Assignment 2
FIT2086 - Modelling for Data Analysis

Question 1:

1. To be able to calculate an estimate of the average number of dog-bites for days on which there was a full moon, we must first sort out our data in the given *dogbites.fullmoon.csv* based on the value of *is.full.moon == 1*. We could do this by using the following code in R.

```
dogbites_fullmoon = read.csv("dogbites.fullmoon.csv", header = TRUE)
full_moon = dogbites_fullmoon[dogbites_fullmoon$is.full.moon==1,]
```

The result of the code above would give us the list of the number of dog bites only when there is a full moon, we could see this by using the following code, which will show us some the top of the list full_moon as the whole list would not fit in this report.

```
> head(full_moon)
    daily.dogbites is.full.moon
8                0            1
38               3            1
67               4            1
97               3            1
126              6            1
156              3            1
```

To be able to get the estimate of the average number of dog bites based on this data, we could use the equation :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

We could get the value of n by looking at :

```
> nrow(full_moon)
[1] 13
```

We could get the value of the summation of the entire number of *daily.dogbites* by using the code:

```
> sum(full_moon$daily.dogbites)
[1] 55
```

Based on the data we gained, we could use calculate the formula above.

$$\hat{\mu} = \frac{1}{13}55$$
$$\hat{\mu} = 4.2308$$

To find the 95% confidence interval, we must first search the α by using:

$$95 = 100(1-\alpha)$$
$$95 = 100 - 100\alpha$$
$$5 = 100\alpha$$
$$\alpha = 0.05$$

Then we could search the interval by using the equation :

$$(\hat{\mu} - t_{\frac{n}{2}}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + t_{\frac{n}{2}}\frac{\sigma}{\sqrt{n}})$$

To be able to use this formula, we must first get the $t_{\frac{n}{2}}$ value, and to get this value we could either use the t-score table, by firstly searching the value of :

$$v = n - 1$$
$$v = 13 - 1$$
$$v = 12$$

or we could use the following code in R by implementing the previous values.

```
> t_val = qt(0.975,12)
> t_val
[1] 2.178813
```

For the σ value, we could use the R code

```
> sdev = sqrt(var(full_moon$daily.dogbites))
> sdev
[1] 2.554533
```

And now we could calculate the interval.

$$(4.2308 - 2.178813\frac{2.554533}{\sqrt{13}}, 4.2308 + 2.178813\frac{2.554533}{\sqrt{13}})$$
$$(4.2308 - 2.178813(0.7085), 4.2308 + 2.178813(0.7085))$$
$$(4.2308 - 1.5437, 4.2308 + 1.5437)$$
$$(2.687, 5.774)$$

In conclusion, the estimate average number of dog-bites for days with a full moon (sample size *n* = 13) is 4.2308. We are 95% confident that the population mean number of dog-bites for this group is between 2.687 and 5.774

2. To calculate the estimated mean difference in mean dog bite occurrences between full-moon days and non-full moon days, we must first search the estimate number of dog-bites for days on which there was no full moon. Like the previous question, we must first sort out our data in the given *dogbites.fullmoon.csv* based on the value of *is.full.moon == 0*. We could do this by using the following code in R.

```
13  no_moon = dogbites_fullmoon[dogbites_fullmoon$is.full.moon==0,]
```

The result of the code above would give us the list of the number of dog bites only when there is a full moon, we could see this by using the following code, which will show us some the top of the list full_moon as the whole list would not fit in this report.

```
> head(no_moon)
  daily.dogbites is.full.moon
1              1            0
2              0            0
3              0            0
4              0            0
5              1            0
6              0            0
```

To be able to get the estimate of the average number of dog bites based on this data, we could use the equation :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

We could get the value of n by looking at :

```
> nrow(no_moon)
[1] 365
```

We could get the value of the summation of the entire number of *daily.dogbites* by using the code:

```
> sum(no_moon$daily.dogbites)
[1] 1648
```

Based on the data we gained, we could use calculate the formula above.

$$\hat{\mu} = \frac{1}{365}1648$$
$$\hat{\mu} = 4.5151$$

Then we could now calculate estimated mean difference in mean dog bite occurrences between full-moon days and non-full moon days by subtracting the û for the number of dog-bites in days with full- moon and days without full moon

$$\Delta\hat{\mu} = \hat{\mu}_{full-moon} - \hat{\mu}_{no\ full-moon}$$
$$\Delta\hat{\mu} = 4.2308 - 4.5151$$
$$\Delta\hat{\mu} = -0.2843$$

To find the 95% confidence interval, we must first search the α by using:

$$95 = 100(1-\alpha)$$
$$95 = 100 - 100\alpha$$
$$5 = 100\alpha$$
$$\alpha = 0.05$$

Then we could search the difference interval by using the equation:

$$(\hat{\mu}_A - \hat{\mu}_B - z_{\frac{n}{2}}\sqrt{\frac{{\sigma_A}^2}{n_A} + \frac{{\sigma_B}^2}{n_B}}\ , \hat{\mu}_A - \hat{\mu}_B + z_{\frac{n}{2}}\sqrt{\frac{{\sigma_A}^2}{n_A} + \frac{{\sigma_B}^2}{n_B}})$$

To be able to use this formula, we must first get the $z_{\frac{n}{2}}$ value, and to get this value we could either use the z-score table or we could use the following code in R by implementing the previous values.

```
> z_val = qnorm(0.975)
> z_val
[1] 1.959964
```

For the ${\sigma_A}^2$value, we could use the R code

```
> var_a = var(full_moon$daily.dogbites)
> var_a
[1] 6.525641
```

For the ${\sigma_B}^2$value, we could use the R code

```
> var_b = var(no_moon$daily.dogbites)
> var_b
[1] 12.72299
```

And now we could calculate the interval.

$$(4.2308 - 4.5151 - 1.96\sqrt{\frac{6.5256}{13} + \frac{12.723}{365}}, 4.2308 - 4.5151 + 1.96\sqrt{\frac{6.5256}{13} + \frac{12.723}{365}})$$
$$(4.2308 - 4.5151 - 1.96(0.7237), 4.2308 - 4.5151 + 1.96(0.7237)$$
$$(4.2308 - 4.5151 - 1.436, 4.2308 - 4.5151 + 1.436)$$
$$(-1.7204, 1.1518)$$

In conclusion, the estimate average number of dog-bites for days without a full moon (sample size n = 365) is 4.5151. Here, the difference of average number of dog bites for days with and without full moon is -0.2843, showing that there is more dog-bites happening in days without full-moon. We are 95% confident that the difference of the population mean number of bites for the two groups is between -1.7204 and 1.1518.

3. The hypothesis we are testing could be written as follows:

$$H_0 : \hat{\mu}_A - \hat{\mu}_B \leq 0 : full.moon \text{ does not affect } daily.dogbites$$
$$H_1 : \hat{\mu}_A - \hat{\mu}_B > 0 : full.moon \text{ affects } daily.dogbites$$

To be able to check on these hypotheses, we must find the p-value using the approximate hypothesis test for differences in means with unknown variances. To do this, we must find the z-score of the differences with the equation :

$$z(\hat{\mu}_A - \hat{\mu}_B) = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{{\sigma_A}^2}{n_A} + \frac{{\sigma_B}^2}{n_B}}}$$
$$z(\hat{\mu}_A - \hat{\mu}_B) = \frac{4.2308 - 4.5151}{\sqrt{\frac{6.5256}{13} + \frac{12.723}{365}}}$$
$$z(\hat{\mu}_A - \hat{\mu}_B) = \frac{-0.2843}{0.7327}$$
$$z(\hat{\mu}_A - \hat{\mu}_B) = -0.388$$

From the obtained z-score, we could use the pnorm function in R:

$$p\_value = pnorm(-z)$$

And it will return the following in R:

```
> p_value = pnorm(-(-0.388))
> p_value
[1] 0.650992
```

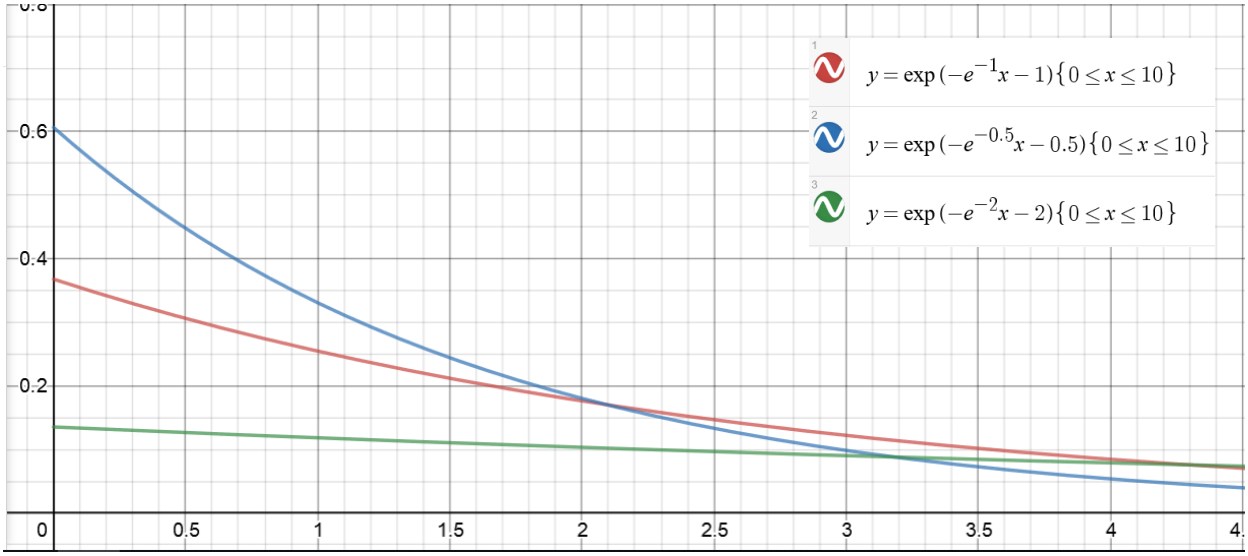
when the p value is < α, we therefore have to reject the null hypothesis, and based on the given confidence level of 95% (α = 0.05), and the acquired p value:

$$0.650992 > 0.05$$

Therefore, we can conclude that we must accept the null hypothesis (H_0), which declares that *full.moon* does not affect the number of *daily.dogbites.*

Question 2 :

1. We could plot the graph of the given equation by using https://www.desmos.com, which will give us the following graph, given with the legend.



2. The joint probability of this sample of data, under the assumption that it came from an exponential distribution with log-scale parameter *v* could be calculated with the following formula :

$$P\left(\frac{y}{\theta}\right) = \prod_{i=1}^{n} P\left(\frac{y_i}{v}\right)$$

$$where\ P\left(\frac{y_i}{v}\right) = \exp\left(-e^{-v}y_i - v\right)$$

$$P\left(\frac{y}{\theta}\right) = \prod_{i=1}^{n} \exp\left(-e^{-v}y_i - v\right)$$

$$P\left(\frac{y}{\theta}\right) = \exp(-e^{-v}y_1 - v).\exp(-e^{-v}y_2 - v) \ldots . \exp\left(-e^{-v}y_n - v\right)$$

$$P\left(\frac{y}{\theta}\right) = \exp\left((-e^{-v}y_1 - v) + (-e^{-v}y_2 - v) + \cdots + (-e^{-v}y_n - v)\right)$$

$$P\left(\frac{y}{\theta}\right) = \exp\left(-e^{-v}\sum_{i=1}^{n} y_i - v\right)$$

3. The negative logarithm of the likelihood expression and the negative log-likelihood of the data *y* under the exponential model with log-scale *v* could be calculated with the following:

$$\left(\frac{y}{v}\right) = -\ln\left(\exp\left(-e^{-v}\sum_{i=1}^{n} y_i - v\right)\right.$$

$$\left(\frac{y}{v}\right) = -(-e^{-v}\sum_{i=1}^{n} y_i - v)$$

$$\left(\frac{y}{v}\right) = e^{-v}\sum_{i=1}^{n} y_i - v$$

4. Derive the maximum likelihood estimator by using the following formula and steps:

$$\frac{d}{dv}(e^{-v}\sum_{i=1}^{n} y_i - v) = \sum_{i=1}^{n} y_i \cdot \frac{d}{dv}(e^{-v}) + n \cdot \frac{d}{dv}(v)$$

$$\frac{d}{dv}(e^{-v}\sum_{i=1}^{n} y_i - v) = n + e^{-v}.\frac{d}{dv}(-v).\sum_{i=1}^{n} y_i$$

$$\frac{d}{dv}(e^{-v}\sum_{i=1}^{n} y_i - v) = \left(-\frac{d}{dv}(v)\right)\left(\sum_{i=1}^{n} y_i\right)e^{-v} + n$$

$$\frac{d}{dv}(e^{-v}\sum_{i=1}^{n} y_i - v) = n - (\sum_{i=1}^{n} y_i)e^{-v}$$

5. To determine the approximate bias and variance of the maximum likelihood estimator $\hat{v}$ of v for the exponential distribution, we could use the following equation.

$$f(Y) = n - \sum_{i=1}^{n} y_i\, e^{-v} = n - Y(e^{-v})$$

$$\frac{df(Y)}{dY} = \frac{d}{dy}n - \frac{d}{dy}Y(e^{-v})$$

$$\frac{df(Y)}{dY} = 0 - (e^{-v})$$

$$\frac{df(Y)}{dY} = -(e^{-v})$$

$$\frac{df^2(Y)}{dY^2} = 0$$

$$bias(\hat{v}) = E(n - Ye^{-v}) - v$$

$$E\big(f(Y)\big) = f(vy) + \left[\frac{d^2f(Y)}{dY^2}\middle|\mathrm{y} = \mathrm{Vy}\right]\frac{(a^2v)}{2}$$

$$E\big(f(Y)\big) = n - Vye^{-v} + 0$$

$$bias = n - Vye^{-v} - v$$

$$Var(\hat{v}) = v(f(Y))$$

$$v\big(f(Y)\big) = \left[\frac{df(Y)}{dY}\middle|\mathrm{y} = \mathrm{Vy}\right]\alpha^2 y$$

$$v\big(f(Y)\big) = (-e^{-v})^2\alpha^2 y$$

$$Var(\hat{v}) = -e^{-2v}\alpha^2 y$$

Question 3:

1. As we need the estimate of the people turning their heads to the right, we could determine :

$$\hat{p} = \frac{80}{124} \; and \; \hat{q} = \frac{44}{124}$$

To provide the 95% confidence interval, we need the α value.

$$95 = 100(1-\alpha)$$
$$95 = 100 - 100\alpha$$
$$5 = 100\alpha$$
$$\alpha = 0.05$$

To find the confidence interval of the given data, we could use the equation:

$$(\hat{p} - z_{\frac{n}{2}}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)} \, , \hat{p} + z_{\frac{n}{2}}\sqrt{\left(\frac{\hat{p}\hat{q}}{n}\right)})$$

0To be able to use this formula, we must first get the $z_{\frac{n}{2}}$ value, and to get this value we could either use the z-score table or we could use the following code in R by implementing the previous values.

```
> z_val = qnorm(0.975)
> z_val
[1] 1.959964
```

With the data derived, we could now insert the values into the equation:

$$(\frac{80}{124} - 1.96\sqrt{\left(\frac{\frac{80}{124}\cdot\frac{44}{124}}{124}\right)} \, , \frac{80}{124} + 1.96\sqrt{\left(\frac{\frac{80}{124}\cdot\frac{44}{124}}{124}\right)})$$

$$(0.6452 - 1.96\sqrt{\left(\frac{0.2289}{124}\right)} \, , 0.6452 + 1.96\sqrt{\left(\frac{0.2289}{124}\right)})$$
$$(0.6452 - 1.96(0.4785) \, , 0.6452 + 1.96(0.4785) \,)$$
$$(0.6452 - 0.08422 \, , 0.6452 + 0.08422)$$
$$(0.5609 \; , 0.7294)$$

The estimated preference for humans turning their heads to the right when kissing (with the number of unique samples = 124) is 0.6452. We are 95% confident the population preference I between 0.5609 and 0.7294.

2. The hypothesis we are testing could be written as follows:

$$H_0 : \; P_0 = 0.5 : there \; is \; no \; certain \; preference$$
$$H_1 : \; P_1 \neq 0.5 : there \; is \; a \; certain \; preference$$

We want to check if we should accept or H_0 , so we could assume that *P = 0.5* and based on the previous calculation, we could obtain $\hat{p} = \frac{80}{124}$. To be able to check if we should accept H_0 or not, we must first obtain the z-score with the following equation :

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \,.\, (1 - p_0)}{n}}}$$

$$Z = \frac{\frac{80}{124} - 0.5}{\sqrt{\frac{0.5 \,.\, (1 - 0.5)}{124}}}$$

$$Z = \frac{\frac{18}{24}}{\sqrt{\frac{0.25}{124}}}$$

$$Z = \frac{0.1452}{\sqrt{0.002016}}$$

$$Z = 3.2329$$

To achieve the p-value, we could use the pnorm function in R or we could use the p-value table. We could derive it with the function :

$$p_{value} = 2 \,.\, pnorm(-z)$$

The following will be returned in R:

```
> p_value = 2 * pnorm(-3.2329)
> p_value
[1] 0.001225404
```

when the p value is < α, we therefore have to reject the null hypothesis, and based on the given confidence level of 95% (α = 0.05), and the acquired p value:

$$0.001225404 < 0.05$$

Therefore, we can conclude that we must reject the null hypothesis (H_0), which declares that there is a preference in humans for tilting their head to one side when kissing.

3. To calculate an exact p-value to test the above hypothesis using R, we could use the following code :

```
> binom.test(80,124,0.5)

        Exact binomial test

data:  80 and 124
number of successes = 80, number of trials = 124, p-value = 0.001565
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5542296 0.7289832
sample estimates:
probability of success
             0.6451613
```

Based on the data given by the code above, the p-value = 0.001565 which in this case, is still lower than 0.05, giving the same result as previous, where based on the condition if the p-value is lower than 0.05, then the null-hypothesis must still be rejected. Thus, based on this data, we can conclude that there is a preference in humans for tilting their head to one side when kissing.

4. The hypothesis we are testing could be written as follows:

$$H_0 : P_A \neq P_B : right - handedness\ and\ right\ head\ turning\ preference\ are\ not\ related$$
$$H_1 : P_A = P_B : right - handedness\ and\ right\ head\ turning\ preference\ are\ related$$

We want to check if we should accept or H_0 , so we could assume that *P = 0.5* and based on the previous calculation, we could obtain $\widehat{p_A} = \frac{80}{124}\ and\ n_A = 124$. To be able to check if we should accept H_0 or not, we must first obtain the z-score with the following equation :

$$Z = \frac{\widehat{p_A} - \widehat{P_B}}{\sqrt{\frac{\hat{P}_p \,.\, (1 - \hat{P}_p)}{n_A + n_B}}}$$

However, we need $\widehat{P_B}\ and\ \hat{P}_p$, and we could calculate both by using:
Because we are looking for the correlation of the right-head turning with righthandedness, then $\widehat{P_B}$ would be:

$$\widehat{P_B} = \frac{83}{83 + 17}$$
$$\widehat{P_B} = \frac{83}{100}$$

And we could achieve $\hat{P}_p$ by using the following equation:

$$\hat{P}_p = \frac{80 + 83}{124 + 100}$$
$$\hat{P}_p = \frac{163}{224}$$

Based on the achieved values, we could calculate the z-score.

$$Z = \frac{\widehat{p_A} - \widehat{P_B}}{\sqrt{\frac{\hat{P}_p \,.\, (1 - \hat{P}_p)}{n_A + n_B}}}$$
$$Z = \frac{\frac{80}{124} - \frac{83}{100}}{\sqrt{\frac{\frac{163}{224} \cdot (\frac{61}{224})}{224}}}$$
$$Z = \frac{-0.1848}{\sqrt{\frac{0.1982}{224}}}$$
$$Z = -6.2132$$

To achieve the p-value, we could use the pnorm function in R or we could use the p-value table. We could derive it with the function :

$$p_{value} = 2\,.\,pnorm(-|z|)$$

The following will be returned in R:

```
> p_value = 2 * pnorm(-6.2132)
> p_value
[1] 5.19163e-10
```

when the p value is < α, we therefore have to reject the null hypothesis, and based on the given confidence level of 95% (α = 0.05), and the acquired p value:

$$5.19163e - 10 < 0.05$$

Therefore, we can conclude that we must reject the null hypothesis (H_0), which declares that any preference for head turning to the right/left could be simply a product of right/left handedness.

5. The problem with our conclusions is that the data was collected in a biased environment, meaning that the sample does not cover / portrait the entire population sample, in the sense that the environment is more on right-handed people and they are not really checking if those who turn their heads to the right are right-handed or not. The correlation between these two are not presented in the data; however, the entire population is still generalized based on the biased data.

Question 4:

1. To be able to fit the multiple linear model to the fuel efficiency data using R, we could use the following R code :

```
> fuel = read.csv("fuel2017-20.csv", header = TRUE)
> fit = lm(Comb.FE~., fuel)
> summary(fit)

Call:
lm(formula = Comb.FE ~ ., data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2229 -0.9985 -0.0975  0.7149 11.4355

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.003e+02  7.241e+01  -2.766  0.00573 **
Model.Year               1.074e-01  3.588e-02   2.993  0.00279 **
Eng.Displacement        -1.287e+00  8.674e-02 -14.832  < 2e-16 ***
No.Cylinders             2.569e-03  5.767e-02   0.045  0.96447
AspirationOT            -2.471e-01  6.343e-01  -0.390  0.69692
AspirationSC            -1.015e+00  1.995e-01  -5.089 3.94e-07 ***
AspirationTC            -1.268e+00  1.085e-01 -11.685  < 2e-16 ***
AspirationTS            -1.183e+00  4.215e-01  -2.807  0.00506 **
No.Gears                -1.745e-01  2.534e-02  -6.888 7.58e-12 ***
Lockup.Torque.ConverterY -7.859e-01  9.506e-02  -8.267 2.48e-16 ***
Drive.SysA              -3.829e-02  1.294e-01  -0.296  0.76725
Drive.SysF               1.512e+00  1.438e-01  10.511  < 2e-16 ***
Drive.SysP              -4.435e-01  2.427e-01  -1.827  0.06781 .
Drive.SysR               9.319e-02  1.243e-01   0.750  0.45349
Max.Ethanol             -6.993e-03  2.490e-03  -2.808  0.00503 **
Fuel.TypeGM              5.696e-01  3.752e-01   1.518  0.12913
Fuel.TypeGP              5.024e-01  1.163e-01   4.321 1.63e-05 ***
Fuel.TypeGPR             2.066e-01  1.199e-01   1.723  0.08500 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.619 on 1982 degrees of freedom
Multiple R-squared:  0.6639,    Adjusted R-squared:  0.661
F-statistic: 230.3 on 17 and 1982 DF,  p-value: < 2.2e-16
```

Based on the code, we could achieve the p-values of the predictors with regards to *Comb.FE* (which in this case is our target, the fuel efficiency). To determine which predictors could be

correlated to *Comb.FE,* we could see those predictors that seem to be in borderline ($p < 0.1$*)*, and because for some of the predictors there exist sub-categories, our number of predictors would increase from 9 into 17. Thus, the possible predictors based on the data would be *Model.Year, Eng.Displacement, AspirationSC, AspirationTC, AspirationTS, No.Gears, Lockup.Torque.ConverterY, Drive.SysF, Drive.SysP, Max.Ethanol, Fuel.TypeGP,* and *Fuel.TypeGPR.*

However, in picking the top 3, we could see which among all those possible predictors are those whose p-value are close to 0 and having a significance level of 95% - 99.9%. These are *Eng.Displacement, AspirationTC,* and *Drive.SysF*.

2. Given the following hypothesis :

$$H_0 : \ the\ predictor\ is\ unimportant$$
$$H_1 : the\ predictor\ is\ important$$

According to the Bonferroni procedure, we should reject the H_0 if :

$$p_{value} < \frac{\alpha}{p}$$

And based on the values derived previously,

$$\frac{\alpha}{p} = \frac{0.05}{17}$$
$$\frac{\alpha}{p} = 0.00294$$

Thus, only for those predictors with the values less than 0.00294 that we can reject the null hypothesis and confirm that those predictors are in face important and relevant to the data. We notice that there is a major change in the threshold, (from 0.1 into 0.00294), then we can be positive that there would be a change in the possible predictors, as the number of predictors would decrease, adjusting to the new threshold.

3. Based on the multiple linear model to the fuel efficiency, we can see that *Model.Year* has a positive relationship with the *Comb.FE,* where we can see based on the data that every single increasing unit in *Model.Year*, we have a positive *Comb.FE* of 1.074e-01.

For the number of gears (*No.Gears),* it has a negative relationship with *Comb.FE,* as we could see from the data that for every single increasing unit of *No.Gears*, we have a negative value of *Comb.FE* of -1.745e-01.

4. We could derive the Stepwise BIC model by using the following R code :

```
> fit.sw.bic = step(fit, k = log(length(fuel$Comb.FE)))
Start:  AIC=2044.97
Comb.FE ~ Model.Year + Eng.Displacement + No.Cylinders + Aspiration +
    No.Gears + Lockup.Torque.Converter + Drive.Sys + Max.Ethanol +
    Fuel.Type

                          Df Sum of Sq    RSS    AIC
- No.Cylinders             1      0.01 5192.6 2037.4
- Fuel.Type                3     54.09 5246.6 2042.9
<none>                                 5192.6 2045.0
- Max.Ethanol              1     20.66 5213.2 2045.3
- Model.Year               1     23.47 5216.0 2046.4
- No.Gears                 1    124.29 5316.8 2084.7
- Lockup.Torque.Converter  1    179.06 5371.6 2105.2
- Aspiration               4    367.46 5560.0 2151.3
- Drive.Sys                4    613.78 5806.3 2238.0
- Eng.Displacement         1    576.36 5768.9 2247.9

Step:  AIC=2037.37
Comb.FE ~ Model.Year + Eng.Displacement + Aspiration + No.Gears +
    Lockup.Torque.Converter + Drive.Sys + Max.Ethanol + Fuel.Type

                          Df Sum of Sq    RSS    AIC
- Fuel.Type                3     54.36 5246.9 2035.4
<none>                                 5192.6 2037.4
- Max.Ethanol              1     20.66 5213.2 2037.7
- Model.Year               1     23.49 5216.0 2038.8
- No.Gears                 1    124.85 5317.4 2077.3
- Lockup.Torque.Converter  1    179.52 5372.1 2097.8
- Aspiration               4    405.29 5597.9 2157.3
- Drive.Sys                4    614.83 5807.4 2230.8
- Eng.Displacement         1   2981.13 8173.7 2937.2

Step:  AIC=2035.4
Comb.FE ~ Model.Year + Eng.Displacement + Aspiration + No.Gears +
    Lockup.Torque.Converter + Drive.Sys + Max.Ethanol

                          Df Sum of Sq    RSS    AIC
<none>                                 5246.9 2035.4
- Model.Year               1     25.59 5272.5 2037.5
- Max.Ethanol              1     29.25 5276.2 2038.9
- No.Gears                 1    109.64 5356.6 2069.2
- Lockup.Torque.Converter  1    193.74 5440.7 2100.3
- Aspiration               4    394.26 5641.2 2149.9
- Drive.Sys                4    600.34 5847.3 2221.7
- Eng.Displacement         1   3035.61 8282.5 2940.8
```

The summary of the above could be extracted by :

```
> summary(fit.sw.bic)

Call:
lm(formula = Comb.FE ~ Model.Year + Eng.Displacement + Aspiration +
    No.Gears + Lockup.Torque.Converter + Drive.Sys + Max.Ethanol,
    data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1799 -1.0033 -0.0835  0.6849 11.4237

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.097e+02  7.261e+01  -2.887 0.003927 **
Model.Year                1.120e-01  3.598e-02   3.113 0.001881 **
Eng.Displacement         -1.253e+00  3.698e-02 -33.897  < 2e-16 ***
AspirationOT             -1.014e-01  6.294e-01  -0.161 0.872034
AspirationSC             -7.208e-01  1.866e-01  -3.863 0.000116 ***
AspirationTC             -1.093e+00  9.018e-02 -12.116  < 2e-16 ***
AspirationTS             -1.100e+00  4.098e-01  -2.685 0.007309 **
No.Gears                 -1.606e-01  2.493e-02  -6.442 1.47e-10 ***
Lockup.Torque.ConverterY -7.999e-01  9.341e-02  -8.563  < 2e-16 ***
Drive.SysA                7.188e-02  1.242e-01   0.579 0.562843
Drive.SysF                1.545e+00  1.401e-01  11.027  < 2e-16 ***
Drive.SysP               -5.454e-01  2.376e-01  -2.295 0.021813 *
Drive.SysR                1.689e-01  1.231e-01   1.372 0.170300
Max.Ethanol              -8.184e-03  2.460e-03  -3.327 0.000893 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 1986 degrees of freedom
Multiple R-squared:  0.6603,    Adjusted R-squared:  0.6581
F-statistic:   297 on 13 and 1986 DF,  p-value: < 2.2e-16
```

Based on the summary above, we could obtain the regression equation :

$$y = \beta_0 + \beta_1(x) + \beta_2(x2) + \cdots$$

$$Comb.FE = -2.097e^{02} + 1.120e^{-01}(Model.Year) - 1.253(Eng.Displacement) - 1.014e^{-01}(AspirationOT) - 7.208e^{-01}(AspirationSC) - 1.083(AspirationTC) - 1.100(AspirationTS) - 1.606e^{-01}(No.Gears) - 7.999e^{-01}(Lockup.Torque.ConverterY) + 7.188e^{-02}(Drive.SysA) + 1.545(Drive.SysF) - 5.45e^{-01}(Drive.SysP) + 1.689e^{-01}(Drive.SysR) - 8.184e^{-03}(Max.Ethanol)$$

5. Based on the BIC data, if we want to have a higher *Comb.FE,* then we must increase those predictors that have positive relationship (as explained before) with the *Comb.FE* and decrease the values that has a negative relationship (as stated earlier) with *Comb.FE*. The key predictors that have positive relationship with *Comb.FE,* such as *Model.Year, Drive.SysA, Drive.SysF, Drive.SysR,* whereas all the other aspects like *Eng.Displacement, AspirationSC, AspirationTC, AspirationTS, No.Gears, Lockup.Torque.ConverterY , Drive.SysP, Max.Ethanol, Fuel.TypeGP,* and *Fuel.TypeGPR* would have to be lowered.

6. The answers for task 6:

   i. We could get the prediction of the mean fuel efficiency for this new car by BIC method using the following code:

```
> fuel.test = read.csv("fuel2017-20.test.csv")
> first_row = fuel.test[1,] #to take the first row of the list
> prediction = predict(fit.sw.bic, first_row, interval="confidence")
> prediction
       fit      lwr      upr
1 8.467534 8.260339 8.674729
```

   ii. We have been given that the current car efficiency rate is at 8.5 km/l, and because this value is still within the CI, then we have a weak evidence to say that the new car has a better fuel efficiency.