Stat2170 Assignment Joel Gregory 46425187

```
surg = read.table("data/surg.dat", header = TRUE)
```

Question 1

A medical research team wants to investigate the survival time of patients that have a particular type of liver operation as part of their treatment.

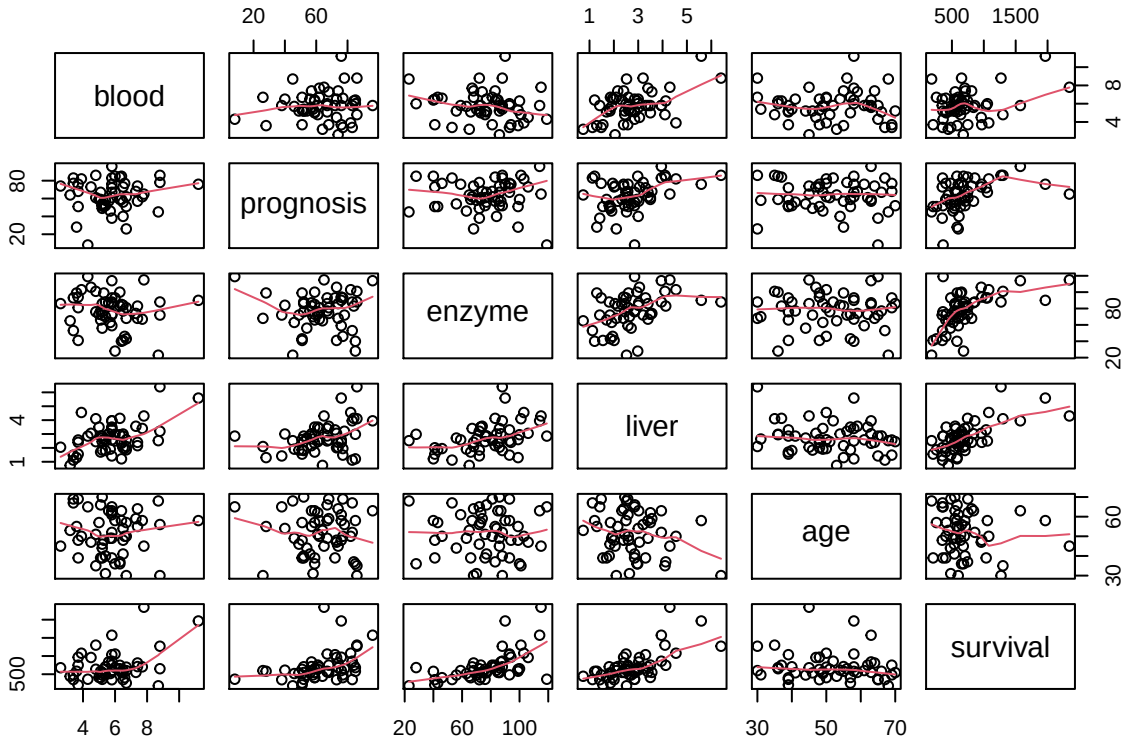
A) Adjusting the data

You will need to remove the gender variable to do this.

```
surgnew= surg[-6]
summary(surgnew)
```

```
##      blood          prognosis         enzyme           liver
##  Min.   : 2.600   Min.   : 8.00   Min.   : 23.00   Min.   :0.740
##  1st Qu.: 5.025   1st Qu.:52.50   1st Qu.: 67.25   1st Qu.:2.020
##  Median : 5.800   Median :63.00   Median : 79.00   Median :2.595
##  Mean   : 5.783   Mean   :63.24   Mean   : 77.11   Mean   :2.744
##  3rd Qu.: 6.500   3rd Qu.:76.00   3rd Qu.: 89.50   3rd Qu.:3.275
##  Max.   :11.200   Max.   :96.00   Max.   :119.00   Max.   :6.400
##       age           survival
##  Min.   :30.00   Min.   : 181.0
##  1st Qu.:44.25   1st Qu.: 482.0
##  Median :51.50   Median : 605.5
##  Mean   :51.61   Mean   : 702.1
##  3rd Qu.:60.50   3rd Qu.: 750.5
##  Max.   :70.00   Max.   :2343.0
```

```
plot(surgnew,panel=panel.smooth)
```



For proper analysis of correlation matrix we need to remove the categorical variable (gender). We do this as observing the relationship gender has on each variable independently does not benefit the model.

From the plot we can see that there is some relationship between variables, all except age had linear relationship with survival. Definitively no strong linear relationship. A lot of horizontal lines that represent no relationships between variables

B) The Correlation matrix

```
cor(surgnew)
```

```
##                blood   prognosis      enzyme      liver         age   survival
## blood      1.00000000  0.09011973 -0.14963411  0.5024157 -0.02068803  0.3465497
## prognosis  0.09011973  1.00000000 -0.02360544  0.3690256 -0.04766570  0.4204810
## enzyme    -0.14963411 -0.02360544  1.00000000  0.4164245 -0.01290325  0.5782260
## liver      0.50241567  0.36902563  0.41642451  1.0000000 -0.20737776  0.6741950
## age       -0.02068803 -0.04766570 -0.01290325 -0.2073778  1.00000000 -0.1191715
## survival   0.34654968  0.42048097  0.57822600  0.6741950 -0.11917146  1.0000000
```

Generally speaking from analyzing the correlation matrix we can see there is an even mix between variables of positive and negative. One outlier to show is age. in all cases it is compared it is negatively correlated to its partnered variable

C) fitting of the regression model to predict survival

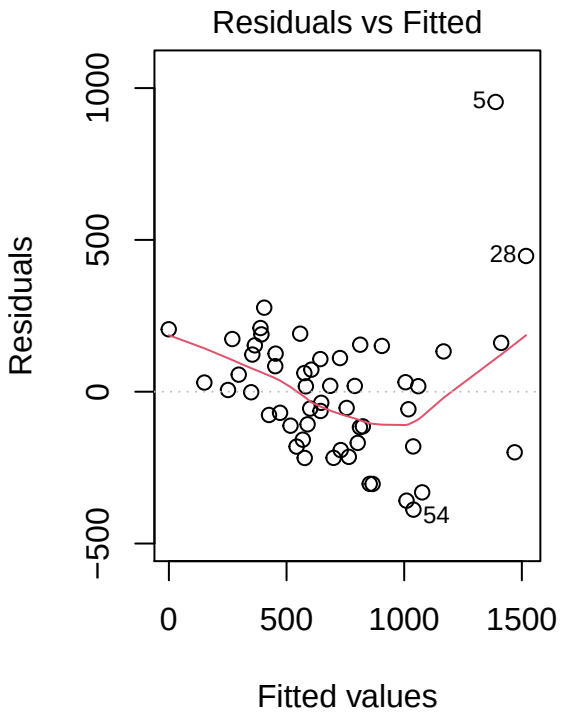
fitting the model:

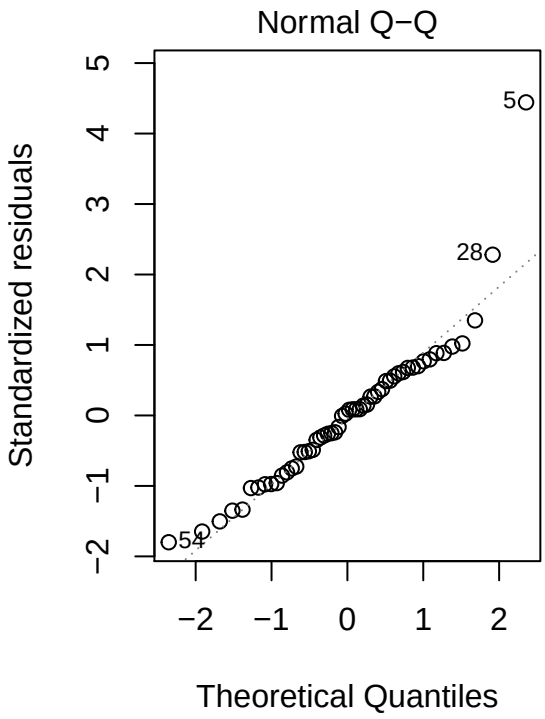
```
lm_surg = lm(survival~ age + liver + enzyme + prognosis + blood, data = surgnew)
summary(lm_surg)
```

```
##
## Call:
## lm(formula = survival ~ age + liver + enzyme + prognosis + blood,
##     data = surgnew)
##
## 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 ***
## age            -2.340      2.969  -0.788 0.434514
## liver          38.554     49.251   0.783 0.437595
## enzyme         11.124      1.958   5.683 7.62e-07 ***
## prognosis       8.501      2.137   3.978 0.000234 ***
## blood          86.630     26.905   3.220 0.002302 **
## ---
## 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
```

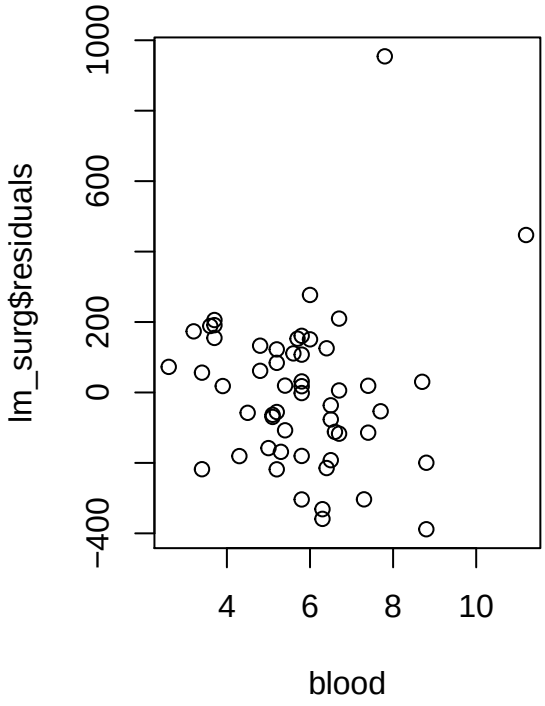
checking the Assumptions:

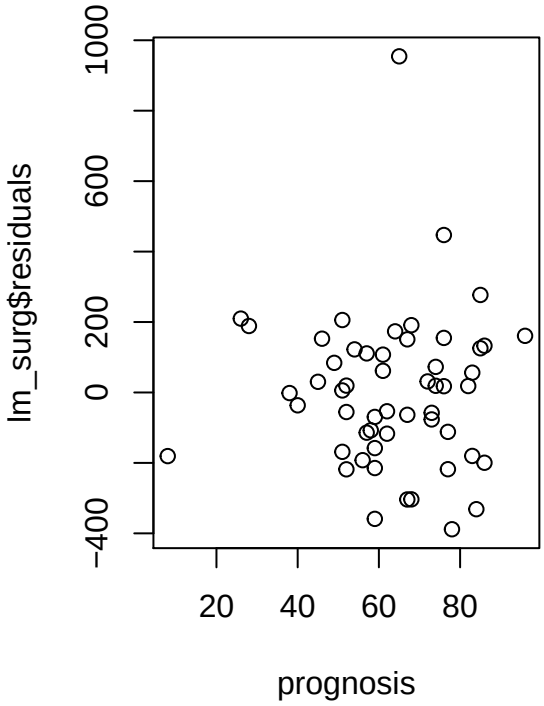
```
par(mfrow=c(1,2))
plot(lm_surg,which = 1:2)
```

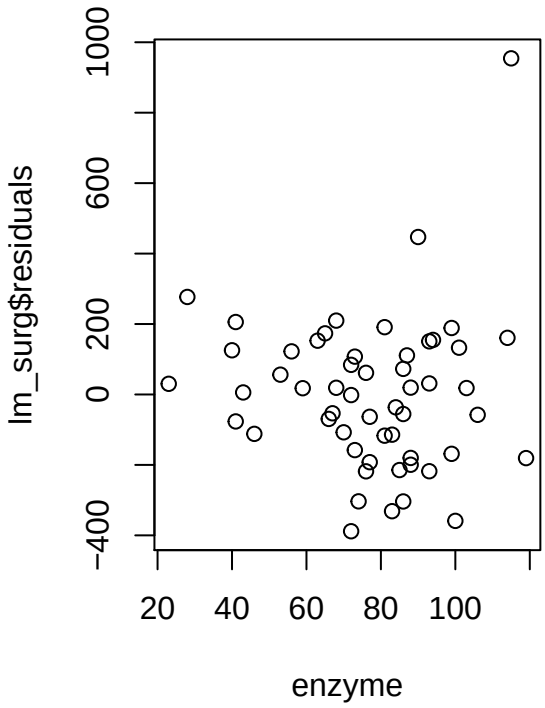


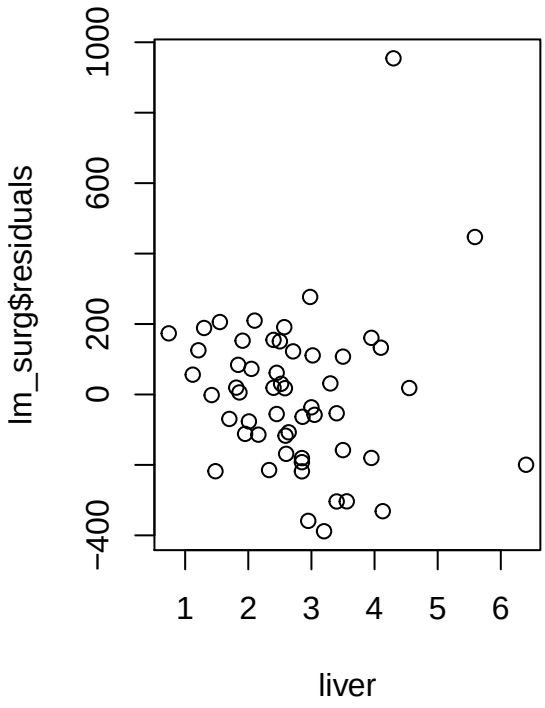


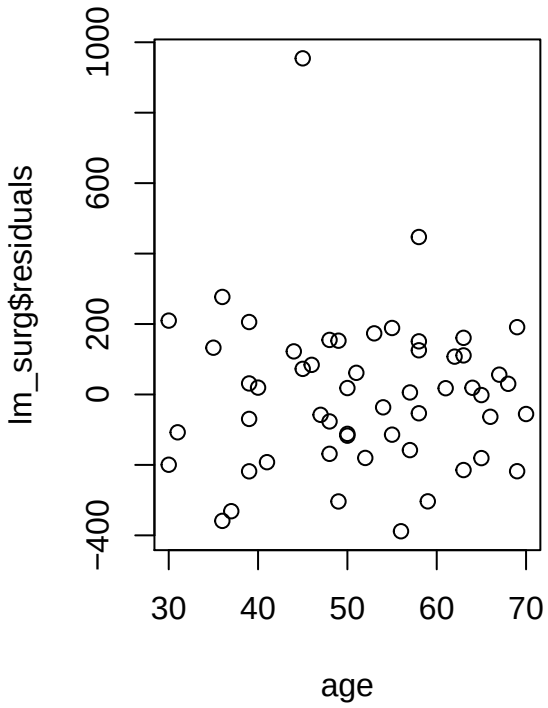

```
plot(lm_surg$residuals~ blood + prognosis + enzyme + liver + age,data = surgnew)
```


lm_surg$residuals
1000
600
200
0
−400
4
6
8
10
blood


lm_surg$residuals
1000
600
200
0
−400
20
40
60
80
prognosis


lm_surg$residuals
1000
600
200
0
−400
20
40
60
80
100
enzyme


lm_surg$residuals
1000
600
200
0
−400
1
2
3
4
5
6
liver



Assumptions not exactly met: * Evident curve in variance suggest non linear relationship. * QQ plot is mostly linear, However data portrays an outlier * signs pattern/ clustering thus linear assumption is questionable

The mathematical multiple regression model for this situation

predicted: survival = -1179.367 + (-2.340 x age) + (38.554 x Liver) (11.124 x enzyme) + (8.501 x prognosis) + (86.630 x blood)

Hypotheses

Null hypothesis: variables are all statistically significant in predicting survival

Alternative Hypothesis: They are not statistical indicators of survival

Analysis of varience

Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient).

```
surgaov= aov(survival~blood + prognosis + enzyme + liver + age,data = surgnew)
anova(surgaov)
```

```
## 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 seem to be insignificant to the model

Full model regression = 5 816 714

F statistic for this test = 4.04

Computing the p value:

```
1-pf(4.04,1,48)
```

```
## [1] 0.05007225
```

p value = 0.05, insignificant

#conclusion: statisically/contextually

Analysis Statistically shows the predictors as just insignificant in predicting survival and fail to pass the regression assumptions. contextually this is counter intuitive as something like **age** for example has historically been a significant indicator of surgery survival rate. This suggests that the data given is inadequate or is an anomaly.

D) Using model selection procedures discussed in the course, find the best multiple regression model that explains the data.

from the summary we can see that two are insignificant variables (liver and age) we remove them from the data as they do not statistically benefit to the model.

We can also remove the outlier that is evident in the qqnorm and seems to skew the data

```
lm_surgnew = lm(survival~ enzyme + prognosis + blood, data = surgnew)
summary(lm_surgnew)
```

```
##
## Call:
## lm(formula = survival ~ enzyme + prognosis + blood, data = surgnew)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -432.4 -134.3  -19.1  111.9  961.1
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1410.847    209.118  -6.747 1.50e-08 ***
## enzyme         12.128      1.503   8.069 1.30e-10 ***
## prognosis       9.382      1.876   5.000 7.43e-06 ***
## blood         101.054     20.005   5.052 6.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.9 on 50 degrees of freedom
## Multiple R-squared:  0.6841, Adjusted R-squared:  0.6652
## F-statistic:  36.1 on 3 and 50 DF,  p-value: 1.469e-12
```

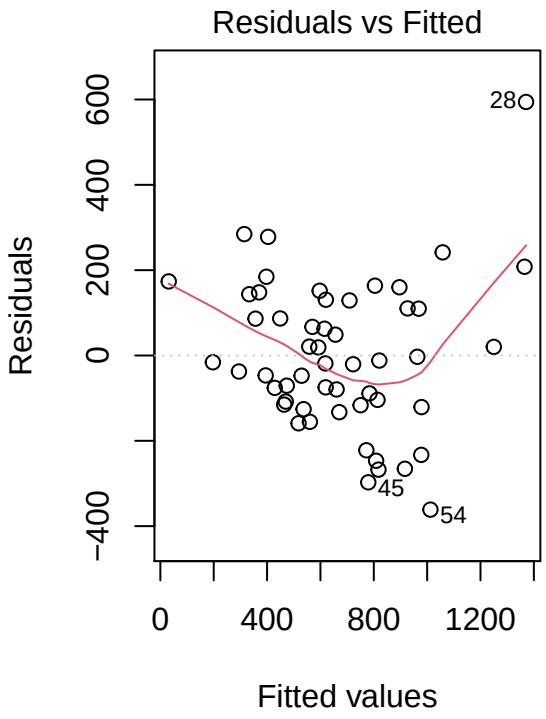
Removal of the outlier

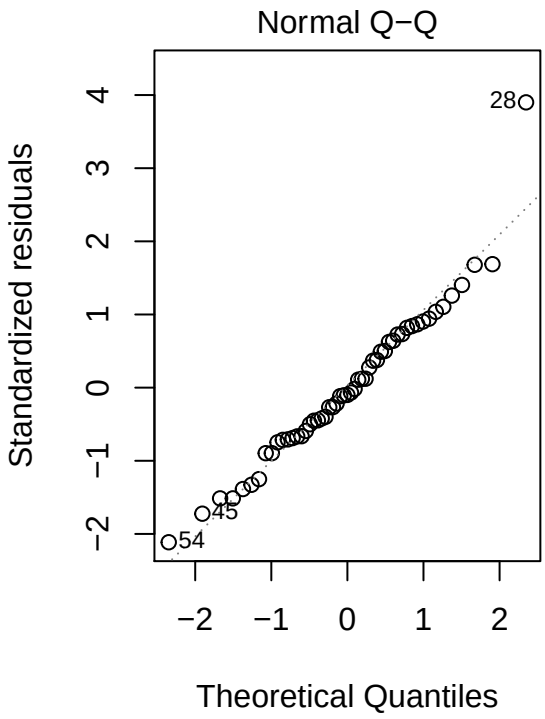
```
outlier = which(lm_surgnew$residuals==max(lm_surgnew$residuals))
surgnew = surgnew[-outlier,]

lm_surgnew = lm(survival~ enzyme + prognosis + blood, data = surgnew)
summary(lm_surgnew)
```

```
##
## Call:
## lm(formula = survival ~ enzyme + prognosis + blood, data = surgnew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -361.27 -115.14  -15.91  128.99  594.23
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1162.643    169.513  -6.859 1.10e-08 ***
## enzyme         10.166      1.227   8.286 7.00e-11 ***
## prognosis       9.368      1.470   6.373 6.21e-08 ***
## blood          80.940     16.064   5.039 6.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 180.1 on 49 degrees of freedom
## Multiple R-squared:  0.7174, Adjusted R-squared:  0.7001
## F-statistic: 41.46 on 3 and 49 DF,  p-value: 1.733e-13
```

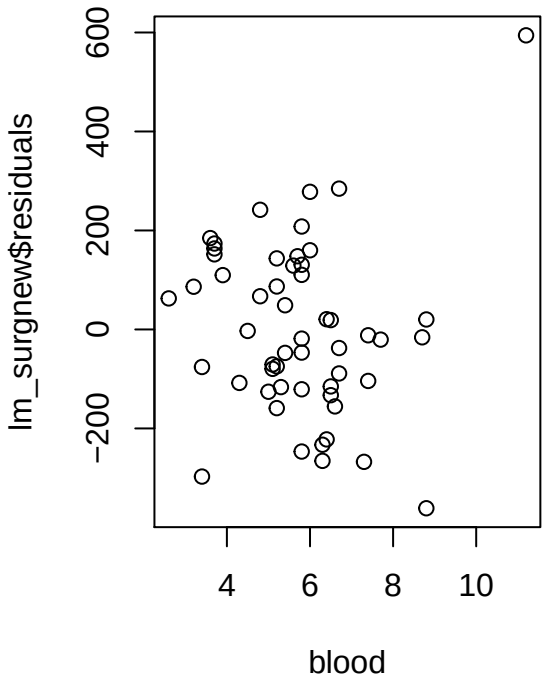
Selection process: removing the liver and age variable: make all the other significant indicators of survival.

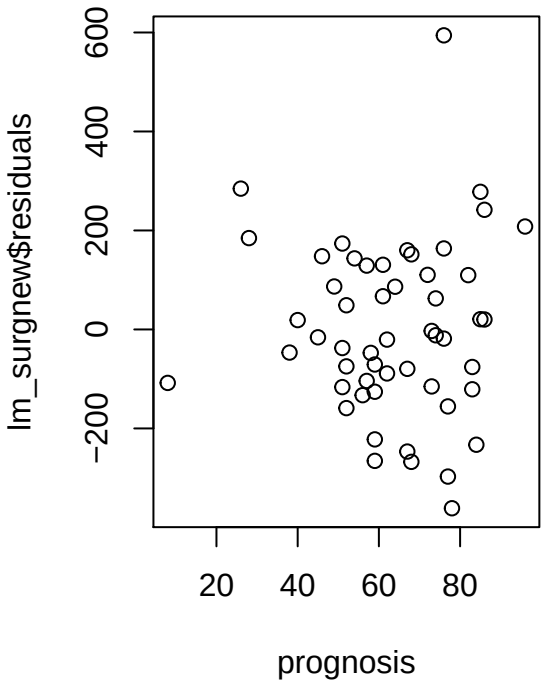
```
par(mfrow=c(1,2))
plot(lm_surgnew,which = 1:2)
```

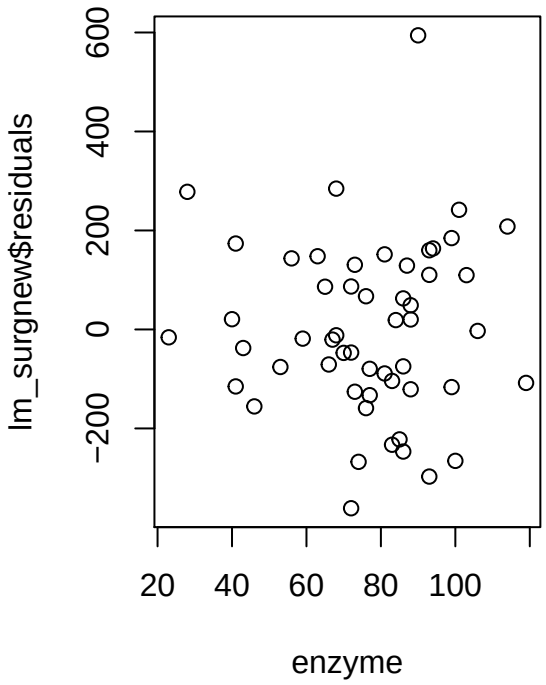





```
plot(lm_surgnew$residuals~ blood + prognosis + enzyme ,data = surgnew)
```


lm_surgnew$residuals
−200
0
200
400
600
4
6
8
10
blood


lm_surgnew$residuals
−200
0
200
400
600
20
40
60
80
prognosis



#Validate your final model and comment why it is not appropriate to use the multiple regression model to explain the survival time

Although the data is statistically significant the models to not pass the tests. Correlation does not always mean causation:

- Between variables plot of the data shows very minimal if any relationships
- Linearity test shows aspects of grouping
- Normality and variance test shows an obvious outlier
- Evident curve and pattern in the model

The Assumptions are hardly passed. This leads us to believe that this data is stretching to be as definitive causation.

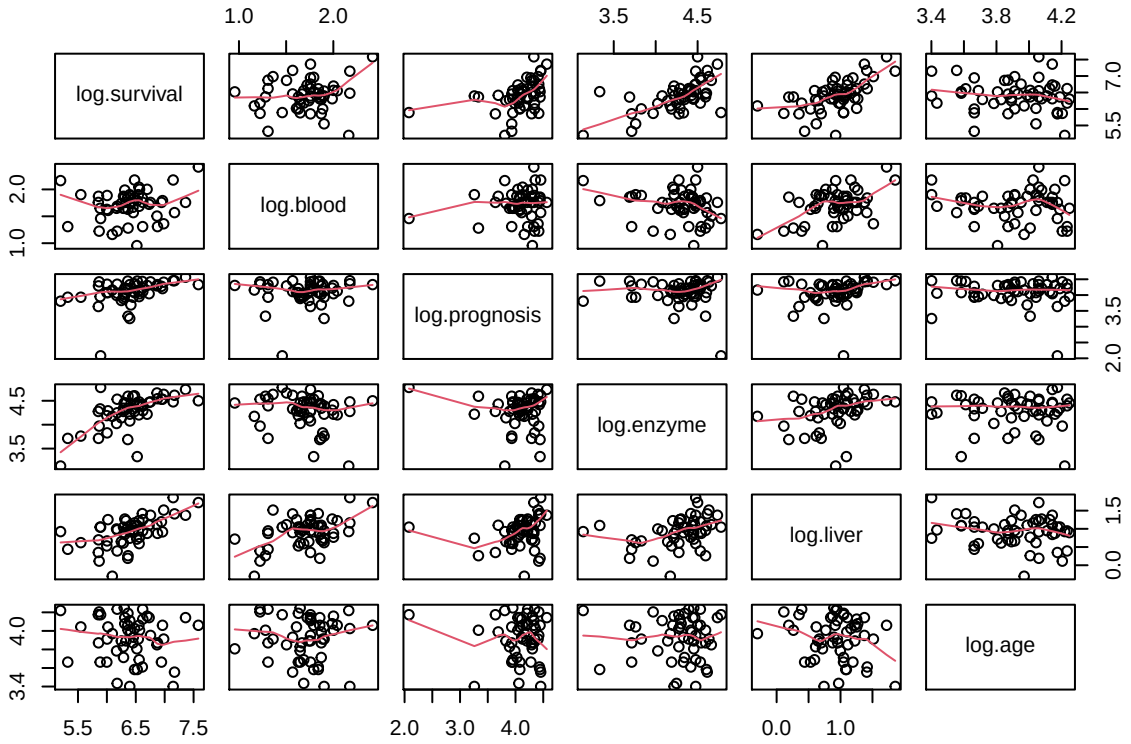
#Re-fit the model using log(survival) as the new response variable

```
logsurg = data.frame(log.survival= log(surgnew$survival), log.blood= log(surgnew$blood),log.prognosis= l

head(logsurg)
```

```
##   log.survival log.blood log.prognosis log.enzyme log.liver  log.age
## 1     6.543912  1.902108      4.127134   4.394449 0.9516579 3.912023
## 2     5.998937  1.629241      4.077537   4.189655 0.5306283 3.663562
## 3     6.565265  2.001480      4.043051   4.418841 0.7701082 4.007333
## 4     5.855072  1.871802      4.290459   3.713572 0.6981347 3.871201
```

```
## 5     5.852202  1.757858      3.637586   4.276666 0.3506569 4.174387
## 6     6.249975  1.740466      3.828641   4.143135 0.6471032 3.891820
```

```
plot(logsurg,panel=panel.smooth)
```



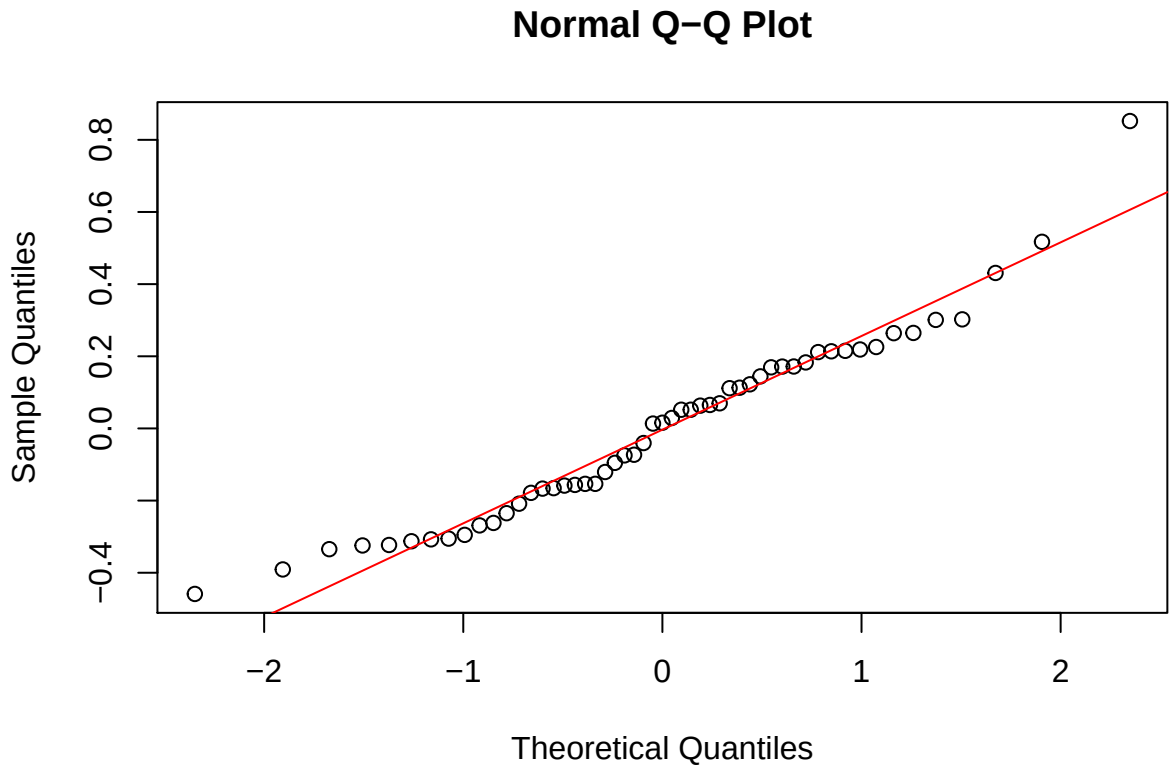

```
log_lm_surg = lm(log.survival ~ log.blood+ log.prognosis+ log.enzyme,data = logsurg)

summary(log_lm_surg)
```

```
##
## Call:
## lm(formula = log.survival ~ log.blood + log.prognosis + log.enzyme,
##     data = logsurg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45876 -0.17854  0.01583  0.17168  0.85207
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.91966    0.71497  -1.286  0.20439
## log.blood      0.45066    0.13380   3.368  0.00148 **
## log.prognosis  0.54861    0.09556   5.741 5.85e-07 ***
## log.enzyme     1.00469    0.11371   8.835 1.04e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2656 on 49 degrees of freedom
## Multiple R-squared:  0.686,  Adjusted R-squared:  0.6668
## F-statistic: 35.69 on 3 and 49 DF,  p-value: 2.238e-12
```

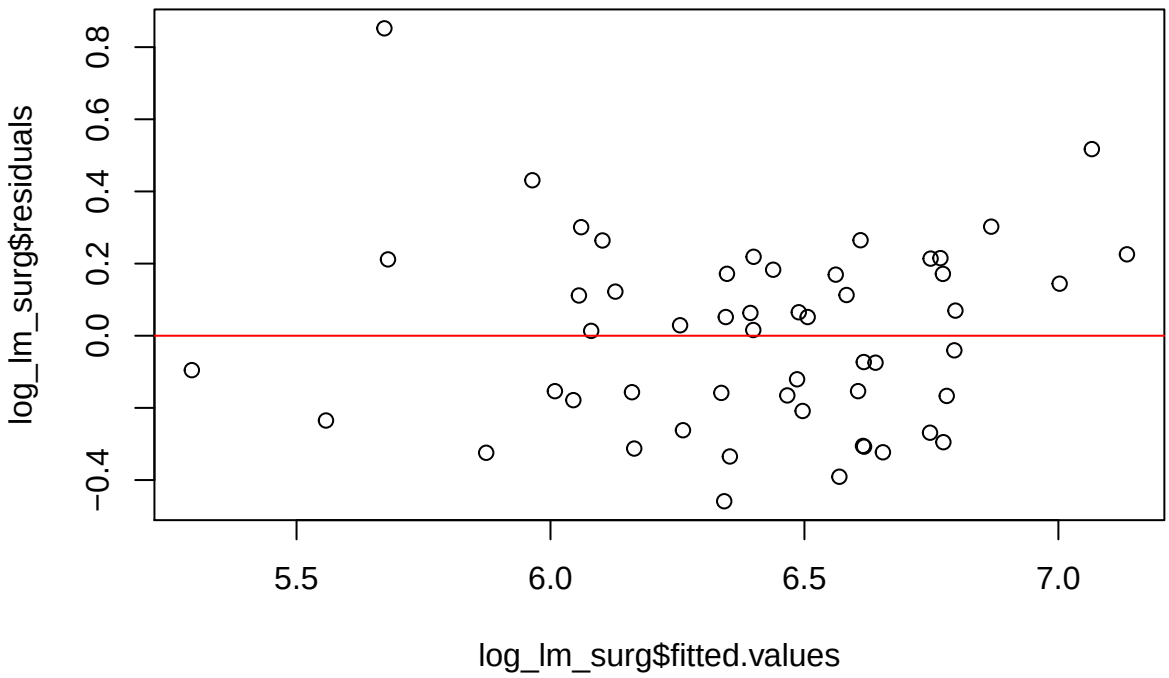
re-evaluation of assumptions with log transformations: Normality: Strong relationship

```
qqnorm(log_lm_surg$residuals)
qqline(log_lm_surg$residuals,col= "red")
```



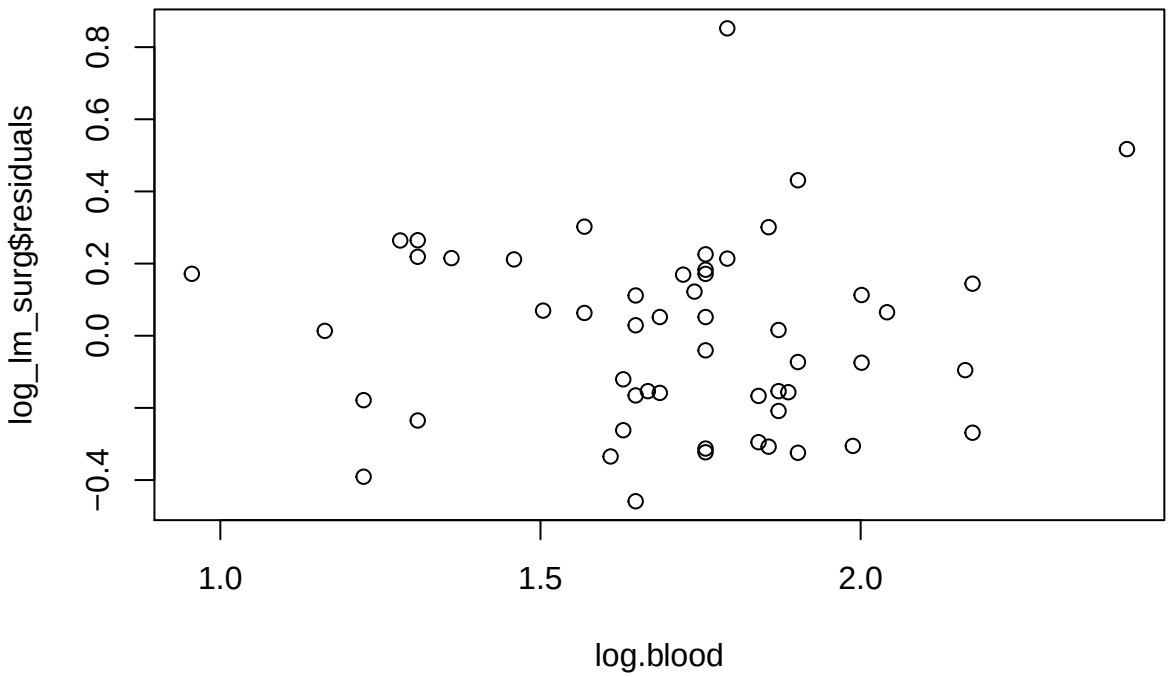
Variance: Even spread above and below

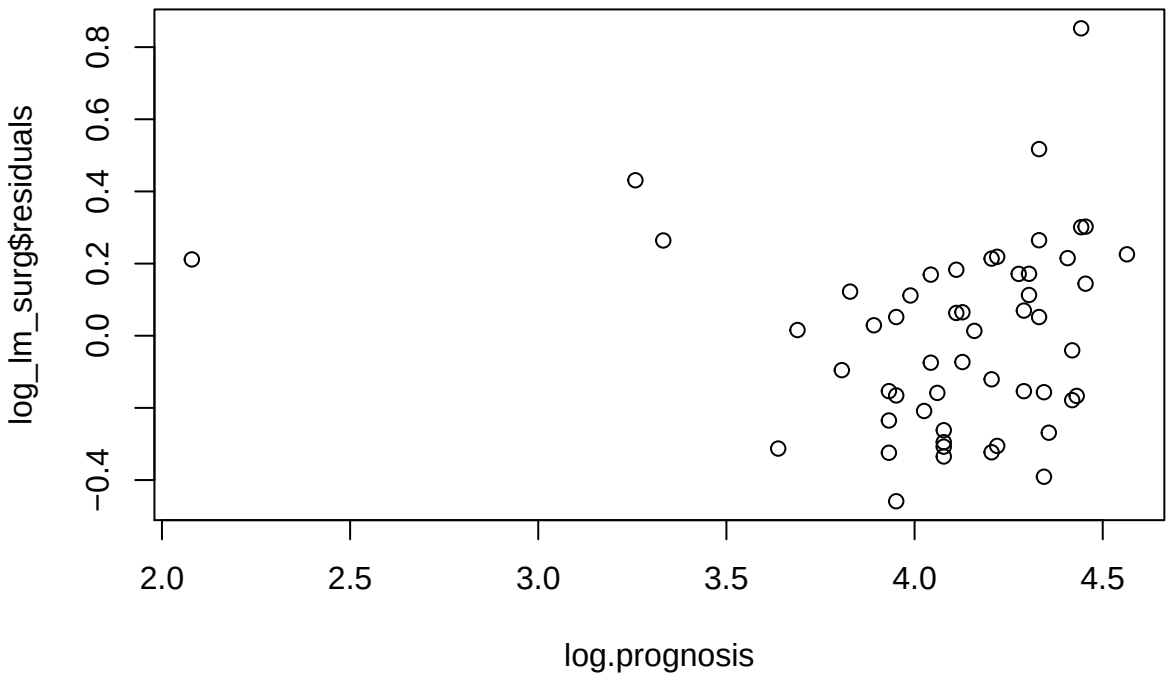
```
plot(log_lm_surg$fitted.values,log_lm_surg$residuals)
abline(h=0,col="red")
```

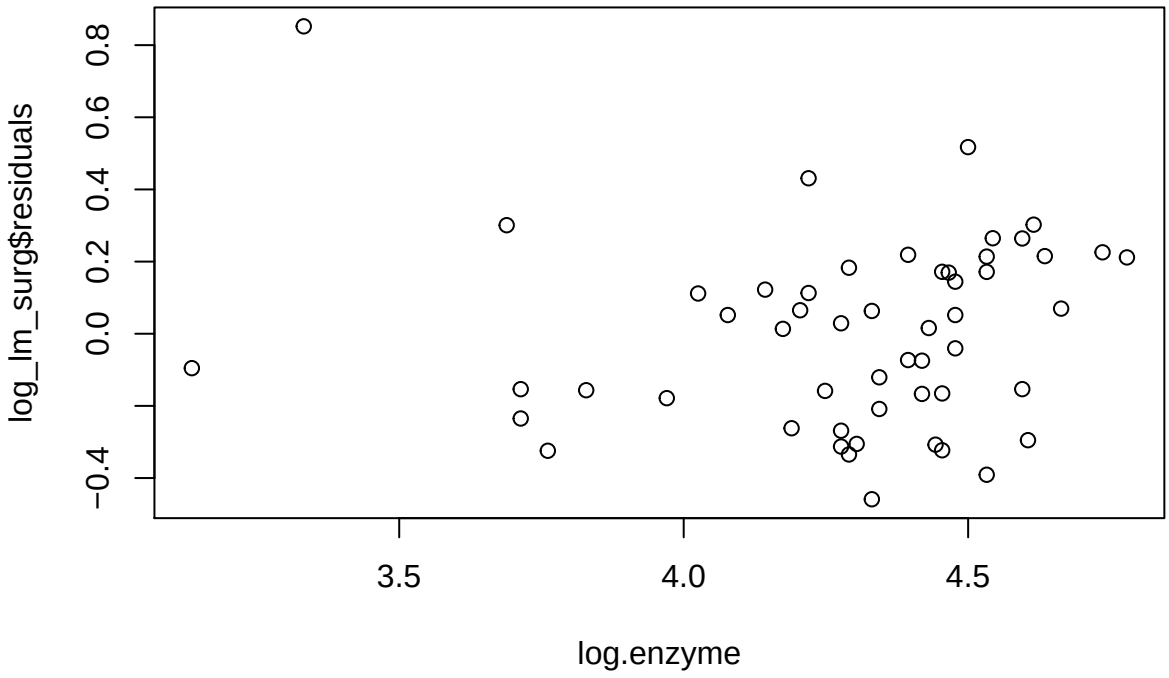


Linearity:

```
plot(log_lm_surg$residuals~ log.blood + log.prognosis + log.enzyme ,data = logsurg)
```


log_lm_surg$residuals
0.8
0.6
0.4
0.2
0.0
−0.4
1.0
1.5
2.0
log.blood


log_lm_surg$residuals
0.8
0.6
0.4
0.2
0.0
−0.4
2.0
2.5
3.0
3.5
4.0
4.5
log.prognosis



G.) validate my model: Explain why the regression model with log(survival) response variable is superior to the model with the survival response variable

The log transforms the data to fit the respected assumptions better.

- Normality assumption makes the outliers less profound and fits the data closer to the line.
- Variance: there is less clustering to one side no signs of grouping
- linearity: no evident sign of curve, signs of being linear

Question 2:

A car manufacturer wants to study the fuel efficiency of a new car engine. It wishes to account for any differences between the driver and production variation. The manufacturer randomly selects 5 cars from the production line and recruits 4 different test drivers.

```
kml= read.table("data/kml.dat", header = TRUE)
```

##A.) Design of the data

```
with(kml,table(car,driver))
```

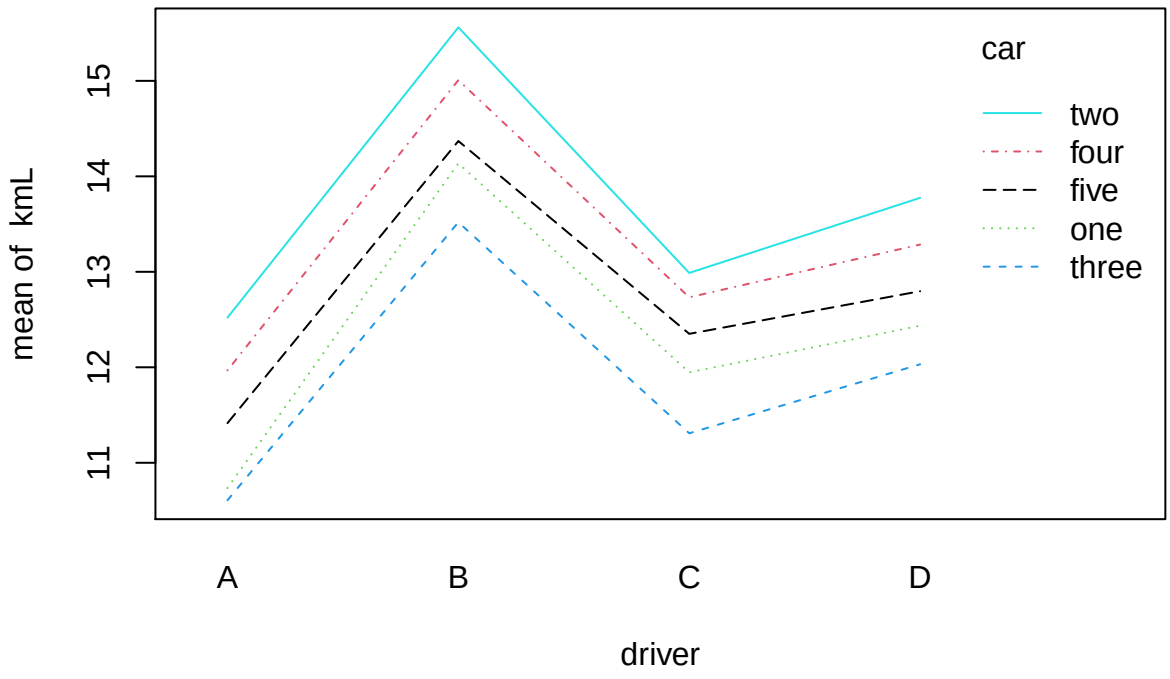
```
##        driver
## car     A B C D
##   five  2 2 2 2
##   four  2 2 2 2
##   one   2 2 2 2
##   three 2 2 2 2
##   two   2 2 2 2
```

This is a balanced design, having equal sample size number per category

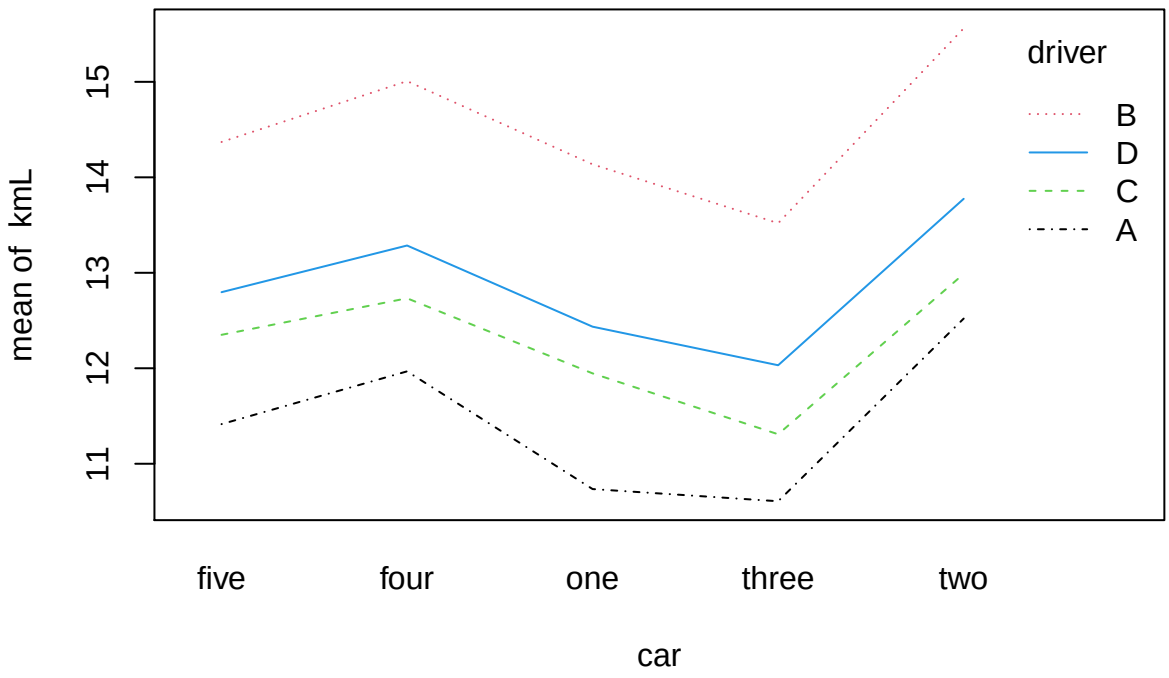
##B.)Preliminary graphs

Plot 1

```
with(kml,interaction.plot(driver,car,kmL,col= 1:5))
```



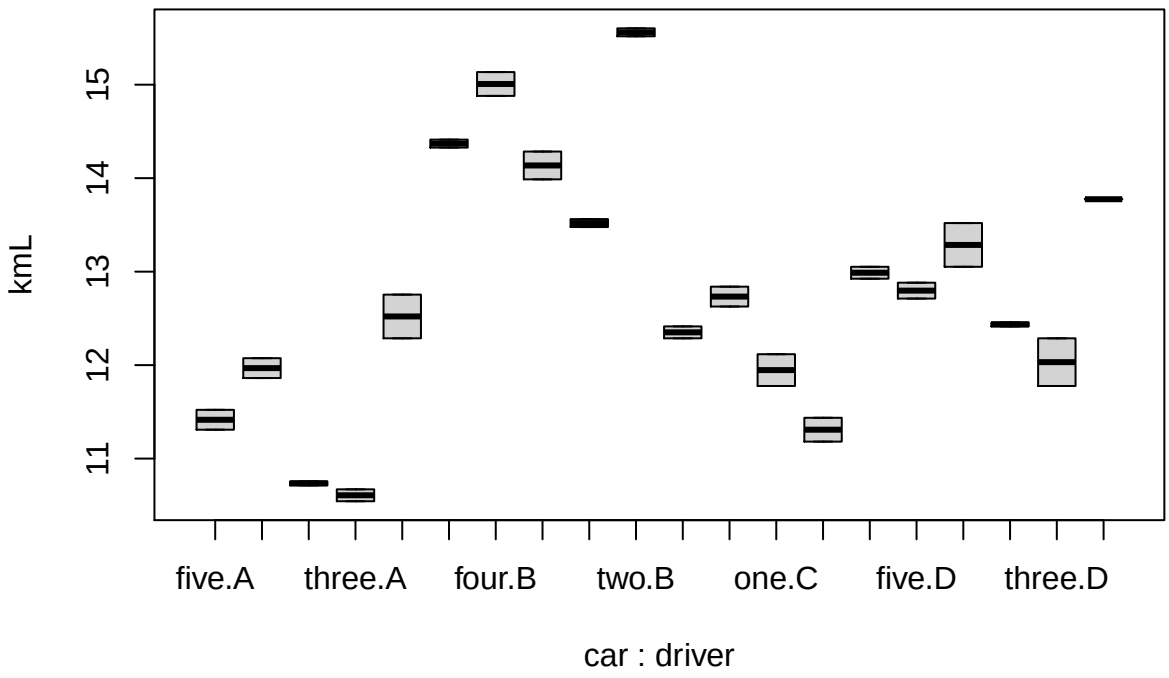

```
with(kml,interaction.plot(car,driver,kmL,col= 1:5))
```



Lines are not parallel, data suggests there could be an interaction

###Plot 2

```
boxplot(kmL~car+driver, data = kml)
```



Box spreads are not equally sized. Driver B has consistently higher scores than other drivers. A pattern emerges and can be seen

##C.)Analyse the data,

H0: kmL is the same for every car randomly pulled off production line H1: is not the same.

```
aov_km = aov(kmL~driver+car,data = kml)
summary(aov_km)
```

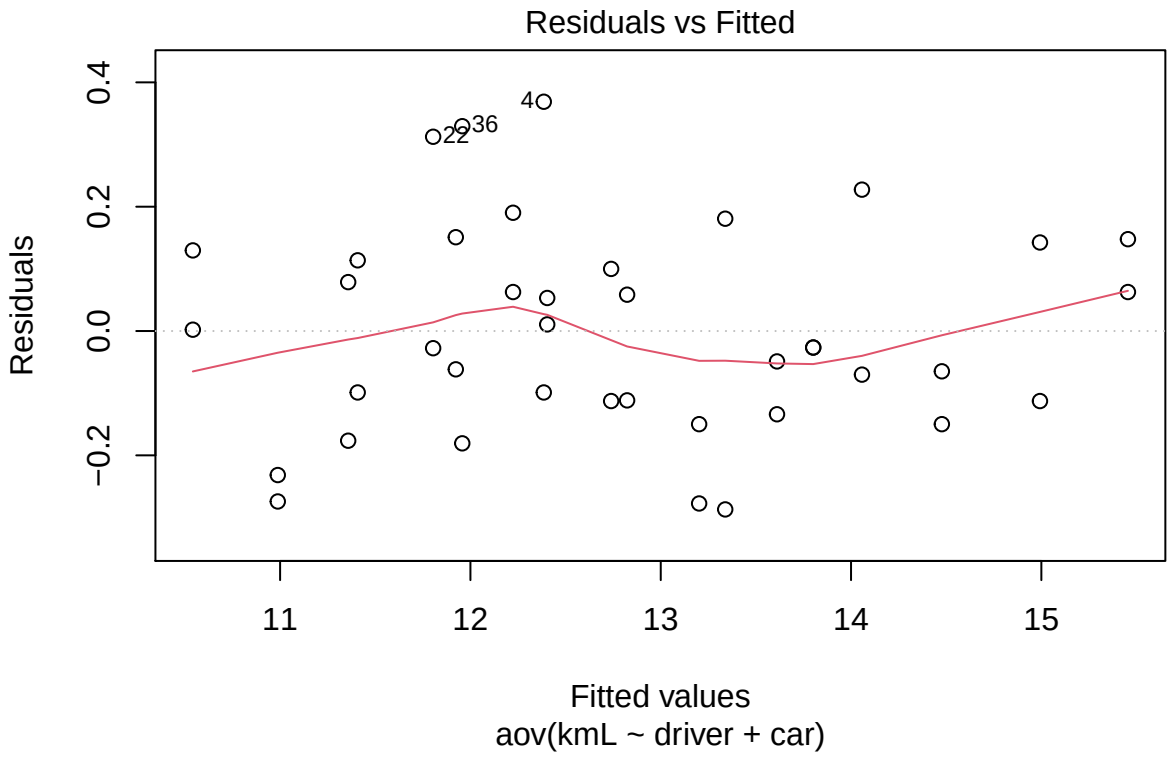
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## driver       3  50.66  16.887   501.5 <2e-16 ***
## car          4  17.12   4.280   127.1 <2e-16 ***
## Residuals   32   1.08   0.034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

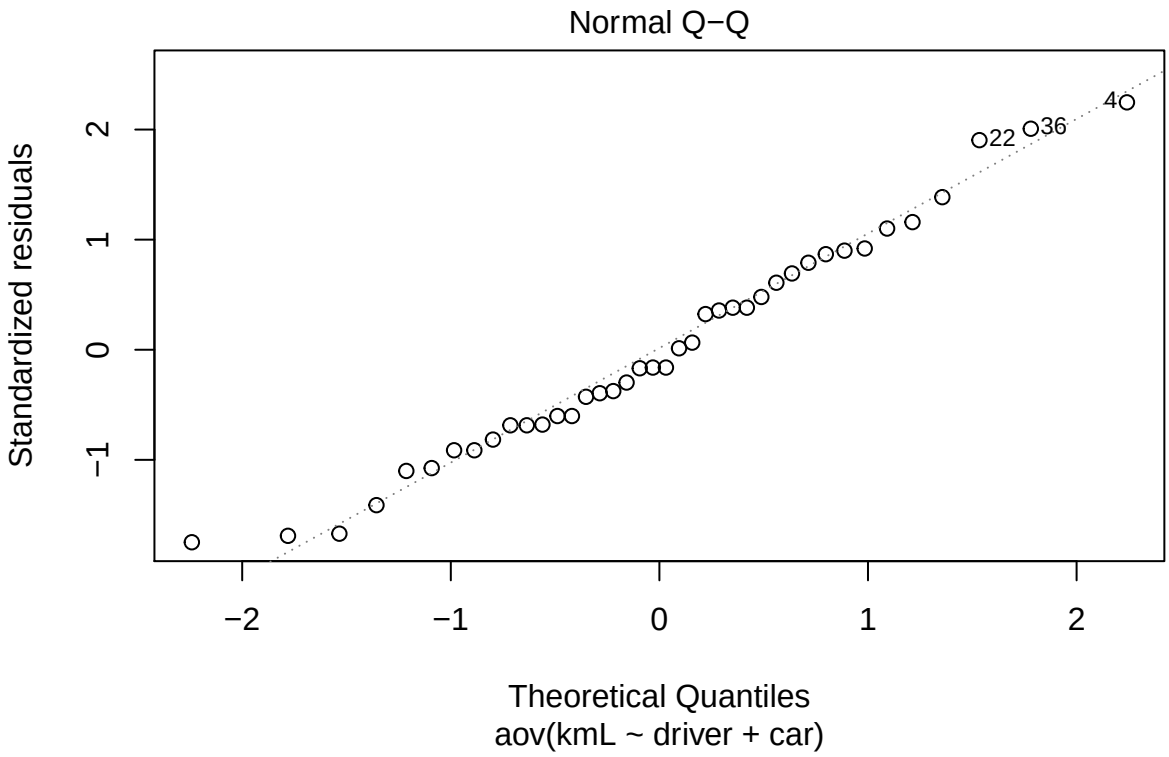
```
aov_km2 = aov(kmL~driver*car,data = kml)
summary(aov_km2)
```

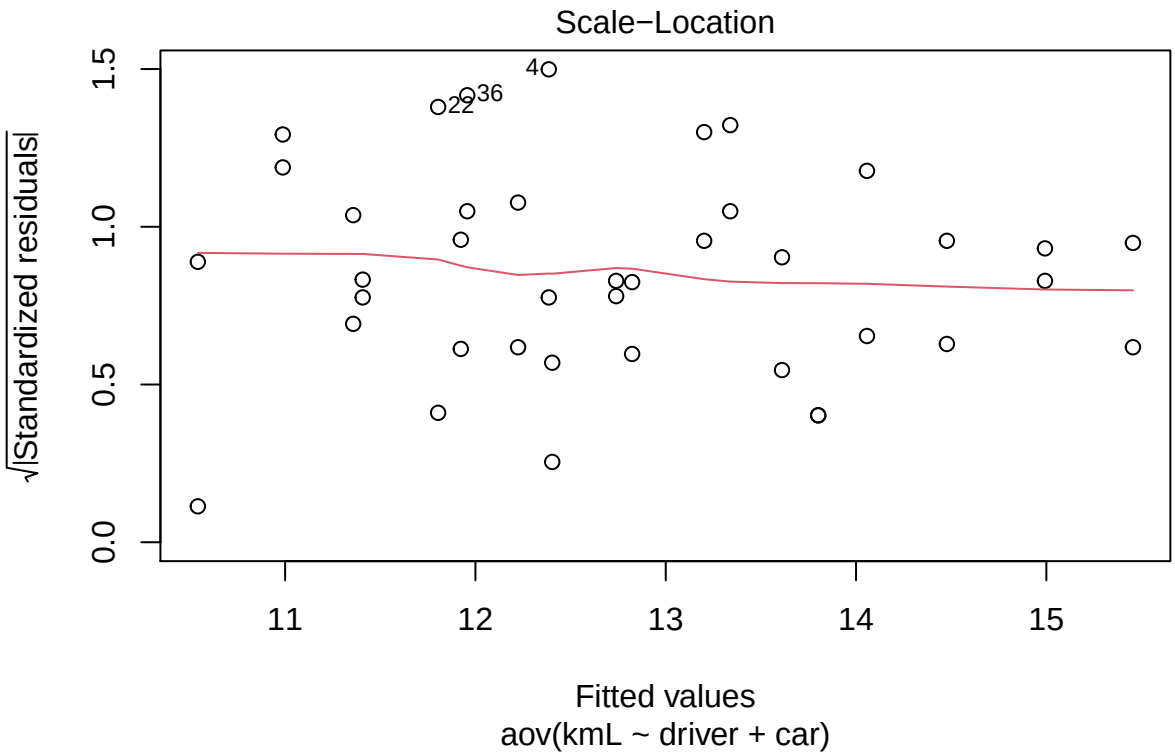
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## driver       3  50.66  16.887  531.60  < 2e-16 ***
## car          4  17.12   4.280  134.73 3.66e-14 ***
## driver:car  12   0.44   0.037    1.16    0.371
## Residuals   20   0.64   0.032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova test shows both variables are statistically significant to the model ,however no significant interaction affect

```
plot(aov_km, which = 1:3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(kmL ~ driver + car)
22
36
4
−2
−1
0
1
2



Assumptions pass: - Residuals vs fitted: About even spread above and below - QQ norm: Strong linear relationship - Scale location: very little clustering

Conclusion

From observation the data seems to suggest that there is a relationship between kmL recorded and the particular driver /car chosen. The drivers kmL placement stayed consistent no matter what car. The car chosen at random also held a consistent placement between drivers

key stats: - driver B always finished with the highest kmL between cars - driver A always finished with the lowest kmL between cars - Car 2 always had the highest kmL - Car 3 always had the lowest kmL