Piyakorn Munegan (Mona) StudentID: 46318461

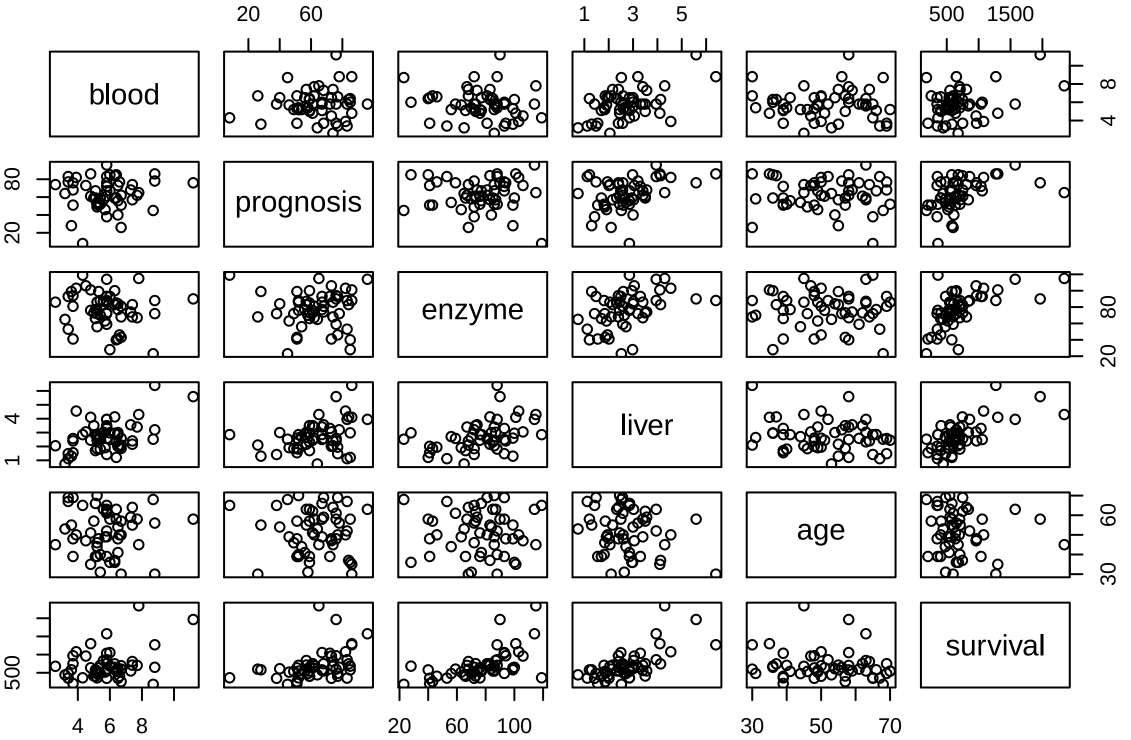
Assignment Semester 1

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?

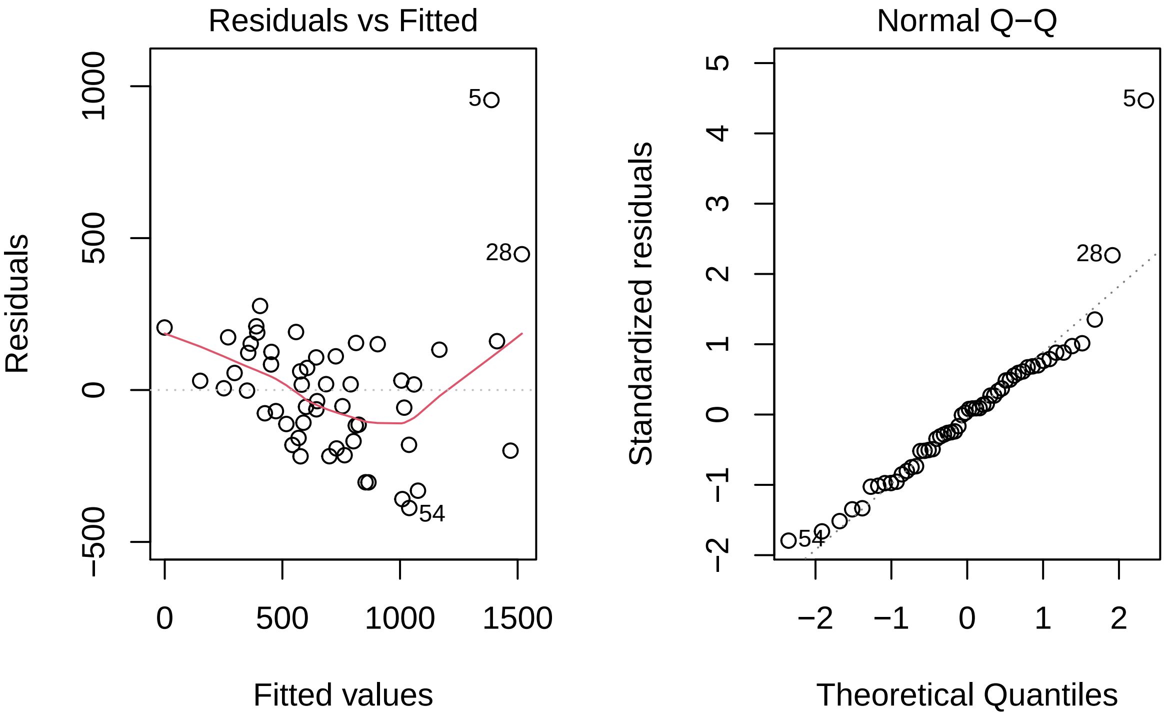
Because the gender variable is a categorical variables. To compute the correlation matrix, every variables must be numeric.

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



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

```
##   (Intercept)         blood     prognosis        enzyme         liver 
## -1179.1888797    86.6437068     8.5012606    11.1245627    38.5068155 
##           age        gender 
##    -2.3408883    -0.2201138
```

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

$$\hat{survival} = -1179.1888797 + 86.6437068 blood + 8.5012606 prognosis + 11.1245627 enzyme$$
$$+38.5068155 liver - 0.2201138 age - 0.2201138 gender$$

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

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0;$$
$$H_1 : \beta_i \neq 0 \quad \text{for at least one i (not all } \beta_i \quad \text{parameters are zero)}$$

- 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 1005152 1005152 18.5060 8.502e-05 ***
## prognosis  1 1278496 1278496 23.5385 1.387e-05 ***
## enzyme     1 3442172 3442172 63.3742 2.915e-10 ***
## liver      1   57862   57862  1.0653    0.3073
## age        1   33032   33032  0.6082    0.4394
## gender     1       1       1  0.0000    0.9974
## Residuals 47 2552807   54315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Compute the F statistic for this test.

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

$FullRegSS = 1005152 + 1278496 + 3442172 + 57862 + 33032 + 1 = 5816715$

$RegMS = \frac{RegSS}{k} = \frac{5816715}{6} = 969452.5$

Test statistic: $F_{obs} = \frac{RegMS}{ResMS} = \frac{969452.5}{54315} = 17.84871$

- 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] 1.190218e-10
```

P-Value: $P(F_{6,47} >= 17.84871) = 1.190218e - 10 < 0.05$

- State your conclusion (both statistical conclusion and contextual conclusion).
  P-value is 1.190218e-10, 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.

```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + liver +
##     age + gender, data = surg.new)
##
## 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 ***
## blood          86.6437    27.4920   3.152 0.002825 **
## 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
## gender         -0.2201    67.5146  -0.003 0.997413
## ---
## 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
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove gender P-value = 0.997413

```
##                 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
```

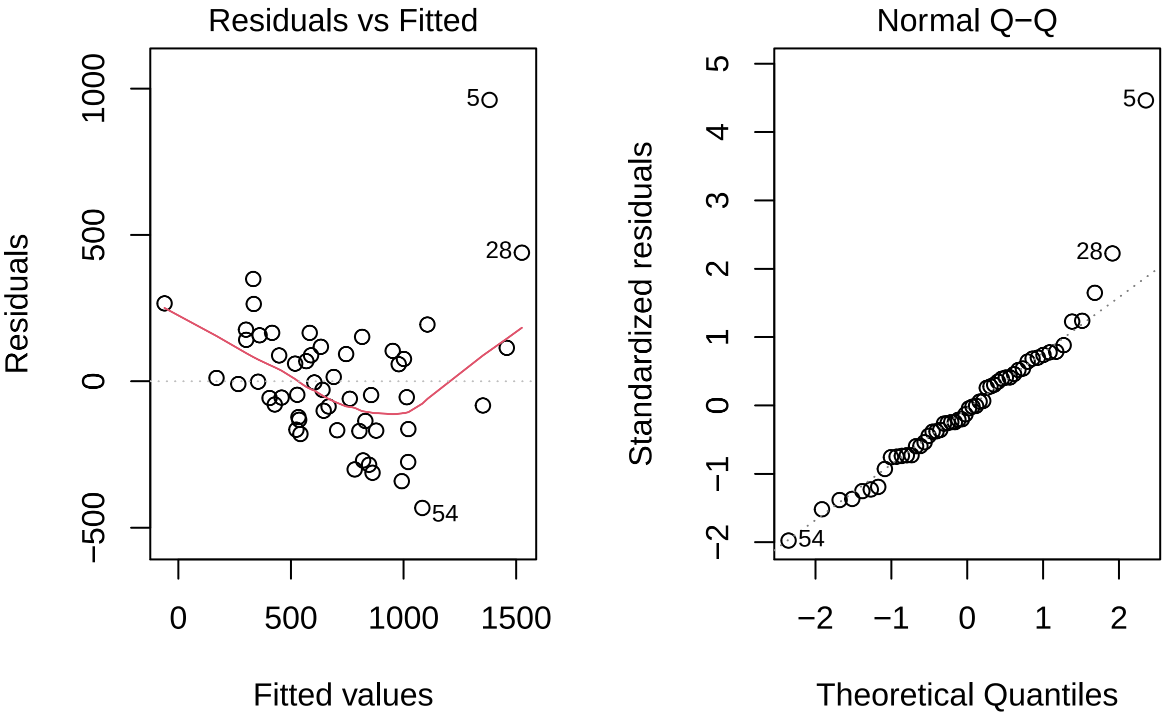
Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove age P-value = 0.4345

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1279.241625 243.808376 -5.246914 3.299829e-06
## blood          82.988246  26.402146  3.143239 2.835553e-03
## prognosis       8.345872   2.119706  3.937278 2.604351e-04
## enzyme         10.869607   1.923218  5.651783 8.012244e-07
## liver          49.346241  47.125921  1.047115 3.001845e-01
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove liver P-value = 0.30018

```
##                 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
```

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)  4.100996997 0.302780511 13.54445497 7.665031e-18
## blood        0.094858430 0.029328278  3.23436749 2.233610e-03
## prognosis    0.013019518 0.002304419  5.64980388 9.084175e-07
## enzyme       0.016245445 0.002114393  7.68326550 7.593288e-10
## liver       -0.003132326 0.055256339 -0.05668718 9.550347e-01
## age         -0.004863398 0.003215405 -1.51253027 1.370946e-01
## gender      -0.066139894 0.072024114 -0.91830208 3.631495e-01
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove liver P-value = 0.95503

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  4.105132163 0.290794570 14.1169492 1.039336e-18
## blood        0.093737869 0.021439405  4.3722235 6.576975e-05
## prognosis    0.012959509 0.002025519  6.3981183 6.163295e-08
## enzyme       0.016170082 0.001626984  9.9386870 3.096659e-13
## age         -0.004810137 0.003042976 -1.5807348 1.205066e-01
## gender      -0.065009685 0.068487175 -0.9492242 3.472615e-01
```

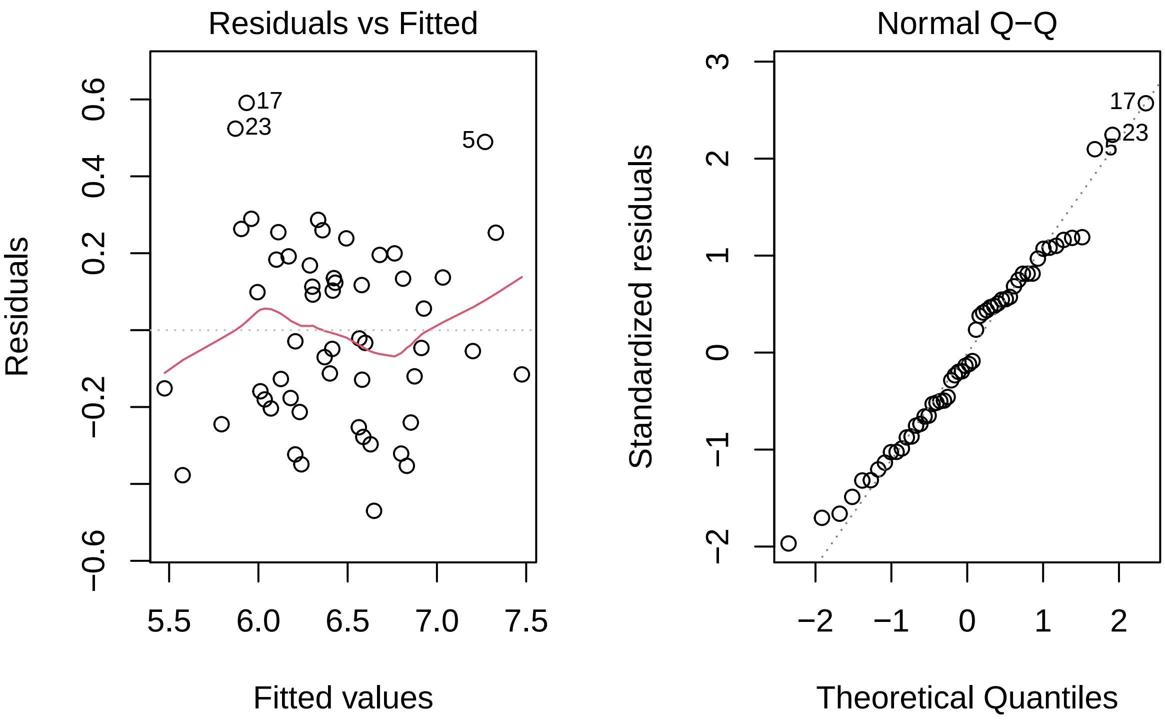
Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove gender P-value = 0.3472615

```
##                 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
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove age P-value = 0.1232646

```
##               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
```

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 is close to normally distributed.

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.

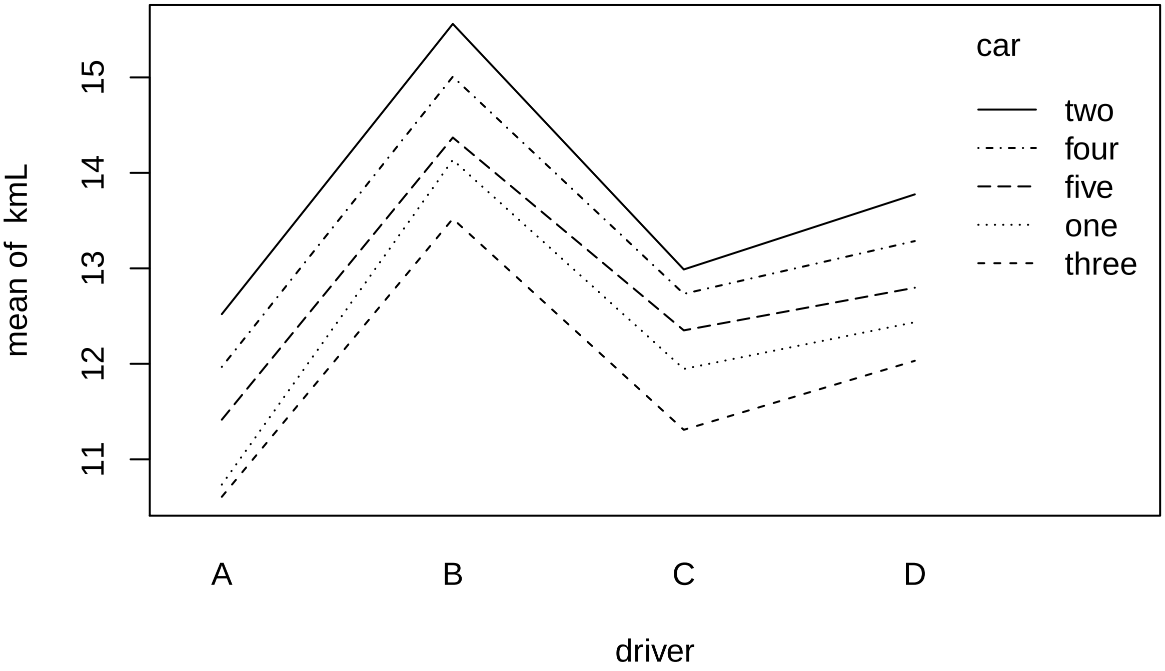
kmL	The observed efficiency of the car in km/L over a standard course
car	The specific car (labelled 1, 2, 3, 4 or 5)
driver	The driver of the car (labelled A, B, C, D)

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

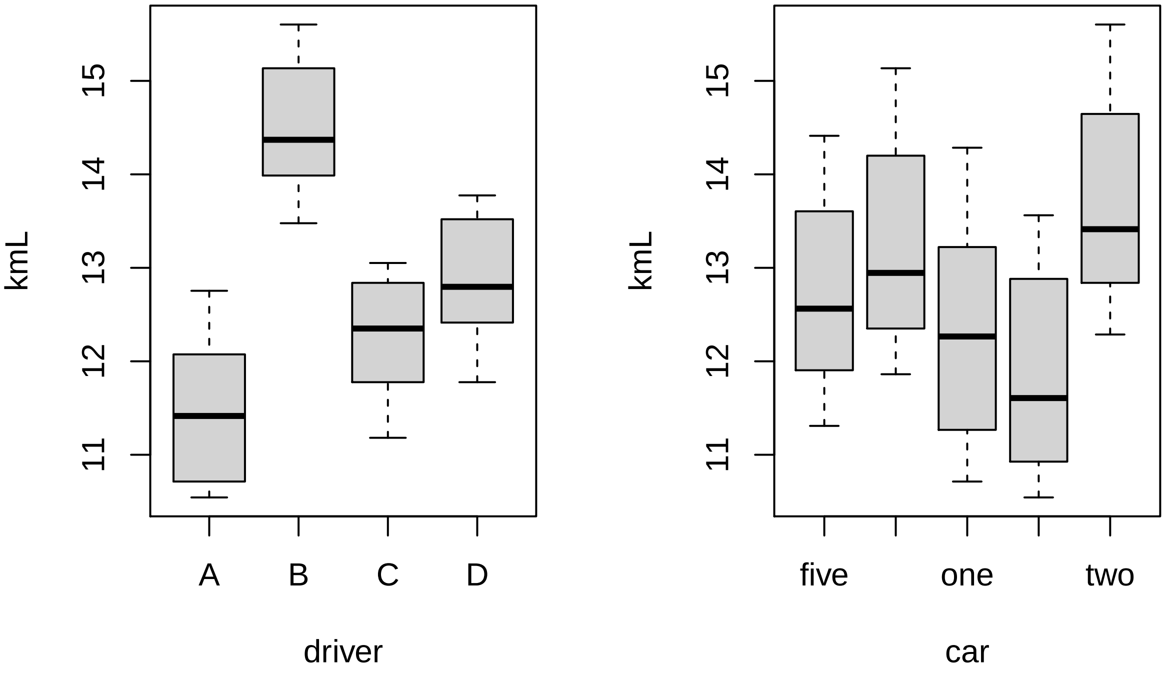
```
##       car
## driver five four one three two
##      A    2    2   2     2   2
##      B    2    2   2     2   2
##      C    2    2   2     2   2
##      D    2    2   2     2   2
```

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

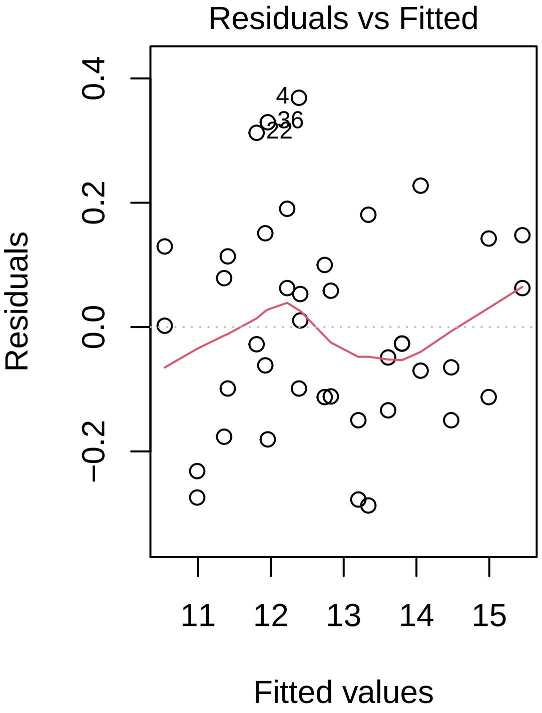
Because P-value $= 0.371 > 0.05, \quad \gamma_{ij}$ is not significant.
No evidence to suggest that the two factors (driver and car) are not independent.
As interaction is not significant, re-fit the model with main effects only.

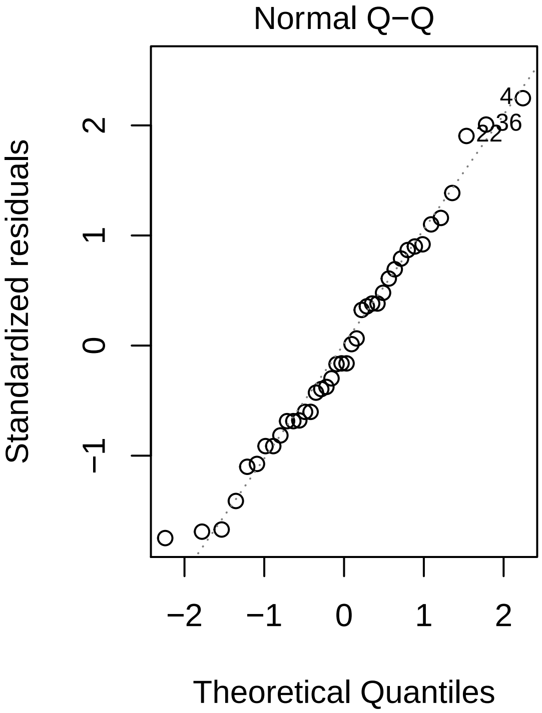
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## driver       3  50.66  16.887   501.5 <2e-16 ***
## car          4  17.12   4.280   127.1 <2e-16 ***
## Residuals   32   1.08   0.034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.