Assignment 1

Raghav Setiya

5/27/2021

Question 1

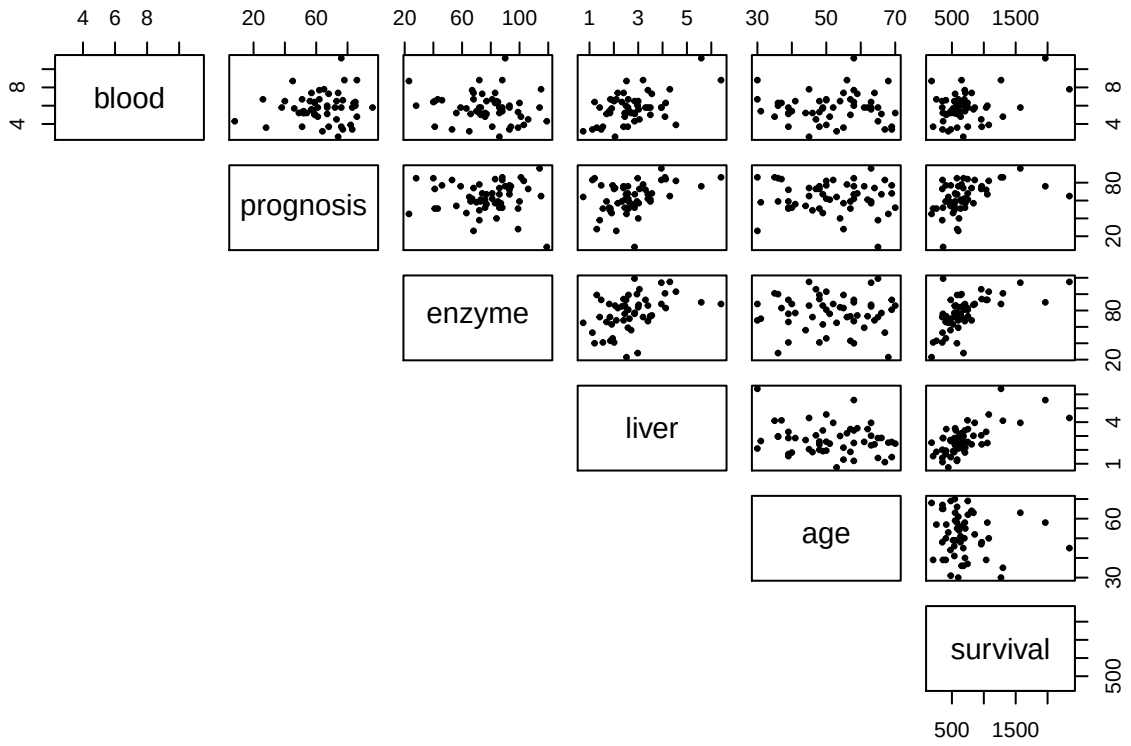
```
surg_data <- read.table("surg.txt", header = T)
summary(surg_data)
```

```
##      blood          prognosis         enzyme           liver
##  Min.   : 2.600   Min.   : 8.00   Min.   : 23.00   Min.   :0.740
##  1st Qu.: 5.025   1st Qu.:52.50   1st Qu.: 67.25   1st Qu.:2.020
##  Median : 5.800   Median :63.00   Median : 79.00   Median :2.595
##  Mean   : 5.783   Mean   :63.24   Mean   : 77.11   Mean   :2.744
##  3rd Qu.: 6.500   3rd Qu.:76.00   3rd Qu.: 89.50   3rd Qu.:3.275
##  Max.   :11.200   Max.   :96.00   Max.   :119.00   Max.   :6.400
##       age           gender             survival
##  Min.   :30.00   Length:54          Min.   : 181.0
##  1st Qu.:44.25   Class :character   1st Qu.: 482.0
##  Median :51.50   Mode  :character   Median : 605.5
##  Mean   :51.61                      Mean   : 702.1
##  3rd Qu.:60.50                      3rd Qu.: 750.5
##  Max.   :70.00                      Max.   :2343.0
```

A)

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

```
ns <- subset(surg_data, select = -c(gender))
pairs(ns, pch = 19,  cex = 0.5,
      lower.panel=NULL)
```



To make the scatter plot,we need to use the Pairs function. This function is compatible with numeric variables while character variables are not compatible with this,therefore,we need to remove character variables like gender. From the above graph, we can see that there is almost linear relationship with every predictor. All the variables are predictors while survival is the response variable.

B)

Compute the correlation matrix of the dataset and comment.

```
cor(ns)
```

```
##                 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
```

From the above correlation matrix, we can see that some values are negative and some are positive. Values with more than 0.5 are known as highly correlated variables,whereas negative values shows negative correlation. All the predictors have positive correlation with response variable except age.

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. - 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). - Compute the F statistic for this test. - State the Null distribution. - Compute the P-Value - State your conclusion (both statistical conclusion and contextual conclusion).

```
#####################################################################
lin_mod1 <- lm(survival ~., data = surg_data)
lin_mod2 <- lm(survival ~ . -blood, data = surg_data)
lin_mod3 <- lm(survival ~ liver+age, data = surg_data)
lin_mod4 <- lm(survival ~ liver+age+gender+prognosis, data = surg_data)
lin_mod5 <- lm(survival ~ blood+prognosis+enzyme+liver, data = surg_data)

#####################################################################
#H0: liver age and gender are not significant variables to predict survival
#H1: liver age and gende are significant variables to predict survival

#####################################################################
anova_mod <- anova(lin_mod1)
anova_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
```

```
#####################################################################
anova_mod$`F value`
```

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

```
#####################################################################
# Null Distribution
# Our null distribution is then Normal. Now that we have a null distribution, we need to dream up a tes

#####################################################################
t_value = (mean(surg_data$survival) - 10) / (sd(surg_data$survival) / sqrt(length(surg_data$survival)))
t_value
```

```
## [1] 12.7982
```

```
p_value = 2*pt(-abs(t_value), df=length(surg_data$survival)-1)
p_value
```

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

From the correlation matrix,it is found that variables like blood, prognosis and enzyme are highly significant. The final P-value is about 7.659093e-18 which is less than overall p-value significance level (0.05). The maximum variance in model is about 69%. As Statistical conclusion, our p-value is about 1.19*10^-10 which is also less than 0.05 and therefore,we reject null hypothesis.

D)

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

```
summary(lin_mod1)
```

```
##
## Call:
## lm(formula = survival ~ ., data = surg_data)
##
## 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(lin_mod2)
```

```
##
## Call:
## lm(formula = survival ~ . - blood, data = surg_data)
##
## 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(lin_mod3)
```

```
##
## Call:
## lm(formula = survival ~ liver + age, data = surg_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -485.67 -170.01  -43.62  155.08 1254.01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -29.1387   244.7808  -0.119    0.906
## liver       251.9656    39.2328   6.422 4.45e-08 ***
## age           0.7706     3.7754   0.204    0.839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 299.1 on 51 degrees of freedom
## Multiple R-squared:  0.455,  Adjusted R-squared:  0.4336
## F-statistic: 21.29 on 2 and 51 DF,  p-value: 1.899e-07
```

```
summary(lin_mod4)
```

```
##
## Call:
## lm(formula = survival ~ liver + age + gender + prognosis, data = surg_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -459.15 -160.20  -49.67  133.18 1270.02
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -266.2630   283.9558  -0.938   0.3530
## liver        228.0779    43.3629   5.260 3.16e-06 ***
```

```
## age            0.6392     3.7378   0.171   0.8649
## genderM       26.0504    84.5586   0.308   0.7593
## prognosis      4.6722     2.5826   1.809   0.0766 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 295.2 on 49 degrees of freedom
## Multiple R-squared:  0.4899, Adjusted R-squared:  0.4482
## F-statistic: 11.76 on 4 and 49 DF,  p-value: 8.952e-07
```

```
summary(lin_mod5)
```

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

Model 5 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.

The linear model 1 and final model is about same. Therefore, we can say that blood, prognosis, enzyme and liver are the most efficient variables in terms of regression model.

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.

```
lin_mod6 <- lm(log(survival) ~., data = surg_data)
lin_mod7 <- lm(log(survival) ~ . -blood, data = surg_data)
lin_mod8 <- lm(log(survival) ~ liver+age, data = surg_data)
```

```
lin_mod9 <- lm(log(survival) ~ liver+age+gender+prognosis, data = surg_data)
lin_mod10 <- lm(log(survival) ~ blood+prognosis+enzyme+liver, data = surg_data)
```

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(lin_mod6)
```

```
##
## Call:
## lm(formula = log(survival) ~ ., data = surg_data)
##
## 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(lin_mod7)
```

```
##
## Call:
## lm(formula = log(survival) ~ . - blood, data = surg_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44926 -0.18289 -0.04884  0.17867  0.63901
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.594056   0.286228  16.050  < 2e-16 ***
## prognosis    0.011098   0.002436   4.556 3.59e-05 ***
## enzyme       0.012740   0.001986   6.414 5.83e-08 ***
## liver        0.117326   0.044661   2.627   0.0115 *
```

```
## age         -0.002959   0.003459  -0.856   0.3965
## genderM     -0.031673   0.077936  -0.406   0.6863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.272 on 48 degrees of freedom
## Multiple R-squared:  0.7226, Adjusted R-squared:  0.6937
## F-statistic: 25.01 on 5 and 48 DF,  p-value: 2.584e-12
```

```
summary(lin_mod8)
```

```
##
## Call:
## lm(formula = log(survival) ~ liver + age, data = surg_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15753 -0.18840  0.01151  0.28354  0.86327
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.6400962  0.3119077  18.083  < 2e-16 ***
## liver        0.2970863  0.0499917   5.943 2.52e-07 ***
## age         -0.0004812  0.0048108  -0.100    0.921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3811 on 51 degrees of freedom
## Multiple R-squared:  0.4216, Adjusted R-squared:  0.3989
## F-statistic: 18.59 on 2 and 51 DF,  p-value: 8.654e-07
```

```
summary(lin_mod9)
```

```
##
## Call:
## lm(formula = log(survival) ~ liver + age + gender + prognosis,
##     data = surg_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00111 -0.20100  0.00417  0.19750  0.94563
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.3378925  0.3529386  15.124  < 2e-16 ***
## liver        0.2452388  0.0538973   4.550 3.55e-05 ***
## age         -0.0009641  0.0046458  -0.208   0.8365
## genderM     -0.0400536  0.1051009  -0.381   0.7048
## prognosis    0.0077627  0.0032100   2.418   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3669 on 49 degrees of freedom
```

```
## Multiple R-squared:  0.4849, Adjusted R-squared:  0.4428
## F-statistic: 11.53 on 4 and 49 DF,  p-value: 1.126e-06
```

```
summary(lin_mod10)
```

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

Model 6 gives the best significance.

Question 2

```
kml_data <- read.table("kml.txt", header = T)
summary(kml_data)
```

```
##       kmL           driver              car
##  Min.   :10.54   Length:40          Length:40
##  1st Qu.:11.84   Class :character   Class :character
##  Median :12.67   Mode  :character   Mode  :character
##  Mean   :12.77
##  3rd Qu.:13.62
##  Max.   :15.60
```

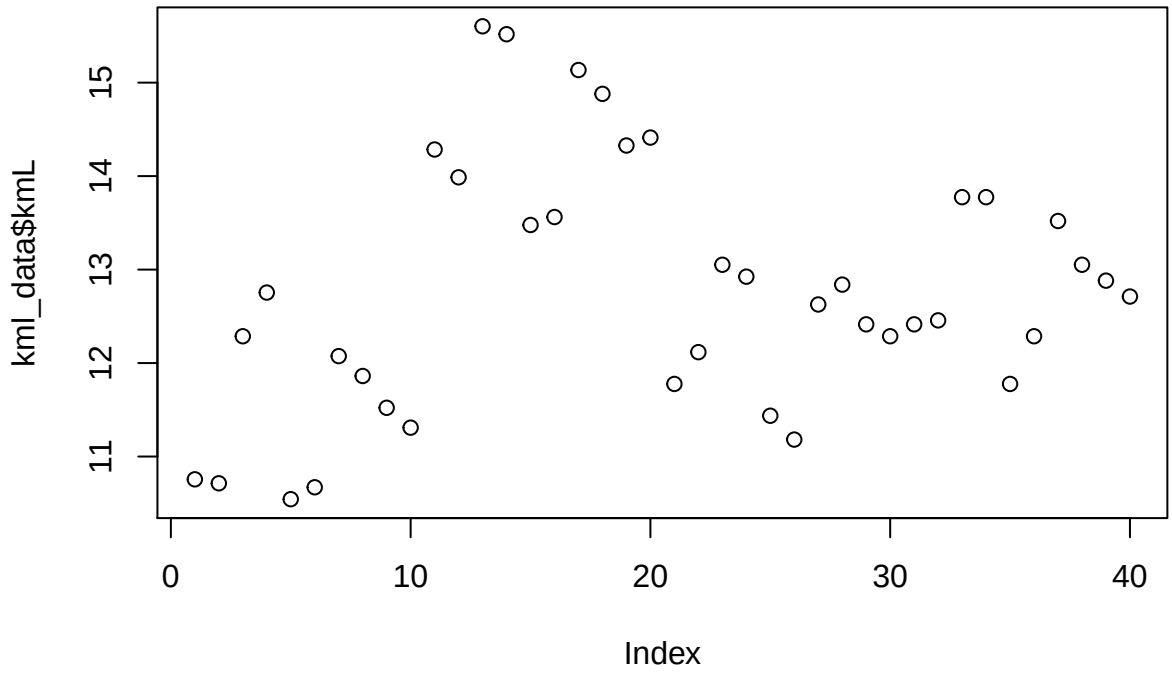
A)

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

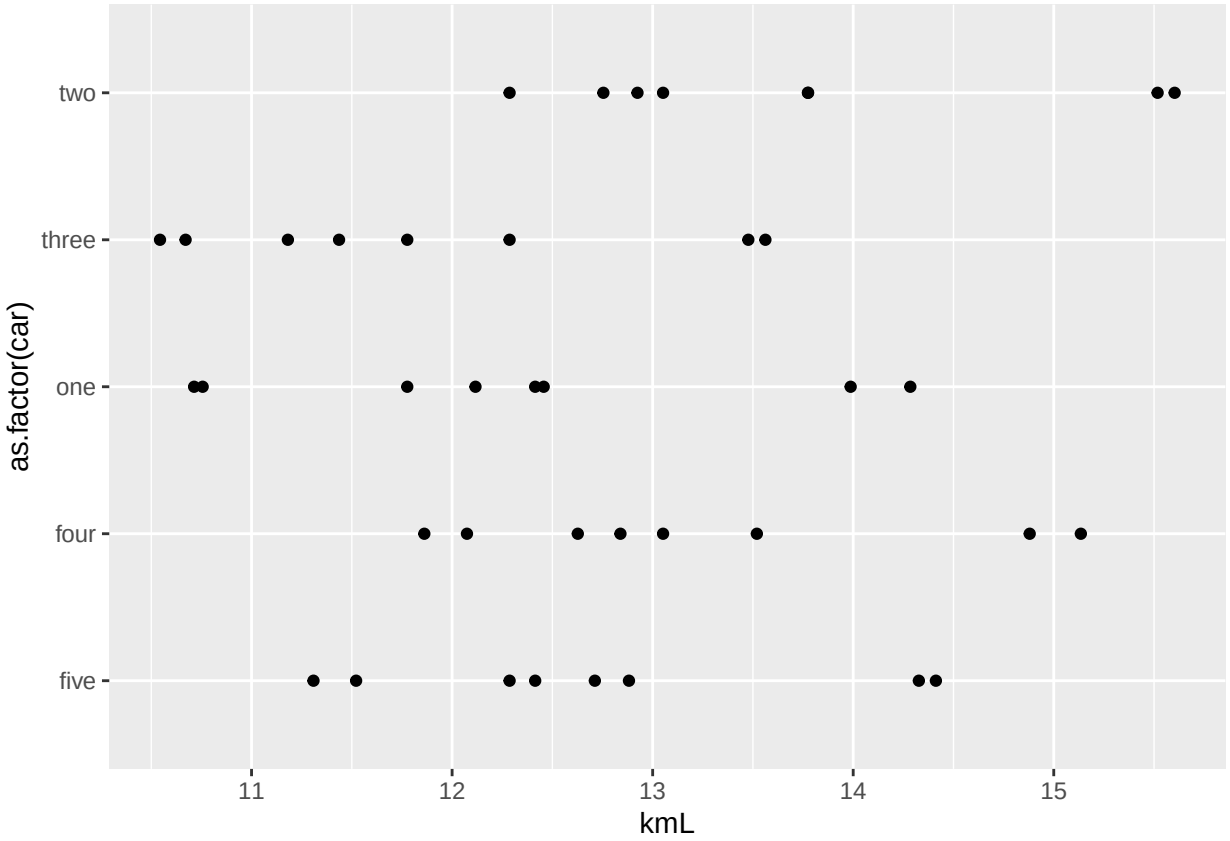
```
summary(kml_data$kmL)
```

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

```
plot(kml_data$kmL)
```




```
ggplot(kml_data, aes(x=kmL, y= as.factor(car))) +
  geom_point()
```

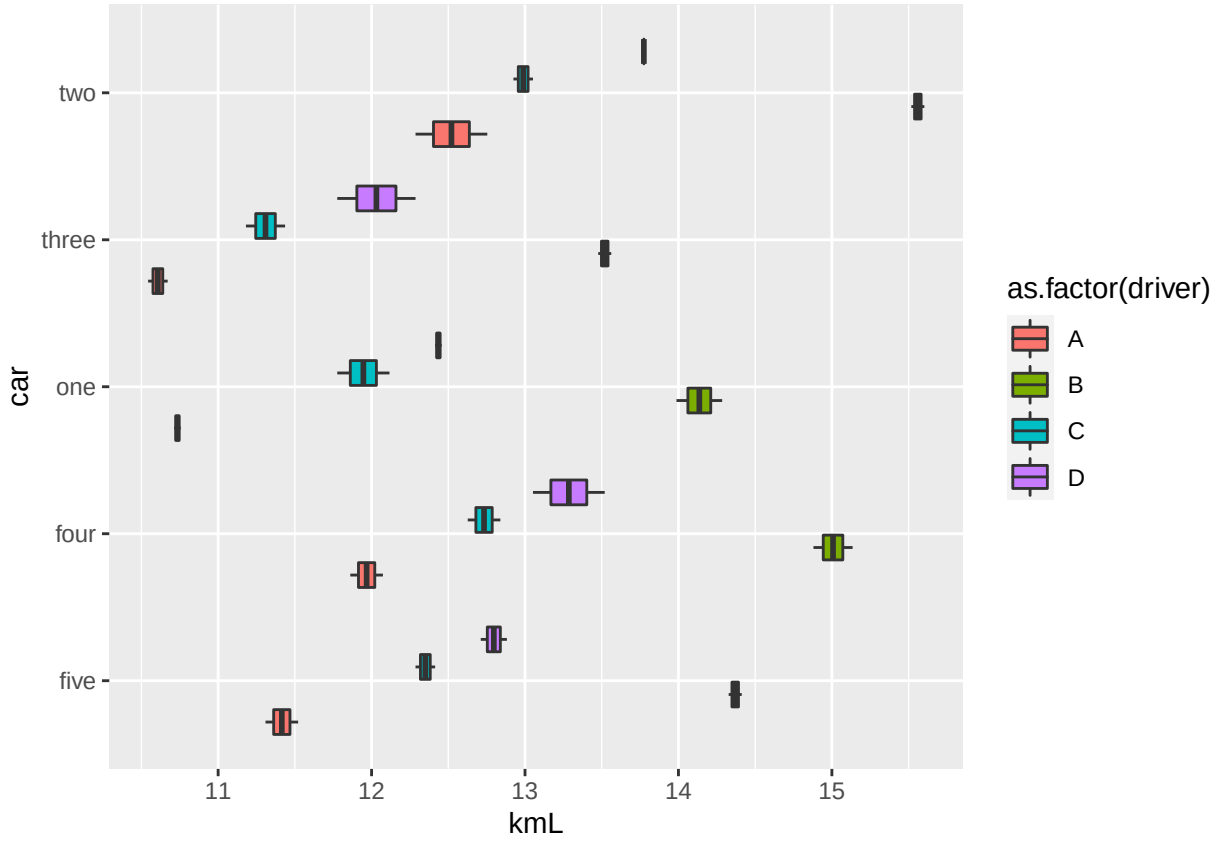


From the above plot, we can see that data is distributed simultaneously in whole graph. This shows that the design is balanced.

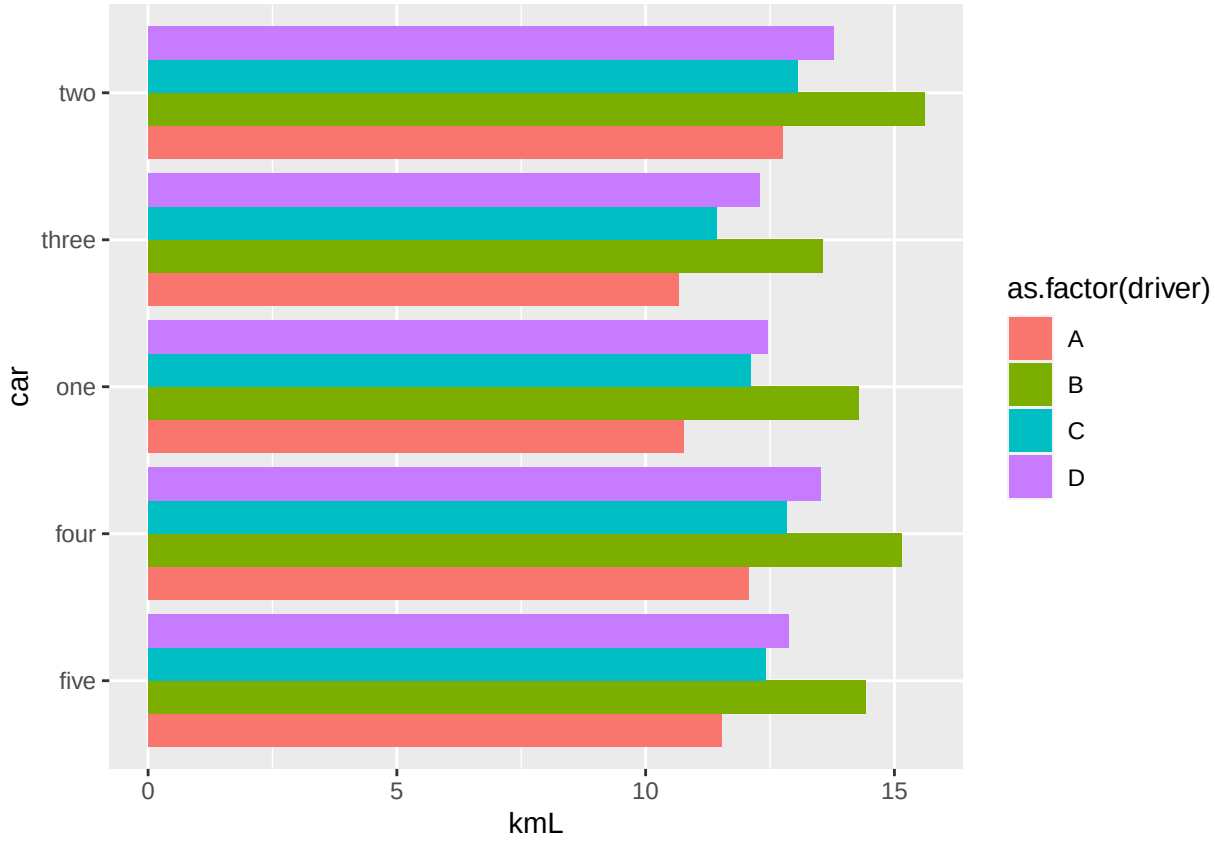
B)

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

```
ggplot(kml_data, aes(x=kmL, y=as.factor(car), fill=as.factor(driver))) +
  geom_boxplot() + xlab("kmL") + ylab("car")
```




```
ggplot(kml_data, aes(x=kmL, y=as.factor(car), fill=as.factor(driver))) +
  geom_bar(position="dodge", stat="identity")+ xlab("kmL") + ylab("car")
```



Each driver is working on different cars. The driver B is riding most efficient cars. It has high km/l efficiency,therefore, the plot stacked above shows that driver b drives efficient cars.

C)

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

```
#H0: Driver & car are not effecting the kmL
#H1: Driver & car are effecting the kmL
str(kml_data)
```

```
## '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" ...
```

```
summary(kml_data)
```

```
##       kmL           driver              car
##  Min.   :10.54   Length:40          Length:40
##  1st Qu.:11.84   Class :character   Class :character
##  Median :12.67   Mode  :character   Mode  :character
##  Mean   :12.77
##  3rd Qu.:13.62
##  Max.   :15.60
```

```r
model_check <- lm(kmL~. , data = kml_data)
summary(model_check)
```

```
##
## Call:
## lm(formula = kmL ~ ., data = kml_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28697 -0.11266 -0.02657  0.11771  0.36881
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.40768    0.08206 139.009  < 2e-16 ***
## driverB      3.06954    0.08206  37.404  < 2e-16 ***
## driverC      0.81628    0.08206   9.947 2.58e-11 ***
## driverD      1.41573    0.08206  17.251  < 2e-16 ***
## carfour      0.51549    0.09175   5.618 3.29e-06 ***
## carone      -0.41983    0.09175  -4.576 6.79e-05 ***
## carthree    -0.86623    0.09175  -9.441 9.08e-11 ***
## cartwo       0.97783    0.09175  10.657 4.67e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1835 on 32 degrees of freedom
## Multiple R-squared:  0.9844, Adjusted R-squared:  0.9809
## F-statistic: 287.6 on 7 and 32 DF,  p-value: < 2.2e-16
```

Model gives 98% variance showing that it is working perfectly. P value is less than 0.05 showing that null hypothesis is rejected. So, alternative hypothesis is considered and driver and the car are therefore effecting the kmL is accepted. By doing a hypothesis test.

```r
anova(model_check)
```

```
## Analysis of Variance Table
##
## Response: kmL
##           Df Sum Sq Mean Sq F value    Pr(>F)
## driver     3 50.661 16.8869   501.5 < 2.2e-16 ***
## car        4 17.119  4.2798   127.1 < 2.2e-16 ***
## Residuals 32  1.078  0.0337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It looks that driver is more significant than car and therefore, car efficiency depends more on the driver.

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.

We used the results of the above research to create a scatter plot to assess the design balance, which revealed that the design is balanced. Part b shows that the B car driver is more efficient, with an efficiency of over 15 percent. In part three, the linear model and anova hypothesis testing are used to evaluate kmL. The 98 percent variance is calculated using a linear regression model, which demonstrates the significance of both the driver and the car variables. Driver variable is more significant than the car variable according to the anova results.