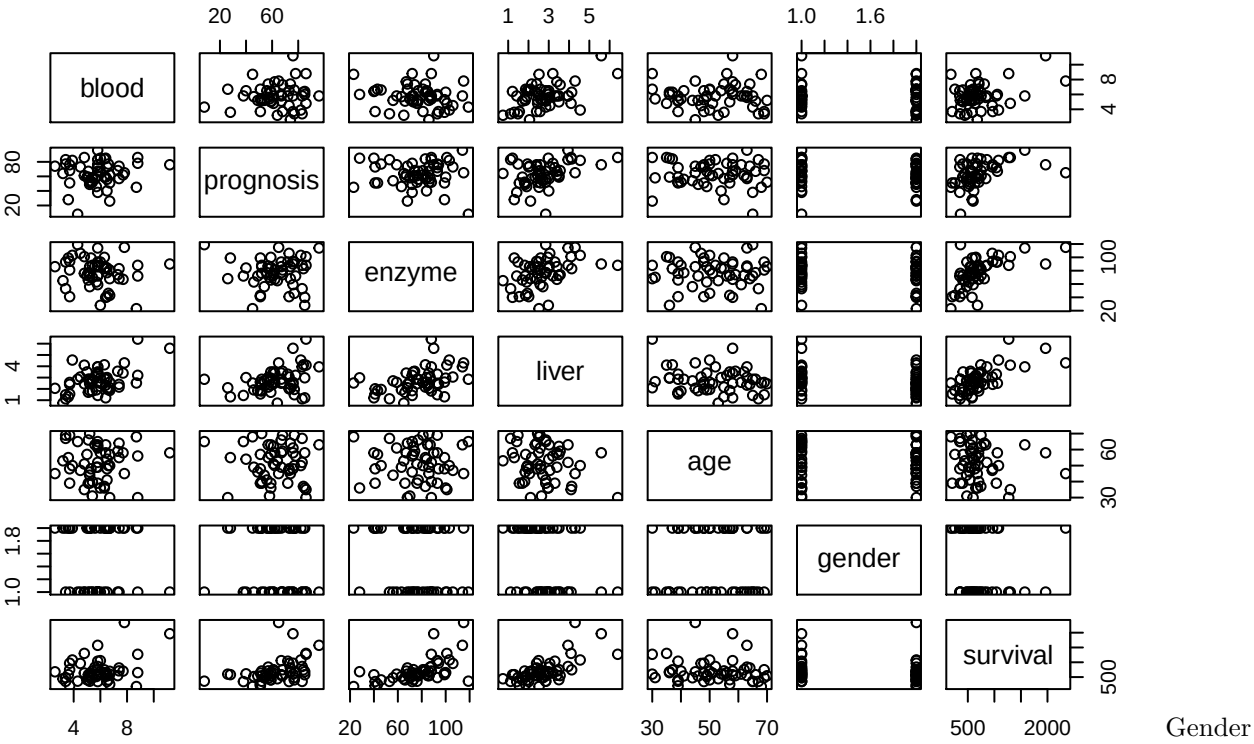
title: “assignment” author: “jackson lane” date: “17/05/2021” output: pdf_document



Gender variable appears to be categorical based off the scatterplot. Liver and survival have a slightly linear relationship. also enzyme and survival also have a slight linear relationship.

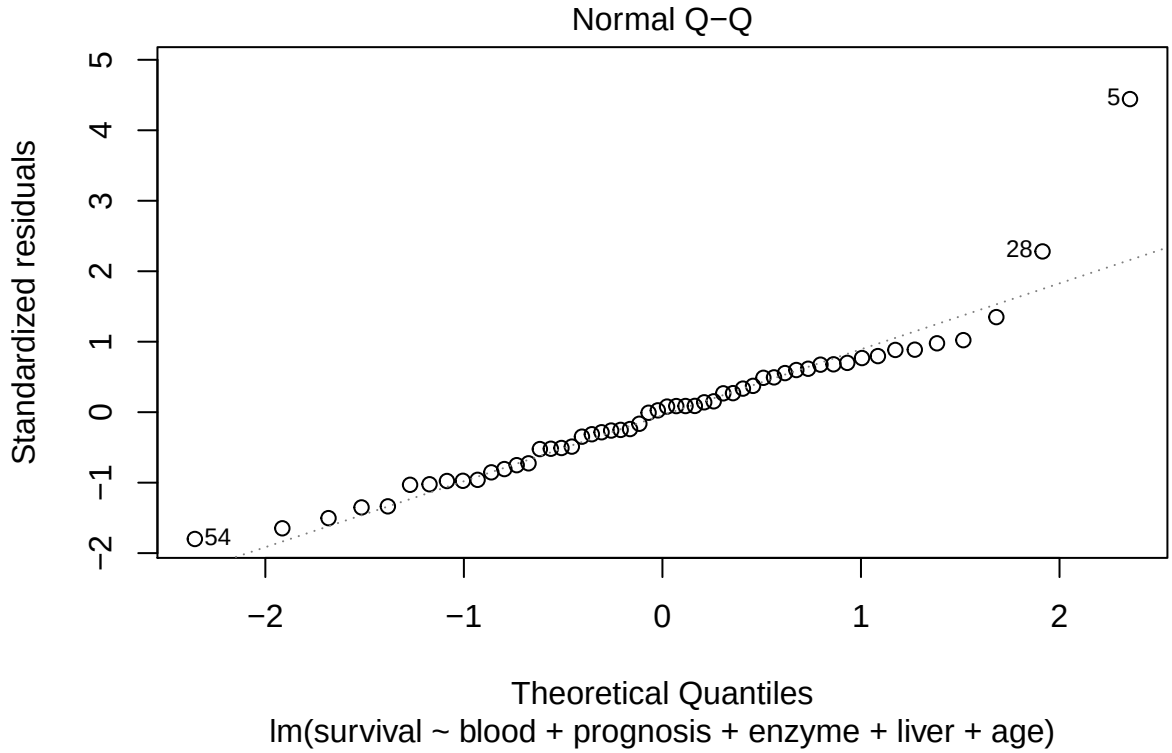
the correlation matrix will not work because gender is categorical variable and correlation matrix requires numeric variables.

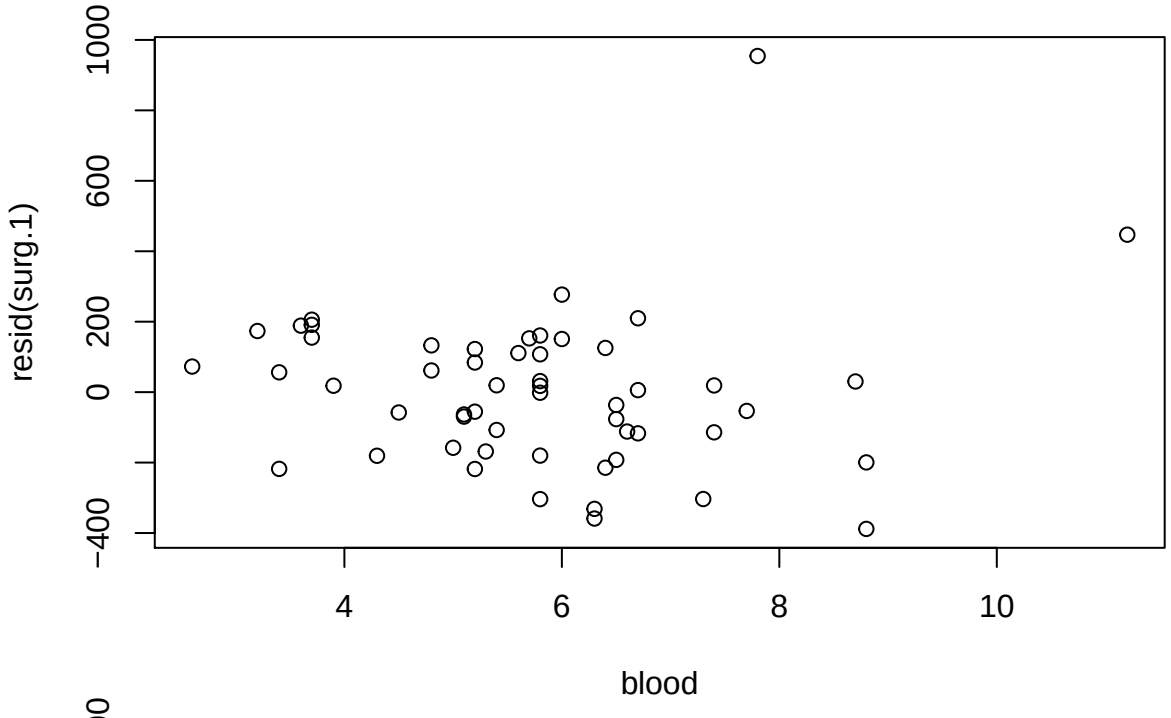
There is a strong positive correlation between survival and every variable besides age that has a negative correlation.

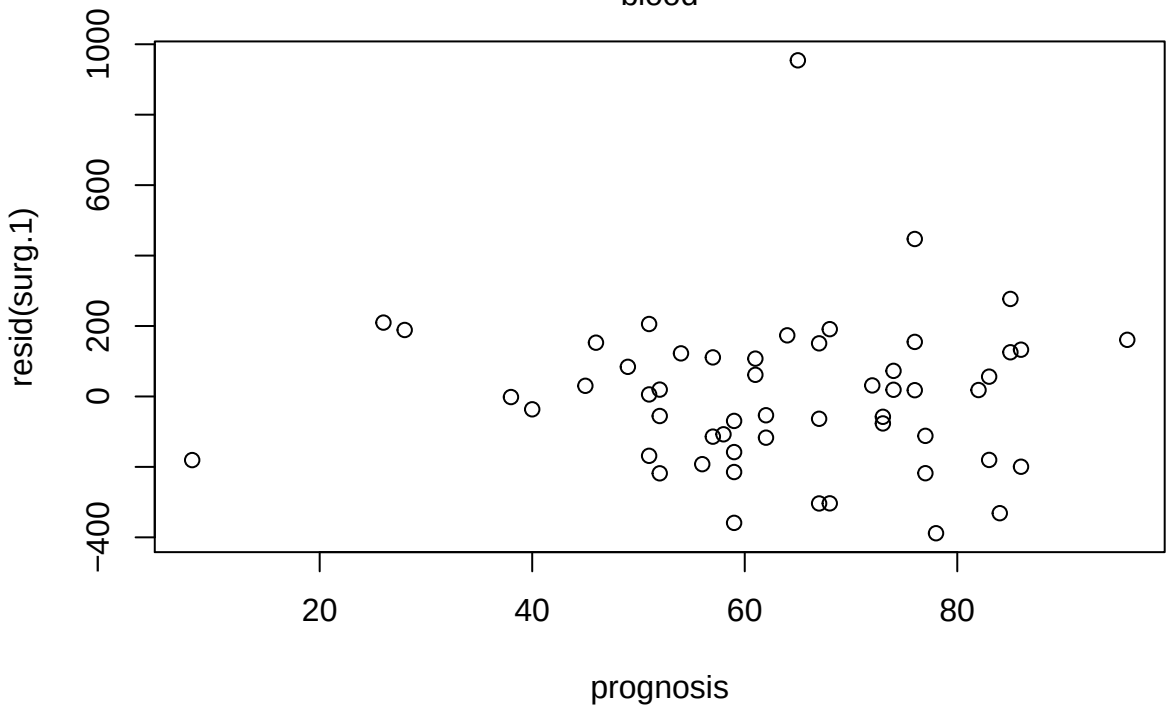
```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + liver +
##     age, data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.34 -147.74   11.74  124.67  954.32
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1179.367    275.619  -4.279 8.91e-05 ***
## blood          86.630     26.905   3.220 0.002302 **
## prognosis       8.501      2.137   3.978 0.000234 ***
## enzyme         11.124      1.958   5.683 7.62e-07 ***
## liver          38.554     49.251   0.783 0.437595
## age            -2.340      2.969  -0.788 0.434514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.6 on 48 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.6632
```

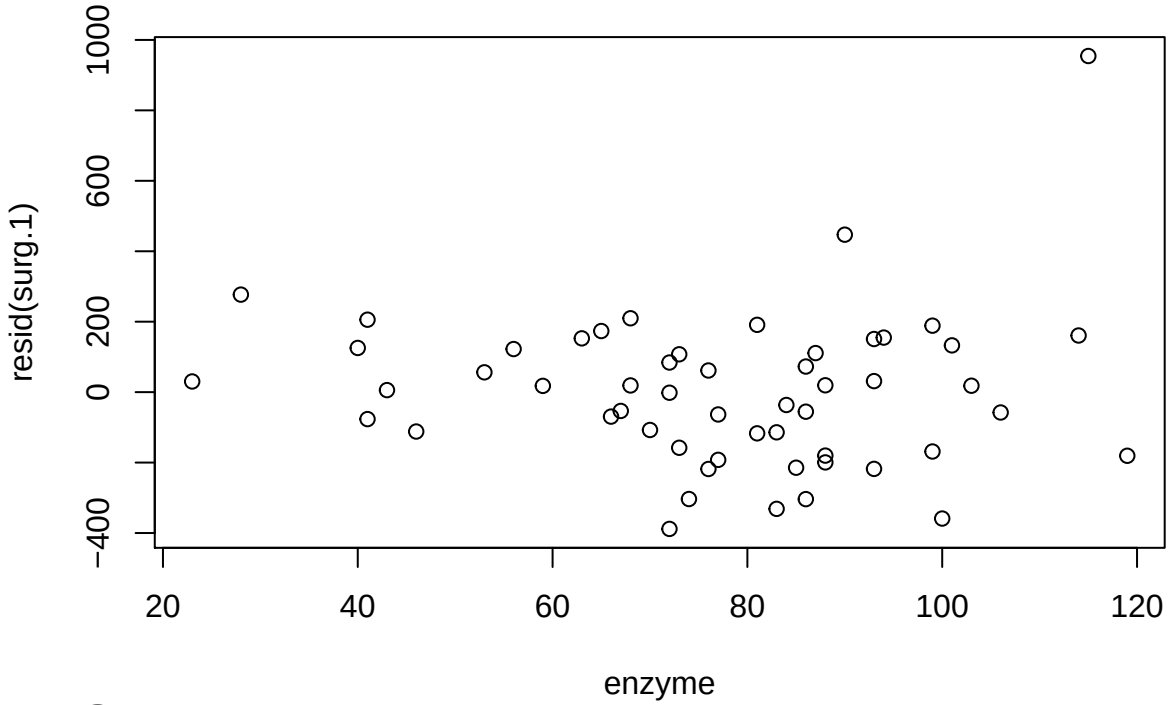
```
## F-statistic: 21.87 on 5 and 48 DF,  p-value: 2.386e-11
```

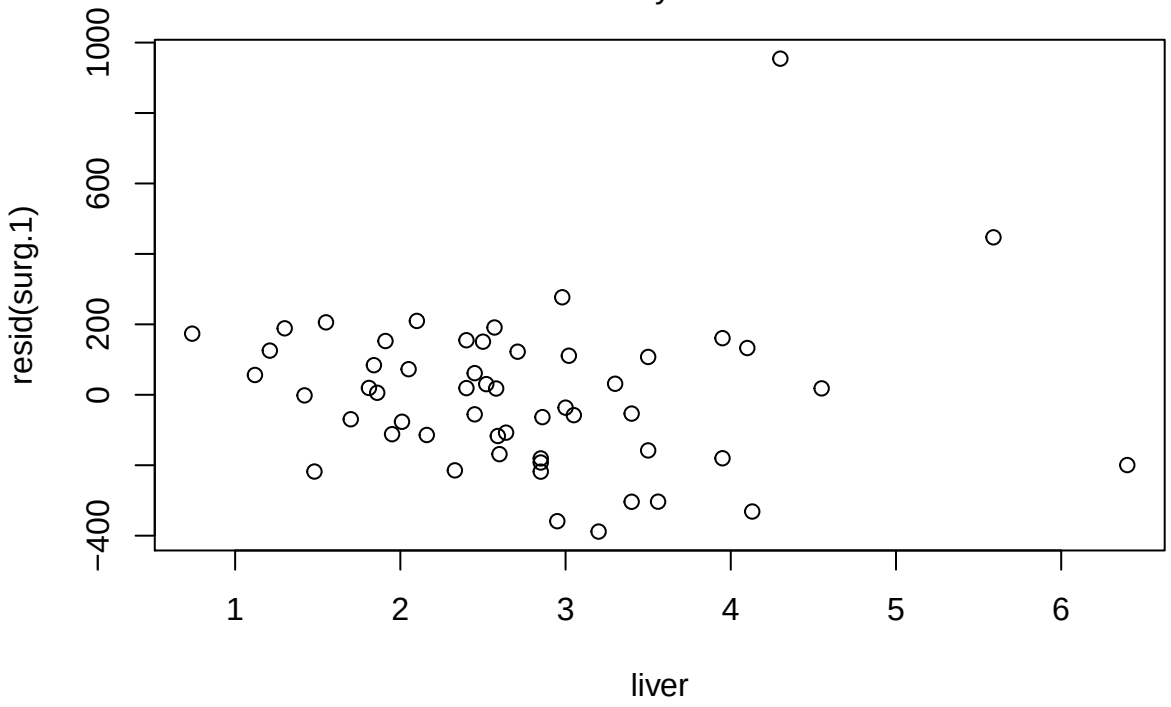
```
plot(surg.1, which = 1:2)
```

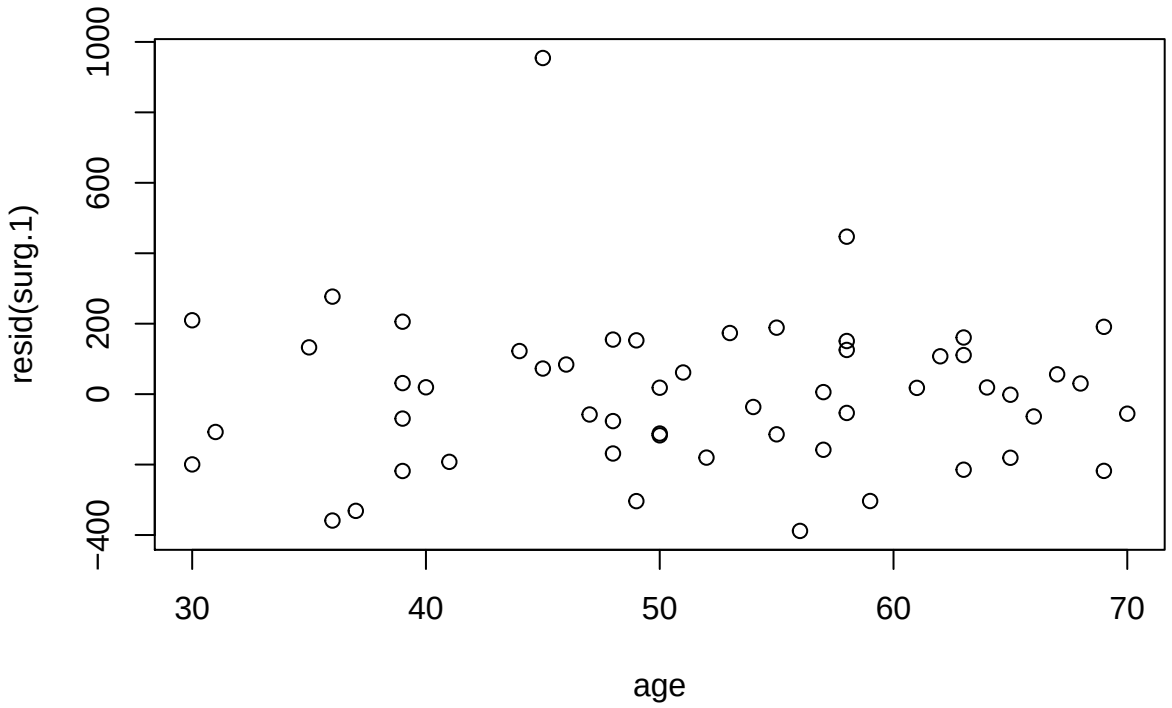



resid(surg.1)
1000
600
200
0
−400
4
6
8
10
blood


resid(surg.1)
1000
600
200
0
−400
20
40
60
80
prognosis


resid(surg.1)
1000
600
200
0
−400
20
40
60
80
100
120
enzyme


resid(surg.1)
1000
600
200
0
−400
1
2
3
4
5
6
liver



checking assumptions. The Fitted vs residuals has a significant curvature indicating it may not be normally distributed The normal QQ plot also has a slight curvature indicating the relationship may not be linear, however there is an outlier and this may be the cause for the curvature. the residuals vs predictors look concentrated could be a potential pattern.

mathematical equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon; \epsilon \sim N(0, \sigma^2)$$

defining parameters

$$Y_i = survival$$
$$X_1 = blood$$
$$X_2 = prognosis$$
$$X_3 = enzyme$$
$$X_4 = liver$$
$$X_5 = age$$
$$Hypothesis H\ 0\ : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 H\ 1\ : \beta_i \neq 0$$

produce anova table

```
## Analysis of Variance Table
##
## Response: survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## blood      1 1005152 1005152 18.8997 7.133e-05 ***
## prognosis  1 1278496 1278496 24.0393 1.121e-05 ***
## enzyme     1 3442172 3442172 64.7226 1.883e-10 ***
## liver      1   57862   57862  1.0880    0.3021
## age        1   33032   33032  0.6211    0.4345
## Residuals 48 2552807   53183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Liver and age don't have significant p values. will remove liver because has the highest p value and check diagnostics.

```
totalregSS = 1005152+1278496+3442172+57862+33032
```

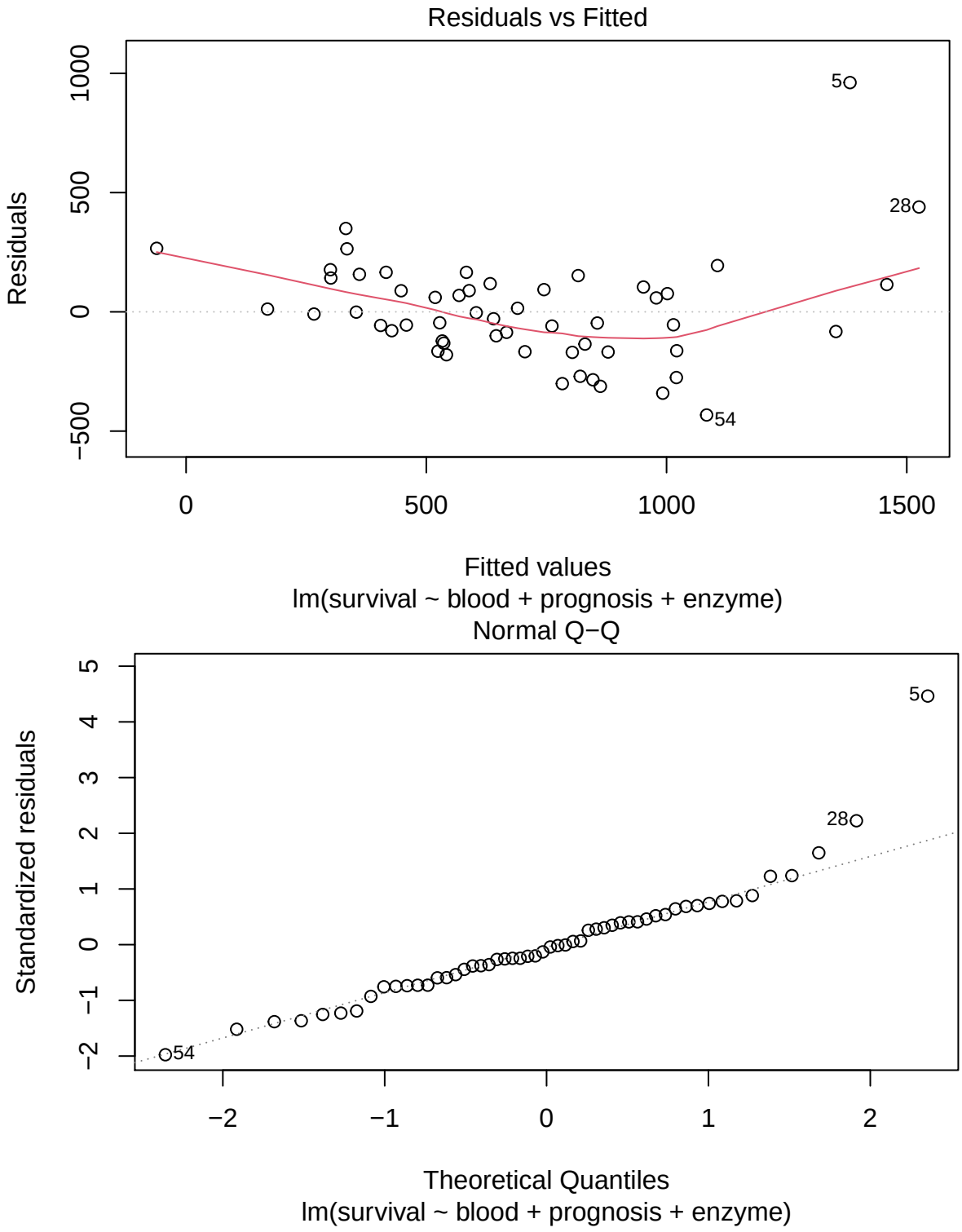
```
## Analysis of Variance Table
##
## Response: survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## blood      1 1005152 1005152  19.050 6.559e-05 ***
## prognosis  1 1278496 1278496  24.231 1.009e-05 ***
## enzyme     1 3442172 3442172  65.238 1.457e-10 ***
## age        1   58305   58305   1.105    0.2983
## Residuals 49 2585396   52763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

age still does not have a significant p value need to remove.

```
anova(surg.3)
```

```
## Analysis of Variance Table
##
## Response: survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## blood      1 1005152 1005152  19.010 6.484e-05 ***
## prognosis  1 1278496 1278496  24.180 9.883e-06 ***
## enzyme     1 3442172 3442172  65.101 1.303e-10 ***
## Residuals 50 2643701   52874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

checking diagnostics



the Residauls vs fitted and QQplot still have curvatures.

compute F stat

```
## [1] 21.87434
```

the null distribution is that none of the predictor #variables have an effect on survival time from liver #operation.

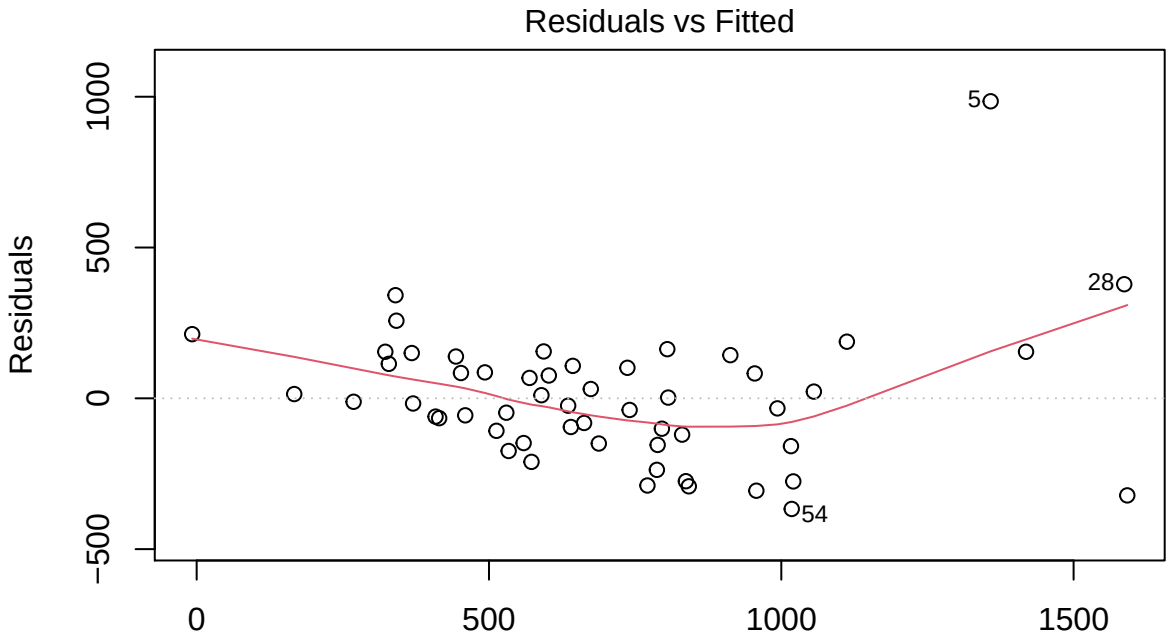
conclusion. After backwise step method all predictor variables have significant p values. The F static is greater than the p value meaning we need #to reject null hypothesis. Contextually the removal of liver and age variables is not correct because they are dependent variables the outliers from survival look to have created a curvature in the residuals vs fitted and normal QQ plot.

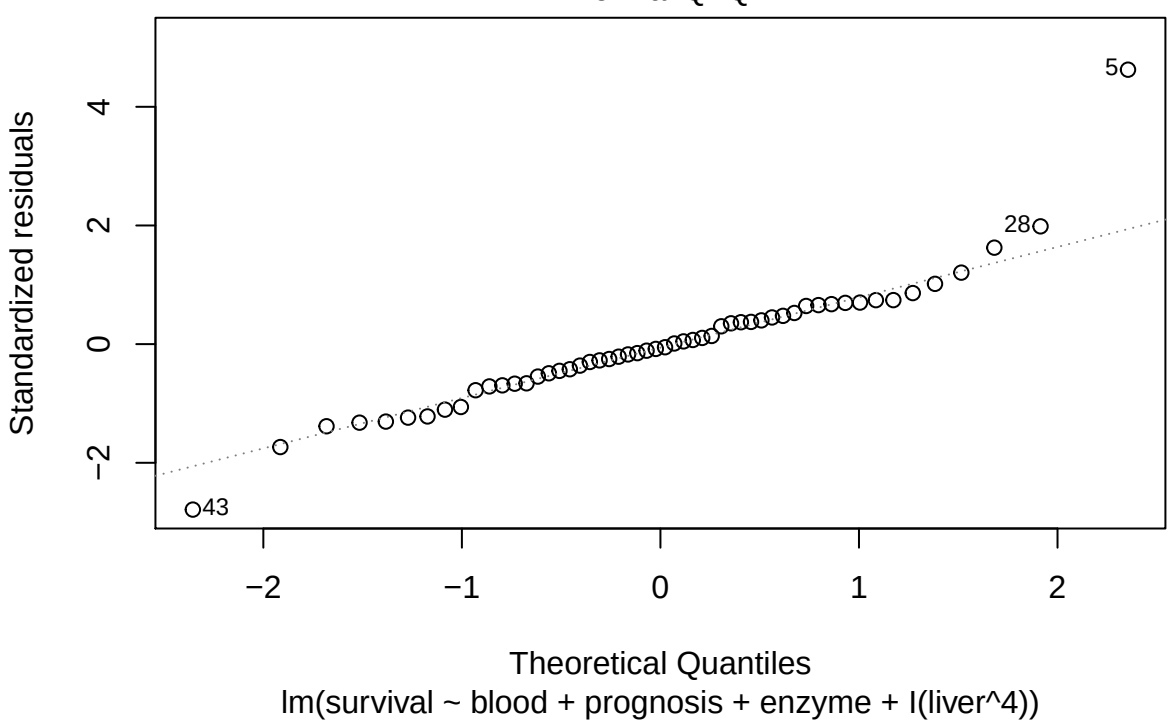
```
## Analysis of Variance Table
##
## Response: survival
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## liver       1 3804272 3804272 52.3321 2.881e-09 ***
## I(liver^2)  1  107925  107925  1.4846   0.22888
## I(liver^3)  1  435679  435679  5.9933   0.01799 *
## I(liver^4)  1  459596  459596  6.3223   0.01526 *
## Residuals  49 3562048   72695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: survival
##           Df  Sum Sq Mean Sq F value Pr(>F)
## age        1  118863  118863  0.7023 0.4063
## I(age^2)   1   47526   47526  0.2808 0.5987
## I(age^3)   1  144451  144451  0.8534 0.3603
## I(age^4)   1    1920    1920  0.0113 0.9156
## I(age^5)   1    7231    7231  0.0427 0.8371
## I(age^6)   1   94392   94392  0.5577 0.4589
## Residuals 47 7955138  169258
```

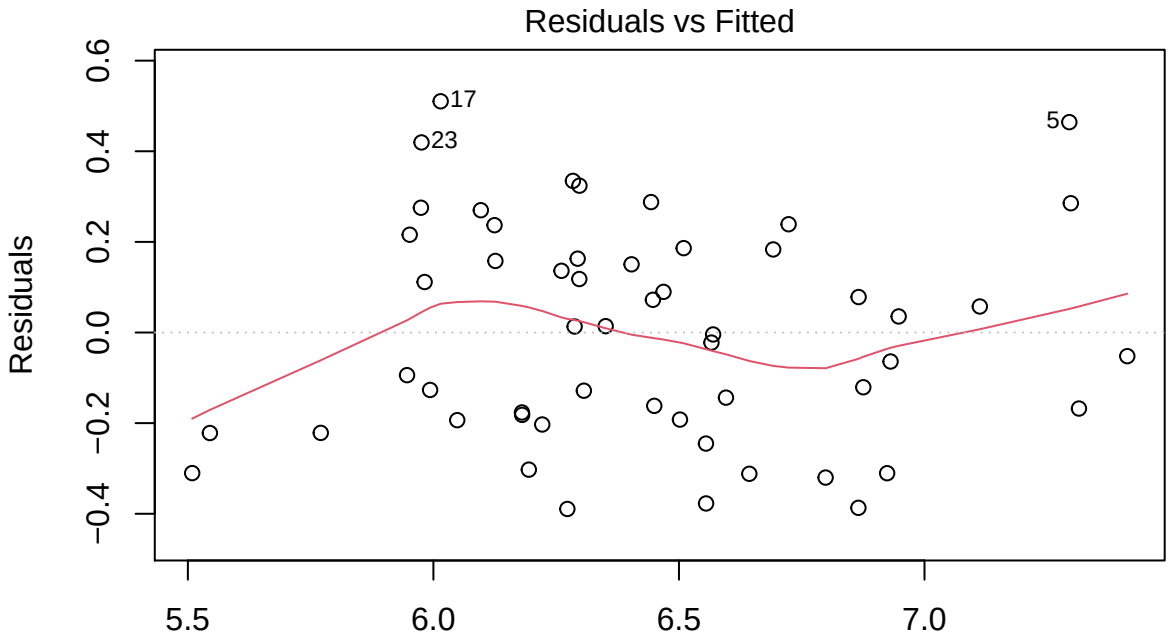
fitting the model with quadratic model

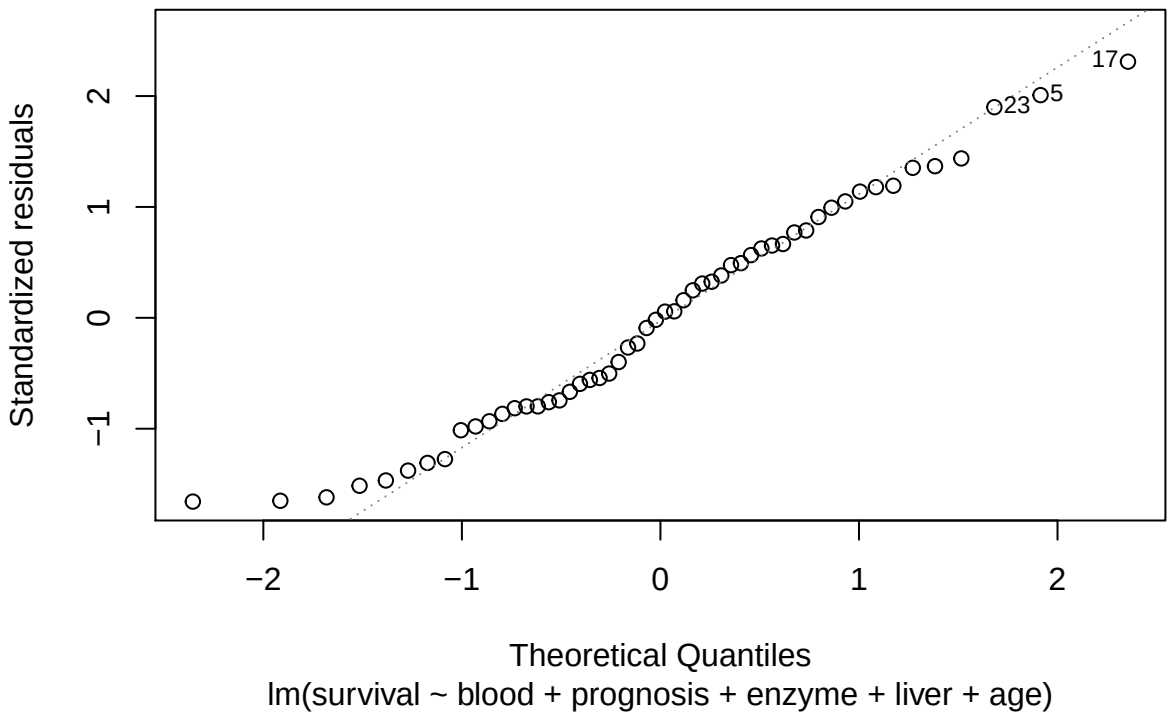
```
## Analysis of Variance Table
##
## Response: survival
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## blood       1 1005152 1005152 19.3045 5.961e-05 ***
## prognosis   1 1278496 1278496 24.5543 9.019e-06 ***
## enzyme      1 3442172 3442172 66.1089 1.207e-10 ***
## I(liver^4)  1   92361   92361  1.7738    0.1891
## Residuals  49 2551340   52068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

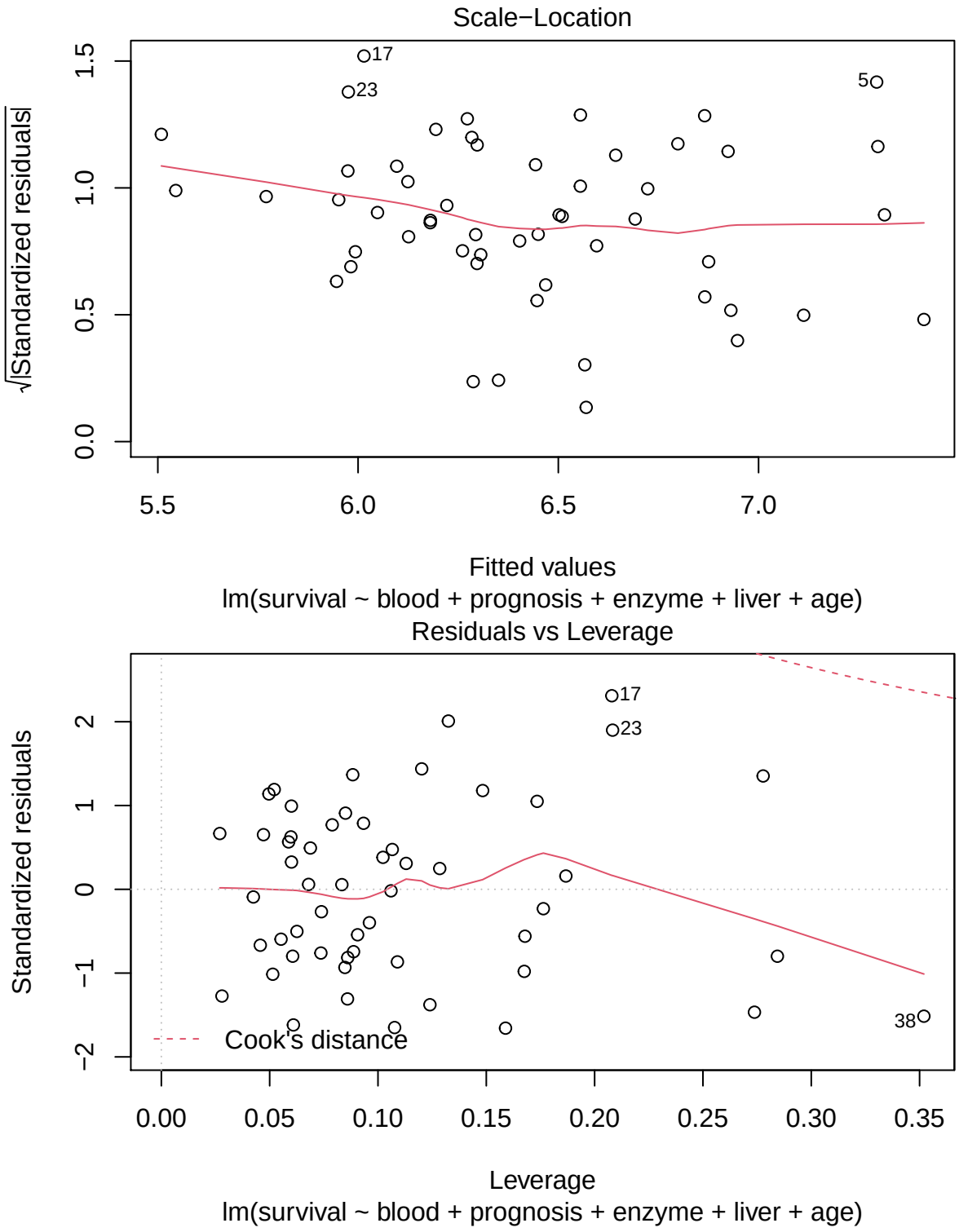


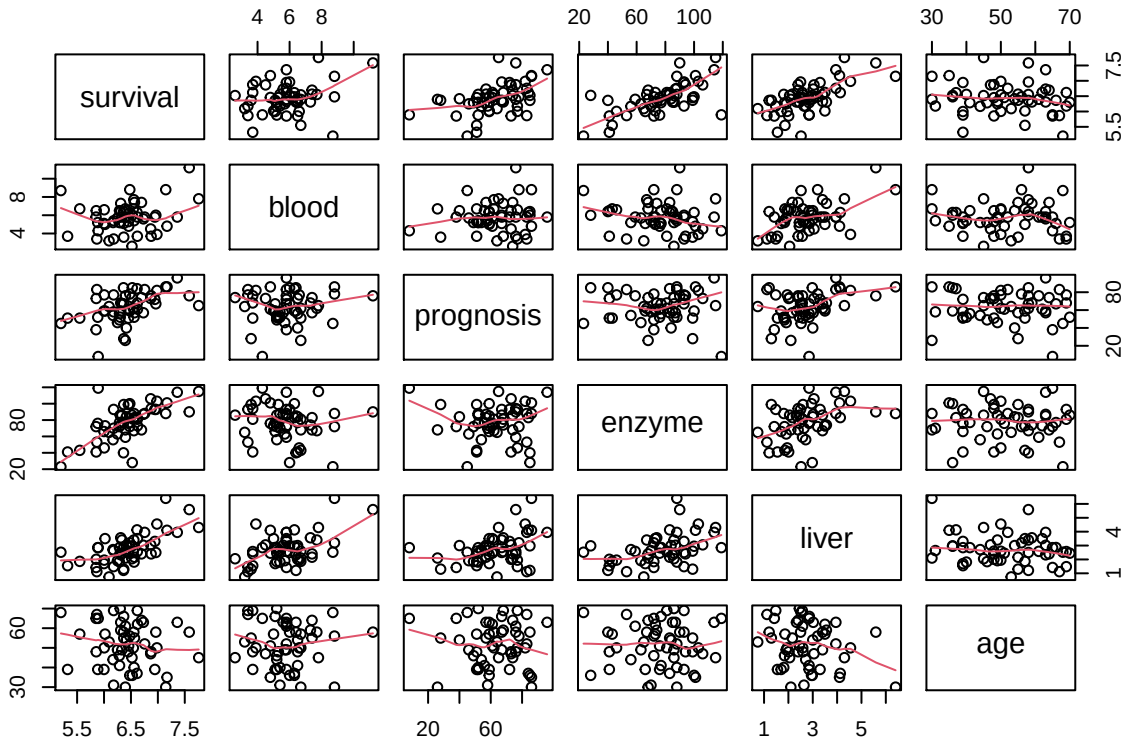


It is not appropriate to use the multiple regression model because the assumptions have not been met all of the QQplots are not linear and the residuals vs fitted all have curvature.

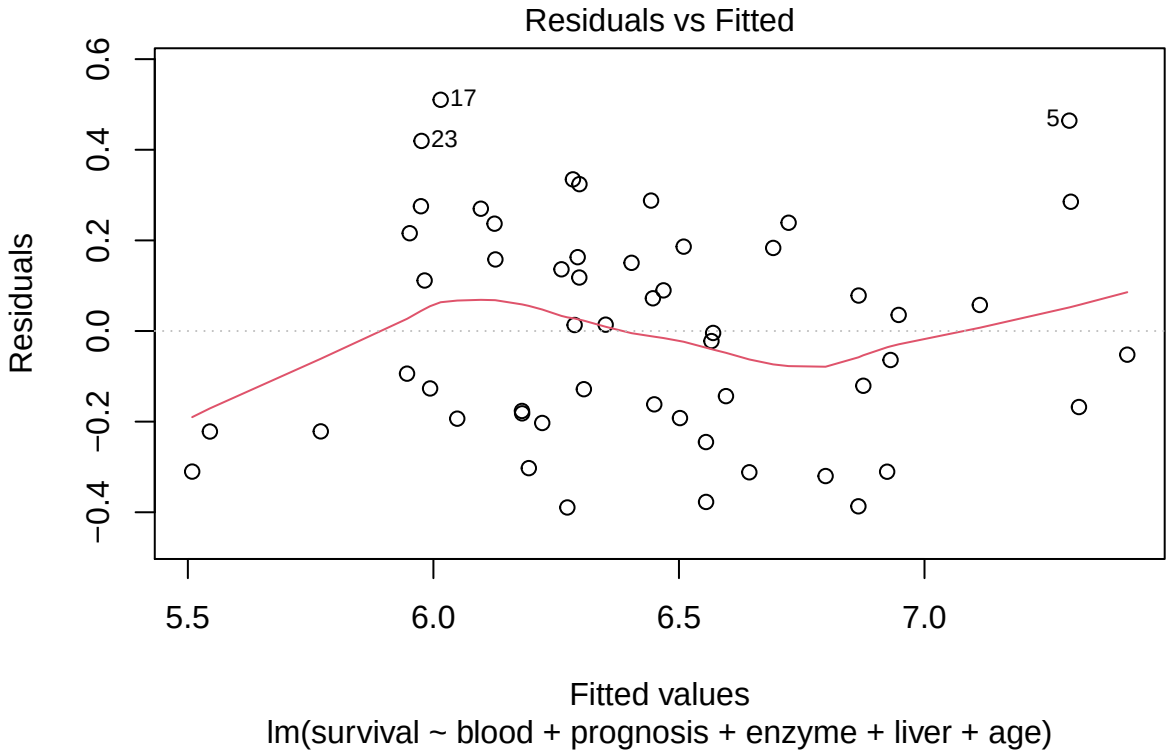

Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.2
−0.4
17
23
5
5.5
6.0
6.5
7.0
Fitted values
lm(survival ~ blood + prognosis + enzyme + liver + age)

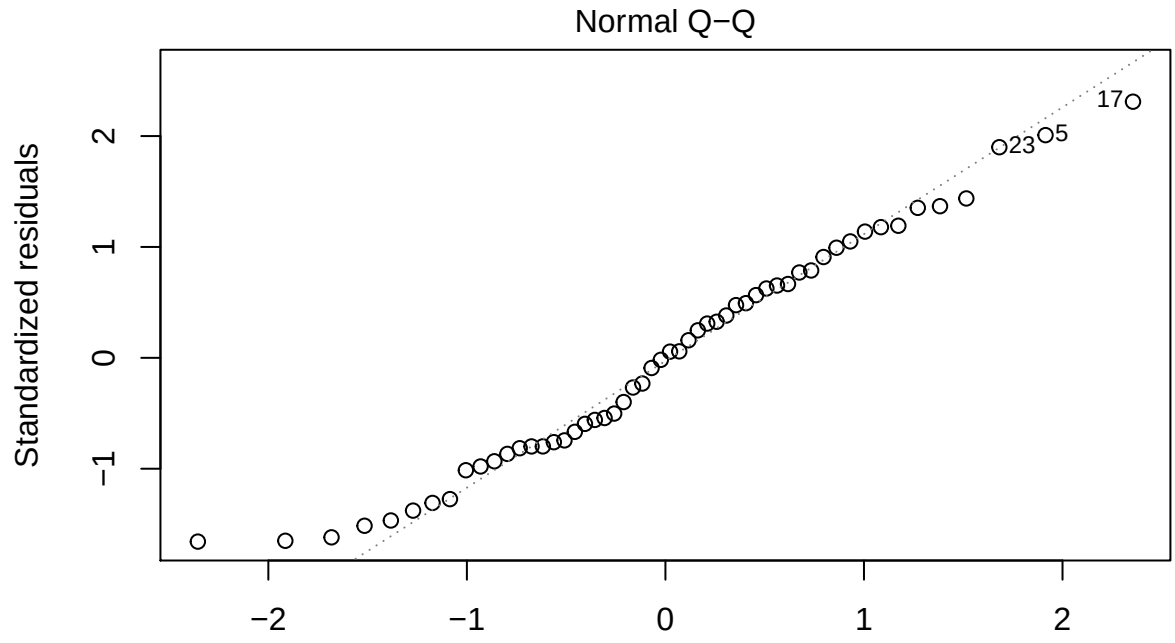

Normal Q−Q
Standardized residuals
2
1
0
−1
17
23
5
−2
−1
0
1
2
Theoretical Quantiles
lm(survival ~ blood + prognosis + enzyme + liver + age)

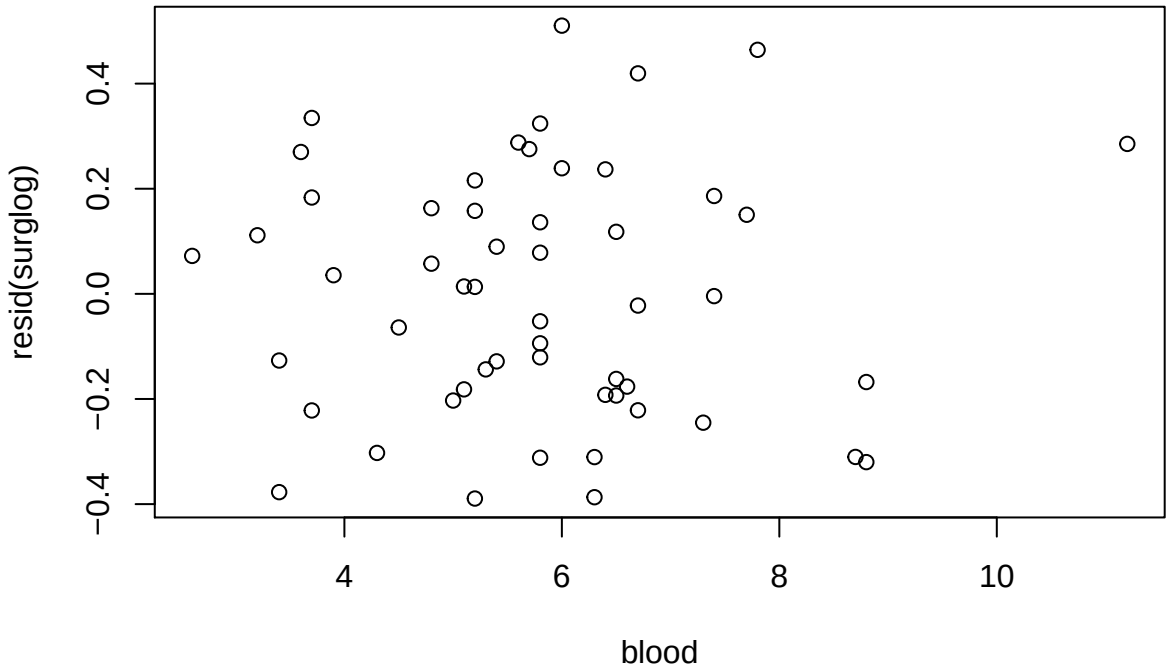

Scale−Location
√|Standardized residuals|
17
23
5
0.0
0.5
1.0
1.5
5.5
6.0
6.5
7.0
Fitted values
lm(survival ~ blood + prognosis + enzyme + liver + age)
Residuals vs Leverage
Standardized residuals
17
23
38
0.5
Cook's distance
−2
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Leverage
lm(survival ~ blood + prognosis + enzyme + liver + age)

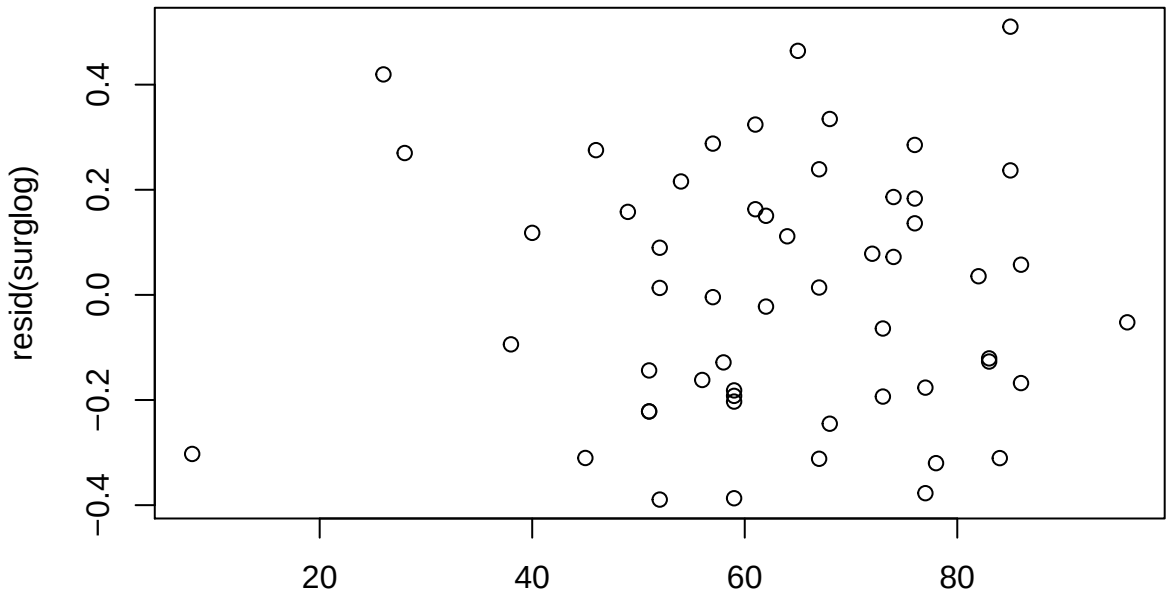


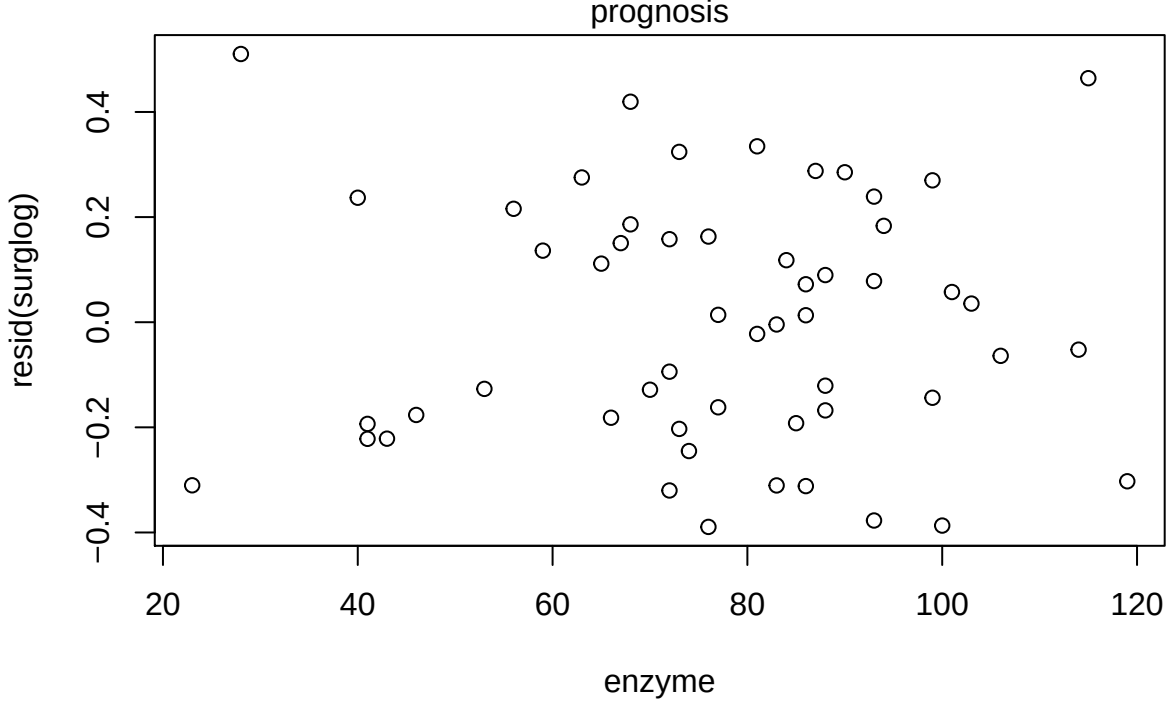

```
par(mfrow = c(2, 2))
```

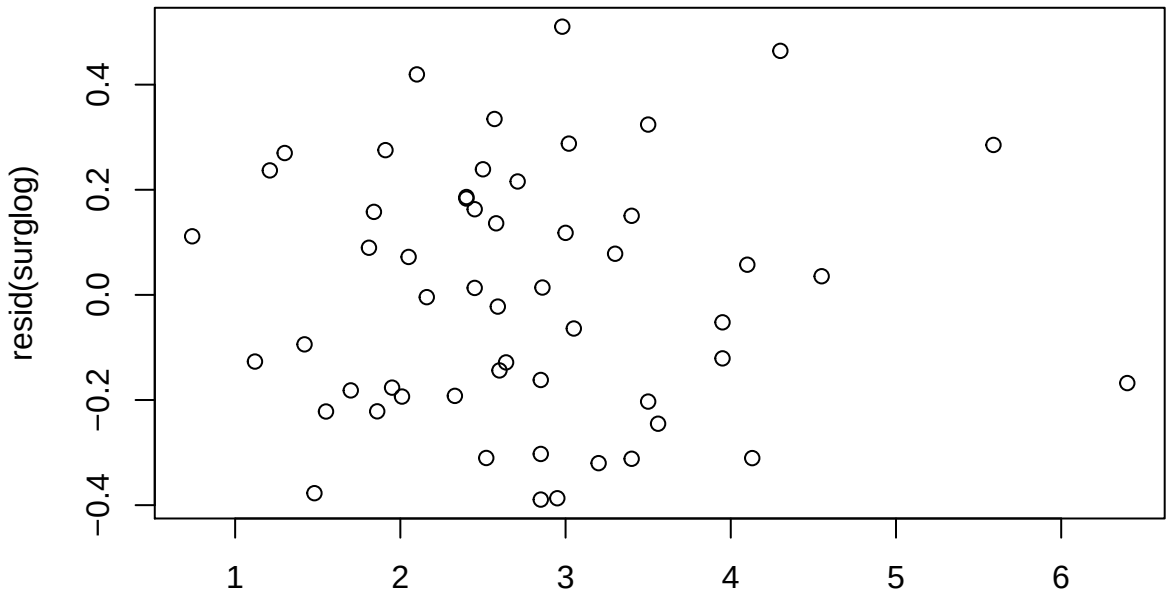


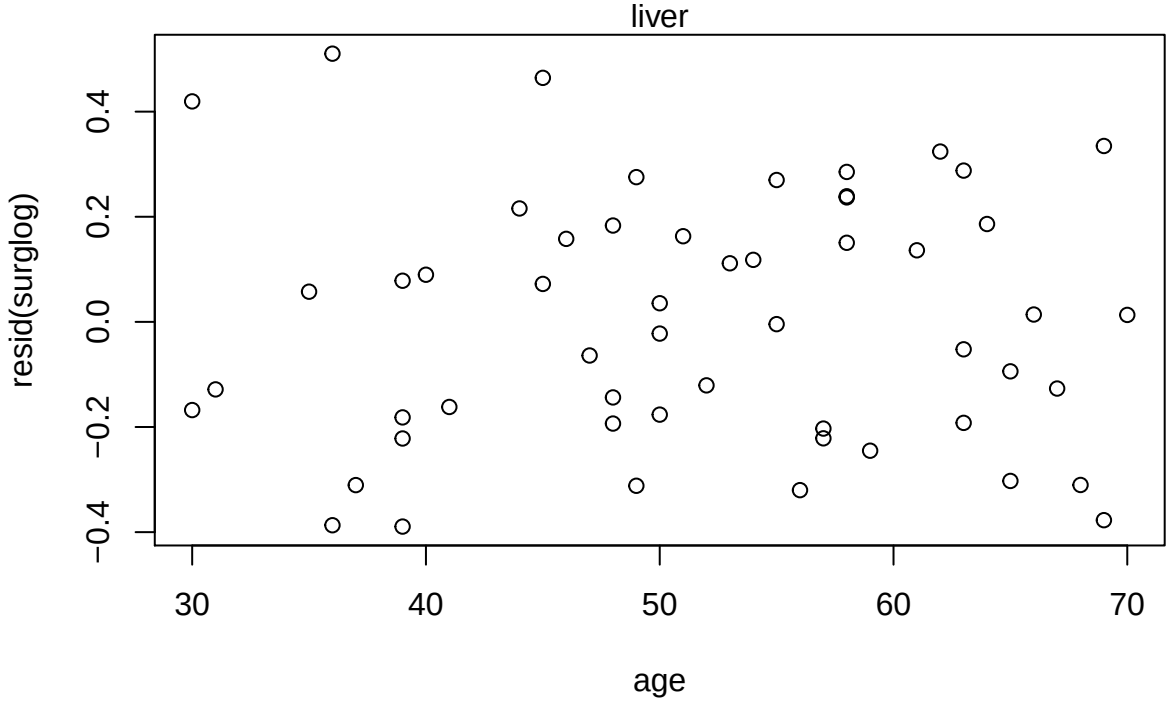

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(survival ~ blood + prognosis + enzyme + liver + age)
17
23
5


resid(surglog)
blood


resid(surglog)
0.4
0.2
0.0
−0.2
−0.4
20
40
60
80
prognosis


resid(surglog)
0.4
0.2
0.0
−0.2
−0.4
20
40
60
80
100
120
enzyme





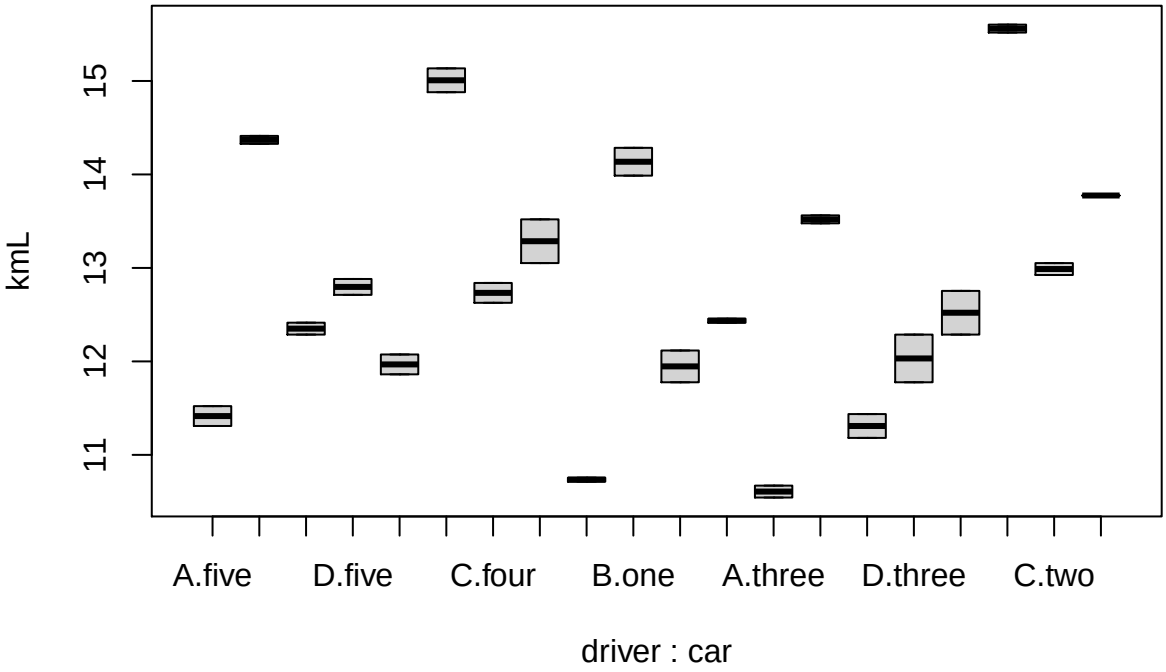

```
## Analysis of Variance Table
##
## Response: survival
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## blood      1 0.7770  0.7770  12.6107 0.0008711 ***
## prognosis  1 2.5904  2.5904  42.0446 4.548e-08 ***
## enzyme     1 6.3286  6.3286 102.7187 1.630e-13 ***
## liver      1 0.0244  0.0244   0.3963 0.5319831
## age        1 0.1268  0.1268   2.0573 0.1579579
## Residuals 48 2.9573  0.0616
## ---
```

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

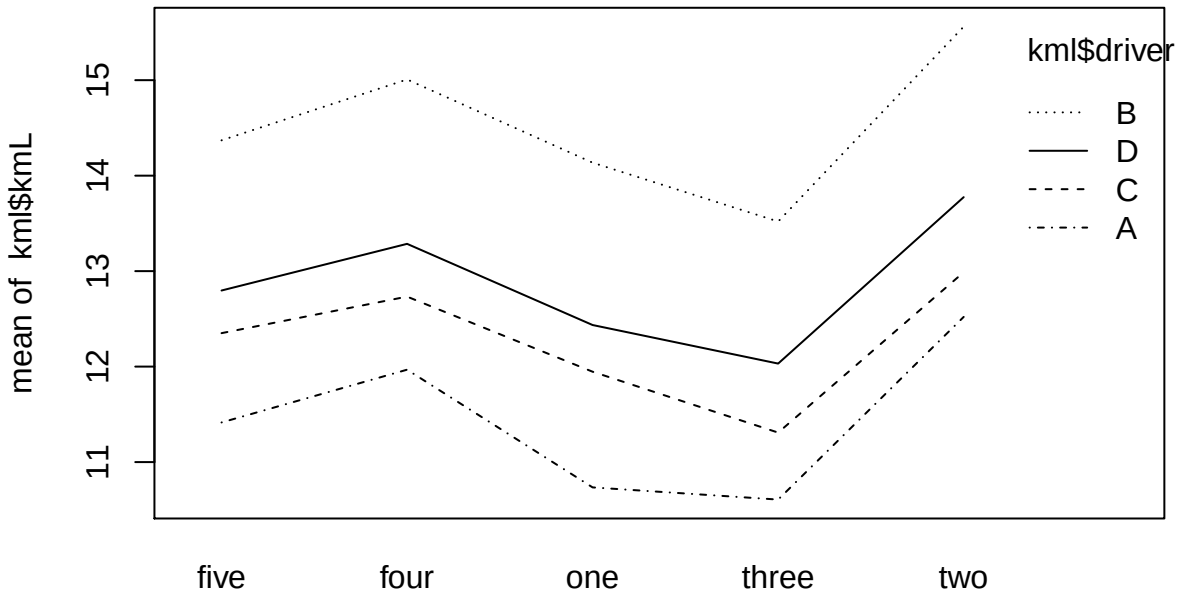
The use of logarithm helps by clarifying the exponential data increase instead of using the survival data set that has significant outlier. This is shown in the residuals where all three assumptions homoscedasticity, normality and un-correlatedness have been met.

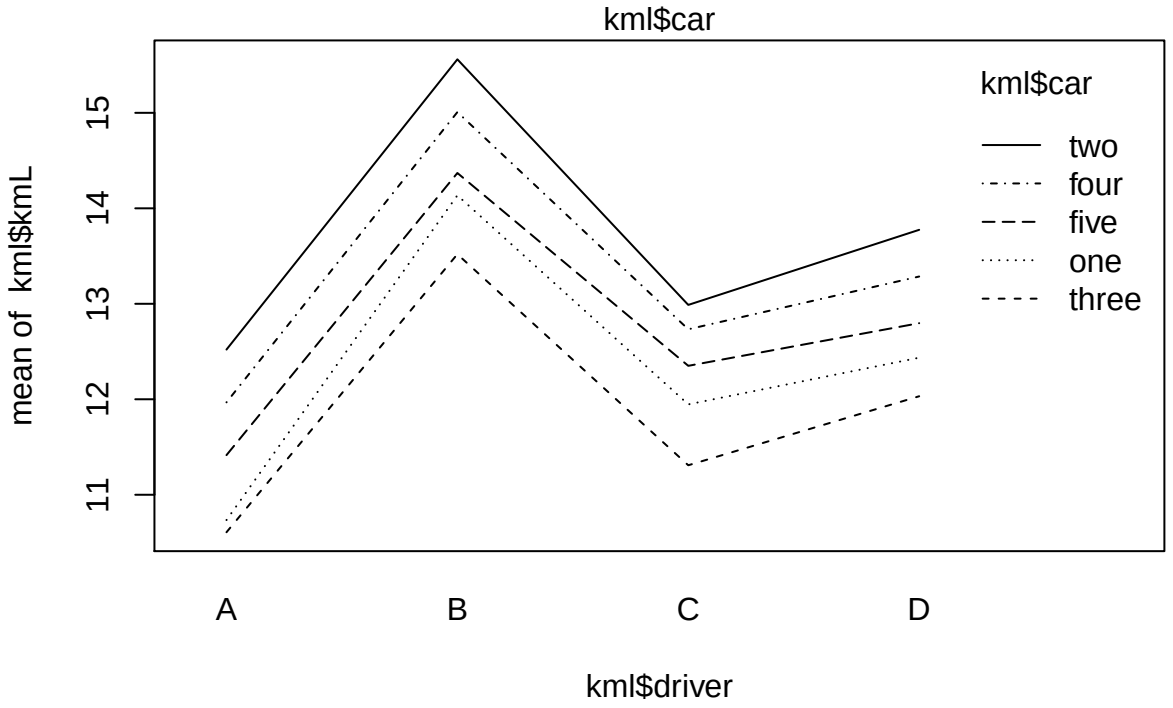
```
##       car
## driver five four one three two
##      A    2    2   2     2   2
##      B    2    2   2     2   2
##      C    2    2   2     2   2
##      D    2    2   2     2   2
```

design is balanced we can see from looking at the table there is an equal amount of variables.



variances do not look to be equal, and the averages have a large variance between variables.





The lines are parallel indicating a lack of interaction occurring between variables.

anova two way model

$$Yijk \ = \mu + \alpha_i + \beta_j + \gamma_i j + \epsilon_i jk$$

defining parameters

$$ij = interaction\ of\ driver\ and\ car$$
$$i = driver$$
$$j = car$$

Three tests to be conducted

The interaction

$$H\ 0\ : \gamma_i j = 0 H\ A\ : \gamma_i j \neq 0$$

Main effect of driver

$$H\ 0\ : \alpha_i = 0 H\ A\ : \alpha_i \neq 0$$

main effect of car

$$H\ 0\ : \beta_j = 0 H\ A\ : \beta_j \neq 0$$

```
anova(kmlanova)
```

```
## Analysis of Variance Table
##
## Response: kmL
##            Df Sum Sq Mean Sq F value    Pr(>F)
## driver      3 50.661 16.8869  531.60 < 2.2e-16 ***
## car         4 17.119  4.2798  134.73 3.664e-14 ***
## driver:car 12  0.442  0.0368    1.16    0.3715
## Residuals  20  0.635  0.0318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

interaction effect

$$model = \mu + \alpha_i + \beta_j + \gamma_i j + \epsilon$$

$$Hypotheses : H\ 0\ : \gamma_i j = 0 H\ A\ : \gamma_i j \neq 0$$

P-value $= 0.3715 > 0.05$

interaction is not significant because p value is not significant

```
## Analysis of Variance Table
##
## Response: kmL
##           Df Sum Sq Mean Sq F value    Pr(>F)
## driver     3 50.661 16.8869   501.5 < 2.2e-16 ***
## car        4 17.119  4.2798   127.1 < 2.2e-16 ***
## Residuals 32  1.078  0.0337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Driver effect

$$model = \mu + \alpha_i + \beta_j + \gamma_i j$$

$$HypotheseH\ 0\ : \alpha_i = 0 H\ A\ : \alpha_i \neq 0$$

P value $= 0.00000000000000022 > 0.05$

driver type is significant Car effect

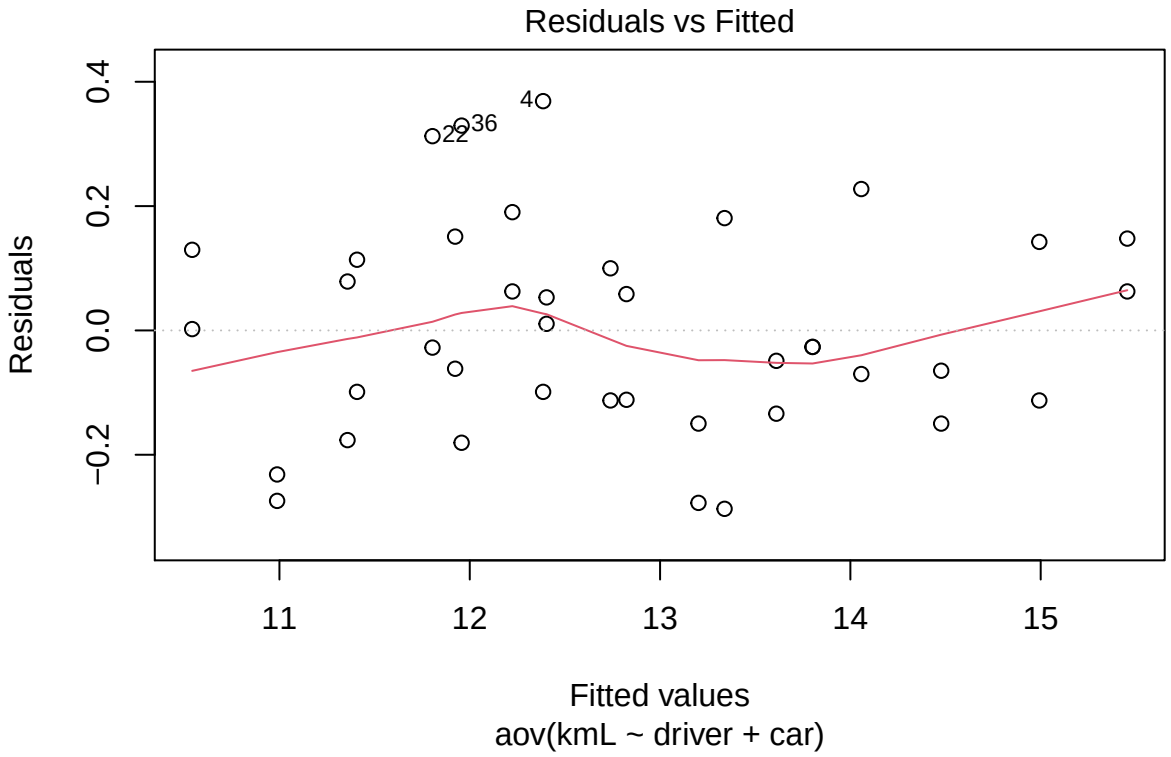
$$model = \mu + \alpha_i + \beta_j + \gamma_i j$$

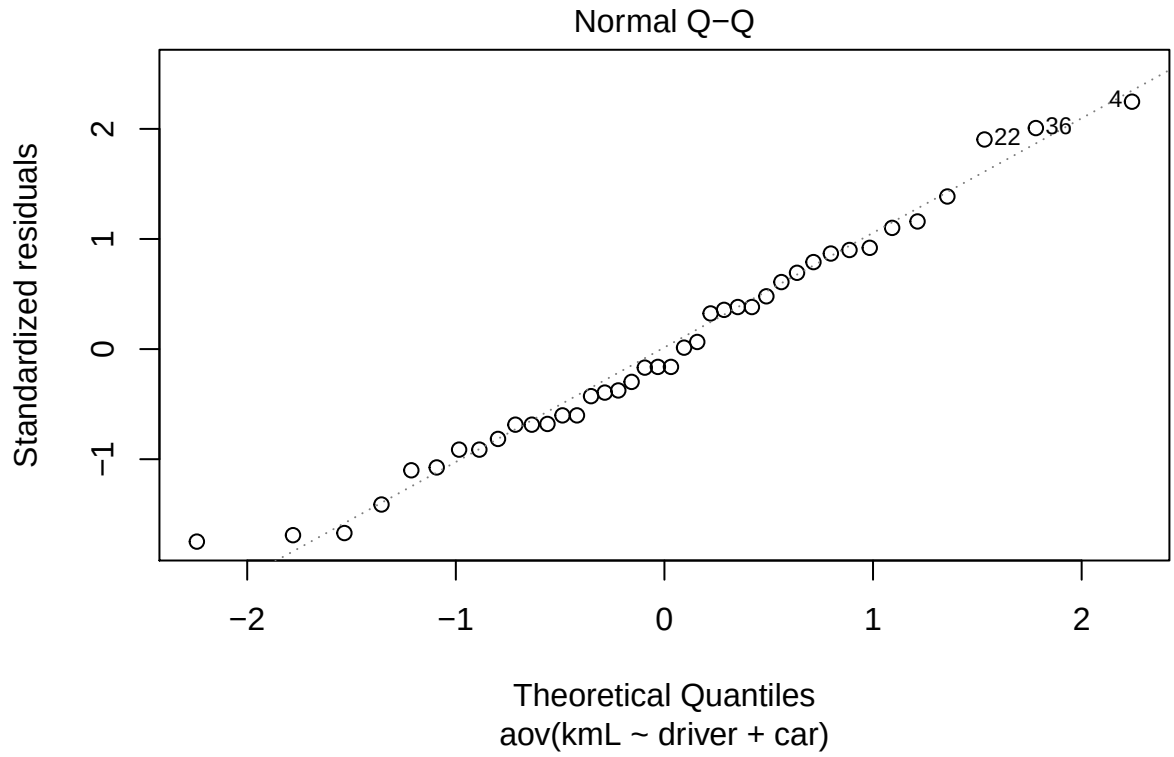
$$Hypothese H\ 0\ : \beta_j = 0 H\ A\ : \beta_j \neq 0$$

p value = 0.00000000000000022 > 0.05

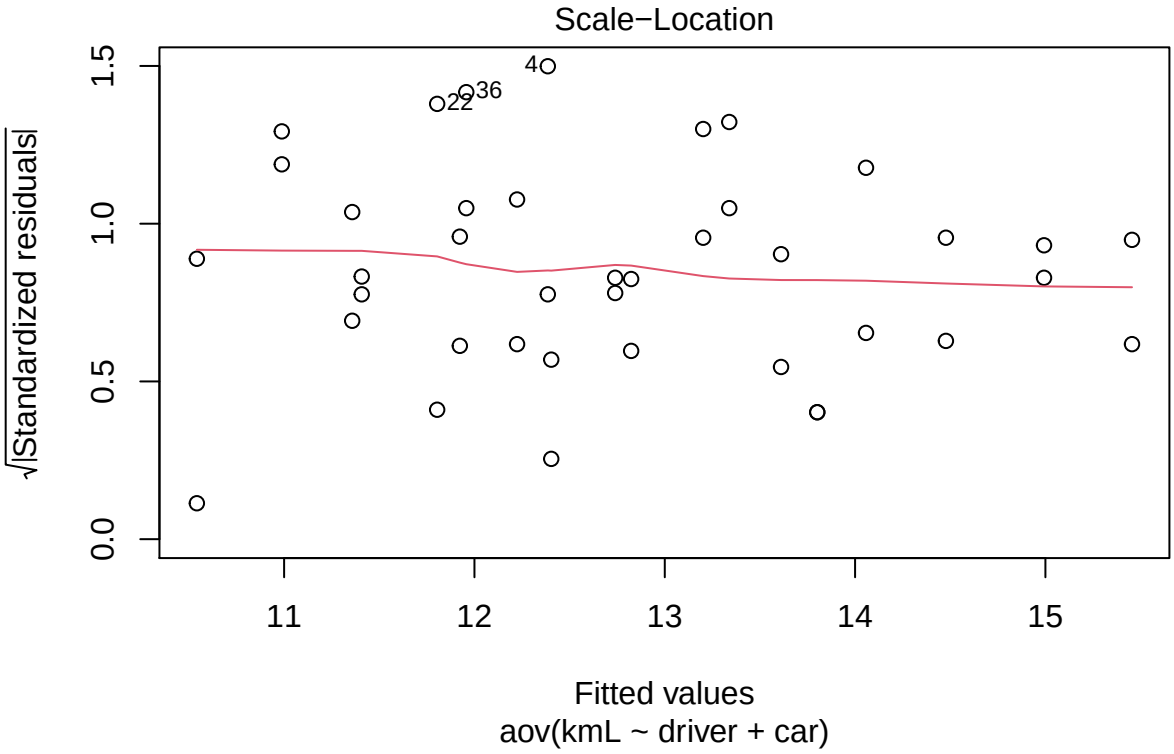
car type is significant

```
plot(kmlanova.1)
```






```
## hat values (leverages) are all = 0.2
##  and there are no factor predictors; no plot no. 5
```



residuals look to be normally distributed. assumption of normal distribution looks to be valid.

p value is insignificant assumption not valid.

conclusion The effect of driver and car on fuel efficiency insignificant however individual interactions of car and drive separately on fuel efficiency are significant.

The interaction plot shows that there is no strong interaction between variables.

would need to reject alternate hypothesis for overall effect of car and driver on fuel efficiency.