School of Computing, Faculty of Engineering
University Technology Malaysia

SECI 2143 PROBABILITY AND STATISTICAL DATA ANALYSIS

Project II

Name : Chua Kek An (A19EC0039)

Section : 04

Data Set : Framingham Heart Study

TABLE OF CONTENT

Introduction

In order to carry out this project, a secondary dataset is obtained from the Kaggle website (https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset). This data was collected by an individual with the name of Aman Ajmera on the topic of the heart study in Framingham. This data had helped to predict the most relevant factors of heart disease. World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. Thus, this data is considered important as many people who are suffering from heart disease can be saved if the risk of heart disease is predicted beforehand.

From this dataset, I would like to conduct a hypothesis testing on the mean of the systolic blood pressure. The purpose of this testing is to verify whether the people in Framingham has normal systolic blood pressure or not. Next, the study on the correlation between age and systolic blood pressure and regression between age and diastolic blood pressure can also be carried out. Furthermore, the existence of relationship between gender and smoking status can also be tested.

Result:

Hypothesis testing (1-Sample)

In this project, hypothesis testing on 1 sample is conduct on the mean of systolic blood pressure from the dataset. In this testing, we want to test that whether the population mean for systolic blood pressure is on the normal reading or not. For your information, the systolic blood pressure for healthy human being is between 90mm/Hg to 120mm/Hg.

The null hypothesis, H_0 and alternative hypothesis, H_1 are formed in order to carry out this testing.

H_0: $\mu = 120$

H_1: $\mu > 120$

Where μ is the population mean of the systolic blood pressure. This hypothesis statement will be tested by using significance level of 5% where $\alpha = 0.05$.

As the population variances are unknown and the size of sample is relatively large ($n = 50$), this sample is assumed to be normally distributed. Thus, Z value will be used in this testing to determine which hypothesis statement will be accepted.

```
> n = 50
> s = sd(data$sysBP)
> xbar = mean(data$sysBP)
> mu = 120
> alpha = 0.05
> z=(xbar-mu)/(s/sqrt(n))
> z
[1] 4.712821
> z.alpha = qnorm((1-alpha))
> z.alpha
[1] 1.644854
>
```

As we can see, the test statistics, Z-value that we got from Rstudio is 4.712821 while the critical value is 1.644854. For left-tailed test, if test statistic is greater than critical value, the null hypothesis, H_0 is rejected. Thus, since the test statistic (4.712821) is greater than critical value (1.644854), the null hypothesis is rejected at 5% significance level. There is sufficient evidence at 5% significance level to conclude that the population mean of systolic blood pressure is greater than 120 mm/Hg.

Correlation

In this project, correlation test is conduct on the age and systolic blood pressure variables in order to investigate whether there exist to be a linear relationship between these two variables at 5% significance level.

H_0: $\rho = 0$

H_1: $\rho \neq 0$

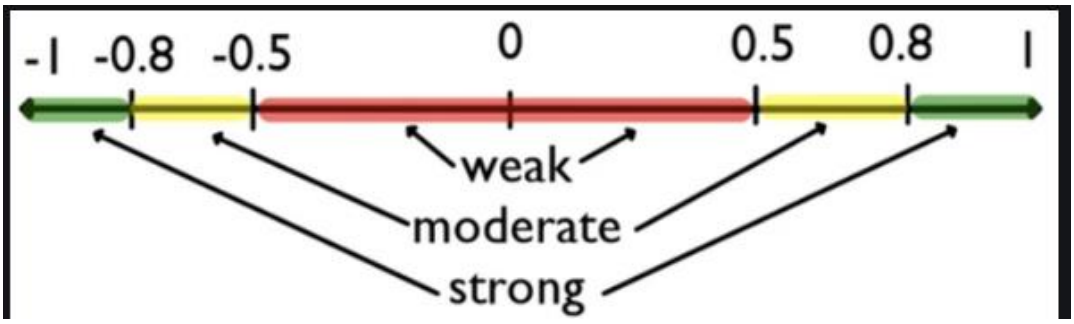
Where H_0 means there is no linear correlation between age and systolic blood pressure while H_1 means there is a linear correlation between age and systolic blood pressure.

```
> plot(data$age, data$sysBP, xlab = "Age", ylab = "Systolic Blood Pressure")
> z = cor.test(data$age, data$sysBP, method = "pearson")
> z

        Pearson's product-moment correlation

data:  data$age and data$sysBP
t = 3.1766, df = 48, p-value = 0.002605
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1566013 0.6228701
sample estimates:
      cor
0.4167817

> |
```

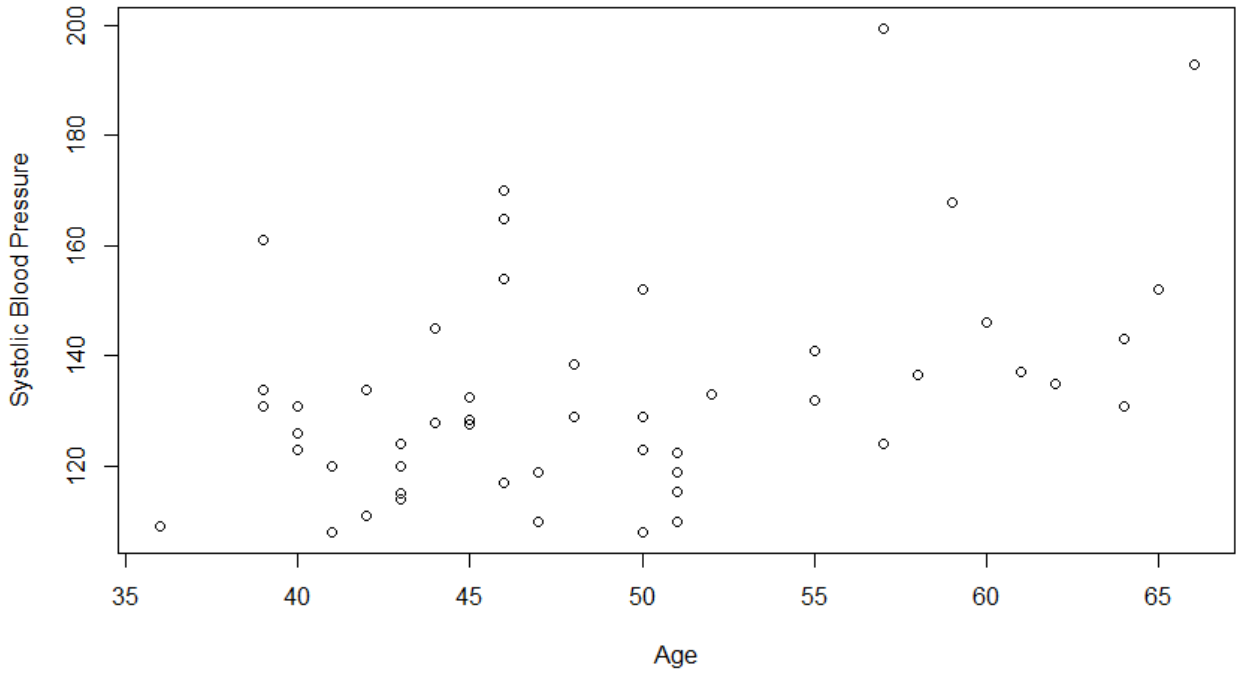
From above figure, the Pearson's correlation coefficient, r which is calculated by using Rstudio is equal to 0.4167817. Next, the test statistic calculated is equal to 3.1766 and the p-value is 0.002605. On the other hand, the critical value at 5% significance level with degree of freedom of 48 (50-2) for two-tailed test is equal to ±2.011. If the test statistic is greater than 2.011 or less than -2.011, the null hypothesis is rejected.

Thus, since the test statistic (3.1766) is greater than critical value (2.011), the null hypothesis is rejected. There is sufficient evidence at 5 % significance level to conclude that there is a linear correlation relationship between age and systolic blood pressure.

Based on the value of Pearson's correlation coefficient, r, we can conclude that the age and systolic blood pressure variable have a weak positive linear correlation.



Graph:



Regression

In this project, regression test is conduct on the age and diastolic blood pressure variables in order to determine if there is a linear relationship between these two variables at 5% significance level.

Independent variable (x): Age

Dependent variable (y): Diastolic blood pressure

H_0: $\beta_1 = 0$

$H_{1:}$ $\beta_1 \neq 0$

Where H_0 means there is no linear relationship while H_1 means linear relationship does exist between age (independent variable) and diastolic blood pressure (dependent variable).

```
> summary(model)

Call:
lm(formula = data$diaBP ~ data$age)

Residuals:
    Min      1Q  Median      3Q     Max
-17.137  -5.961  -1.608   6.084  24.275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.1982     8.8939   7.780 4.74e-10 ***
data$age      0.2941     0.1794   1.639    0.108
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.934 on 48 degrees of freedom
Multiple R-squared:  0.05298,   Adjusted R-squared:  0.03325
F-statistic: 2.685 on 1 and 48 DF,  p-value: 0.1078
```

From the calculation of Rstudio, we can form an estimated regression model equation.

$$\hat{y} = 69.1982 + 0.2941\ x$$

From the equation, we can know that the diastolic blood pressure increases by 0.2941 for every increase in age by one.

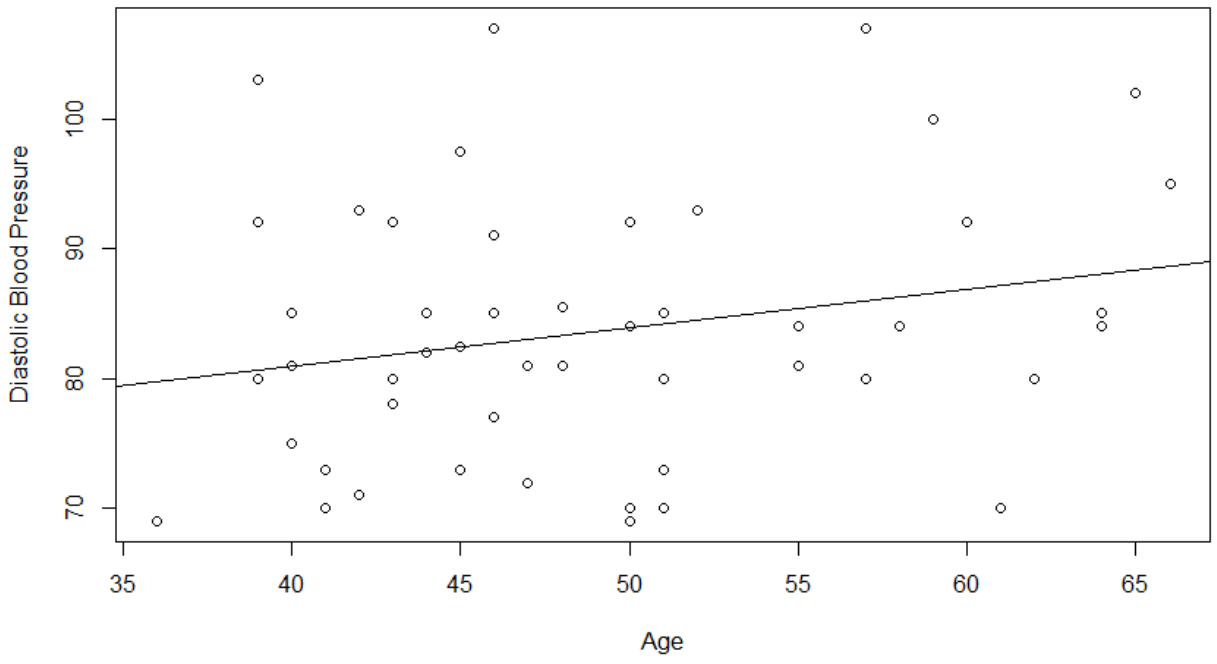
From Rstudio, we can also identify the coefficient of determination.

$$\mathbf{R^2 = 0.05298}$$

Based on the value of coefficient of determination that we get from Rstudio, we can know that 5.298 % of the variation in diastolic blood pressure is explained by variation in age.

From the calculations in Rstudio, we can know the value for the p-value for the regression testing. Since this test is carry out at 5% significance level and p-value is equal to 0.1078 which is greater than 0.05, the null hypothesis is failed to be rejected. Thus, there is sufficient evidence at 5 % significance level to conclude that there is no linear relationship between age and diastolic blood pressure which means that the diastolic blood pressure is not dependent on the age. The value of Pearson's correlation coefficient, r = 0.2302 which is the square root of the coefficient of determination also shows that there is only weak relationship between these two variables.

Graph:



Chi-square Test of Independence

In this project, chi-square independence test is conduct on the gender and smoking status variables in order to determine whether these two variables are dependent on each other or not at 5 % significance level.

H_0: Smoking status is independent of gender.

$H_{1:}$ Smoking status is not independent of gender.

```
> x <- table(data$male, data$currentSmoker)
> chisq.test(x)

        Pearson's Chi-squared test with Yates' continuity correction

data:  x
X-squared = 0.45197, df = 1, p-value = 0.5014

> alpha = 0.05
> x.alpha = qchisq(1-alpha, df = 1)
> x.alpha
[1] 3.841459
>
```

From the calculation of Rstudio, we can know that the test statistic, χ^2 is equal to 0.45197 and the p-value of it is equal to 0.5014. The degree of freedom is also calculated based on the formula:

Degree of freedom = (row - 1)(column - 1)

Gender / Smoker	Yes	No
Male	11	7
Female	15	17

From the table, as there are two rows and two columns, the degree of freedom is equal to 1 based on the formula given. If test statistic is greater than critical value, null hypothesis is rejected as chi-square test is always right-tailed.

At 5% significance level, the critical value with degree of freedom of 1 is equal to 3.841459. Thus, since test statistic (0.45197) is less than critical value (3.841459), null hypothesis is failed to reject. There is sufficient evidence at 5% significance level to conclude that the smoking status is independent of gender.

Discussion and Conclusion

In conclusion, we have learnt about the relationships among the variables of the dataset by conducting various types of hypothesis testing. First of all, the hypothesis testing on 1-sample which has been carried out on the mean of systolic blood pressure at 5 % significance level proves that the population mean of systolic blood pressure of the Framingham's people exceed 120 mm/Hg which is not on a healthy level. Thus, the people in Framingham must be informed to perform healthy lifestyle so that the risk of coronary heart diseases can be decreased.

Next, the correlation test conducted has proved that there is a positive linear relationship between the age and systolic blood pressure, which means as the age increases, the systolic blood pressure of that individual will also increase. Thus, although it is only a weak relationship exists between these two variables, we must always get medical check up when we are older so that we will not suddenly collapse due to high blood pressure. However, when the regression test is conducted on age and diastolic blood pressure, it is proved that the diastolic blood pressure is not dependent on the age. The result of this test is not exactly the same as what I have expected. I assumed that the systolic blood pressure and diastolic blood pressure will be increased when the age increases.

Lastly, the chi-square test of independence has proved that smoking status is independent of the gender. Thus, the stereotype about the smokers are mostly man is proved wrong from the result of this testing.