FIT5197 Assignment 3 Semester 1, 2020

Do & Don't

1. Some of these policies are similar to your assignment 1; however, please read carefully as there are some differences between this assignment and the first assignment.
2. These questions are meant for you to solve independently, we encourage students to figure out the questions themselves as it would be good for their understandings of the topics; however, please feel free to consult your tutors if needed. Plagiarism (either from using online sources or copying the answers from your classmates) will be penalised accordingly.
3. Requests for special consideration or extension must be submitted at least 2 days BEFORE THE DEADLINE. The due date is on Sunday, so the latest day you can ask for extension is on Friday (the last official working day of the week for the teaching team). Please follow Monash guideline to request for extensions (medical certificates, doctor or GP letter, etc). Emergencies are to be adjusted individually.
4. Please show all working when answering questions, you will not get full marks for a question if you don't comply.
5. Assignments need to be submitted in PDF and ipynb file format. Failure to comply will result in 20% penalty on each missing file.
6. Filename format for submitting the assignment should be “Assignment3_StudentId.pdf” and "Assignment3_StudentId.ipynb". Files with the wrong format incurs 20% penalty each.
7. This assignment has 10 marks for presentation, this includes presenting your explanation in **Markdown**, writing and commenting on code efficiently, creating good plots with clear labels on the axis, etc.
8. Only answers with correct methodology will be considered for consequential marks. Meaning if you attempt the question and your answer is incorrect, but your methodology is correct, you will still receive partial marks for subsequent questions. However, answers with incorrect methodology (misunderstanding the questions) will generate no marks for subsequent questions.
9. Challenge questions are for students aiming to get a HD for the assignment. We don't advise for students to spend time on these questions before finishing all the other parts in the assignment. This assignment is designed in a way that students can get up to 80 (HD) without attempting the challenge questions.
10. Please don’t send emails to tutors asking for suggestions, we have Moodle and consultations for that, In writing your inquiries on Moodle please try to be clear in your problem and not revealing your working to others as this might be counted as plagiarism on your part. A good format for inquiry topic would be e.g. “Assignment 3 – Tutorial 10 (your tutorial slot) – Question about median"
11. Handwritten answers incur a penalty of 10% on your assignment, you have Markdown, please learn how to use it as it will be an useful skill for you going through the degree as well as in real life situation.
12. This assignment will contribute towards 20% of your total score.
13. Late submission is 5% per day, after 10 days you will be given no marks. Late submission is calculated as follows: If you get 70% on this assignment and you are late for 2 days (you submit on Tuesday), your score is now 70% -10% (2x5% per day) = 60%. This is done to ensure that the teaching team can release your result as soon as possible so that you can review on your mistakes and have a better study experience.
14. Assignments shall be marked completely in two weeks’ time according to Monash Policies. If there are any changes to the marking time, we will duly inform you. Solutions will not be released for this assignment; you can come to the tutorial and ask for explanation about how to solve the questions after scores are released.

Part 1 - Hypothesis Testing & Confidence Interval (45 Marks)

No R codes, libraries of any kinds should be used in this part, using libraries here results in 0 marks

Question 1. (5 Marks)

An online article believe that teenagers currently spends at least 10h online per week (should be significantly higher than this). the principal of one highschool believes that her students spend significantly less time online; thus, she would like to perform an experimental testing on the subject by collecting 100 surveys from her students. The results came back that the average time the students claim spending online is roughly 8.5 hours/week with the standard deviation of 2h.

Based on these findings, What would be your conclusion on the matter? Use a 0.05 level of significance.

Given :
Number of surveys : = 100
Mean : $\mu = 8.5$
Testing Value :

$$\mu_i = 10$$

Standard Deviation : $\sigma = 2$

Solution :
Part 1: Calculating z value

$$\begin{aligned} z_u &= \frac{(\mu - \mu_0)}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{(8.5 - 10)}{\frac{2}{\sqrt{100}}} \\ &= \frac{(-1.5)}{0.2} \\ &= -7.5 \end{aligned}$$

Part 2: Calculating p value

$$\begin{aligned} p &= 1 - p(-|z_u| < z < |z_u|) \\ &= 1 - p(-7.5 < z < 7.5) \\ &= 0 \end{aligned}$$

The z score can be approximated to 0.

Conclusion on the matter :

- Here, the p-value is really small.
- This indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
  - Thus, the hypothesis that teenagers spend around 10 hours online per week can be rejected.
  - Hence, the online article's hypothesis is wrong and the principal's hypothesis is correct!

Question 2. (10 Marks)

The light bulb in Monash university normally has a lifetime of about 1500 hours with a standard deviation of 100 hours. How many of these light bulbs should Monash stock up so that it can guarrantee that the light will be on for at least 7200 hours with a probability of at least 98%?

Given :

Let the number of light bulbs be **n**
Mean : $\mu = 1500$
Testing Value :

$$\mu_i = (7200/n)$$

Standard Deviation : $\sigma = 100$

Solution :
Part 1: Calculating z value

$$z_u = \frac{(\mu - \mu_0)}{\frac{\sigma}{\sqrt{n}}}$$
$$= \frac{(1500 - \frac{7200}{n})}{100}$$

Part 2: P value is given as 0.98

$$1 - p(-|z_u| < z < |z_u|) = 0.98$$
$$2(z < -|z_u|) = 0.98$$
$$(z < -|z_u|) = 0.49$$
$$z_u = -2.05$$

Part 2: Solving for n from Part 1 and Part 2

$$\frac{(1500 - \frac{7200}{n})}{100} = -2.05$$
$$(15 - \frac{72}{n}) = -2.05$$
$$n = 5.56$$

Answer: Hence, the number of light bulbs that Monash can stock up is 6.

Question 3. (10 Marks)

The error for the production of a machine is uniformly distribute over [-0.75, 0.75] unit. Assuming that there are 100 machines working at the same time, approximate the probability that the final production differ from the exact production by more than 4.5 unit?

YOUR ANSWER HERE

Given: Error is Uniformly Distributed over [-0.75,0.75]
a = -0.75
b = 0.75
Number of Machines (n) = 100

To Find: P(X > 4.5)

Solution:
Since, it is a Uniform Distribution,
Mean and Variance can be calculated as follows:
Step 1 : Calculating μ :

$$\begin{aligned}\mu &= \frac{a+b}{2} \\ &= \frac{-0.75+0.75}{2} \\ &= 0\end{aligned}$$

Step 2 : Calculating σ :

$$\begin{aligned}\sigma^2 &= \frac{(b-a)^2}{12} \\ &= \frac{(1.5)^2}{12} \\ &= \frac{3}{16} \\ &= 0.433\end{aligned}$$

Step 3:
For Large Samples, i.e for n = 100

$$\begin{aligned}\sigma^{'} &= \sigma(\sqrt{n}) \\ &= 0.433 \cdot \sqrt{100} \\ &= 4.33\end{aligned}$$

Step 4:
P(X > 4.5)

$$\begin{aligned}P(X > 4.5) &= P(\frac{x-\mu}{\sigma} > \frac{4.5-\mu}{\sigma}) \\ &= P(z > \frac{4.5-0}{4.33}) \\ &= P(z > 1.04) \\ &= 1 - P(z <= 1.04) \\ &= 1 - 0.8508 \\ &= 0.1492\end{aligned}$$

Answer:
Probability that the final production differ from the exact production by more than 4.5 unit is **0.1492**

Question 4. (5 Marks)

A very successful car washing shop has roughly **2.4 million dollar** revenue every year. Since this is quite a large amount of money for a car washing shop, the **IRS (tax people)** wants to check whether this establishment is laundering money (sounds familiar?); however, due to the limited resources, they want to be **95% confident** in their decision; thus, they send an investigator to record the daily number of customers and record down how much that each customers have to pay on average for the service. The investigator came back and reported that there are roughly **2,000 customers for the month** and each one of them paying roughly **80 dollars on average** for the services with a **standard deviation of 30 dollars**. Given this information, what is the approximate probability that the car washing shop can achieve it current claimed revenue and what would be your conclusion here about the legitimacy of this car washing shop **(i.e whether they are laundering money or not)**?

YOUR ANSWER HERE

Given:
Mean (μ) : 80
Standard Deviation (σ) : 30
Confidence Interval (α) : 0.05
Z(α/2) = 1.96
Number of Customers (n) : 2000

Solution:

Step 1:
For Large Samples, i.e for n = 100

$$\begin{aligned}\sigma^{'} &= \sigma(\sqrt{n}) \\ &= 30 \cdot \sqrt{2000} \\ &= 0.6708\end{aligned}$$

Step 2:
Now, 95% Confidence Interval is given by:

$$\begin{aligned}CI &= (\mu - Z(\alpha/2) \cdot \frac{\sigma}{\sqrt{n}}, \mu + Z(\alpha/2) \cdot \frac{\sigma}{\sqrt{n}}) \\ &= (80 - 1.96 \cdot \frac{30}{\sqrt{2000}}, 80 + 1.96 \cdot \frac{30}{\sqrt{2000}}) \\ &= (80 - 1.3148, 80 + 1.3148) \\ &= (78.6852, 81.3148)\end{aligned}$$

Step 3:
Now, finding the maximum values of the company's revenue :
Maximum Income in a month :

$$\begin{aligned}&= (2000) \cdot (81.3148) \\ &= 162,629.8\end{aligned}$$

Maximum Income in a year :

$$\begin{aligned}&= (12) \cdot (162,629.8) \\ &= 1,951,555.2\end{aligned}$$

Conclusion:

- Thus, we can see that the minimum value charged by the shop is 78 Dollars.
- Now, with an average of 80 Dollars and Standard Deviation of 30 Dollars,
  - The Range that they should charge should be - Min : 50 Dollars, Max : 110 Dollars.
- However, the minimum charged is around 78 Dollars.
- Also :
  - Reported Income : **1,951,555.2 Dollars**
  - Actual Income : **2,400,000 Dollars**
- Hence, we can conclude and say that the company is **laundering money**

Question 5. (10 Marks)

Note that this is a challenge question, only attempt if you are comfortable with your progress. You should not use consultations to ask about this question as the tutors will prioritize answering queries about other questions.

Let $X \sim Binom(n, p = 0.6)$. What would be the lowest value for n such that $Pr(\frac{X}{n} > \frac{1}{4}) \geq 0.99$

YOUR ANSWER HERE

Solution:
Mean $(\mu) = p = 0.6$
Variance $(\sigma^2) = npq = 0.24n$

Hence, Standard Deviation $(\sigma) = \sqrt{0.24n}$

Proceeded by Normalisation Method:

$$
\begin{aligned}
Pr(\frac{X}{n} > \frac{1}{4}) &>= 0.99 \\
1 - Pr(\frac{X}{n} < \frac{1}{4}) &>= 0.99 \\
1 - Pr(\frac{z}{n} < \frac{0.24 - \mu}{\sigma}) &>= 0.99 \\
1 - Pr(\frac{z}{n} < \frac{0.24 - 0.6}{\sqrt{0.24n)}} &>= 0.99 \\
1 - Pr(\frac{z}{n} < \frac{-0.35}{\sqrt{0.24n)}} &>= 0.99 \\
Pr(\frac{z}{n} < \frac{-0.35}{\sqrt{0.24n)}} &>= -0.01 \\
\frac{-0.35}{\sqrt{0.24n)}} &>= \frac{-2.33}{n} \\
\frac{0.35n}{2.33} &>= \sqrt{0.24n} \\
n &>= 10.6 \\
n &\approx 11
\end{aligned}
$$

Hence, Lowest Value of n = 11

Question 6. (5 Marks)

Note that this is a challenge question, only attempt if you are comfortable with your progress. You should not use consultations to ask about this question as the tutors will prioritize answering queries about other questions.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with mean λ. Thus, $Y = \sum_{i=1}^{n} X_i$ has a Poisson distribution with mean $n\lambda$. Moreover, by the Central limit Theorem, $\bar{X} = Y/n$ has, approximately, a Normal $(\lambda, \lambda/n)$ distribution for large n. Show that for large values of n, the distribution of

$$2\sqrt{n}\left(\sqrt{\frac{Y}{n}} - \sqrt{\lambda}\right)$$

is independent of λ.

Given:
$X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution.
Mean: λ
$Y = \sum_{i=1}^{n} X_i$
Mean : $n\lambda$

To Prove:
Show that for large values of n, the distribution of

$$2\sqrt{n}\left(\sqrt{\frac{Y}{n}} - \sqrt{\lambda}\right)$$

is independent of λ.

Proof: Step 1:
Here, $\bar{X}$ follows Normal $(\lambda, \lambda/n)$ distribution
Hence,

- Mean = λ
- Variance = $\frac{\lambda}{n}$
- Implies, Standard Deviation = $\sqrt{\sigma}$

Step 2:
Using the Central Limit Theorem, Using Z Transformation on $\bar{X}$, we get :

$$= \frac{\bar{X} - \mu}{\sigma}$$
$$= \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$
$$= \frac{\frac{Y}{n} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

Step 3:

- Using Weak Law of Large Numbers,
- Now, For large values of n, Y/n tends to the mean value i.e λ
- Hence, $\frac{Y}{n}$ follows a Normal Distribution with :
  - Mean = 0
  - Variance = 1

Step 4:

- Now, I have assumed the following -
- Our Required Distribution, follows the same pattern as $\bar{X}$
- $2\sqrt{n}\left(\sqrt{\frac{Y}{n}} - \sqrt{\lambda}\right) \approx \frac{\frac{Y}{n} - \lambda}{\sqrt{\frac{\lambda}{n}}}$
- Hence, For large values of n, $2\sqrt{n}\left(\sqrt{\frac{Y}{n}} - \sqrt{\lambda}\right)$ follows a Normal Distribution i.e Normal(0,1)
- Hence, it would be **independent** of λ

Part 2 - Modelling and Data Analysis (20 Marks)

No libraries should be used in this part

This question will require you to analyse a regression dataset. In particular, you will be looking at predicting the fuel efficiency of a car (in kilometers per litre) based on characteristics of the car and its engine. This is clearly an important and useful problem. The dataset **fuel2017-20.csv** contains $n = 2,000$ observations on $p = 9$ predictors obtained from actual fuel efficiency tables for car models available for sale during the years 2017 through to 2020. The target is the fuel efficiency of the car measured in kilometers per litre. The higher this score, the better the fuel efficiency of the car. Provide working/R code/justifications for each of these questions as required.

Question 1 (4 Marks)

Fit a multiple linear model to the fuel efficiency data using R. Using the results of fitting the linear model, which predictors do you think are possibly associated with fuel efficiency, and why? Which three variables appear to be the strongest predictors of fuel efficiency, and why? $[2 \text{ marks }]$

In [129]:

```
# R Code

# Reading in the data into the dataframe fuel
fuel <- read.csv("fuel2017-20.csv")
#summary(fuel)

# Our Target : Fuel Efficiency

# Fitting in a multiple linear model
# "Comb.FE ~ ." expression indicates all variables other than Comb.FE that will be used to
fit <- lm(Comb.FE ~ ., fuel)

# Details of the Linear Model
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
```

Predictors that are possibly associated with fuel efficiency, and why? :

Eng.Displacement

AspirationSC

AspirationTC

No.Gears

Lockup.Torque.ConverterY
Drive.SysF
Fuel.TypeGP

Reason:

- Looking at the p values, these predictors have their p the values less than 0.001.
- Hence, these are most possibly associated with fuel efficiency.

Three variables appear to be the strongest predictors of fuel efficiency, and why? :
AspirationTC
Drive.SysF
Eng.Displacement

Reason:

- Looking at the p values of these predictors, these are the ones which have the least p values.
- Hence, strongest predictors of fuel efficiency.

Question 2 (5 Marks)

Describe what effect the year of manufacture (Model. Year) appears to have on the mean fuel efficiency. Describe the effect that the number of gears (No. Gears) variable has on the mean fuel efficiency of the car.

In [130]:

```
# Effect of (Model.Year) on the mean fuel efficiency
fit_1 <- lm(Comb.FE ~ Model.Year, fuel)

# Details of the Linear Model
summary(fit_1)
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5695 -1.9177 -0.5266  1.2473 15.7159 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -59.40519  122.42170  -0.485    0.628
Model.Year    0.03465    0.06066   0.571    0.568

Residual standard error: 2.78 on 1998 degrees of freedom
Multiple R-squared:  0.0001633, Adjusted R-squared:  -0.0003372 
F-statistic: 0.3262 on 1 and 1998 DF,  p-value: 0.5679
```

Effect of (Model.Year) on the mean fuel efficiency: :

- The p value of (Model.Year) is 0.5679
- Here, Model.Year does have an impact on the Fuel Efficiency when compared to
  - No.Cylinders
  - AspirationOT
  - AspirationSC

    - Drive.SysA
    - Drive.SysP
    - Drive.SysR
    - Fuel.TypeGM
    - Fuel.TypeGPR
- However, when you compare Model.Year with the following, it's impact is significantly lesser :
    - Eng.Displacement
    - AspirationSC
    - AspirationTC
    - No.Gears
    - Lockup.Torque.ConverterY
    - Drive.SysF
    - Fuel.TypeGP
- Also, by looking at the **Estimate Value**, we can see that (Model.Year) does has a very less effect on Fuel Efficiency.
- Thus, Model.Year is **not that great a predictor** for the fuel efficiency.
- There are many other predictors which have a better,significant and a greater impact on the Fuel Efficiency Values.

In [131]:

```
# Effect of (No.Gears) on the mean fuel efficiency
fit_2 <- lm(Comb.FE ~ No.Gears, fuel)

# Details of the Linear Model
summary(fit_2)
```

```
Call:
lm(formula = Comb.FE ~ No.Gears, data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7465 -1.6772 -0.0824  1.4878 15.0521 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 15.31544    0.23027   66.51   <2e-16 ***
No.Gears    -0.69053    0.03216  -21.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.506 on 1998 degrees of freedom
Multiple R-squared:  0.1875,    Adjusted R-squared:  0.1871 
F-statistic: 461.1 on 1 and 1998 DF,  p-value: < 2.2e-16
```

Effect of (No.Gears) on the mean fuel efficiency: :

- The p value of (No.Gears) is < 2.2e-16
- It comes in the list of predictors which are most likely to be associated with Fuel Efficiency along with the following :
    - Eng.Displacement
    - AspirationSC
    - AspirationTC
    - No.Gears

    - Lockup.Torque.ConverterY
    - Drive.SysF
    - Fuel.TypeGP
- No.Gears does have a **good** impact while predicting the **Fuel Efficiency** when compared to most of the variables which are made available to us.
- Also, when we look at the **estimate**, we cab observe that :
    - Fuel Efficiency is **INVSERSELY** related to (No.Gears)

Question 3 (3 Marks)

Use the stepwise selection procedure with the BIC penalty to prune out potentially unimportant variables. Write down the final regression equation obtained after pruning.

In [132]:

```
# stepwise selection procedure
# with the BIC penalty to prune out potentially unimportant variables

fullmod <- glm(Comb.FE ~ . , data=fuel, family=gaussian)
back.fit = step(fullmod, trace = 0, k = log(nrow(fuel)), direction = "both")

# Details
summary(back.fit)
```

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

Deviance 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

(Dispersion parameter for gaussian family taken to be 2.641955)

    Null deviance: 15447.9  on 1999  degrees of freedom
Residual deviance:  5246.9  on 1986  degrees of freedom
AIC: 7634.7

Number of Fisher Scoring iterations: 2
```

In [133]:

```
# final regression equation obtained after pruning

# Obtaining the coefficients for the Regression Equation
back.fit$coefficients
```

(Intercept): -209.657883335339 **Model.Year:** 0.111995007976626 **Eng.Displacement:** -1.25348170242984 **AspirationOT:** -0.101401989270943 **AspirationSC:** -0.720841697007822 **AspirationTC:** -1.0926229099929 **AspirationTS:** -1.10036735799902 **No.Gears:** -0.160634635566795 **Lockup.Torque.ConverterY:** -0.799916434115236 **Drive.SysA:** 0.0718813593093842 **Drive.SysF:** 1.54476755878251 **Drive.SysP:** -0.545365788393033 **Drive.SysR:** 0.168883804163226 **Max.Ethanol:** -0.00818396661069438

Final Regression Equation:

Comb.FE =
-209.66 + 0.11(**Model.Year**) -1.25(**Eng.Displacement**) - 0.101(**AspirationOT**)

-0.72(**AspirationSC**) -1.09(**AspirationTC**) -1.1(**AspirationTS**) - 0.16*(*No.Gears**)**

-0.79(Lockup.Torque.ConverterY**) + 0.07(**Drive.SysA**) + 1.54(**Drive.SysF**) - 0.54(**Drive.SysP**)**

+0.16(Drive.SysR**) - 0.008(**Max.Ethanol**)**

Question 4 (4 Marks)

If we wanted to improve the fuel efficiency of our car, what does this BIC model suggest we could do?

In order to improve the fuel efficiency of our car, this BIC model suggests the following:

- As per the BIC Model,
  - Following can be **"INCREASED" :**
    - Drive.SysA, Drive.SysF, Drive.SysR, Model.Year, Drive.SysF
    - From the above information on p values,
      - We can see that **Drive.SysF** has the most impact on Fuel Efficiency
      - As it has the least p value among these four
  - Following can be **"DECREASED"** :
    - Max.Ethanol, Eng.Displacement, AspirationOT, AspirationTS, AspirationSC, Lockup.Torque.ConverterY, Drive.SysP, AspirationTC, No.Gears
    - From the above information on p values,
      - We can see that **No.Gears** has the most impact on Fuel Efficiency
      - As it has the least p value among these values
        - Variables which **DO NOT** have **ANY IMPACT** :
        - Fuel.GDR
        - No.Cylinders
        - Fuel.TypeGM
        - Fuel.TypeGP

In Conclusion:
In order to INCREASE the value of Fuel Efficiency in the most EFFECTIVE manner :
INCREASE the value of **Drive.SysF**
DECREASE the value of **No.Gears** \

BIC Model:

- This model has dropped 4 variables.
- Hence, the resulting model is a bit more **less complex**
- Reason :
  - To **PREDICT** the fuel efficiency of the car
  - We now need **ONLY** 13 variables
- Thus, this model is better at predicting the **Fuel Efficieny**.

Question 5 (4 Marks)

Imagine that you are looking for a new car to buy to replace your existing car. Load the dataset **fuel2017-20.test.csv**. The characteristics of the new car that you are looking at are given by the first row of this dataset.

```
(a) Use your BIC model to predict the mean fuel efficiency for this new car. Provi
de a 95% confidence interval for this prediction. [ 2 mark]
(b) The current car that you own has a mean fuel efficiency of 8.5 km/l (measured
 over the life time of your ownership). Does your model suggest that the new car w
ill have better fuel efficiency than your current car? [2 mark]
```

In [134]:

```
# Loading the data
fuel_test <- read.csv("fuel2017-20.test.csv")

# Characteristics of the new car that you are looking at are given by the first row of this
row1 = head(fuel_test,1)
#row1

# BIC model to predict the mean fuel efficiency for this new car
fit <- lm(Comb.FE ~ ., fuel)
pred_1 = predict(back.fit, row1, interval = "confidence")
pred_1
```

1: 8.46753367714036

a) Predicted Mean with the BIC Model : 8.467 km/l

b) According to my model,

- Fuel Efficiency of the current car : 8.5 km/l
- Predicted Fuel Efficiency of the new car : 8.46 km/l
- Thus, The new car **will/will not** have better **Fuel Efficiency** than my current car.

Part 3 - Simulation (25 Marks)

No libraries should be used in this part

This part of the assignment will deal with **Rejection Sampling & Inverse Sampling**.

Question 1 - Inverse Sampling (10 Marks)

Given the following distribution:

$$\text{PDF(x)} = e^{-x^2\pi} \text{ for x } [-\infty, +\infty]$$

Question 1.a (4 Marks)

Perform the Inverse Sampling Process as you have learnt from the tutorial and lecture

Step 1 :
Calculating the CDF from PDF :

$$\begin{aligned} CDF(x) &= \int_{-\infty}^{x} f(t) \cdot dt \\ &= \int_{-\infty}^{x} e^{-x^2\pi} \cdot dt \end{aligned}$$

where $-\infty < x < \infty$

Solving the above equation, we get :

$$CDF(x) = \frac{1 + erf(\sqrt{\pi} \cdot x)}{2}$$

where $x >= 0$

$$\begin{aligned} CDF(x) &= \frac{1 + erfc(\sqrt{\pi} \cdot x)}{2} \\ &= \frac{1 + 1 - erf(\sqrt{\pi} \cdot x)}{2} \\ &= \frac{2 - erf(\sqrt{\pi} \cdot x)}{2} \end{aligned}$$

where $x < 0$

Step 2:
Generating the Quantile Function from CDF:
We know that Quantile Function is nothing but inverse PDF.
Part a : where $x >= 0$

$$\begin{aligned} CDF(x) &= y \\ y &= \frac{1 + erf(\sqrt{\pi} \cdot x)}{2} \\ 2y - 1 &= erf(\sqrt{\pi} \cdot x) \\ erf^{-1}(2y - 1) &= \sqrt{\pi} \cdot x \\ F^{-1}(y) &= \frac{erf^{-1}(2y - 1)}{\sqrt{\pi}} \end{aligned}$$

Part b : where $x < 0$

$$
\begin{aligned}
CDF(x) &= y \\
y &= \frac{2 - erf(\sqrt{\pi} \cdot x)}{2} \\
2 - 2y &= erf(\sqrt{\pi} \cdot x) \\
erf^{-1}(2 - 2y) &= \sqrt{\pi} \cdot x \\
F^{-1}(y) &= \frac{erf^{-1}(2 - 2y)}{\sqrt{\pi}}
\end{aligned}
$$

Question 1.b (3 Marks)

Write the code for this inverse sampling process based on the result you obtain from the previous part. ie. You should write a function that take in a vector of number and return the corresponding samples for these inputs. The better your function is (errors handling, comments, variable names, etc) the higher the score you will get for this particular part.

The vector of number that you should use to test your function will be provided below, please kindly print the result obtained from your function in your answer (if you don't print the final result to the terminal, you automatically lose half the mark)

In [40]:

```
# Code for Inverse Sampling Process

# Defining the Inverse of our Error Function
# As these functions are not available in r,
# I have created these functions which are closer to the error function

erf_inverse <- function(x1) qnorm((x1 + 1)/2)/sqrt(2)
erfc_inverse <- function(x2) qnorm(x2/2,lower = FALSE )/sqrt(2)
```

In [32]:

```
# Please use this vector to test your function
v <- c(0.265508661956518, 0.572853363366483, 0.201681933540029, 0.944675271666033, 0.629114
print(v)
```

```
 [1] 0.26550866 0.57285336 0.20168193 0.94467527 0.62911404 0.20597457
 [7] 0.68702285 0.76984142 0.71761851 0.38003518 0.93470523 0.65167376
[13] 0.26722067 0.01339034 0.86969085 0.48208011 0.49354130 0.82737332
[19] 0.79423986 0.72371094 0.82094630 0.78293276 0.52971958 0.02333120
[25] 0.73231374 0.47761962 0.43809711 0.07067905 0.31627170 0.66200508
```

In [26]:

```
# Creating our Required Inverse Function
inverse_function <- function(y) {
    erf_inverse( (2*y) - 1 )/ sqrt(pi)
    }

# Function to take in our vector and return the samples
samples = inverse_function(v)

# Printing the samples
print(samples)
```

```
 [1] -0.24991891  0.07326309 -0.33336758  0.63642496  0.13145458 -0.32731953
 [7]  0.19445606  0.29454894  0.22970353 -0.12183234  0.60311343  0.15552495
[13] -0.24783913 -0.88353742  0.44878251 -0.01792592 -0.00645898  0.37653617
[19]  0.32761986  0.23693235  0.36661893  0.31202724  0.02974712 -0.79363650
[25]  0.24727470 -0.02239213 -0.06215341 -0.58674531 -0.19075427  0.16673456
```

Question 1.c (3 Marks)

Please provide 100 sample for this distribution using all the answers you have completed so far. Please set your random number seed to 1 (**You should not ask your tutor what setting random number means, if you don't know what this means, go back and read materials from your tutorials**)

YOUR ANSWER HERE

In [128]:

```
# Setting the seed pf the random number generator to 1
set.seed(1)

# Generating 100 Samples
X = runif(100)

# Presenting the 100 Samples to our Inverse Function
new_sample = inverse_function(X)

print(new_sample)
```

```
  [1] -0.249918910 -0.130148283  0.073263087  0.530513149 -0.333367588
  [6]  0.507625846  0.636424950  0.165417997  0.131454586 -0.614351166
 [11] -0.327319525 -0.370444656  0.194456059 -0.117576450  0.294548940
 [16] -0.002300771  0.229703523  0.959317914 -0.121832343  0.304629710
 [21]  0.603113433 -0.318758576  0.155524951 -0.457848898 -0.247839137
 [26] -0.115478458 -0.883537448 -0.119369562  0.448782506 -0.164169073
 [31] -0.017925915  0.100622601 -0.006458975 -0.355825053  0.376536163
 [36]  0.173812489  0.327619861 -0.493706408  0.236932352 -0.089469944
 [41]  0.366618926  0.150559074  0.312027239  0.053193510  0.029747123
 [46]  0.320825197 -0.793636476 -0.022782312  0.247274693  0.200910514
 [51] -0.022392130  0.433159304 -0.062153414 -0.275650698 -0.586745309
 [56] -0.512481000 -0.190754270  0.018641044  0.166734558 -0.094033315
 [61]  0.542034722 -0.216581077 -0.041006367 -0.172865804  0.154658597
 [66] -0.259101703 -0.021465104  0.289931601 -0.549367288  0.459546300
 [71] -0.165558881  0.395814722 -0.157298934 -0.171350936 -0.023662622
 [76]  0.494011857  0.438846638 -0.111443045  0.304463065  0.701301854
 [81] -0.065635420  0.223705324 -0.101076686 -0.180634672  0.278048165
 [86] -0.331937247  0.222076492 -0.465394316 -0.274773763 -0.425109136
 [91] -0.282249726 -0.623858788  0.145447159  0.461391508  0.306600273
 [96]  0.331938775 -0.044819651 -0.090691027  0.351511127  0.106173447
```

Question 2 - Rejection Sampling (15 Marks)

Using the same distribution as the previous question:

$$\text{PDF(x)} = e^{-x^2\pi} \text{ for x } [-\infty, +\infty]$$

please complete these following questions:

YOUR ANSWER HERE

Question 2.a (3 Marks)

In Rejection Sampling, you would sample from the proposal within a certain range. Your task here includes finding an appropriate range to sample from for the proposal distribution. Giving the range without proper explanation receives 0 mark.

Proposal Distribution:

- Here, the proposal distribution is **Uniform Distribution.**
- It completely overlaps the **Target Distribution.**
- Reason :
  - Here, maximum value of the Target Distribution is 1, at x = 0.

    - Hence, all the values of the Target Distribution fall under the Proposal Distribution.
- Also, the constant with which the Proposal Distribution needs to be multiplied is 1 as all samples of the Target Distribution lie under the Proposal Distribution.

Range:

- Clearly, the **appropriate range** to sample from would be from [-1,1]

Question 2.b (3 Marks)

Please write down rejection sampling process for this distribution here, let's say that the range is $[-2, 2]$ for the proposal distribution. (if you think using this answer for the previous part is a good idea, it is not)

YOUR ANSWER HERE

Writing Down the Rejection Sampling Process:

Step 1:

- Here, we first plot our Target Distribution.
- Hence, we plot this -

$$\text{PDF(x)} = e^{-x^2\pi} \text{ for x } [-\infty, +\infty]$$

Step 2:

- We then, plot our **Proposal Distribition** Function.
- Here, the Proposal Distribution Function taken as the **Uniform Distribution** Funtion.

Step 3:

- We now check if our Proposal Distribution completely overlaps all values of Target Distribution.
- Clearly, our Proposal Distribution, clearly encompasses all values of the Target Distribution.
- Hence, it is **sufficient** and we do not need to multiply it by a constant.

Step 4:

- We now generate random variables that lie in the Uniform Distriution : U(0,1)

Step 5:

- We then compare the following :
- We **ACCEPT** the sample x_i if
    - $$\frac{PDF(x_i)}{X(x_i)} >= U_i$$
- Else, we **REJECT** the sample.

Step 6:

- Finally, to check if the Rejection Sampling is correct,
- we check if the **ACCEPTED** Samples follow the Target Distribution or not.

Question 2.c (4 Marks)

What would be the acceptance and rejection percentage here?

YOUR ANSWER HERE

Now, we need to satisfy the following:-

- Proposed Distribution is the Uniform Distribution = g(X) = 1
- We now need to find a C that satisfies -

$$\text{for x } [-2, +2]$$

Hence, we can write the following:

$$C \cdot f(x) <= g(x)$$
$$C \cdot f(x) <= 1$$
$$C <= \frac{1}{\hat{f}}$$

where $\hat{f}$ is any upper bound on f(x) $\text{ for x } [-2, +2]$

- Here, **maximum value** of f(x) = 1 at x=0
- Hence, an appropriate value of the Constant would be 3

Proportion of Samples Accepted : C = 1/3
Proportion of Samples Rejected : 1 - C = 2/3

Hence,
Acceptance percentage : 33.33 %
Rejection percentage : 66.66 %

Question 2.d (5 Marks)

Note that this is a challenge question, only attempt if you are comfortable with your progress. You should not use consultations to ask about this question as the tutors will prioritize answering queries about other questions.

Prove that your acceptance/rejection percentage is correct (i.e simulate it) using your own function; furthermore, also provide a graphical display of your final results. The better the display (including interactive, better graphic, etc), the higher the mark you will get for this question.

YOUR ANSWER HERE

In [116]:

```
# Checking the above percentages through Simulation

# Considering 10000 Samples

# Proposal Distribution X
# Here, the Proposal Distribution chosen is the Uniform Distribution

# Range : [-2,2]
X = runif(10000, -2, 2)

# Uniform Distribution U
U = runif(10000, 0, 1)

# Creating a function for calculating PDF
pdf_of_x <- function(x) exp(-x^(2)*pi)

# Array for storing all the Accepted variables
accept = c()
# Array for storing all the Rejected variables
reject = c()

for(count in 1:length(X)){

    # Generating random variables from the Uniform Distribution
    test_u = U[count]

    # Calculating the PDF of the Random Variables generated
    # From the Proposal Distribution
    test_x = pdf_of_x( X[count] )

    # Final Ratio for testing
    ratio = test_x/test_u

    # If Ui <= PDF(Xi)/X(xi), we accept the sample
    if (U[count] <= ratio){
        # Store it in the accept vector
        accept = rbind(accept, X[count])
        count = count + 1
    }
    else {
        # Store it in the reject vector
        reject = rbind(reject, X[count])
        count = count + 1
    }

    }
```

In [117]:

```
acceptance_percentage = (length(accept)/10000)*100
rejection_percentage = (length(reject)/10000)*100

cat(" \n Values from Simulation:")
cat( " \n Acceptance Percentage :",
        round( acceptance_percentage, digits = 2), "%" )
cat( "\n Rejection Percentage :",
        round( rejection_percentage, digits = 2), "%" )
```

```
 Values from Simulation:
 Acceptance Percentage : 35.5 %
 Rejection Percentage : 64.5 %
```

- In every simulation, the count value keeps on changing.
- This is because the number of samples **REJECTED** in every simulation are different are slightly different.
- However, roughly -
  - Acceptance Percentage is around - 33%
  - Rejection Percentage is around - 66%

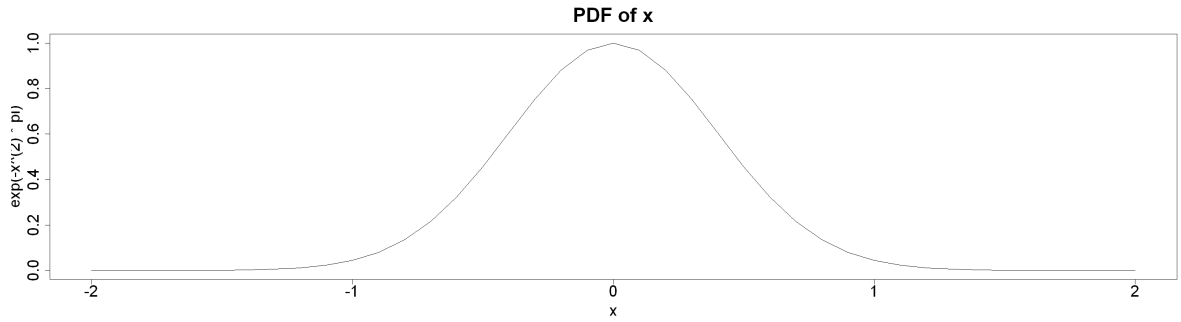
Thus, our CALCULATED VALUES match our SIMULATED VALUES

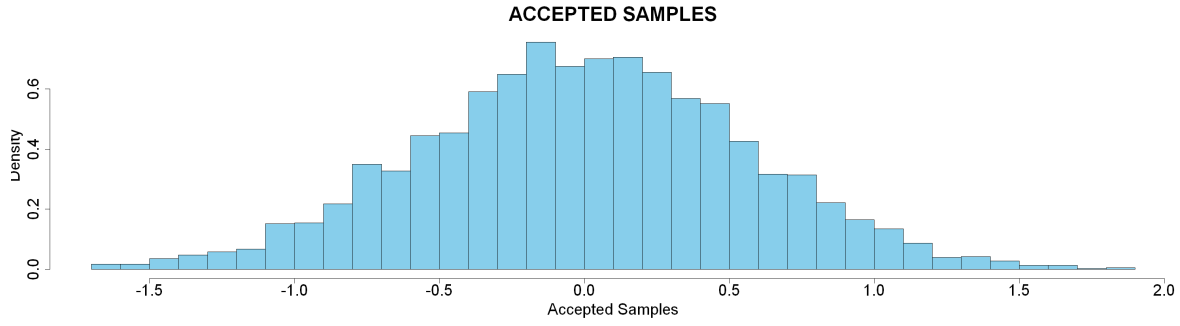
In [108]:

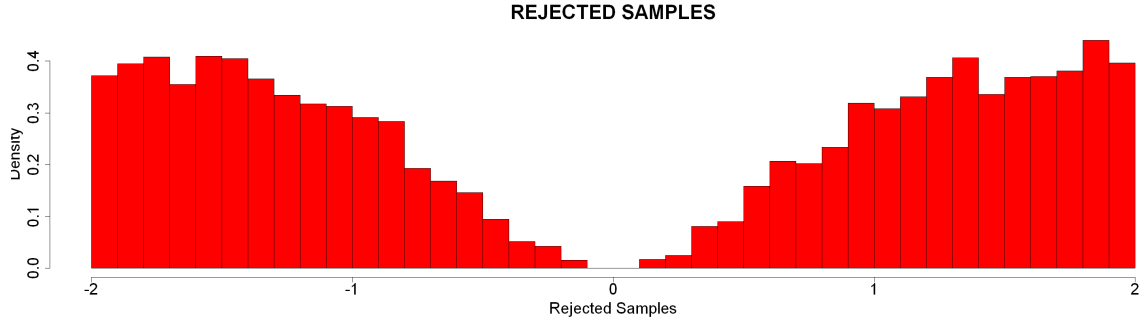
```
# Now, these are the Results for PDF in the interval : [-2,2]
# Making the Process Interactive

cat("\n Options for Display :")
cat("\n 1) Accepted and Rejected Values ")
cat("\n 2) PDF along with the Accepted Values")
cat("\n 3) PDF along with the Rejected Values")

user_choice = readline(prompt = "\n Enter Your Choice:")

# Plotting our Target Distribution
x <- seq(-2, 2, by = 0.1)
plot(x,
    exp(-x^(2)*pi),
    cex.main = 2.5,
    cex.lab = 2,
    cex.axis = 2,
    cex.sub = 1,
    col = "black",
    main = sprintf("PDF of x"),
    typ = "l")

if (user_choice == 1) {
    # Plotting the Accepted Values
    h = hist(
            accept,
            cex.main = 2.5,
            cex.lab = 2,
            cex.axis = 2,
            cex.sub = 1,
            xlab = "Accepted Samples",
            breaks = (bins = 40),
            col = 'skyblue',
            freq = FALSE,
            main=sprintf("ACCEPTED SAMPLES"))
    # Plotting the Rejected Values
    h = hist(
            reject,
            cex.main = 2.5,
            cex.lab = 2,
            cex.axis = 2,
            cex.sub = 1,
            xlab = "Rejected Samples",
            breaks = (bins = 40),
            col = 'red',
            freq = FALSE,
            main=sprintf("REJECTED SAMPLES"))
}

if (user_choice == 2) {
    # Plotting the Accepted Values
    h = hist(
            accept,
            cex.main = 2.5,
            cex.lab = 2,
            cex.axis = 2,
            cex.sub = 1,
            xlab = "Accepted Samples",
            breaks = (bins = 40),
```

```
            col = 'skyblue',
            freq = FALSE,
            main=sprintf("ACCEPTED SAMPLES") )
}

if (user_choice == 3) {
    # Plotting the Rejected Values
    h = hist(
            reject,
            cex.main = 2.5,
            cex.lab = 2,
            cex.axis = 2,
            cex.sub = 1,
            xlab = "Rejected Samples",
            breaks = (bins = 40),
            col = 'red',
            freq = FALSE,
            main=sprintf("REJECTED SAMPLES"))
}
```

```
Options for Display :
1) Accepted and Rejected Values
2) PDF along with the Accepted Values
3) PDF along with the Rejected Values
Enter Your Choice:1
```







Proof that my Acceptance/Rejection percentage is correct:

Correctness from the graph:

- From my histogram of values from the Accept Vector, we can clearly see that :
  - It looks similar to the PDF.
  - It follows the Target Distribution.
  - Clealy, we can see that the histogram looks as though it comes from the PDF.
  - Thus, our values are correctly **SAMPLED** from our **Target Distribution.**
  - And only **30-35% of the samples are selected.**
- From my histogram of values from the Reject Vector, we can clearly see that :
  - It looks like an inverse of the the PDF.
  - It **does not** follow the Target Distribution.
  - Also, we can see that a large portion, around **60-66% of the samples are rejected.**
- **Hence,** our Rejection Sampling Process has successfully worked!