Piyakorn Munegan (Mona) StudentID: 46318461

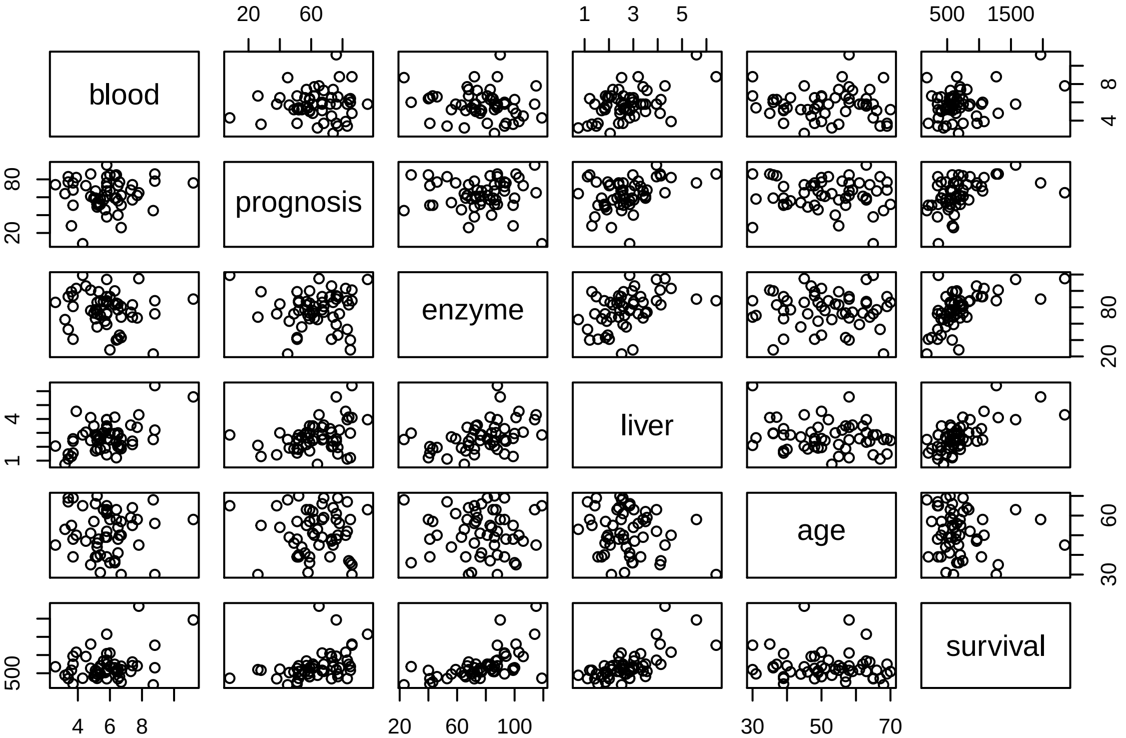
Assignment Semester 1

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. For each patient in the study, the following variables were recorded:

blood	Blood clotting Index
prognosis	Prognosis Index
enzyme	Enzyme function Index
liver	Liver function Index
age	Age of the patient, in years
gender	Gender of the patient, (Male of Female)
survival	Survival time of the patient after surgery (in days)

a. Produce a scatterplot of the data and comment on the features of the data and possible relationships between the response and predictors and relationships between the predictors themselves.



There are moderate correlation between survival and enzyme and liver. Slight correlation between survival and prognosis.

Why it is necessary to remove the gender variable to compute the correlation matrix?

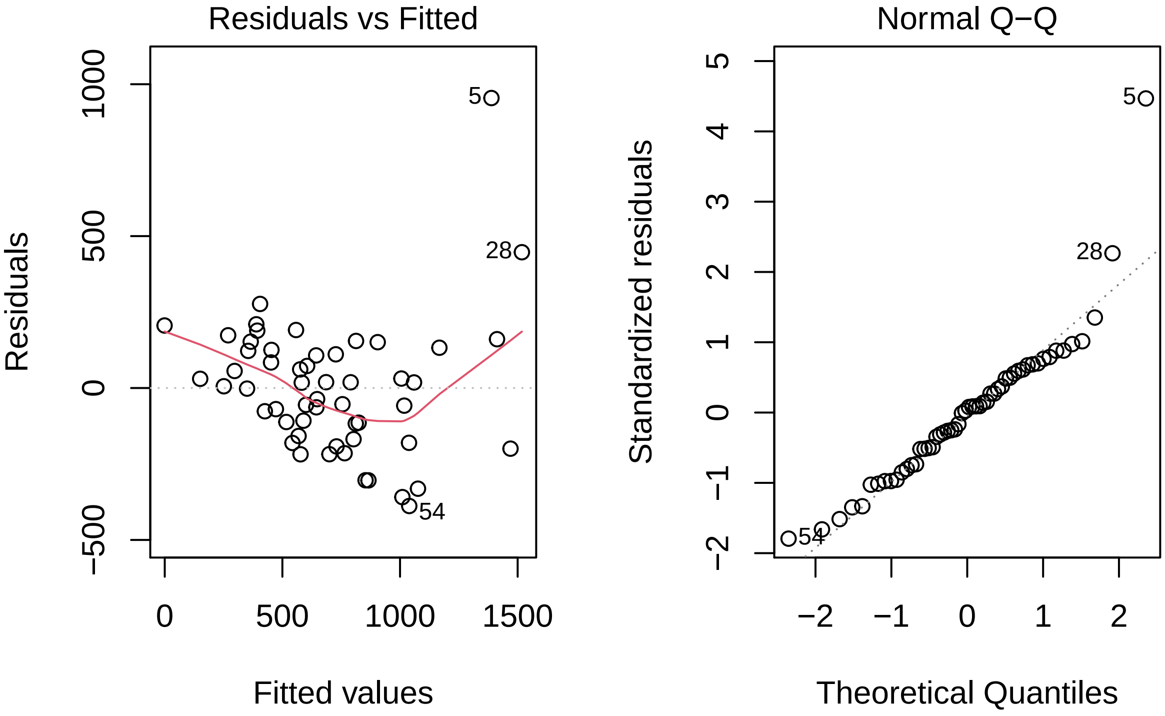
Because the gender variable is a categorical variables. To compute the correlation matrix, every variables must be numeric.

b. Compute the correlation matrix of the dataset and comment.

```
##           blood prognosis enzyme liver   age survival
## blood      1.00      0.09  -0.15  0.50 -0.02     0.35
## prognosis  0.09      1.00  -0.02  0.37 -0.05     0.42
## enzyme    -0.15     -0.02   1.00  0.42 -0.01     0.58
## liver      0.50      0.37   0.42  1.00 -0.21     0.67
## age       -0.02     -0.05  -0.01 -0.21  1.00    -0.12
## survival   0.35      0.42   0.58  0.67 -0.12     1.00
```

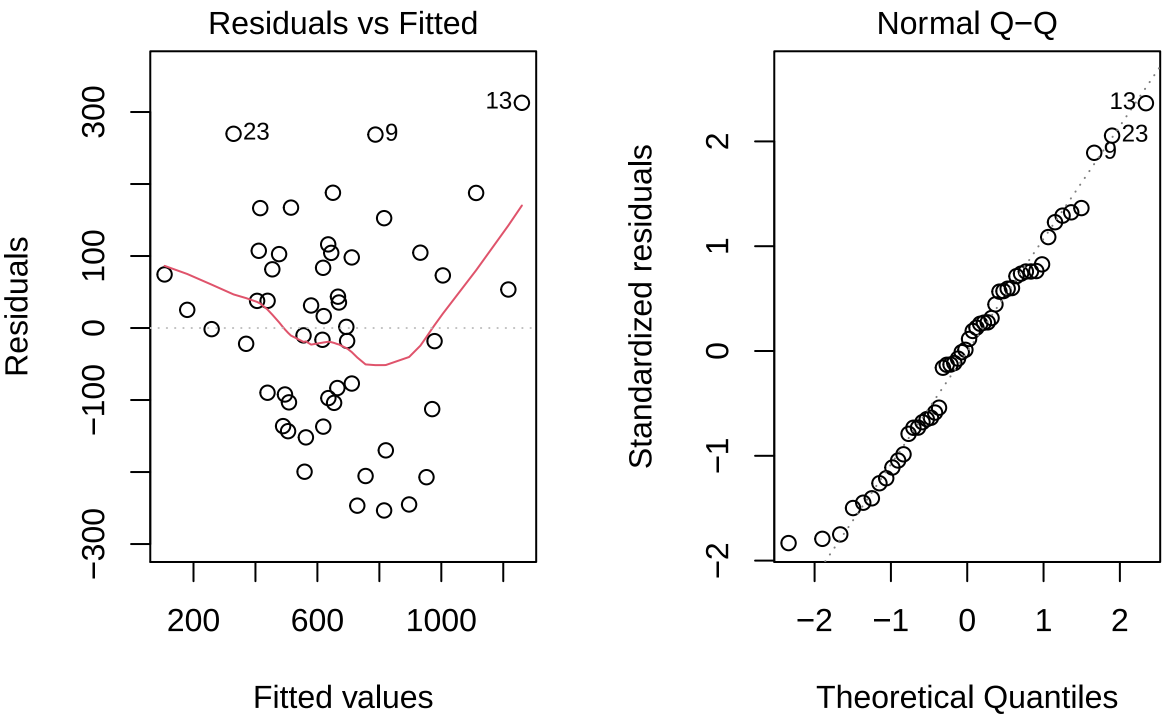
The correlation matrix shows that there are moderate correlation between survival and liver(0.67) and enzyme(0.58). Low correlation between survival and blood(0.35) and prognosis(0.42). The correlation between survival and age is -0.12, which is close to 0, indicates that no linear relationship between these variables.

c. Fit a model using all the predictors to explain the survival response



5 and 28 are significant outliers, need to be remove.

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



After cleaning outliers, The normal Q-Q plot looks better as each point falls quite close to the line. Although there is a pattern in residuals vs fitted plot.

```
## (Intercept)       blood   prognosis      enzyme       liver         age 
## -595.240879   31.568219    8.059842    8.087624   31.028299   -2.316786 
##      gender 
##  -42.186475 
```

- Write down the mathematical multiple regression model for this situation, defining all appropriate parameters.

$$\hat{survival} = -595.240879 + 31.568219 blood + 8.059842 prognosis + 8.087624 enzyme$$
$$+31.028299 liver - 2.316786 age - 42.186475 gender$$

- Write down the Hypotheses for the Overall ANOVA test of multiple regression.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0;$$
$$H_1 : \beta_i \neq 0 \quad \text{for at least one i (not all } \beta_i \quad \text{parameters are zero)}$$

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

```
## Analysis of Variance Table
## 
## Response: survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)    
## blood      1   12670   12670  0.5765    0.4517    
## prognosis  1 1118269 1118269 50.8816 6.470e-09 ***
## enzyme     1 1692481 1692481 77.0085 2.637e-11 ***
## liver      1   58863   58863  2.6783    0.1087    
## age        1   28117   28117  1.2793    0.2640
## gender     1   20450   20450  0.9305    0.3399
## Residuals 45  989003   21978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Compute the F statistic for this test.

$$FullRegSS = RegSS_{blood} + RegSS_{prognosis|blood} + RegSS_{enzyme|blood\&prognosis}$$

$$+RegSS_{liver|blood\&prognosis\&enzyme}+RegSS_{age|blood\&prognosis\&enzyme\&liver}+RegSS_{gender|blood\&prognosis\&enzyme\&liver\&}$$

$$FullRegSS = 12670 + 1118269 + 1692481 + 58863 + 28117 + 20450 = 2930850$$

$RegMS = \frac{RegSS}{k} = \frac{2930850}{6} = 488475$

Test statistic: $F_{obs} = \frac{RegMS}{ResMS} = \frac{488475}{21978} = 22.22563$

- State the Null distribution.

$$H_0 : \beta_{blood} = \beta_{prognosis} = \beta_{enzyme} = \beta_{liver} = \beta_{age} = \beta_{gender} = 0;$$
$$H_1 : \text{not all} \quad \beta_i = 0$$

- Compute the P-Value

```
## [1] 5.788194e-12
```

P-Value: $P(F_{6,45} >= 22.22563) = 5.788194e - 12 < 0.05$

- State your conclusion (both statistical conclusion and contextual conclusion).
  P-value is 5.788194e-12, Reject H0.
- There is a significant linear relationship between survival and at least one of the five predictor variables.

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

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -595.240879 196.435416 -3.0302116 4.041342e-03
## blood         31.568219  19.082773  1.6542783 1.050307e-01
## prognosis      8.059842   1.375161  5.8610151 5.019129e-07
## enzyme         8.087624   1.312128  6.1637443 1.783352e-07
## liver         31.028299  33.030251  0.9393903 3.525456e-01
## age           -2.316786   1.935688 -1.1968799 2.376209e-01
## gender       -42.186475  43.734458 -0.9646050 3.398981e-01
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value).
Remove liver P-value = 0.35255

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -632.140041 192.222012 -3.288593 1.935157e-03
## blood         42.206773  15.339322  2.751541 8.457038e-03
## prognosis      8.649116   1.222188  7.076747 6.991298e-09
## enzyme         8.824440   1.050543  8.399885 7.663650e-11
## age           -2.857956   1.845612 -1.548514 1.283519e-01
## gender       -53.049188  42.124109 -1.259355 2.142552e-01
```

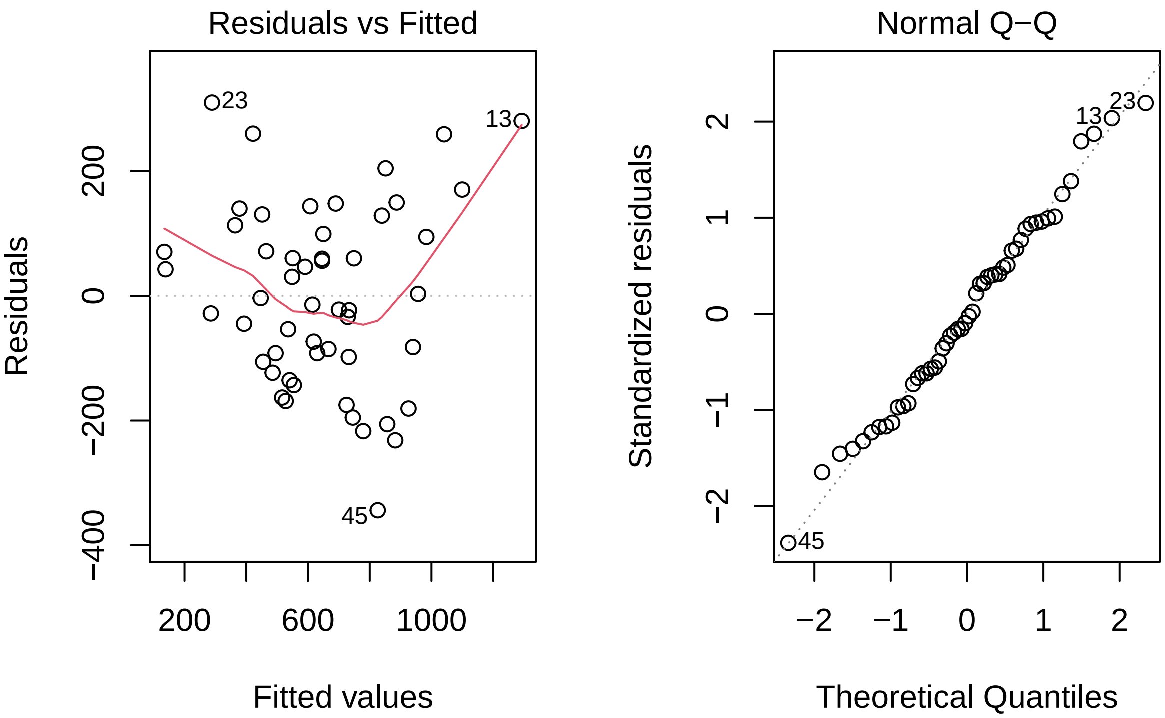
Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove gender P-value = 0.21426

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -693.564656 187.086448 -3.707188 5.520570e-04
## blood         42.987489  15.422037  2.787407 7.642694e-03
## prognosis      8.833336   1.220943  7.234846 3.602947e-09
## enzyme         9.058673   1.040372  8.707145 2.290531e-11
## age           -2.876862   1.857020 -1.549182 1.280461e-01
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove age P-value = 0.128046

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -857.150664 156.673701 -5.470929 1.594387e-06
## blood         44.191481  15.625441  2.828175 6.810261e-03
## prognosis      8.934785   1.236837  7.223897 3.338716e-09
## enzyme         9.083856   1.055306  8.607792 2.698794e-11
```

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



From above picture show the residual vs fitted plot and the normal Q-Q plot for the final model of the survival response(surg.lm4). The linearity of the points in the normal Q-Q plot suggests that the data are close to normally distributed. However, the residuals vs fitted plot has a fan pattern. This is the reason why this final model is not appropriate to use the multiple regression model to explain the survival time. In this case, In this transformation the response variable is needed.

f. Re-fit the model using log(survival) as the new response variable. In your answer,

- Use the model selection procedure discussed in the course starting with log(survival) as the response and start with all the predictors.

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  4.425952674 0.308684319 14.3381196 2.161848e-18
## blood        0.064057552 0.029987224  2.1361614 3.814085e-02
## prognosis    0.012776384 0.002160969  5.9123413 4.212536e-07
## enzyme       0.014605843 0.002061917  7.0836241 7.660019e-09
## liver       -0.007667289 0.051904696 -0.1477186 8.832248e-01
## age         -0.004908609 0.003041796 -1.6137207 1.135796e-01
## gender      -0.087503259 0.068725597 -1.2732266 2.094758e-01
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove liver P-value = 0.8832

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  4.435070690 0.299216888 14.822261 3.890930e-19
## blood        0.061428698 0.023877516  2.572659 1.338436e-02
## prognosis    0.012630771 0.001902484  6.639094 3.174326e-08
## enzyme       0.014423771 0.001635298  8.820274 1.883443e-11
## age         -0.004774882 0.002872919 -1.662031 1.033073e-01
## gender      -0.084819014 0.065571287 -1.293539 2.022827e-01
```

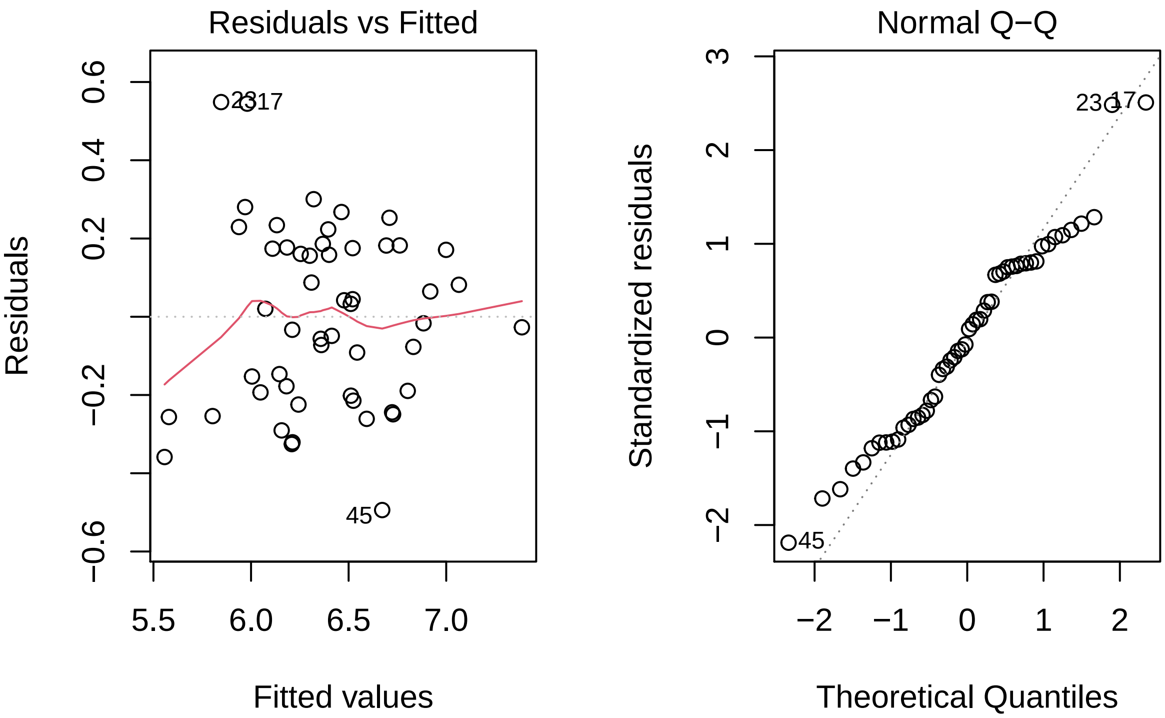
Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove gender P-value = 0.2023

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  4.336860417 0.291489678 14.878264 2.103075e-19
## blood        0.062676965 0.024028275  2.608467 1.215567e-02
## prognosis    0.012925314 0.001902288  6.794614 1.675317e-08
## enzyme       0.014798281 0.001620950  9.129386 5.563495e-12
## age         -0.004805111 0.002893327 -1.660756 1.034202e-01
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove age P-value = 0.1034

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 4.06362903 0.244988435 16.587024 1.649035e-21
## blood       0.06468795 0.024433280  2.647534 1.093678e-02
## prognosis   0.01309476 0.001934025  6.770730 1.652388e-08
## enzyme      0.01484034 0.001650167  8.993236 7.255893e-12
```

g. Validate your final model with the log(survival) response. In particular, in your answer,



- Explain why the regression model with log(survival) response variable is superior to the model with the survival response variable
  From above picture show the residual vs fitted plot and the normal Q-Q plot for the log(survival) response. Comparing with the final model of the survival response, the normality assumption within

the log(survival) model is better as this can be shown in the residuals vs fitted plot.

The normal Q-Q plot has some linearity but not the fitted value one….

Overall, by comparing the multiple regression assumptions of both log(survival) and survival response, it clarifies the reason why log(survival) is superior to the other response.

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.

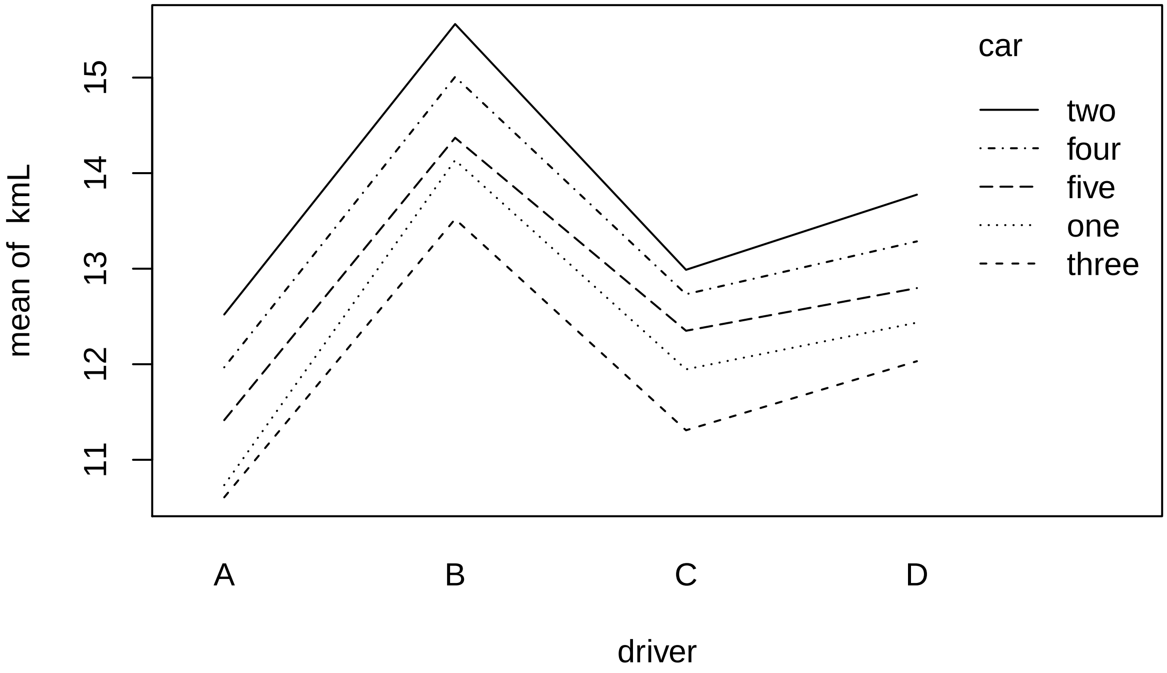
kmL	The observed efficiency of the car in km/L over a standard course
car	The specific car (labelled 1, 2, 3, 4 or 5)
driver	The driver of the car (labelled A, B, C, D)

a. For this study, is the design balanced or unbalanced? Explain why.

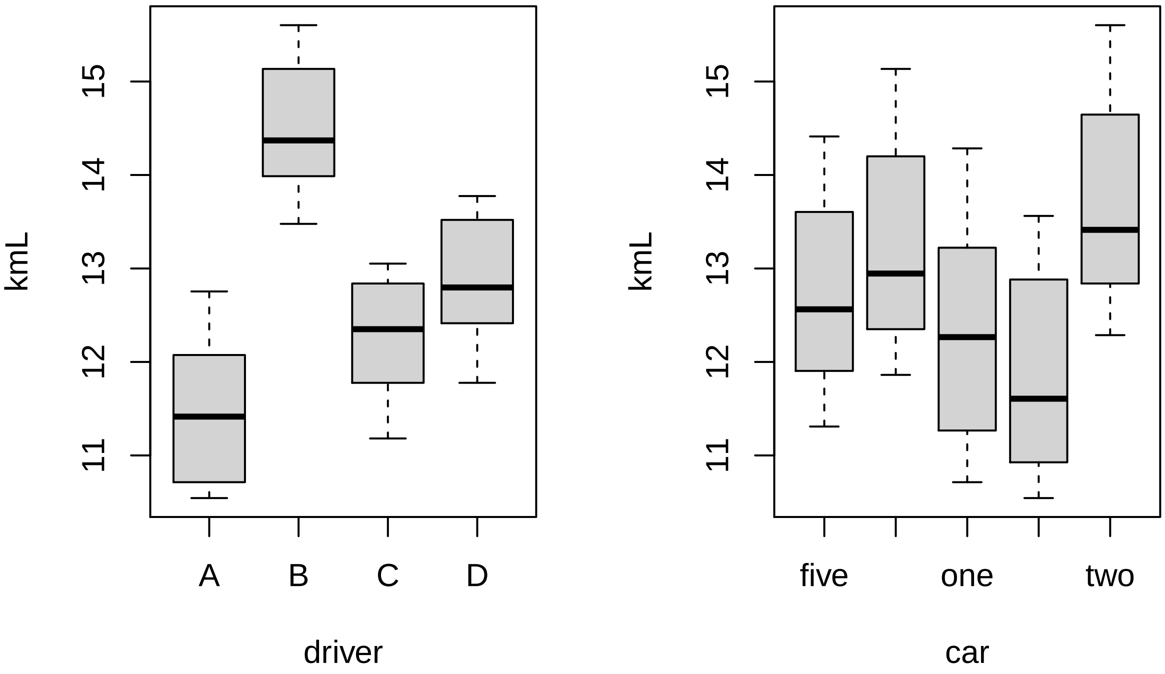
```
##        car
## driver five four one three two
##      A    2    2   2     2   2
##      B    2    2   2     2   2
##      C    2    2   2     2   2
##      D    2    2   2     2   2
```

This is a balanced design because there is the same no. of replicates for each treatment combinations.

b. Construct two different preliminary graphs that investigate different features of the data and comment.



As the lines are not parallel, interaction could be there.



For the boxplots, there are similar spread for driver and car.

c. Analyse the data, stating null and alternative hypothesis for each test, and check assumptions.

```
##             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
```

Model: $Y = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$

where ϵ_{ijk} are $N(0, \sigma^2)$ random variables

μ : overall population mean

α_i : main effect on driver

β_j : main effect on car

γ_{ij} : interaction effect between driver and car

ϵ_{ijk} : error term

Hypotheses $\quad H_0 : \gamma_{ij} = 0 \quad$ against $\quad H_1$: at least one $\quad \gamma_{ij}$ non-zero

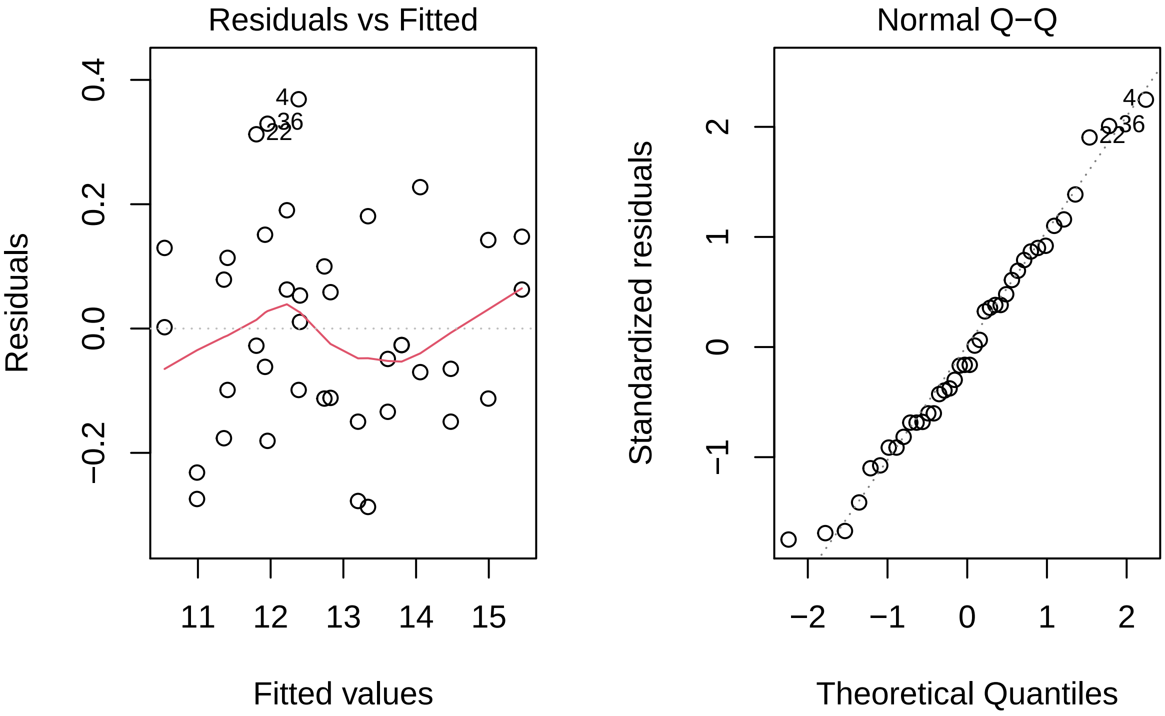
Because P-value $= 0.371 > 0.05$, $\quad \gamma_{ij}$ is not significant.
No evidence to suggest that the two factors (driver and car) are not independent.
As interaction is not significant, re-fit the model with main effects only.

```
##             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
```

Model: $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

Hypotheses : $\quad H_0 : \beta_j = 0 \quad$ against $\quad H_1$: at least one $\quad \beta_j$ non-zero

Both the driver and car effects are highly significant (P-Value < 0.001)



- Residuals vs Fitted plot shows a negligible pattern, variability among residuals vs fitted is not constant. There are several outliers, with residuals close to 0.4.
- The normal Q-Q plot of residuals follows a linear trend, residuals look close to normally distributed.

d. State your conclusions about the effect of driver and car on the efficiency kmL. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests in
e. and the preliminary plots in b.. You do not need to statistically examine the multiple comparisons between contrasts and interactions.

The p-value for the effects of driver and car are less than the significant level 0.05, we have evidence to reject H_0. Moreover, the normal Q-Q plot follows a linear trend and residual plots have no pattern, suggesting linear model adequate.