UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROJECT 2 STASTISCAL AND DATA ANALYSIS :

STUDY OF DEATH IN UNITED STATES

Name: Muhammad Mukhlis Bin MOHD Feris Halmy

Matrics: A19EC5225

Lecturer: DR. Ernie Nazira binti Bazin

Table of Contents

INTRODUCTION

As we know, there will always death happen in our daily life. It can said that about 150,000 people die a day. It is a large number for people if It in small population.

For this project, I will be studying and make analysis on death in United states. I use the dataset from website called https://www.kaggle.com/ which contain a lot of dataset. My dataset name is 'Leading Causes of Death in the USA' which data from 1900 to 2013. The data set consist of variables such as number of death , year, gender and age adjusted death rate (AADR).

Main objective is to analyse and see the trends of death in U.S. We will get the data whether the date rate is getting bigger by year of it is biased by gender.

HYPOTHESIS TESTING(TWO SAMPLES)

H_o: The average number of deaths for year 1999 is equal 2005.

$H_{1:}$ The average number of deaths is greater than 2005.

Calculation using R

-Hyphothesis testing two tailed tests.

- confidence level is 95%

```
        Welch Two Sample t-test

data:  YEAR == "1999" and YEAR == "2005"
t = 0, df = 26518, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006004824  0.006004824
sample estimates:
 mean of x  mean of y
0.06666667 0.06666667
```

- t_{calc} = 0.0
- $T_{26518,\ 0.05}$ = 1

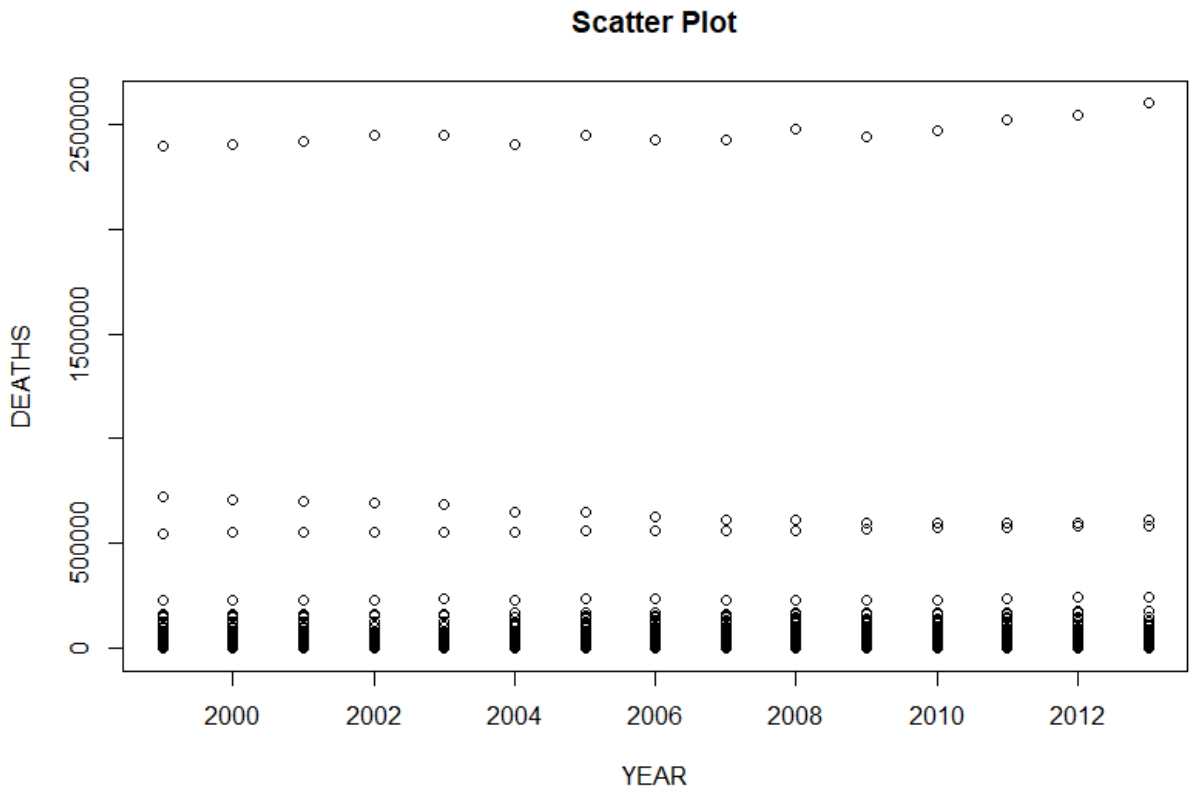
Since t_{calc} = 0.0 < $T_{26518,\ 0.05}$ = 1 , we fail to to reject the null hypothesis. There is sufficient evidence to support and conclude that the average numbet of death for year 1999 is equal 2005.

CORRELATION

YEAR AGAINST DEATH

H_0 : Y = 0 (There is no linear correlation)

H_1 : Y ≠ 0 (There is linear correlation)



Calculation using R

```
        Pearson's product-moment correlation

data:  YEAR and DEATHS
t = 0.13046, df = 13243, p-value = 0.8962
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01589723  0.01816388
sample estimates:
         cor
0.001133651
```

- $\mathbf{t_{calc}}$ = 0.13046
- Z critical value = 0.8962

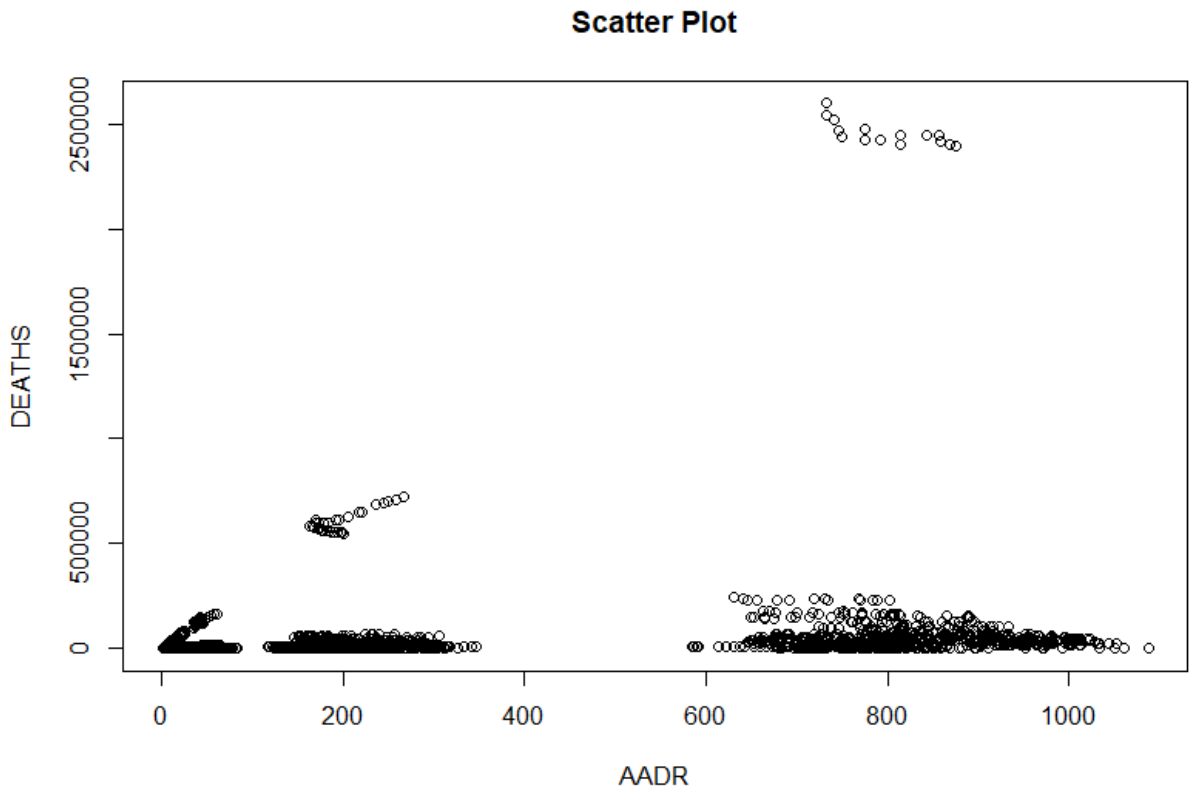
Since the Z critical value = 0.8962 > tcalc = 0.13046 , we reject H_0 , null hypothesis. There is sufficient evidence to support that there exists a linear relationship between the year and death.

- r = 0.001133651

The correlation efficient is far from 1. So, this is positve linear relationship because the value is postive and it is weak linear relationship because its range from 0.0 to 0.5 .

REGGRESSION

AGE ADJUSTING RATES(AADR) AGAINST DEATHS



Calculations using R

```
Residuals:
    Min      1Q  Median      3Q     Max
-712.86  -74.41  -64.27  -37.40 1002.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.207e+01  1.639e+00   50.08   <2e-16 ***
DEATHS      5.247e-04  1.820e-05   28.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 186.8 on 13153 degrees of freedom
  (1035420 observations deleted due to missingness)
Multiple R-squared:  0.05946,   Adjusted R-squared:  0.05939
F-statistic: 831.5 on 1 and 13153 DF,  p-value: < 2.2e-16

> mod$coefficients
 (Intercept)       DEATHS
8.206962e+01 5.247189e-04
```

From the image above, the linear equation :

$$\hat{y} = 8.206962\text{x}10\text{^}1 + 5.247189\text{x}10\text{^} - 4\boldsymbol{x}$$

From above equation, we can say that, $b_0 = 8.206962\text{x}10\text{^}1$ is the death that is unexplained by AADR. The slope coefficient tell us that $\mathbf{b_1} = 5.247189\text{x}10\text{^} - 4$, the average AADR increase on death.

CHI-SQUARE TEST OF INDEPENDENCE

RELATIONSHIP BETWEEN GENDER AND DEATH

H_0 : There is no relationship between Gender and Death.

H_1 : There is relationship between Gender and Death.

CALCULATION USING R

i. Two – way contigency table: -

```
> table(Gender,DEATHS)
        DEATHS
Gender   10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
        DEATHS
Gender   49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87
        DEATHS
Gender   88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
        DEATHS
Gender   120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148
```

ii. Test statistics

```
        Pearson's Chi-squared test

data:  data
X-squared = 5288.4, df = 5312, p-value = 0.5882

>
```

From above calculation,

- $X^2 = 5288.4$
- Degree of freedom = 5312
- $X^2_{5312,0.05} =$ invalid

Discussion

In progress of doing this project, I have obatained and learn new things. First of all , I learn on how to done the hypothesis testing , correlation, reggression, and chi-square test. The hypothesis testing shown that there is no different from for the average death data that I have obtain. There is just sligthly different from other.

For the correlation, I do the relationship between year and death. This is to study that whether the year is affecting the death. And it is proven that it it related to each other. So, we can see that there is slighty different in scatterplot in correlation.

For regression, we want to understand if the average death can affect the age adjusted rates(AADR). So, from the analysis and calcualtion above we can says that AADR can increase depend on average death.

Lastly, for chi-square test of independence is based on gender and deaths. For this one, I don't get the desire analysis on it because the degree of freedom is too high. So, we cannot conclude whether to support the H_0 or H_1 . The two ways contigency table is not the table that we wanted and we cannot find solution to it.

Notes:

Age Adjusted Rates (AADR) is a way to make fairer comparisons between groups with different age distributions. For example, a county having a higher percentage of elderly people may have a higher rate of death or hospitalization than a county with a younger population, merely because the elderly are more likely to die or be hospitalized. (The same distortion can happen when comparing races, genders, or time periods.) Age adjustment can make the different groups more comparable.
A "standard" population distribution is used to adjust death and hospitalization rates. The age-adjusted rates are rates that would have existed if the population under study had the same age distribution as the "standard" population. Therefore, they are summary measures adjusted for differences in age distributions.

CONCLUSION

This study teach me on how we can clearly read and analyse the dataset. It has make me more understanding on reading the dataset. Moreover , I can see that my skill on using R-studio has increasing. Before, I can only doing basic thing like barplot, pie cahrt and scatter plot. Now I can even do the calculation for the hypothesis sample , correlation, reggression and also chi-square test. In the near future, maybe by doing this project it can help me strive for better understanding in doing analysis and hypothesis.

Not to forget that we sould always stay healthy in hope that the death average will decreased. If we got any disease or injury , do take fast action by going to hospital and get treatment.