Piyakorn Munegan (Mona) StudentID: 46318461

Assignment Semester 1

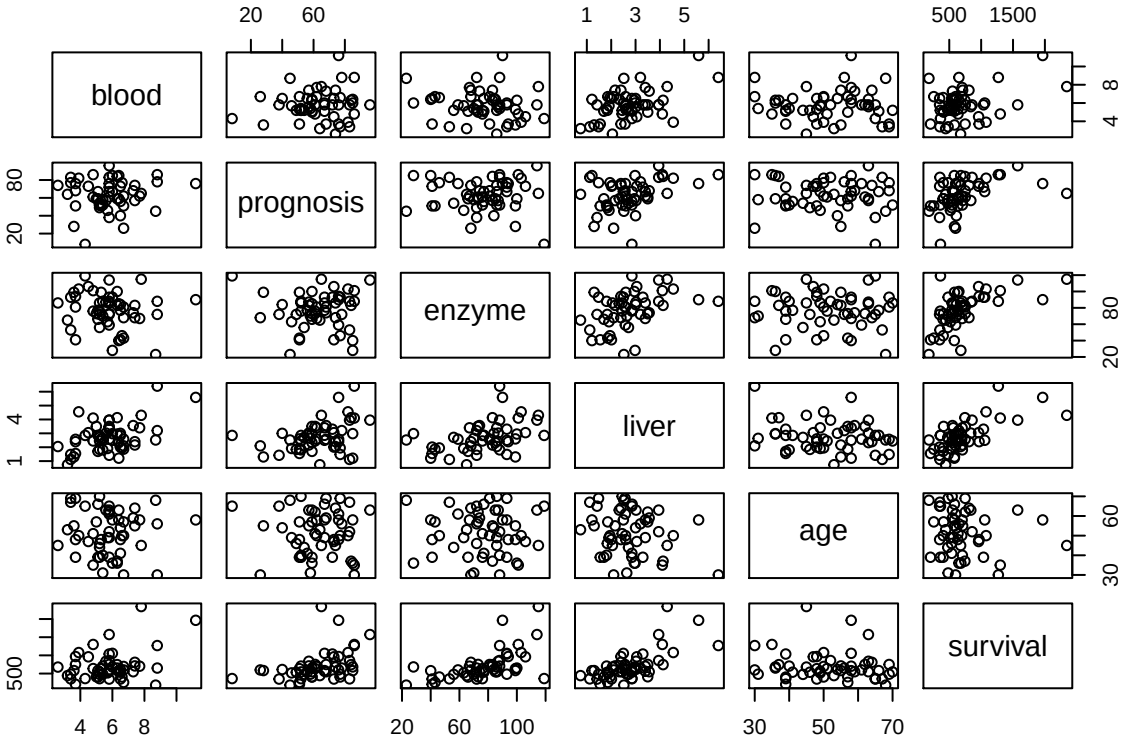
Question 1

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.

```
##   blood prognosis enzyme liver age survival
## 1   6.7        62     81  2.59  50      695
## 2   5.1        59     66  1.70  39      403
## 3   7.4        57     83  2.16  55      710
## 4   6.5        73     41  2.01  48      349
## 5   7.8        65    115  4.30  45     2343
## 6   5.8        38     72  1.42  65      348
```



From the Scatterplot above.....

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 liver(0.67) and enzyme(0.58) low correlation blood(0.35) 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

- Write down the mathematical multiple regression model for this situation, defining all appropriate parameters.
- Write down the Hypotheses for the Overall ANOVA test of multiple regression.
- 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.

```
##
## 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
```

- State the Null distribution.
- Compute the P-Value
- State your conclusion (both statistical conclusion and contextual conclusion).

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

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

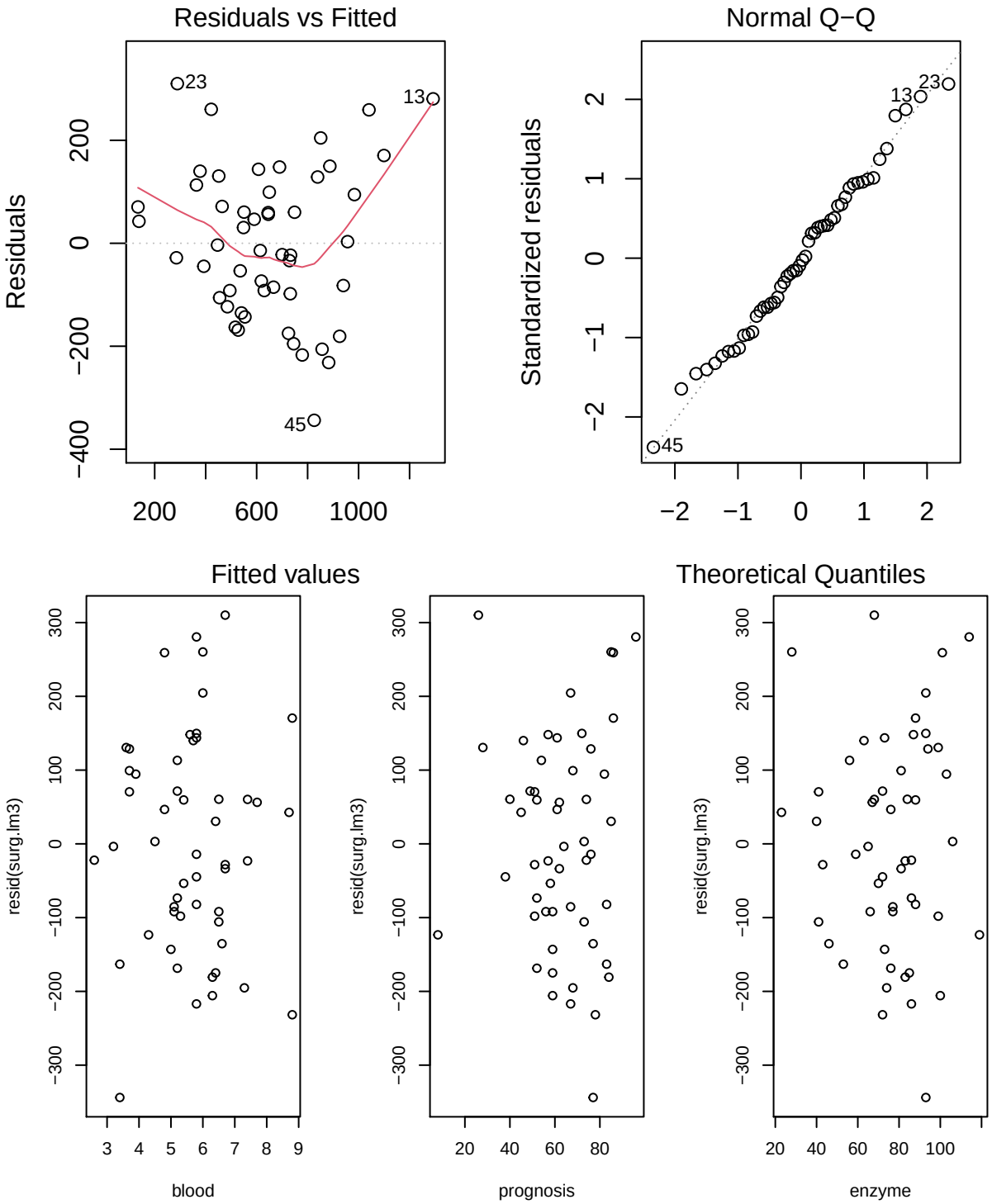
```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + age, data = surg.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -294.71 -116.28   -5.66  102.88  317.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -693.565    187.086  -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.



surg.lm3 is final model

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

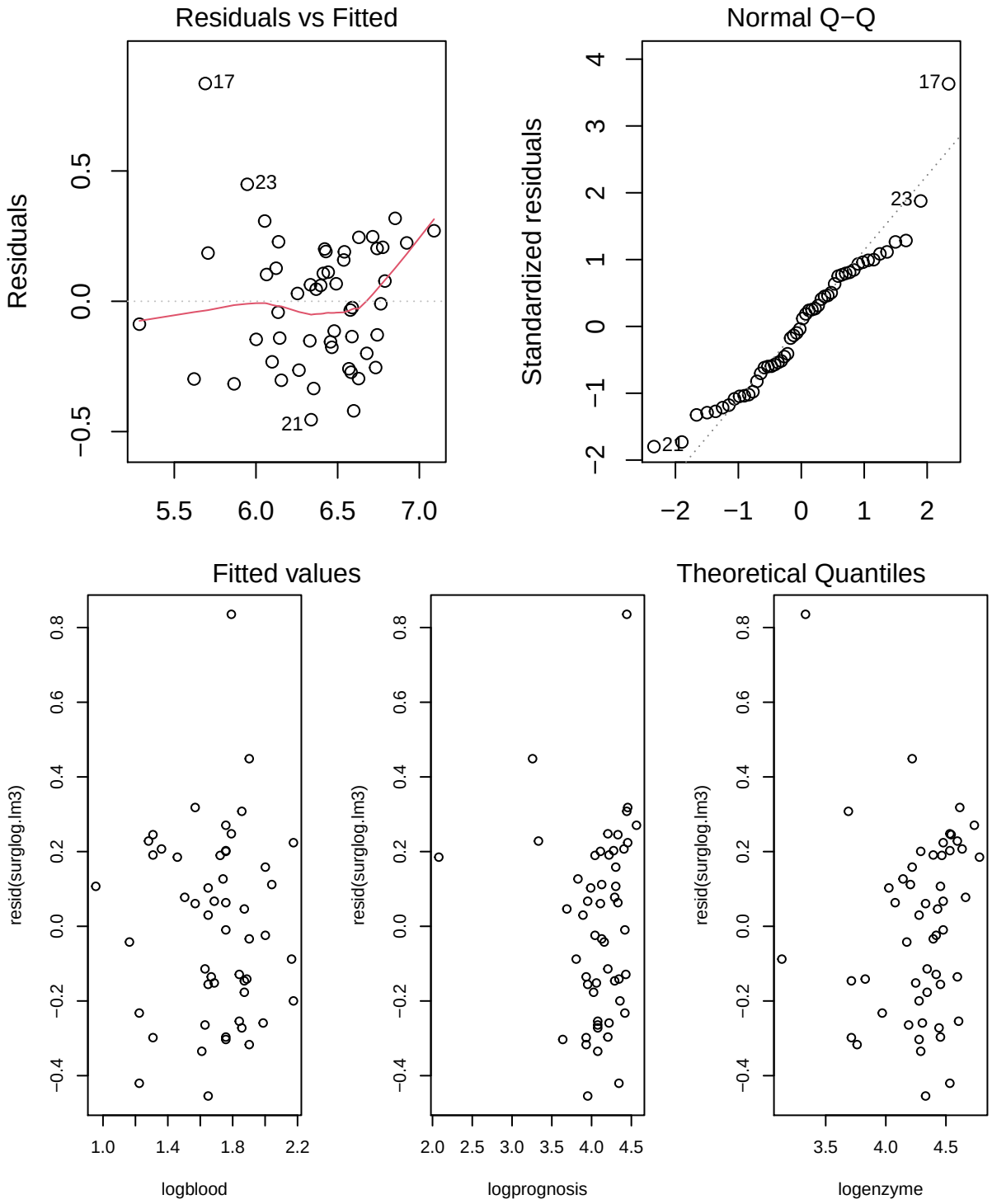
```
##
## Call:
## lm(formula = logsurvival ~ logblood + logprognosis + logenzyme +
##     logliver, data = surg.new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49346 -0.17295  0.04336  0.14248  0.70862
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (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 final model is surglog.lm3

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.

```
##         kmL driver   car
## 1  10.75614      A   one
## 2  10.71363      A   one
## 3  12.28666      A   two
## 4  12.75432      A   two
## 5  10.54357      A three
## 6  10.67111      A three
## 7  12.07409      A  four
## 8  11.86152      A  four
## 9  11.52140      A  five
## 10 11.30883      A  five
## 11 14.28484      B   one
## 12 13.98724      B   one
## 13 15.60278      B   two
## 14 15.51776      B   two
## 15 13.47706      B three
## 16 13.56209      B three
## 17 15.13513      B  four
## 18 14.88004      B  four
## 19 14.32735      B  five
## 20 14.41238      B  five
## 21 11.77649      C   one
## 22 12.11660      C   one
## 23 13.05192      C   two
## 24 12.92438      C   two
## 25 11.43637      C three
## 26 11.18129      C three
## 27 12.62678      C  four
## 28 12.83935      C  four
## 29 12.41420      C  five
## 30 12.28666      C  five
## 31 12.41420      D   one
## 32 12.45672      D   one
## 33 13.77467      D   two
## 34 13.77467      D   two
## 35 11.77649      D three
## 36 12.28666      D three
## 37 13.51958      D  four
## 38 13.05192      D  four
## 39 12.88186      D  five
## 40 12.71181      D  five
```

This study is balanced because .....

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

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

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.