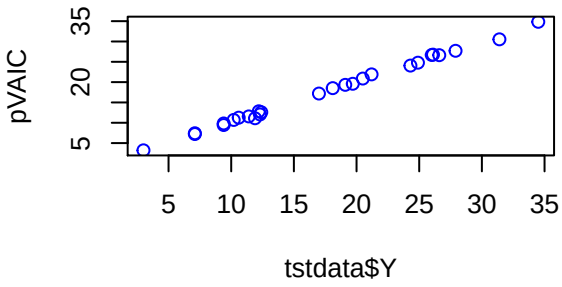
```
## Step:  AIC=139.65
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X11 + X12 + X13
##
##        Df Sum of Sq    RSS    AIC
## - X4    1       0.2  380.6 137.78
## - X7    1       0.8  381.1 138.10
## - X5    1       0.8  381.1 138.11
## - X13   1       0.9  381.3 138.21
## - X3    1       2.0  382.3 138.82
## - X6    1       2.0  382.3 138.83
## - X12   1       3.0  383.3 139.40
## <none>                380.4 139.65
## - X11   1       3.8  384.1 139.89
## - X2    1       4.6  385.0 140.38
## - X1    1    3697.4 4077.8 673.77
##
## Step:  AIC=137.78
## Y ~ X1 + X2 + X3 + X5 + X6 + X7 + X11 + X12 + X13
##
##        Df Sum of Sq    RSS    AIC
## - X5    1       0.8  381.4 136.26
## - X13   1       0.9  381.5 136.32
## - X7    1       0.9  381.5 136.34
## - X3    1       1.7  382.3 136.82
## - X6    1       2.2  382.7 137.06
## - X12   1       2.8  383.4 137.46
## <none>                380.6 137.78
## - X11   1       3.7  384.3 137.98
## - X2    1       4.5  385.1 138.45
## - X1    1    3727.2 4107.8 673.42
##
## Step:  AIC=136.26
## Y ~ X1 + X2 + X3 + X6 + X7 + X11 + X12 + X13
##
##        Df Sum of Sq    RSS    AIC
## - X13   1       0.6  382.0 134.63
## - X7    1       0.9  382.3 134.77
## - X3    1       1.2  382.6 134.98
## - X6    1       2.0  383.4 135.42
## - X12   1       3.3  384.7 136.20
## <none>                381.4 136.26
## - X11   1       3.4  384.8 136.26
## - X2    1       3.9  385.3 136.56
## - X1    1    3851.9 4233.3 678.22
##
## Step:  AIC=134.63
## Y ~ X1 + X2 + X3 + X6 + X7 + X11 + X12
##
##        Df Sum of Sq    RSS    AIC
## - X7    1       0.7  382.7 133.02
## - X3    1       1.4  383.5 133.48
## - X6    1       2.4  384.4 134.02
## - X12   1       2.7  384.7 134.22
## - X11   1       3.2  385.2 134.53
```

```
## <none>               382.0 134.63
## - X2    1       3.8  385.8 134.85
## - X1    1    3889.0 4271.0 678.23
##
## Step:  AIC=133.02
## Y ~ X1 + X2 + X3 + X6 + X11 + X12
##
##        Df Sum of Sq    RSS    AIC
## <none>               382.7 133.02
## - X12   1       3.4  386.1 133.04
## - X6    1       3.8  386.4 133.23
## - X11   1       4.1  386.8 133.43
## - X3    1       4.9  387.6 133.91
## - X2    1       5.6  388.3 134.30
## - X1    1    7173.0 7555.6 805.15
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X6 + X11 + X12, data = trndata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4487 -0.4107 -0.1295  0.2513 14.9838
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.531e+02  8.106e+00  55.900   <2e-16 ***
## X1          -4.134e+02  6.453e+00 -64.070   <2e-16 ***
## X2           1.410e-02  7.887e-03   1.788   0.0751 .
## X3           1.506e-02  8.969e-03   1.679   0.0945 .
## X6           4.013e-02  2.736e-02   1.467   0.1438
## X11         -1.178e-01  7.687e-02  -1.533   0.1268
## X12         -6.907e-02  4.928e-02  -1.401   0.1625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.322 on 219 degrees of freedom
## Multiple R-squared:  0.9757, Adjusted R-squared:  0.975
## F-statistic:  1465 on 6 and 219 DF,  p-value: < 2.2e-16
```




```
##       X1       X2       X3       X6       X11      X12
## 1.943975 1.256325 8.987419 6.905175 1.849686 2.916400
```

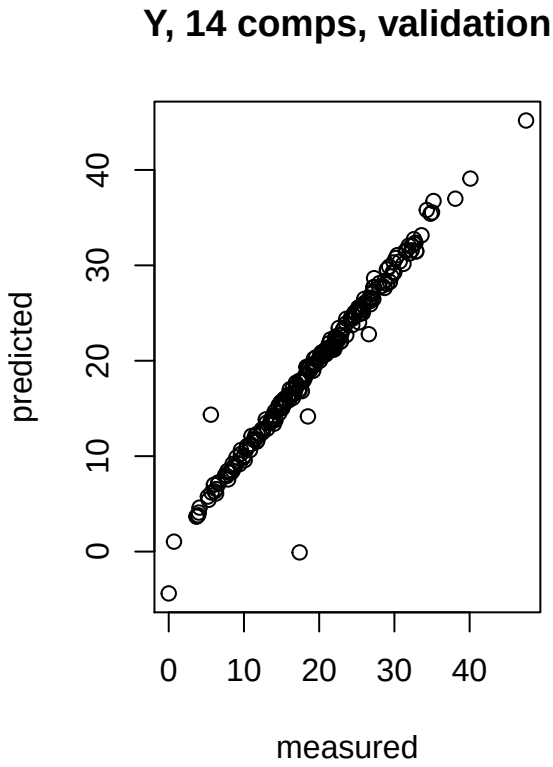
```
## [1] 0.9988102
```

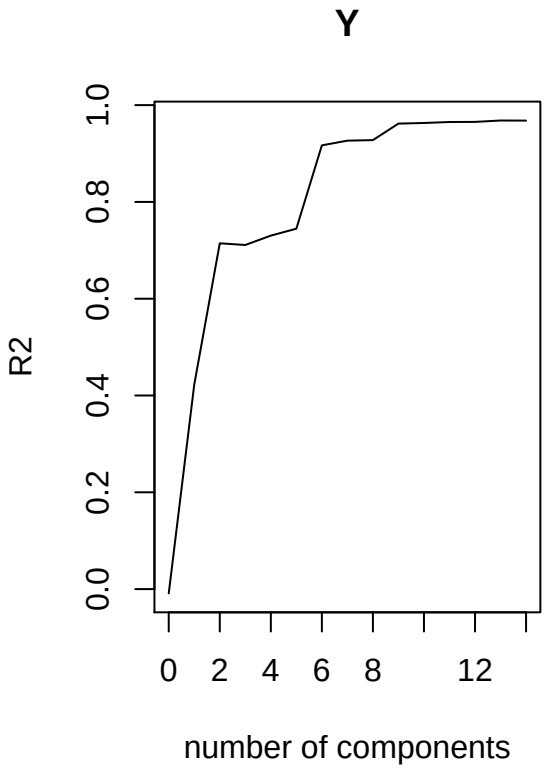


Principle Component Regression

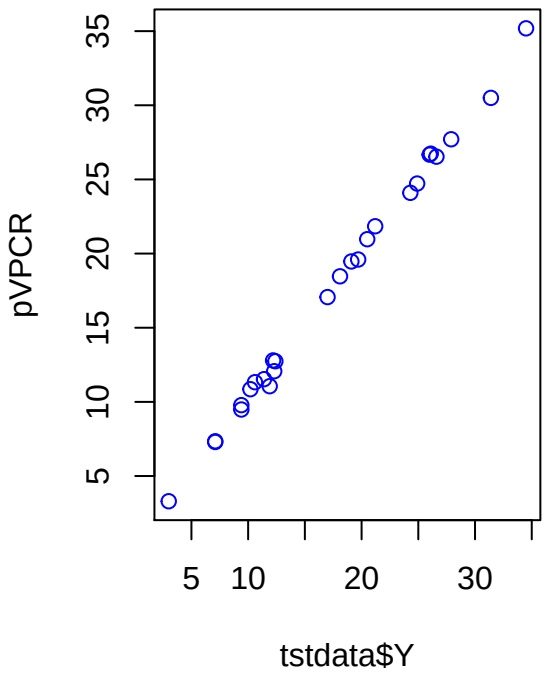
```
## Data:    X dimension: 226 14
##  Y dimension: 226 1
## Fit method: svdpc
## Number of components considered: 14
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           8.382    6.341    4.459    4.486    4.333    4.217    2.405
## adjCV        8.382    6.333    4.446    4.460    4.314    4.190    2.371
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       2.262    2.244    1.630     1.603     1.559     1.554     1.488
## adjCV    2.240    2.236    1.552     1.586     1.548     1.543     1.478
##        14 comps
## CV        1.493
## adjCV     1.482
##
## TRAINING: % variance explained
##    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X    60.29    71.54    79.13    83.83    88.29    91.49    93.57    95.39
## Y    43.78    73.53    77.26    78.52    81.83    92.98    93.56    93.59
##    9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X    96.79     98.09     98.99     99.55     99.85    100.00
```

```
## Y    96.83     97.19     97.35     97.37     97.57     97.59
```





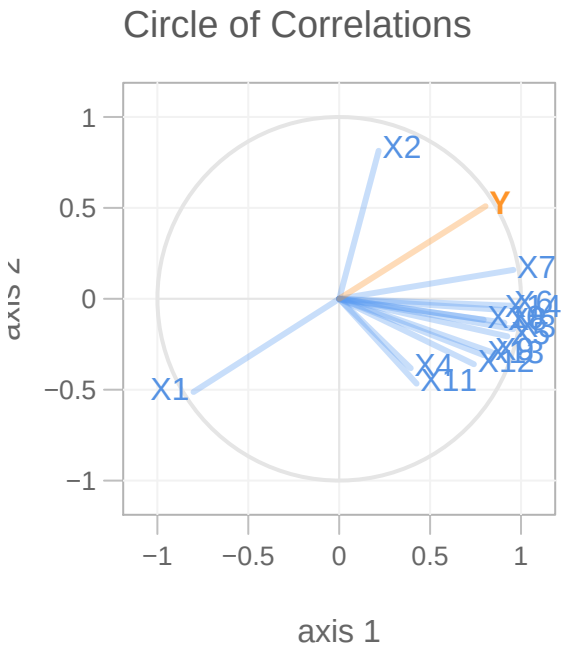

```
## [1] 0.9986404
```

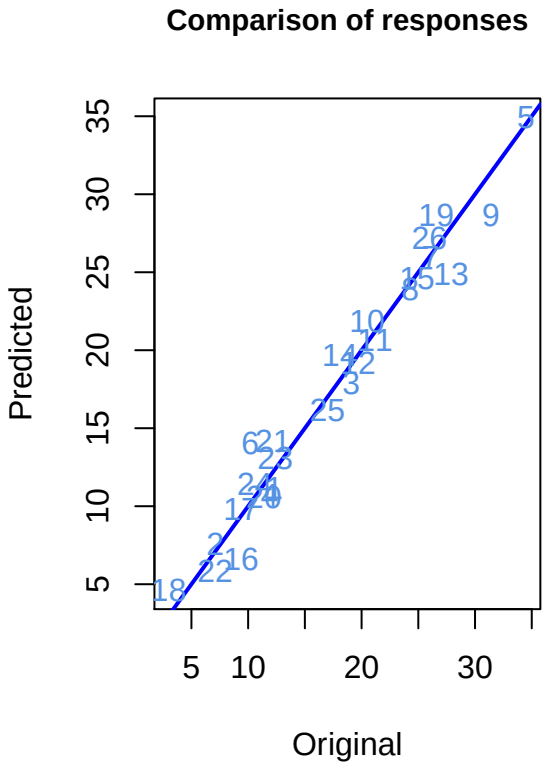


R^2 value close to 1, therefore the model is a good fit.

Partial Least Squares

```
##         t1         t2         t3
## 0.64759751 0.25897025 0.05614655
```

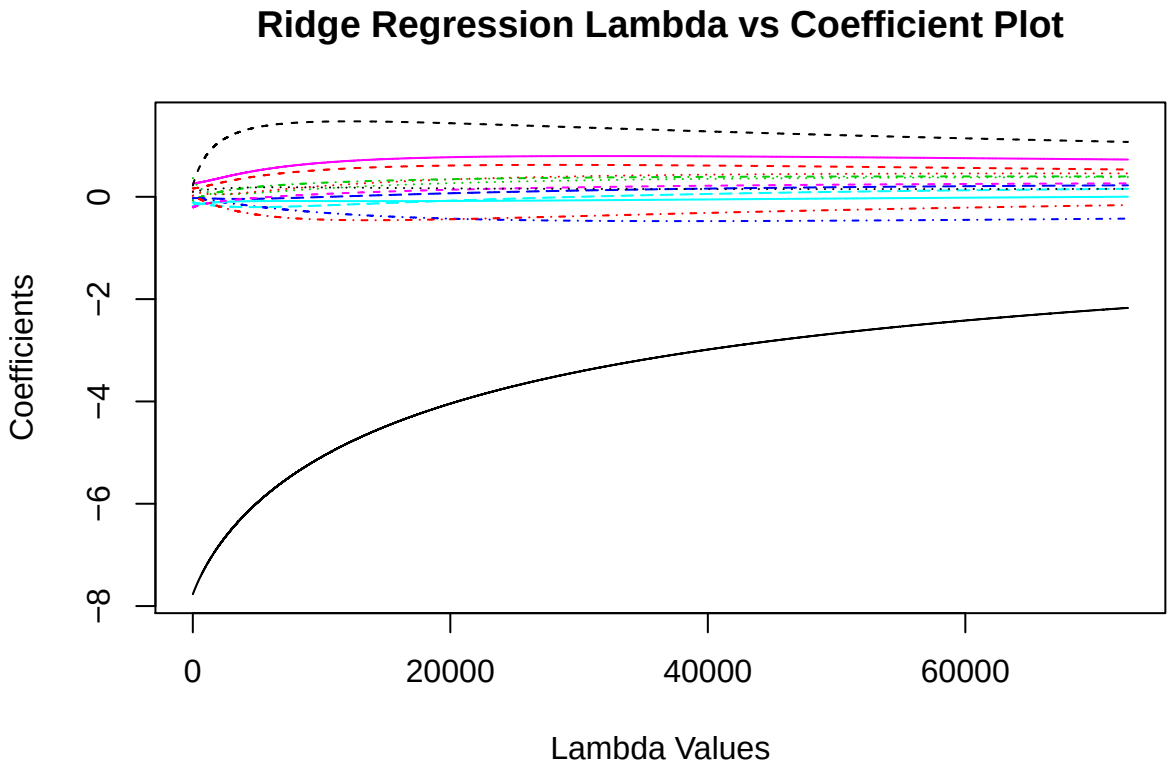

Circle of Correlations
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
axis 1
X2
Y
X7
X1
X4
X11
X12



The R^2 values of the model are lower than the other, therefore not the best model.

Ridge Regression

```
##        Length  Class  Mode
## coef   1016834 -none- numeric
## scales      14 -none- numeric
## Inter        1 -none- numeric
## lambda   72631 -none- numeric
## ym           1 -none- numeric
## xm          14 -none- numeric
## GCV      72631 -none- numeric
## kHKB         1 -none- numeric
## kLW          1 -none- numeric
```



Regression values

```
round(ridgered$coef[, which(ridgered$lambda ==.005)], 2)
```

```
##    X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11   X12
## -7.77  0.16  0.37 -0.04 -0.11  0.26  0.22  0.11 -0.07 -0.03 -0.18 -0.21
##   X13   X14
##  0.09  0.02
```

```
round(ridgered$coef[, which(ridgered$lambda ==0)], 2)
```

```
##    X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11   X12
## -7.77  0.16  0.37 -0.04 -0.11  0.26  0.22  0.11 -0.07 -0.03 -0.18 -0.21
##   X13   X14
##  0.09  0.02
```

Since the values are the same, Ridge regression performs as well as ordinary least square method. Linear regression with variables selected using AIC performs the best with the Ra^2 value = 0.975.