Assignment Dinh Nghi Dung Le 46150641

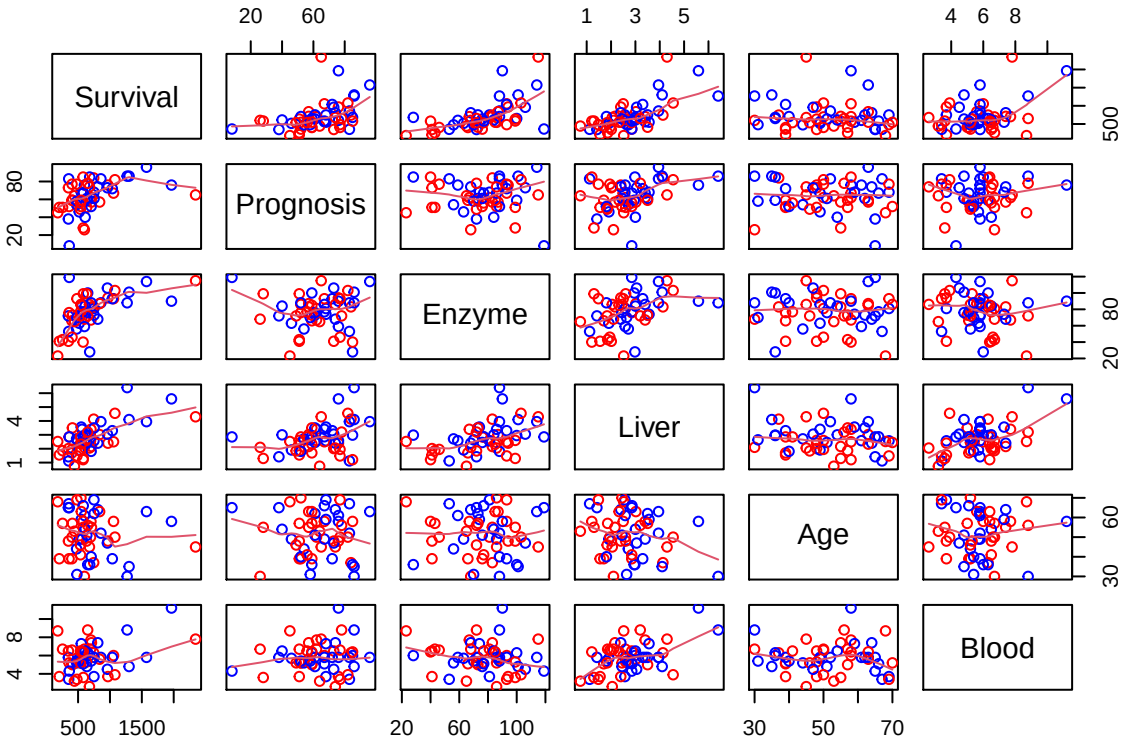
Question 1

```
surg$Blood = surg$blood # Add copy of column blood
surg$blood = surg$survival # Swap column survival to blood
colnames(surg)[1] = "Survival" # Rename column
surg$blood <- NULL # Remove one of the blood column
surg$survival <- NULL # Remove one of the survival column
newNames = c("Survival","Prognosis","Enzyme","Liver","Age","Gender","Blood")
colnames(surg) = newNames # Change column name
```

a) Checking scatter plot matrix

```
mycols = c("blue","red")[as.factor(surg$Gender)] # Put color for gender factor
levels(as.factor(surg$Gender)) # Checking level of gender, so we have blue=F and red=M
```

```
## [1] "F" "M"
```



Gender variable need to be removed because it is a categorical variable. Gender is not a continuous variable. It is an independent variable and has no strength to affect dependent variable.

Since the correlation matrix is numeric summary, the categorical variable should be removed for the correlation matrix computation.

Scatter plot comment: Survival is correlated with Prognosis, Enzyme, Liver, Age and Blood but every single predictor has correlation with other predictors. There is no abnormal observations. The scatterplot spreads similarly.

b) Compute the correlation matrix

```
cor(surg[,-6]) #Compute correlation matrix without gender variable
```

```
##              Survival   Prognosis      Enzyme      Liver         Age       Blood
## Survival   1.0000000  0.42048097  0.57822600  0.6741950 -0.11917146  0.34654968
## Prognosis  0.4204810  1.00000000 -0.02360544  0.3690256 -0.04766570  0.09011973
## Enzyme     0.5782260 -0.02360544  1.00000000  0.4164245 -0.01290325 -0.14963411
## Liver      0.6741950  0.36902563  0.41642451  1.0000000 -0.20737776  0.50241567
## Age       -0.1191715 -0.04766570 -0.01290325 -0.2073778  1.00000000 -0.02068803
## Blood      0.3465497  0.09011973 -0.14963411  0.5024157 -0.02068803  1.00000000
```

The diagonals are 1.0000 and every single variable perfectly correlated with itself, off diagonals values include correlations among different variables. The matrix shows high level of correlation between predictors that indicate multi-collinearity.

c) Fit model to explain relationship between the response Survival and other predictors

Mathematical multiple regression model

$$Y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + \varepsilon$$

The response Y: dependent variable (Survival variable)

β_0: intercept

x_1: the first independent variable (Prognosis variable)

x_2: the second independent variable (Enzyme variable)

x_3: the third independent variable (Liver variable)

x_4: the fourth independent variable (Age variable)

x_5: the fifth independent variable (Gender variable)

x_6: the sixth independent variable (Blood variable)

Intercept:

$$b_0 = \bar{y} - \bar{x_1}b_1 - \bar{x_2}b_2 - ... - \bar{x_6}b_6$$

Mathematical model for the overall ANOVA

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y_i} - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y})^2$$

Hypotheses for Overall ANOVA

H0 : $\beta_{Prognosis} = \beta_{Enzyme} = \beta_{Liver} = \beta_{Age} = \beta_{Gender} = \beta_{Blood} = 0$;

H1 : $\beta_i \neq 0$ for at least one i (not all β_i parameters are zero)

Overall ANOVA table for Multiple Regression

```
surg.lm=lm(Survival ~ Prognosis + Enzyme + Liver + Age + Gender + Blood, data=surg)
summary(surg.lm)
```

```
##
## Call:
## lm(formula = Survival ~ Prognosis + Enzyme + Liver + Age + Gender +
##     Blood, data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.25 -147.61   11.72  124.67  954.44
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1179.1889   283.8232  -4.155 0.000136 ***
## Prognosis       8.5013     2.1601   3.936 0.000273 ***
## Enzyme         11.1246     1.9820   5.613 1.03e-06 ***
## Liver          38.5068    51.7967   0.743 0.460926
## Age            -2.3409     3.0141  -0.777 0.441257
## GenderM        -0.2201    67.5146  -0.003 0.997413
## Blood          86.6437    27.4920   3.152 0.002825 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 233.1 on 47 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.656
## F-statistic: 17.85 on 6 and 47 DF,  p-value: 1.19e-10
```

```
anova(surg.lm)
```

```
## Analysis of Variance Table
##
## Response: Survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Prognosis  1 1479767 1479767 27.2441 3.989e-06 ***
## Enzyme     1 2896818 2896818 53.3336 2.842e-09 ***
## Liver      1  885709  885709 16.3069 0.0001975 ***
## Age        1    3027    3027  0.0557 0.8144163
## Gender     1   11906   11906  0.2192 0.6418084
## Blood      1  539487  539487  9.9326 0.0028253 **
## Residuals 47 2552807   54315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full RegSS = 1479767 + 2896818 + 885709 + 3027 + 11906 + 539487 = 5816714.

RegM.S $= \frac{Reg.S.S}{k} = \frac{5816714}{6} = 969452.3$.

H0: $\beta_{Prognosis} = \beta_{Enzyme} = \beta_{Liver} = \beta_{Age} = \beta_{Gender} = \beta_{Blood} = 0$;

H1: $\beta_i \neq 0$ for at least one i (not all β_i parameters are zero).

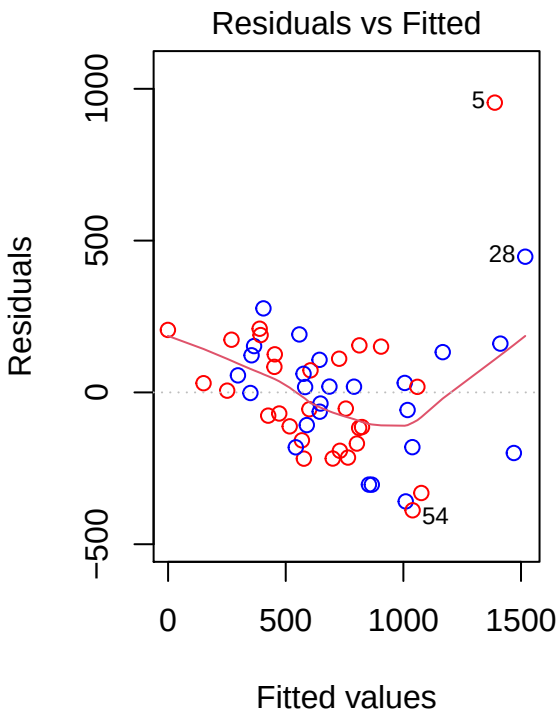
Test statistic: $F_{obs} = \frac{Reg.M.S}{Res.M.S} = \frac{969452.3}{54315} = 17.8487$.

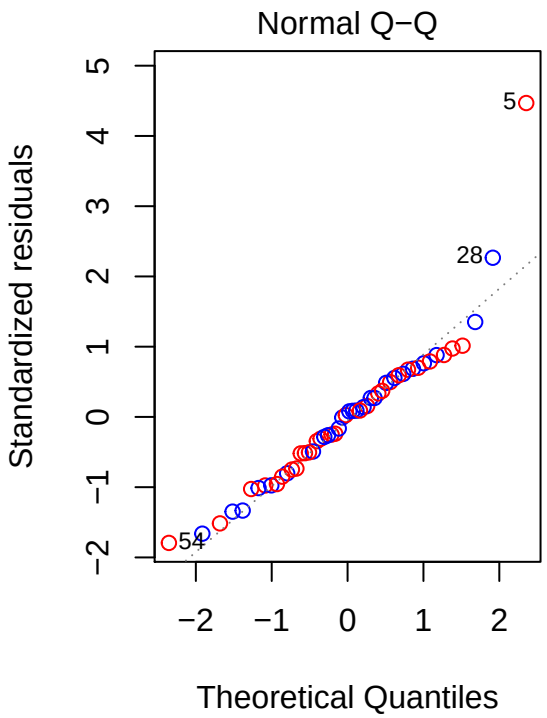
P-value $= \mathrm{P}(F_{6,47} \geq 17.8487) = 1.190228\text{e-}10 < 0.01$.

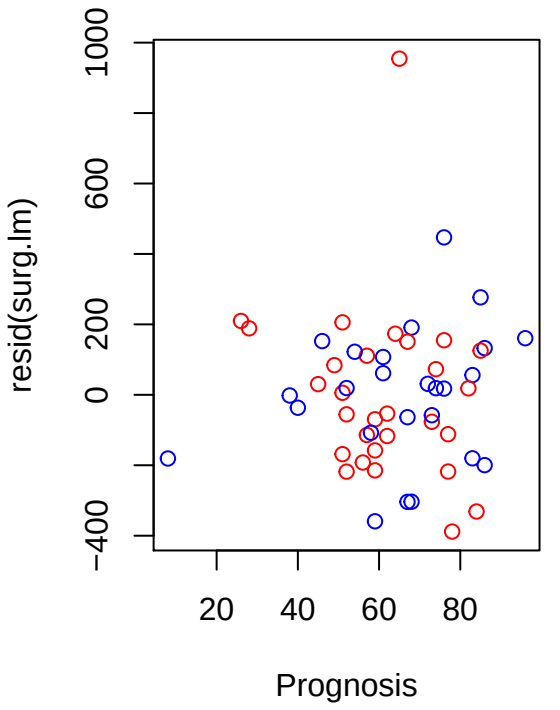
Reject H0 at 5% significant level. There is significant linear relationship between Survival response and at least one of the six predictors.

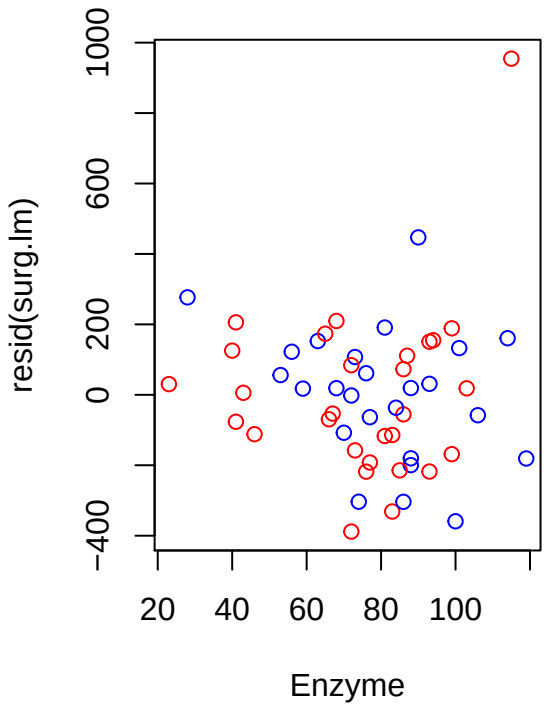
d) The most suitable regression model for the data

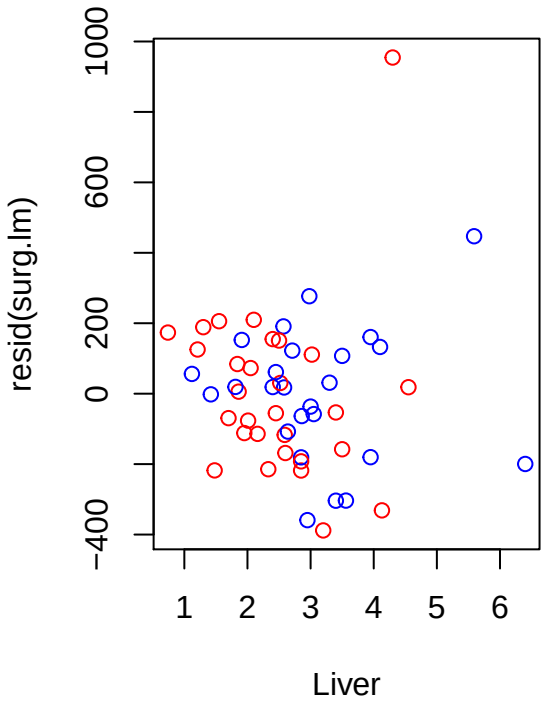
Diagnostic check

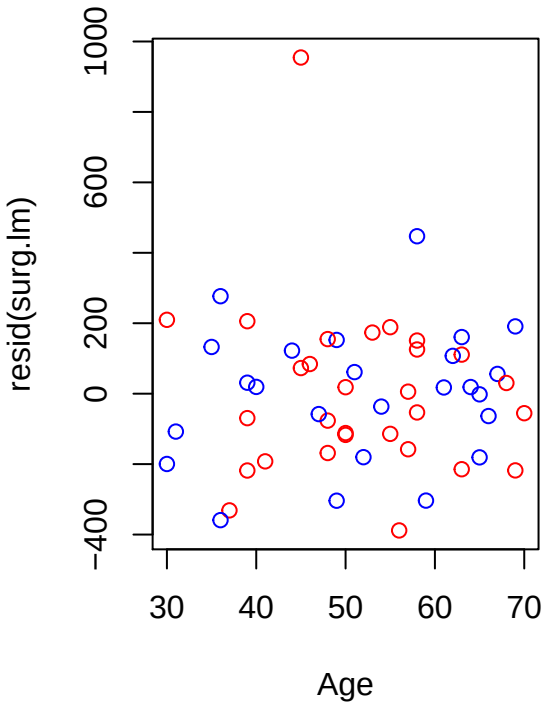


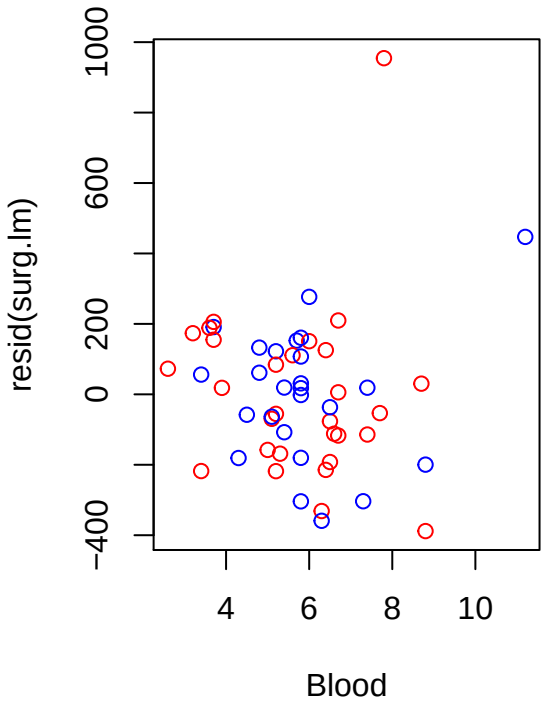












Normal Quantile-Quantile plot of residuals has concave up shape, which indicates skewness. The residuals vs fitted plot shows curvature. There might be slight curvature evidence in residuals vs liver function Index that motivate

a multiplicative model. It is shown that log transformation is necessary in this situation because of data skewness and curvature.

Full multiple regression model starts with all predictors

```
surg.1 = lm(Survival ~ ., data = surg)
summary(surg.1)
```

```
##
## Call:
## lm(formula = Survival ~ ., data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.25 -147.61   11.72  124.67  954.44
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1179.1889   283.8232  -4.155 0.000136 ***
## Prognosis       8.5013     2.1601   3.936 0.000273 ***
## Enzyme         11.1246     1.9820   5.613 1.03e-06 ***
## Liver          38.5068    51.7967   0.743 0.460926
## Age            -2.3409     3.0141  -0.777 0.441257
## GenderM        -0.2201    67.5146  -0.003 0.997413
## Blood          86.6437    27.4920   3.152 0.002825 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 233.1 on 47 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.656
## F-statistic: 17.85 on 6 and 47 DF,  p-value: 1.19e-10
```

Predictor Gender has the largest P-value (P-value = 0.997413). Drop Gender predictor.

Reduced model after dropping Gender

```
surg.2 = lm(Survival ~ Prognosis + Enzyme + Liver + Age + Blood, data = surg)
summary(surg.2)
```

```
##
## Call:
## lm(formula = Survival ~ Prognosis + Enzyme + Liver + Age + Blood,
##     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 ***
## 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
## 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
```

Predictor Liver has the largest P-value (P-value = 0.437595). Drop Liver predictor.

Reduced model after dropping Liver

```
surg.3 = lm(Survival ~ Prognosis + Enzyme + Age + Blood, data = surg)
summary(surg.3)
```

```
##
## Call:
## lm(formula = Survival ~ Prognosis + Enzyme + Age + Blood, data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -416.92 -142.56  -13.98  138.10  943.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1246.655    260.835  -4.779 1.64e-05 ***
## Prognosis       9.291      1.876   4.951 9.14e-06 ***
## Enzyme         12.101      1.502   8.058 1.56e-10 ***
## Age            -2.986      2.841  -1.051    0.298
## Blood         100.660     19.987   5.036 6.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.7 on 49 degrees of freedom
## Multiple R-squared:  0.6911, Adjusted R-squared:  0.6659
## F-statistic: 27.41 on 4 and 49 DF,  p-value: 5.68e-12
```

Predictor Age has the largest P-value (P-value = 0.298). Drop Age predictor.

Final model after dropping Gender, Liver and Age

```
surg.4 = lm(Survival ~ Prognosis + Enzyme + Blood, data = surg)
summary(surg.4)
```

```
##
## Call:
## lm(formula = Survival ~ Prognosis + Enzyme + Blood, data = surg)
##
```

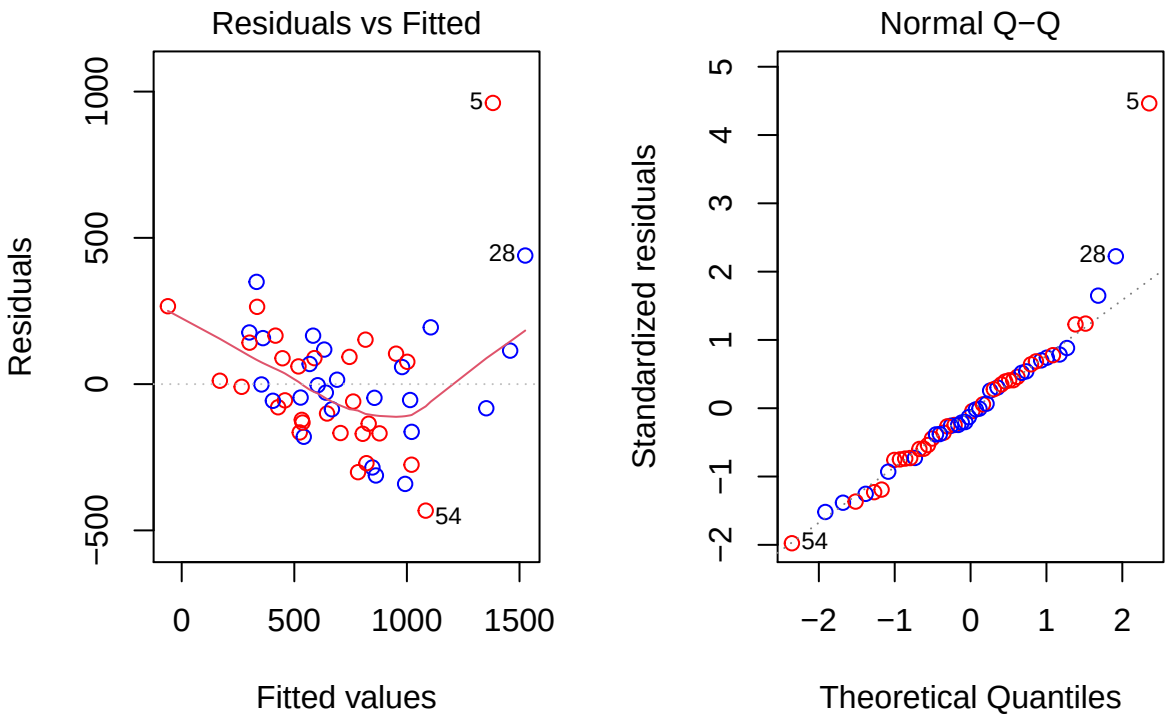
```
## Residuals:
##    Min     1Q Median     3Q    Max
## -432.4 -134.3  -19.1  111.9  961.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1410.847    209.118  -6.747 1.50e-08 ***
## Prognosis       9.382      1.876   5.000 7.43e-06 ***
## Enzyme         12.128      1.503   8.069 1.30e-10 ***
## 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
```

The most appropriate multiple regression model is the reduced model after dropping insignificant predictors such as Gender, Liver and Age. At this stage, remaining predictors are all significant and need to be remained in the model.

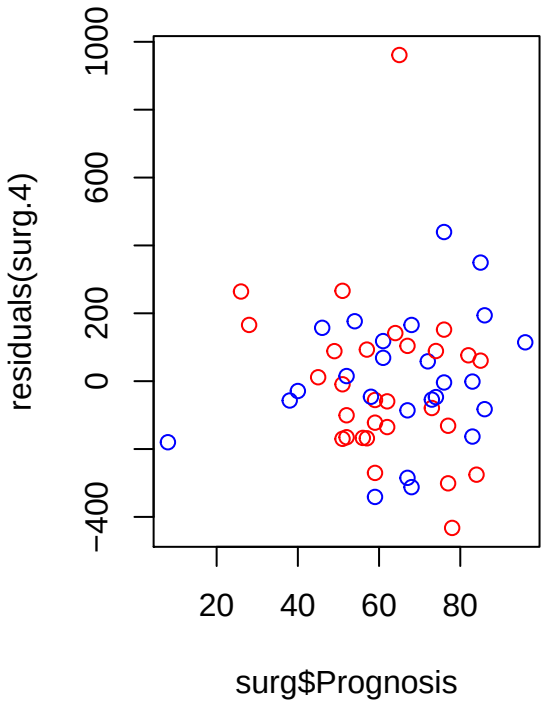
e) Validate Final Model and explanation for inappropriate multiple regression model

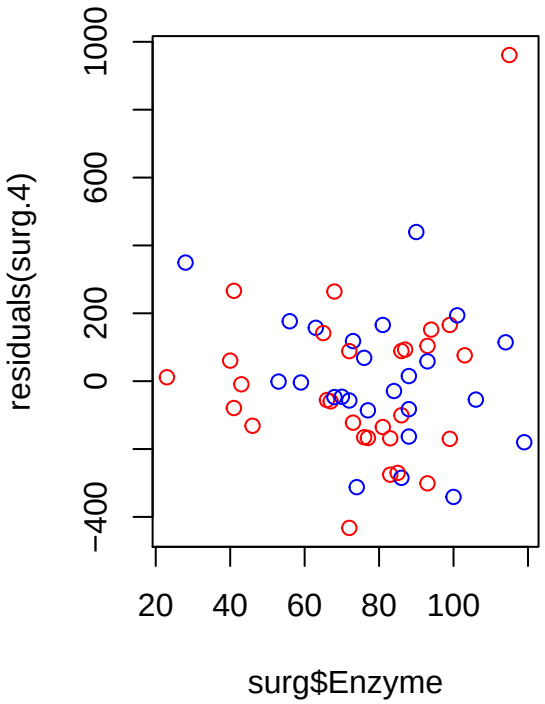
The final model is the reduced model after dropping three insignificant predictors such as Gender, Liver and Age.

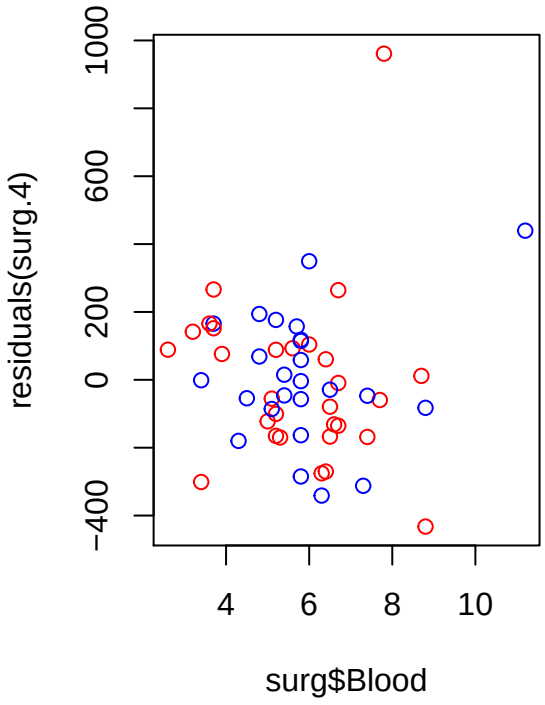
Check Diagnostics



Check residuals against predictors







As can be seen in the plots after applying the final multiple regression model, the Normal Quantile-Quantile plot of residuals still has concave up shape, which indicates skewness. The residuals vs fitted plot shows curvature. It is shown that log transformation is necessary in this situation because the data still shows skewness and curvature. Additionally, when using multiple regression model, it is shown that intercept is negative. It means that the expected number on Survival response will be less than 0 when other predictors will be set to 0. Therefore, it is inappropriate to apply multiple regression model to this study.

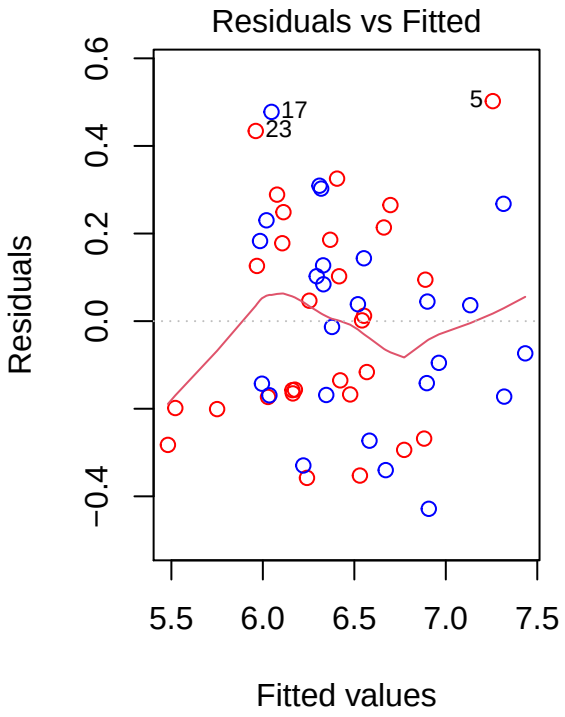
f) Re-fit the model using transformation

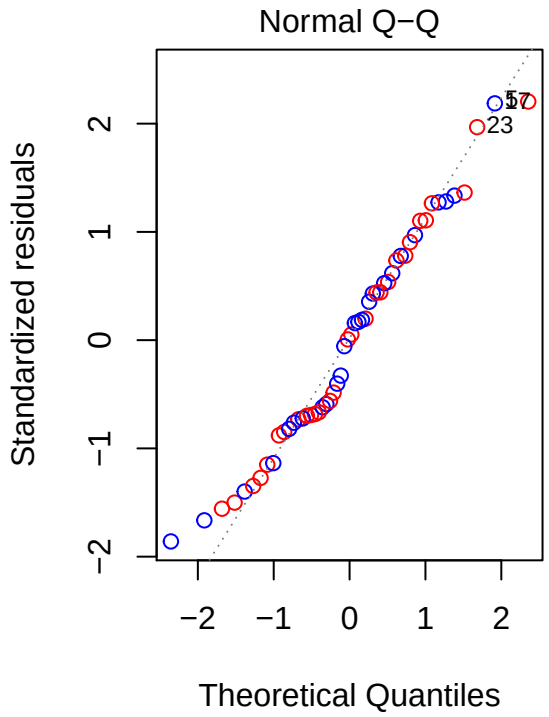
Log transform start with all predictors

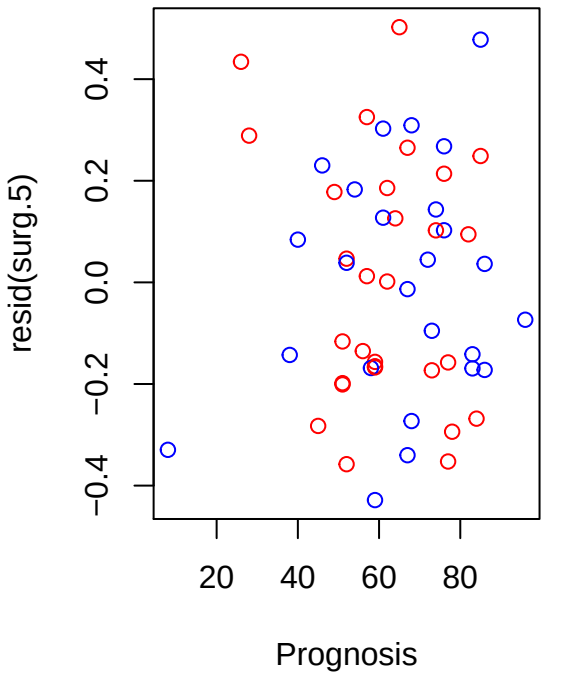
```
surg.5 = lm( log(Survival) ~ Prognosis + Enzyme + Liver + Age + Gender + Blood, data = surg)
summary(surg.5)
```

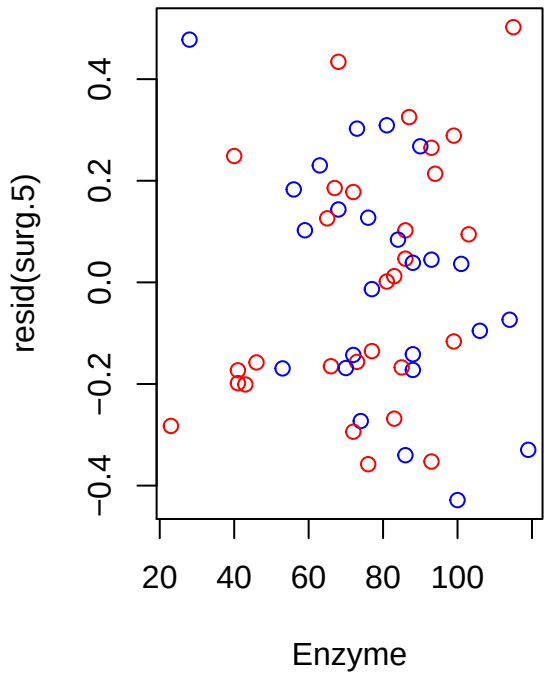
```
##
## Call:
## lm(formula = log(Survival) ~ Prognosis + Enzyme + Liver + Age +
##     Gender + Blood, data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42847 -0.16913  0.00696  0.18167  0.50226
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.100997   0.302781  13.544  < 2e-16 ***
## Prognosis    0.013020   0.002304   5.650 9.08e-07 ***
## Enzyme       0.016245   0.002114   7.683 7.59e-10 ***
## Liver       -0.003132   0.055256  -0.057  0.95503
## Age         -0.004863   0.003215  -1.513  0.13709
## GenderM     -0.066140   0.072024  -0.918  0.36315
## Blood        0.094858   0.029328   3.234  0.00223 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2486 on 47 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7441
## F-statistic: 26.69 on 6 and 47 DF,  p-value: 1.391e-13
```

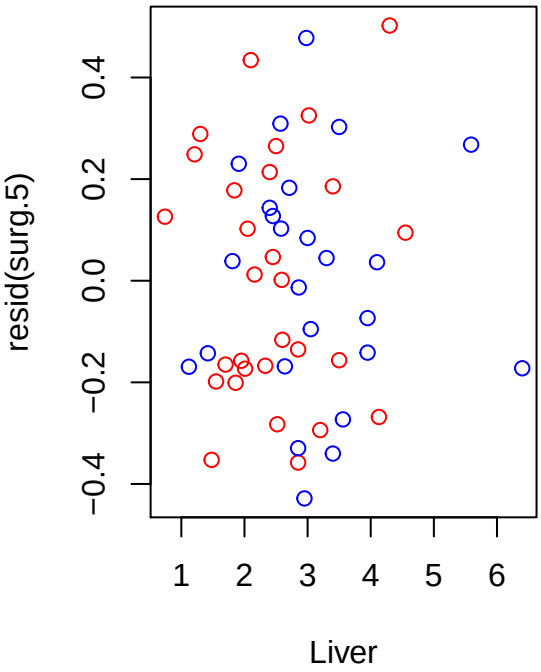
Predictor Liver has the largest P-value (P-value = 0.95503). Drop Liver Predictor.

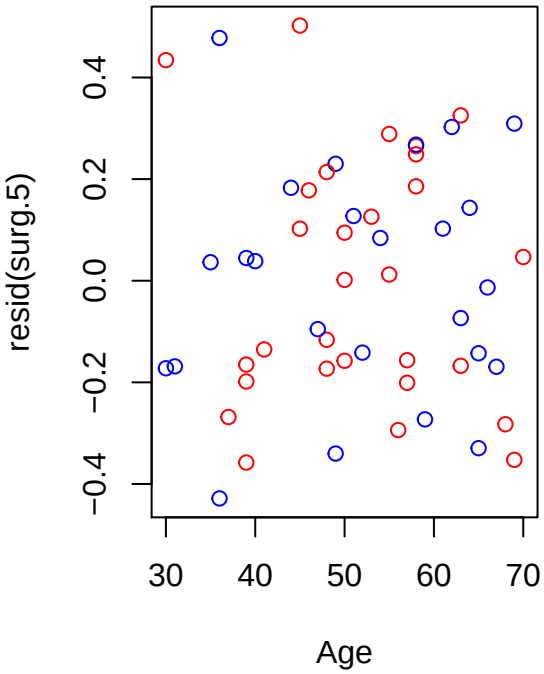

Residuals vs Fitted
Residuals
Fitted values
17
23
5
−0.4
0.0
0.2
0.4
0.6
5.5
6.0
6.5
7.0
7.5

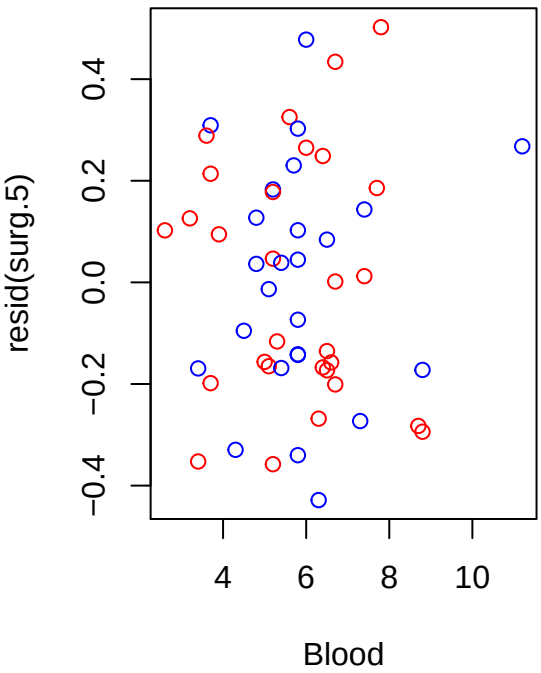

Normal Q−Q
Standardized residuals
Theoretical Quantiles
17
23
5
−2
−1
0
1
2


resid(surg.5)
Prognosis
−0.4
−0.2
0.0
0.2
0.4
20
40
60
80


resid(surg.5)
Enzyme
−0.4
−0.2
0.0
0.2
0.4
20
40
60
80
100







The normal quantile-quantile plot of residuals is more linear, which closer meets the requirements of the model. There is slight curvature in the residual versus Liver and Age predictors. It is possible to use quadratic term. In

both the original model and log transformed model, Liver, Gender and Age is insignificant values and need to be dropped.

Remove Liver predictor

```
surg.6 = update(surg.5, . ~ . - Liver)
summary(surg.6)
```

```
##
## Call:
## lm(formula = log(Survival) ~ Prognosis + Enzyme + Age + Gender +
##     Blood, data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42563 -0.16780  0.00911  0.18059  0.50244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.105132   0.290795  14.117  < 2e-16 ***
## Prognosis    0.012960   0.002026   6.398 6.16e-08 ***
## Enzyme       0.016170   0.001627   9.939 3.10e-13 ***
## Age         -0.004810   0.003043  -1.581    0.121
## GenderM     -0.065010   0.068487  -0.949    0.347
## Blood        0.093738   0.021439   4.372 6.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.246 on 48 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7495
## F-statistic: 32.71 on 5 and 48 DF,  p-value: 2.291e-14
```

Predictor Gender has the largest P-value (P-value = 0.347). Drop Gender Predictor.

Remove Gender predictor

```
surg.7 = update(surg.6, . ~ . - Gender)
summary(surg.7)
```

```
##
## Call:
## lm(formula = log(Survival) ~ Prognosis + Enzyme + Age + Blood,
##     data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39491 -0.18866 -0.00045  0.17491  0.51787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.028531   0.279090  14.434  < 2e-16 ***
## Prognosis    0.013199   0.002008   6.574 3.04e-08 ***
```

```
## Enzyme       0.016402   0.001607  10.208 1.01e-13 ***
## Age         -0.004767   0.003040  -1.568    0.123
## Blood        0.094845   0.021386   4.435 5.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2458 on 49 degrees of freedom
## Multiple R-squared:  0.7688, Adjusted R-squared:   0.75
## F-statistic: 40.74 on 4 and 49 DF,  p-value: 5.171e-15
```

Predictor Age has the largest P-value (P-value = 0.123). Drop Age Predictor.

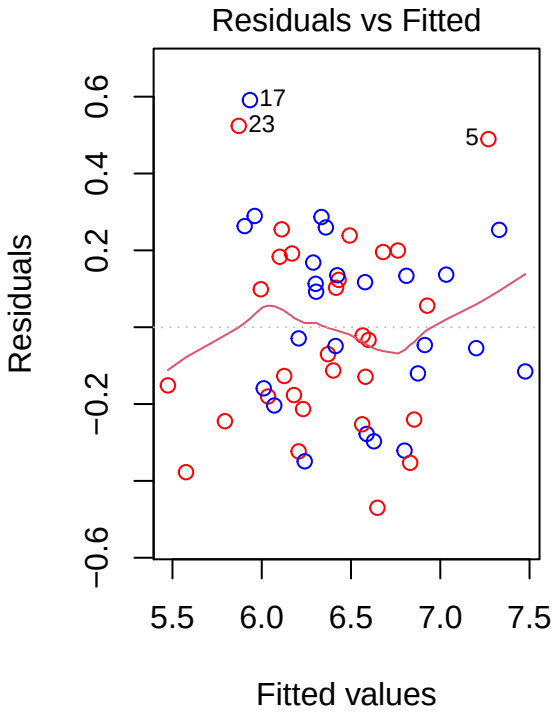
Remove Age predictor

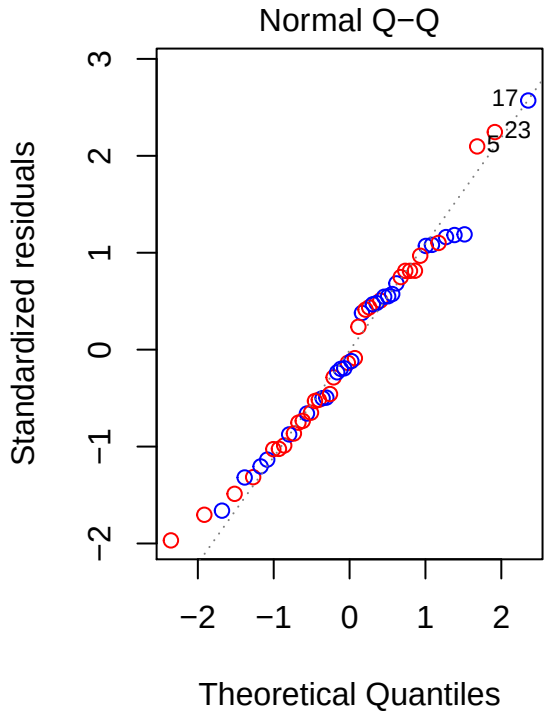
```
surg.8 = update(surg.7, . ~ . - Age)
summary(surg.8)
```

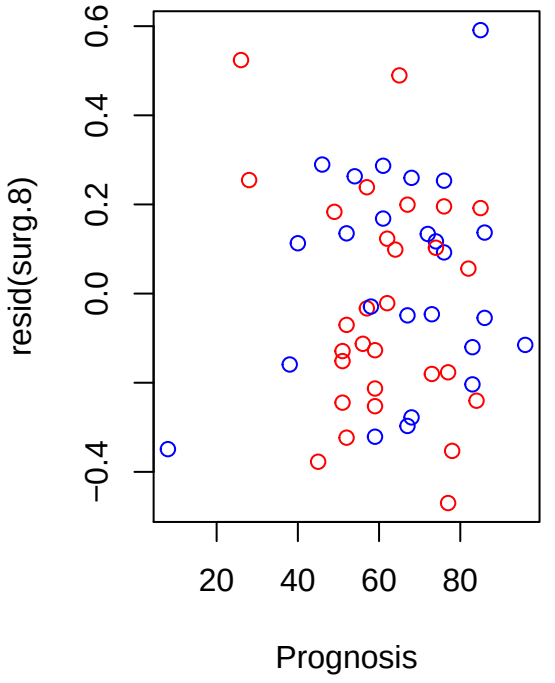
```
##
## Call:
## lm(formula = log(Survival) ~ Prognosis + Enzyme + Blood, data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46994 -0.17938 -0.03116  0.17959  0.59105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.766441   0.226757  16.610  < 2e-16 ***
## Prognosis   0.013344   0.002035   6.558 2.95e-08 ***
## Enzyme      0.016444   0.001630  10.089 1.19e-13 ***
## Blood       0.095475   0.021692   4.401 5.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2493 on 50 degrees of freedom
## Multiple R-squared:  0.7572, Adjusted R-squared:  0.7427
## F-statistic: 51.99 on 3 and 50 DF,  p-value: 2.137e-15
```

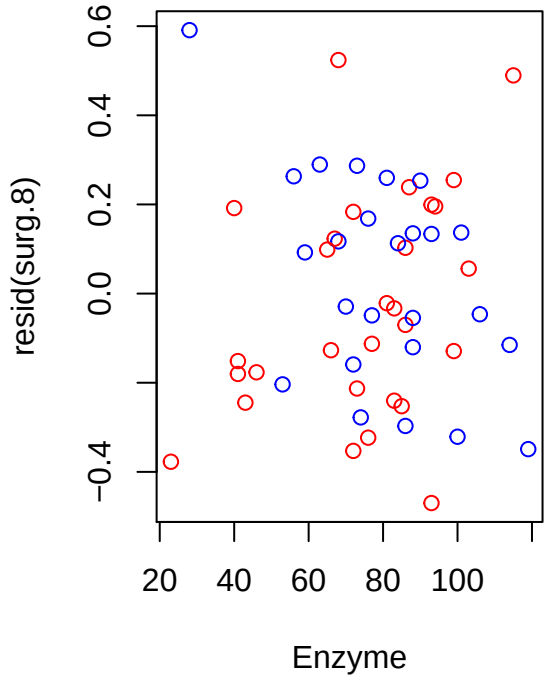
At this stage, remaining predictors are all significant and need to be remained in the model.

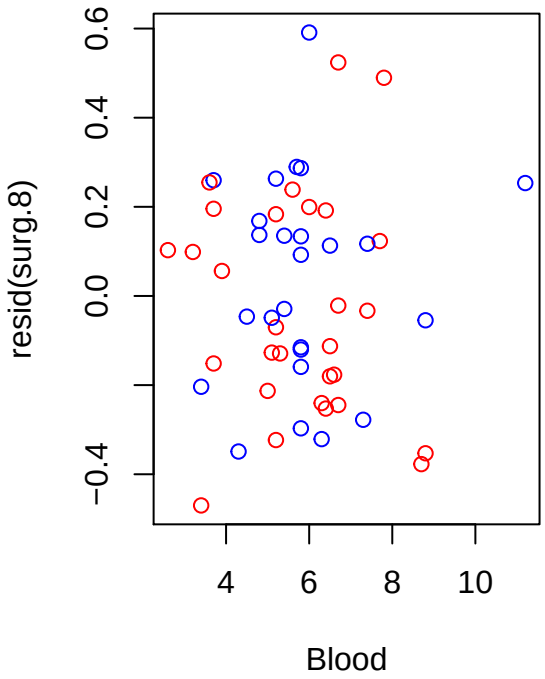
g) Validate final model with log(survival) response











The normal quantile-quantile plot of residuals is linear and show no skewness which meets the model requirements. There is no curvature in the residual versus all the predictors.

The regression model with log(survival) response is more appropriate than the regression model with original data. Because log transformation follows linear regression framework, reduce skewness of the data and its inferences would be reliably used.

Question 2

a) Explain the design study

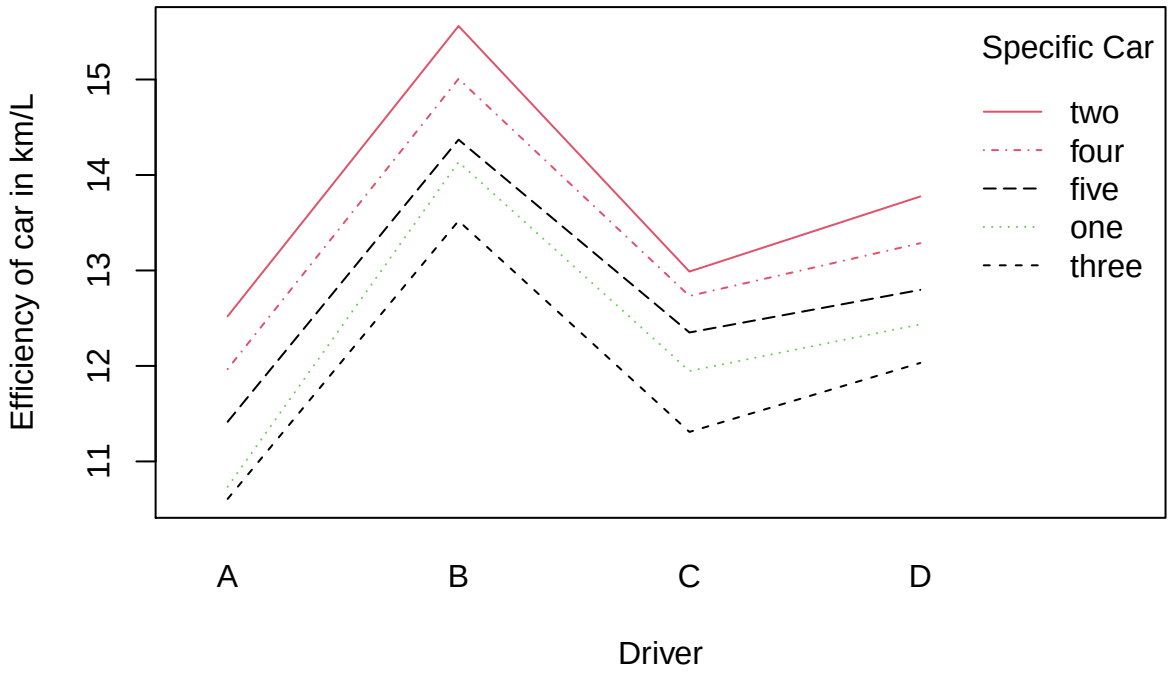
```
with(kml,table(driver,car)) # Check the number of pairs
```

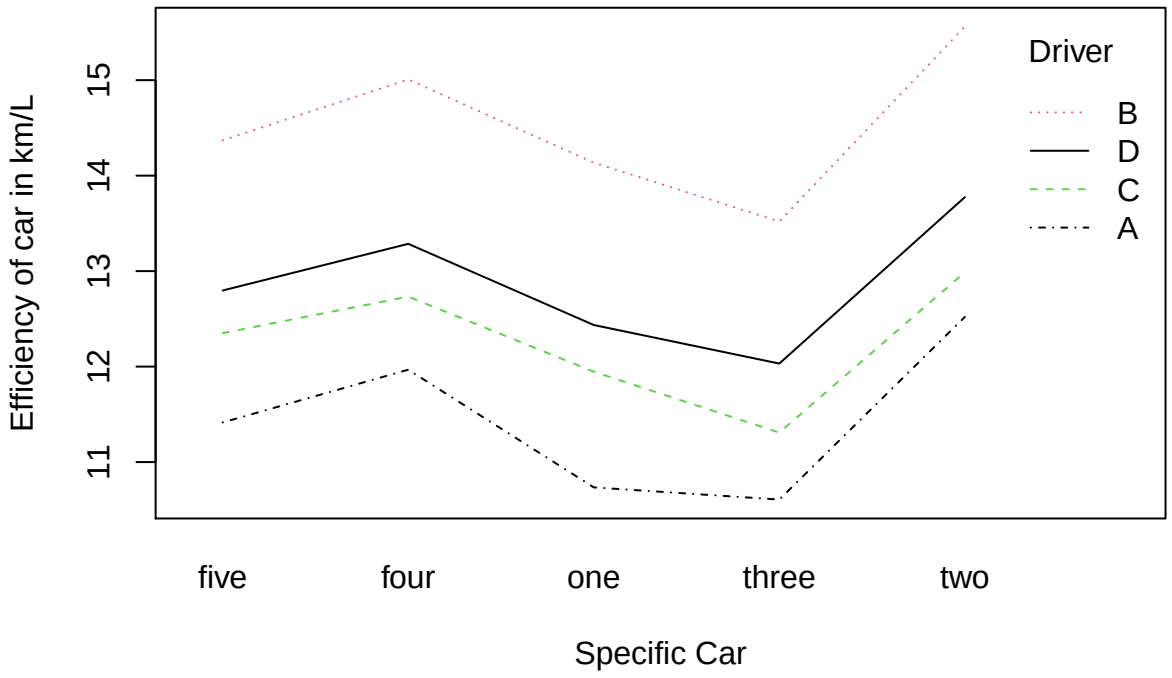
```
##       car
## driver five four one three two
##      A    2    2   2     2   2
##      B    2    2   2     2   2
##      C    2    2   2     2   2
##      D    2    2   2     2   2
```

The design of the study is balanced. Because there are equal number of observations in all cells and for all possible pairs of factor levels for factor driver (labeled A, B, C and D) and car (labeled one, two, three, four and five).

b) Preliminary investigation

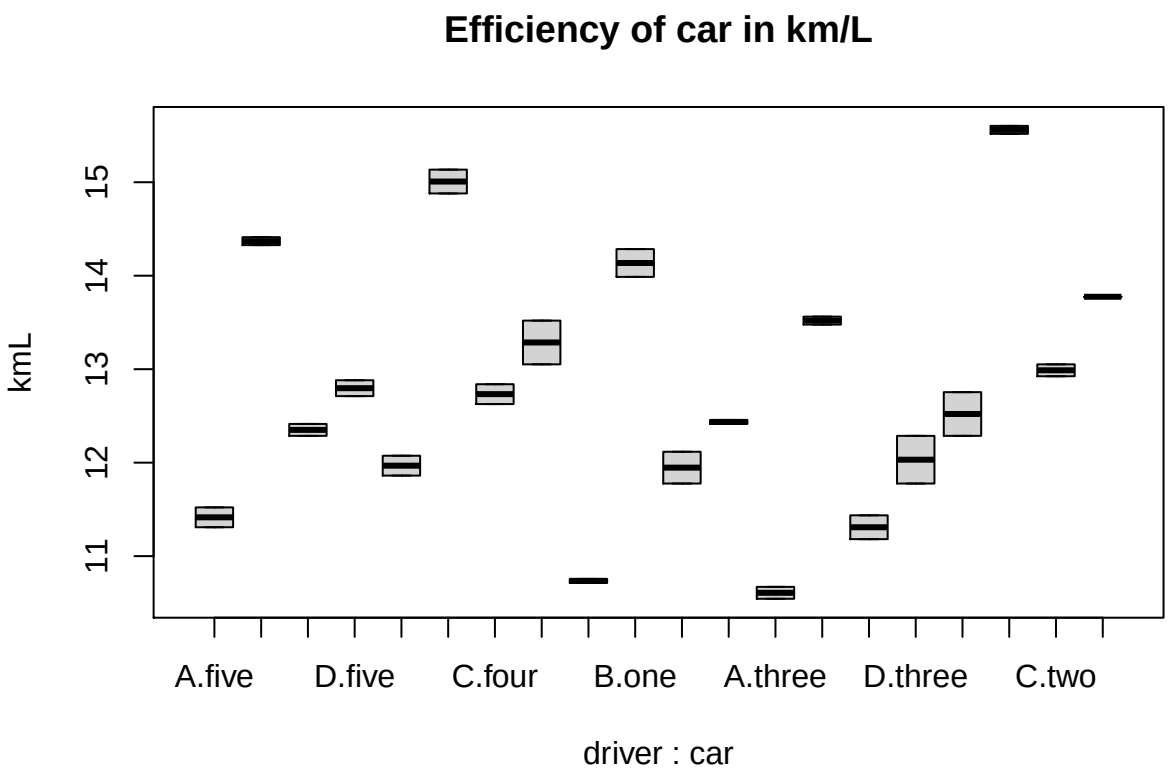
Interaction plot





The lines are parallel, so there is no interaction between Factor driver and Factor car. Factor car has a constant effect on the efficiency of car in km/L (which is irrespective of Factor driver)

Boxplot



The boxplot contains huge number of cells which is difficult to interpret.

c) Balanced Design Test

Model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}$$

There are 3 test types

1. Interaction test

H0: $\gamma_{ij} = 0$ for all i, j; **HA:** not all $\gamma_{ij} = 0$

2. Main effect Driver

H0: $\alpha_i = 0$ for all i; **HA:** not all $\alpha_i = 0$

3. Main effect Car

H0: $\beta_j = 0$ for all j; **HA:** not all $\beta_j = 0$

Fit full model with interaction

```
kml.1 = lm(kmL ~ car * driver, data = kml)
summary(kml.1)$coefficients
```

```
##                    Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      11.4151164  0.1260285 90.5756837 1.276357e-27
## carfour           0.5526872  0.1782312  3.1009566 5.632645e-03
## carone           -0.6802304  0.1782312 -3.8165619 1.079967e-03
## carthree         -0.8077736  0.1782312 -4.5321673 2.029804e-04
## cartwo            1.1053744  0.1782312  6.2019132 4.662143e-06
## driverB           2.9547508  0.1782312 16.5781910 3.752449e-13
## driverC           0.9353168  0.1782312  5.2477727 3.898917e-05
## driverD           1.3817180  0.1782312  7.7523915 1.888067e-07
## carfour:driverB   0.0850288  0.2520570  0.3373396 7.393755e-01
## carone:driverB    0.4464012  0.2520570  1.7710329 9.179512e-02
## carthree:driverB -0.0425144  0.2520570 -0.1686698 8.677505e-01
## cartwo:driverB    0.0850288  0.2520570  0.3373396 7.393755e-01
## carfour:driverC  -0.1700576  0.2520570 -0.6746792 5.076044e-01
## carone:driverC    0.2763436  0.2520570  1.0963537 2.859503e-01
## carthree:driverC -0.2338292  0.2520570 -0.9276839 3.646325e-01
## cartwo:driverC   -0.4676584  0.2520570 -1.8553678 7.834370e-02
## carfour:driverD  -0.0637716  0.2520570 -0.2530047 8.028469e-01
## carone:driverD    0.3188580  0.2520570  1.2650235 2.204044e-01
## carthree:driverD  0.0425144  0.2520570  0.1686698 8.677505e-01
## cartwo:driverD   -0.1275432  0.2520570 -0.5060094 6.183821e-01
```

F-test for interaction term, ANOVA table for the full model

```
anova(kml.1)
```

```
## Analysis of Variance Table
##
## Response: kmL
##            Df Sum Sq Mean Sq F value    Pr(>F)
## car         4 17.119  4.2798  134.73 3.664e-14 ***
## driver      3 50.661 16.8869  531.60 < 2.2e-16 ***
## car:driver 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
```

Model:

$$Y = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon$$

Hypotheses:

H0: $\gamma_{ij} = 0$;

H1: at least one $\gamma_{ij} \neq 0$.

P-Value $= 0.3715 > 0.05$.

The interaction is not significant. Therefore, reduced model with main effects only need to be fit

Fit reduced model without interaction (only main effects)

```
kml.2 = update(kml.1, . ~ . - car:driver)
anova(kml.2)
```

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

Main Effects: Driver

Model: $Y = \mu + \alpha_i + \beta_j + \varepsilon$.

Hypotheses:

H0: $\beta_j = 0$;

H1: at least one $\beta_j \neq 0$. P-Value = 2.2e-16 < 0.05.

Driver type is significant

Main Effects: Car

Model: $Y = \mu + \alpha_i + \beta_j + \varepsilon$.

Hypotheses:

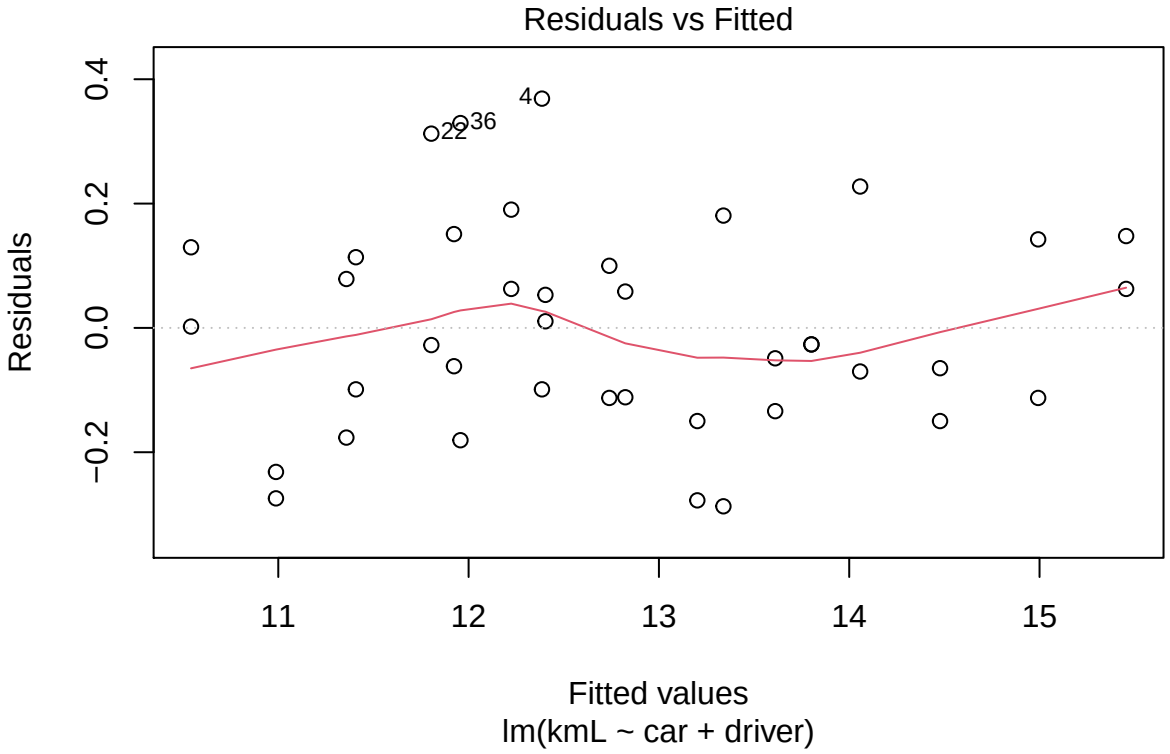
H0: $\alpha_i = 0$;

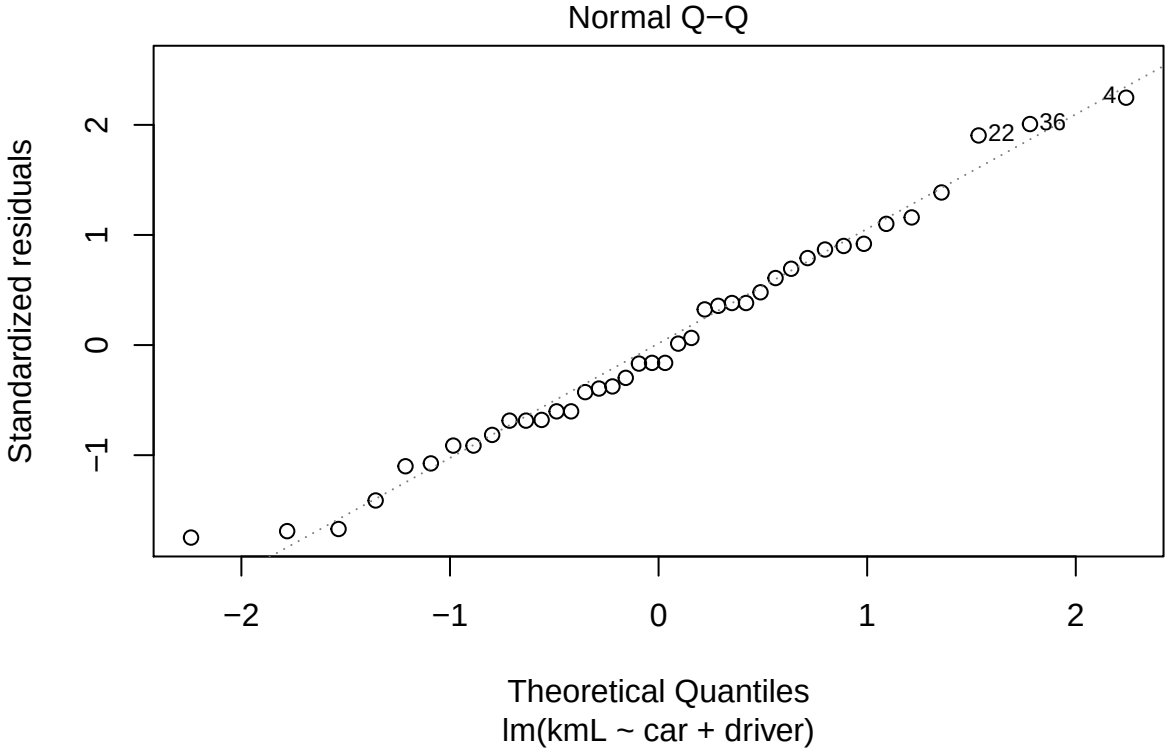
H1: at least one $\alpha_i \neq 0$.

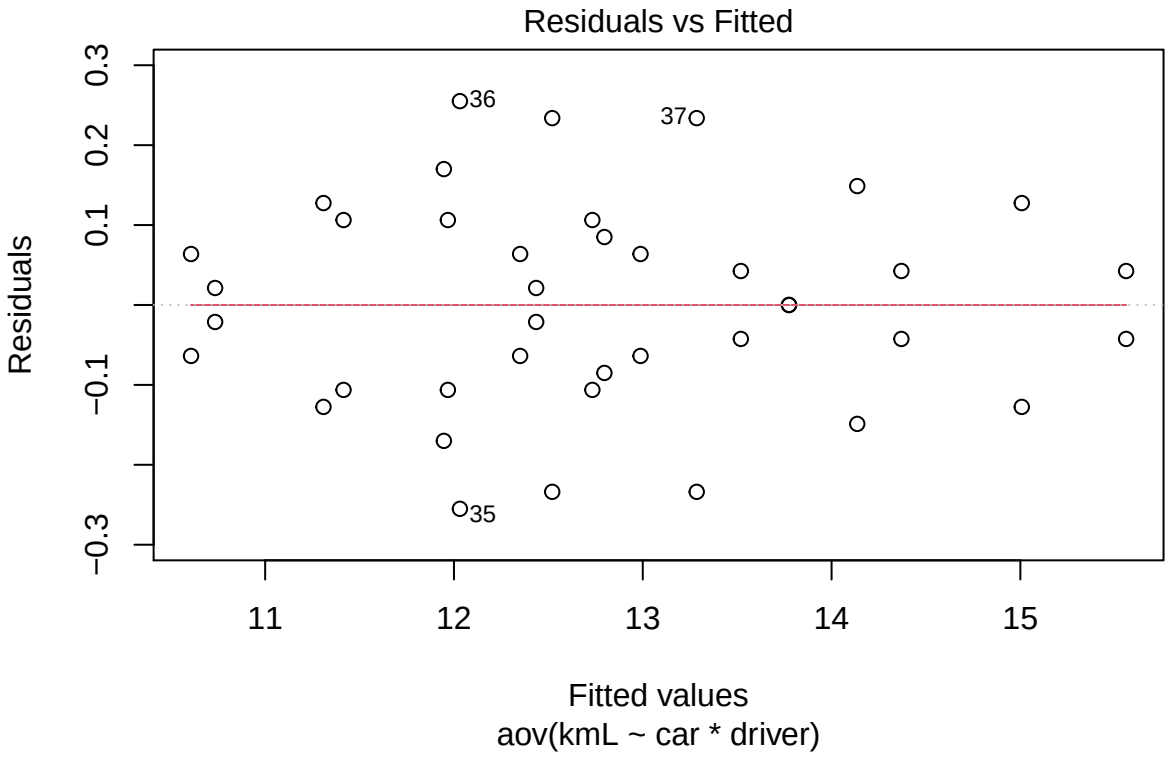
P-Value = 2.2e-16 < 0.05.

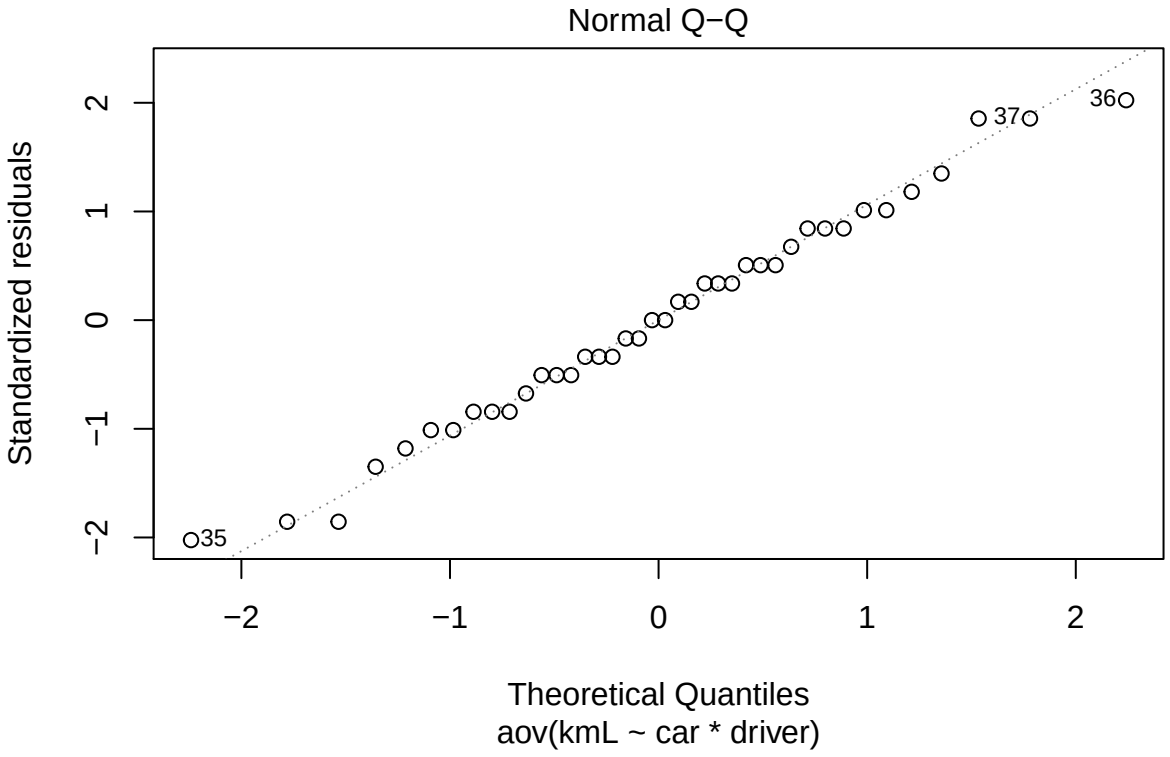
Car type is significant

Checking Assumptions









No curvature in the normal quantile plot of residuals. The points is scattered evenly above and below the line.

d) Conclusion

For the fit model with interaction, since the result of p-value is insignificant, so the effect of factor Driver on the efficiency of the car in km/L is independent of factor Car and there is no interaction between the two factors.

For the fit model with main effects, since the p-value for both factor Driver and Car are significant, so at least one population mean of the efficiency of the car in km/L is different from others for all levels of factor Driver and factor Car.