UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

Project 2 (Individual)

SECI2143 PROBABILITY & STATISTICAL DATA ANALYSIS

SEMESTER II, SESSION 2019/2020

Title: The horsepower, weight, miles per gallon, and engine size of cars among some brand of Japanese car

Lecturer: Dr. Chan Weng Howe

Name	Student ID
SEE WEN XIANG	A19EC0206

Section: 06

Programme: Bachelor of Computer Science (Software Engineering)

Table of Contents

1.0 Introduction and Background

The topic of this project is the horsepower, weight, miles per gallon, and engine size of cars among some brand of Japanese car.

The aim of study is to investigate whether there is a relationship between the horsepower, weight, city miles per gallon, and engine size of cars among some brand of Japanese car. Several statistical methods are used to estimate the relationships between the each of the variables.

2.0 Methodology

The data set used are obtained from online source, thus it is secondary data. A population of 205 data of cars is chosen. Then, 76 data are randomly picked from the population, which is Honda, Toyota, Nissan, Mitsubishi car. The data included type of car, horsepower, acceleration, car weight, fuel system, model, and others. A few variables are picked from the list for testing purpose, that is type of car, horsepower, weight, city mpg and engine size. Then, hypothesis testing, correlation, regression, chi square test for independence and ANOVA are used to test the sample data. The ways of analysis of data are by using RStudio to generate graphical presentation of data, and to do some of the basic calculation. Then, the conclusions are drawn.

3.0 Data Analysis and Results

3.1 Hypothesis testing 2 samples

Based on the data obtained, we want to study whether the weight of Honda Nissan is same as the weight of Toyota car. So, hypothesis test for 2 independent sample are carried out. The population variances are assumed to be unknown. In this case, since the number of data in the sample is 30 and 32 respectively,

the test statistic formula is $T_0^* = \dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$; degree of freedom is $v = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$

Statement: The weight of Honda Nissan car is different from weight of Toyota car

The null hypothesis is

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

where μ_1 represent population mean weight of Honda Nissan car, μ_2 represent population mean weight of Toyota car. The sample mean weight of Honda and Nissan car, $\bar{X}_1 = 2273.0645$, while

the sample mean weight of Toyota car is $\bar{X}_2 = 2441.0938$. The sample standard deviation is calculated using formula $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}$ for both of the samples.

So, the results are $s_1 = 464.6801$ and $s_2 = 354.5106$.

```
> library(readxl)
> hondanissan <- read_excel("AutoData.xlsx", sheet = "HT2", range = "A1:B32")
> toyota <- read_excel("AutoData.xlsx", sheet = "HT2", range = "D1:E33")
> xbar1<-mean(hondanissan$weight)
> xbar2<-mean(toyota$weight)
> s1<-sd(hondanissan$weight)
> s2<-sd(toyota$weight)
> n1<-nrow(hondanissan)
> n2<-nrow(toyota)
> v<-((s1^2/n1)+(s2^2/n2))^2/(((s1^2/n1)^2/(n1-1))+((s2^2/n2)^2/(n2-1)))
> t0<-(xbar1-xbar2-0)/(sqrt((s1^2/n1)+(s2^2/n2)))
> pvalue<-pt(t0,floor(v))
> #assume alpha=0.05
> alpha<-0.05
> t.alpha = qt(alpha/2,floor(v))
> criticalregion<-c(t.alpha,-t.alpha)
> cat("The test statistic is: ", t0)
The test statistic is:  -1.609957
> cat("The P-value is: ", pvalue)
The P-value is:  0.05651521
> cat("The critical value is: ", t.alpha)
The critical value is:  -2.003241
> cat("The critical region is: ", criticalregion)
The critical region is:  -2.003241 2.003241
```

Figure 3.1.1: The result of hypothesis testing of 2 sample, and test statistic, T_0^* by using RStudio

A significant level of 0.05 is used to test this claim. Since it is two-tail test, the critical value of 0.05 significance level is ±2.003241.

Hence, we reject H_0 if value of test statistic falls within the rejection area which is t<-2.003241 or t>2.003241. Since $T_0^* = -1.609957, -2.003241 < T_0^* < 2.003241$, thus we failed to reject H_0. We can conclude that we have strong evidence that the weight of Honda and Nissan car is same with the weight of Toyota car.

3.2 Correlation

In the correlation test, we measure the relationship between the weight of Toyota cars and the horsepower of Toyota cars in the sample size of 32. We use Pearson's technique to calculate a correlation coefficient since both data are ratio type.

Sample correlation coefficient:

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

where,

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

We use n=32 ,x=weight of Toyota cars and y=Horsepower of Toyota cars to calculate correlation coefficient by using RStudio.

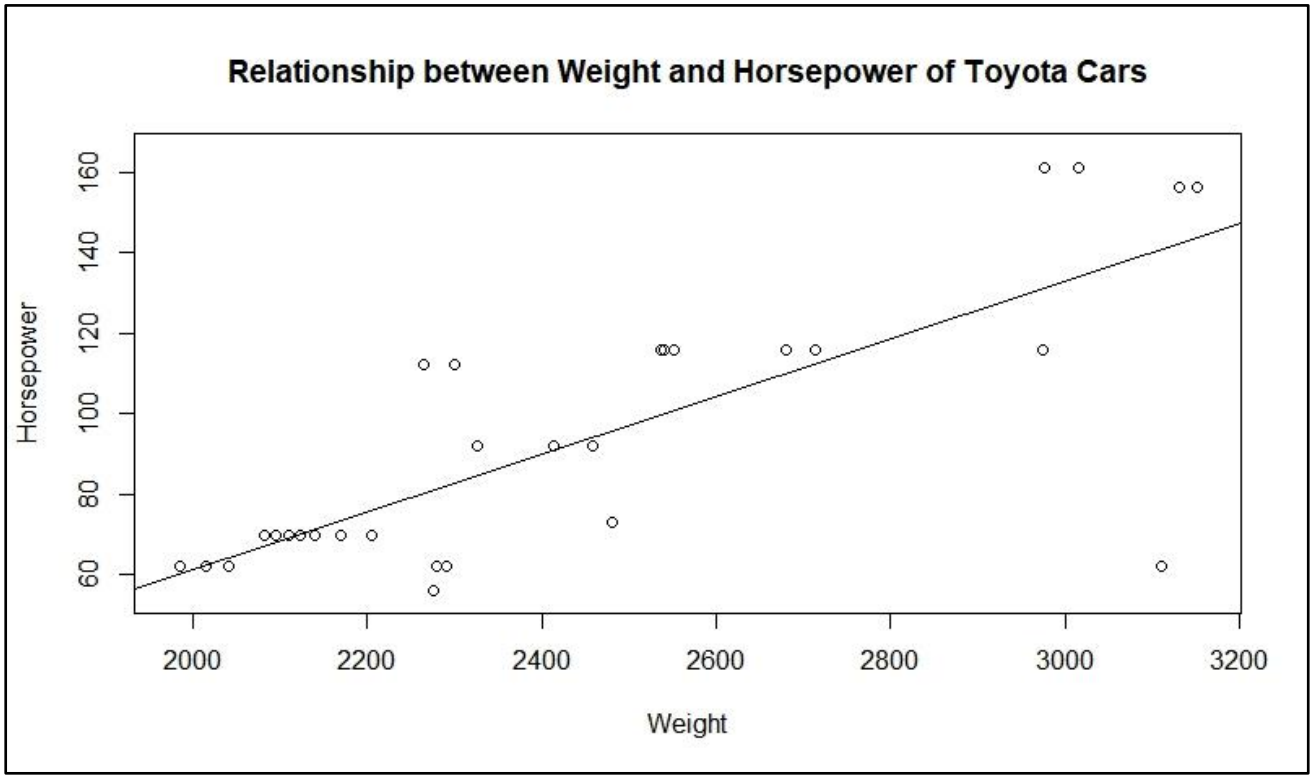

Figure 3.2.1: Scatter plot of horsepower of Toyota cars against weight of Toyota cars by using RStudio

```
> library(readxl)
> toyota <- read_excel("AutoData.xlsx", sheet = "CR", range = "A1:C33")
> x<-c(toyota$weight)
> y<-c(toyota$horsepower)
> plot(x,y,main = "Relationship between Weight and Horsepower of Toyota Cars",
xlim=c(1980,3155),ylim=c(55,165),xlab="Weight",ylab="Horsepower")
> abline(lm(y ~ x))
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 6.6064, df = 30, p-value = 2.595e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5756653 0.8818144
sample estimates:
      cor
0.7698304
```

Figure 3.2.2: The result of correlation coefficient, r and test statistic, t by using RStudio

Based on Figure 3.2.2, the correlation coefficient, r = 0.7698304. A scatter plot and correlation coefficient, r indicates that there is positive relationship between the weight and horsepower of Toyota cars. Besides, this is a moderate positive relationship because r falls within $0.576 < r < 0.882$. This means that the when the weight of car increase, the horsepower also increases.

Significance Test for correlation

We decide to test whether is there any evidence of a linear relationship between the weight of Toyota cars and horsepower of Toyota cars at the 0.05 level of significance.

Null hypothesis, H_0: $\rho = 0$ (no linear correlation)

Alternative hypothesis, H_1: $\rho \neq 0$ (linear correlation exists)

$\alpha = 0.05$, degrees of freedom, df = 32-2 = 20

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Based on Figure 3.2.2, the test statistic, t is 6.6064. P-value is the significance level of the t-test. P-value is 2.595×10^{-07}. Confidence interval of the correlation coefficient at 95% is (0.575665, 0.881814).

Conclusion: Since P-value from Figure 3.2.2 is 2.595×10^{-07} is less than the significance level of 0.05, so we reject the null hypothesis. There is sufficient evidence of a linear relationship between the weight and horsepower of Toyota cars at the 0.05 significance level.

3.3 Regression

In the regression test, we measure the relationship between the weight of Toyota cars and the horsepower of Toyota cars in the sample size of 32. The independent variable is weight and the dependent variable is horsepower. The sample regression line provides an estimate of the population regression line.

Estimated Regression Model:

$$Y = b_o + b_1 x$$

where:

Y = Estimated (or predicted) Y value

b_o = Estimate of the regression intercept

b_1 = Estimate of the regression slope

X = Independent variable

We use n=32, x=weight of Toyota cars and y= horsepower of Toyota cars to calculate estimated regression model by using RStudio.

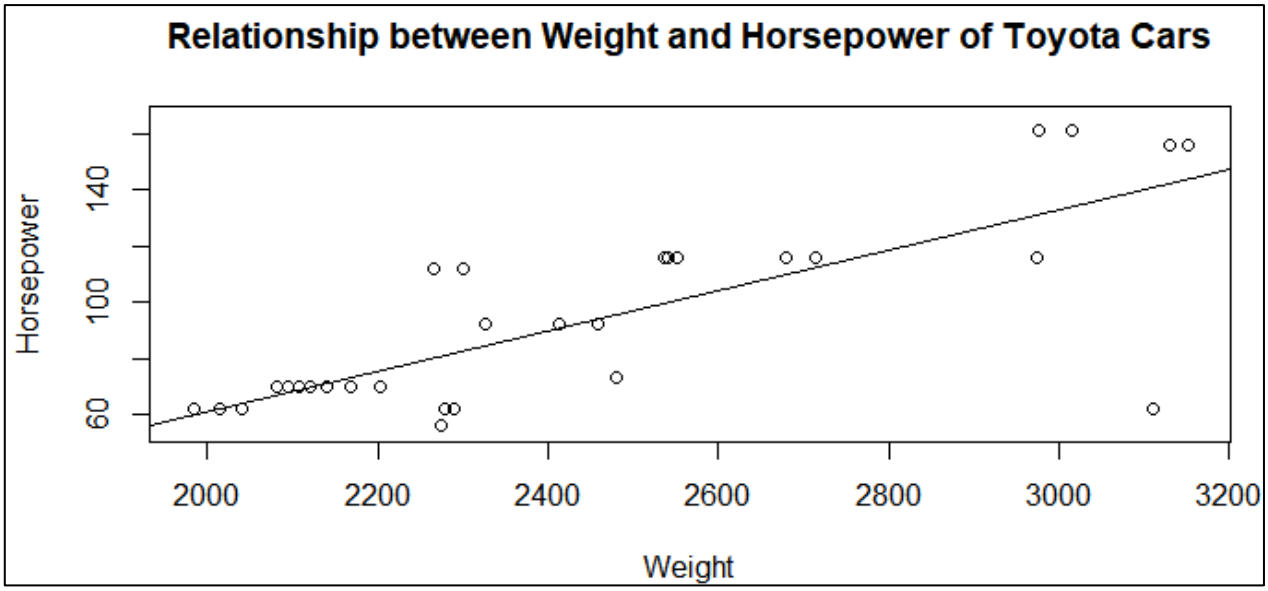

Figure 3.3.1: Scatter plot of horsepower of Toyota cars against weight of Toyota cars by using RStudio

```
> library(readxl)
> toyota <- read_excel("AutoData.xlsx", sheet = "CR", range = "A1:C33")
> x<-c(toyota$weight)
> y<-c(toyota$horsepower)
> plot(x,y,main = "Relationship between Weight and Horsepower of Toyota Cars",
xlim=c(1980,3155),ylim=c(55,165),xlab="Weight",ylab="Horsepower")
> abline(lm(y ~ x))
> (lm(y ~ x))

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
  -81.97367      0.07159

> summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-78.667  -3.929   1.158  12.755  31.825

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.97367   26.72117  -3.068  0.00454 **
x             0.07159    0.01084   6.606  2.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.39 on 30 degrees of freedom
Multiple R-squared:  0.5926,    Adjusted R-squared:  0.5791
F-statistic: 43.64 on 1 and 30 DF,  p-value: 2.595e-07
```

Figure 3.3.2: The summary of the graph by using RStudio

From the summary, we can get the formula for estimated regression model is:

Y= -81.974 + 0.072X

From the formula, we get that when X (weight of Toyota cars) is zero, we get negative value of horsepower of Toyota cars. This shows that there is no such thing exist.

b_1 measures the estimated change in the average value of Y because of a one-unit change in X

Here, b_1 = 0.072 tells us that the average value of horsepower of Toyota cars increases by 0.072, for each additional 1kg weight of Toyota cars.

The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{sum of square explaind by regression}}{\text{total sum of squares}}$$

From Figure 5, we get that the coefficient of determination, R^2 = 0.593.

Since, 0 <R^2< 1, it shows weaker linear relationship between x and y:

Some but not all the variation in y is explained by variation in x.

Test Statistical of Regression

$H_0 = \beta_1 = 0$ (no linear relationship)

$H_1 = \beta_2 \neq 0$ (linear relationship)

Test statistic, t = $\frac{b_1 - \beta_1}{S_{b_1}}$

Degree of freedom = n-2

Where:

b_1= Sample regression slope coefficient

β_1= Hypothesized slope

S_{b_1}=Estimator of the standard error of the slope

Based on Figure 3.3.2, test statistic, t = 6.06. P-value is the significance level of the t-test. P-value is 2.595×10^{-07}. Since P-value is 0.000000295 is less than significance level 0.05, therefore we reject the null hypothesis. So, there is sufficient evidence that the weight of Toyota cars affects the horsepower of Toyota cars.

3.4 Chi-Square Test of Independence

Chi-Square test of Independence is to test for relation between two nominal variables. The data can be represented in a contingency table where each row represents a category for one variable and the column represents another variable. In this test, we test whether there is a relationship between car's city miles per gallon and the brand of the car.

H_0: Car's city miles per gallon is independent with the brand of car
H_1: Car's city miles per gallon is dependent with the brand of car

Two-ways contingency table is drawn using R-programming.

```
> print(x2.alpha)
[1] 16.91898
> output$statistic
X-squared
 16.76501
> output$parameter
df
 9
> output$observed

            16<x<=25 25<x<=34 34<x<=43 43<x<=52
  Honda            2        9        1        1
  Mitsubishi      10        2        1        0
  Nissan           6       11        0        1
  Toyota          10       19        3        0
> output$expected

             16<x<=25  25<x<=34  34<x<=43  43<x<=52
  Honda       4.789474  7.013158 0.8552632 0.3421053
  Mitsubishi  4.789474  7.013158 0.8552632 0.3421053
  Nissan      6.631579  9.710526 1.1842105 0.4736842
  Toyota     11.789474 17.263158 2.1052632 0.8421053
```

Figure 3.4.1: Calculation and contingency table drawn using RStudio

The formula of test statistic is:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

From the calculation, the test statistic, chi-square value is 16.76501. Using 0.05 significance level and degree of freedom of 4, the critical value calculated is 16.91898.

Since the chi-square value is not in the critical region, we fail to reject the null hypothesis. Therefore, we have strong evidence that there is not a relationship between car's city miles per gallon and the brand of the car.

3.5 ANOVA

The purpose of ANOVA in the test is to test for significance differences between means of engine size among 4 brands of car (Honda, Toyota, Mitsubishi, Nissan) at the 0.05 significance level.

$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$ (all brand of car has the same mean of engine size)
$H_1 =$ at least one mean is different

$$F = \frac{variance\ between\ samples}{variance\ within\ samples}$$

```
> library(readxl)
> AutoData <- read_excel("AutoData.xlsx", sheet = "ANOVA")
> res.aov <- aov(Enginesize ~ Brand, data = AutoData)
> summary(res.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
Brand         3   3603  1201.1   3.687 0.0181 *
Residuals    48  15638   325.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.5.1: The result of ANOVA

Based on the Figure 3.5.1, the numerator = k-1 =4-1=3. The denominator = k (n - 1) = 4(13-1) = 48. The test statistic, F = 3.687. P-value is the significance level of the f-test. P-value is 0.0181.

Since P-value of F =0.0181 is less than the significance level 0.05, therefore we reject the null hypothesis. There is insufficient evidence to claim that the different types of cars among 4 brands of car have the same mean of engine size.

4.0 Discussion and Conclusion

From the hypothesis test, we can conclude that the weight of Honda and Nissan car is different than weight of Toyota car. From the correlation test, it shows the result that the weight and horsepower of Toyota cars have a positive relationship, that is, if the car is heavier, it has larger horsepower. From the regression test, for Toyota car, it shows that for each increase of 1kg of weight of car, the horsepower also increased by 0.072. From the Chi-Square test, it shows that there is a relation between car's city miles per gallon and the brand of the car, meaning that car from different brands has different city mpg. From the ANOVA test, it shows that the mean engine size of each brand is different.

5.0 References

Rushan, M. (24 January, 2020). *kaggle*. Retrieved from Automobile Dataset: https://www.kaggle.com/mrushan3/automobile-dataset