# Using Multiple Linear Regression & Two Way ANOVA

Dom Nasrabadi

May 2021

In this document, we will be exploring 2 data sets in order to use 2 statistical methods and display how they can help us make inferences about the data. These 2 datasets are explored below. This document is a mini project to demonstrate my experience using statistical methods, using the R programming language with RStudio, as well as presenting to an audience using RMarkdown.

Table 1: Dataset 1: Fuel Efficiency of Cars & Drivers

| colname | description |
|---|---|
| kmL | the observed fuel efficiency of the car in km/L over a standard course |
| car | the specific car model |
| driver | the driver of the car |

Table 2: Dataset 2: Surgical procedure Results

| colnames | description |
|---|---|
| blood | blood clotting index |
| prognosis | prognosis index |
| enzyme | enzyme function index |
| liver | liver function index |
| age | age of patient, in years |
| gender | Gender of the patient |
| survival | survival time of the patient after surgery (in days) |

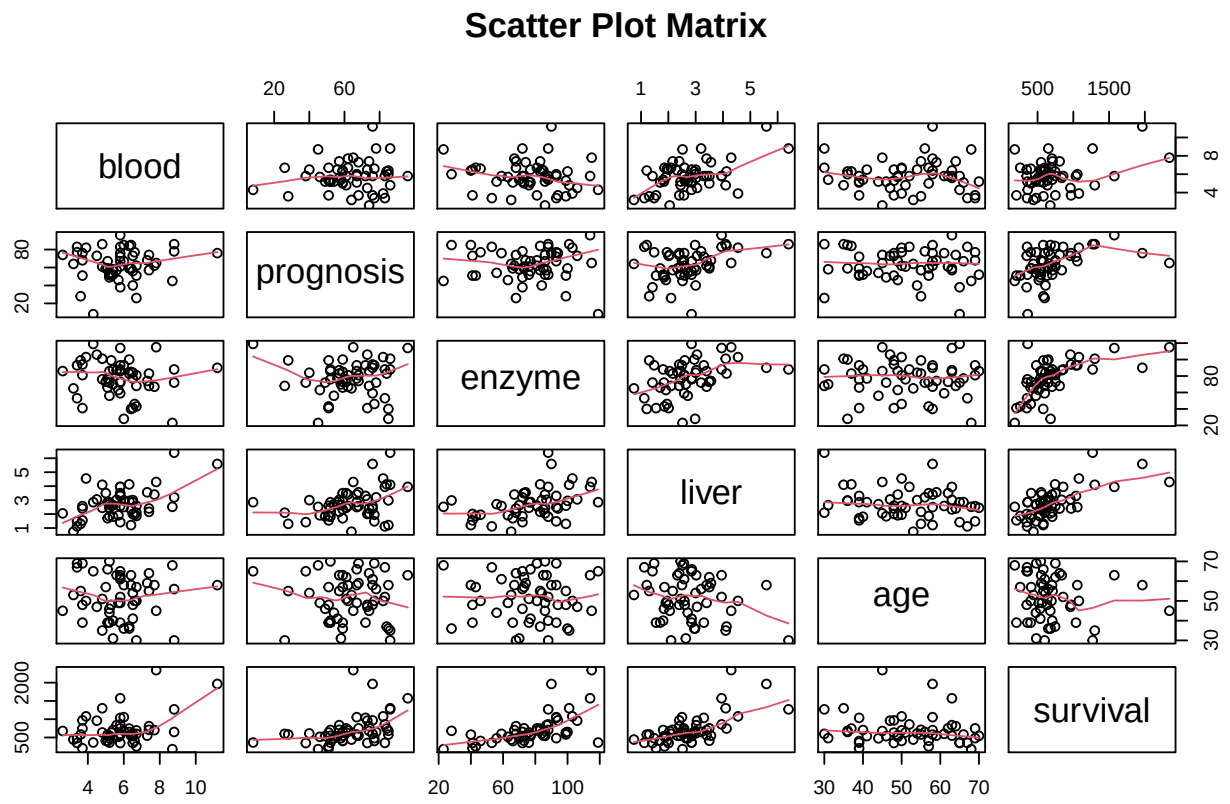
We will explore a variety of questions with regards to statistical procedures on the way. This includes exploratory data analysis, checking assumptions of the statistical tests, stating hypotheses, calculating test statistics and p-values and ultimately giving conclusions.

Enjoy.

# Question 1 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."*

```
# Bring in packages needed
library(dplyr)
# Read data in & remove gender since we are only concrned with true numeric variables
surg <- read.table("surg.dat", header = T)
surg.wo.gen <- select(surg, -gender)
# Produce scatter plot matrix
plot(surg.wo.gen, panel = panel.smooth, main="Scatter Plot Matrix")
```




Since Gender is a dichotomous/binary variable with a value of either 0 or 1, or rather Male or Female, it would not show any useful linear relationship, which means we shall exclude it from 1a and 1b. Let's examine each pair of variables, considering **correlation/linear trends, unusual observations and constant variance**.

**survival (response) & age (predictor)**

- linear pattern present although it is not strong in either direction
- there seems to be a few outliers that had survived longer than the majority
- there is constant variance about the line, excluding those couple of outliers

**survival (response) & liver (predictor)**

- there is a weak to moderate positive linear relationship here
- there are a couple of outliers among the tail end of the scatter plot (quite reasonable since better liver function should lead to larger survival times)
- there is also even spread

**survival (response) & enzyme (predictor)**

- probably our strongest linear relationship of the predictors with the response. There is a moderate linear association present
- possibly could argue there is a couple of unusual observations present however nothing too discerning
- there seems to be constant variance among the line too

**survival (response) & prognosis (predictor)**

- a weak to moderate positive correlation is evident, similarly to survival & liver and survival & enzyme which could be a sign that higher values on these 3 indexes leads to larger survival times
- there is some unusual observations present
- excluding the outliers, the spread is reasonable throughout

**survival (response) & blood (predictor)**

- there is a very slight linear trend that is possibly influenced by a couple of outliers
- a quite obvious outlier is present, some observations had rather large blood values
- again, the variance here is quite evenly spread

**age (predictor) & liver (predictor)**

- possibly a very slight negative correlation can be observed but nothing significant
- there is one outlier present in the corner of the plot
- we can also say there is constant variance here

**age (predictor) & enzyme (predictor)**

- no noteworthy correlation is present between these 2 variables
- there seems to be no unusual observations here
- the constant variance feature is probably most satisfied here

**age (predictor) & prognosis (predictor)**

- a very small negative linear slope is present here
- some slightly unusual observations with a low prognosis index
- constant variance is mostly satified

**age (predictor) & blood (predictor)**

- no clear linear association
- one outlier with a high blood clotting index
- even spread throughout the plot

**liver (predictor) & enzyme (predictor)**

- small linear pattern present which naively seems reasonable as greater liver function would mostly accompany greater enzyme function
- no very unusual observations, however 2 observations present with high liver indexes relative to the rest of the data
- spread seems reasonable throughout

**liver (predictor) & prognosis (predictor)**

- a small positive trend is visible
- no obvious outliers
- mostly even spread throughout

**liver (predictor) & blood (predictor)**

- similarly to liver and other predictors, a small positive slope found which indicates weak multicollinearity
- a couple of unusual observations here too
- variance is not completely constant

**enzyme (predictor) & prognosis (predictor)**

- no clear linear trend
- a few unusual observations on both sides of the plot
- variance is mostly constant

**enzyme (predictor) & blood (predictor)**

- no apparent linear trend or strong correlation here
- some outliers evident towards the higher end of the blood scale
- spread seems constant

**prognosis (predictor) & blood (predictor)**

- similarly to enzyme and blood, no obviously strong correlation
- some outliers evident towards the extremities of each axis
- variance is largely constant

# Question 1 b)

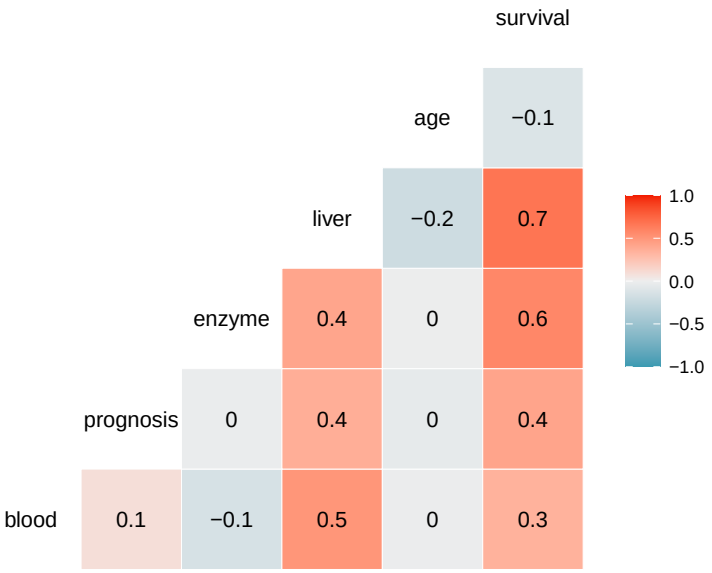
*"Compute the correlation matrix of the dataset and comment."*

NOTE: The arguments in the correlation matrix function (cor) require the inputs to be a numeric vector. We can either remove gender therefore or alternatively, change it to a numeric data type. We will proceed by leaving it out.

```
# Correlation Matrix
cor(surg.wo.gen)
```

```
##                 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
```

```
# We can also use the GGally package to check the correlation quite nicely in a visual way
library(GGally)
ggcorr(surg.wo.gen, label = T)
```




As seen above, the correlation matrix shows a numeric summary of association between pairs of variables. It measures both the strength and direction (positive or negative) of the linear relation between 2 variables.

In regards to our 'surg' data set, it seems like the correlation for some pairs is positive while a few have negative correlation (less than 0). The additional correlation visualization (from the GGally package) confirms this with the variation in orange and teal shades.

Our response variable (survival) is most correlated to the predictor variable - liver with a correlation coefficient of 0.67. Predictor variables, enzyme and prognosis also have moderate strength positive correlations with our response variable, survival. Additionally, there is weak to moderate multicollinearity between some predictor variables, namely, between liver and 3 other predictors: enzyme, prognosis and blood.

Thus, our correlation matrix allows us to assume this data is suitable for a multiple linear regression model. Most notably, this is due to our moderate linear association between our response and predictor variables.

# Question 1 c)

Fitting a Multiple-Linear Regression model, using Survival as our response variable, and all other variables as predictor variables. We will also use a significance level of 0.05 for this excercise.

### Multiple Regression Model

Let our response variable, $\hat{Y}$ be **survival** Our predictor, or $X_i$ variables shall be:

- $X_1$ blood: *blood clotting index*
- $X_2$ prognosis: *prognosis index*
- $X_3$ enzyme: *enzyme function index*
- $X_4$ liver: *liver function index*
- $X_5$ age: *age of the patient, in years*
- $X_6$ gender: *gender of the patient, (male of female)*

Our regression equation will then be:

$\hat{Y} = \beta_0 \ + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$

- $\beta_0$ is the y intercept
- $\beta_1 \dots \beta_k$ are the partial regression coefficients
- $\epsilon \sim$ i.i.d $N(0, \sigma^2)$ is the random error term, which is normally distributed

### Hypothesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

$H_1 :$ not all $\beta_i$ parameters are 0

### ANOVA Table

```
# model with all other variables as predictors
surg.lm <- lm(survival ~ . , data = surg)
surg.aov <- anova(surg.lm)

TotalRegSS <- sum(surg.aov$'Sum Sq'[1:6])
surg.aov
```

```
## 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
```

```
TotalRegSS # This is our combined regression SS
```

```
## [1] 5816714
```

Full Regression SS = 1005152+1278496+3442172+57862+33032+1 = 5816714

### F-Test

$F_{obs} = \frac{Reg.MS}{Res.MS} = \frac{969452}{54315} = 17.8487$

- Reg.MS = $\frac{Reg.SS}{k} = \frac{5816714}{6} = 969452$
- Res.MS = $\frac{Res.SS}{n-k-1} = \frac{2552807}{47} = 54315$

### Null Distribution

Assuming $H_0$ is true, $F_{obs}$ behaves like a $F_{k,n-k-1} = F_{6,47}$ distribution

### P-Value

```
# finding the p-value
1 - pf(17.85, df1 = 6, df2 = 47)
```

```
## [1] 1.188863e-10
```

P( $F_{6,47} \geq 17.85$ )= $0.0000000001188863 < 0.05$

### Contexual & Statistical Conclusion

Therefore, as examined from our very large F statistic and our extremely small P-Value, we can fairly reject the null hypothesis at the 5% and the 1% significance levels, indicating that at least one partial coefficient is not equal to 0. This means that there is a significant linear relationship between the response variable (*survival*) and at least one of the 6 predictor variables.

Lastly, to summarize, our estimated regression line is:

$\hat{survival}$ = -1179.1889 + 86.6437*blood* + 8.5013*prognosis* + 11.1246*enzyme* + 38.5068*liver* - 2.3409*age* - 0.2201*gender*

# Question 1 d)

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

In this example, we can either choose to use forward model selection (where we start with one predictor and then subsequently add others) or backwards model selection (which regresses with all predictors and subsequently removes those that are not significant) to achieve a parsimonious model. We shall use backwards model selection in this example.

```
# initial model with all predictors
surg.lm <- lm(survival ~ . , data = surg)
summary(surg.lm)
```

```
##
## Call:
## lm(formula = survival ~ ., data = surg)
##
## 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
## genderM        -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
```

Seems as though the Gender variable has the highest p-value of them all. We shall drop this from our model and regress with the reduced model.

```
# reduced model after dropping gender
surg.lm.2 <- lm(survival ~ .-gender , data = surg)
summary(surg.lm.2)
```

```
##
## Call:
## lm(formula = survival ~ . - gender, data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.34 -147.74   11.74  124.67  954.32
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -1179.367    275.619  -4.279 8.91e-05 ***
## blood          86.630     26.905   3.220 0.002302 **
## prognosis       8.501      2.137   3.978 0.000234 ***
## enzyme         11.124      1.958   5.683 7.62e-07 ***
## liver          38.554     49.251   0.783 0.437595
## age            -2.340      2.969  -0.788 0.434514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.6 on 48 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.6632
## F-statistic: 21.87 on 5 and 48 DF,  p-value: 2.386e-11
```

Predictor variables liver and age both have rather high p-values around 0.43, but liver is slightly higher so it will be dropped in our next iteration.

```
# reduced model after dropping liver
surg.lm.3 <- lm(survival ~ .-gender - liver , data = surg)
summary(surg.lm.3)
```

```
##
## Call:
## lm(formula = survival ~ . - gender - liver, data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -416.92 -142.56  -13.98  138.10  943.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1246.655    260.835  -4.779 1.64e-05 ***
## blood         100.660     19.987   5.036 6.83e-06 ***
## prognosis       9.291      1.876   4.951 9.14e-06 ***
## enzyme         12.101      1.502   8.058 1.56e-10 ***
## age            -2.986      2.841  -1.051    0.298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.7 on 49 degrees of freedom
## Multiple R-squared:  0.6911, Adjusted R-squared:  0.6659
## F-statistic: 27.41 on 4 and 49 DF,  p-value: 5.68e-12
```

While all the other predictors have now become more significant, the age variable is still not significant at $\alpha = 0.05$ and it will be dropped. This last iteration should result in our final model with all signifcant predictor variables.

```
# reduced model after dropping age
surg.lm.4 <- lm(survival ~ .-gender - liver - age, data = surg) # or can re-write as ...
surg.lm.4 <- lm(survival ~ blood + prognosis + enzyme, data = surg)
summary(surg.lm.4)
```

```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme, data = surg)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -432.4 -134.3  -19.1  111.9  961.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1410.847    209.118  -6.747 1.50e-08 ***
## blood         101.054     20.005   5.052 6.22e-06 ***
## prognosis       9.382      1.876   5.000 7.43e-06 ***
## enzyme         12.128      1.503   8.069 1.30e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.9 on 50 degrees of freedom
## Multiple R-squared:  0.6841, Adjusted R-squared:  0.6652
## F-statistic:  36.1 on 3 and 50 DF,  p-value: 1.469e-12
```

We have now come to our final model where all predictors are significant in explaining our new transformed response variable. This is after removing the predictors : gender, liver and age which were all not significantly improving our model output. This final model can be written as:

$\hat{Y} = \beta_0 \ + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

- $X_1$ blood: *blood clotting index*
- $X_2$ prognosis: *prognosis index*
- $X_3$ enzyme: *enzyme function index*
- $\beta_0$ is the intercept term
- $\epsilon \sim$ i.i.d $N(0, \sigma^2)$

While our line of best fit is:

$\hat{survival}$ = -1410.847 + 101.054(*blood*) + 9.382(*prognosis*) + 12.128(*enzyme*)

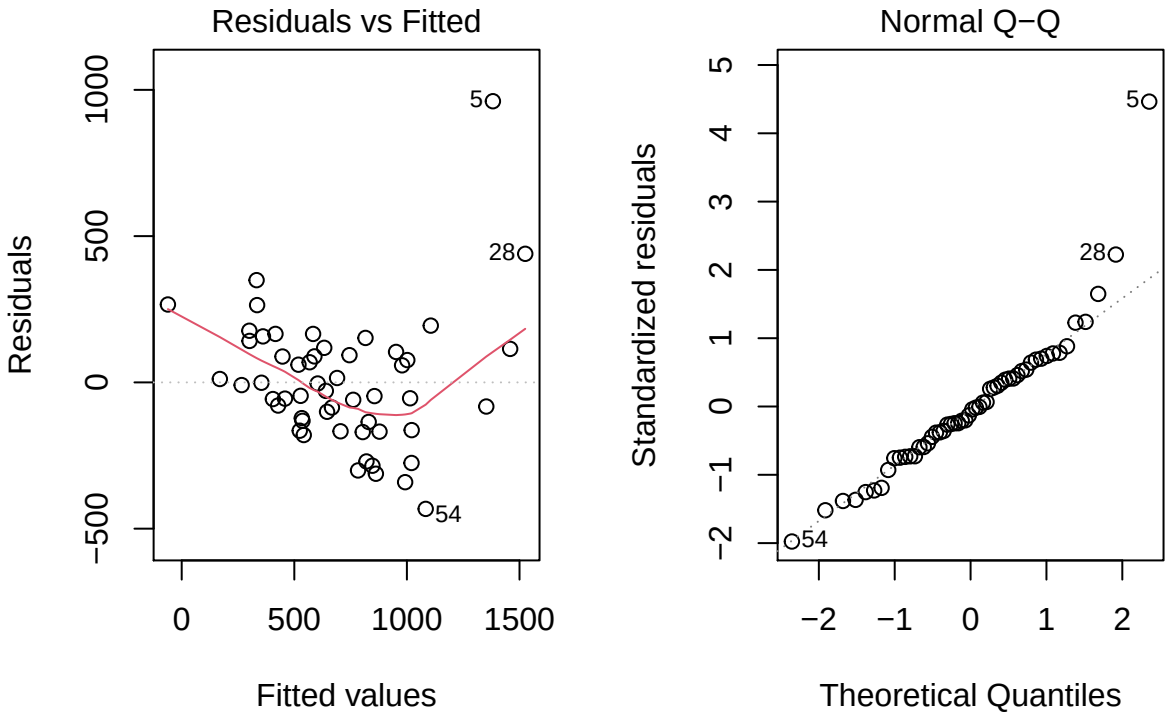
## Question 1 e)

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

After finalising our model, we must now re-evaluate if it still meets our model assumptions for a multiple linear regression to hold. 3 such assumptions include:

- Constant variance among residuals: checked using a Residuals vs Fitted Plot
- Constant variance for predictors: checked using a Residuals vs Fitted Plot for each predictor
- Normal distribution among residuals: checked using a Normal Q-Q Plot

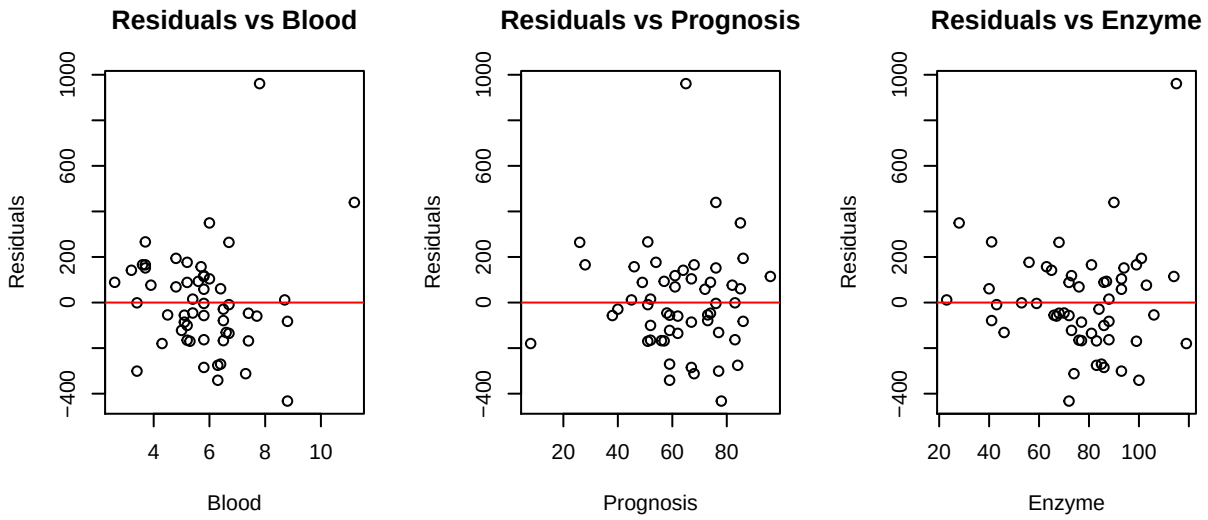
```
par(mfrow=c(1,2))
plot(surg.lm.4, which = 1:2) # plot our Residuals vs Fitted and Normal Q-Q Plot
```




Looking at our Residuals vs Fitted Plot, we can see a slight curvature or dip towards the centre of the fitted values. The curve upwards could also be caused by some unusual observations (observations 5, 28 & 54). Since there is some curvation present and a change in variation, our linearity and constant variance assumptions do not hold.

Furthermore, checking our Normal Q-Q Plot, it seems to be following the diagonal line quite well. However, similarly to our Residuals vs Fitted Plot, observations 5 and 28 are deviating from the line considerably.

```r
# plot our residuals vs predictors
par(mfrow=c(1,3))
plot(surg$blood,surg.lm.4$residuals, xlab = "Blood", ylab = "Residuals",
     main = "Residuals vs Blood")
abline(h=0, col="red")
plot(surg$prognosis,surg.lm.4$residuals, xlab = "Prognosis", ylab = "Residuals",
     main = "Residuals vs Prognosis")
abline(h=0, col="red")
plot(surg$enzyme,surg.lm.4$residuals, xlab = "Enzyme", ylab = "Residuals",
     main = "Residuals vs Enzyme")
abline(h=0, col="red")
```




The Residuals vs Predictor plots seem to be the "best-looking" of all 3 groups of plots. This is because on first glance, they largely seem to be distributed randomly about the line which indicates mostly constant variance among the predictors. Looking closer however, there is more observations below the horizontal 0 line while there are also some outliers much higher than the line too.

We could argue that in the blood predictor plot, the variance is not completely constant with more values clustered towards the left with outliers above and below. This fan shape pattern indicates large observations taking greater residual values.

**Summary**: Thus, after checking the diagnostics of our model, it can be argued that the 3 linear regression assumptions are not completely satisfied due to the curvature in the Residuals vs Fitted Plot, as well as some deviations in the Normal Q-Q Plot. Our Residuals vs Predictors Plots also show a slightly uneven spread with some outliers. We might however be able to improve these by transforming some of our variables (next page).

## Question 1 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 predictors"*

Since we are transforming our response variable (*survival*) to a log transformed version, we must begin our testing by including all variables again similarly to what we did in 1d. This is because log(survival) is not the same as survival when we interpret our results and it will influence our model output. Additionally, we will again be using backwards model selection here to achieve a parsimonious final model.

```
# initial model with all predictors & log(response)
log.surg.lm <- lm(log(survival) ~ . , data = surg) # or re-write as ...
log.surg.lm <- lm(log(survival) ~ blood + prognosis + enzyme + liver + age + gender ,data=surg)
summary(log.surg.lm)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme + liver +
##     age + gender, data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42847 -0.16913  0.00696  0.18167  0.50226
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.100997   0.302781  13.544  < 2e-16 ***
## blood        0.094858   0.029328   3.234  0.00223 **
## prognosis    0.013020   0.002304   5.650 9.08e-07 ***
## enzyme       0.016245   0.002114   7.683 7.59e-10 ***
## liver       -0.003132   0.055256  -0.057  0.95503
## age         -0.004863   0.003215  -1.513  0.13709
## genderM     -0.066140   0.072024  -0.918  0.36315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2486 on 47 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7441
## F-statistic: 26.69 on 6 and 47 DF,  p-value: 1.391e-13
```

Interestingly, in our first iteration of model selection, liver is not significant with a p-value of 0.955. We will regress with all other predictors while removing liver.

```
# model version 2 after removing liver variable
log.surg.lm.2 <- lm(log(survival) ~ blood + prognosis + enzyme + age + gender ,data=surg)
summary(log.surg.lm.2)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme + age +
##     gender, data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42563 -0.16780  0.00911  0.18059  0.50244
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.105132   0.290795  14.117  < 2e-16 ***
## blood        0.093738   0.021439   4.372 6.58e-05 ***
## prognosis    0.012960   0.002026   6.398 6.16e-08 ***
## enzyme       0.016170   0.001627   9.939 3.10e-13 ***
## age         -0.004810   0.003043  -1.581    0.121
## genderM     -0.065010   0.068487  -0.949    0.347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.246 on 48 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7495
## F-statistic: 32.71 on 5 and 48 DF,  p-value: 2.291e-14
```

While all p-values are now more significant than before, gender is still insignificant at a p-value of 0.347. Let's drop it from the model and regress with the others.

```
# model version 3 after dropping gender variable
log.surg.lm.3 <- lm(log(survival) ~ blood + prognosis + enzyme + age ,data=surg)
summary(log.surg.lm.3)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme + age,
##     data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39491 -0.18866 -0.00045  0.17491  0.51787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.028531   0.279090  14.434  < 2e-16 ***
## blood        0.094845   0.021386   4.435 5.20e-05 ***
## prognosis    0.013199   0.002008   6.574 3.04e-08 ***
## enzyme       0.016402   0.001607  10.208 1.01e-13 ***
## age         -0.004767   0.003040  -1.568    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2458 on 49 degrees of freedom
## Multiple R-squared:  0.7688, Adjusted R-squared:   0.75
## F-statistic: 40.74 on 4 and 49 DF,  p-value: 5.171e-15
```

Age is still above 0.05 so we will also remove this from our model.

```
# model version 4 after dropping age variable
log.surg.lm.4 <- lm(log(survival) ~ blood + prognosis + enzyme ,data=surg)
summary(log.surg.lm.4)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme, data = surg)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46994 -0.17938 -0.03116  0.17959  0.59105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.766441   0.226757  16.610  < 2e-16 ***
## blood       0.095475   0.021692   4.401 5.66e-05 ***
## prognosis   0.013344   0.002035   6.558 2.95e-08 ***
## enzyme      0.016444   0.001630  10.089 1.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2493 on 50 degrees of freedom
## Multiple R-squared:  0.7572, Adjusted R-squared:  0.7427
## F-statistic: 51.99 on 3 and 50 DF,  p-value: 2.137e-15
```

We have now come to our final model where all predictors are significant in explaining our response variable. This is after removing, liver, gender and age respectively. We are left with just 3 significant predictor variables, namely, blood, prognosis and enzyme. This final model can be written as:

$\hat{Y} = \beta_0 \ + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

- $\hat{Y}$ log(survival): the log of the survival time of the patient after surgery (in days)
- $X_1$ blood: *blood clotting index*
- $X_2$ prognosis: *prognosis index*
- $X_3$ enzyme: *enzyme function index*
- $\beta_0$ is the intercept term
- $\epsilon \sim$ i.i.d $N(0, \sigma^2)$

While our line of best fit is:

$\hat{log(survival)} = 3.766441 + 0.095475(blood) + 0.013344(prognosis) + 0.016444(enzyme)$

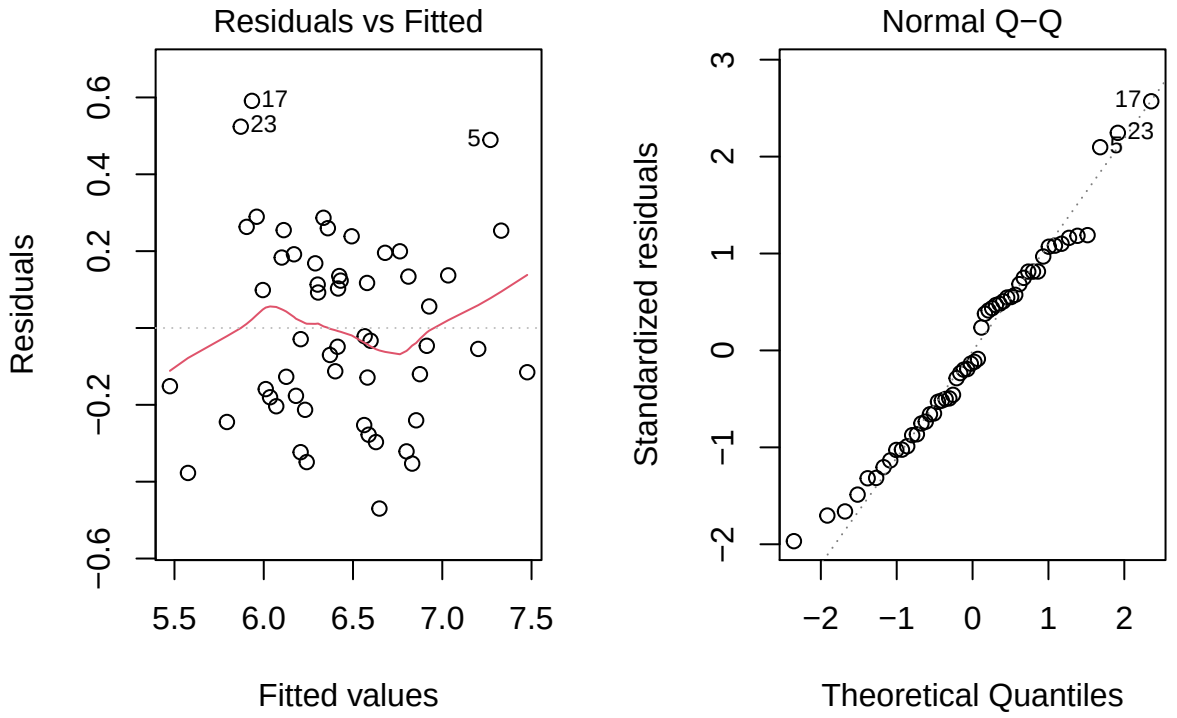
Note: If we wanted to interpret this in terms of the original units of *survival* we would need to reverse the log transformation using an exponential function to get it back.

## Question 1 g)

*"Validate your final model with the log(survival) time response. In your answer, explain why the regression model with the log(survival) response variable is superior to the model with the survival response variable"*

Just like we did in 1e, let's check our 3 types of plots to validate our model has satisfied the assumptions.

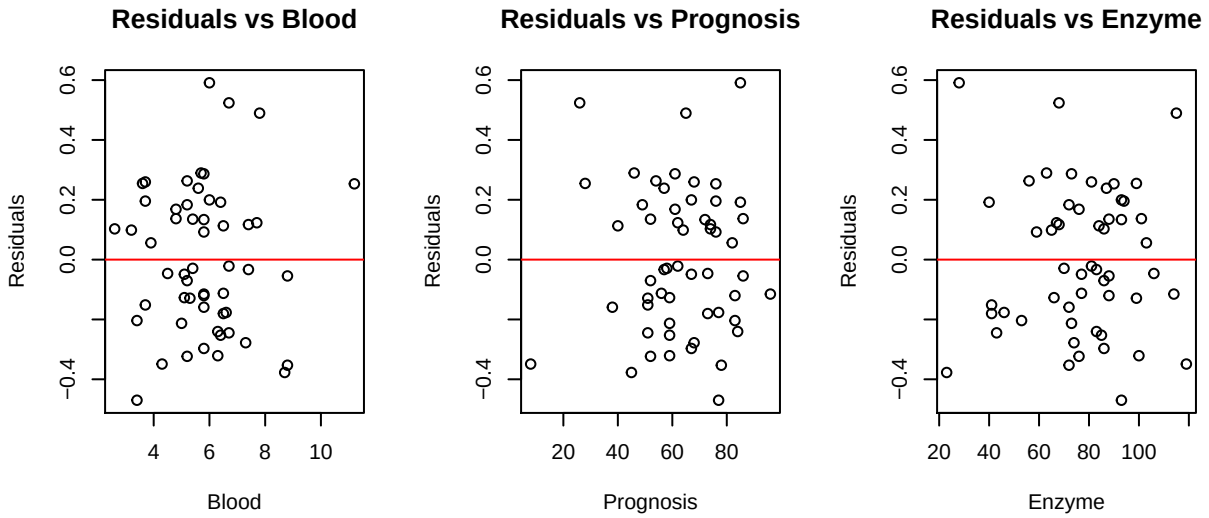
```
par(mfrow=c(1,2))
plot(log.surg.lm.4, which = 1:2) # plot our Residuals vs Fitted and Normal Q-Q Plot
```




The Residuals vs Fitted plot is displaying a large improvement compared to the one we saw earlier. Mainly due to how evenly our points are spread about the zero line in a random pattern.

The Normal Q-Q Plot is also better than last time with the effect of the outliers now removed since log transformations penalize large values quite heavily. There is an ever so slight deviation in the higher parts of the line, however it is mostly uniform around the line with no discernible curvature. Our normality assumption is therefore valid.

```
# Residuals vs Predictors Plots
par(mfrow=c(1,3))
plot(surg$blood,log.surg.lm.4$residuals, xlab = "Blood", ylab = "Residuals",
     main = "Residuals vs Blood")
abline(h=0, col="red")
plot(surg$prognosis,log.surg.lm.4$residuals, xlab = "Prognosis", ylab = "Residuals",
     main = "Residuals vs Prognosis")
abline(h=0, col="red")
plot(surg$enzyme,log.surg.lm.4$residuals, xlab = "Enzyme", ylab = "Residuals",
     main = "Residuals vs Enzyme")
abline(h=0, col="red")
```




Yet again, the residual vs predictor plots here seem better than previously with more even distribution and spread around the line with no discernible pattern (e.g. diamond or fan shape).

**Summary**: After checking our diagnostics for the new model where our response variable is log(survival), we can safely say it is superior to the previous model where the response variable is only *survival*. All 3 types of plots show improvements compared to before with significantly less outliers and no odd patterns.

## Question 2 a)

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

```
library(knitr)
# read data in
fuel.eff <- read.table("kml.dat", header = T)
# compute summary table for number of replicates
kable(table(fuel.eff$car, fuel.eff$driver),caption = "Number of Replicates")
```

Table 3: Number of Replicates

| | A | B | C | D |
|---|---|---|---|---|
| five | 2 | 2 | 2 | 2 |
| four | 2 | 2 | 2 | 2 |
| one | 2 | 2 | 2 | 2 |
| three | 2 | 2 | 2 | 2 |
| two | 2 | 2 | 2 | 2 |

It appears we have a balanced study since there are an equal number of replicates within each treatment group. That is, each car type has 8 replicates, while each driver has 10 replicates. This results in each pair of combinations having 2 observations. This is a rather small amount and we will proceed with caution when interpreting our results.

## Question 2 b)

*"Construct two different preliminary graphs that investigate different features of the data set and comment."*

Let's explore 2 useful plots for this sort of problem. They are the interaction plot and boxplots. We will first need to convert our car and driver variables to the factor data type since they are currently as the character data type. We will also use the help of the ggplot2 package to aid our visualisation.

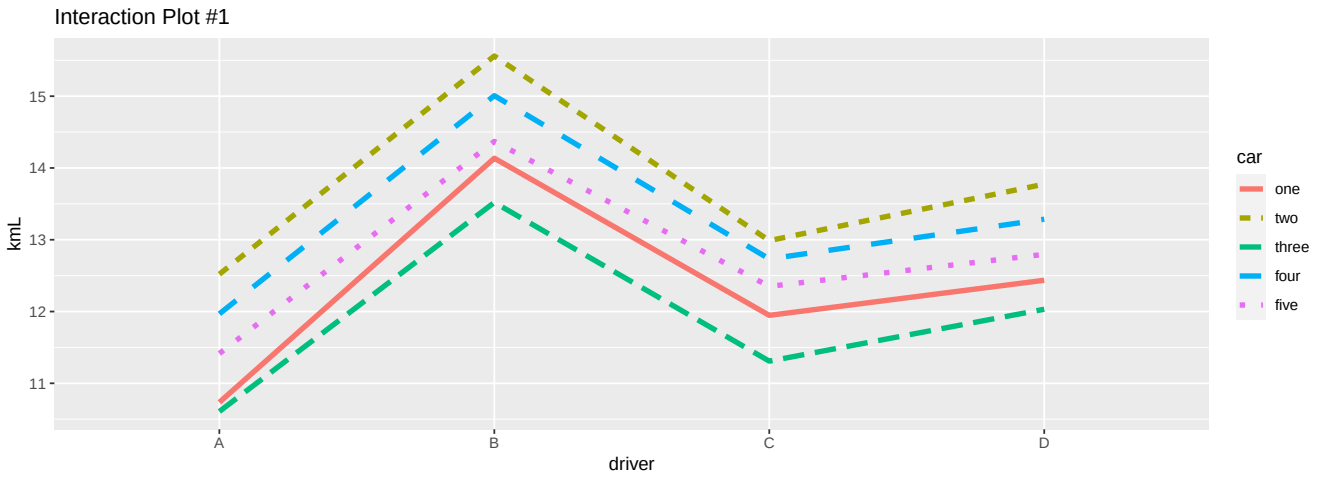
Since ggplot2 does not have a native interaction plot function like base R we will need to create a ANOVA Means table or wide summary. We will also append a column to make sure our boxplot captures all pairs of our 2 factors (car and driver).

```
# load packages
library(ggplot2)
# create wide summary for interaction plot input
fuel.eff.means <- aggregate(kmL ~ driver + car, data = fuel.eff, "mean")
# convert to factor
fuel.eff$car <- factor(fuel.eff$car)
fuel.eff$driver <- factor(fuel.eff$driver)
# append columns for interaction between car and driver
fuel.eff$interaction <- interaction(fuel.eff$driver, fuel.eff$car)
# show mean table
group.means <- data.frame(tapply(fuel.eff$kmL, list(fuel.eff$car,fuel.eff$driver), mean))
group.means$rwnms <- factor(rownames(group.means))
x <- factor(c("one","two","three","four","five"))
sorted.group.means <- group.means[match(x,group.means$rwnms),] # order row names correctly
sorted.group.means <- select(sorted.group.means,-rwnms)
kable(sorted.group.means, caption = "Treatment Means")
```

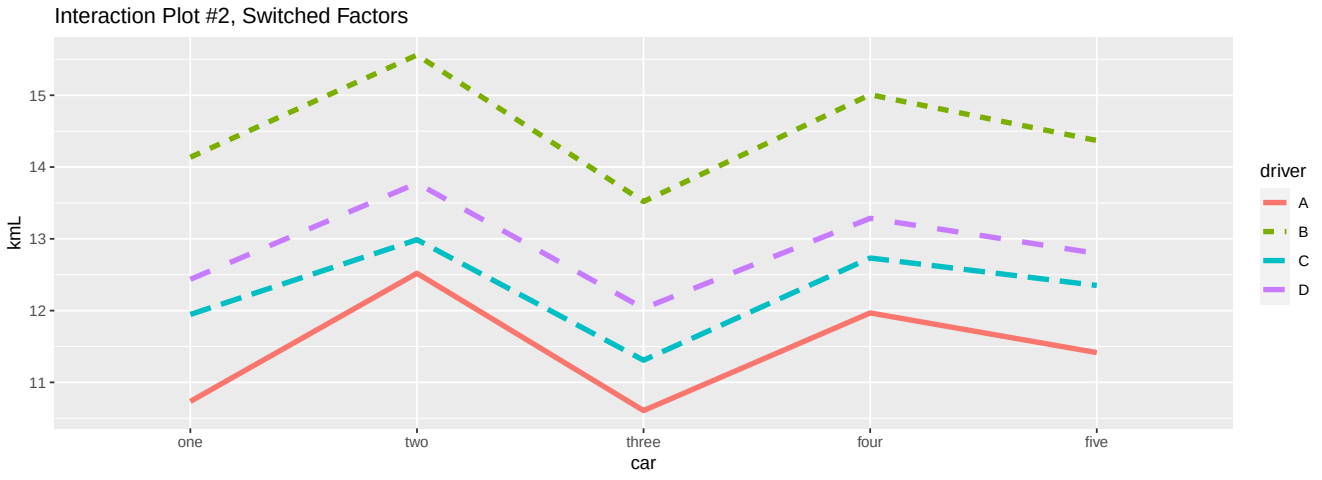
Table 4: Treatment Means

| | A | B | C | D |
|---|---|---|---|---|
| one | 10.73489 | 14.13604 | 11.94655 | 12.43546 |
| two | 12.52049 | 15.56027 | 12.98815 | 13.77467 |
| three | 10.60734 | 13.51958 | 11.30883 | 12.03158 |
| four | 11.96780 | 15.00758 | 12.73306 | 13.28575 |
| five | 11.41512 | 14.36987 | 12.35043 | 12.79683 |

```
# create interaction plot + order car variable
fuel.eff.means$car <- factor(fuel.eff.means$car, levels = c("one","two","three","four","five"))
ggplot(fuel.eff.means, aes(x=driver,y=kmL))+
   geom_line(size = 1.5, aes(group=car,color=car,linetype=car)) + ggtitle("Interaction Plot #1")
```




```
# create interaction plot, switch factor order
ggplot(fuel.eff.means, aes(x=car,y=kmL))+
   geom_line(size = 1.5, aes(group=driver,color=driver,linetype=driver)) + ggtitle("Interaction Plot #2, S
```

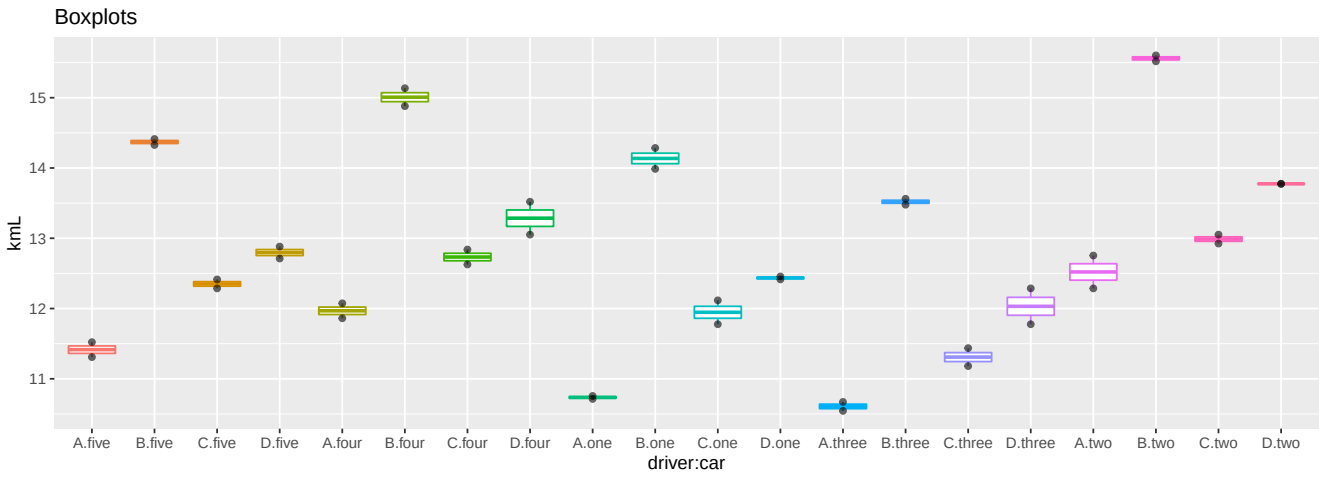



Our first type of plot, the interaction plot is great for interpreting the response change with different levels of our 2 factors. We have created 2 plots since it is suggested to try both factors in different order and examine.

If there is no interaction, then the lines would be parallel indicating that a change in response to a change in one factor will be the same irrespective of the other factor's level.

That is not the case here though, since the lines are not exactly parallel and there is a difference in the slope of some lines indicating there could be some interaction between the two factors. We would not say however that this interaction is extremely significant since there is no crossing of lines or large differences in slopes, rather, the interactions are subtle.

```
# create boxplot
ggplot(fuel.eff, aes(x=interaction,y=kmL)) +
  geom_boxplot(aes(color=interaction))+
  geom_point(width = 0.1, alpha = 0.6) + ggtitle("Boxplots") + xlab("driver:car") +
  scale_color_discrete(guide = FALSE)
```




Comparitive boxplots are another type of graph that are helpful in visualising the effects. Here, we have all our pairs of combinations displayed which is useful for checking the relative size + variability of effects and any outliers present in the data.

One thing we must keep in consideration is that the number of replicates in each pair is very small (2) and will make comparisons less noteworthy than if our sample size was much larger. It is also hard to gauge the underlying dynamics of the dataset with these graphs alone, thus, we will continue with caution.

# Question 2 c)

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

### Two Way ANOVA Model

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

- $\gamma_{ijk}$ is the kmL response
- $\mu$ is the overall population mean
- $\alpha_i$ is the main effect of Car where i takes the values of 1,2,3,4,5 (5 total levels)
- $\beta_j$ is the main effect of Driver where j takes the values of 1,2,3,4 (4 total levels)
- $\gamma_{ij}$ is the interaction effect of the $i^{th}$ and $j^{th}$ combination of the two factors
- $\epsilon_{ijk}$ unexplained variation for each replicated observation ~ i.i.d $N(0, \sigma^2)$

Using our 'kmL' data set, this model takes form in:

$kmL_{ijk} = \mu + \alpha(Car_i) + \beta(Driver_j) + \gamma(Car : Driver_{ij}) + \epsilon_{ijk}$

### Hypothesis

In factorial ANOVA we break down our Treatment Sum of Squares into 3 groups. They include:

- Factor A Variability
- Factor B Variability
- Interaction Variability

This means we will have 3 types of tests and therefore, hypotheses to examine. The first one is always the interaction effect.

**Testing Interaction (car:driver)**

- $H_0$ : $\gamma_{ij} = 0$ for all i's and j's
- $H_1$ : not all $\gamma_{ij} = 0$

**Testing Main Effect of Factor A (car)**

- $H_0$ : $\alpha_i = 0$ for all i's
- $H_1$ : not all $\alpha_i = 0$

**Testing Main Effect of Factor B (driver)**

- $H_0$ : $\beta_i = 0$ for all j's
- $H_1$ : not all $\beta_j = 0$

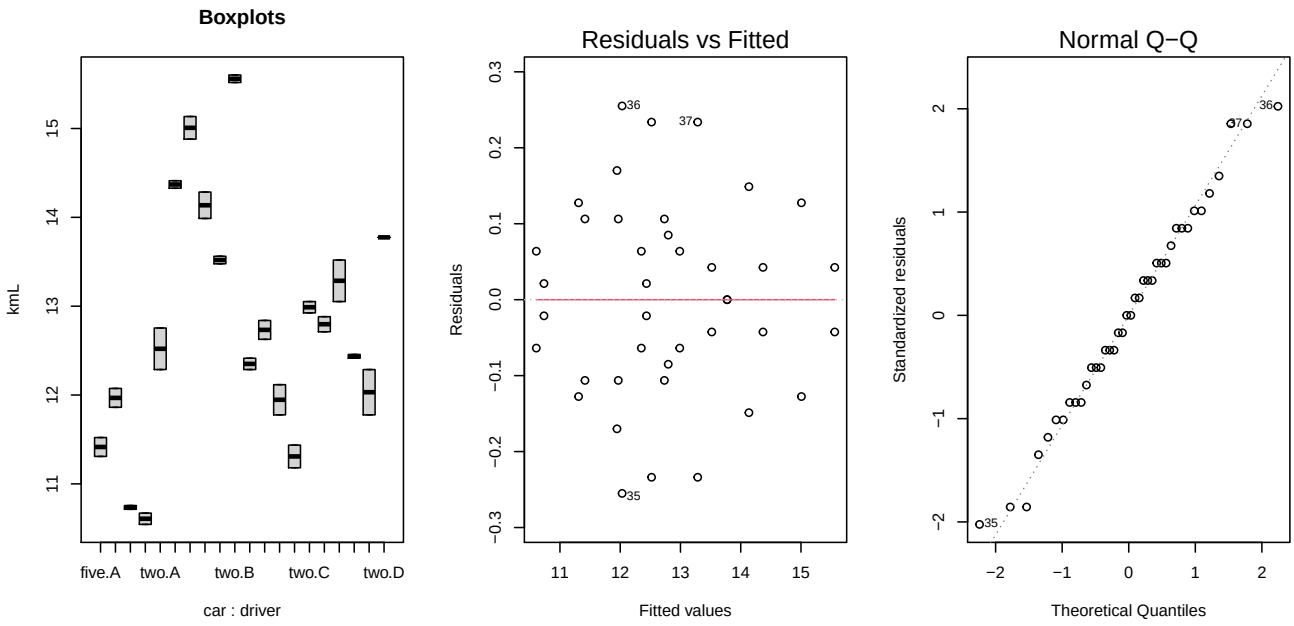
### Checking Assumptions

Similar to a One-Way ANOVA, we have 3 assumptions in order for our test results to hold.

- Residuals should be normally distributed: checked with a Normal Q-Q Plot
- Constant variability among groups: checked by looking at the IQR of boxplots and if we are really keen to use the rule of thumb
- Residuals have contant variance: checked using a Residuals vs Fitted Plot

We also prefer to have a balanced design (which we do in this example), but if not, we can compute this using a slightly longer method.

```
par(mfrow=c(1,3))
kmL.aov <- aov(kmL ~ car + driver + car:driver, data = fuel.eff)
boxplot(kmL ~ car + driver, data = fuel.eff, main = "Boxplots")
plot(kmL.aov, which = 1:2)
```




Having a look at our 3 plots, it seems like our assumptions are mostly satisfied. The Residuals vs Fitted plot has no discernible pattern and mostly even spread, while the Normal Q-Q Plot is also matching the diagonal line very well. The only thing we might be concerned about is the assumption of constant variability since our boxplots vary a fair bit in their size. This however is also expected since we only have 2 observations in each pair of combinations and we should not be overly worried about these differences. We may continue, albeit with caution.

### Two Way ANOVA Analysis

```
# check interaction effect + main effects in one model
kmL.aov <- aov(kmL ~ car + driver + car:driver, data = fuel.eff)
summary(kmL.aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## car           4  17.12   4.280  134.73 3.66e-14 ***
## driver        3  50.66  16.887  531.60  < 2e-16 ***
```

```
## car:driver  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
```

Judging from our ANOVA table above, the interaction effect is found to be statistically **insignificant** with a p-value of around 0.37. This is higher than our generic signifcance level of 0.05 so we do not reject the null for our interaction test hypothesis. We are therefore saying that the interaction effect of car:driver is insignificant and can be dropped from the model.

**Hypothesis (Interaction Effect)** : Do not reject $H_0$ where $\gamma_{ij} = 0$ for all i's and j's

**P-Value** = $0.371 > 0.05$

That also means, we can move on to testing the main effects of Factor A (car) and Factor B (driver) in a reduced model. As seen above, both car type and driver type are significant with much smaller p-values than our $\alpha$ of 0.05. Thus, concluding for both factors, that not all factor levels are = 0 (i.e. they both have an effect on the response kmL) meaning we will reject the null and accept the alternative hypothesis.

**Hypothesis (Main Effects)** : $H_1$ : not all $\alpha_i = 0$ & $H_1$ : not all $\beta_j = 0$

**P-Value** = both are less than 0.05

# Question 2 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 c. and the preliminary plots in b. You do not need to statistically examine the multiple comparisons between contrasts and interactions"*

In summary, judging from our analysis above and the preliminary graphs examined in 2b, we can make a couple of inferences about this data set.

Our hypothesis tests found there to be no significant interaction effect, rather *car* and *driver* were both significant on their own as main effects.

Interaction Plot #1 displayed that the driver of the car had an influence on the fuel efficiency (kmL). This is particularly evident in the spike around driver B which had higher values for each car level than the other drivers. Driver A constrastingly seemed to have a lower average than the others.

Additionally, checking Interaction Plot #2, the plot shows 2 humps at car two and four. This indicates that the cars had different fuel efficiency (kmL) means suggesting not all levels were equal with no effect. Because of these 2 events, we rejected the null for both main effects.

Rounding off, our comparitive boxplots also showed alot of variability with means/medians very disparate and spread out. While the small sample size played a part in this, it is clear that not all groups were similar e.g. boxplots including driver B were had much higher kmL values relative to the rest.