Assignment

Isha

5/19/2021

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:

```
str(surg)
```

```
## 'data.frame':    54 obs. of  7 variables:
##  $ blood    : num  6.7 5.1 7.4 6.5 7.8 5.8 5.7 3.7 6 3.7 ...
##  $ prognosis: int  62 59 57 73 65 38 46 68 67 76 ...
##  $ enzyme   : int  81 66 83 41 115 72 63 81 93 94 ...
##  $ liver    : num  2.59 1.7 2.16 2.01 4.3 1.42 1.91 2.57 2.5 2.4 ...
##  $ age      : int  50 39 55 48 45 65 49 69 58 48 ...
##  $ gender   : chr  "M" "M" "M" "M" ...
##  $ survival : int  695 403 710 349 2343 348 518 749 1056 968 ...
```

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.

- You will need to remove the gender variable to do this.

```
str(surg)
```

```
## 'data.frame':    54 obs. of  7 variables:
##  $ blood    : num  6.7 5.1 7.4 6.5 7.8 5.8 5.7 3.7 6 3.7 ...
##  $ prognosis: int  62 59 57 73 65 38 46 68 67 76 ...
##  $ enzyme   : int  81 66 83 41 115 72 63 81 93 94 ...
##  $ liver    : num  2.59 1.7 2.16 2.01 4.3 1.42 1.91 2.57 2.5 2.4 ...
##  $ age      : int  50 39 55 48 45 65 49 69 58 48 ...
##  $ gender   : chr  "M" "M" "M" "M" ...
##  $ survival : int  695 403 710 349 2343 348 518 749 1056 968 ...
```

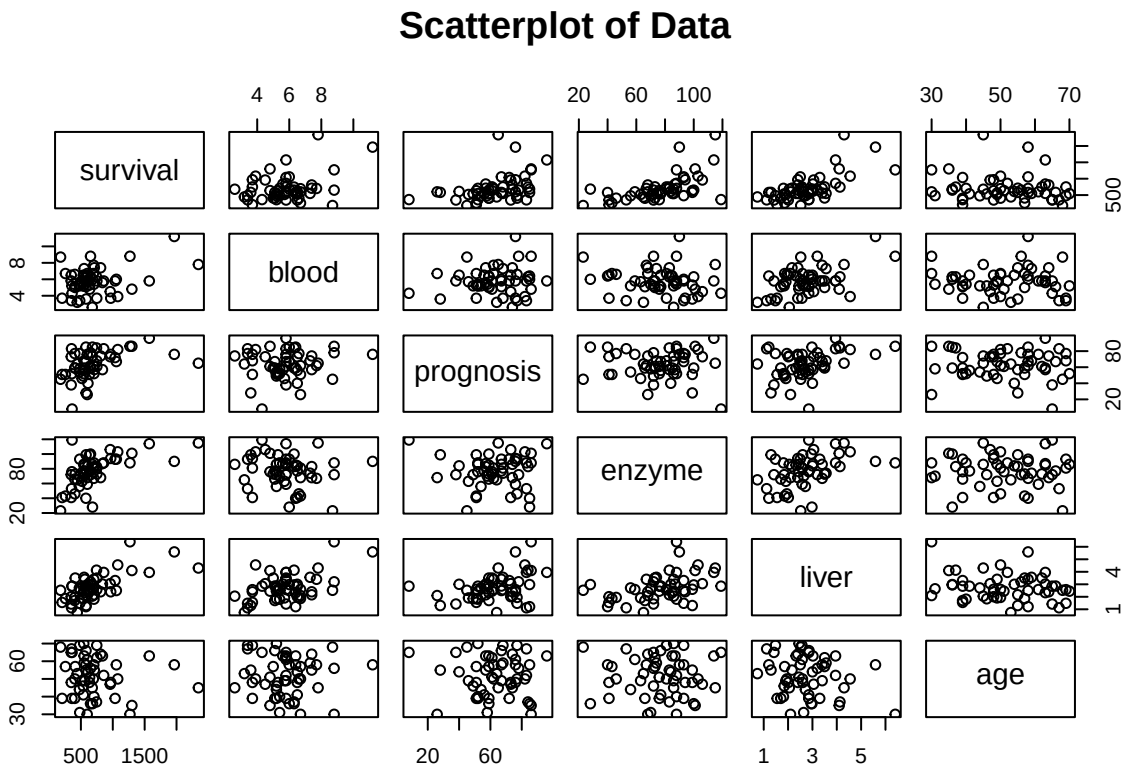
```
surg1 <- subset(surg, select = -c(gender) )
```

There are total 7 features in the data with 54 records. Survival is response variable while other 6 are predictor variables.

- Comment on why it is necessary to remove the gender variable to compute the correlation matrix.

For correlation matrix, only numeric variables are taken. While gender variable is of character type. So, with subset, gender variable is removed.

```
pairs(~survival+blood+prognosis+enzyme+liver+age,data=surg1,
   main="Scatterplot of Data")
```



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

```
cor_mat <- cor(surg1)
cor_mat
```

```
##                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
```

To evaluate the correlation between variables, correlation matrix is used. When the correlation between variables is above 0 it means positive correlation and if it's less than 1, it means negative correlation. From above analysis we can see that correlation between survival and liver is highest which is 0.67. Next most important predictor is enzyme with 0.57 correlation. After that prognosis and blood have correlation of about 0.42 and 0.34 respectively with response variable survival.

c. Fit a model using all the predictors to explain the survival response. Conduct an F-test for the overall regression i.e. is there any relationship between the response and the predictors. In your answer:

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

```
# with all variables
model <- lm(survival ~., data = surg)
model1 <- lm(survival ~ . -blood, data = surg)
model2 <- lm(survival ~ blood+prognosis, data = surg)
model3 <- lm(survival ~ blood+prognosis+enzyme+liver, data = surg)
```

- Write down the Hypotheses for the Overall ANOVA test of multiple regression.
- **Null Hypothesis H0: Liver, prognosis, blood and enzyme are not significant variables to predict survival**
- **Alternative Hypothesis H1: Liver, prognosis, blood and enzyme are significant variables to predict survival**
- Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient).

```
aov_mod <- anova(model)
aov_mod
```

```
## 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.

```
aov_mod$`F value`
```

```
## [1] 1.850596e+01 2.353853e+01 6.337420e+01 1.065305e+00 6.081592e-01
## [6] 1.062916e-05           NA
```

- State the Null distribution.

The null distribution defines the sets of data under null hypothesis. If the p-value is less than 0.05 which is significant level. Here from overall p-value we can see that `model3` is significant. While blood prognosis and enzyme are significant in terms of survival.

- Compute the P-Value

```
t.value = (mean(surg1$survival) - 10) / (sd(surg1$survival) / sqrt(length(surg1$survival)))
p.value = 2*pt(-abs(t.value), df=length(surg1$survival)-1)
p.value
```

```
## [1] 7.659093e-18
```

- State your conclusion (both statistical conclusion and contextual conclusion).

From the above analysis, we can see that model have some significant variables which are blood, prognosis and enzyme. The p-value of these variables is less than 0.05 which is proof of it's significance. The model gives the r-squared value of 0.69 which shows that it is giving 69% variance in model. The overall p-value is 1.19e-10 which shows that overall model is significant. If we talk about contextual conclusion we can see that p-value for survival is almost 7.659093e-18 which is showing it's significance.

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

```
summary(model)
```

```
##
## 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
```

```
summary(model1)
```

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -372.89 -147.52   -3.25   79.94 1079.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -728.8286   267.0699  -2.729 0.008853 **
## prognosis      6.7465     2.2731   2.968 0.004664 **
## enzyme         7.9223     1.8533   4.275 9.04e-05 ***
## liver        148.5332    41.6720   3.564 0.000837 ***
## age           -0.6013     3.2271  -0.186 0.852973
## genderM       31.2622    72.7195   0.430 0.669192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 253.8 on 48 degrees of freedom
## Multiple R-squared:  0.6305, Adjusted R-squared:  0.592
## F-statistic: 16.38 on 5 and 48 DF,  p-value: 2.072e-09
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = survival ~ blood + prognosis, data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -577.79 -190.77  -62.79  156.83 1469.11
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -327.523    240.851  -1.360  0.17986
## blood         77.141     29.721   2.595  0.01231 *
## prognosis      9.226      2.819   3.273  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 345.4 on 51 degrees of freedom
## Multiple R-squared:  0.2729, Adjusted R-squared:  0.2443
## F-statistic: 9.569 on 2 and 51 DF,  p-value: 0.0002961
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + liver, data = surg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -391.55 -144.81   -8.34  129.51  970.26
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1279.242    243.808  -5.247 3.30e-06 ***
## blood          82.988     26.402   3.143  0.00284 **
## prognosis       8.346      2.120   3.937  0.00026 ***
## enzyme         10.870      1.923   5.652 8.01e-07 ***
## liver          49.346     47.126   1.047  0.30018
## ---
## 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:  0.691,  Adjusted R-squared:  0.6658
## F-statistic:  27.4 on 4 and 49 DF,  p-value: 5.704e-12
```

From the above analysis, we can see that model3 is best model.

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

model3 is giving best variance i.e. 69% which is almost equal to model1 but the problem is model1 using all the variables while model3 is using just 4 positively related variables. Multiple regression model includes independent variables like continuous and categorical which is not appropriate for explanation of survival time.

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.

```
model4 <- lm(log(survival) ~., data = surg)
model5 <- lm(log(survival) ~ blood + prognosis, data = surg)
model6 <- lm(log(survival) ~ blood + prognosis + enzyme + liver, data = surg)
model7 <- lm(log(survival) ~ enzyme + liver + gender + age, data = surg)
```

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

```
summary(model4)
```

```
##
## Call:
## lm(formula = log(survival) ~ ., 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
```

```
summary(model5)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis, data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17639 -0.23716  0.05288  0.27434  1.17839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.23536    0.29992  17.456  < 2e-16 ***
## blood        0.06305    0.03701   1.704 0.094546 .
## prognosis    0.01313    0.00351   3.742 0.000465 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4302 on 51 degrees of freedom
## Multiple R-squared:  0.263,  Adjusted R-squared:  0.2341
## F-statistic: 9.099 on 2 and 51 DF,  p-value: 0.0004175
```

```
summary(model6)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme + liver,
##     data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43514 -0.17436 -0.02156  0.18475  0.56054
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.851933   0.266263  14.467  < 2e-16 ***
## blood       0.083739   0.028834   2.904  0.00551 **
## prognosis   0.012671   0.002315   5.474 1.50e-06 ***
## enzyme      0.015627   0.002100   7.440 1.38e-09 ***
```

```
## liver       0.032056   0.051466   0.623  0.53627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2509 on 49 degrees of freedom
## Multiple R-squared:  0.7591, Adjusted R-squared:  0.7395
## F-statistic: 38.61 on 4 and 49 DF,  p-value: 1.398e-14
```

```
summary(model7)
```

```
##
## Call:
## lm(formula = log(survival) ~ enzyme + liver + gender + age, data = surg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00378 -0.16995  0.03273  0.17638  0.66837
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.185026   0.302233  17.156  < 2e-16 ***
## enzyme       0.010808   0.002299   4.702 2.13e-05 ***
## liver        0.198753   0.048482   4.100 0.000155 ***
## genderM     -0.038040   0.092304  -0.412 0.682052
## age         -0.002187   0.004092  -0.534 0.595420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3222 on 49 degrees of freedom
## Multiple R-squared:  0.6027, Adjusted R-squared:  0.5702
## F-statistic: 18.58 on 4 and 49 DF,  p-value: 2.382e-09
```

Here `model4` is best model which gives the 77% variance. With log transformation survival variable will change into normal distribution. This will gives a model more accuracy.

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.

```
str(kml)
```

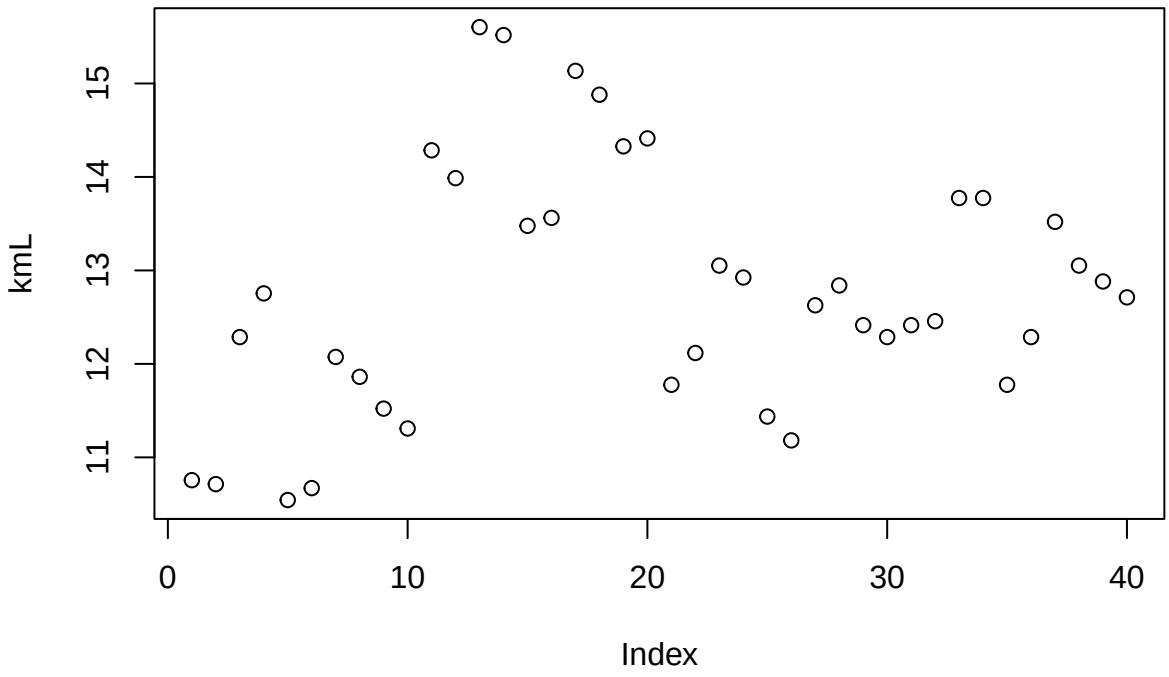
```
## 'data.frame':    40 obs. of  3 variables:
##  $ kmL   : num  10.8 10.7 12.3 12.8 10.5 ...
##  $ driver: chr  "A" "A" "A" "A" ...
##  $ car   : chr  "one" "one" "two" "two" ...
```

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

```
summary(kml$kmL)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.54   11.84   12.67   12.77   13.62   15.60
```

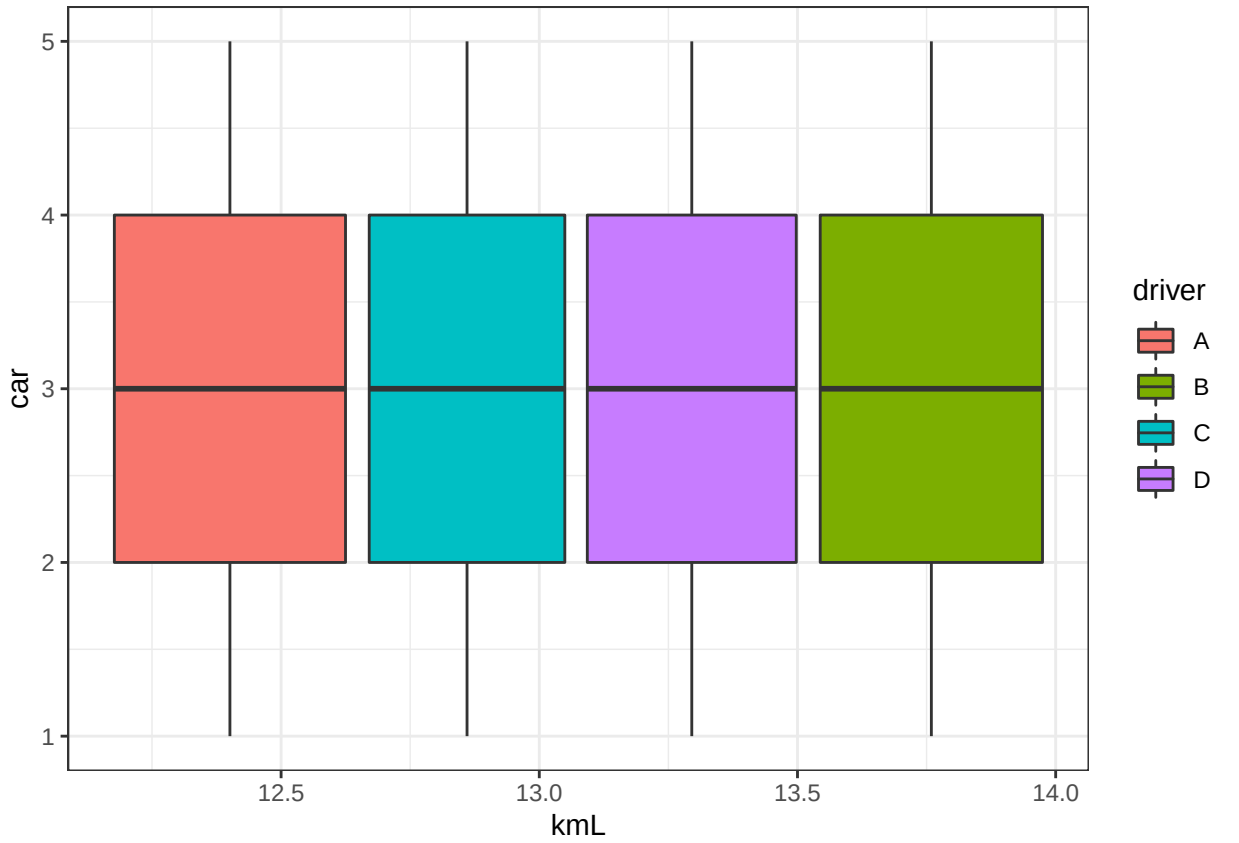
```
plot(kml$kmL, ylab = "kmL")
```



From the above plot and summary statistics, we can see that mean value is 12.77 and in plot we can see that values are almost equally distributes above and below this point with 10.54 minimum value and 15.60 maximum value.

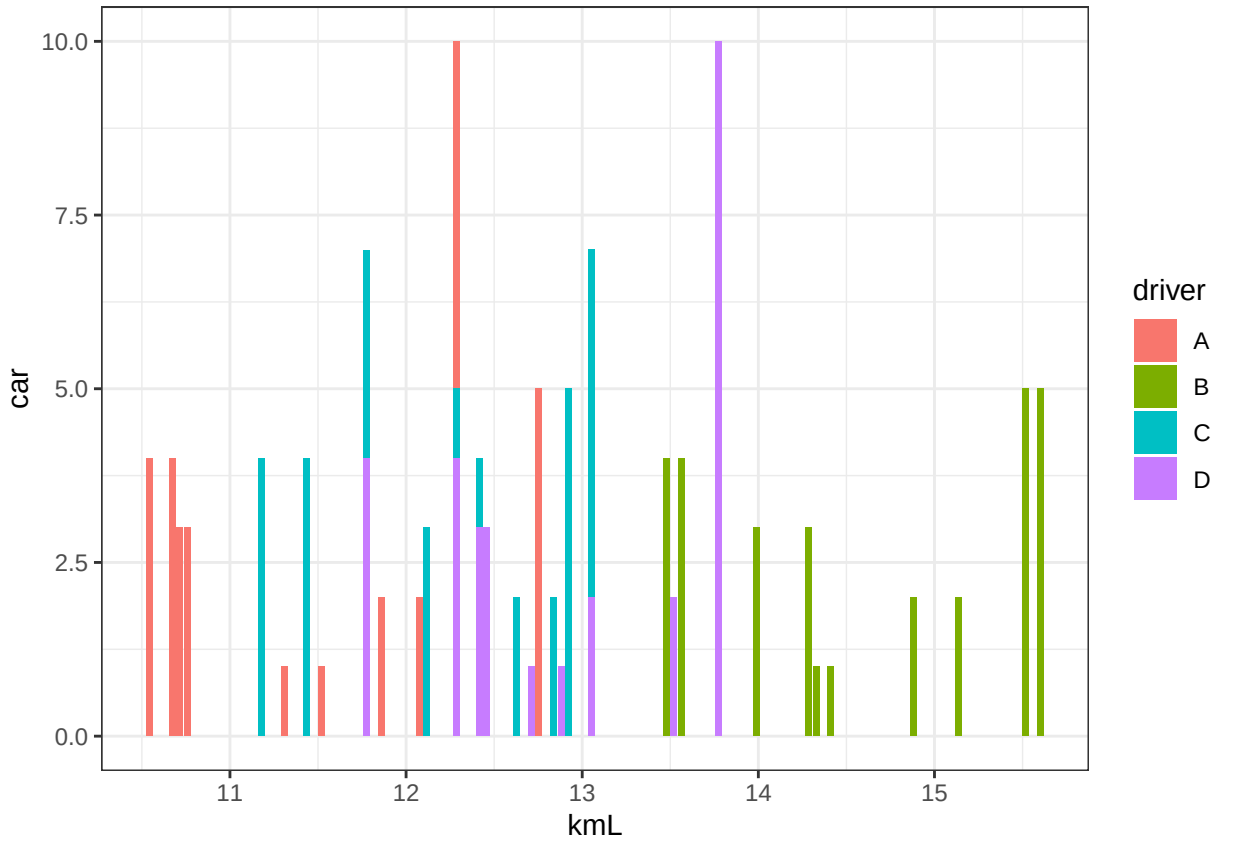
b. Construct two different preliminary graphs that investigate different features of the data and comment.

```
kml$driver <- as.factor(kml$driver)
kml$car <- as.numeric(as.factor(kml$car))
ggplot(kml, aes(x=kmL, y=car, fill=driver)) +
  geom_boxplot() + theme_bw()
```



From the above plot we can see that four drivers have different observed efficiency car's. Driver B is getting high efficient car with more than 13.5 km/L.

```
ggplot(data=kml, aes(x=kmL, y = car, fill=driver)) +
  geom_bar(stat="identity")+
  theme_bw()
```



From this plot, we can see that average efficiency of cars is between 12 to 13. While most people select driver A cars while car's with driver B has high efficiency of more than 15 here.

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

- **Null Hypothesis H0: Driver and car are not effecting the kmL**
- **Alternative Hypothesis H1: Driver and car are effecting the kmL**

```
summary(kml)
```

```
##       kmL        driver      car
##  Min.   :10.54   A:10   Min.   :1
##  1st Qu.:11.84   B:10   1st Qu.:2
##  Median :12.67   C:10   Median :3
##  Mean   :12.77   D:10   Mean   :3
##  3rd Qu.:13.62          3rd Qu.:4
##  Max.   :15.60          Max.   :5
```

```
mod <- lm(kmL~. , data = kml)
summary(mod)
```

```
##
## Call:
```

```
## lm(formula = kmL ~ ., data = kml)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14576 -0.55747  0.07015  0.64144  1.19040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.27694    0.32997  34.176  < 2e-16 ***
## driverB      3.06954    0.32012   9.589 2.51e-11 ***
## driverC      0.81628    0.32012   2.550   0.0153 *
## driverD      1.41573    0.32012   4.423 9.05e-05 ***
## car          0.05739    0.08003   0.717   0.4780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7158 on 35 degrees of freedom
## Multiple R-squared:  0.7396, Adjusted R-squared:  0.7098
## F-statistic: 24.85 on 4 and 35 DF,  p-value: 8.302e-10
```

From the above model we can see that p-value is less than 0.05 significant level so we reject null hypothesis and consider alternative hypothesis.

Hypothesis Test

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: kmL
##           Df Sum Sq Mean Sq F value    Pr(>F)
## driver     3 50.661 16.8869 32.9579 2.663e-10 ***
## car        1  0.264  0.2635  0.5143     0.478
## Residuals 35 17.933  0.5124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above anova test we can see that driver is more significant than car.

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.

From the above analysis, we first made scatter plot to check the design balance which gives the result that design is balanced. In part b, B driver car is more efficient with more than 15 efficiency. In part three, kmL is tested using linear model and anova hypothesis testing. Linear regression model gives the 73% variance and shows the significance of both driver and car variables. Variable driver is more significant according to model and also according to anova test as the p-value is less than 0.05 significant level.