SCSI2143 / SCSI2143

Probability & Statistical Data Analysis
2019/2020 – Semester 2

Project 2

NAME	:	EYU SI XIONG
MATRIC NUMBER	:	A19EC0044
SECTION	:	04
COURSE NAME	:	BACHELOR OF COMPUTER SCIENCE (COMPUTER NETWORK & SECURITY
LECTURER’S NAME	:	DR. SUHAILA MOHAMAD YUSUF
SUBMISSION’S DATE	:	27TH June 2020

Table of Content

Introduction

Dataset that I used in this project is the Road Safety Statistic book. All the information in this book was renewed at 22 July 2019. The formal data and non-formal data form newspapers, news report and others were collected and published by the Road Safety Department of Malaysia. Inside this book is publishing about different kind of data such as vehicles and driver's data, road accident's data and also some analysis and comparison. The purpose that I use this statistic as my project is to show the conditions and relationship of the road accident happened in Malaysia. The selected variables in data for this project are year (2011-2018), age, population in Malaysia, Number of vehicles in each state and in Malaysia and accident's death.

There are 4 studies that I have done in this project and these studies are targeting the people who involved in the road accident:

- To find out whether the number of vehicles will affect the number of accident's death in Malaysia or not from year 2011 to year 2018
- To study the relationship between the number of vehicles and the number of deaths in accident in different states in 2017
- To investigate the proportion of road accident's death in 2017 is less than in 2016
- To study whether the total of road accident's death in 2017 is different in all grouped ages or not

Hence, the proposed analysis that I will used for the studies are regression, correlation, 1-sample hypothesis test and goodness of fit test. All the test statistics will be calculated by using R studio.

Analysis and Result (Hypothesis testing)

Regression

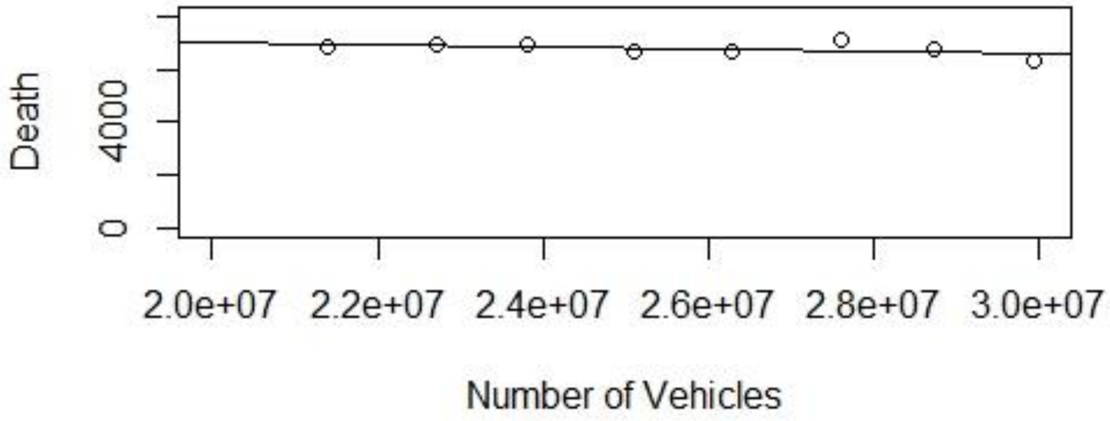
1st study: To find out whether the number of vehicles, x will affect the number of accident's death, y in Malaysia or not from year 2011 to year 2018

Significance level used: 0.10

H0:B1 = 0 (the number of vehicles, x will not affect the number of accident's death, y in Malaysia)

H1:B1 != 0 (the number of vehicles, x will affect the number of accident's death, y in Malaysia)

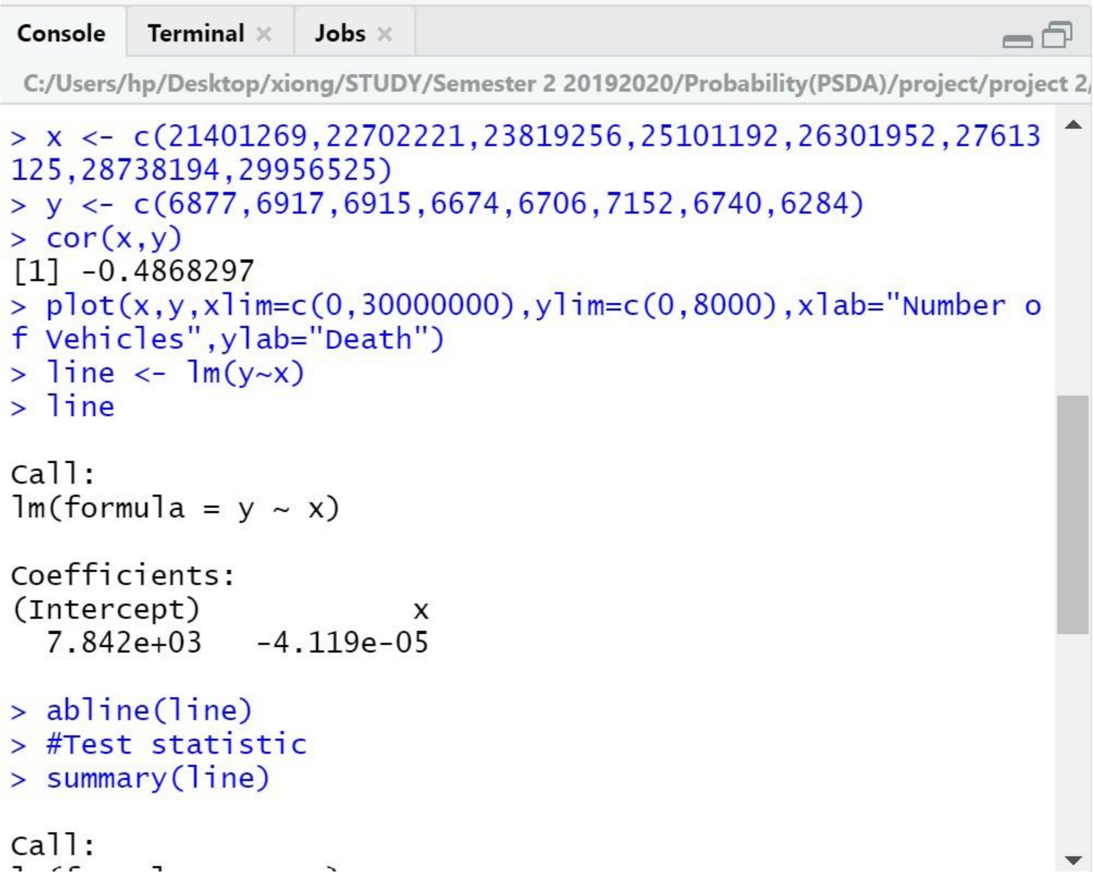
Year	Number of vehicles, x	Death, y
2011	21401269	6877
2012	22702221	6917
2013	23819256	6915
2014	25101192	6674
2015	26301952	6706
2016	27613125	7152
2017	28738194	6740
2018	29956525	6284



Test statistic:

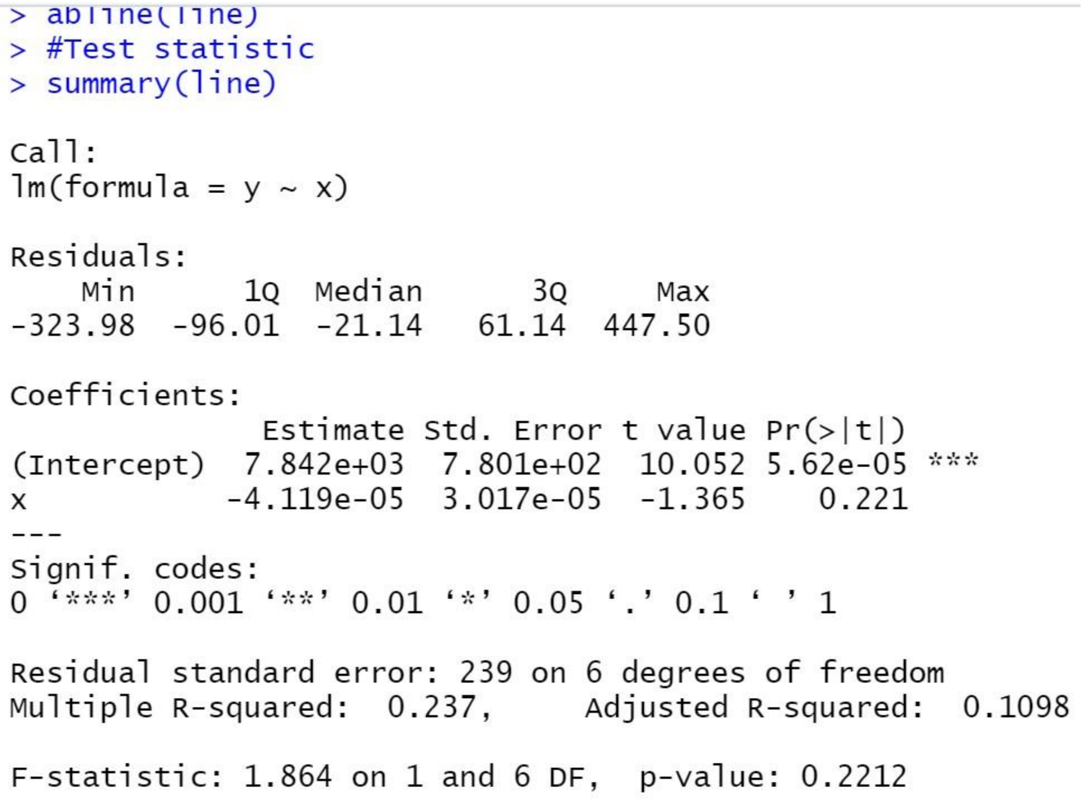
r = -0.487

p-value = 0.2212


Console
Terminal
Jobs
C:/Users/hp/Desktop/xiong/STUDY/Semester 2 20192020/Probability(PSDA)/project/project 2/

```
> x <- c(21401269,22702221,23819256,25101192,26301952,27613
125,28738194,29956525)
> y <- c(6877,6917,6915,6674,6706,7152,6740,6284)
> cor(x,y)
[1] -0.4868297
> plot(x,y,xlim=c(0,30000000),ylim=c(0,8000),xlab="Number o
f Vehicles",ylab="Death")
> line <- lm(y~x)
> line

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
  7.842e+03   -4.119e-05

> abline(line)
> #Test statistic
> summary(line)

Call:
```

Critical value: p-value = 0.10



```
> abline(line)
> #Test statistic
> summary(line)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-323.98  -96.01  -21.14   61.14  447.50 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.842e+03  7.801e+02  10.052 5.62e-05 ***
x           -4.119e-05  3.017e-05  -1.365    0.221    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 239 on 6 degrees of freedom
Multiple R-squared:  0.237,	Adjusted R-squared:  0.1098 

F-statistic: 1.864 on 1 and 6 DF,  p-value: 0.2212
```

Decision: H0 is failed to reject since 0.2212>0.10

Conclusion: There is insufficient evidence to prove that the number of vehicles will affect the number of accident’s death in Malaysia or not from year 2011 to year 2018 at 90% confidence level

Correlation

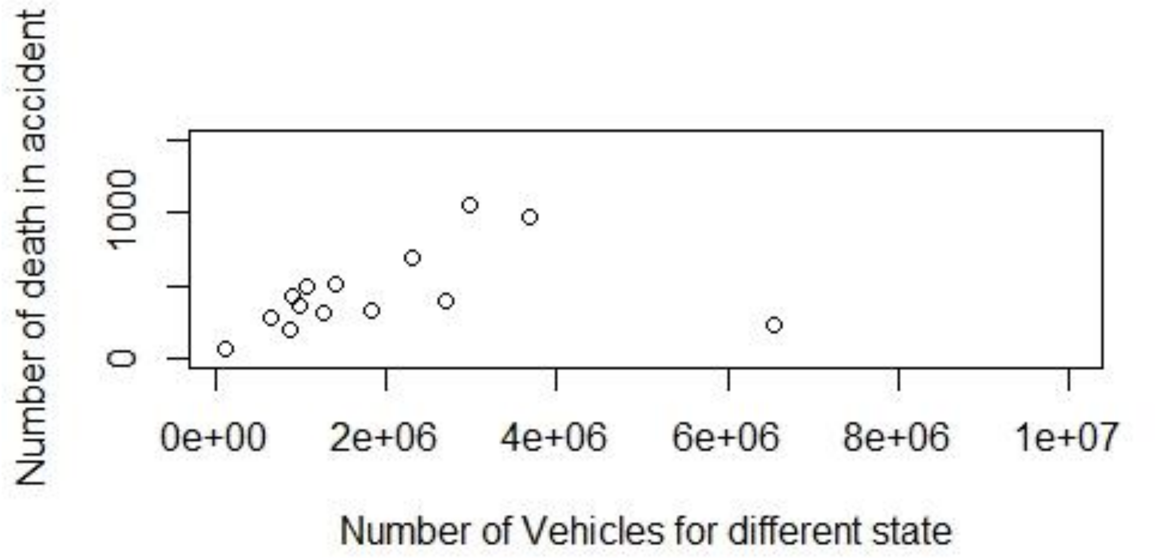
2nd study: To study the relationship between the number of vehicles and the number of deaths in accident in different states (2017)

Significance level used: 0.01

H0: p = 0 (No relationship between the number of vehicles and the number of deaths in accident in different states (2017))

H1: p != 0 (has relationship between the number of vehicles and the number of deaths in accident in different states (2017))

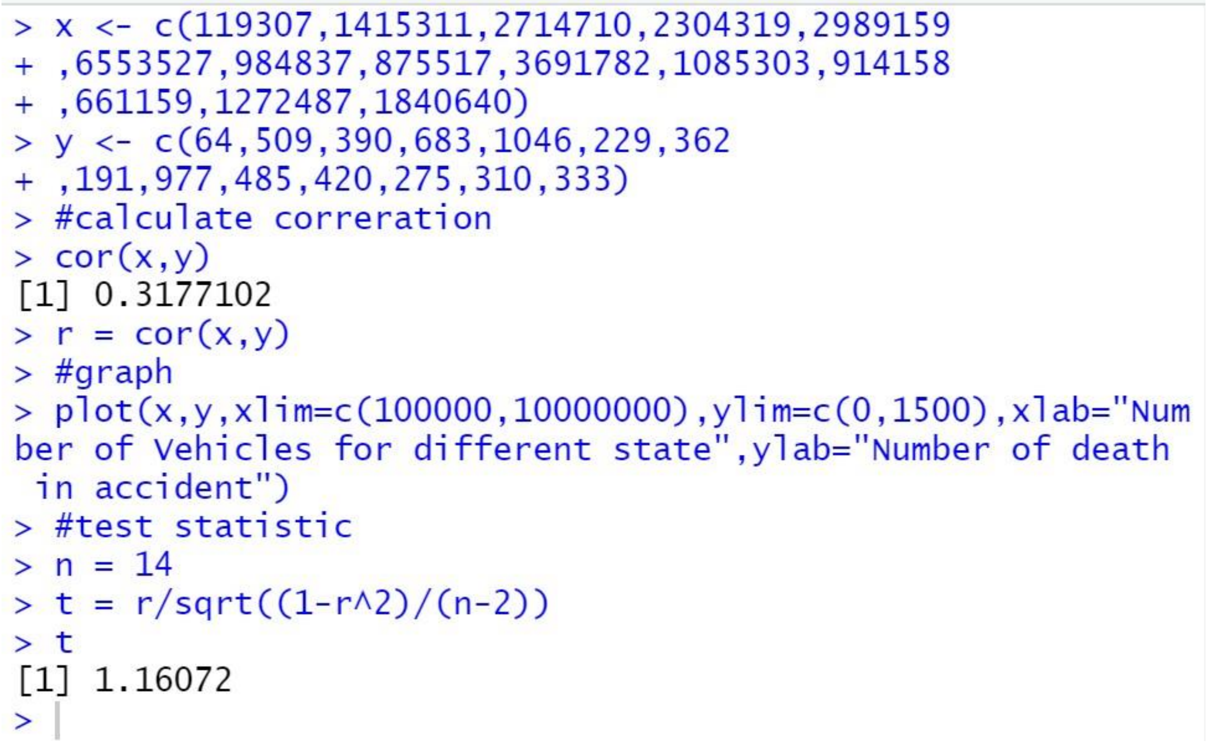
State	Death	Vehicle
Perlis	64	119307
Kedah	509	1415311
Pulau Pinang	390	2714710
Perak	683	2304319
Selangor	1046	2989159
Wilayah Persekutuan	229	6553527
Negeri Sembilan	362	984837
Melaka	191	875517
Johor	977	3691782
Pahang	485	1085303
Kelantan	420	914158
Terengganu	275	661159
Sabah	310	1272487
Sarawak	333	1840640



Test Statistic:

r = 0.3177

t value = -1.161



```
> x <- c(119307,1415311,2714710,2304319,2989159
+ ,6553527,984837,875517,3691782,1085303,914158
+ ,661159,1272487,1840640)
> y <- c(64,509,390,683,1046,229,362
+ ,191,977,485,420,275,310,333)
> #calculate correration
> cor(x,y)
[1] 0.3177102
> r = cor(x,y)
> #graph
> plot(x,y,xlim=c(100000,10000000),ylim=c(0,1500),xlab="Num
ber of Vehicles for different state",ylab="Number of death
 in accident")
> #test statistic
> n = 14
> t = r/sqrt((1-r^2)/(n-2))
> t
[1] 1.16072
> 
```

Critical value:

Df = 14-2 = 12

t (0.01,12) = (-3.054, 3.054)

Decision: H0 is not rejected since -1.161 > -3.054

Conclusion: There is insufficient evidence of a linear relationship between the number of vehicles and the number of deaths in accident in different states (2017) at 99% confidence level.

Hypothesis 1-sample test

3rd study: To investigate the proportion of road accident's death in 2017 is less than in 2016

Significance level used: 0.05

p2016 = proportion of road accident's death in 2016

p2017 = proportion of road accident's death in 2017

H0: p2017 >= p2016

H1: p2017 < p2016

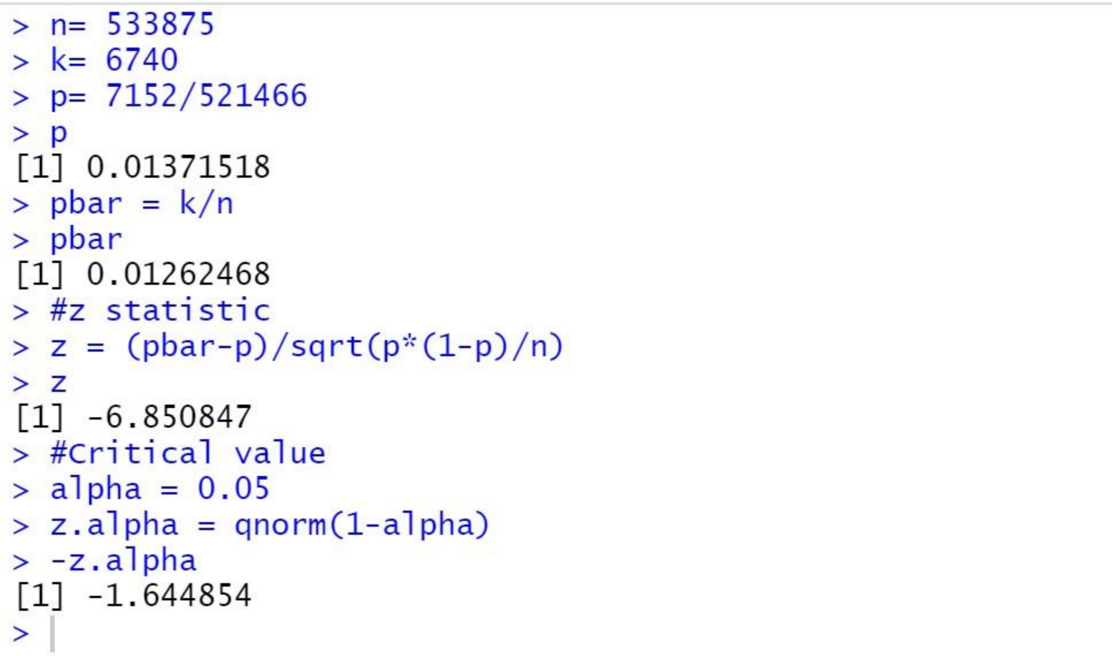
Year	Number of Accident	Death
2016	521466	7152
2017	533875	6740

p2016 = 0.0137

p2017 = 0.0126

Test Statistic: z = -6.851

Critical value: z (0.05) = -1.645



```
> n= 533875
> k= 6740
> p= 7152/521466
> p
[1] 0.01371518
> pbar = k/n
> pbar
[1] 0.01262468
> #z statistic
> z = (pbar-p)/sqrt(p*(1-p)/n)
> z
[1] -6.850847
> #Critical value
> alpha = 0.05
> z.alpha = qnorm(1-alpha)
> -z.alpha
[1] -1.644854
> 
```

Decision: H0 is rejected since -6.851 < -1.645

Conclusion: There is sufficient evidence to conclude that the proportion of road accident's death in 2017 is less than in 2016 at 95% confidence level.

Goodness of Fit Test

4th study: To study whether the total of road accident's death in 2017 is different in all grouped ages or not

Significance level used: 0.01

H0: p1, p2, p3, …, p16 have the same proportion value

H1: At least one of the proportion values is different from others

Ages	Death
0 to 5	75
6 to 10	73
11 to 15	417
16 to 20	1090
21 to 25	989
26 to 30	614
31 to 35	534
36 to 40	428
41 to 45	382
46 to 50	389
51 to 55	370
56 to 60	362
61 to 65	295
66 to 70	263
71 to 75	214
greater than 75	245
Total	6740

Test Statistic:

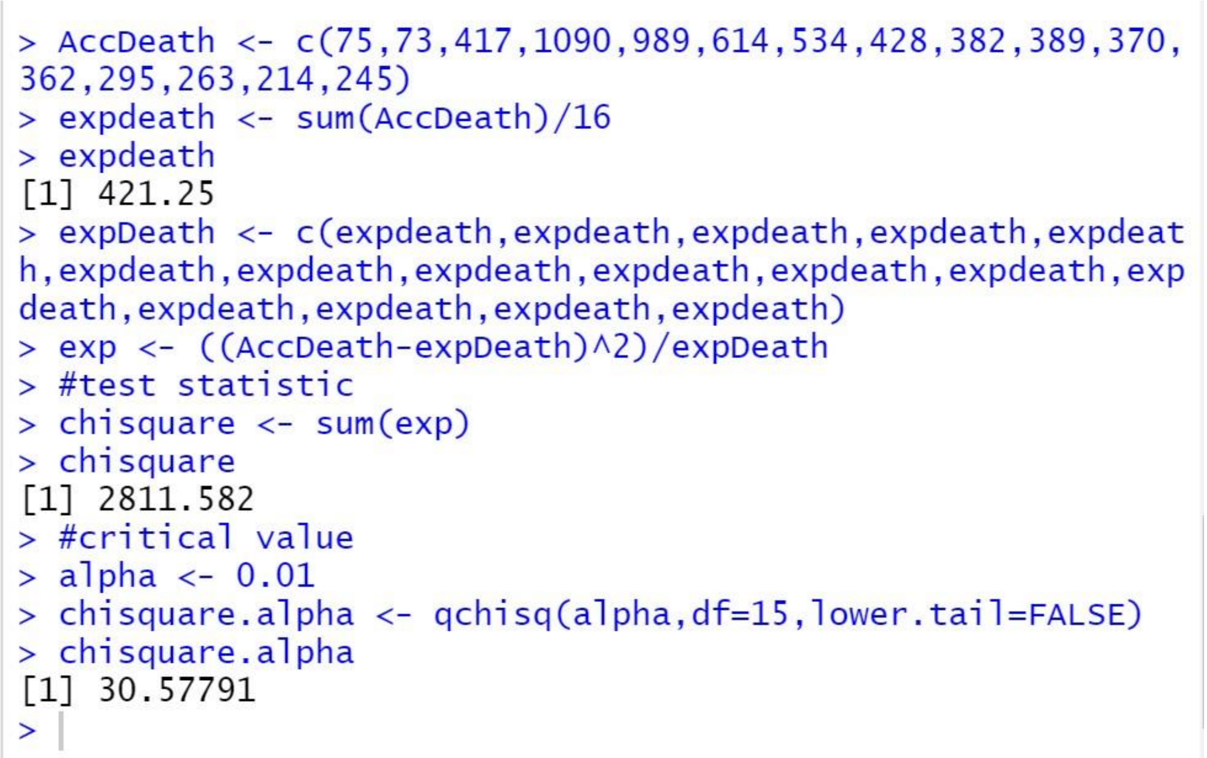
Expected death value for each grouped age= 421.25

X2 value = 2811.582

Critical value:

Df = 16-1 = 15

X2(0.01,15) = 30.578



```
> AccDeath <- c(75,73,417,1090,989,614,534,428,382,389,370,
362,295,263,214,245)
> expdeath <- sum(AccDeath)/16
> expdeath
[1] 421.25
> expDeath <- c(expdeath,expdeath,expdeath,expdeath,expdeat
h,expdeath,expdeath,expdeath,expdeath,expdeath,expdeath,exp
death,expdeath,expdeath,expdeath,expdeath)
> exp <- ((AccDeath-expDeath)^2)/expDeath
> #test statistic
> chisquare <- sum(exp)
> chisquare
[1] 2811.582
> #critical value
> alpha <- 0.01
> chisquare.alpha <- qchisq(alpha,df=15,lower.tail=FALSE)
> chisquare.alpha
[1] 30.57791
> 
```

Decision: H0 is rejected since 2811.582 > 30.578

Conclusion: There is sufficient evidence to prove that the total of road accident’s death in 2017 is different in all grouped ages at 99% confidence level.

Discussion and Conclusion

Now, I would like to discuss about the result obtained from the 4 studies above. The first study is I uses the regression method to test and conclude that the number of vehicles does not affect the number of accident’s death in Malaysia. Next, the second study shows that the number of vehicles in the state does not have the relationship with the number of accident’s death in the state.

After that, the proportion of number of deaths in 2017 is less than in 2016. This means that the number of accident’s death started to decrease. The last study is the number of accident’s death in 2017 are different in different grouped age. This shows that people in age between 16-25 has the greatest number of accident’s death since mostly they like go for outplay and also some people with their immature thinking.

In my opinion, I think the most important is the action of the drivers. No matter how many numbers of vehicles, if most of the drivers has a good driving ethics and discipline habits then it can reduce in the frequency of road accidents happened. Moreover, follow the road safety rules and mostly you can avoid from the road accidents.

In conclusion, I hope that the rate of road accidents will be decreased year by year.

References

1. BUKU-STATISTIK-KEMALANGAN-JALAN-RAYA. Retrieved from: http://www.jkjr.gov.my/ms/muat_turun/Statistik---Statistic/BUKU-STATISTIK-KEMALANGAN-JALAN-RAYA-(Kemaskini-22-07-2019)/lang,ms-my/

2. R Part 5 (Correlation and Regression)