SCSI/SECI2143(04)
PROBABILITY & STATISTICAL DATA ANALYSIS 2020/2021 SEMESTER 2

PROJECT 2

LECTURER'S NAME: DR NOR AZIZAH ALI

NAME	MATRIC NUMBER
MUHAMMAD AFIQ ZAKWAN BIN ANUAR	A18KE0139
MUHAMMAD AMIR SYAFIQ BIN AHMAD RAZALI	A18KM0200
SHARTESWARY A/P BOJARAJOO	A20EC0225
NUR SYAKIRAH BINTI MOHD SHUKRI	A19EM0384

Introduction

This time for our PSDA Project 2, we conducted a study based on cars from various models and brands. The purpose of our study is to find their average acceleration.

Our dataset was taken from dataset public site, which is https://perso.telecom-paristech.fr/eagan/class/igr204/datasets. There are many datasets available there. The dataset that we conducted our study from is on topic Cars, which its direct download link is https://perso.telecom-paristech.fr/eagan/class/igr204/data/cars.csv.

This dataset contained raw data of 400 cars with various models and brands. Eight different aspects were measured from these 400 cars. One of the aspects is acceleration. We conducted our whole study for this project around that particular aspect. This raw data set at first, was not arranged properly. This made it harder for us to make our analysis. Thus we use https://www.convertcsv.com/csv-viewer-editor.htm to make our dataset more presentable and easier for us to analyze.

Other seven aspects that contained in this dataset are MPG (miles per gallon), that show petrol consumption, Cylinder, that show number of cylinders that particular cars have, Displacement, that show engine displacement value, Horsepower, that show maximum horsepower that cars have, Weight, that show the weight of that cars, Model, that show the year of that model released, and Origin, that show the origin for that cars. All of aspects definitely informative for us, but to make our analysis for our study easier, we chose acceleration aspect, as its range are not too large compare to other aspects.

In order for us to finish our study, we use Microsoft Excel and R Studio. This two software really help us to analysis our data set.

Dataset

The dataset used in this project is about the list of cars with their specific scop. The data have 9 column and 406 rows. The 406 rows are corresponding to the data of the cars and also the title and type of title. The 9 columns stand for each title or attributes from left to right. The first start from the left is Cars, MPG, Cylinders, Displacement, Horsepower, Weight, Acceleration, Model and Origin. For this project, we only take 40 out of 406 rows which are the data of cars for the sample in this analysis. The car is chosen from row 3 until 42. For the attributes, we only take two attributes only from the 9 listed which is the cars and the acceleration.

Data analysis

i. Hypothesis Testing 2 Sample

Famous car brand always tries to deliver the best product for their client. An observation was conducted where 406 types of cars were selected and their acceleration is recorded. It is of interest to determine if there is evidence, with significance level 5%, to support a claim that the mean acceleration exceeds 15mph.

$$H_0: \mu = 15$$
$$H_1: \mu > 15$$

to test this hypothesis, we use test statistic for population mean with variance unknown:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

We put read our value into our R script file and compute the formula in software R

```
#Hypothesis testing on population mean with unknown variance - one tailed

n = 406
s = sd(x)
xbar = mean(x)
mu = 15

# Calculate t statistics
t = (xbar-mu)/(s/sqrt(n))
```

Then we calculate the t value with respect to alpha 0.05 and degree of freedom 405 in R script file

```
# Calculate critical value
alpha = 0.05
t.alpha = qt(1-alpha, df=n-1)
```

After comparing both value

t	3.73543126636051
t.alpha	1.64862469317651

We concluded that, we reject μ=15 at the 0.05 level of significance, that the mean acceleration exceeds 15.

ii. Correlation test

To determine whether there is a correlation between two variables, we have to find the value of correlation coefficient. In this test, we are finding the correlation between variable Horsepower (hp) and Acceleration (mph). We read our value into software R and run our code to find the correlation coefficient.

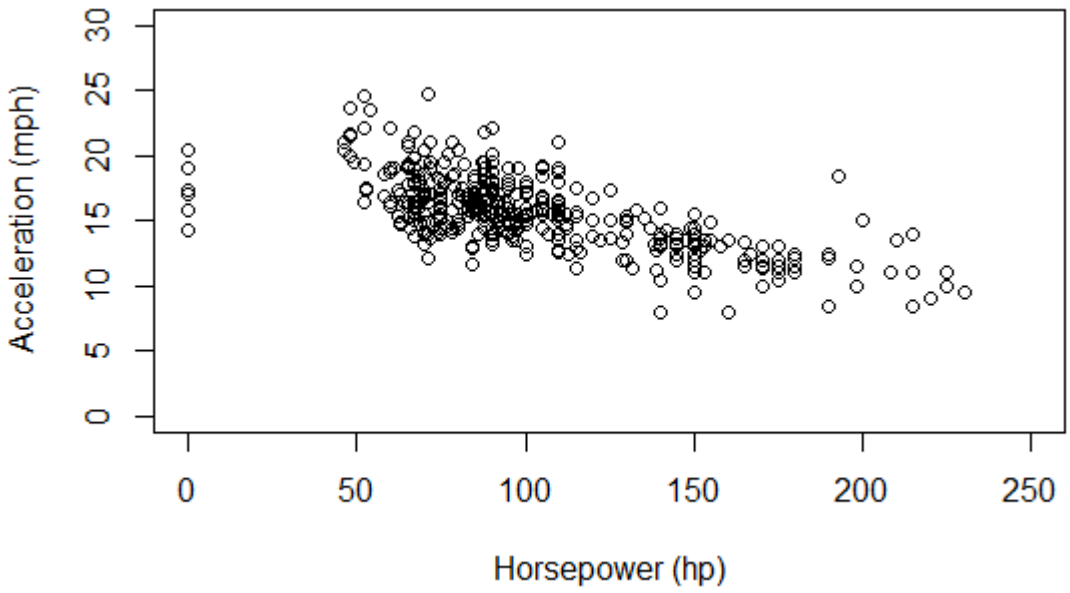
```
#load column 5(horsepower) to x as assumed to be an independent variable
x<-proj2[[5]]
#load column 7(acceleration) to y as assumed to be a dependent variable
y<-proj2[[7]]

#Calculate correlation coefficient
cor(x,y)
```

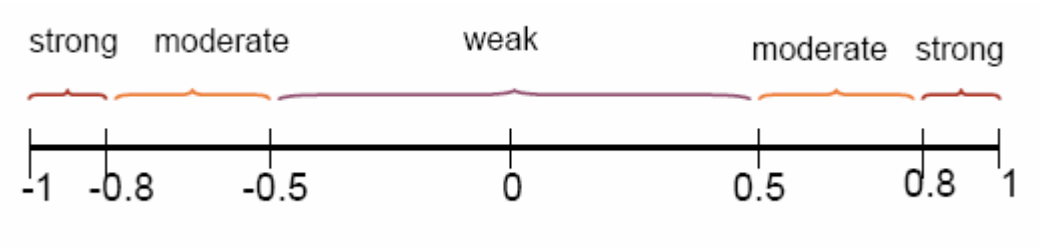
We found that the value in the console area of software R

```
> #Calculate correlation coefficient
> cor(x,y)
[1] -0.6820468
```

We also can plot the graph and see the scatter plot



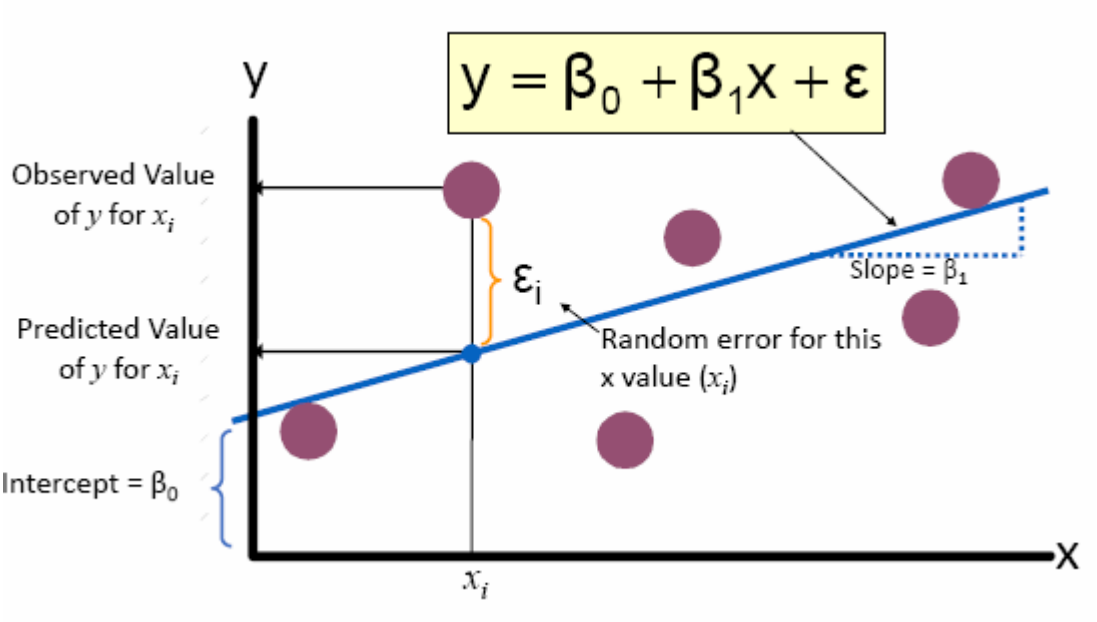
It can be seen that the acceleration decreases as the horsepower increases. A scatter plot and correlation analysis of the data indicated that there is negative relationship between the Horsepower (hp) and Acceleration (mph). For the correlation coefficient, we obtained -0.6820468



From this diagram we can conclude that horsepower and acceleration has a moderately negative linear relationship.

iii. Regression test

Regression test helps us to find the most suitable linear relationship for a scatter plot.



We can use this relationship and formula to find the regression line equation between horsepower and acceleration. To do this we insert our value into software R and run our code.

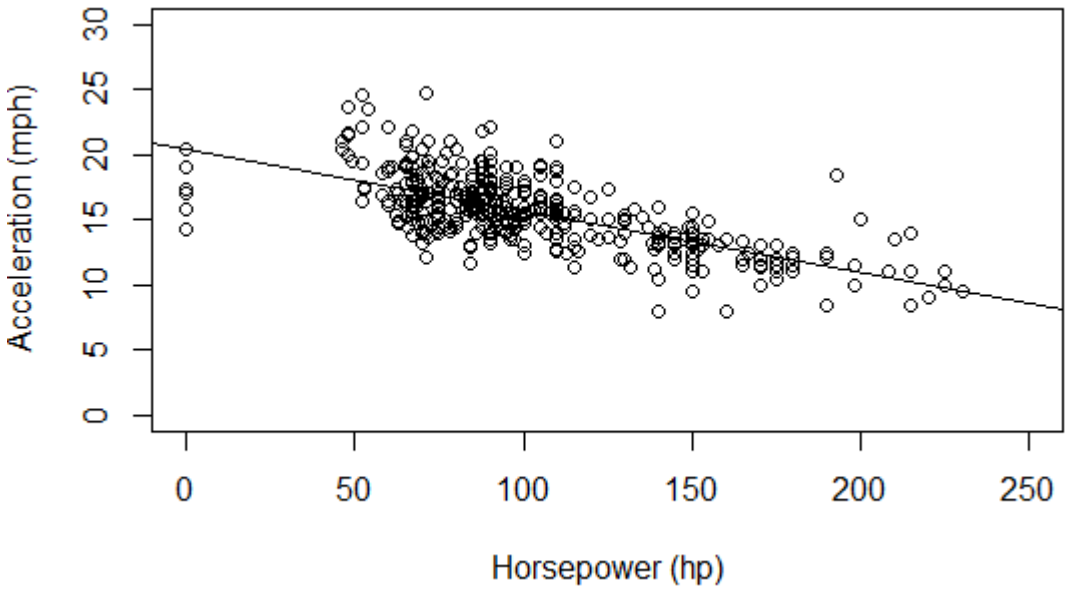
```
#load .csv file into proj2 dataframe with separator
file="E:/Hopes and Dreams/Computer Network & Computer Security/20202021-2/
proj2<-read.csv2(file, header = TRUE, sep = ";", quote = "\"", dec = ".")

#load column 5(horsepower) to x as assumed to be an independent variable
x<-proj2[[5]]
#load column 7(acceleration) to y as assumed to be a dependent variable
y<-proj2[[7]]


#Calculating Linear Regression
model<-lm(y~x)
model


#Plotting the Regression Line
plot(x,y, xlim=c(0,250), ylim=c(0,30), xlab="Horsepower (hp)", ylab="Accel
abline(model)
```

This code also plotted out regression line equation into the graph.



The regression equation obtained from software R is y=20.40489 -0.04719x. It means that no car has 0 horsepower, so 20.40489 indicates that, for car within the range of horsepower observed, 20.40489mph is the part of the car acceleration not explained by horsepower.

The -0.04719 tells us that the average acceleration of a car decreases by -0.04719(1000mph) =-47.19mph. On average, for each additional horsepower.

iv. Goodness-of-Fit Test

For this test we use a new dataset which contains 5 car company, with an average acceleration of their random 5 car.

Brand	Frequency	Average Acceleration
Chevrolet	5	11.5
Buick	5	11.8
Plymouth	5	10.7
AMC	5	12.4
Ford	5	11.1
Pontiac	5	12.1

$$H_0: p1 = p2 = p3 = p4 = p5 = p6$$

$$H_1: at\ least\ 1\ of\ the\ 5\ proportion\ is\ different$$

$$\chi^2 = \sum \frac{(O-E)^2}{E}:$$

To use the formula above we run our goodness-of-fit test in software R.

```
x <- proj1[[3]]
expprob <- sum(x)/6
expAcc <- c(expprob, expprob, expprob, expprob, expprob, expprob)

#Calculate x2 for each category
exp <- ((x-expAcc)^2)/expAcc

#Sum up to a x2 statistics
x2 <- sum(exp)
```

Then we would find the $x^2$with respect to alpha 0.05, and degree of freedom 5

```
#Critical value
alpha <- 0.05

#To get the x2 statistics for the corresponding alpha values/p-values
x2.alpha <- qchisq(alpha, df=5, lower.tail=FALSE)
```

The result of both value is as follows.

x2	0.172413793103448
x2.alpha	11.0704976935164

From this, since x2=0.1724 < x2 5,0.05=11.0705 we accept the hypothesis null. That is, we accept claim that the average acceleration of each of the company's car is of equal proportions (frequency) on their 5 random cars respectively.

Conclusion

After we finished our project, we become more understand about what we learn in class. All the test results show that we reject our hypothesis null.

For our first test at (i), we decided to use hypothesis test for one sample. For this test, we edit our Microsoft Excel document. As stated in the introduction, we focus our project on acceleration, thus we delete the other aspects data from our Excel document. We made this adjustment so we would not get distracted, as we were doing this test using R Studio. The result for this test show that we reject our hypothesis null.

For our second test at (ii), correlation test, we made the relationship between acceleration and horsepower. We were also doing this test using R Studio. For our dataset, 400 cars, this was more efficient for us, to get the result for our correlation test directly.

As for our test at (iii), which is regression test, we only take 40 sample data of cars from the 400 population of cars. As we did not make this test using R Studio, this is more efficient for us to make our analysis. The only tool that we used to make this test is Microsoft Excel. Using Microsoft Excel, we successfully make our scatterplot. Then by using the usual steps in regression test, we obtained our regression equation. From our regression model, we can see that the relationship for our dataset for the first 40 cars is negative linear.

For our last test on our dataset at (iv), which is Anova, we also only take 40 sample data of cars from the 400 population of cars. The result for this test is we rejected our hypothesis null, as Ftest statistic vale is larger than Fcritical value.

From all of our tests we can say that our study has a weakness. Acceleration aspect actually can be affected by various factors. Some of the various factors are stated in the whole dataset, which is the other aspects. For our project, we only focused on some of the aspects to relate it to the acceleration.

As we are finishing our project, we find quite a few interesting knowledges. One of them is the practical use of R Studio. In order to use R Studio, even we see our data in Excel document is overlapping with each, even it is in the correct column and row, R Studio still can process the data correctly. As all the example dataset online, all of them have a large number of data. The use of R Studio, make it possible for us to finish our project.