Question 1

1. Calculate the estimate of the average number of dog bites when there is a full moon and the 1. 95% confidence interval.

R code:

```
dogbites_fullmoon


dogbites= factor(dogbites_fullmoon$daily.dogbites)
fullmoon= factor(dogbites_fullmoon$is.full.moon)

# Question 1 part 1

fullmooncounter=0
fullmooncounter2=0
dogbitescounter=0
dogbitescounter2=0

for (i in 1:nrow(dogbites_fullmoon)) {
  if (dogbites_fullmoon[i,"is.full.moon"]==1) {
    fullmooncounter = fullmooncounter + 1
    dogbitescounter = dogbitescounter + dogbites_fullmoon[i, "daily.dogbites"]
  }

}

mean = dogbitescounter / fullmooncounter
mean
dogbitescounter


for (j in 1:nrow(dogbites_fullmoon)) {
  if (dogbites_fullmoon[j,"is.full.moon"]==1) {
    dogbitescounter2 = dogbitescounter2 + (dogbites_fullmoon[j, "daily.dogbites"]- mean)**2
  }

}

variance=dogbitescounter2/(fullmooncounter-1)
variance


tvalue=qt(1-0.05/2,12)
n=13

lowerbound= mean - tvalue * (sqrt(variance)/(sqrt(n)))
upperbound= mean + tvalue * (sqrt(variance)/(sqrt(n)))

lowerbound
upperbound
```

In order to get the mean for the average number of dog bites when there is full moon, the formula is stated below.

$$\hat{\mu} \equiv \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Mean for average number of dog bites when there is full moon = 4.230769

The formula below is used to find unknown variance:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Unknown variable = 6.525641

The formula below is used to find confidence interval by using the Student t-distribution. The $\alpha = 0.05$, $\mu = 4.230769$, $\sigma = 6.525641$, $n = 13$

$$\left(\hat{\mu} - t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}\right)$$

Working:

```
> tvalue=qt(1-0.05/2,12)
> tvalue
[1] 2.178813
> n=13
> lowerbound= mean - tvalue * (sqrt(variance)/(sqrt(n)))
> upperbound= mean + tvalue * (sqrt(variance)/(sqrt(n)))
> lowerbound
 daily.dogbites
        2.68708
> upperbound
 daily.dogbites
       5.774458
```

$$(4.230769 - 2.178813\frac{\sqrt{6.525641}}{\sqrt{13}}, 4.230769 + 2.178813\frac{\sqrt{6.525641}}{\sqrt{13}})$$

$$= (4.230769 - 2.178813(0.708500095), 4.230769 + 2.178813(0.708500095))$$

$$= (4.230769 - 1.543689217, 4.230769 + 1.543689217)$$

$$= (2.687079783, 5.774458217)$$

Answer: The mean for average number of dog bites when there is full moon is 4.230769 and the confidence interval falls between 2.68708 and 5.774458. It is 95% confident that the probability of having dog bites when there is full moon will fall between the range 0.560945238 and 0.729377341.

2. Calculate the estimated mean difference in mean dog bite occurrences between full moon days and non-full moon days, and a 95% confidence interval for this difference.

R code:

```
total=0
dogbitescounter3=0
for (i in 1:nrow(dogbites_fullmoon)) {
    total = total + dogbites_fullmoon[i, "daily.dogbites"]

}
nonfullmooncounter= 378-fullmooncounter
mean2= nonfullmoon_dogbites/nonfullmooncounter
mean2
daily.dogbites
      4.515068
for (j in 1:nrow(dogbites_fullmoon)) {
  if (dogbites_fullmoon[j,"is.full.moon"]==0) {
    dogbitescounter3 = dogbitescounter3 + (dogbites_fullmoon[j, "daily.dogbites"]- mean2)**2
  }

}
total
daily.dogbites
         1703
dogbitescounter3
daily.dogbites
      4631.167
nonfullmoon_dogbites= total - dogbitescounter
nonfullmoon_dogbites
daily.dogbites
         1648
variance2=dogbitescounter3/(nonfullmooncounter-1)
variance2
daily.dogbites
      12.72299
```

Mean for average number of dog bites when there is full moon = 4.515068
Unknown variance = 12.72299

```
mean = dogbitescounter / fullmooncounter
mean
daily.dogbites
      4.230769
```

Mean for average number of dog bites when there is full moon = 4.230769

By using R, we can get the mean and unknown variance and we can use the formula below to find the exact mean difference:

$$\hat{\mu}_A - \hat{\mu}_B$$

$\mu_a = 4.230769$

$\mu_b = 4.515068$

The mean μ_a is the same mean as in question 1 part 1 which is the mean for average number of dog bites when there is full moon and μ_b is the mean for average number of dog bites when there is no full moon.

Mean difference = $\mu_a - \mu_b$

= 4.230769 – 4.515068

= - 0.284299

In order to find the difference of Normal means, the formula below is used:

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

$\mu_a - \mu_b = -0.284299,\ \sigma_a = 6.525641, \sigma_b = 12.72299, n_a = 13, n_b = 365$

$$(-0.284299 - 1.959964\sqrt{\frac{6.525641}{13} + \frac{12.72299}{365}}, -0.284299 + 1.959964\sqrt{\frac{6.525641}{13} + \frac{12.72299}{365}})$$

$$= (-0.284299 - 1.959964(0.732686762), -0.284299 + 1.959964(0.732686762))$$

$$= (-0.284299 - 1.436039677, -0.284299 + 1.436039677)$$

$$= (-1.720338677, 1.151740677)$$

Answer: The mean difference is - 0.284299 and the confidence interval for difference of normal means falls between $-1.720338677\ and\ 1.151740677$. It is 95% confident that the probability of the estimated mean difference in mean dog bite occurrence between full moon day and non-full moon day is fall between the range $-1.720338677\ and\ 1.151740677$.

3. Test the hypothesis

H_0 = dogs bite more frequently on full moon days

H_A = dogs bite more frequently on non-full moon days

$H_0 : \mu_x \geq \mu_y$

$H_A : \mu_x < \mu_y$

$\mu_x = mean\ of\ dog\ bites\ for\ days\ on\ which\ there\ was\ a\ full\ moon$

$\mu_y = mean\ of\ dog\ bites\ for\ days\ on\ which\ there\ was\ not\ a\ full\ moon$

$\mu_x - \mu_y = -0.2842993, \sigma_x = 6.525641, \sigma_y = 12.72299, n_x = 13, n_y = 365$

The formula to find the z value is:

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

$$Z\ (\mu_x - \mu_y) = \frac{-0.2842993}{\sqrt{\frac{6.525641}{13} + \frac{12.72299}{365}}}$$

$$= \frac{-0.2842993}{0.732686762}$$

$$= -0.388022978$$

Since we know that $H_0 : \mu_x \geq \mu_y$ and $H_A : \mu_x < \mu_y$ and therefore we used the equation $P(Z < z(\mu_x - \mu_y))$ as it can be seen below:

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|) & \text{if } H_0 : \mu_x = \mu_y \text{ vs } H_A : \mu_x \neq \mu_y \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \leq \mu_y \text{ vs } H_A : \mu_x > \mu_y \\ \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \geq \mu_y \text{ vs } H_A : \mu_x < \mu_y \end{cases}.$$

R code:

By using the pnorm() function, the p value is 0.3489995.

```
> pnorm(-0.388022978)
[1] 0.3489995
```

Answer: p-value = 0.3489995. The null hypothesis states that there is preference that dogs will bite more frequently on a full moon day than on a non-full moon day but from the p-value that we've gotten from above, the p-value is not < 0.01 which have strong evidence supporting the null hypothesis. Therefore, we can conclude by saying that there is a preference that dogs will bite more frequently on a full moon day.

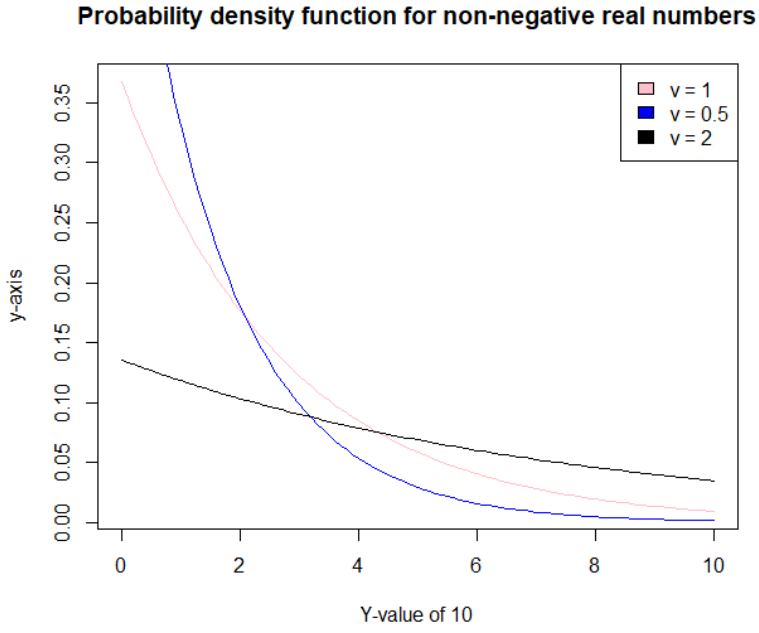
Question 2

1. Plot the exponential probability density function (1) for the values y ∈ (0, 10), for v = 1, v = 0.5 and v = 2

R code:

```
> equation=function(x) {exp((-exp(1)**(-1))*x-1)}
> plot(equation,xlim=c(0,10),xlab="Y-value of 10",ylab="y-axis",main="Probability density function for non-negative real numbers",col="pink")
> equation=function(x) {exp((-exp(1)**(-0.5))*x-0.5)}
> plot(equation,xlim=c(0,10),add=TRUE,col="blue")
> equation=function(x) {exp((-exp(1)**(-2))*x-2)}
> plot(equation,xlim=c(0,10),add=TRUE,col="black")
> legend("topright",c("v = 1","v = 0.5","v = 2"),fill=c("pink","blue","black"))
```

The diagram below shows the probability density function for non-negative real numbers:



2. Joint probability of the sample data from an exponential distribution.

y = (y1, . . ., yn)

The joint probability p(y | v) is:

$\exp(-e^{-v}y_1 - v) * \exp(-e^{-v}y_2 - v) * \exp(-e^{-v}y_3 - v) \ldots \exp(-e^{-v}y_n - v)$

$= \exp((-e^{-v}y_1 - v) + (-e^{-v}y_2 - v) + (-e^{-v}y_3 - v) \ldots (-e^{-v}y_n - v))$

$= \exp((-e^{-v}y_1) + (-e^{-v}y_2) + (-e^{-v}y_3) \ldots (-e^{-v}y_n) - nv)$

$= \exp(-e^{-v}\sum_{i=1}^{n} y_i - nv)$

Answer: $\exp(-e^{-v}\sum_{i=1}^{n} y_i - nv)$

3. Negative loglikelihood of the data

By using the formula below to get the negative likelihood:

$$L(\mathbf{y} \mid \boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta})$$

L(y | v)

$= -\sum_{i=1}^{n} \ln p(y_i \mid v)$

$= -\ln \exp(-e^{-v}\sum_{i=1}^{n} y_i - nv)$

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

$= e^{-v}(\sum_{i=1}^{n} y_i) + nv$

Answer: $e^{-v}(\sum_{i=1}^{n} y_i) + nv$

4. Derive the maximum likelihood estimator v for v and find the value of v that minimises the negative log-likelihood.

By using the formula below to find the maximum likelihood estimator:

$$\begin{aligned}\frac{dL(\mathbf{y} \mid \lambda)}{d\lambda} &= -\sum_{i=1}^{n} y_i \frac{d}{d\lambda} \log \lambda + n \\ &= -\frac{\sum_{i=1}^{n} y_i}{\lambda} + n\end{aligned}$$

We equal the formula $\frac{dL(y \mid v)}{dv}$ to 0 to get the maximum/minimum:

$$\frac{dL(y \mid v)}{dv} = 0$$

We will then differentiate the equation $e^{-v}(\sum_{i=1}^{n} y_i) + nv$ and equate it to 0:

$-e^{-v}(\sum_{i=1}^{n} y_i) + n = 0$

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

$-e^{-v} = -\frac{n}{\sum_{i=1}^{n} y_i}$

$-\ln e^{-v} = -\ln \frac{n}{\sum_{i=1}^{n} y_i}$

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

$-(-v) * (1) = -\ln \frac{n}{\sum_{i=1}^{n} y_i}$

$v = -\ln \frac{n}{\sum_{i=1}^{n} y_i}$

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

$v = \ln \frac{\sum_{i=1}^{n} y_i}{n}$

$v_{ML} = \ln \frac{\sum_{i=1}^{n} y_i}{n}$

Answer: Therefore, $v_{ML} = \ln \frac{\sum_{i=1}^{n} y_i}{n}$

5. Determine the approximate bias and variance of the maximum likelihood estimator v.

Let X = log $\frac{\sum_{i=1}^{n} y_i}{n}$, X ~exp(v), E[X] = e^v and V[X] = e^{2v}.

F(X) = ln(X)

$$\frac{df(X)}{dX} = \frac{1}{X}$$

$$\frac{d^2f(X)}{dX^2} = -\frac{1}{X^2}$$

Since we know the formula to find E[f(X)] is:

$$\mathbb{E}\left[f(X)\right] \approx f(\mu_X) + \left[\left.\frac{d^2f(x)}{dx^2}\right|_{x=\mu_X}\right]\frac{\sigma_X^2}{2}$$

We can use the formula above to find E[f(X)].

E[f(X)]

= f(μ_x) + [- $\frac{1}{X^2}$ | $_{X=\mu_x}$] $\frac{\sigma_x^2}{2}$

= f(μ_x) + [- $\frac{1}{(\mu_x)^2}$] $\frac{\sigma_x^2}{2}$

= ln (e^v) + [- $\frac{1}{e^{2v}}$] $\frac{e^{2v}}{2}$

= v - $\frac{1}{2}$

Since we know the formula to find V[f(X)] is:

$$\mathbb{V}\left[f(X)\right] \approx \left[\left.\frac{df(x)}{dx}\right|_{x=\mu_X}\right]^2 \sigma_X^2$$

We can use the formula above to find E[f(X)].

V[(f(X))

= [$\frac{1}{X}$ | $_{X=\mu_x}$] ^{2} σ_x^2

= [$\frac{1}{\mu_x}$] ^{2} σ_x^2

= [$\frac{1}{e^v}$] ^{2} e^{2v}

= ($\frac{1}{e^{2v}}$) e^{2v}

= 1

Bias of v(y)

= E[v(y)] – v

= E[ln($\frac{\sum_{i=1}^{n} y_i}{n}$)] – v

= E[ln(X)] – v

= (v - $\frac{1}{2}$) – v

= - $\frac{1}{2}$

Variance of v(y)

= V[v(y)]

= V[ln($\frac{\sum_{i=1}^{n} y_i}{n}$)]] – v

= V[f(X)]

= 1

Answer: The appropriate bias is - $\frac{1}{2}$ and the variance of the maximum likelihood is 1.

Question 3

1. Calculate the estimate of the preference for humans turning their heads to the right when kissing and the 95% confidence interval

Mean for turning the head to the right = 80 / 124 = 0.64516129

$$\left(\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right).$$

Working:

$$(0.64516129 + 1.96(\sqrt{\frac{(0.64516129)(1-0.64516129)}{124}}), 0.64516129 + 1.96(\sqrt{\frac{(0.64516129)(1-0.64516129)}{124}}))$$

$$= (0.64516129 + 1.96(\sqrt{\frac{(0.64516129)(0.35483871)}{124}}), 0.64516129 + 1.96(\sqrt{\frac{(0.64516129)(0.35483871)}{124}}))$$

$$= (0.64516129 + 1.96(\sqrt{0.228928199}), 0.64516129 + 1.96(\sqrt{0.228928199}))$$

$$= (0.64516129 + 0.084216051, 0.64516129 + 0.084216051)$$

$$= (0.560945238, 0.729377341)$$

Answer: The estimated mean for humans turning their heads to the right is 0.64516129 and the confidence interval falls between 0.560945238 and 0.729377341. It is 95% confident that the probability of the human turning their heads to the right is fall between the range 0.560945238 and 0.729377341.

2. Test the hypothesis

θ_x = Humans do not turn their heads when kissing

θ_y = Humans turn their heads to one side when kissing

$$H_0 : \theta = \theta_0$$
$$\text{vs}$$
$$H_A : \theta \neq \theta_0$$

$H_0 : \theta_x = \frac{1}{2}$

vs

$H_A : \theta_x \neq \frac{1}{2}$

$\theta = 0.64516129$

θ_0 = 0.5

By using the formula below to get the z-value:

$$z_{\hat{\theta}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$$

z-score = $\frac{(0.64516129-0.5)}{\sqrt{\frac{(0.5)(1-0.5)}{124}}}$

= $\frac{0.14516129}{0.044901325}$

= 3.232895429

Since we know that $H_0 : \mu_x \neq \mu_y$ and $H_A : \mu_x = \mu_y$ and therefore we used the equation $2P(Z < -|\, z\left(\mu_x - \mu_y\right)|\,)$ as it can be seen below:

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|) & \text{if } H_0 : \mu_x = \mu_y \text{ vs } H_A : \mu_x \neq \mu_y \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \leq \mu_y \text{ vs } H_A : \mu_x > \mu_y \\ \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \geq \mu_y \text{ vs } H_A : \mu_x < \mu_y \end{cases}.$$

R code:

```
> 2*pnorm(-abs(3.232895429))
[1] 0.001225424
```

p-value: 0.001225424

Answer: p-value = 0.001225424. The null hypothesis states that there is no preference that humans turn their heads when kissing but from the p-value that we've gotten from above, the p-value is < 0.01 which have strong evidence against the null hypothesis proving the null hypothesis is wrong. Therefore, we can conclude by sayi ng that there is a preference that humans do turn their heads to either left or right side when kissing.

3. Calculate the exact p-value.

```
> binom.test(x=80, n=124, p=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
```

Exact p-value by using R = 0.001565

Answer: The p-value is still < 0.01 but it is still larger than the approximate test that we've gotten from the previous question. This value proves that the null hypothesis is wrong as the p-value is < 0.01 but the evidence is not as strong as the one we got from the approximate test because the exact p-value is 0.001565 whereas the p-value that we got from the appropriate test is just 0.001225424

4. Testing two Bernoulli populations

Total number of people = 83 + 17 = 100

Right-handed = 83

Left-handed = 17

Θ_x = Right-handed / Total number of people = $\frac{83}{100}$

Θ_y = Left-handed / Total number of people = $\frac{17}{100}$

$$H_0 \quad : \quad \theta_x = \theta_y$$

vs

$$H_A \quad : \quad \theta_x \neq \theta_y$$

M_x, M_y = the number of successes in the two samples

M_x, M_y = the total number of trials

By using the formula below to find the Θ_p:

$$\Theta_p = \frac{M_x + M_y}{N_x + N_y}$$

$$\Theta_p = \frac{80 + 83}{124 + 100}$$

$$= \frac{163}{224}$$

$$= 0.727678571$$

By using the formula to find z-score for the difference of the probabilities:

$$z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{\hat{\theta}_x - \hat{\theta}_y}{\sqrt{\hat{\theta}_p(1 - \hat{\theta}_p)(1/n_x + 1/n_y)}}$$

$$Z(\Theta_x - \Theta_y) = \frac{\frac{80}{124} - \frac{83}{100}}{\sqrt{\frac{163}{224}\left(1 - \frac{163}{224}\right)\left(\frac{1}{124} + \frac{1}{100}\right)}}$$

$$= \frac{-0.184838709}{0.05983067}$$

= -3.089363837

Since we know that $H_0 : \Theta_x \neq \Theta_y$ and $H_A : \Theta_x = \Theta_y$ and therefore we used the equation $2P(Z < -|\, z\left(\mu_x - \mu_y\right)|\,)$ as it can be seen below:

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|) & \text{if } H_0 : \mu_x = \mu_y \text{ vs } H_A : \mu_x \neq \mu_y \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \leq \mu_y \text{ vs } H_A : \mu_x > \mu_y \\ \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu_x \geq \mu_y \text{ vs } H_A : \mu_x < \mu_y \end{cases} .$$

$2P(Z < -3.089363837) = \quad 0.002005856$

p-value = 0.002005856

Answer: p-value = 0.002005856. The p-value is < 0.01 which have strong evidence against null hypothesis. Therefore, we can conclude by saying that the rate of right-handedness in the population is the same as the preference for turning their heads to the right when kissing.

5. Identify the problems with the conclusions based on the way in which the data was collected.

Answer: The sample population given are too small which is less accurate compare to using larger sample population. The next problem is that the sample population could not represent the general population as the sample population is too small and the sample population do not represent all types of cases and people. For example, in this case if some people have experience in kissing might tend to turn their heads when kissing compared to some people who doesn't and therefore having these problems to the kissing test.

Question 4

1. Fit a multiple linear model to the fuel efficiency data and find the three variables that appear to be the strongest predictors.

R code:

```
fuel=fuel2017_20

linear=lm(Comb.FE~.,fuel)
linear$coefficient
summary(linear)
```

By using the R code above, it can get the coefficients table as it's shown below. The predictors Eng. Displacement, AspirationTC and Drive.SysF are associated with fuel efficiency because if p-value is < 0.1 is suggesting that the data is at it's odds with the null hypothesis of no association and we should potentially consider the 3 predictors Eng. Displacement, AspirationTC and Drive.SysF as being associated with the target.

```
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 .
---
```

Answer: The predictors Eng. Displacement, AspirationTC and Drive.SysF are associated with fuel efficiency because if p-value is < 0.1 is suggesting that the data is at its odds with the null hypothesis of no association and we should potentially consider the 3 predictors Eng. Displacement, AspirationTC and Drive.SysF as being associated with the target.

2. Using the Bonferroni procedure to determine which predictors are associated to change.

Bonferroni procedure with $\alpha = 0.05$ where α is the significance level and p=17 where p is the number of estimators.

We should only reject the null hypothesis only if:

$$\text{P-value} < \frac{\alpha}{p}$$

p-value = 2e-16, α = 0.05, p = 17

2e-16 < $\frac{0.05}{17}$

Answer: No, the assessment of the predictors would not change as it is proven above that the p-value is smaller than the value $\frac{\alpha}{p}$. Therefore, the predictors will still affect the fuel efficiency.

3. Describe the effect the year of manufacture and the number of gears had on the mean fuel efficiency.

The effect that the year of manufacture (Model.Year) and the number of gears (No.Gears) variable has on mean fuel efficiency of the car is shown below by using the R code:

```
> result=lm(Comb.FE~Model.Year,fuel)
> result$coefficient
  (Intercept)    Model.Year
-59.40519408    0.03464502
> result=lm(Comb.FE~No.Gears,fuel)
> result$coefficient
(Intercept)     No.Gears
 15.3154421   -0.6905314
```

Answer: The effect for the year of manufacture (Model.Year) appears to have on the mean fuel efficiency is 0.03464502 when the year of manufacture changes by 1 unit. The effect for the number of gears (No.Gears) variable has on the mean fuel efficiency of the car is -0.6905314 for every 1 number of gear changes.

4. Write down the final regression equation

R code:

```
# Question 4 part 4
result=lm(Comb.FE~No.Gears,fuel)

fit.sw.bic = step(linear, k = log(length(fuel$Comb.FE)))
summary(fit.sw.bic)

fit.sw.bic$coefficients
```

By using the formula:

```
Comb.FE ~ Model.Year + Eng.Displacement + No.Cylinders + Aspiration +
    No.Gears + Lockup.Torque.Converter + Drive.Sys + Max.Ethanol +
    Fuel.Type
```

And applying the values from the coefficients table into the formula above:

(Intercept)	Model.Year	Eng.Displacement	AspirationOT	AspirationSC	AspirationTC
-2.096579e+02	1.119950e-01	-1.253482e+00	-1.014020e-01	-7.208417e-01	-1.092623e+00
AspirationTS	No.Gears	Lockup.Torque.ConverterY	Drive.SysA	Drive.SysF	Drive.SysP
-1.100367e+00	-1.606346e-01	-7.999164e-01	7.188136e-02	1.544768e+00	-5.453658e-01
Drive.SysR	Max.Ethanol				
1.688838e-01	-8.183967e-03				

Answer: Comb.FE = -209.6579 + 0.111995 * Model Year + (-1.253482) * Eng.Displacement + (-0.1014020) * AspirationOT + (-0.7208417) * AspirationSC + (-1.092623) * AspirationTC + (-1.100367) * AspirationTS + (-0.1606346) * No.Gears + (-0.7999164) * Lockup.Torque.ConverterY + 0.07188136 * Drive.SysA + 1.544768 * Drive.SysF + (-0.5453658) * Drive.SysP + 0.1688838 * Drive.SysR + (-0.008183967) * MaxEthanol

5. BIC model to improve the fuel efficiency

Answer: We can improve the fuel efficiency of our car by increasing the ModelYear which means that getting a new car. Increasing the Drive.SYsA and DriveSysF and DriveSysR will increase the fuel efficiency too. Furthermore, decreasing the Eng.Displacement, AspirationOT, AspirationSC, AspirationTC, AspirationTS, No.Gears, Lockup.Torque.ConverterY, Drive.SysP and MaxEthanol will also improve the fuel efficiency because it is a negative value which means that the smaller the value, the bigger the fuel efficiency.

6a. BIC model to predict the mean fuel efficiency for this new car and provide 95% confidence interval

R code:

```
> fuel=fuel2017_20
> fuel_test= fuel2017_20_test
> ModelYear=fuel_test$Model.Year[1]
> Eng.Displacement=fuel_test$Eng.Displacement[1]
> fuel_test$Aspiration[1]
[1] "TC"
> No.Gears=fuel_test$No.Gears[1]
> Lockup.Torque.Converter=fuel_test$Lockup.Torque.Converter[1]
> fuel_test$Drive.Sys[1]
[1] "R"
> MaxEthanol=fuel_test$Max.Ethanol[1]
> fit.sw.bic = step(linear, k = log(length(fuel$Comb.FE)))
```

```
> mean_fuel_efficiency = -209.6579 + 0.111995 * ModelYear + (-1.253482) * Eng.Displacement + (-0.1014020) * 0 + (-0.7208417) * 0 + (-1.092623) * 1 + (-1.1
00367) * 0 + (-0.1606346) * No.Gears + (-0.7999164) * 1 + 0.07188136 * 0 + 1.544768 * 0 + (-0.5453658) * 0 + 0.1688838 * 1 + (-0.008183967) * MaxEthanol
> mean_fuel_efficiency
[1] 8.4675
> mean= mean(fuel$Comb.FE)
> mean
[1] 10.51936
> variance= ((mean_fuel_efficiency - mean)**(2)) /(nrow(fuel)-1)
> variance
[1] 0.00210611
```

Substitute all the values into the formula:
Comb.FE = -209.6579 + 0.111995 * Model Year + (-1.253482) * Eng.Displacement + (-0.1014020) * AspirationOT + (-0.7208417) * AspirationSC + (-1.092623) * AspirationTC + (-1.100367) * AspirationTS + (-0.1606346) * No.Gears + (-0.7999164) * Lockup.Torque.ConverterY + 0.07188136 * Drive.SysA + 1.544768 * Drive.SysF + (-0.5453658) * Drive.SysP + 0.1688838 * Drive.SysR + (-0.008183967) * MaxEthanol

Therefore, Comb.FE = 8.4675.

By using the formula below to get the confidence interval:

$$\left(\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$\mu = 8.4675,\ \sigma = 0.00210611,\ n = 2000$

$$(8.4675 + 1.959964(\sqrt{\frac{0.00210611}{2000}}), 8.4675 + 1.959964(\sqrt{\frac{0.00210611}{2000}}))$$

$$= (8.4675 - 0.002011285033,\ 8.4675 + 0.002011285033)$$

$$= (8.465488715, 8.469511285)$$

Answer: The mean fuel efficiency is 8.4675 and the confidence interval falls between $8.465488715\ and\ 8.469511285$. It is 95% confident that the probability of mean fuel efficiency for the new car will fall between the range $8.465488715\ and\ 8.469511285$.

6b) Does the model suggest that the new car will have a better fuel efficiency than the current car.

Answer: The current car has a mean fuel efficiency of 8.5km/l and the new car has a mean fuel efficiency of 8.4675km/l. This has proven that the new car does not have better fuel efficiency than the current car as the mean fuel efficiency for the new car is lower than the current car. This will suggest the owner that changing to a new car does not necessary mean that it gives a better fuel efficiency.