Simple Linear Regression and ANOVA

Source One: http://www.r-tutor.com/elementary-statistics/simple-linear-regression

Source Two: https://datascienceplus.com/one-way-anova-in-r/

Simple Linear Regression:

Data

The first thing to do, is to look at the dataset "Faithful" that will be used in this document.

```
summary(faithful)
```

```
##    eruptions        waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

Estimated Simple Regression Equation

Below I create an estimated simple regression equation.

```
eruption.lm = lm(eruptions ~ waiting, data=faithful)
```

Looking at the coefficients we can determine that the Y-intercept is -1.9 and the X coefficient is .075 indicating a positive correlation to eruption times.

```
coeffs = coefficients(eruption.lm); coeffs
```

```
## (Intercept)     waiting
## -1.87401599  0.07562795
```

Creating a Prediction

If we use this to create an estimation, we can determine that if the last eruption was 80 minutes ago, then the next eruption would last 4.2 minutes.

```
newdata = data.frame(waiting=80) # wrap the parameter
predict(eruption.lm, newdata)    # apply predict
```

```
##       1
## 4.17622
```

Coeffient of Determination

The coefficent of determination is the variance of fitted values divided by the variance of the observed values. We can get this value as shown below. This means that 81% of the variance in eruption time can be explained by time since the last eruption.

```
summary(eruption.lm)$r.squared
```

```
## [1] 0.8114608
```

Significace Test for Linear Regression

Null Hypthosis: βB = 0 Alternative Hypthosis: βB != 0

```
summary(eruption.lm)
```

```
##
## Call:
## lm(formula = eruptions ~ waiting, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## waiting      0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Since the p-value is less than .05 we can reject the null hypthesis and determine that there is a significant relationship between X and Y

Confidence Interval

Below we create a 95% confidence interval for the mean of the Y variable if the waiting time is 80 minuted. The 95% confidence interval of the mean eruption duration for the waiting time of 80 minutes is between 4.1 and 4.25 minutes.

```
newdata = data.frame(waiting=80)
predict(eruption.lm, newdata, interval="confidence")
```

```
##       fit      lwr      upr
## 1 4.17622 4.104848 4.247592
```

Prediction Interval

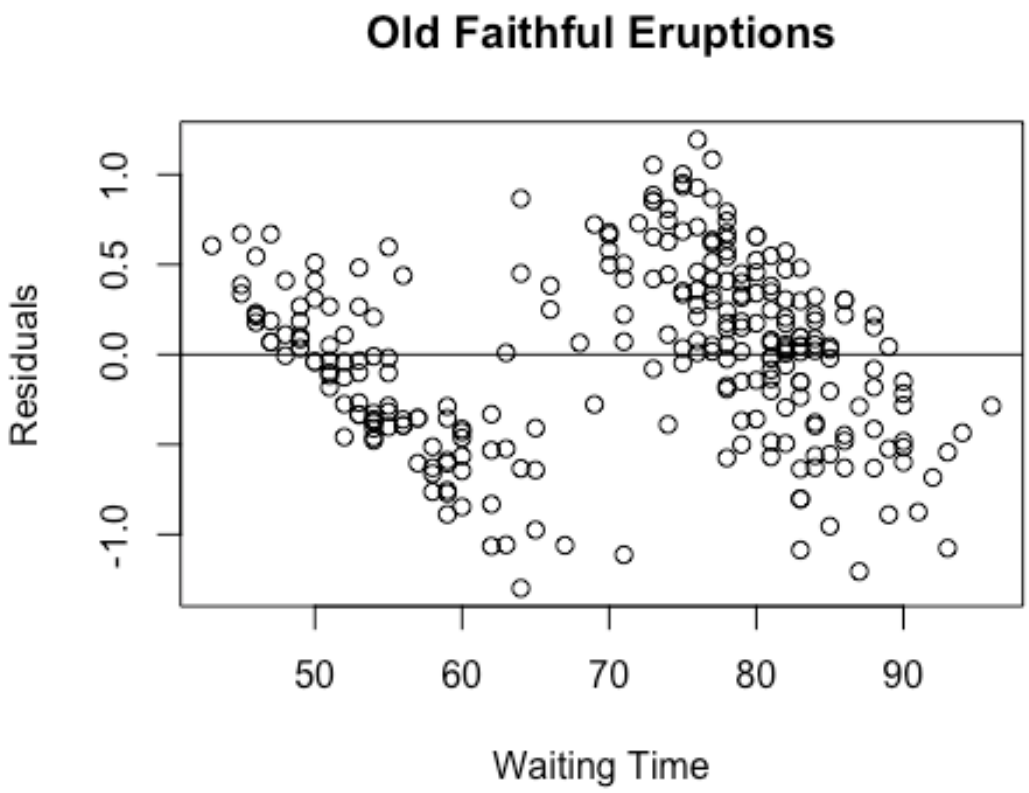
By changing the argument in the predict function from “confidence” to “predict”, we can then see the prediction interval. The 95% prediction interval of the eruption duration for the waiting time of 80 minutes is between 3.2 and 5.2 minutes.

```
predict(eruption.lm, newdata, interval="predict")

##        fit      lwr      upr
## 1 4.17622 3.196089 5.156351
```

Residual Plot

Below we get the difference between the expected values and the obvserved values which is called the residual. We can then plot to see the residual against the observed value.

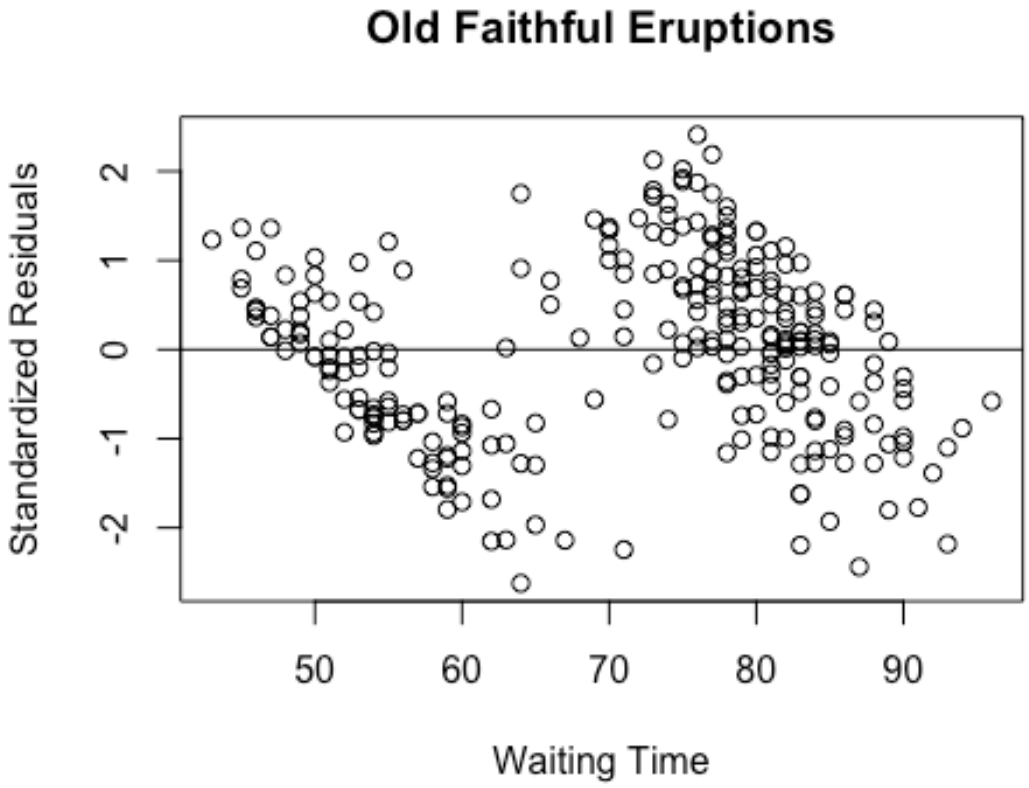
```
eruption.res = resid(eruption.lm)
plot(faithful$waiting, eruption.res,
     ylab="Residuals", xlab="Waiting Time",
     main="Old Faithful Eruptions")
     abline(0, 0)
```



Standardized Residual

We can standardize the graph as well to compare the residuals.

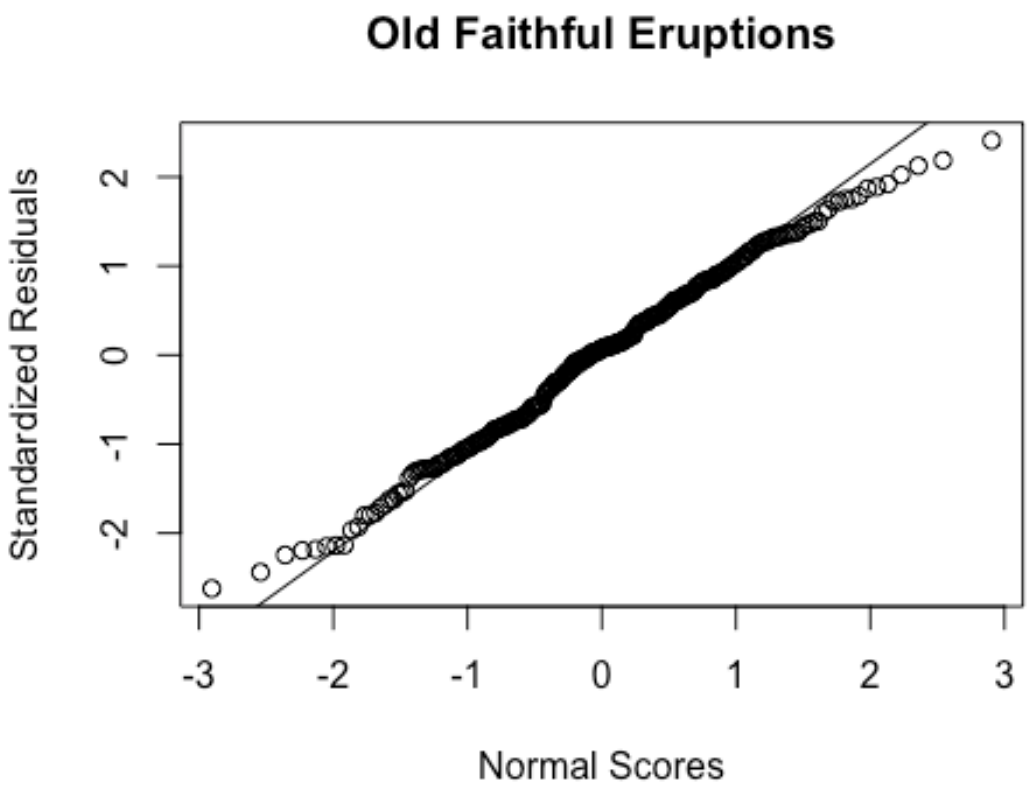
```
eruption.stdres = rstandard(eruption.lm)
plot(faithful$waiting, eruption.stdres,
     ylab="Standardized Residuals",
     xlab="Waiting Time",
     main="Old Faithful Eruptions")
     abline(0, 0)                  # the horizon
```



Normal Probability Plot of Residuals

Below we create a normal probability plot that we can use to determine if the error term in our equation is normally distributed.

```
qqnorm(eruption.stdres,
     ylab="Standardized Residuals",
     xlab="Normal Scores",
     main="Old Faithful Eruptions")
    qqline(eruption.stdres)
```



Simple Linear Regression Questions - Based on tutorial

SLR1: For the eruption.lm model, what are the estimated coefficients and what is the estimated regression equation?

β1 = .075 β0 = -1.9 Y = .075X - 1.9 + Error

SLR2: Interpret the coefficients of β0 and β1 in the context of the problem.

β0 = -1.9: Indicates the y-intercept if the waiting time is 0 β1 = .075: Indicates a positive relationship between waiting time and eruption duration

SLR3: What would the model predict for a waiting time of 65, a 99% significance level with an interval = "prediction"? Interpret.

Running the code below we can interprest that the 99% prediction interval of the eruption duration for the waiting time of 65 minutes is between 1.75 and 4.33 minutes.

```
newdata = data.frame(waiting=65)
predict(eruption.lm, newdata, interval="predict", level=.99)
```

```
##        fit      lwr      upr
## 1 3.041801 1.750955 4.332647
```

SLR4: Interpret the output of the QQ Normal plot (above) against the Normality assumption.

The QQ plot above indicates the the data is mostly normal. This is because the data follows along the horizontal black line. There appears to be non-normality closer to the edges, but overall a normal distribution. Since this is a QQ plot of the residuals, we can tell that our error term is normally distributed.

SLR5: Interpret the Standardized Residuals plot (above) against the Constant Variance assumption.

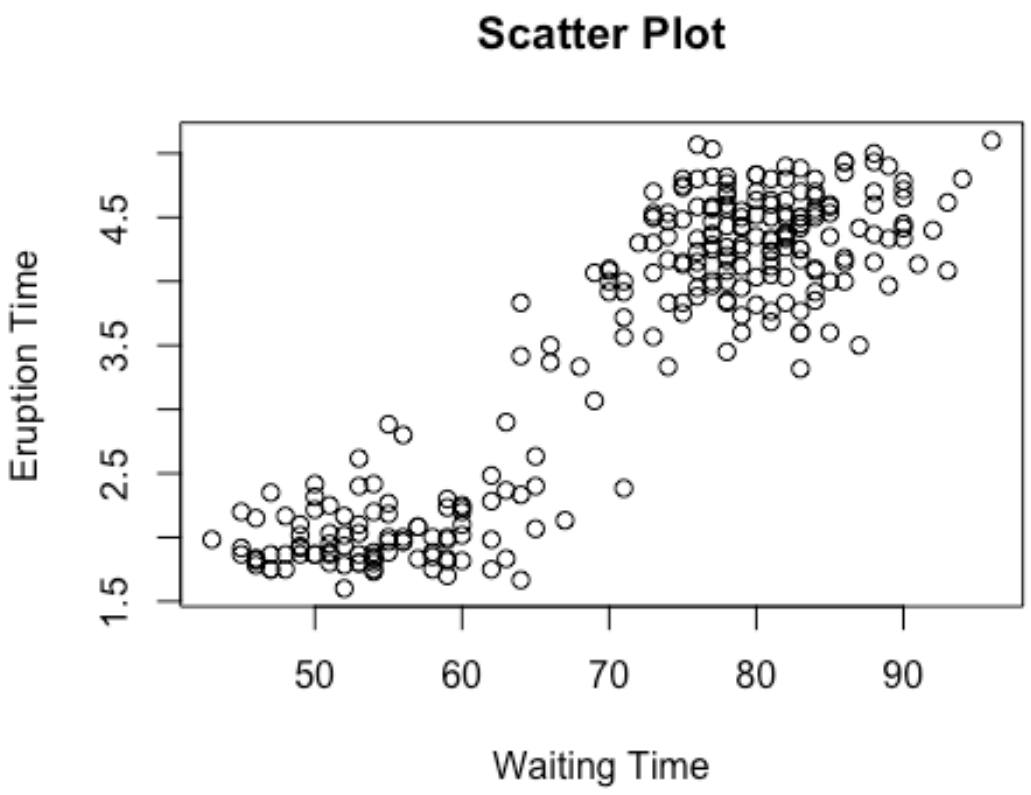
The standardized residuals plot shows that the error term might not be normally distributed. It looks like it is possible that the mean=0 of the residuals but it also looks like there might be a pattern on the graph.

SLR6: Interpret the Standardized Residuals plot (above) against the Independence assumption

Using both graphs, I'm going to say that the error term is mostly normally distributed. Even though there is some deviance on the diagonal in the QQ plot and there appears to be a minor pattern on the Standard Residuals Plot, I'm going to say overall it appears to be normally distributed.

SLR7: Create a scatterplot of the two variables

```
plot(faithful$waiting, faithful$eruption,
     ylab="Eruption Time",
     xlab="Waiting Time",
     main="Scatter Plot")
     abline(0, 0)                  # the horizon
```



SLR8: Interpret the scatterplot against the Linearity assumption Linearity means that the predictor variable in the regression has a straight-line relationship with the response variable. By looking at the scatterplot above we can see that straight line relationship

ANOVA:

Null Hypothesis: The mean lifetime for Apollo, Bridgestone, CEAT and Falken are the same.
Alternative Hypthosis: The mean lifetimes differ.

```
tyre<- read.csv('tyre.csv',header = TRUE, sep = ",")
attach(tyre)
```

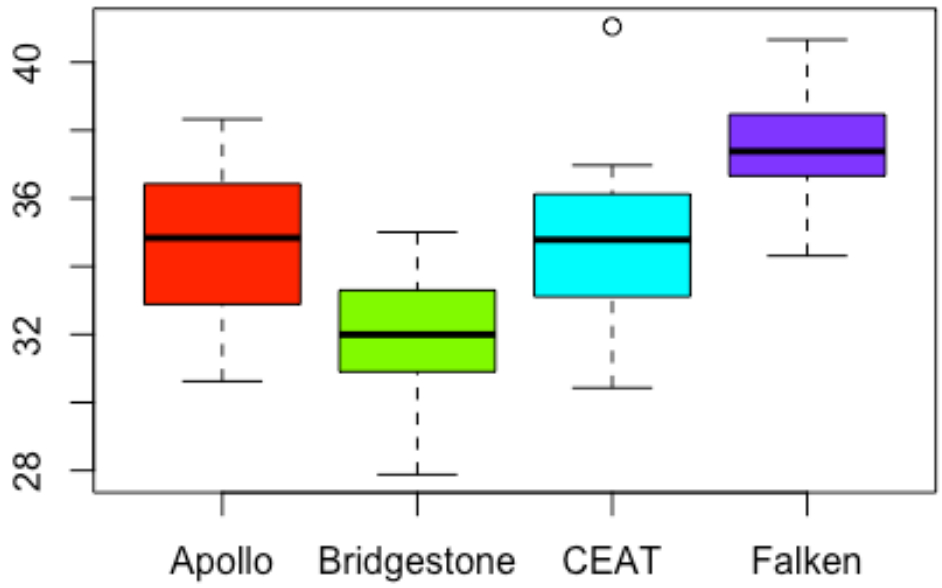
Checking below to ensure that brands is categorical.

```
is.factor(Brands)

## [1] TRUE
```

In order to check for outliers, we use the below boxplot. We can see that there is an outlier for CEAT.

```
boxplot(Mileage~Brands, main="Fig.-1: Boxplot of Mileage of Four Brands of
Tyre", col= rainbow(4))
```

Fig.-1: Boxplot of Mileage of Four Brands of Tyre



In order to see what that one outlier is:

```
boxplot.stats(Mileage[Brands=="CEAT"])
```

```
## $stats
## [1] 30.42748 33.11079 34.78336 36.12533 36.97277
##
## $n
## [1] 15
##
## $conf
## [1] 33.55356 36.01316
##
## $out
## [1] 41.05
```

We can ignore this outlier for now but may come back to it later. Below we are going to fit our model using the aov() function and print out the summary for the model. We can see the F-statistic is 17.94 and the p-value is .01, we can reject the null hypothesis and determine that the average mileage of the four tyre brands are not equal.

```
model1<- aov(Mileage~Brands)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Brands       3  256.3   85.43   17.94 2.78e-08 ***
## Residuals   56  266.6    4.76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we want to determine the brand that differs from the rest. For this we can use the Turkey's HSD test set for 99% confidence.

```
TukeyHSD(model1, conf.level = 0.99)
```

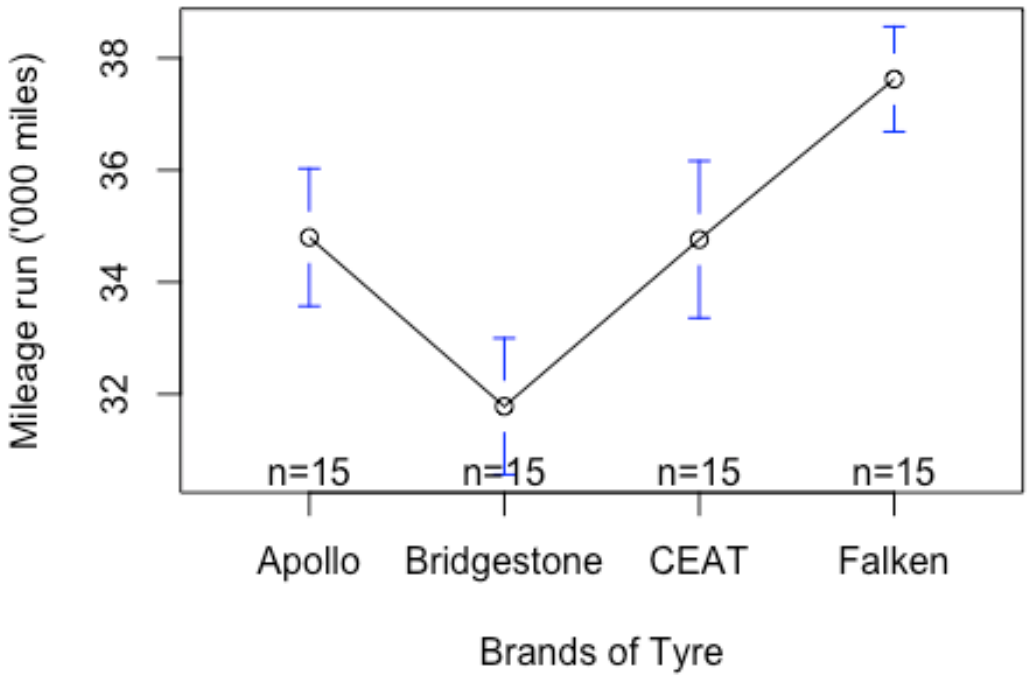
```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = Mileage ~ Brands)
##
## $Brands
##                            diff        lwr        upr     p adj
## Bridgestone-Apollo -3.01900000 -5.6155816 -0.4224184 0.0020527
## CEAT-Apollo        -0.03792661 -2.6345082  2.5586550 0.9999608
## Falken-Apollo       2.82553333  0.2289517  5.4221149 0.0043198
## CEAT-Bridgestone    2.98107339  0.3844918  5.5776550 0.0023806
## Falken-Bridgestone  5.84453333  3.2479517  8.4411149 0.0000000
## Falken-CEAT         2.86345994  0.2668783  5.4600415 0.0037424
```

Plotting the results:

```
plot(TukeyHSD(model1, conf.level = 0.99),las=1, col = "red")
```

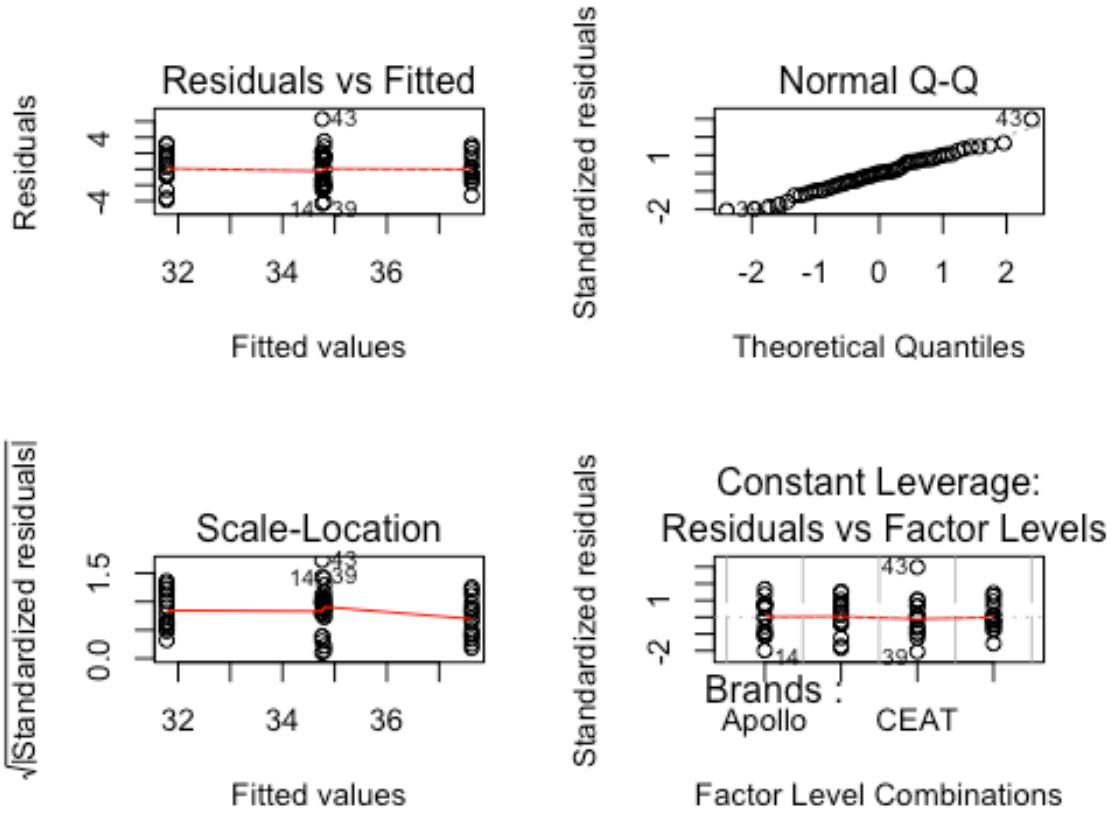

```
library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

plotmeans(Mileage~Brands, main="Fig.-3: Mean Plot with 95% Confidence
Interval", ylab = "Mileage run ('000 miles)", xlab = "Brands of Tyre")
```

Fig.-3: Mean Plot with 95% Confidence Interval



Diagnostics Checking

```
par(mfrow=c(2,2))
plot(model1)
```



Checking for normality: Because the p-value is above the level of significance, it implies that the samples were taken from normal populations.

```
uhat<-resid(model1)
shapiro.test(uhat)

##
##  Shapiro-Wilk normality test
##
## data:  uhat
## W = 0.9872, p-value = 0.7826
```

ANOVA Questions – Based on tutorial

A1: Interpret the Boxplot (above) within the construct of the question: Is the between-variability significantly larger than the within-variability?

The between variability is significantly larger than the within variability. The range of Appollo alone is from 31 to 39 where the range for everyone is from 28-41.

A2: For model1, what is the F-Value and the p-value of the test?

We can see the F-statistic is 17.94 and the p-value is .01, we can reject the null hypothesis and determine that the average mileage of the four tyre brands are not equal.

A3: For model1, what is the value of the total Sum of Squares?

The sum of squares is 266.65

```
sum(resid(model1)^2)
```

```
## [1] 266.6494
```

A4: For model 1, what is the degrees of freedom from the source of variability for treatment?

56

```
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Brands       3  256.3   85.43   17.94 2.78e-08 ***
## Residuals   56  266.6    4.76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A5: For model1, what is the grand mean and individual means of each brand's mileage?

```
mean(tyre$Mileage)
```

```
## [1] 34.74129
```

```
aggregate(tyre$Mileage, list(tyre$Brands), mean)
```

```
##       Group.1        x
## 1      Apollo 34.79913
## 2 Bridgestone 31.78013
## 3        CEAT 34.76121
## 4      Falken 37.62467
```

A6: Interpret the pairwise Falken-Bridgestone comparison. Which brand has more tyre mileage? How many more miles does one brand have over the other?

The pair-wise difference between Falken-Bridgestone is found to be 5.84 which means that Falken has higher mileage than Bridgestone and this is statistically significant.

```
TukeyHSD(model1, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = Mileage ~ Brands)
##
## $Brands
##                           diff        lwr        upr     p adj
## Bridgestone-Apollo -3.01900000 -5.6155816 -0.4224184 0.0020527
## CEAT-Apollo        -0.03792661 -2.6345082  2.5586550 0.9999608
## Falken-Apollo       2.82553333  0.2289517  5.4221149 0.0043198
```

```
## CEAT-Bridgestone    2.98107339  0.3844918  5.5776550 0.0023806
## Falken-Bridgestone  5.84453333  3.2479517  8.4411149 0.0000000
## Falken-CEAT         2.86345994  0.2668783  5.4600415 0.0037424
```