Assignment

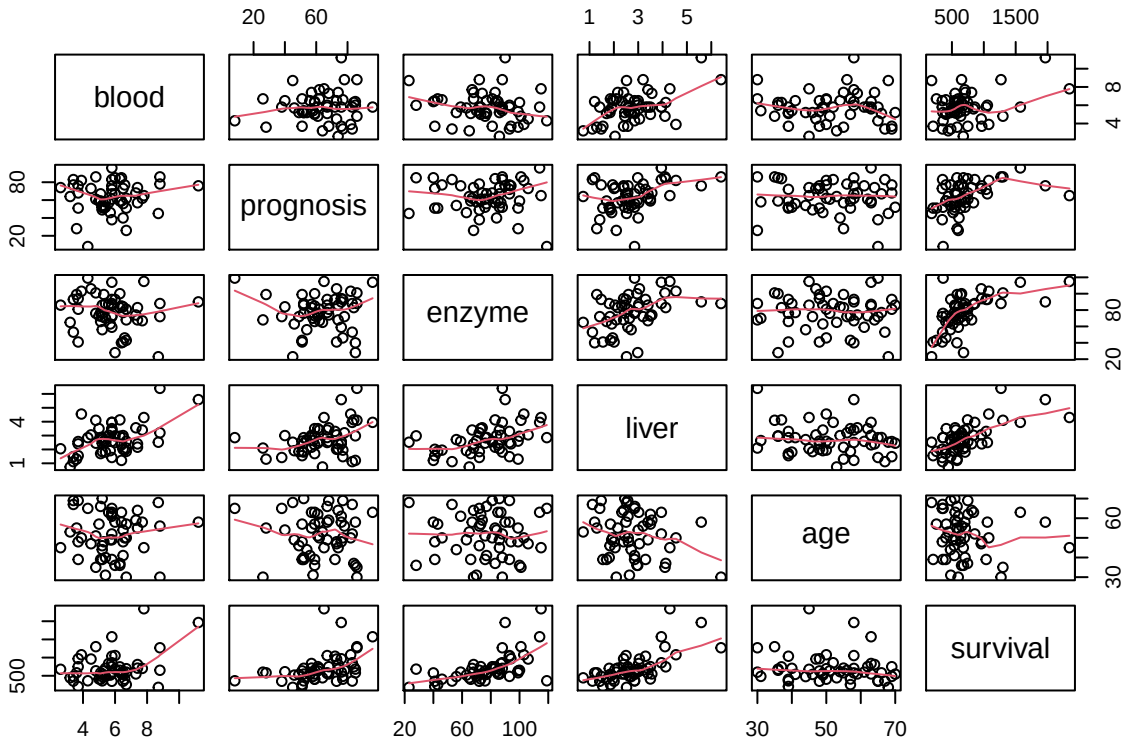
Thi Thu Huong Nguyen

5/19/2021

Question 1

a.

```
surg <- read.csv("~/R/surg.dat", sep="")
surg1= subset(surg, select=-gender)
pairs(surg1, panel=panel.smooth)
```



- Relationship between the response and predictors
  As we can see in the plot, there are possible relationships among survival time and other predictors. While the relationships between Survival with blood, prognosis, enzyme and liver are moderately or slightly positive, its between survival and age is negative. The strongest positive relationship is between survival and liver.
- Relationship among predictors

There are some noticeable positive relationships between liver with other predictors such as blood, prognosis and enzyme. The relationships between age and other predictors can be seen to be weak from the scatter plot.

- We need to remove *gender* variables because it's categorical variable. In order to compute correlation matrix, we need numerical variable.

b.

```
cor(surg1)
```

```
##                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
```

By computing correlation matrix, we can test the correlation strength among the response and predictors. Overall, there are mostly weak correlation, raging from -0.2 to 0.3. However, there are still some moderate correlation. Moderate correlation: survival with enzyme, prognosis and liver; liver with blood and enzyme.

c.
Since the researchers try to use many predictor variables to explain the survival response, multiple regression model is suitable.

```
surg1_lm = lm(survival ~ blood + prognosis + enzyme + liver + age, data=surg1)
```

Mathematics multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon; \epsilon \sim Norm(0, \sigma^2)$$

Y: survival
β_0: Intercept term
$\beta_1, \beta_2, ..., \beta_5$: Partial regression coefficients
$X_1 : blood; X_2 : prognosis; X_3 : enzyme; X_4 : liver; X_5 : age$
ϵ : variation a way from regression line.

Test hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$
H_1 : not all $\beta_i = 0$

```
anova(surg1_lm)
```

```
## 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
```

ANOVA table for the overall multiple regression model

Source	df	Sum of Sq	Mean Sq	F
Regression	5	5816714	1163342.8	21.874
Residual	48	2552807	53183	
Total	53	8369521		

- Test statistic: $F_{obs} = \frac{RegressionM.S}{ResidualM.S} = \frac{1163342.8}{53183} = 21.874$
- Null distribution: $F_{g,n-g-1} : F_{5,48}$
- P-Value: $\mathrm{P}(F_{5,48} \geq 21.874) = 2.386226\text{e-}11 < 0.05$
- Reject at the 5% level.
- There is a significant linear relationship between survival and at least one of the five predictor variables.

d. To find the best multiple regression model, we can apply stepwise backward estimation.

```
summary(surg1_lm)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -1179.366654 275.619347 -4.2789690 8.913076e-05
## blood          86.630445  26.904719  3.2198978 2.302423e-03
## prognosis       8.501113   2.137047  3.9779712 2.337301e-04
## enzyme         11.124165   1.957529  5.6827582 7.623756e-07
## liver          38.553562  49.251408  0.7827911 4.375949e-01
## age            -2.339958   2.969120 -0.7880981 4.345142e-01
```

Drop liver variable since it has the largest p-value (4.375949e-01)

```
summary(lm(survival ~ blood + prognosis + enzyme + age, data= surg1))$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1246.654818 260.835337 -4.779471 1.640827e-05
## blood         100.659551  19.987172  5.036208 6.831794e-06
## prognosis       9.290889   1.876432  4.951359 9.138843e-06
## enzyme         12.101482   1.501730  8.058360 1.555971e-10
## age            -2.986213   2.840741 -1.051209 2.983194e-01
```

Age is insignificant after dropping liver. Other variables are significant.

```
summary(lm(survival ~ blood + prognosis + enzyme, data= surg1))$coefficients
```

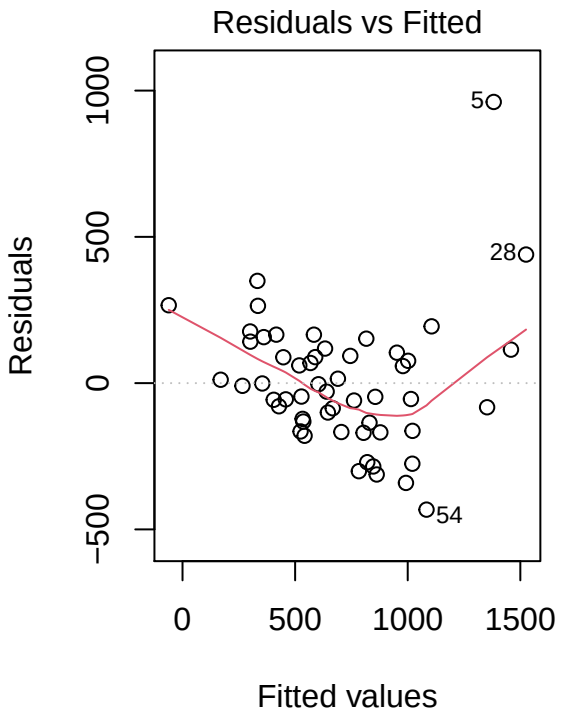
```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1410.846901 209.117946 -6.746656 1.495123e-08
## blood         101.053887  20.004632  5.051525 6.220022e-06
## prognosis       9.381966   1.876399  4.999985 7.433593e-06
## enzyme         12.127807   1.503098  8.068542 1.303361e-10
```

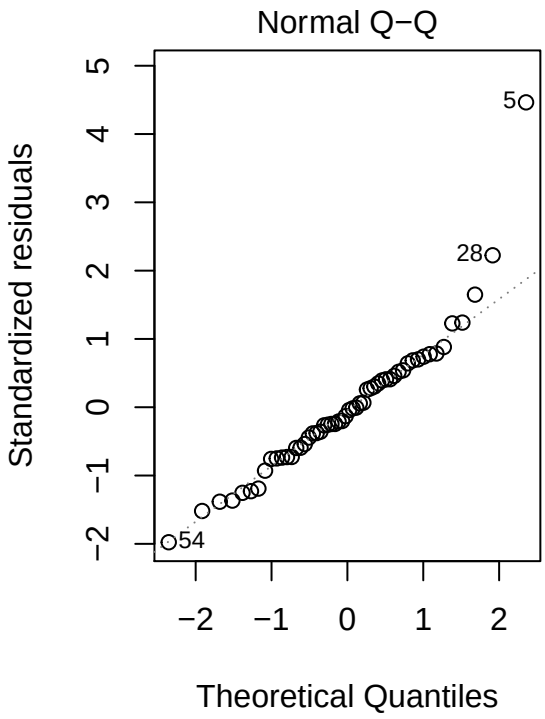
Blood, prognosis and enzyme are significant. This is the best multiple regression model.

e. Validate the best multiple regression model

```
best_model_lm = lm(survival ~ blood + prognosis + enzyme, data= surg1)

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





Comment:

- The QQ plot follows a positive linear, however, there are some observations such as #5, #28 and #54 may cause future problems.
- Residual vs fitted plot shows curvature, where the non-linear relationship cannot be explained by the model.

Therefore, the model is not validate and not appropriate to use the multiple regression model to explain the survival time.

f.

```
new_survival= log(surg$survival)
new_surg1_lm = lm(new_survival ~ blood + prognosis + enzyme + liver+age, data= surg1)
```

```
summary(new_surg1_lm)$coefficients
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  4.047579150 0.296654656 13.6440776 3.873510e-18
## blood        0.090873597 0.028958091  3.1381073 2.905219e-03
## prognosis    0.012975080 0.002300147  5.6409785 8.820988e-07
## enzyme       0.016125812 0.002106928  7.6537081 7.375820e-10
## liver        0.010914028 0.053010283  0.2058851 8.377514e-01
## age         -0.004583761 0.003195724 -1.4343423 1.579579e-01
```

Remove liver since it has the largest p-value.

```
summary(lm(new_survival ~ blood + prognosis + enzyme + age, data= surg1))$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
```

```
## (Intercept)  4.028530718 0.279090496 14.434496 2.820187e-19
## blood        0.094845060 0.021386020  4.434909 5.202486e-05
## prognosis    0.013198656 0.002007759  6.573826 3.035041e-08
## enzyme       0.016402478 0.001606832 10.207960 1.012035e-13
## age         -0.004766708 0.003039557 -1.568224 1.232646e-01
```

After dropping liver, all variables are significant except for age. Age should be dropped.

```
summary(lm(new_survival ~ blood + prognosis + enzyme, data= surg1))$coefficients
```

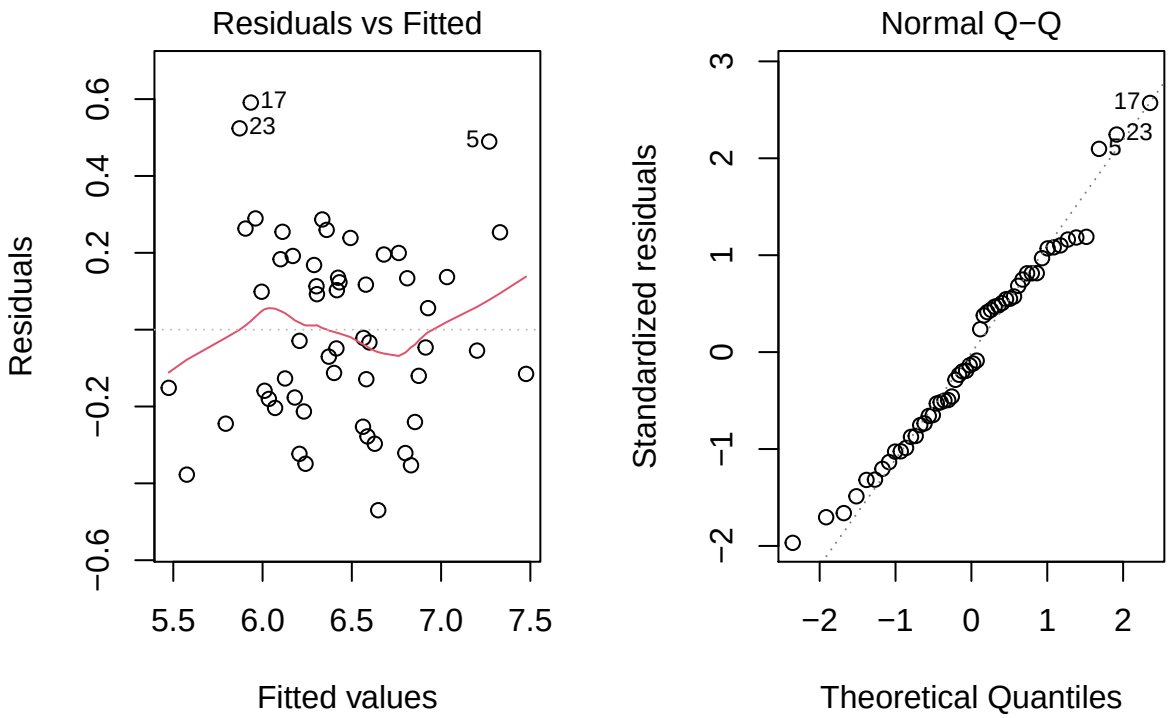
```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 3.76644097 0.226757297 16.610010 5.399369e-22
## blood       0.09547451 0.021692046  4.401360 5.655790e-05
## prognosis   0.01334404 0.002034675  6.558313 2.946869e-08
## enzyme      0.01644450 0.001629886 10.089356 1.190806e-13
```

All variables now are significant. This is the best model.

g. Validate the new best model:

```
new_best_model_lm = lm(new_survival ~ blood + prognosis + enzyme, data= surg1)

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



Comment: In the residuals vs Fitted plot, all the data points are distributed more evenly than the previous best model. Although there are still curves, it is more prone to straight line than the previous one. The normal QQ plot also shows a stronger linear line, which shows an improvement in the new model. Overall, the model with log(survival) is superior to the one with survival as the response.

Question 2

a.

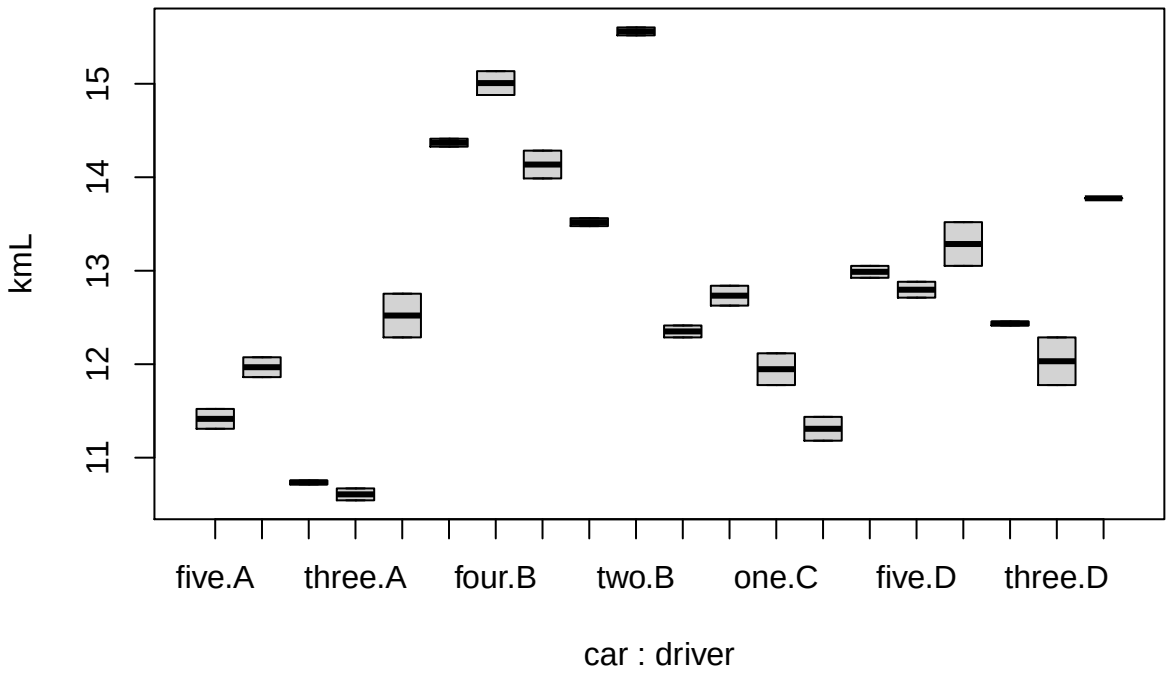
```
kml <- read.csv("~/R/kml.dat", sep="")
table(kml[, c("driver","car")])
```

```
##       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
```

It is a balanced design because the number of driver and car is the same.

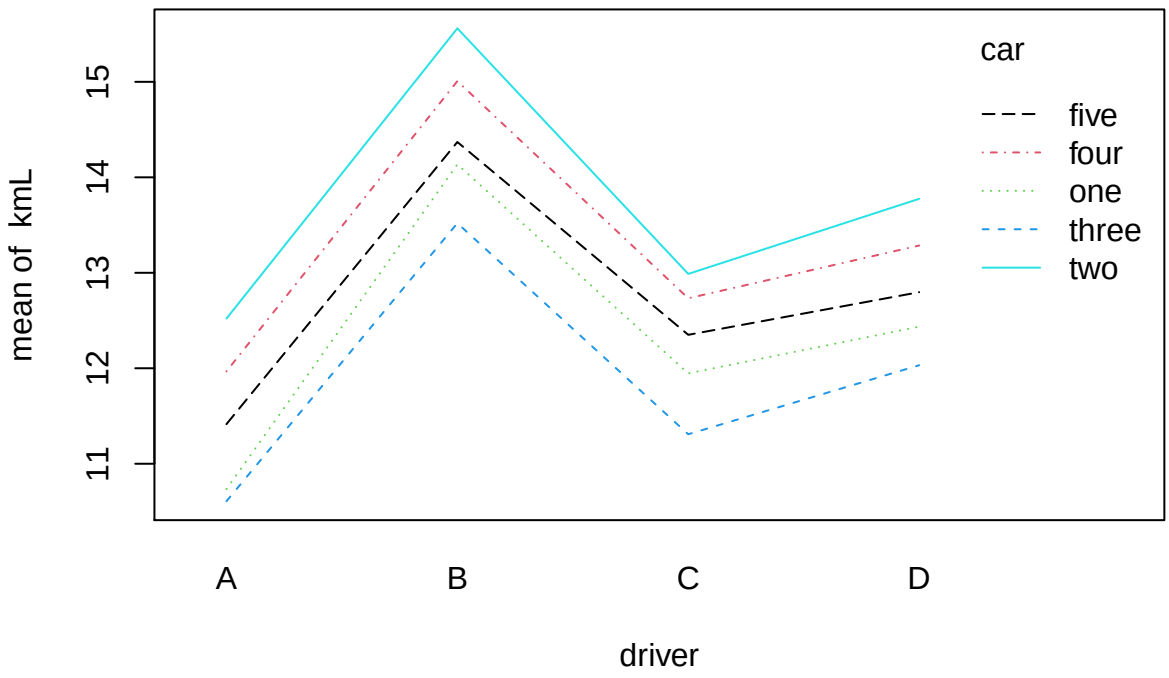
b.

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



With each car and driver, the differences vary in the range from 10 to 16 kmL. Moreover, means of each observation slightly vary while the spread are moderately similar.

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



This plot is able to show clearly the difference among drivers and cars on Kml. While driver A and C seem to share the same fuel efficiency from 10 to 13, driver B shows the greatest efficiency, from 13 to 16. Cars 2, 4, 5, 1 and 3 show the order of efficiency from greatest to lowest, respectively.

c.
Because the question requires to test the interaction between factors, it is relevant to apply Two-way ANOVA.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \epsilon_{ijk} \sim \mathsf{N}(0, \sigma^2)$$

μ = overall population mean.
α_i : base effect of i^{th} level of Factor driver: i = 1,2,. . . ,a.
β_j : base effect of j^{th} level of Factor car: j = 1,2,. . . ,b.
γ_{ij} : effect of the combined effect of the i^{th}, j^{th} combination of the two factors.
ε_{ijk} : unexplained variation for each replicated observation, $\varepsilon \sim \mathsf{N}(0, \sigma^2)$.

- Hypothesis about the interaction term:
  H_0 : $\gamma_{ij} = 0$ for all i,j ; H_A : not all $\gamma_{ij} = 0$

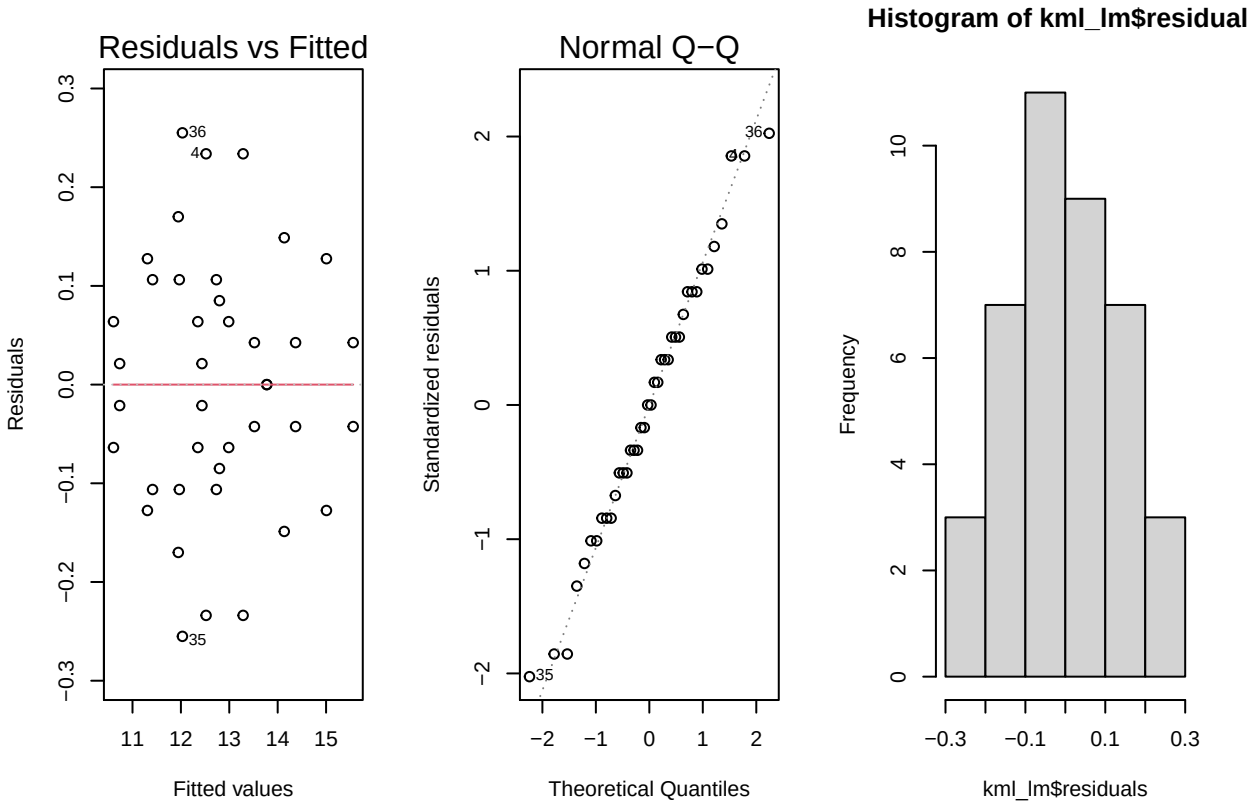
```
kml_lm = lm(kmL ~ driver * car, data = kml)
anova(kml_lm)
```

```
## 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 is not significant (0.3715>0.05). Therefore, we remove the interaction term.

```
par(mfrow=c(1,3))
plot(kml_lm, which = 1:2)
hist(kml_lm$residuals)
```



The residuals vs fitted plot show evenly distributed data points. The QQ plot indicates a straight line and the histogram shows a slightly right-skewed bar graph, but it can be considered to be normally distributed.

d.

The fuel efficiency reaches its greatest with car two and driver B, while reaches it lowest with car three and driver A. As we can see from the model and the graphs, both car and driver can affect the fuel efficiency.