# Piyakorn Munegan (Mona) StudentID: 46318461

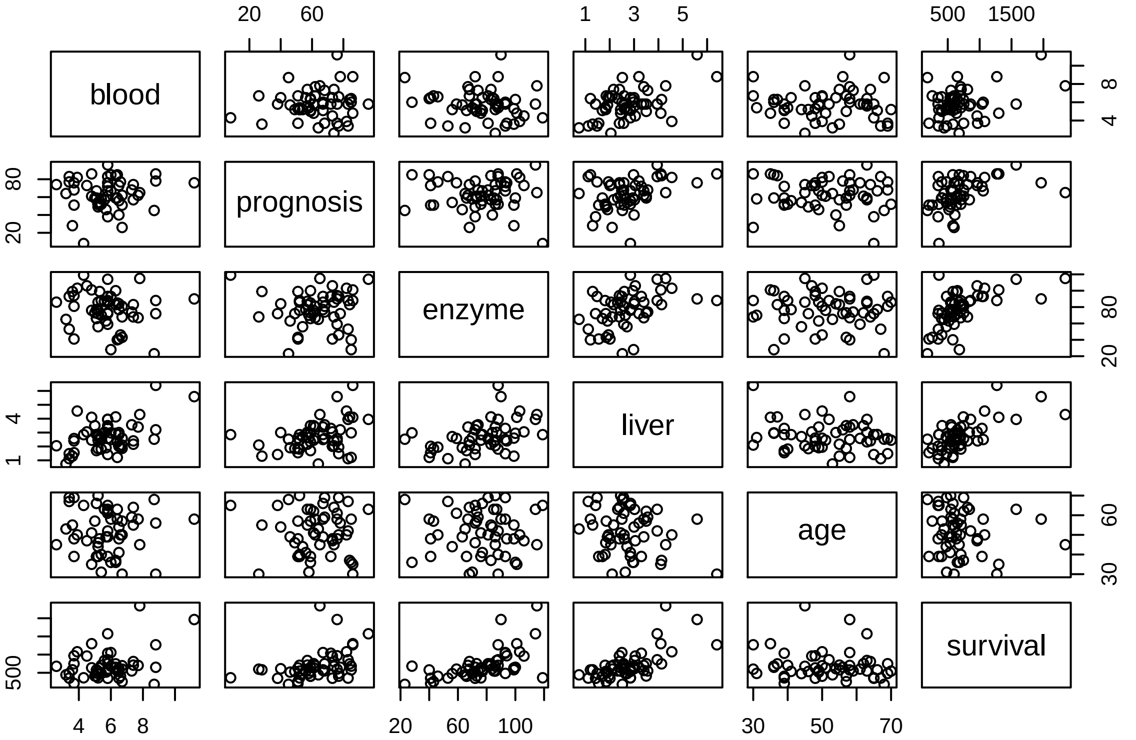
## Assignment Semester 1

### 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 correlatiob matrix shows that there are moderately correlated 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

```
##  (Intercept)        blood    prognosis       enzyme        liver          age 
## -1179.366654    86.630445     8.501113    11.124165    38.553562    -2.339958
```

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

$$\hat{survival} = -630.654734 + 29.257285 blood + 8.036109 prognosis + 8.060826 enzyme$$
$$+39.452521 liver - 2.183841 age$$

- Write down the Hypotheses for the Overall ANOVA test of multiple regression.
  Hypotheses $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ against $H_1 : \beta_i \neq 0$ for at least one i (not all $\beta_i$ 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  0.5774    0.4512    
## prognosis  1 1118269 1118269 50.9587 5.649e-09 ***
## enzyme     1 1692481 1692481 77.1251 2.138e-11 ***
## liver      1   58863   58863  2.6823    0.1083    
## age        1   28117   28117  1.2813    0.2635    
## Residuals 46 1009453   21945                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Compute the F statistic for this test.

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

$FullRegSS = 78517 + 875224 + 1552773 + 56436 + 31888 = 2594838$

$RegMS = \frac{RegSS}{k} = \frac{2594838}{5} = 518967.6$

Test statistic: $F_{obs} = \frac{RegMS}{ResMS} = \frac{518967.6}{23397} = 22.18095$

- State the Null distribution.
  Hypotheses $H_0 : \beta_{blood} = \beta_{prognosis} = \beta_{enzyme} = \beta_{liver} = \beta_{age} = 0$ against $H_1$ : not all $\beta_i = 0$

- Compute the P-Value
  P-Value: $P(F_{5,42} >= 22.18095) = 8.215602e - 11 < 0.05$
- State your conclusion (both statistical conclusion and contextual conclusion).
  P-value is 8.215602e-11, 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, data = surg.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -262.96 -108.97   10.11  102.20  325.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -630.655    192.828  -3.271  0.00204 **
## blood         29.257     18.917   1.547  0.12882
## prognosis      8.036      1.374   5.849 4.87e-07 ***
## enzyme         8.061      1.311   6.149 1.73e-07 ***
## liver         39.453     31.831   1.239  0.22147
## age           -2.184      1.929  -1.132  0.26353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 148.1 on 46 degrees of freedom
## Multiple R-squared:  0.7425, Adjusted R-squared:  0.7145
## F-statistic: 26.52 on 5 and 46 DF,  p-value: 1.65e-12
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value).
Remove liver P-value = 0.437595

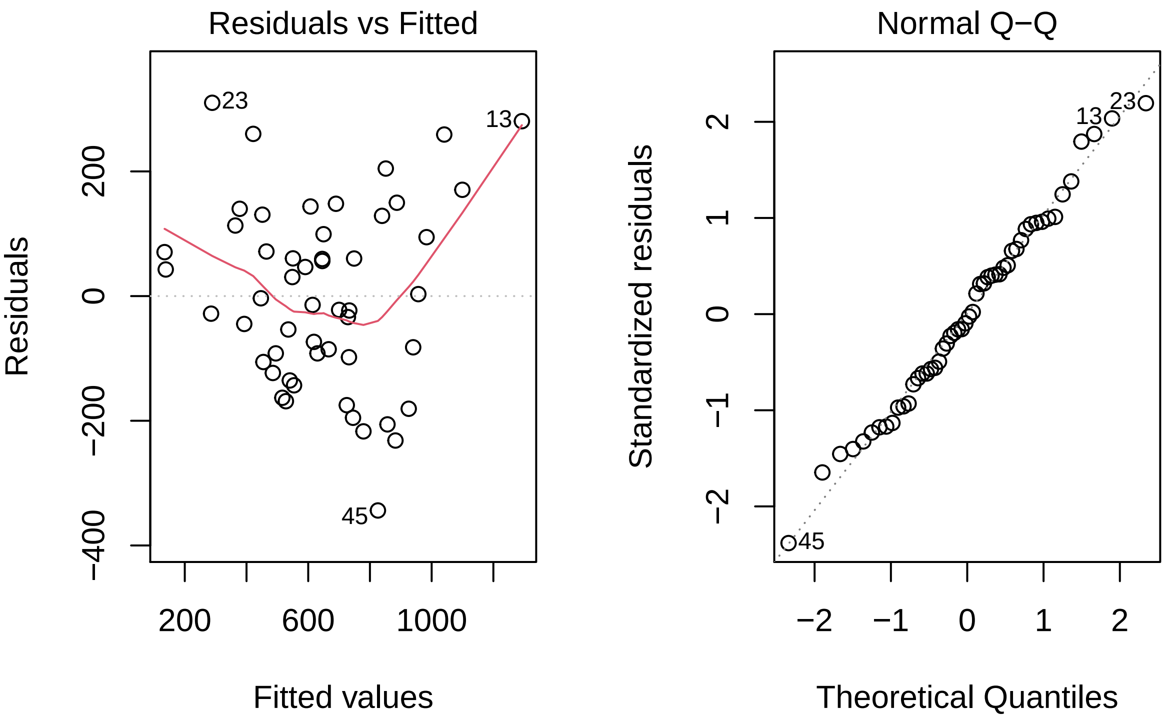
```
##
```

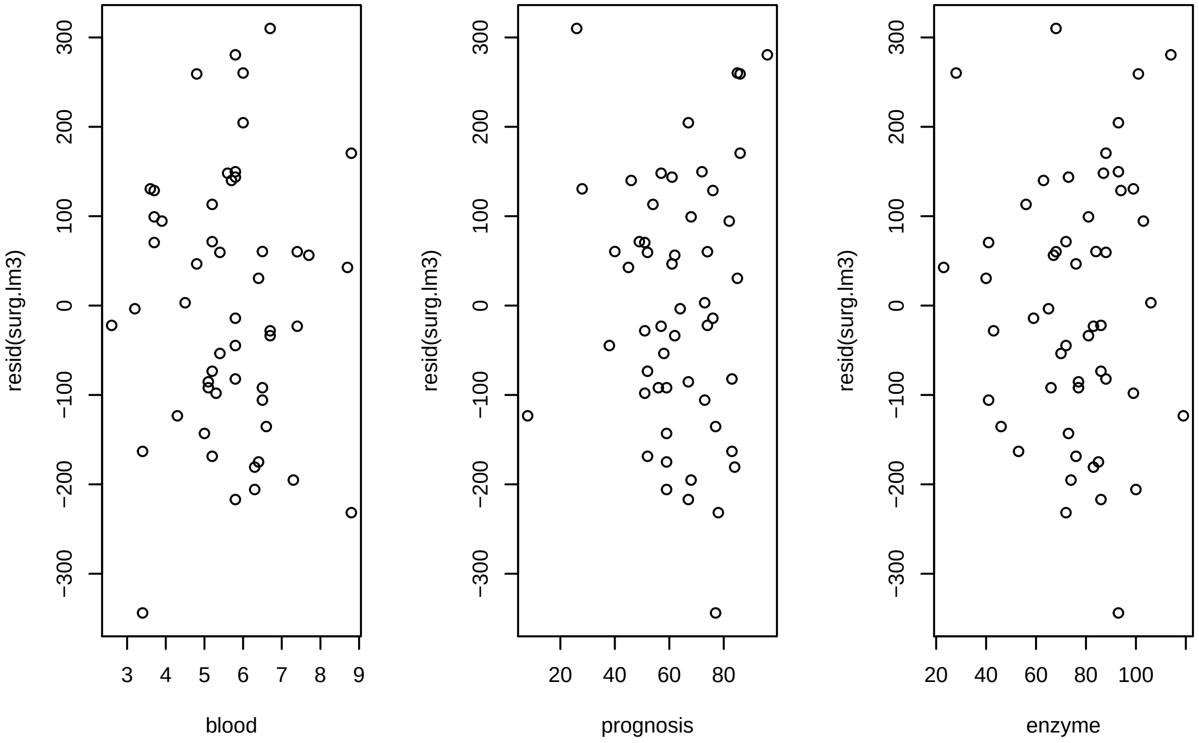
```
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + age, data = surg.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -294.71 -116.28   -5.66  102.88  317.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -693.565    187.086  -3.707 0.000552 ***
## blood         42.987     15.422   2.787 0.007643 **
## prognosis      8.833      1.221   7.235 3.60e-09 ***
## enzyme         9.059      1.040   8.707 2.29e-11 ***
## age           -2.877      1.857  -1.549 0.128046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149 on 47 degrees of freedom
## Multiple R-squared:  0.7339, Adjusted R-squared:  0.7112
## F-statistic:  32.4 on 4 and 47 DF,  p-value: 5.632e-13
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value).
Remove age P-value = 0.298

```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme, data = surg.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -343.88  -99.97   -8.85  102.76  310.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -857.151    156.674  -5.471 1.59e-06 ***
## blood         44.191     15.625   2.828  0.00681 **
## prognosis      8.935      1.237   7.224 3.34e-09 ***
## enzyme         9.084      1.055   8.608 2.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151.1 on 48 degrees of freedom
## Multiple R-squared:  0.7203, Adjusted R-squared:  0.7028
## F-statistic:  41.2 on 3 and 48 DF,  p-value: 2.526e-13
```

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





For the final model(surg.lm3):

1. The Normal Q-Q plot of residuals has slight curvature but close to normally distributed. ???
2. The residuals vs fitted shows fan pattern. ???
3. Residuals vs predictor plots ???

So Transform response. . . .

### Transformation

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.

```
##   blood prognosis enzyme liver age survival logsurvival logblood logprognosis
## 1   6.7        62     81  2.59  50      695    6.543912 1.902108     4.127134
## 2   5.1        59     66  1.70  39      403    5.998937 1.629241     4.077537
## 3   7.4        57     83  2.16  55      710    6.565265 2.001480     4.043051
## 4   6.5        73     41  2.01  48      349    5.855072 1.871802     4.290459
## 6   5.8        38     72  1.42  65      348    5.852202 1.757858     3.637586
## 7   5.7        46     63  1.91  49      518    6.249975 1.740466     3.828641
##   logenzyme  logliver   logage
## 1  4.394449 0.9516579 3.912023
## 2  4.189655 0.5306283 3.663562
## 3  4.418841 0.7701082 4.007333
## 4  3.713572 0.6981347 3.871201
## 6  4.276666 0.3506569 4.174387
## 7  4.143135 0.6471032 3.891820
```

```
## 
## Call:
## lm(formula = logsurvival ~ logblood + logprognosis + logenzyme + 
##     logliver + logage, data = surg.new)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.52566 -0.17518  0.04219  0.15401  0.67504 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.85594    1.01284   0.845    0.402    
## logblood      0.17388    0.16448   1.057    0.296    
## logprognosis  0.48491    0.09384   5.167 5.00e-06 ***
## logenzyme     0.84314    0.12919   6.526 4.69e-08 ***
## logliver      0.19499    0.12207   1.597    0.117    
## logage       -0.13792    0.15508  -0.889    0.378    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.25 on 46 degrees of freedom
## Multiple R-squared:  0.7003, Adjusted R-squared:  0.6677 
## F-statistic: 21.49 on 5 and 46 DF,  p-value: 4.99e-11
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove logage P-value = 0.4551

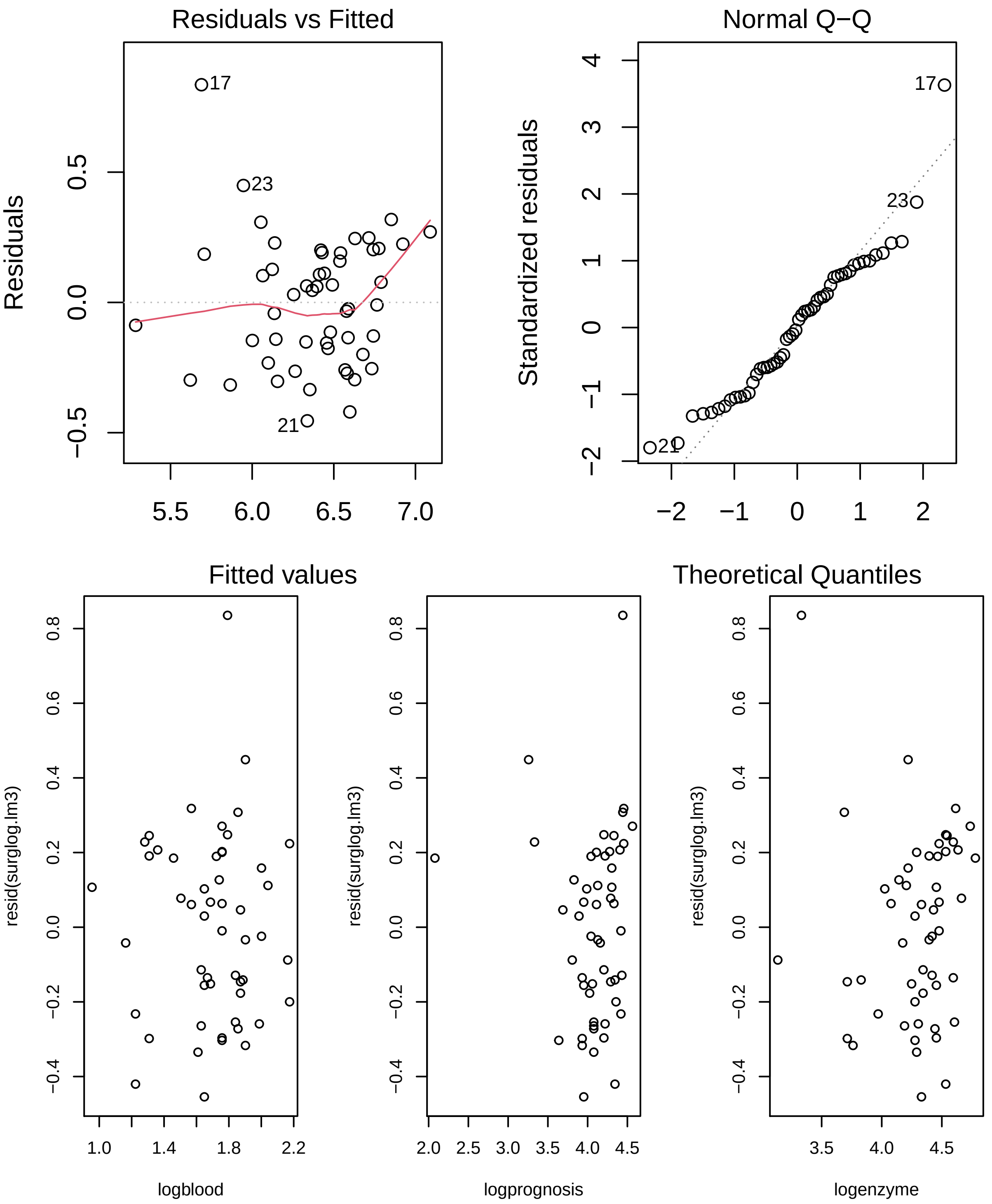
```
## 
## Call:
## lm(formula = logsurvival ~ logblood + logprognosis + logenzyme + 
##     logliver, data = surg.new)
## 
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.49346 -0.17295  0.04336  0.14248  0.70862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.34215    0.83006   0.412   0.6821
## logblood     0.16649    0.16390   1.016   0.3149
## logprognosis 0.48600    0.09362   5.191 4.40e-06 ***
## logenzyme    0.83397    0.12849   6.491 4.85e-08 ***
## logliver     0.21786    0.11906   1.830   0.0736 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2495 on 47 degrees of freedom
## Multiple R-squared:  0.6951, Adjusted R-squared:  0.6692
## F-statistic: 26.79 on 4 and 47 DF,  p-value: 1.308e-11
```

Removing one none significant variable (if there are many none significant vars, pick the largest P-value). Remove logliver P-value = 0.1252

```
##
## Call:
## lm(formula = logsurvival ~ logblood + logprognosis + logenzyme,
##     data = surg.new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45444 -0.18251  0.01012  0.19006  0.83544
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.4763     0.7161  -0.665   0.5092
## logblood      0.3351     0.1388   2.415   0.0196 *
## logprognosis  0.5331     0.0922   5.782 5.40e-07 ***
## logenzyme     0.9595     0.1113   8.623 2.56e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2555 on 48 degrees of freedom
## Multiple R-squared:  0.6734, Adjusted R-squared:  0.653
## F-statistic: 32.99 on 3 and 48 DF,  p-value: 1.01e-11
```

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
  The residuals vs fitted looks better for the log transformed 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.

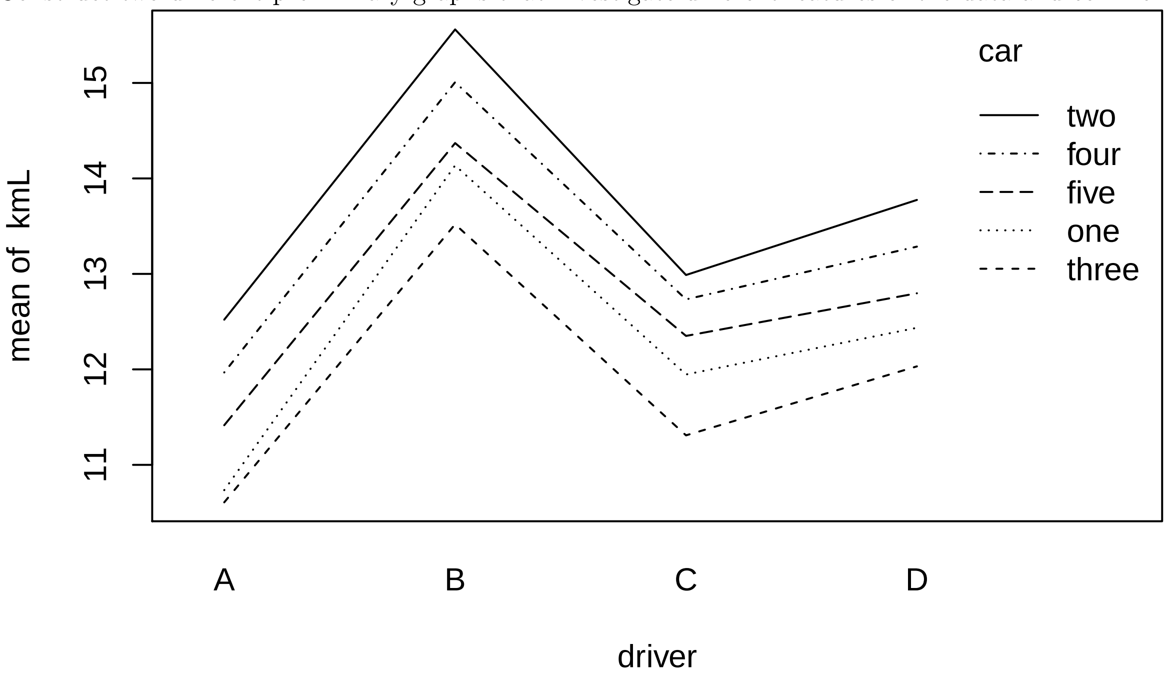
| | |
|---|---|
| kmL | The observed efficiency of the car in km/L over a standard course |
| car | The specific car (labelled 1, 2, 3, 4 or 5) |
| driver | The driver of the car (labelled A, B, C, D) |

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

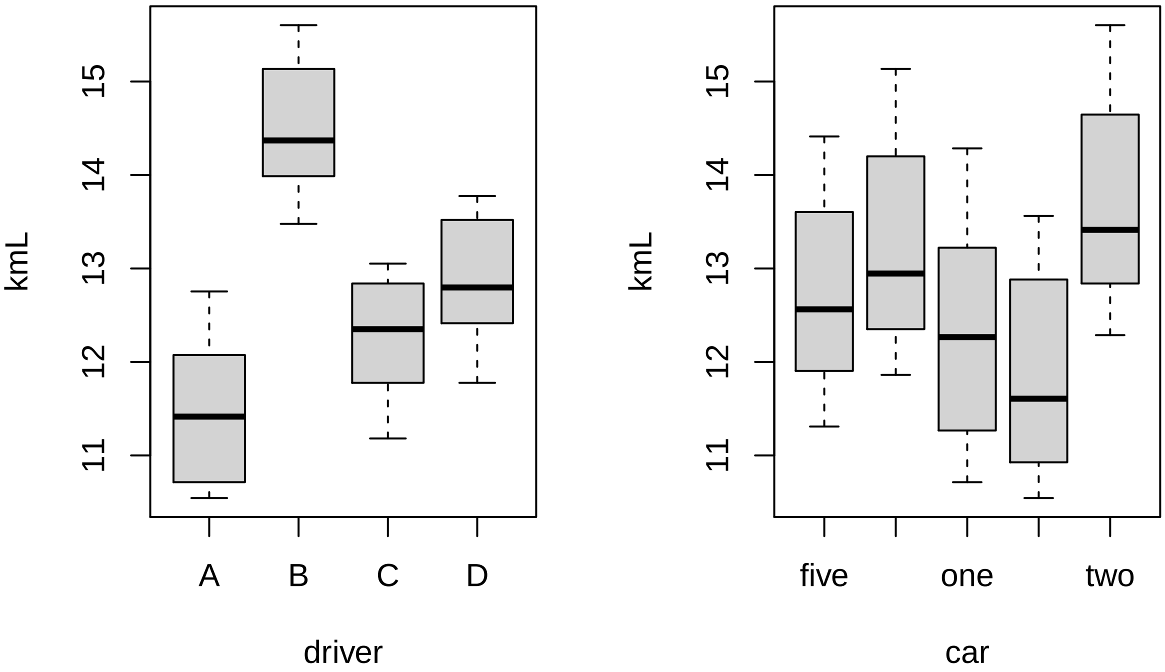
```
##        car
## driver five four one three two
##      A    2    2   2     2   2
##      B    2    2   2     2   2
##      C    2    2   2     2   2
##      D    2    2   2     2   2
```

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

$\mu$ : overall population mean

$\alpha_i$ : main effect on driver

$\beta_j$ : main effect on car

$\gamma_{ij}$ : interaction effect between driver and car

$\epsilon_{ijk}$ : error term

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

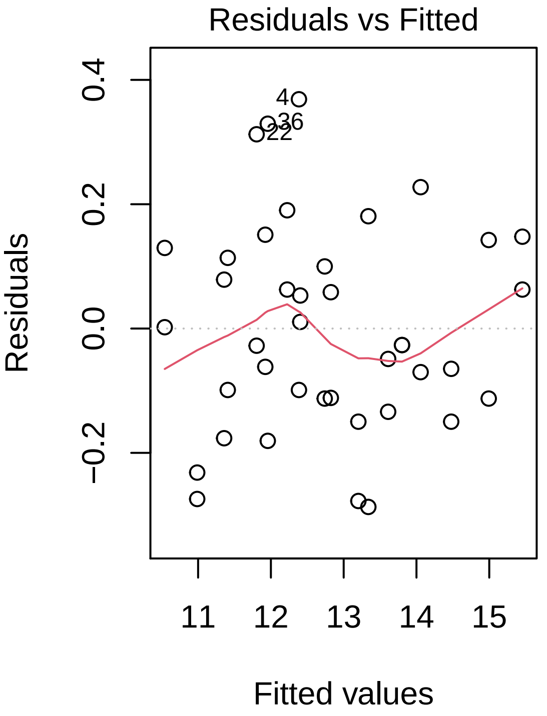
Because P-value $= 0.371 > 0.05$, $\quad \gamma_{ij}$ is not significant.
No evidence to suggest that the two factors (driver and car) are not independent.
As interaction is not significant, re-fit the model with main effects only.

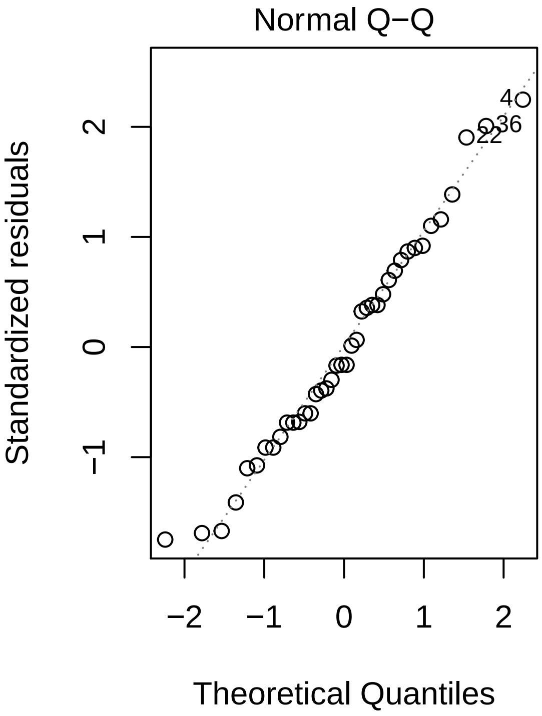
```
##             Df Sum Sq Mean Sq F value Pr(>F)    
## driver       3  50.66  16.887   501.5 <2e-16 ***
## car          4  17.12   4.280   127.1 <2e-16 ***
## Residuals   32   1.08   0.034                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

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





- Residuals vs Fitted plot shows negligible pattern, variability among residuals vs fitted is not constant.
- The quantile 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 $H_0$. and the quantile plot looks linear and residual plots have no pattern, suggesting linear model adequate. ????