Assignment 6

108048110

2022-12-09

Assignment 6

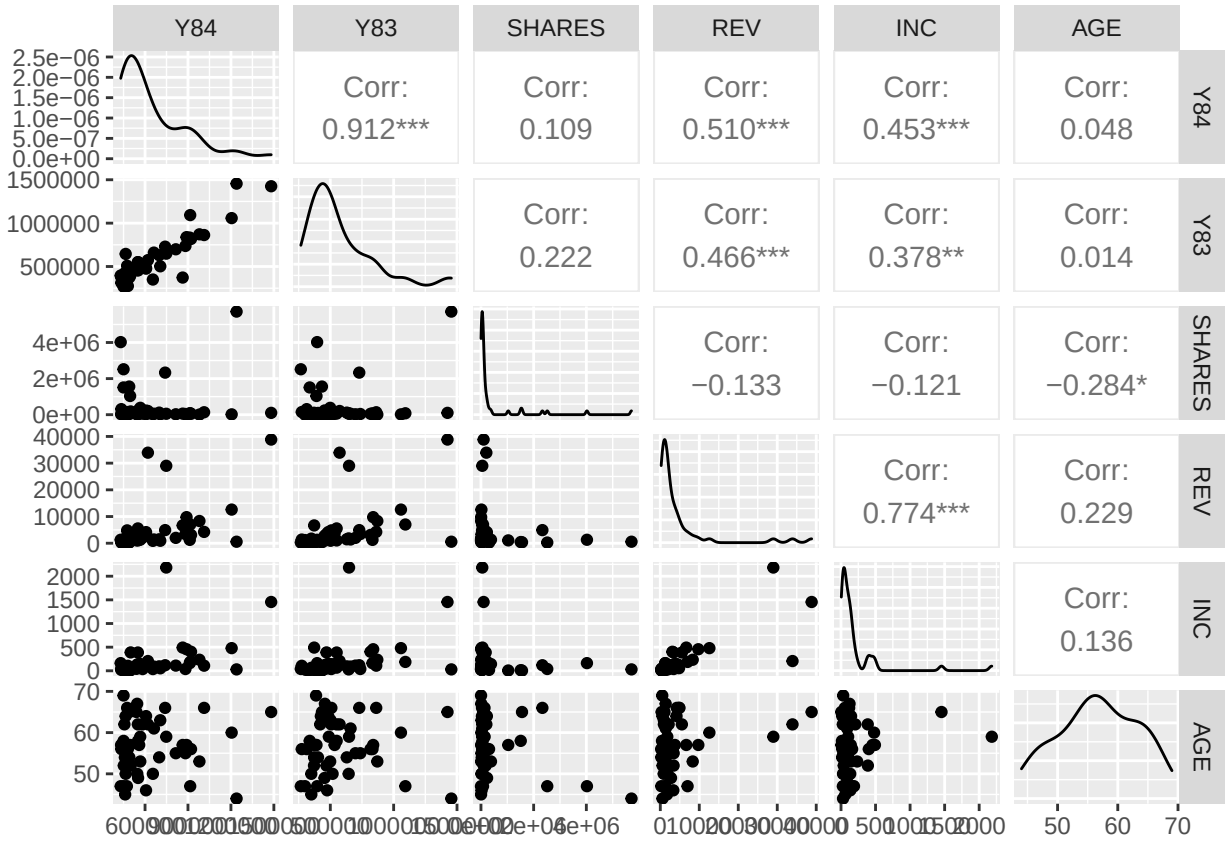
Problem 1

Data Overview

```
summary(data1)
```

```
##       Y84              Y83             SHARES            REV        
##  Min.   : 430000   Min.   : 267510   Min.   :   2000   Min.   :  220  
##  1st Qu.: 473662   1st Qu.: 392276   1st Qu.:  27873   1st Qu.:  922  
##  Median : 555391   Median : 493752   Median :  72506   Median : 1592  
##  Mean   : 650632   Mean   : 571694   Mean   : 445801   Mean   : 4449  
##  3rd Qu.: 746267   3rd Qu.: 656997   3rd Qu.: 171628   3rd Qu.: 4054  
##  Max.   :1481250   Max.   :1455350   Max.   :5713459   Max.   :38828  
##       INC              AGE        
##  Min.   :   4.50   Min.   :44.00  
##  1st Qu.:  36.62   1st Qu.:52.25  
##  Median :  91.75   Median :57.00  
##  Mean   : 192.35   Mean   :56.72  
##  3rd Qu.: 157.57   3rd Qu.:62.00  
##  Max.   :2183.00   Max.   :69.00
```

```
ggpairs(data1)
```



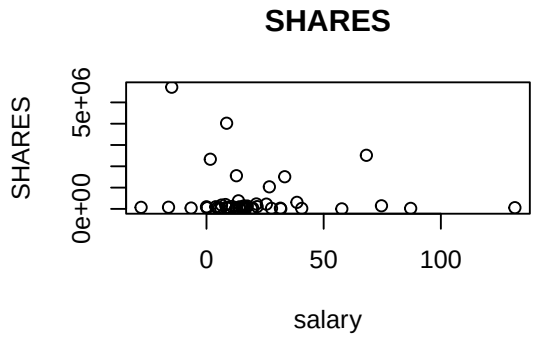
Note. All the variables except for Age seem to be seriously right skewed.

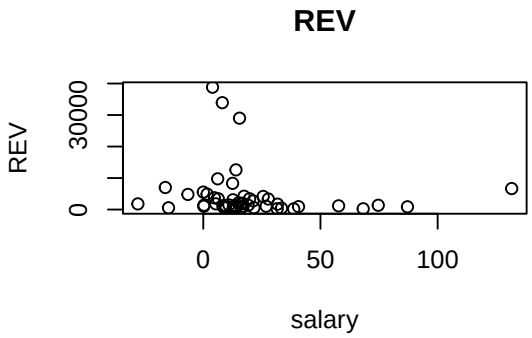
- Transform response variable.

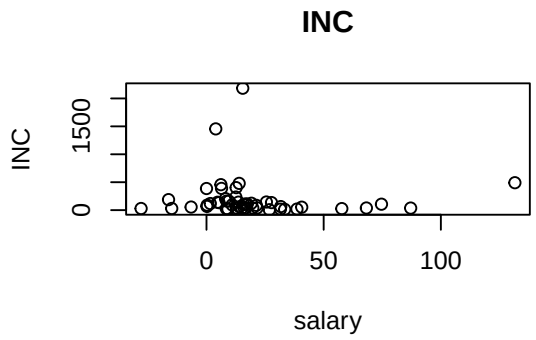
```
salary = 100*(Y84-Y83)/Y83
model1 = lm(salary~SHARES+REV+INC+AGE); summary(model1)
```

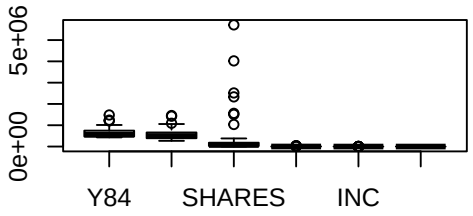
```
##
## Call:
## lm(formula = salary ~ SHARES + REV + INC + AGE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.133 -12.519  -4.066   2.846 109.322
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.509e+01  3.571e+01   1.543    0.130
## SHARES      -3.857e-06  3.717e-06  -1.038    0.305
## REV         -7.237e-04  7.695e-04  -0.940    0.352
## INC          9.744e-03  1.655e-02   0.589    0.559
## AGE         -5.713e-01  6.232e-01  -0.917    0.364
##
## Residual standard error: 26.81 on 45 degrees of freedom
## Multiple R-squared:  0.05754,    Adjusted R-squared:  -0.02623
## F-statistic: 0.6869 on 4 and 45 DF,  p-value: 0.6048
```

- Check for outliers.

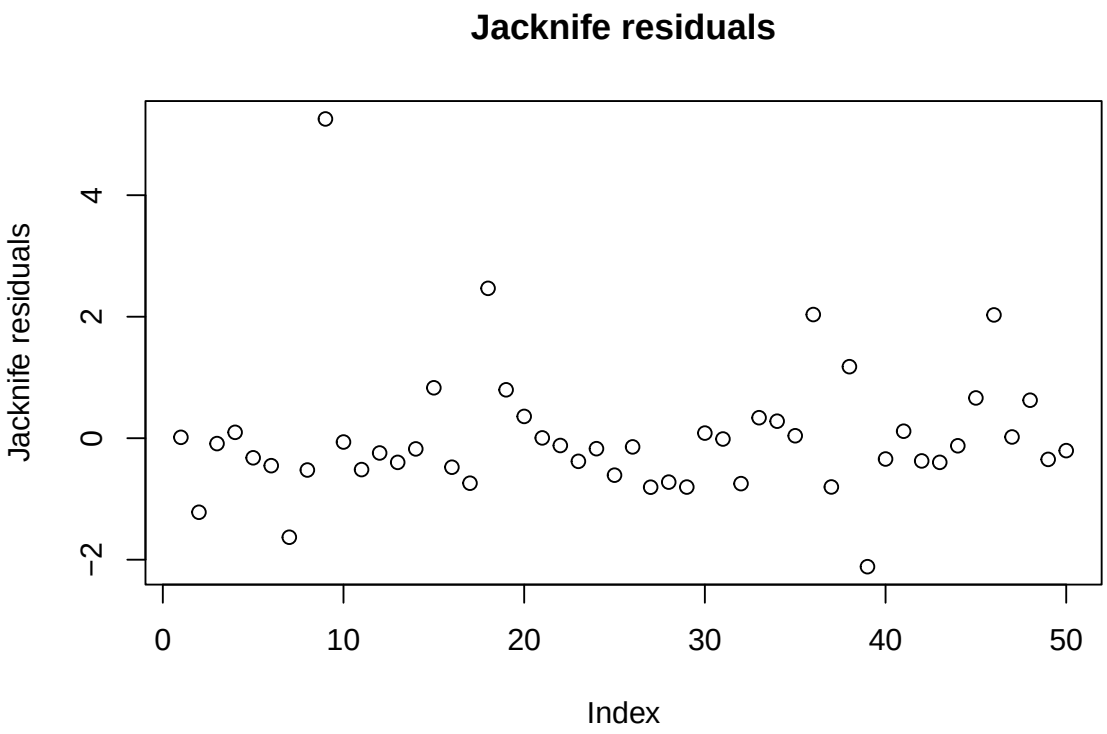








- Exclude i-th observation and recompute the estimates to get $\hat{\beta}_i$ *and* $\hat{\sigma}_i^2$ -> **Jacknife residuals**



- Find out the chair in data that has the largest jacknife residual and compare with critical value.

```
##        9
## 5.255204
```

  - $t_i > t_{n-p-1}(\alpha/2)$

```
## [1] -2.014103
```

  - $t_i > t_{n-p-1}(\alpha/2n)$

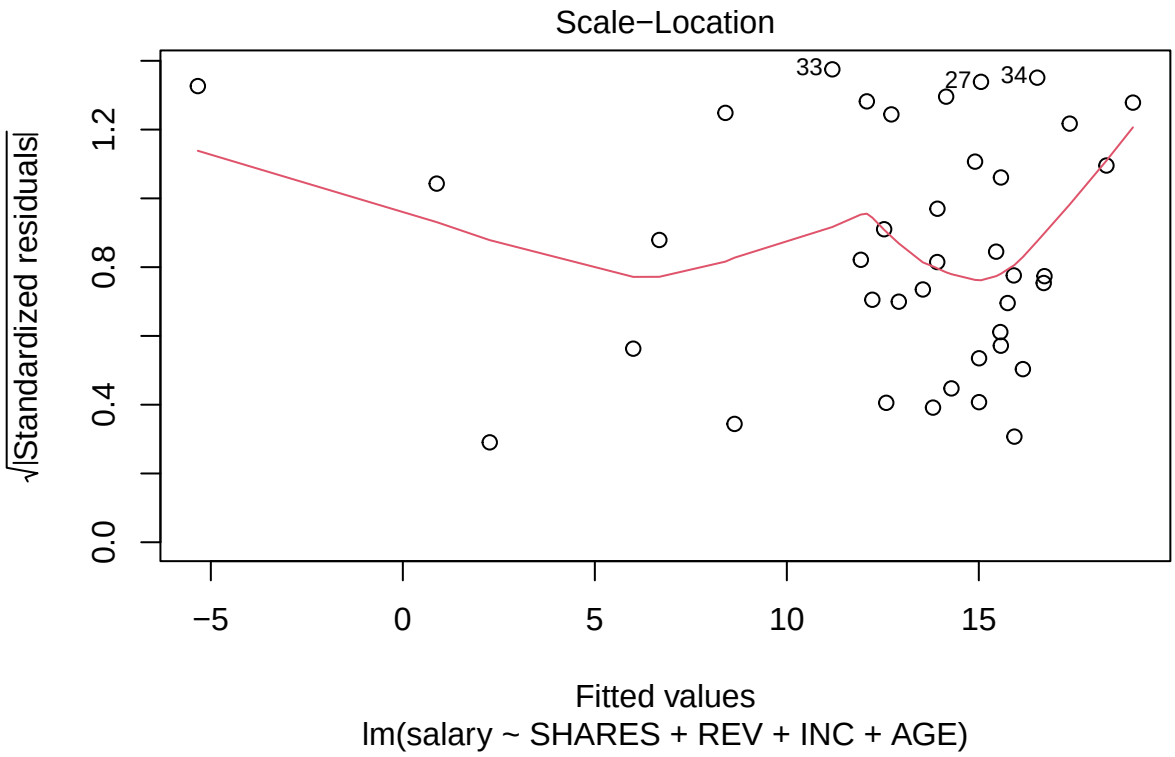
```
## [1] -3.520251
```

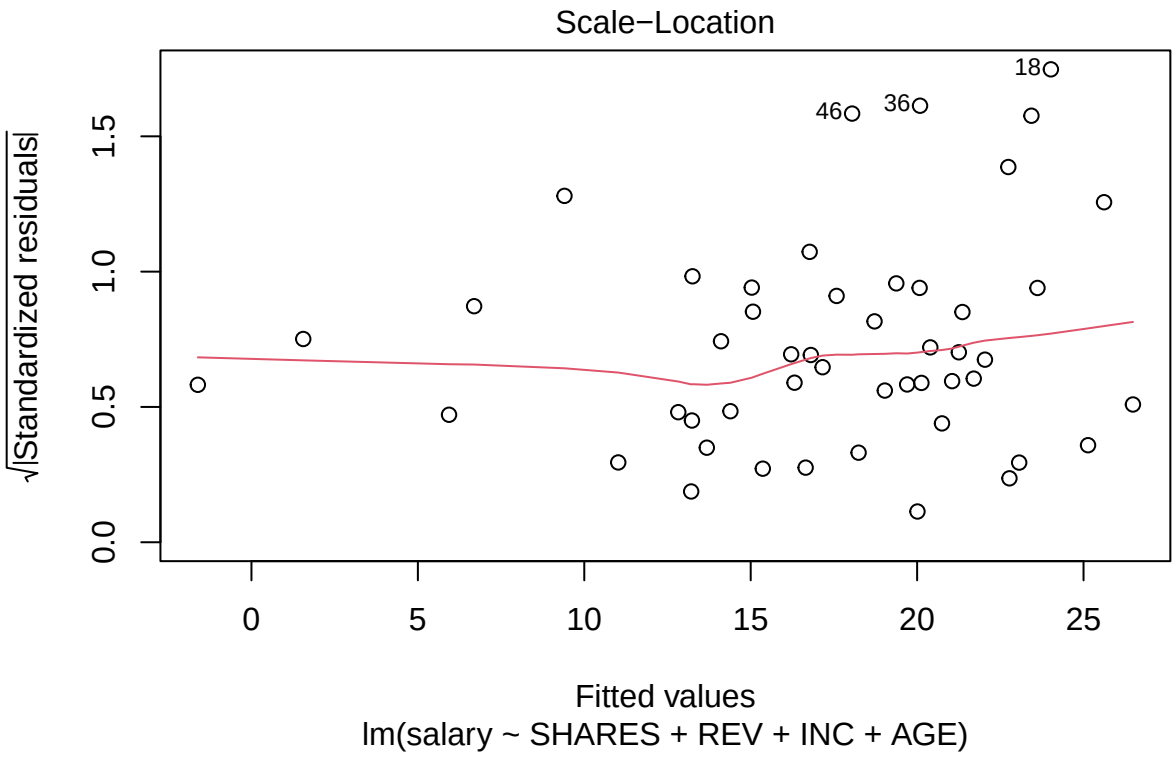
- To conclude, we have enough evidence to reject the null hypothesis, indicating that the 9th chair is an outlier.

```
## [1] "remove: "
##  [1]  9 18 36 46 39 38  7 15 45 48 37

## [1] "remove: "
## [1] 9
```

- Compare t-statistic with Bonferroni critical value we can conclude that there is an outlier in the data.

Residual plots after removing outliers.

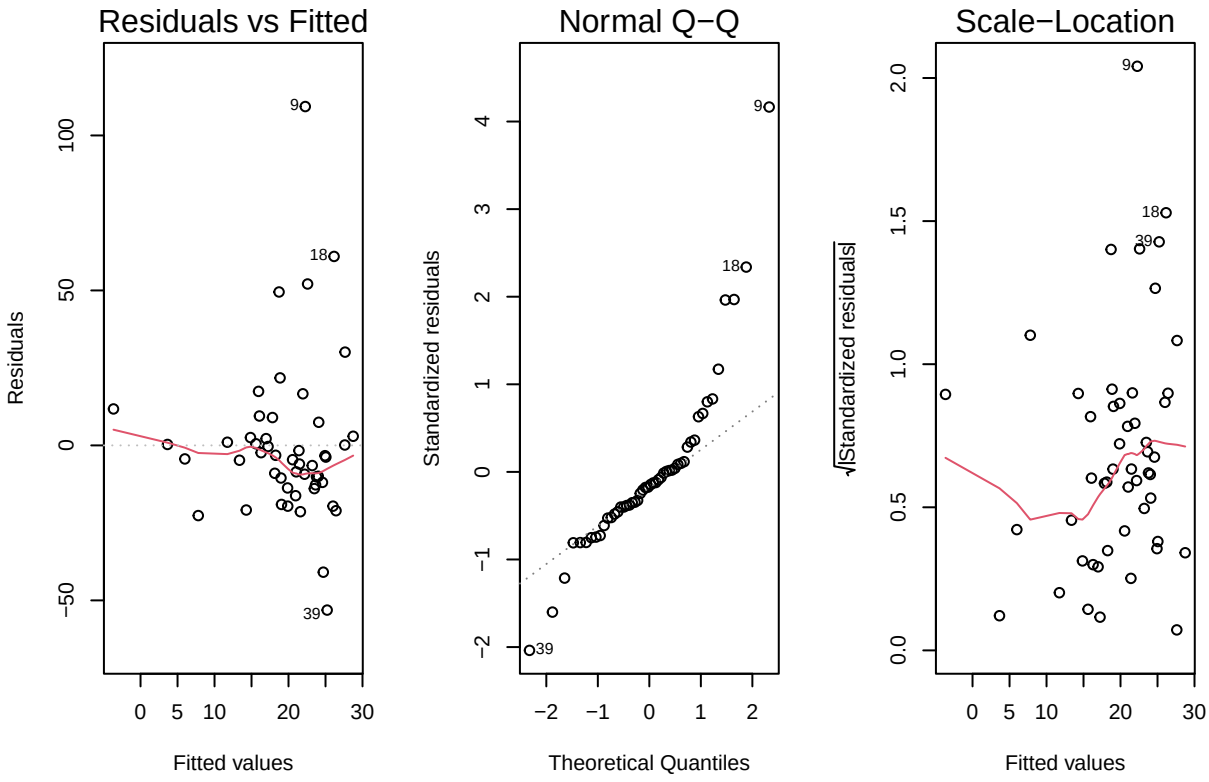

Scale−Location
33
27
34
√|Standardized residuals|
0.0
0.4
0.8
1.2
−5
0
5
10
15
Fitted values
lm(salary ~ SHARES + REV + INC + AGE)



- Obviously, after removing 9th observation from the data, residual appeared to be more like constant.

Do you feel the variances of the raises are equal? If not, what transformation approaches could be applied to improve matters? Make appropriate plots and give your conclusions.

- **Overall pattern**



- Regardless of outliers, the overall pattern do not seems to match the assumption of constant variance.
- Moreover, observed from the third plot, the scale of variance trends upward, indicating non-constant variance.
- And the reason for the non-constant variance is attribute to the large salary percentage. In other words, as the percentage of salary becomes larger, the range salary can vary tends to become larger as well.
- **Transformation**

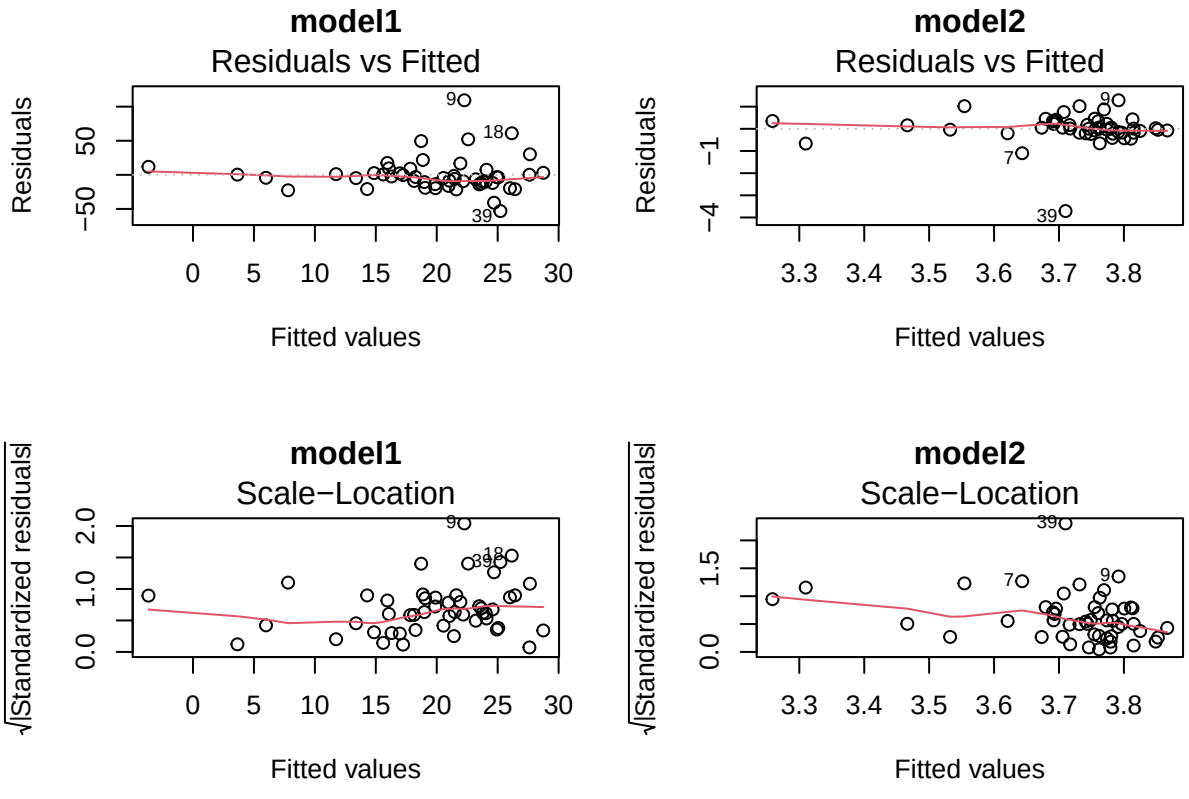
```
salary = 100*(data1$Y84-data1$Y83)/data1$Y83

model1 = lm(salary~SHARES+REV+INC+AGE)
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 4.945166, Df = 1, p = 0.026164
```

- Under the significance level 0.05, we have enough evidence to reject the null hypothesis, error variance changes with the level of the fitted values.
- Since there are negative values in variable *raises*, so we can not directly apply log or sqrt transform on the response. Therefore, I decided to do transformation on *raises+min(raises)+1*, such that response is positive for all observations.

- For the board range of Y, I conduct log transformation on the response and fit model2.

```
##
## Call:
## lm(formula = log(raises) ~ SHARES + REV + INC + AGE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7101 -0.1757  0.0109  0.3015  1.2865
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.439e+00  9.571e-01   3.593 0.000807 ***
## SHARES      -6.844e-08  9.960e-08  -0.687 0.495506
## REV         -1.749e-05  2.062e-05  -0.848 0.400729
## INC          2.679e-04  4.434e-04   0.604 0.548814
## AGE          6.004e-03  1.670e-02   0.359 0.720927
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7184 on 45 degrees of freedom
## Multiple R-squared:  0.02901,    Adjusted R-squared:  -0.0573
## F-statistic: 0.3362 on 4 and 45 DF,  p-value: 0.8522
```



- As we can observe from the plots, model2 seems to have constant variance.

```
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6917002, Df = 1, p = 0.40559
```

- From non-constant variance score test, we conclude under significance level 0.05, we do not have enough evidence to reject the null hypothesis, that is, we do not reject the hypothesis of constant error variance.

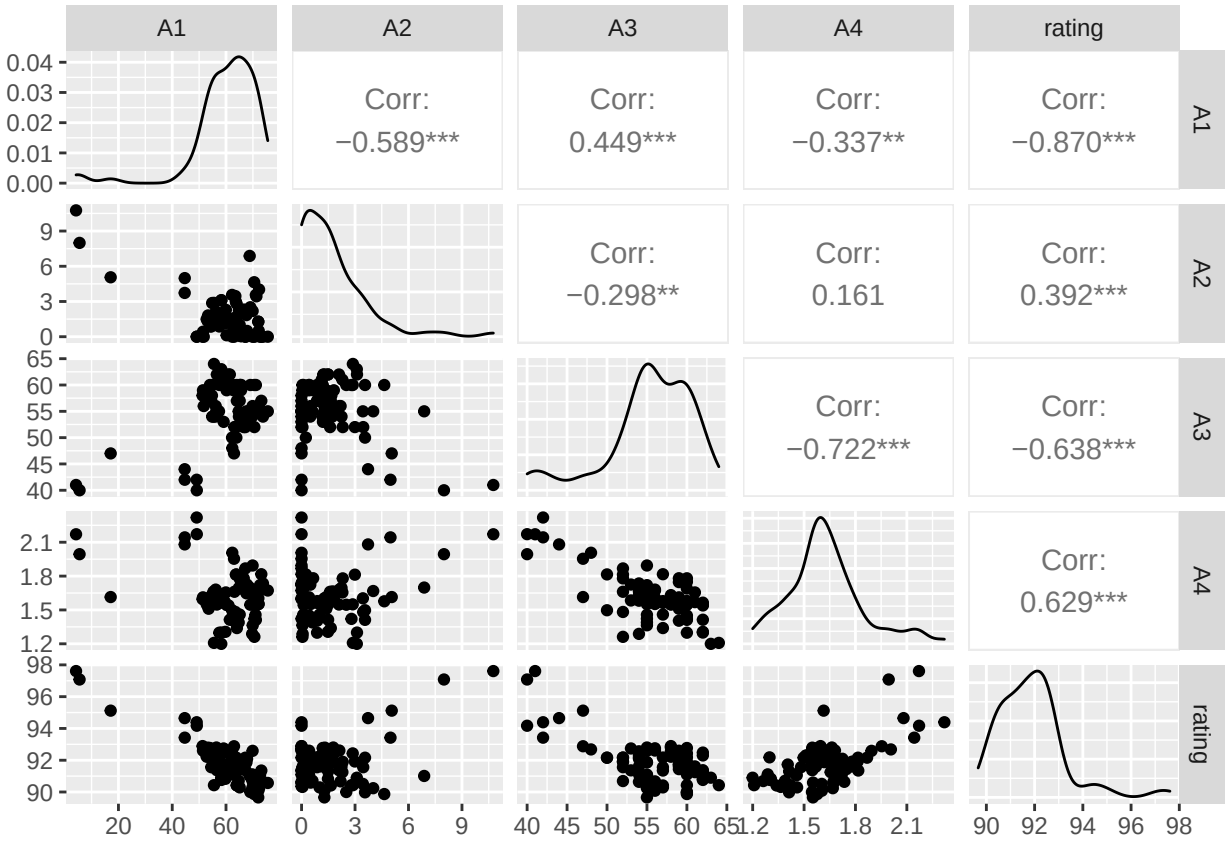
Problem 2

Data Overview

```
summary(data2)
```

```
##        A1              A2               A3              A4
##  Min.   : 4.23   Min.   : 0.000   Min.   :40.00   Min.   :1.200
##  1st Qu.:55.38   1st Qu.: 0.105   1st Qu.:54.00   1st Qu.:1.518
##  Median :62.70   Median : 1.280   Median :56.00   Median :1.604
##  Mean   :60.17   Mean   : 1.664   Mean   :55.46   Mean   :1.627
##  3rd Qu.:67.78   3rd Qu.: 2.277   3rd Qu.:59.75   3rd Qu.:1.723
##  Max.   :75.54   Max.   :10.760   Max.   :64.00   Max.   :2.319
##      rating
##  Min.   :89.66
##  1st Qu.:90.85
##  Median :91.73
##  Mean   :91.85
##  3rd Qu.:92.47
##  Max.   :97.61
```

```
ggpairs(data2)
```



- A2 and A4 have relatively smaller scale, expecting the beta of the predictors to be small as well.
- There exist a lot of blank spaces btw the relation plot in predictors and response, therefore, I suspect the dataset might have non-constant variance.

Fit a model, perform regression diagnostics, present important plots, comment on their meanings and indicate what action should be taken.

- It is acceptable to simply report the outcome of some plots without displaying them.
- Be selective on the plots you present.

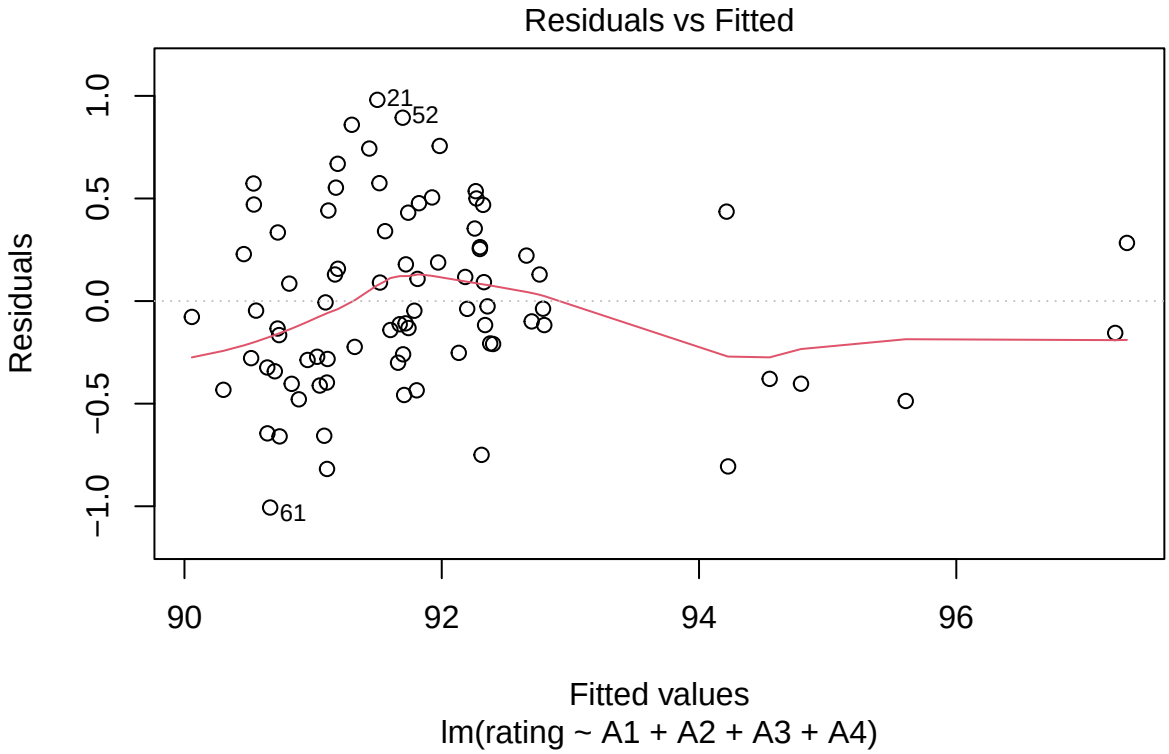
```
model1 = lm(rating~A1+A2+A3+A4)
summary(model1, cor=T)
```

```
##
## Call:
## lm(formula = rating ~ A1 + A2 + A3 + A4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00612 -0.28588 -0.04679  0.32159  0.98069
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 95.853150   1.224877  78.255  < 2e-16 ***
## A1          -0.092821   0.005235 -17.729  < 2e-16 ***
## A2          -0.126798   0.032157  -3.943 0.000176 ***
## A3          -0.025381   0.013971  -1.817 0.073160 .
## A4           1.967603   0.324573   6.062 4.65e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4415 on 77 degrees of freedom
## Multiple R-squared:  0.9056, Adjusted R-squared:  0.9007
## F-statistic: 184.7 on 4 and 77 DF,  p-value: < 2.2e-16
##
## Correlation of Coefficients:
##    (Intercept) A1    A2    A3
## A1 -0.18
## A2 -0.30        0.54
## A3 -0.88       -0.21  0.11
## A4 -0.89        0.08  0.11  0.68
```

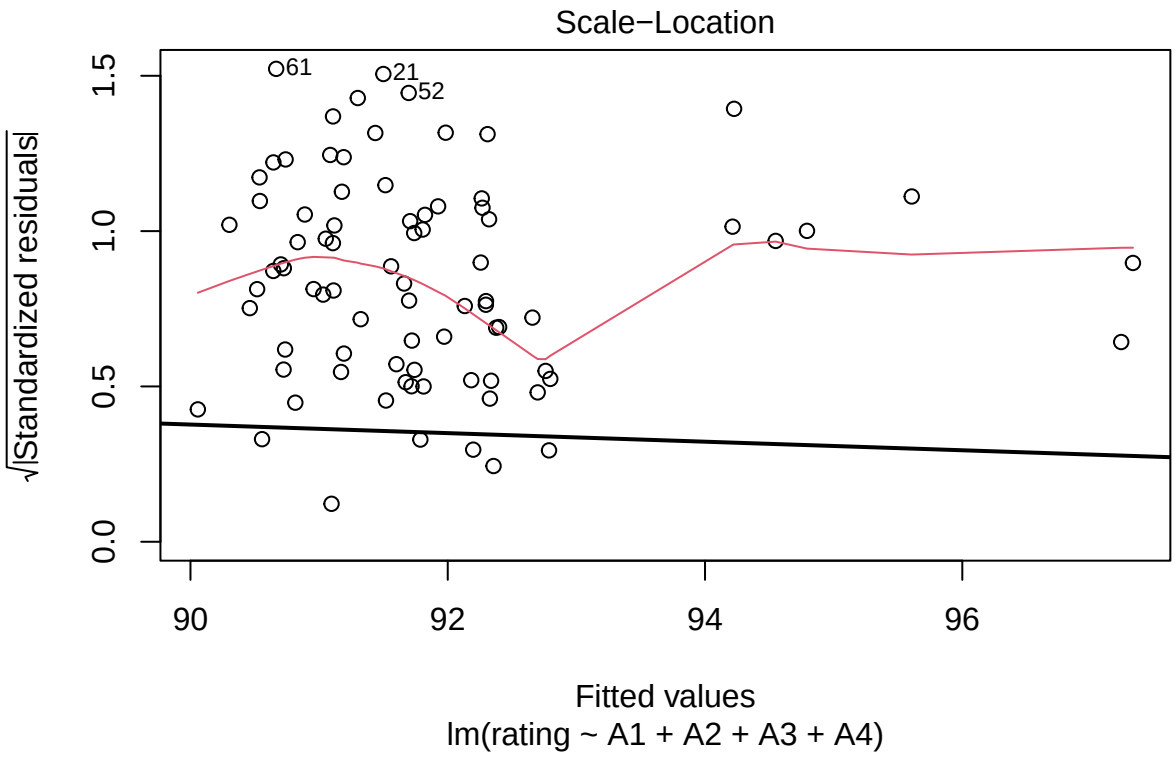
Check on Overall Patterns: variance assumptions.

- **Equal variance**

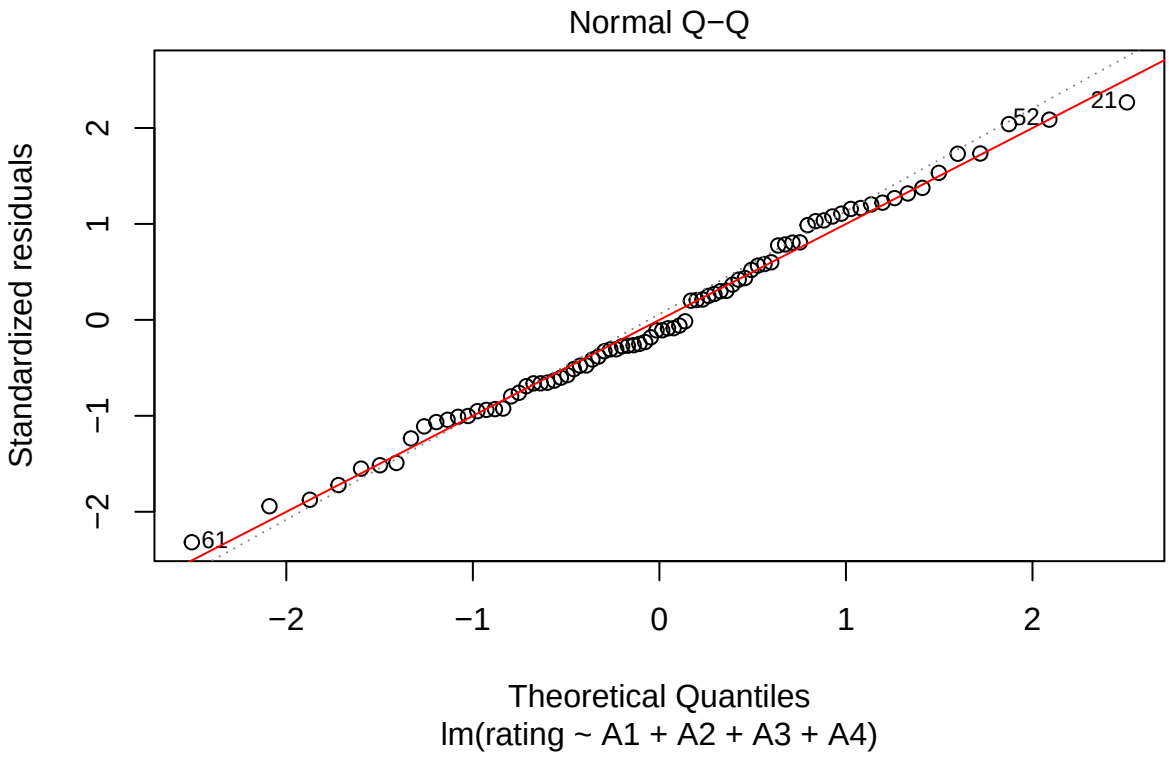


- The model's fitted results have a large proportion assembling in the range 90 to 93, the constant variance assumption does not seem to be seriously violated. Hence, I decided to plot the abs(residual) to obtain a clearer pattern.

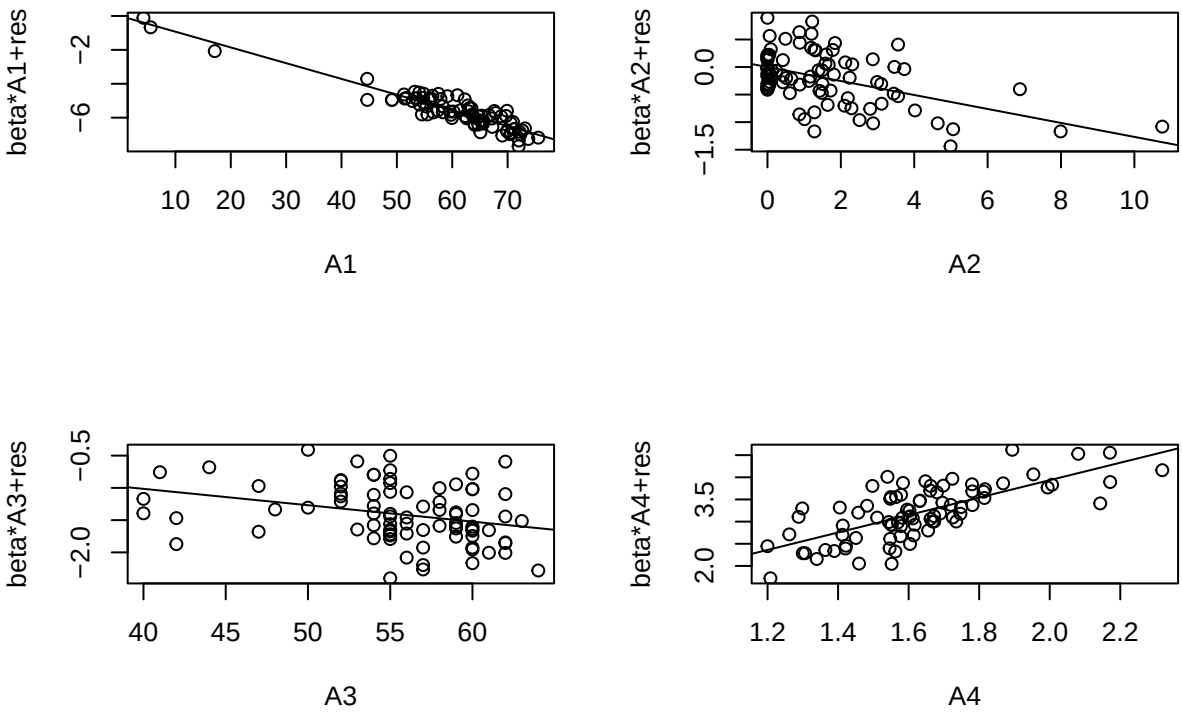
```
## Warning in abline(summary(lm(abs(model1$res) ~ model1$fitted.values)), lwd = 2):
## only using the first two of 8 regression coefficients
```



- Apparently, we can observe some curvature in residual variance plot, which meet our former suspicion. We might need to conduct further transformation on the variables.
- **Normality**



- Null plot, the normality assumption is not violated.
- **Mean curvature examination - Partial Regression Plot**




```
data2[A1<20,]; data2[A2>6,]; data2[(A3<45 & rating>95),]; data2[(A4>2&rating>95),]
```

```
##       A1    A2 A3      A4 rating
## 75  4.23 10.76 41 2.17070  97.61
## 76  5.53  7.99 40 1.99418  97.08
## 77 17.11  5.06 47 1.61437  95.12
```

```
##       A1    A2 A3      A4 rating
## 44 68.81  6.88 55 1.69836  91.01
## 75  4.23 10.76 41 2.17070  97.61
## 76  5.53  7.99 40 1.99418  97.08
```

```
##      A1    A2 A3      A4 rating
## 75 4.23 10.76 41 2.17070  97.61
## 76 5.53  7.99 40 1.99418  97.08
```

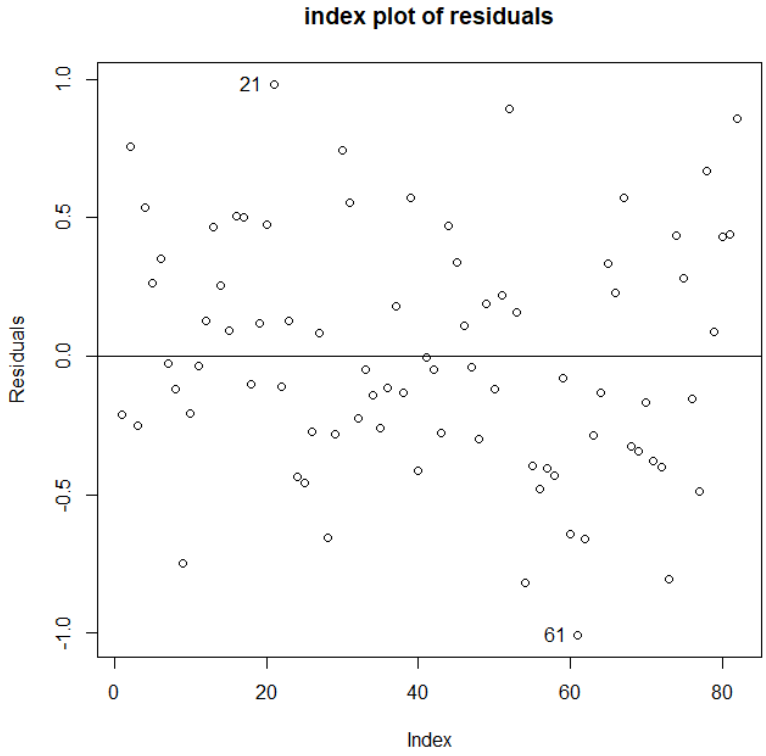
```
##      A1    A2 A3     A4 rating
## 75 4.23 10.76 41 2.1707  97.61
```

- Observe from the plot, some points seem to diverge substantially from the rest of the data.
- Also, response and A3 seems to have some relationship that is not included in model1.
- Furthermore, there exist multiple influential points, which diverge significantly from the rest of the points, would affect the model, I probably need to draw half-normal plots to further check for these extreme values.

Check on Unusual Observations

- Residuals
  - Finding the largest and smallest residual index.

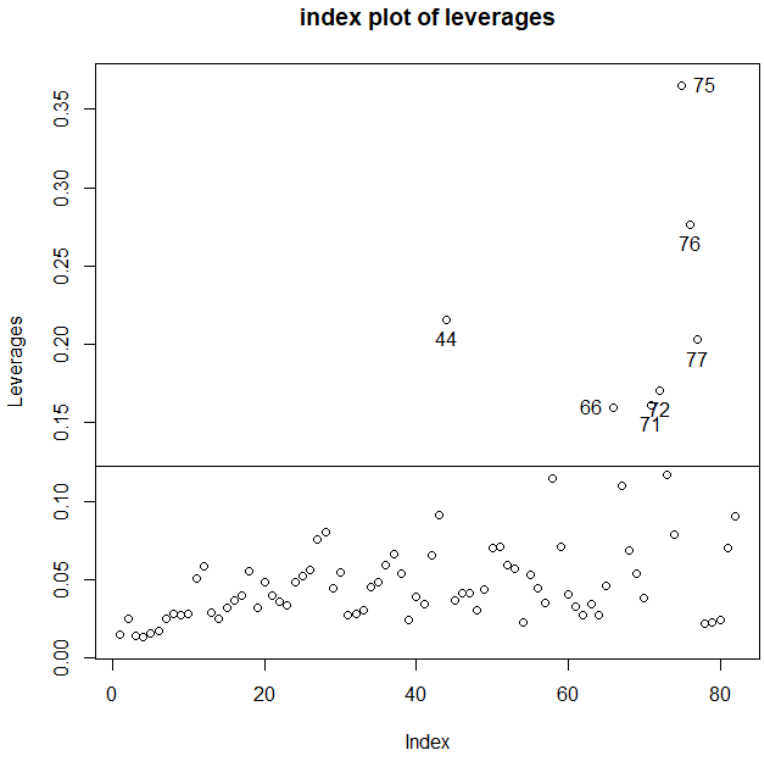
```
##         61         21
## -1.0061236  0.9806932
```



Large residual: 21, 61

- Leverage

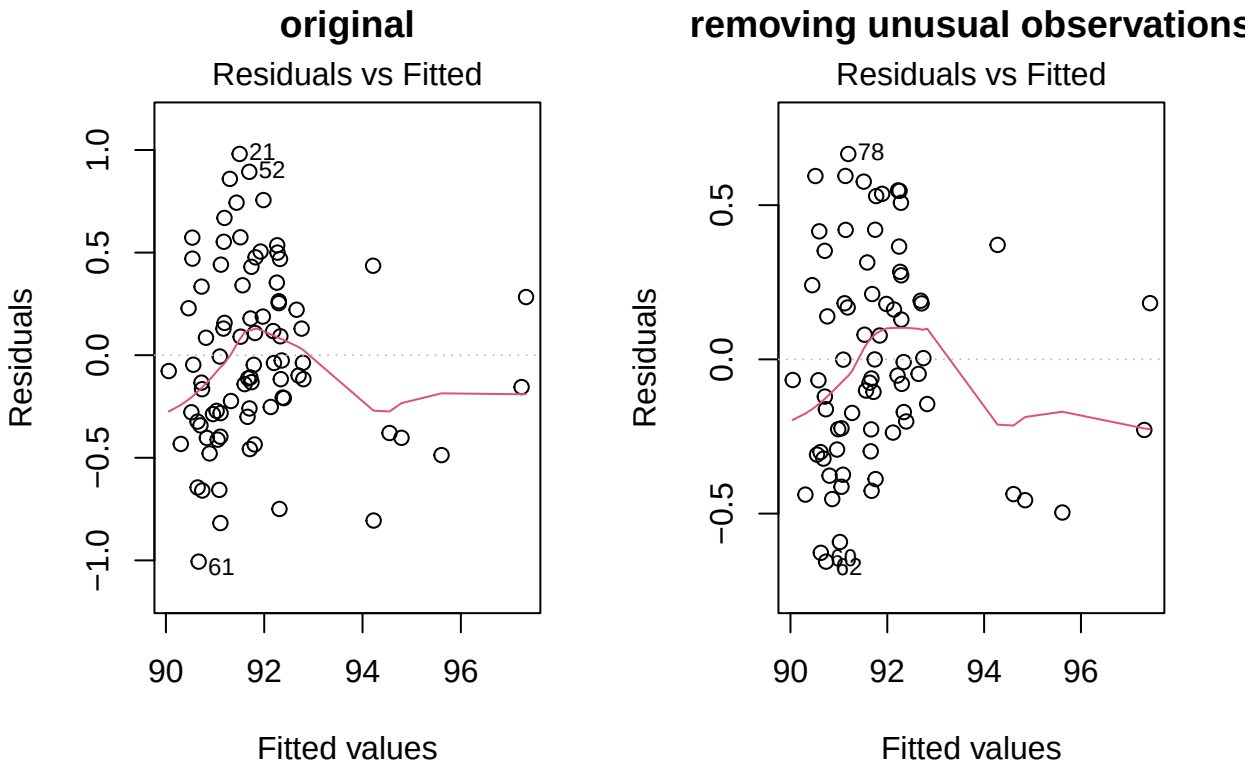
75	76	44	77	72	71	66
0.3648473	0.2762653	0.215549	0.2029067	0.170504	0.1614875	0.1599043



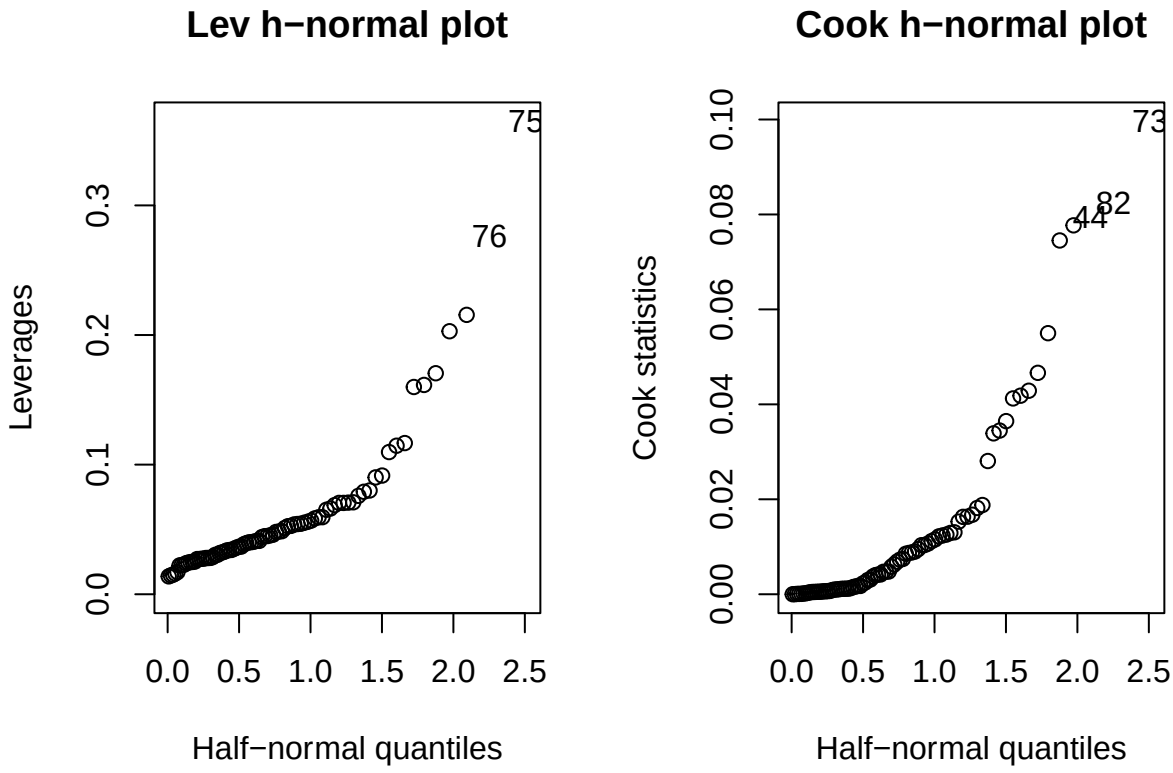
Large leverage: 44, 66, 71, 72, 75, 76, 77

- **Jacknife residuals**

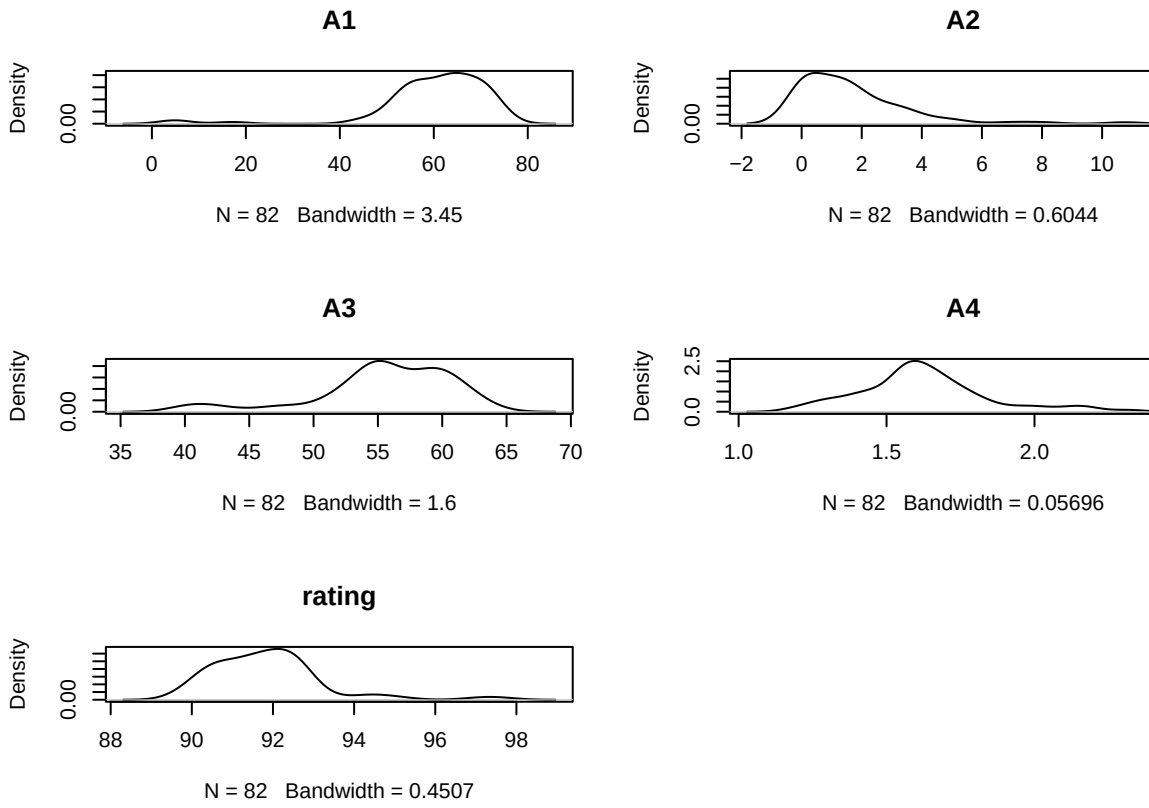
```
## [1] "remove: "
## NULL

## [1] "remove: "
## [1] 61 21 52 73 54 30 82  2  9
```



- As we can observe from the Bonferroni test result and plots, removing unusual observation does not make much difference. Hence, no outliers are indicated when there is no multiple outliers.
- **Half normal plot**
  - Look for extreme values



- Now we know for sure that the residual present a pattern of curvature and that point 75 and 76 can significantly affect the model, likewise, point 42, 73, 82 have relatively large residuals. We should modify the model.



Transforming model

As we can see from the plot, A1 and A3 are left skewed, indicating there are a lot of large values in the variables, so I decided to perform log transformation on both variables to smooth the density curve.

Also, based on the observation from mean curvature examination, I added additional variable to complex the model.

- **Test for lack of fit including point that has *rating* > 94.**

```
model1 = lm(rating~A1+A2+A3+A4)
model2 = lm(rating~log(A1) + A2 + log(A3) + A4 + A1 * A3)
summary(model2)
```

```
##
## Call:
## lm(formula = rating ~ log(A1) + A2 + log(A3) + A4 + A1 * A3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99223 -0.27560 -0.03152  0.28762  1.10119
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 75.683419  28.659297   2.641  0.01008 *
## log(A1)     -1.533107   0.768175  -1.996  0.04964 *
## A2          -0.102587   0.035244  -2.911  0.00476 **
## log(A3)      4.092181  10.276146   0.398  0.69161
```

```
## A4           2.226378   0.324856   6.853 1.84e-09 ***
## A1           0.148222   0.110221   1.345  0.18281
## A3           0.129455   0.265015   0.488  0.62665
## A1:A3       -0.004038   0.001784  -2.264  0.02652 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4231 on 74 degrees of freedom
## Multiple R-squared:  0.9167, Adjusted R-squared:  0.9088
## F-statistic: 116.3 on 7 and 74 DF,  p-value: < 2.2e-16
```

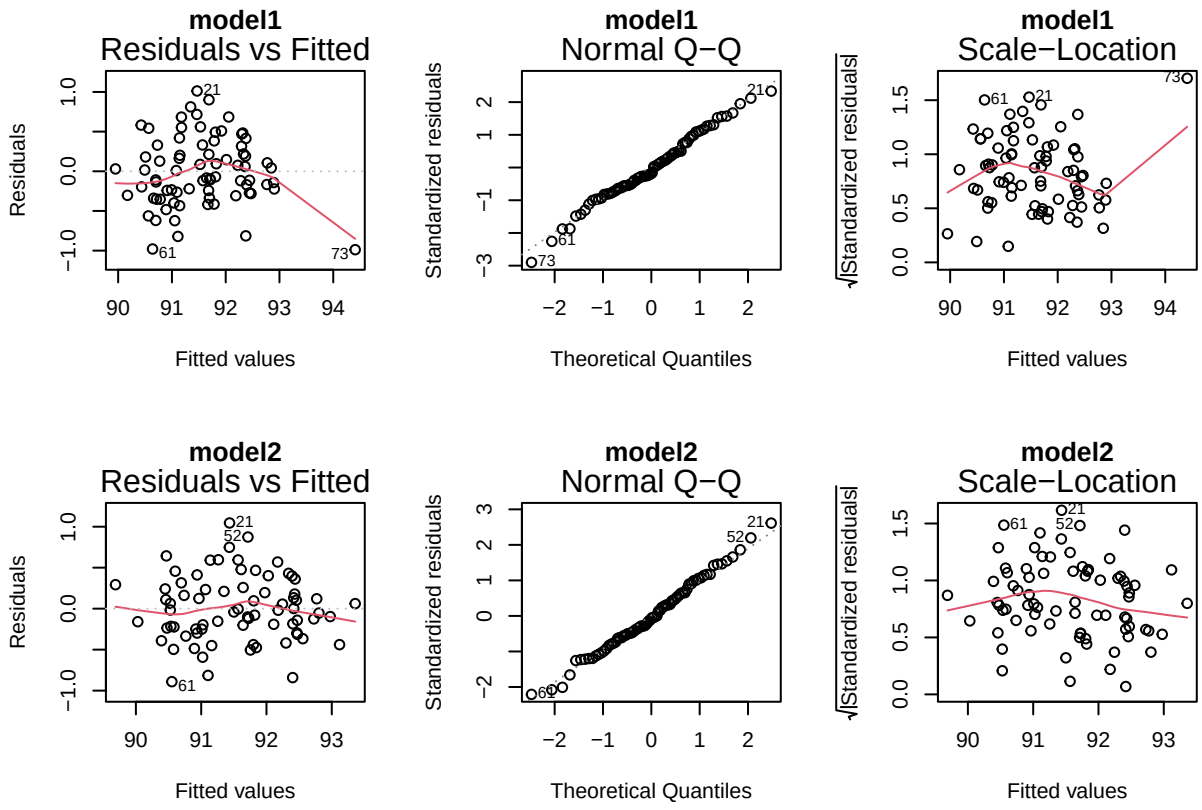
```
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: rating ~ A1 + A2 + A3 + A4
## Model 2: rating ~ log(A1) + A2 + log(A3) + A4 + A1 * A3
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     77 15.006
## 2     74 13.246  3    1.7599 3.2771 0.02565 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- we can observe from the summary table that A1*A3 is a significant additional term.
- However, the anova test's p-value lied in the edge of significance level, it will be subjective to conclude that we should reject the null hypothesis. Hence, to be more specific, I investigated the subset with $rating < 94$.
- **Excluding data where *rating* > 94.**

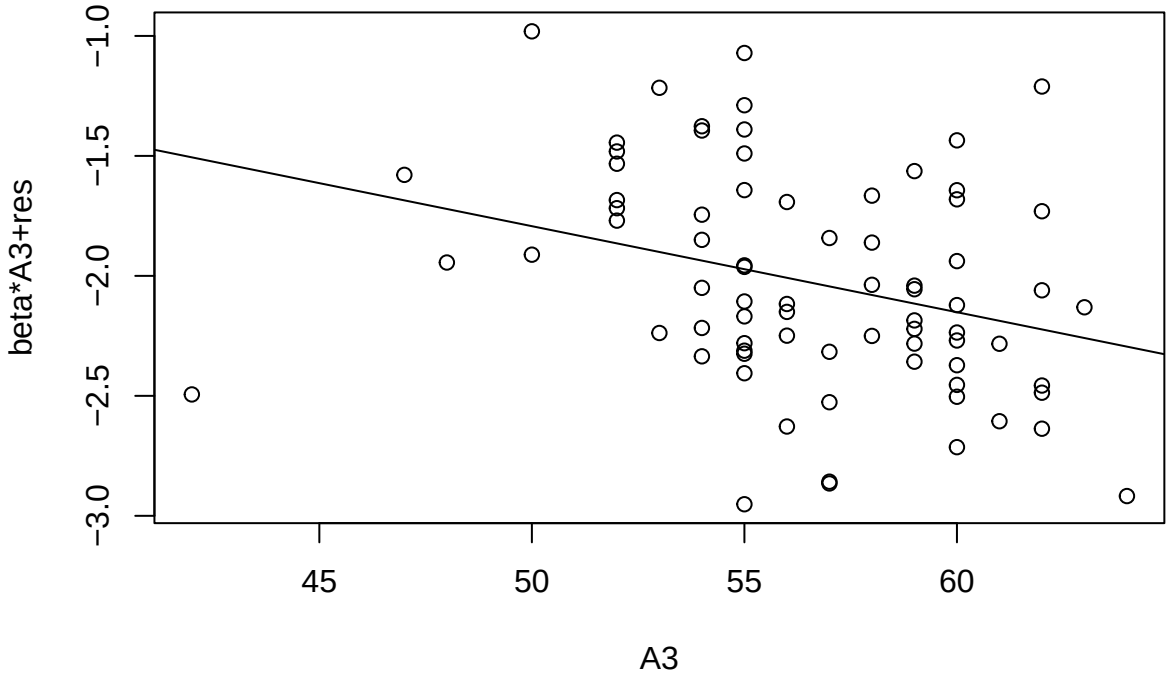
```
## Analysis of Variance Table
##
## Model 1: rating ~ A1 + A2 + A3 + A4
## Model 2: rating ~ log(A1) + A2 + log(A3) + A4 + A1 * A3
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     71 13.935
## 2     68 11.755  3    2.1799 4.2035 0.008677 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

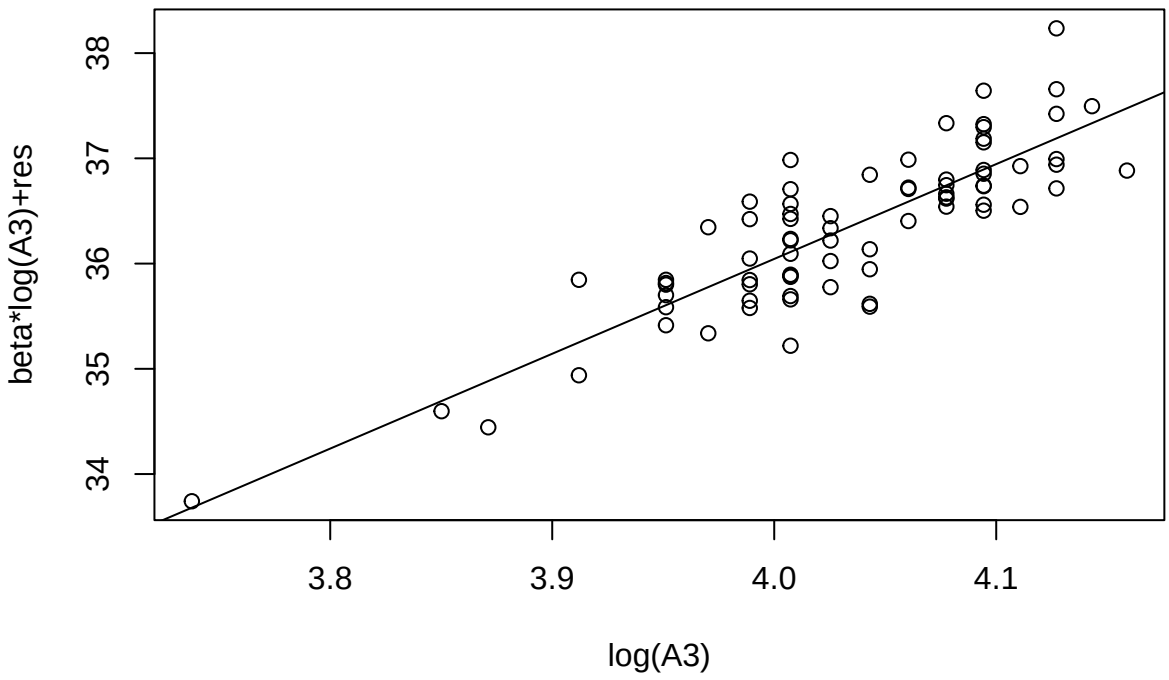
- After excluding points with rating >94 from the data, the model then suggested to reject the null hypothesis, that is, there is lack of fit.
- **Plot**



- Residual plot then suggested constant variance.
- Also the partial regression plot of A3 no longer appear a curvature trend.

```
prplot(model1, 3); prplot(model2, 3)
```


beta*A3+res
−1.0
−1.5
−2.0
−2.5
−3.0
45
50
55
60
A3



Problem 3

Data Overview

```
summary(data3)
```

```
##       ACC              WHP               SP              G         
##  Min.   :0.200   Min.   : 20.50   Min.   : 7.5   Min.   :0.00  
##  1st Qu.:0.700   1st Qu.: 20.50   1st Qu.:22.5   1st Qu.:0.00  
##  Median :1.700   Median : 40.00   Median :30.0   Median :2.00  
##  Mean   :2.606   Mean   : 77.38   Mean   :30.4   Mean   :2.12  
##  3rd Qu.:3.925   3rd Qu.: 84.50   3rd Qu.:45.0   3rd Qu.:2.00  
##  Max.   :8.000   Max.   :257.00   Max.   :60.0   Max.   :6.00
```

```
ggpairs(data3)
```

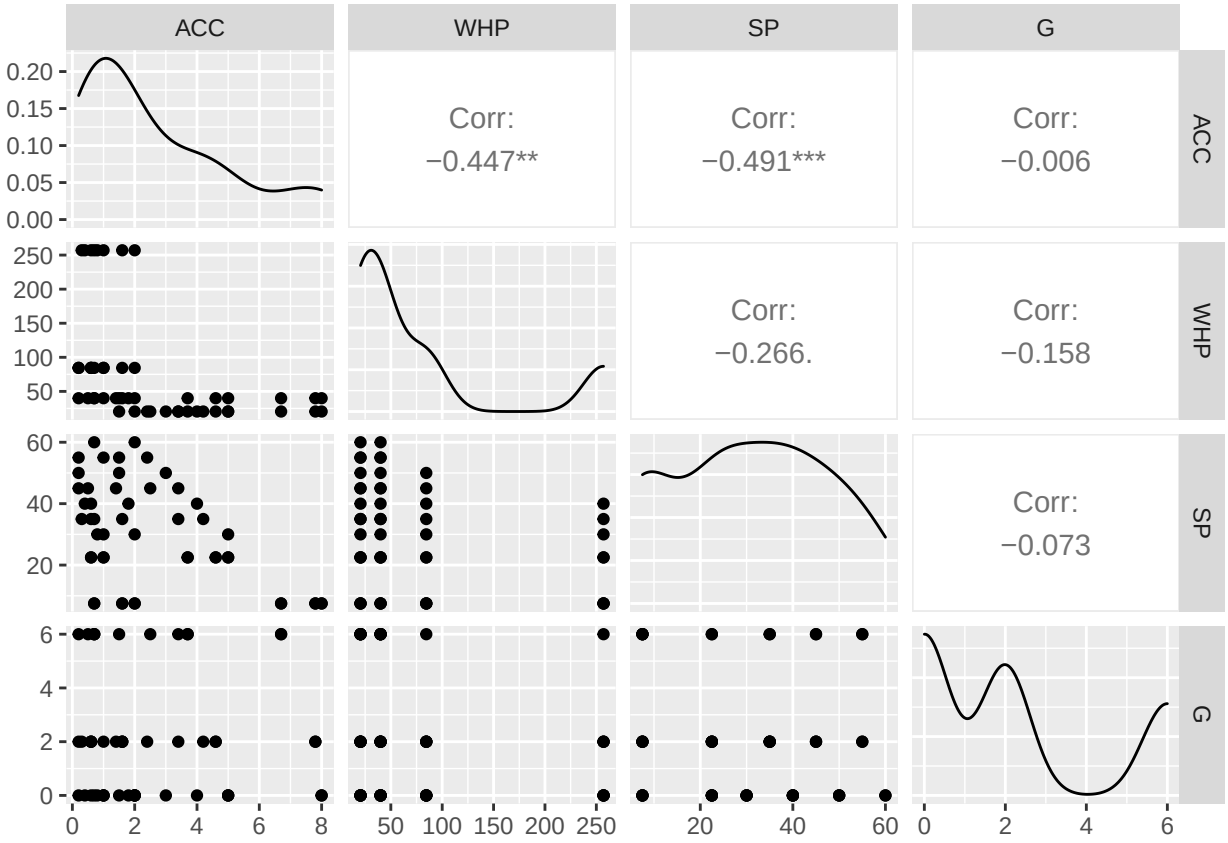


Table 2: Observations on ACC of different vehicles

Item	Variable	Description
1	ACC	acceleration
2	WHP	weight-to-horsepower ratio
3	SP	traveling speed
4	G	grade

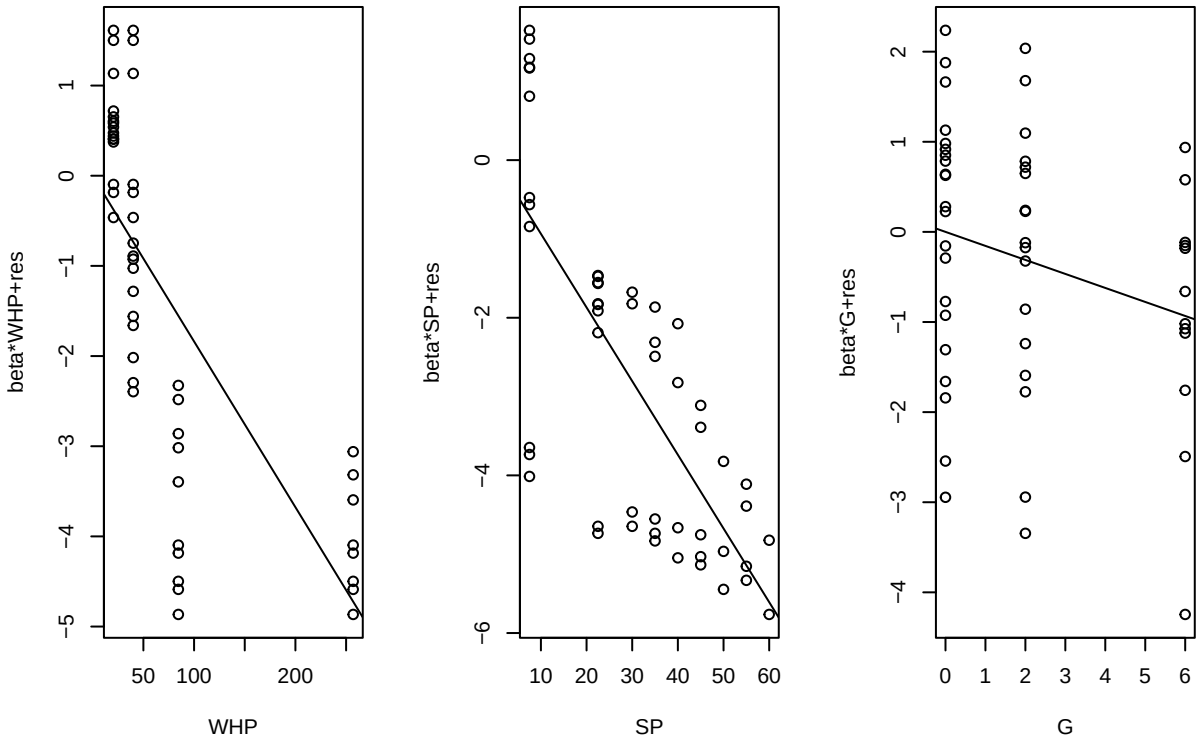
- G=0 implies the road was horizontal.
- plot suggests variables to be quantitative discrete variables.

a. Obtain the partial residual plots.

```
model1 = lm(ACC~WHP+SP+G)
summary(model1)
```

```
##
## Call:
## lm(formula = ACC ~ WHP + SP + G)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3124 -0.9003  0.2486  0.9489  2.3477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.19949    0.60087  11.982 9.57e-16 ***
## WHP         -0.01838    0.00269  -6.833 1.62e-08 ***
## SP          -0.09347    0.01307  -7.149 5.45e-09 ***
## G           -0.15548    0.09040  -1.720   0.0922 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.47 on 46 degrees of freedom
## Multiple R-squared:  0.624,  Adjusted R-squared:  0.5995
## F-statistic: 25.45 on 3 and 46 DF,  p-value: 7.451e-10
```

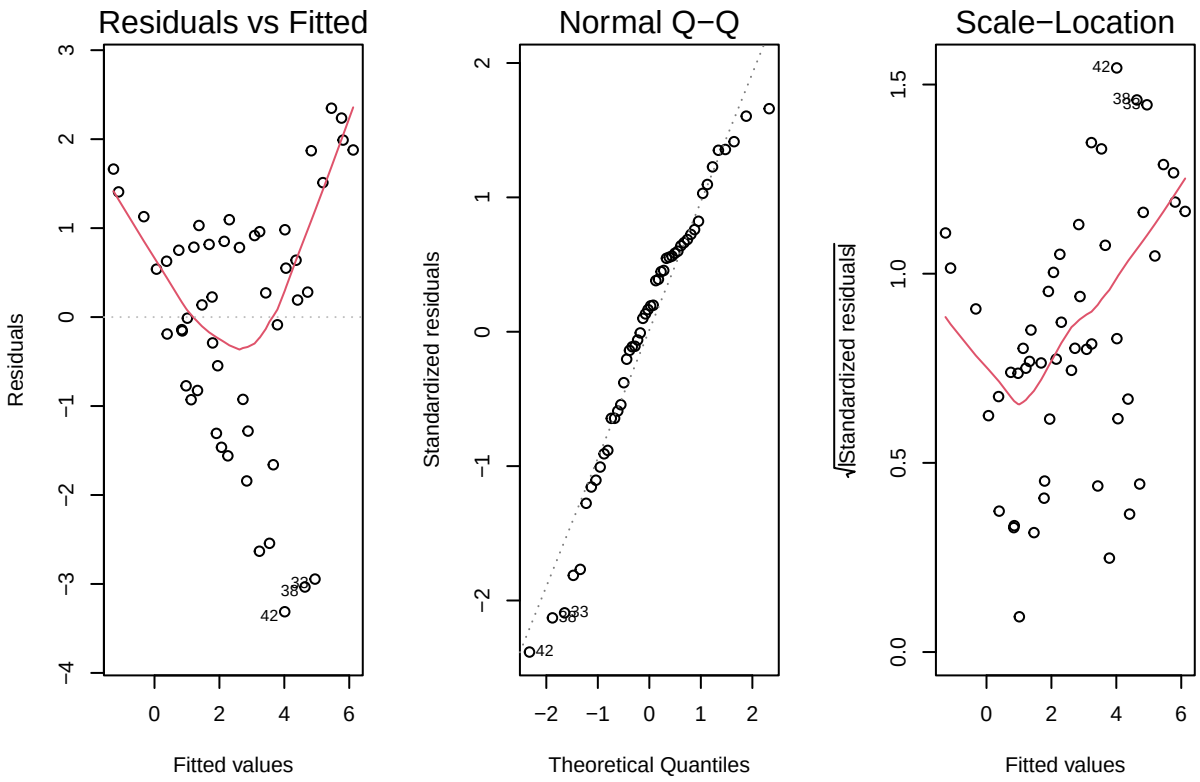


b. Obtain a good fitting model by making whatever changes you think are necessary. Obtain appropriate plots to verify that you have succeeded.

- **Check for non-constant varaince.**

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 9.002484, Df = 1, p = 0.0026961
```

- Small p-value (<0.05) in NCV-test, so under the significance level $\alpha = 0.05$, we reject the null hypothesis, that is, we have enough evidence to conclude the model has non-constant variance.
- **Check for overall pattern.**



- The plot on the left appears to have curvature in the model's mean of residuals.
- QQ plot in the middle indicates the normality assumption is not violated.
- **Mean curvature examination**

  Refer to question a, the added variable plots of both *WHP* and *SP* are proofs of violation of the assumption of constant variance.
- Since there seems to have curvature pattern in the residual plot, I examine every variable to ensure which term should be added into the model.

U = WHP

```
##
## Call:
## lm(formula = ACC ~ WHP + SP + G + U)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70446 -0.64576 -0.05457  0.54006  2.28414
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.011e+01  5.055e-01  19.999  < 2e-16 ***
## WHP         -9.569e-02  9.175e-03 -10.429 1.37e-13 ***
## SP          -1.004e-01  8.188e-03 -12.259 6.08e-16 ***
## G           -2.236e-01  5.689e-02  -3.929  0.00029 ***
## U            2.710e-04  3.162e-05   8.570 5.18e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9161 on 45 degrees of freedom
## Multiple R-squared:  0.8572, Adjusted R-squared:  0.8445
## F-statistic: 67.51 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
## [1] "significant"
```

U = SP

```
##
## Call:
## lm(formula = ACC ~ WHP + SP + G + U)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4960 -0.8530  0.3332  1.0611  2.1472
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.739128   0.801916   9.651 1.56e-12 ***
## WHP         -0.018140   0.002700  -6.720 2.66e-08 ***
## SP          -0.144399   0.051816  -2.787  0.00777 **
## G           -0.162398   0.090626  -1.792  0.07987 .
## U            0.000837   0.000824   1.016  0.31521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.47 on 45 degrees of freedom
## Multiple R-squared:  0.6324, Adjusted R-squared:  0.5998
## F-statistic: 19.36 on 4 and 45 DF,  p-value: 2.527e-09
```

```
## [1] "not significant"
```

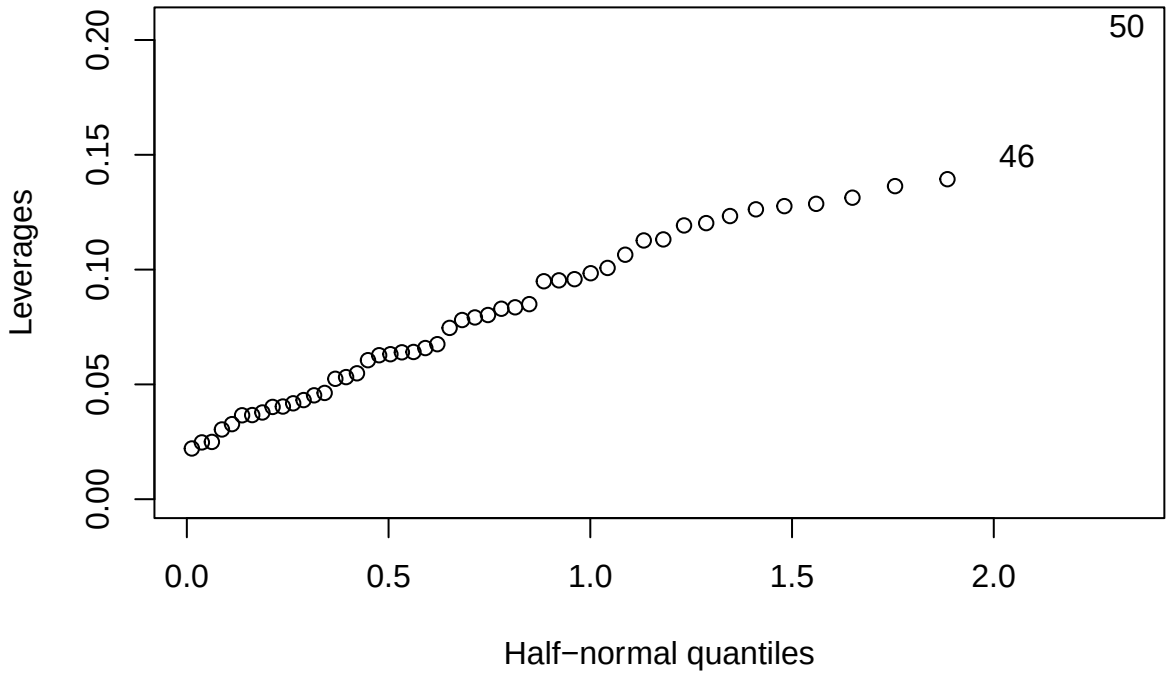
U = G

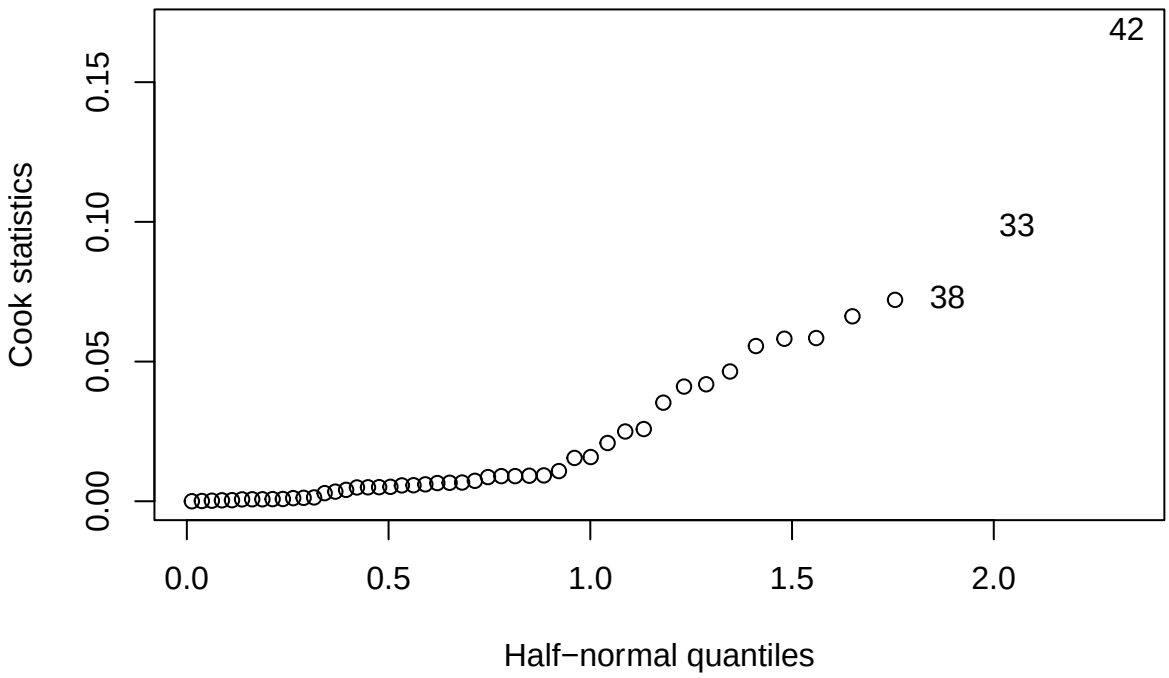
```
##
## Call:
## lm(formula = ACC ~ WHP + SP + G + U)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3022 -0.8902  0.2597  0.9359  2.3277
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.188581   0.627692  11.452 6.29e-15 ***
## WHP         -0.018388   0.002721  -6.757 2.34e-08 ***
## SP          -0.093453   0.013222  -7.068 8.08e-09 ***
## G           -0.132248   0.348496  -0.379    0.706
## U           -0.003838   0.055568  -0.069    0.945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.486 on 45 degrees of freedom
## Multiple R-squared:  0.6241, Adjusted R-squared:  0.5906
## F-statistic: 18.67 on 4 and 45 DF,  p-value: 4.146e-09
```

```
## [1] "not significant"
```

- Hence, I planned on adding the additional variable *WHP^2*.
- **Check for unusual observations.**






```
## [1] "remove: "
## NULL
```

```
## [1] "remove: "
## [1] 42 38 33 39 34 35 19
```

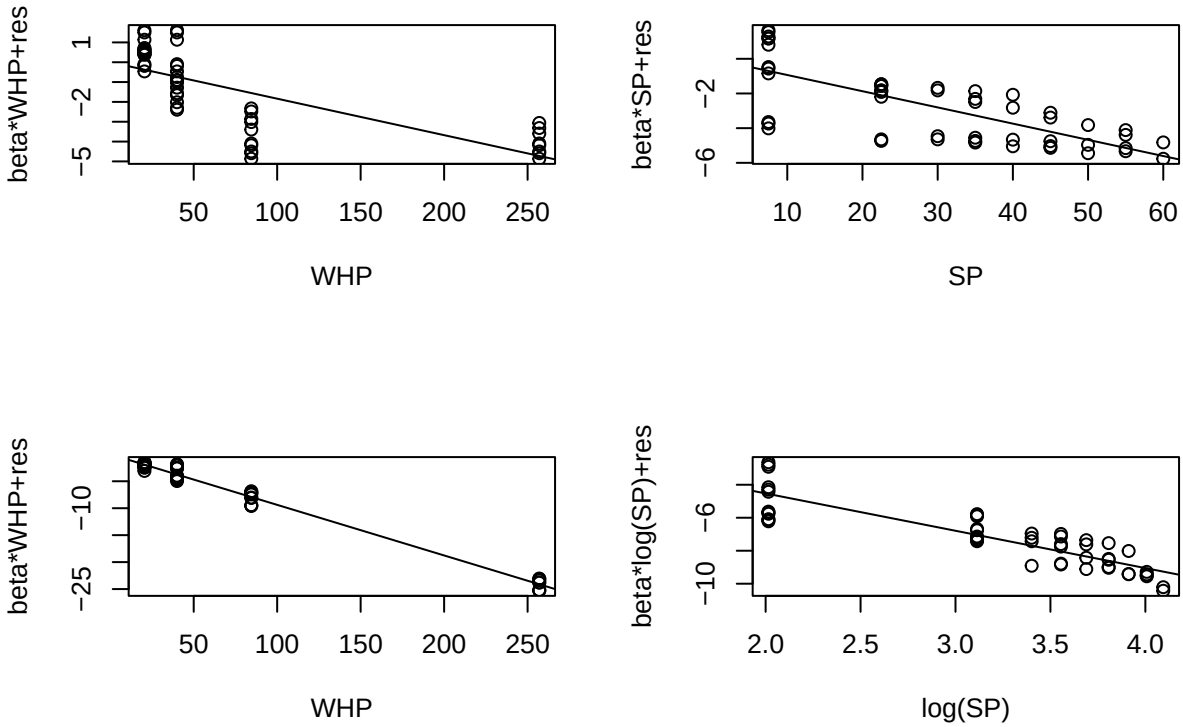
- **Check these experiment**

```
##    ACC  WHP   SP G
## 42 0.7 84.5  7.5 6
## 38 1.6 84.5  7.5 2
## 33 2.0 84.5  7.5 0
## 39 0.6 84.5 22.5 2
## 34 1.0 84.5 22.5 0
## 35 1.0 84.5 30.0 0
## 19 2.0 40.0 30.0 0
```

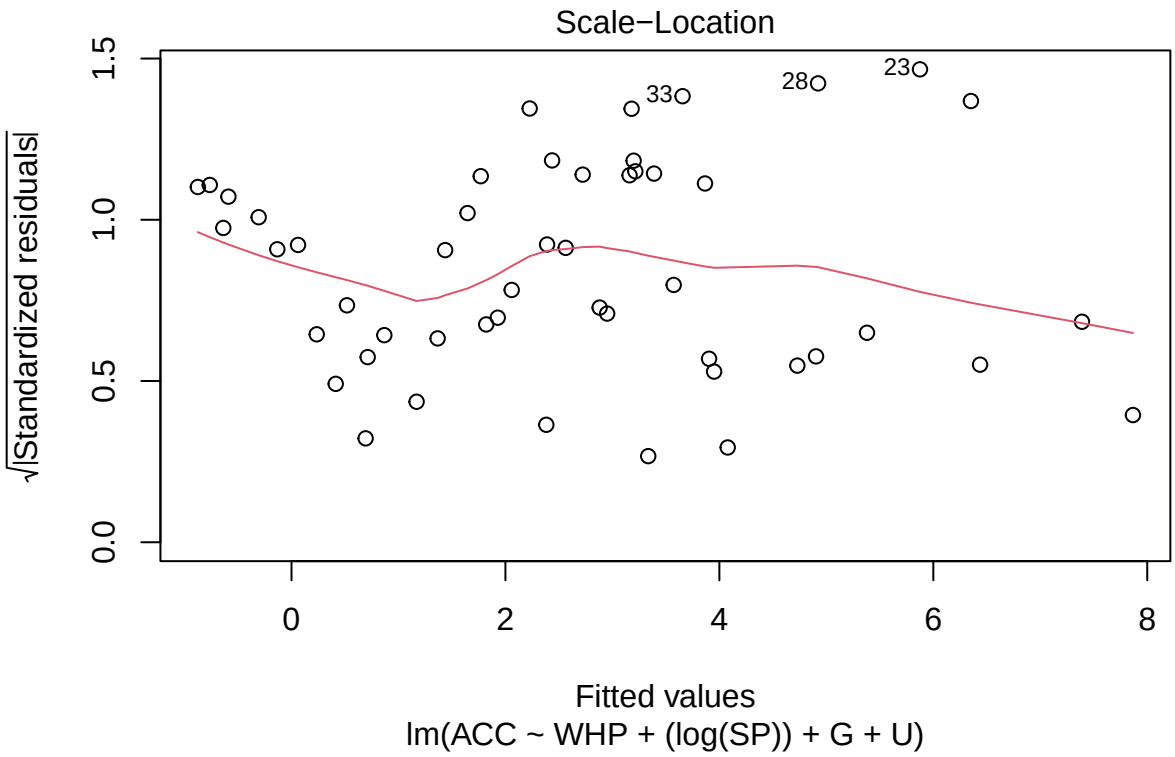
- Many of the data measured with WHP=84.5, so I assume that the unusual results may result from the measurement equipment or driver's condition.
- **Deal with non-constant variance**
  - I think the non-constant occurs in the dataset may result from the lower bound of the WHP variable. Intuitively, the heavier the vehicle, the longer the corresponding accelerate time. However, the *lowest acceleration* ≈ 0, the limited lower bound bounded the room for variance to vary.

– Also, because of the obvious non-constant variance appear in partial regression plots, I perform log transformation on both *WHP* and *SP* independent variables.

- **Transformation**.




```
plot(model2, which=3)
```




```
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.1078, Df = 1, p = 0.29256
```

- The plot and formal test both suggest that we do not reject the hypothesis that model2 has constant variance.
- Moreover, after the transformation (adding WHP^2 and logging SP), the original unusual observations are not regarded as outliers anymore. They no longer possess the same impact on the new fitted model.

```
summary(model2)
```

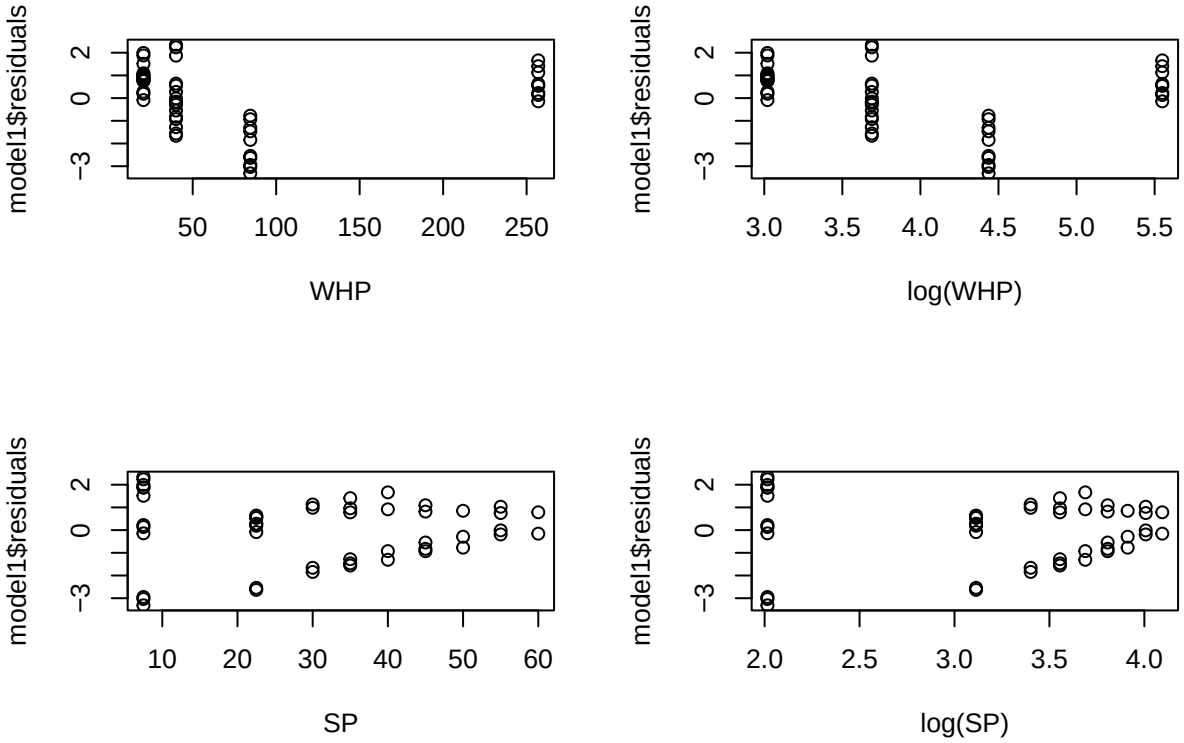
```
##
## Call:
## lm(formula = ACC ~ WHP + (log(SP)) + G + U)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65586 -0.56930 -0.08554  0.66981  1.92460
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.424e+01  7.925e-01  17.965  < 2e-16 ***
```

```
## WHP         -9.382e-02  9.380e-03 -10.002 5.16e-13 ***
## log(SP)     -2.263e+00  1.906e-01 -11.876 1.82e-15 ***
## G           -2.382e-01  5.848e-02  -4.074 0.000185 ***
## U            2.670e-04  3.237e-05   8.248 1.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9386 on 45 degrees of freedom
## Multiple R-squared:  0.8501, Adjusted R-squared:  0.8367
## F-statistic: 63.78 on 4 and 45 DF,  p-value: < 2.2e-16
```

- Compared to the R^2 value obtain from model1, by simply transforming the mean structure, we enhance the proportion of variance interpret by the predictors.

c. Explain why we can eliminate heteroscedasticity in partial residual plots shown in question a by simply transforming the predictors.

- The data for this question gives observations on ACC of different types of vehicles. Based on high school knowledge, acceleration is the change in velocity over the change in time, in other words, it measures how fast velocity changes in meters per second squared.
- Since the equation suggest the existance of certain variable that is not included in this data (time), and the indication that underlying true model has curvature pattern, simply fitting all variable is insufficient to get a desired output.
- That is to say, model1 is too simple to fit the response, the true model is actually more complex than the fitted model. Therefore, variance that is not explained by model1 is then presented on the partial regression plots, showing heteroscedasticity.
- Accordingly, we need to achieve better fitting result by modifying the model's mean structure.



- Hence, by adding additional term and transforming predictors, we made a relatively complex model, model2. The additional term explains part of the variance shown in prplot displayed in question a, thus eliminating the phenomenon of heteroscedasticity.

```
detach(data3)
```