SCSI2143 / SCSI2143

PROBABILITY & STATISTICAL DATA ANALYSIS

2019/2020 – SEMESTER 2

PROJECT 2 - REPORT

ARIF BIN SUHAIMI A19EC0023

Dr. Suhaila Mohamad Yusuf

UNIVERSITI TEKNOLOGI MALAYSIA

TABLE OF CONTENT

INTRODUCTION

This course is designed to introduce some statistical techniques as tools to analyse the data. In the beginning the students will be exposed with various forms of data. The data represented by the different types of variables are derived from different sources; daily and industrial activities. The analysis begins with the data representation visually. The course will also explore some methods of parameter estimation from different distributions. Further data analysis is conducted by introducing the hypothesis testing. Some models are employed to fit groups of data. At the end of course the students should be able to apply some statistical models in analysing data using available software.

Project 2, an individual project that want ask to study any dataset that we ca obtain from any source. Scope of this project is Inferential Statistics and the type of data is secondary data (from any organization and websites). Inferential Statistics Inferential insights are frequently utilized to compare the contrasts between the treatment bunches. Inferential insights utilize estimations from the test of subjects within the try to compare the treatment bunches and make generalizations approximately the bigger populace of subjects. There are numerous sorts of inferential measurements and each is fitting for a investigate plan and test characteristics. Analysts ought to counsel the various writings on test plan and statistics to discover the correct factual test for their try. There are 3 compulsory items: Hypothesis Testing (1 or 2 sample), correlation and Regressions. Other than that, we must choose at least one from ANOVA, Goodness Fit Test and Chi Square Test of Independence. Before start the programming and report, we must send a proposal to Dr Suhaila. The proposal consist of source

of datasets (including description), Variables, description of purpose of study, specification of target population, selection of variables, proposed analysis and expected outcome for analysis.

FOCUS ON TOPIC (CONTENT)

Project 2, I choose dataset from :

https://perso.telecom-paristech.fr/eagan/class/igr204/datasets .

I choose the dataset about cars. This dataset consists of 408 cars from USA, Europe and Japan.

	Car	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model	Origin
2	STRING	DOUBLE	INT	DOUBLE	DOUBLE	DOUBLE	DOUBLE	INT	CAT
3	Chevrolet Chevelle Malibu	18	8	307	130	3504	12	70	US
4	Buick Skylark 320	15	8	350	165	3693	11.5	70	US
5	Plymouth Satellite	18	8	318	150	3436	11	70	US
6	AMC Rebel SST	16	8	304	150	3433	12	70	US
7	Ford Torino	17	8	302	140	3449	10.5	70	US
8	Ford Galaxie 500	15	8	429	198	4341	10	70	US
9	Chevrolet Impala	14	8	454	220	4354	9	70	US
10	Plymouth Fury iii	14	8	440	215	4312	8.5	70	US
11	Pontiac Catalina	14	8	455	225	4425	10	70	US
12	AMC Ambassador DPL	15	8	390	190	3850	8.5	70	US
13	Citroen DS-21 Pallas	0	4	133	115	3090	17.5	70	Europe
14	Chevrolet Chevelle Concour	0	8	350	165	4142	11.5	70	US
15	Ford Torino (sw)	0	8	351	153	4034	11	70	US
16	Plymouth Satellite (sw)	0	8	383	175	4166	10.5	70	US
17	AMC Rebel SST (sw)	0	8	360	175	3850	11	70	US
18	Dodge Challenger SE	15	8	383	170	3563	10	70	US
19	Plymouth 'Cuda 340	14	8	340	160	3609	8	70	US
20	Ford Mustang Boss 302	0	8	302	140	3353	8	70	US

Every car that these three-country produce involve in this dataset. I would like to study the difference characteristics of each car that was written into that dataset. Other than that, I would like to study on how car manufacturer produce car with higher specification but saves more energy. Next, I would like to study which country produces the best car. The variables that I choose from this datasets are Mileage per Gallon (MPG), Displacement, Horsepower, Weight

and Acceleration. These variables will help me to calculate Hypothesis Testing for 2-Sample, Correlation, Regression and ANOVA. The IDE that were use for this project that will helped to calculate is Rstudio that uses R language (a specific language for statistics).

SUPPORT OF TOPIC (RESULTS) :

HYPOTHESIS TESTING 2-SAMPLE

Data	
euro	30 obs. of 1 variable
us	30 obs. of 1 variable
Values	
alpha	0.05
n1	30
n2	30
s1	8.74912146656918
s2	5.3309980519491
t.alpha	-2.01174051372977
t0	-3.08290237848395
v	47.9249536861705
x	int [1:30] 18 15 18 16 17 15 14 14 14 15 ...
xbar1	15.0666666666667
xbar2	20.8333333333333
y	int [1:30] 32 28 24 26 24 26 31 19 18 15 ...

H0 : mean 1 = mean 2

H1 : mean1 != mean 2

Number of variable : 2

Number of data, n : 30

Mean 1 : xbar1 = 15.066667

Mean 2: xbar2 = 20.833333

Standard deviation 1 : s1 = 8.7491214456918

Standard deviation 2 : s2 = 5.3309980519491

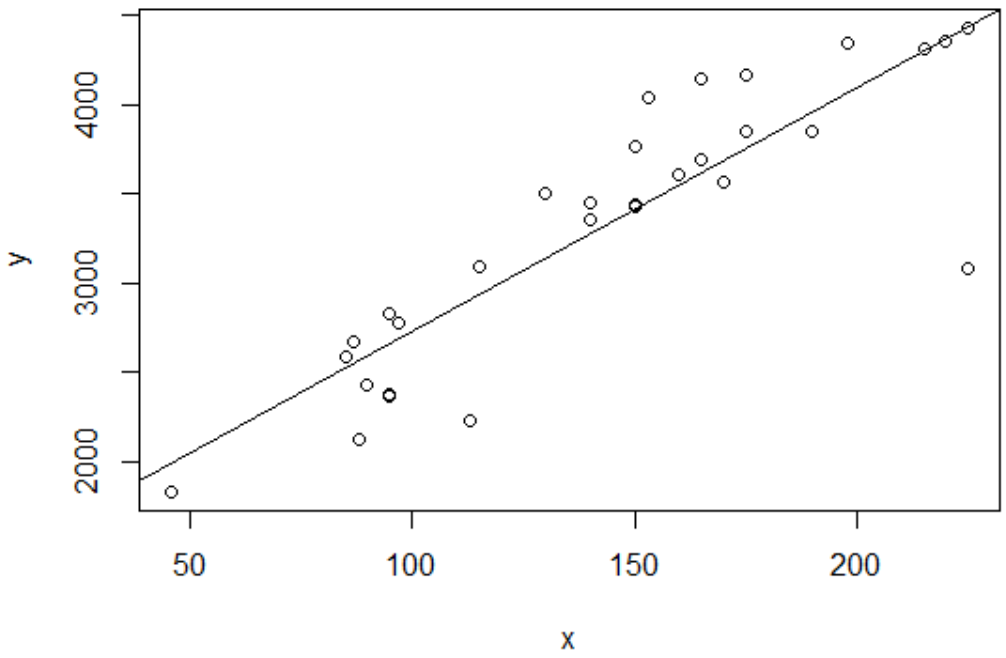
Test Statistic : t0 = -3.08290237848395

Critical value : t.alpha = -2.011740551372977

Decision : Reject H0, there are sufficient evidence that the mean of mpg for euro cars is not equal us.

REGRESSION

```
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    1371.18        13.61

> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1347.82  -118.48    59.39   164.90   580.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1371.176    219.721   6.241 9.60e-07 ***
x             13.612      1.456   9.352 4.14e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 376 on 28 degrees of freedom
Multiple R-squared:  0.7575,    Adjusted R-squared:  0.7488
F-statistic: 87.45 on 1 and 28 DF,  p-value: 4.138e-10
```



X = horsepower , Y = weight

H0 : B1 = 0

H1 : B1 != 0

p-value = 0.000000000413

b1 : x = 13.61

b0 : y = 1371.18

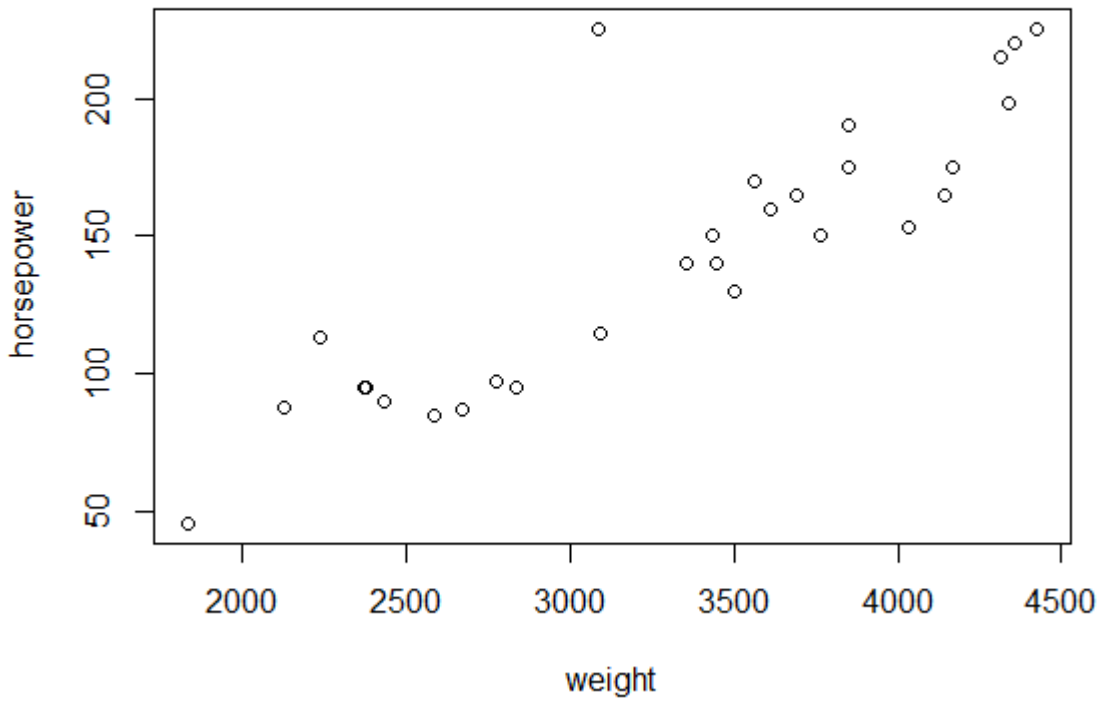
sb1 = 1.456

Test Statistics = (13.61-0)/1.456 = 9.532

Critical Value = 2.048

Decision : Reject H0 , there is sufficient evidence at 95% that weight affects horsepower.

CORRELATION



Cor (x,y) = 0.88703325

Based on the value 0.88703325 > 0.8 and the plotted graph is positive, the relationship between horsepower and weight is strong.

ANOVA

```
> summary(anova_result)
            Df    Sum Sq   Mean Sq F value Pr(>F)
ind          2 210890759 105445380   559.9 <2e-16 ***
Residuals   87  16385100    188334
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

H0 : mean 1= mean2= mean 3

H1 : at least one of them is different

Pvalue = 0.000000000000002

Numerator = 2

Denominator = 87

Test Statistics = 105445380/188334 = 559.9

Critical Value = F(2,87,0.05) = 3.11

Decision, Reject H0.

CONCLUSION

Based on the calculation that I have done, most of the calculation lead to the rejection of H0 in other word, the first mean of a data does not equal to another mean of the data. The plotted graph for correlation show us that the relationship is strong. If we refer t o the scale that Dr Suhaila provide us in the slide, the value of that we calculate is greater than 0.8 and it is poritively plotted at the first quadrant. That is a proof that the Y axis and the x axis has a strong relationship between them. Hypothesis 2-Sample, comparing mpg of us cars manufacturer with Europe manufacturer. The calculation show us that we reject H0 and this proof that the mpg of Europe cars is different form US cars. Most of the findings lead to the same decision rejecting H0.