UNIVERSITI TEKNOLOGI MALAYSIA

JOHOR BAHRU (UTM JB),

81300 JOHOR BAHRU,

JOHOR

SCHOOL OF COMPUTING

FACULTY OF ENGINEERING

PROJECT 2

REPORT TITLE:
Risk Factors of Coronary Heart Disease

PROBABILITY AND STATISTICAL DATA ANALYSIS
SECI2143 – Section 06

NAME	**LIM SIN JIE**
MATRIC NO	**A19EC0073**
LECTURER NAME	**DR. CHAN WENG HOWE**

TABLE OF CONTENTS

1.0 Introduction

The dataset in this study was collected by researchers in the National Heart, Lung and Blood Institute that launched the project, Framingham Heart Study. It is a long-term, ongoing study of risk factors of cardiovascular disease of residents of Framingham, Massachusetts. The findings in this project had articulated the needs for preventing, detecting, and treating risk factors of cardiovascular diseases. The dataset has 4240 observations and 16 variables. It was collected to identify the risk factors that increase the chances of having coronary heart disease which is a type of heart disease that has been the leading cause of death worldwide. The term "risk factors" was actually coined by William Kannell and Roy Dawber from the Framingham Heart Study.

For the dataset, several demographic risk factors are included such as the gender of the patients, the age of the patients, the education level coded as either 1 for some high school, 2 for a high school diploma or GED, 3 for some college or vocational school, and 4 for a college degree. Not only that, the dataset also includes behavioural risk factors associated with smoking in which whether the patient is a current smoker and the number of cigarettes smoked by the person on average in one day. Medical history risk factors are also included, which are whether the patient was on blood pressure medication, whether the patient had previously had a stroke, whether the patient has hypertension, and whether the patient has diabetes. Also, the dataset includes risk factors from the first physical examination of the patient such as the total cholesterol level, systolic blood pressure, diastolic blood pressure, Body Mass Index (BMI), heart rate, and blood glucose level of the patients that were measured. Lastly, the ten-year risk coronary heart disease (CHD) is included.

According to World Health Organization, cardiovascular diseases are the top cause of death globally, taking an estimated 17.9 million lives each year. Coronary heart disease (CHD) is a type of heart disease that develops when the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. Thus, the purpose of this study is to investigate the risk factors that would increase the risk of having coronary heart disease and hence increase public awareness on practising a heart-healthy lifestyle. Since this issue should be placed on the top of agenda, therefore identifying risk factors is one of the crucial actions to reduce population's exposure to those risk factors and help them to prevent heart disease.

Furthermore, it can also help to identify high-risk population and ensure them receive the appropriate treatments. Hence, an individual's risk for cardiovascular disease can be greatly reduced, by changing lifestyle and having preventive treatment. For example, one can change his or her lifestyle by quitting smoking, having healthy diet, doing regular exercise and so on. Preventive medical treatment can include a statin, mini dose aspirin or treatment for hypertension. Therefore, it is vital to be able to predict the risk of an individual patient to coronary heart disease, in order to decide when to initiate lifestyle modifications and have preventive medical treatment.

2.0 Inferential Statistics Analysis

1) Hypothesis Testing - One Sample Test

a) Since obesity is a one of the risk factors which leads to coronary heart disease, hence we will conduct a one sample hypothesis testing on proportion of the coronary heart disease patients who are overweight or obese. BMI is considered as a measure of obesity and a BMI of 25 to 29.9 is considered overweight while BMI of 30 and above is considered obese. According to a study which is "Obesity in Coronary Heart Disease: An Unaddressed Behavioural Risk Factor", over 80% of the patients with coronary heart disease are overweight or obese. Hence, we will claim that the likelihood of overweight or obesity of the patients in this project is different from p = 0.80 using a sample size of 50 and assume that the significance level, α = 0.05.

First, we will state the null hypothesis, H_0 and the alternative hypothesis, H_1.

Hypothesis statement:

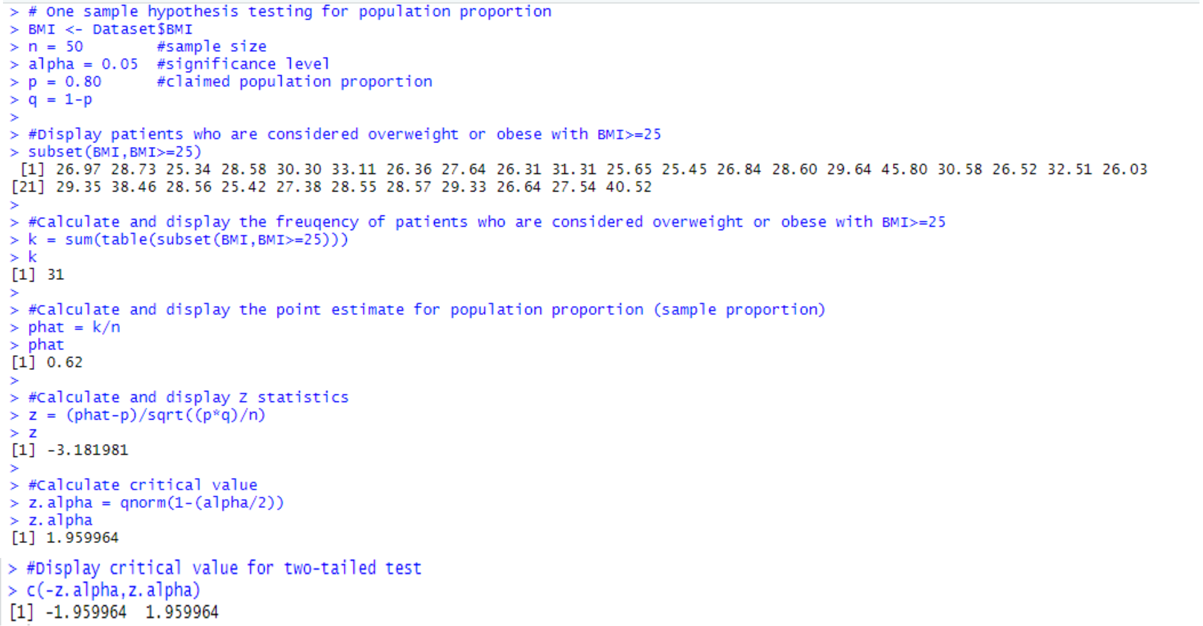
$H_0 : p = 0.80$

$H_1 : p \neq 0.80$

Then, we will calculate the test statistic for proportion by using the formula:

$$z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$$

$\hat{p}$ = point estimate of population proportion (sample proportion)
p = population proportion (claimed value)
q= 1 – p
n = sample size



```
> # One sample hypothesis testing for population proportion
> BMI <- Dataset$BMI
> n = 50         #sample size
> alpha = 0.05  #significance level
> p = 0.80       #claimed population proportion
> q = 1-p
>
> #Display patients who are considered overweight or obese with BMI>=25
> subset(BMI,BMI>=25)
 [1] 26.97 28.73 25.34 28.58 30.30 33.11 26.36 27.64 26.31 31.31 25.65 25.45 26.84 28.60 29.64 45.80 30.58 26.52 32.51 26.03
[21] 29.35 38.46 28.56 25.42 27.38 28.55 28.57 29.33 26.64 27.54 40.52
>
> #Calculate and display the freuqency of patients who are considered overweight or obese with BMI>=25
> k = sum(table(subset(BMI,BMI>=25)))
> k
[1] 31
>
> #Calculate and display the point estimate for population proportion (sample proportion)
> phat = k/n
> phat
[1] 0.62
>
> #Calculate and display z statistics
> z = (phat-p)/sqrt((p*q)/n)
> z
[1] -3.181981
>
> #Calculate critical value
> z.alpha = qnorm(1-(alpha/2))
> z.alpha
[1] 1.959964
> #Display critical value for two-tailed test
> c(-z.alpha,z.alpha)
[1] -1.959964  1.959964
```

>Figure 1

The code in Figure 1 shows how the one sample hypothesis testing for population proportion is conducted in Rstudio. The point estimate which is the sample proportion, $\hat{p}$ is calculated after obtaining frequency of patients who have BMI equal to or more than 25 kg/m^2 which is 31. Thus, $\hat{p}$ = 31/50 = 0.62 since the sample size, n is 50. Then, the test statistic for proportion is calculated, z = -3.181981. Also, since this is a two-tailed test, we will use α/2 = 0.025 to obtain our critical values which are $-z_{0.025} = -1.959964$ and $z_{0.025} = 1.959964$. Please be noted that we will reject H_0 if $z < -z_{0.025} = -1.959964$ or $z > z_{0.025} = 1.959964$.

Therefore, since $z = -3.181981 < z_{0.025} = -1.959964$, hence we reject H_0 at the significance level of 0.05. There is sufficient evidence to support the claim that the likelihood of overweight or obesity of the patients in this project is different from p = 0.80.

b) Since age is one of the risk factors that contribute to the development of coronary heart disease and there is a study that states that the elderly who are more than 65 years old are more likely to develop coronary heart disease. Thus, we will test whether if there is sufficient evidence to support the claim that the population of patients with coronary heart disease age more than 65 years old using significance level, α of 0.05. A simple sample of 50 data is used in the hypothesis testing and the population variance is unknown. It can be said that a one sample hypothesis testing for population mean is conducted with unknown variance and sample size which is more than 30.

After that, the sample mean and sample standard deviation can be calculated using the formula:

$$Sample\ mean, \bar{x} = \sum X / n$$

$$Standard\ deviation, \sigma = \frac{\sqrt{(\sum(X-\bar{x})^2)}}{(n-1)}$$

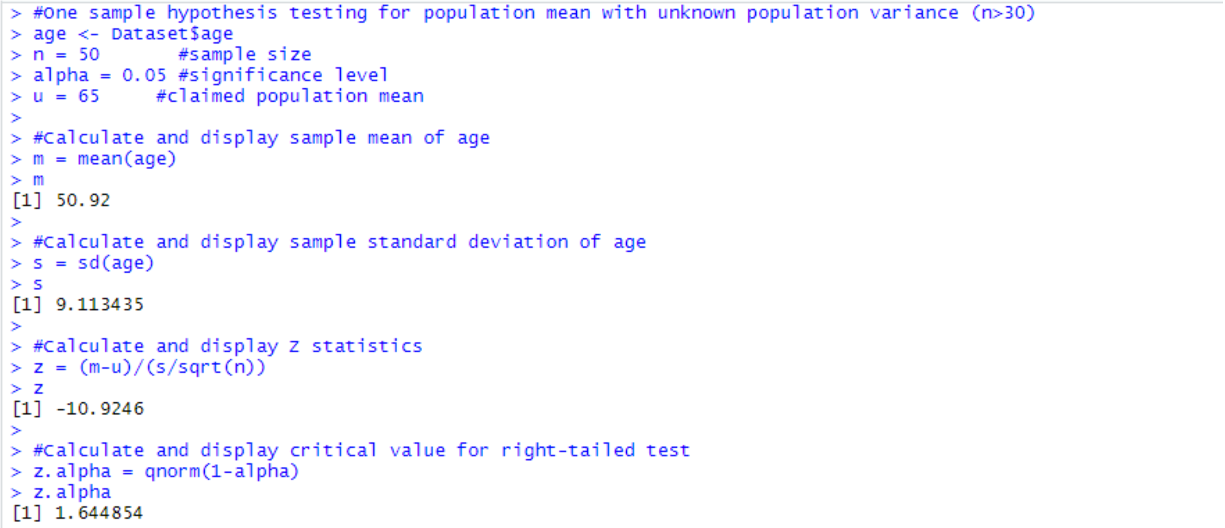
This hypothesis testing is conducted to test whether we reject the null hypothesis, H_0 or not and whether there is sufficient evidence to support the null hypothesis by calculating the test statistic for population mean, z_0 and comparing it to the critical value of the significance level which is 0.05. If the test statistic calculated, z-statistic falls in the critical region, the null hypothesis, H_0 will be rejected, otherwise fail to reject H_0. The z-statistic in the hypothesis testing for population mean with unknown variance and sample size, n which is more than 30 can be calculated by using the formula:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

In this case, hypothesis statement:

$H_0 : \mu = 65$

$H_1 : \mu > 65$



```
> #One sample hypothesis testing for population mean with unknown population variance (n>30)
> age <- Dataset$age
> n = 50        #sample size
> alpha = 0.05 #significance level
> u = 65      #claimed population mean
>
> #Calculate and display sample mean of age
> m = mean(age)
> m
[1] 50.92
>
> #Calculate and display sample standard deviation of age
> s = sd(age)
> s
[1] 9.113435
>
> #Calculate and display z statistics
> z = (m-u)/(s/sqrt(n))
> z
[1] -10.9246
>
> #Calculate and display critical value for right-tailed test
> z.alpha = qnorm(1-alpha)
> z.alpha
[1] 1.644854
```

>Figure 2

The above code in figure 2 shows that how the one sample hypothesis test for population mean is conducted using Rstudio. Also, the sample mean of age is 50.92 and the sample standard deviation of age is 9.113435. Next, the test statistic which is z-statistics is calculated using the formula above and z = -10.9246. The critical value for the significance level of 0.05 is $z_{0.05}$ = 1.644854. For the right-tailed test, we wil only reject H_0 if the z > $z_{0.05}$ = 1.644854.

Since z = -10.9246 < $z_{0.05}$ = 1.644854, hence we fail to reject the null hypothesis, H_0 at the significance level of 0.05. Thus, there is insufficient evidence to support the claim that the population of patients with coronary heart disease age more than 65 years old.

2) Correlation

Correlation is a measure of statistical relationship between two comparable variables or quantities. Also, correlation analysis is used to measure strength of association between two variables which is only concerned with the strength of the relationship. A scatter plot will be used to show the relationship between two variables as well. Since smoking habit is one of the risk factors for coronary heart disease, hence we aim to analyse the strength and relationship between the independent variable chosen which is the number of cigarettes smoked per day and another dependent variable which is heart rate in the correlation test. The sample size that we use is 50.

In this correlation test, we will use Pearson product-moment correlation coefficient that measures the strength of the linear relationship between the two variables because the type of data is ratio. The sample correlation coefficient will be calculated in Rstudio based on the formula below:

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

where: r= Sample correlation coefficient
n= Sample size
x= Value of the independent variable
y= Value of the dependent variable

Significance Test for Correlation

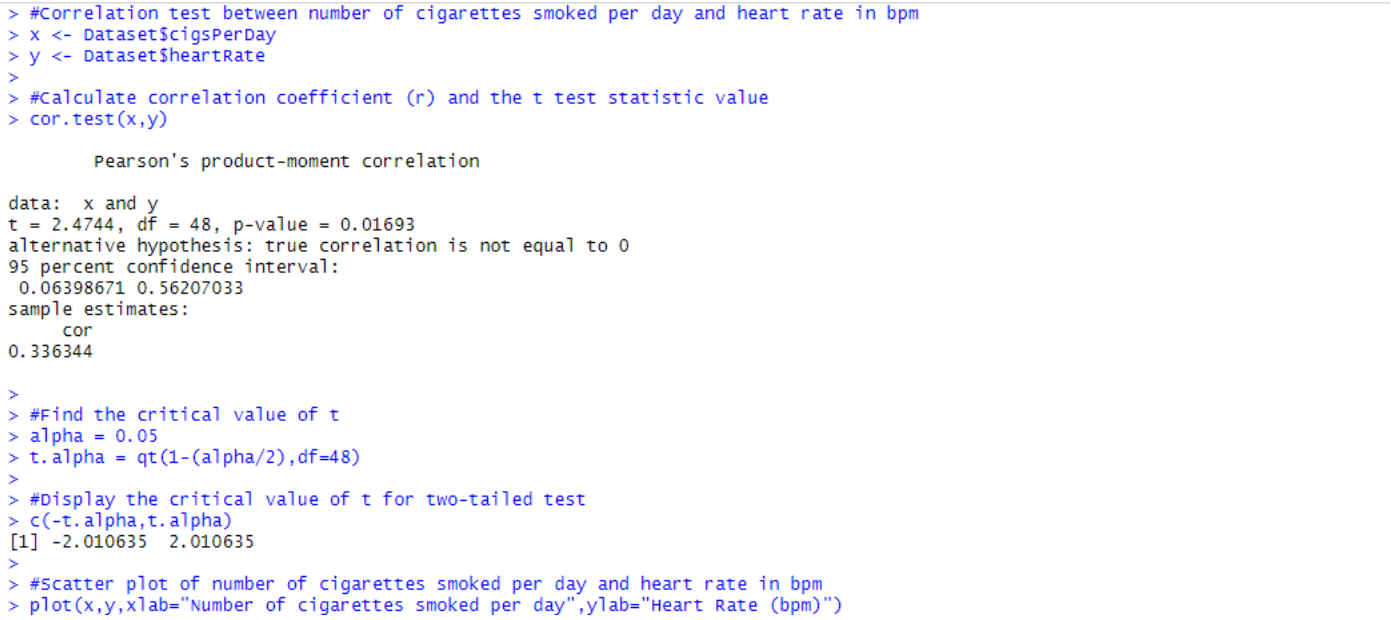
Moreover, the significance test for correlation analysis will be conducted. There will be the significance test for the two variables to investigate whether there is an evidence of the linear relationship between them at significance level, α selected. In this project, we will assume that the significance level, α is 0.05. Hence, the significance test will be carried out where H_0 is assumed to be that the two variables have no linear correlation and H_1 is assumed to be that the two variables tested have linear correlation.

After that, the critical value of t with n-2 degrees of freedom from t distribution table will be determined. The test statistic will be calculated using the formula below to test whether the null hypothesis, H_0 will be rejected or not and to conclude that whether there is sufficient evidence to prove that there is linear correlation between the two variables. The formula for the test statistic, t is shown below:

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

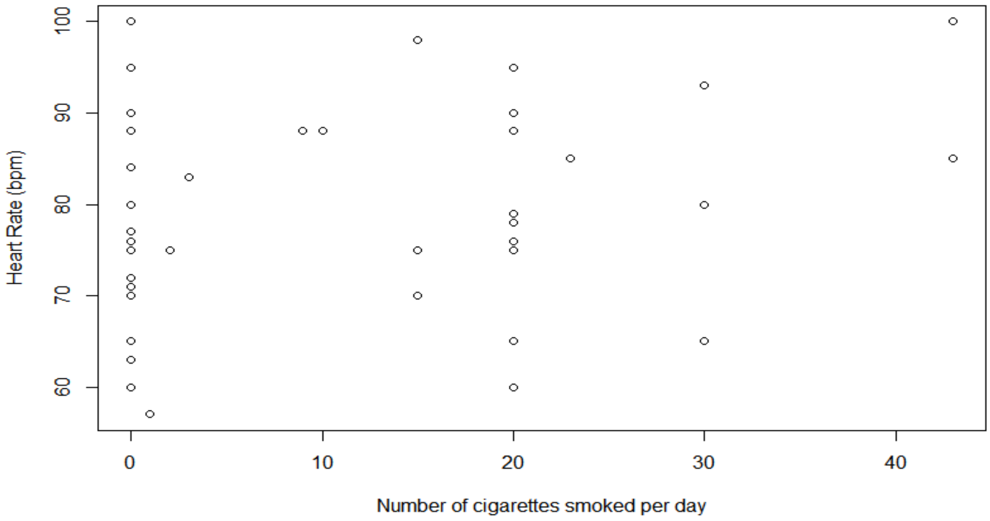
Correlation between number of cigarettes smoked per day and heart rate in bpm

In the correlation test, we will investigate the strength of linear relationship between the independent variable which is number of cigarettes smoked per day and the dependent variable which is heart rate in bpm. Hence, we will calculate the correlation coefficient, r to know the strength of the linear relationship between them based on the formula mentioned before and the scatter plot will also be shown as an evidence of their relationship. In the meanwhile, the significance test for correlation is also conducted to show that whether there is an evidence of a linear relationship between the independent variable which is number of cigarettes smoked per day and the dependent variable which is heart rate in bpm at the significance level, α = 0.05. The sample size, n is 50.



```
> #Correlation test between number of cigarettes smoked per day and heart rate in bpm
> x <- Dataset$cigsPerDay
> y <- Dataset$heartRate
>
> #Calculate correlation coefficient (r) and the t test statistic value
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 2.4744, df = 48, p-value = 0.01693
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06398671 0.56207033
sample estimates:
     cor
0.336344

>
> #Find the critical value of t
> alpha = 0.05
> t.alpha = qt(1-(alpha/2),df=48)
>
> #Display the critical value of t for two-tailed test
> c(-t.alpha,t.alpha)
[1] -2.010635  2.010635
>
> #Scatter plot of number of cigarettes smoked per day and heart rate in bpm
> plot(x,y,xlab="Number of cigarettes smoked per day",ylab="Heart Rate (bpm)")
```

>Figure 3



>Figure 4

The code in Figure 3 has shown the result of correlation test in Rstudio and Figure 4 has shown the scatterplot of number of cigarettes smoked per day and heart rate in bpm. Hence, the correlation coefficient, r value is calculated in Rstudio using cor.test() function with the default method which is Pearson's technique. From the result, we can see that the scatterplot and the correlation coefficient, r = 0.336344 which indicates that there is relatively weak positive linear relationship between number of cigarettes smoked per day and heart rate.

For the significance test for correlation between number of cigarettes smoked per day and heart rate in bpm, we will state our hypothesis statement as below:

H_0 : p = 0 (No linear correlation between number of cigarettes smoked per day and heart rate)

H_1 : p ≠ 0 (Linear correlation exists between number of cigarettes smoked per day and heart rate)

In the significance test for correlation, the test statistic of t-value is also calculated based on the formula shown before using cor.test() function in RStudio which is shown in Figure 3, thus we obtain that t = 2.4744. Not only that, we also obtain that the degrees of freedom, df = 50-2 = 48 which is shown in Figure 3. In addition, since it is a two-tailed test, $\alpha/2 = 0.025$, hence the critical value of t at the significance level, $\alpha = 0.025$ and df =48 is found out to be $-t_{0.025,48} = -2.010635$ and $t_{0.025,48} = 2.010635$. We will reject H_0 if the $t < -t_{0.025,48} = -2.010635$ or $t > t_{0.025,48} = 2.010635$ at the significance level of $\alpha = 0.05$.

Thus, since $t = 2.4744 > t_{0.025,48} = 2.010635$, we will reject the null hypothesis, H_0 at the significance level of $\alpha = 0.05$. There is sufficient evidence of a linear relationship between number of cigarettes smoked per day and heart rate in bpm.

3) Regression

Regression analysis is used to predict the value of a dependent variable based on the value of at least one independent variable and it is also to explain impact of changes in an independent variable on the dependent variable. We will focus on the use of simple linear regression, which is the regression model that involves only one single independent variable and it will describe the relationship between independent variable and dependent variable as a straight line. Here, we will use the estimated regression model for simple linear regression. The sample regression line with formula below will provides an estimate of the population regression line.

$$\hat{y}_i = b_0 + b_1 x$$

$\hat{y}_i$ = Estimated y value

b_0 = Estimate of the regression intercept

b_1 = Estimate of the regression slope

x = Independent variable

The formula for b_0 *and* b_1 are shown below, which will be calculated in RStudio.

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Also, b_0 is the estimated average value of y when the value of x is zero and b_1 is the estimated change in the average value of y as a result of a one-unit change in x.

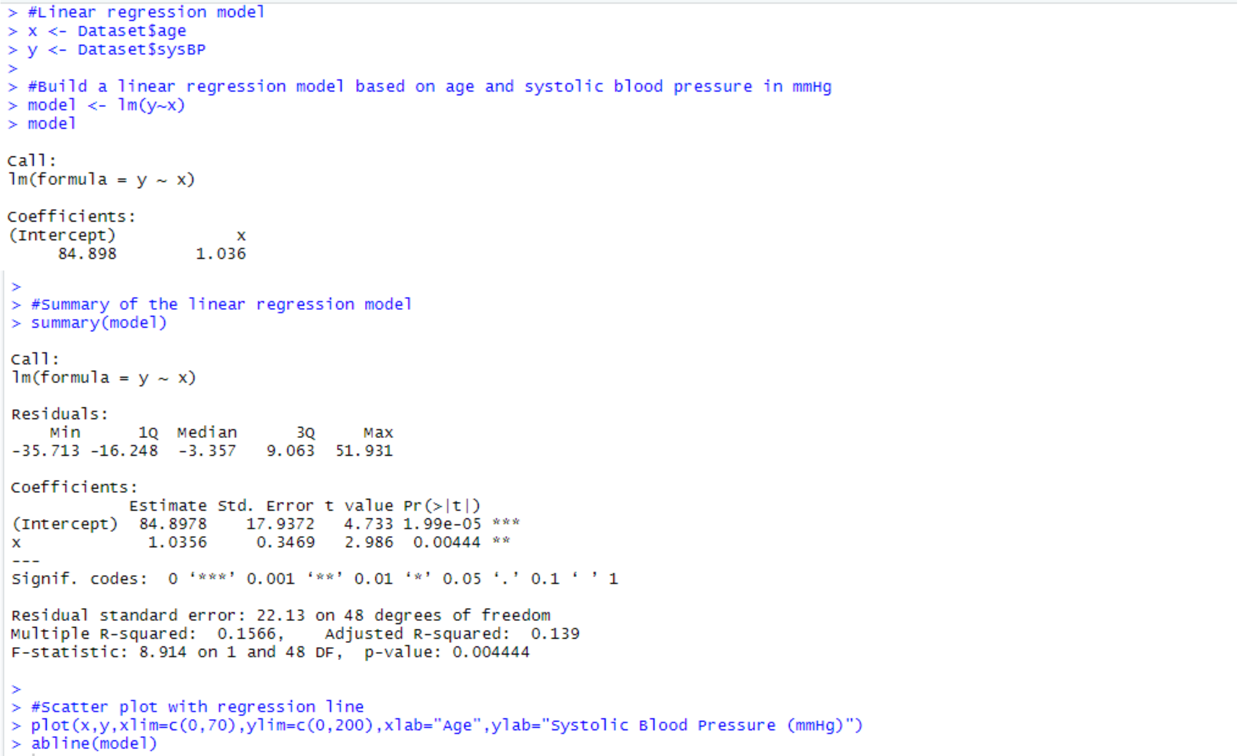
Furthermore, the coefficient of determination, R^2 which is the portion of the total variation in the dependent variable that is explained by variation in the independent variable, will be calculated as well using the formula below:

$$R^2 = \frac{\text{SSR}}{\text{SST}} \qquad \text{where } 0 \leq R^2 \leq 1$$

SSR, regression sum of squares = $\sum(\hat{y} - \bar{y})^2$

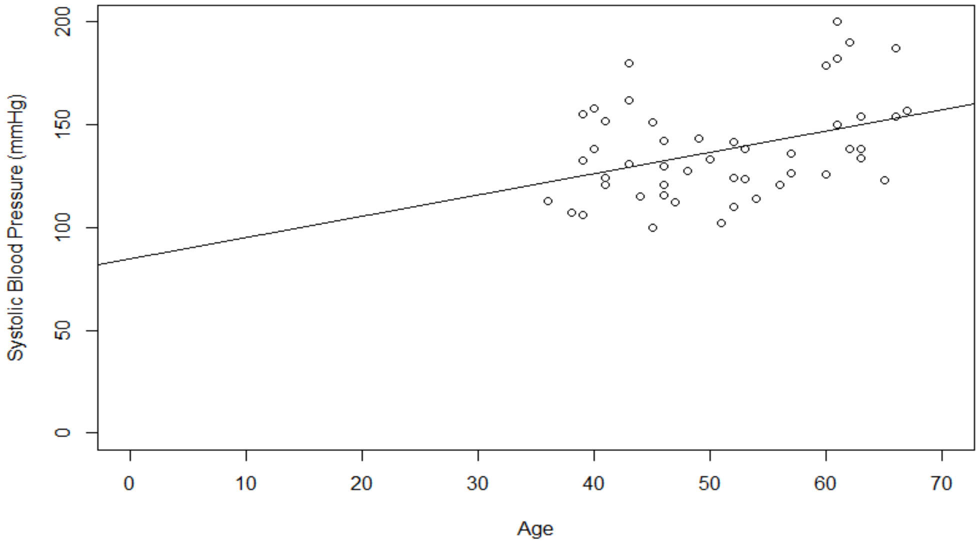
SST, total sum of squares = $\sum(y_i - \bar{y})^2$

In our project, we will build an estimated regression model for age and systolic blood pressure in mmHg to investigate the relationship between them. Hence, the regression analysis can be conducted to predict the value of systolic blood pressure in mmHg based on the value of age and we can also explain the impact of changes in age on systolic blood pressure in mmHg. The independent variable, x is age and the dependent variable, y is systolic blood pressure in mmHg. The sample size is 50.



```
> #Linear regression model
> x <- Dataset$age
> y <- Dataset$sysBP
>
> #Build a linear regression model based on age and systolic blood pressure in mmHg
> model <- lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     84.898        1.036

>
> #Summary of the linear regression model
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-35.713 -16.248  -3.357   9.063  51.931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.8978    17.9372   4.733 1.99e-05 ***
x             1.0356     0.3469   2.986  0.00444 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.13 on 48 degrees of freedom
Multiple R-squared:  0.1566,	Adjusted R-squared:  0.139
F-statistic: 8.914 on 1 and 48 DF,  p-value: 0.004444

>
> #Scatter plot with regression line
> plot(x,y,xlim=c(0,70),ylim=c(0,200),xlab="Age",ylab="Systolic Blood Pressure (mmHg)")
> abline(model)
```

> Figure 5



>Figure 6: Scatter plot of age and systolic blood pressure in mmHg

The code in Figure 5 has shown that how we build an estimated regression model based on age and systolic blood pressure in mmHg using RStudio and hence we obtain the estimated regression equation for our sample which is shown below:

$$\hat{y} = 84.898 + 1.036x$$

Also, we have the regression equation in the graphical presentation with the scatter plot and regression line which is shown in Figure 6.

Since we have obtained that the estimated regression equation, thus from the equation, we can know that the estimate of regression intercept, b_0 = 84.898 and the estimate of regression slope, b_1 = 1.036. Also, we can interpret the regression intercept, b_0 and slope, b_1 using the values obtained. Since the estimate of regression intercept, b_0 = 84.898, it indicates that systolic blood pressure is 84.898 mmHg when the age of the patients is 0 years old. Also, b_1 = 1.036 tells us that the average value of systolic blood pressure increases by 0.3609 mmHg, on average, for each additional 1 year old of age.

Moreover, Figure 5 also shows the value of coefficient of determination, R^2 calculated in RStudio in which $R^2 = 0.1566$. Since $0 < R^2 < 1$, this indicates that there is weaker linear relationship between age and systolic blood pressure in mmHg and there is some but not all of the variation in systolic blood pressure in mmHg is explained the age. From the value of coefficient of determination, R^2, we can say that only 15.66% of the variation in systolic blood pressure in mmHg is explained by the variation in the age of patients.

Regression slope test: t test

Furthermore, we will have a t-test for a population slope as well to determine that whether there is a linear relationship between x which is the age and y which is systolic blood pressure in mmHg. The hypothesis statement is stated below:

H_0 : $\beta_1 = 0$ (No linear relationship between age and systolic blood pressure in mmHg)

H_1 : $\beta_1 \neq 0$ (Linear relationship exists between age and systolic blood pressure in mmHg)

Also, the test statistic of t-value can be calculated using the formula as shown below:

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

where:
b_1= Sample regression slope coefficient
β_1 = Hypothesized slope
s_{b_1} = Estimator of the standard error of the slope

Hence, we have obtained the value for the test statistic in RStudio which is shown in Figure 5 and the test statistic is t = 2.986. We also know that the degrees of freedom, df = 50-2 = 48 using the summary of the estimated regression model. Also, we have already obtained the value for the critical value of t at the significance level of α = 0.05 before. Since it is a two-tailed test, α/2 = 0.025, hence the critical value of t for α = 0.025 and df =48 is found out to be $-t_{0.025,48} = -2.010635$ and $t_{0.025,48} = 2.010635$. And we will reject H_0 if the $t < -t_{0.025,48} = -2.010635$ or $t > t_{0.025,48} = 2.010635$ at the significance level of α = 0.05.

Since $t = 2.986 > t_{0.025,48} = 2.010635$, hence we will reject the null hypothesis, H_0 at the significance level of α = 0.05. There is sufficient evidence that the age does affect the systolic blood pressure in mmHg.

4) Goodness-of-fit Test

In general, goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits some claimed distributions. There are two categories for the test, which are equal probabilities and unequal probabilities.

If all the expected frequencies, E are equal, then the formula which is E = n / k is used to calculate the value of expected frequencies, E where n is the sum of all observed frequencies, O and k is the number of categories. However, if all the expected frequencies, E are not equal, then the formula which is E = n * p is used to calculated the value of expected frequencies, E where n is the sum of all observed frequencies, O and p is the probability of the respective category.

Then, the test statistic value, χ^2 will be calculated using the formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

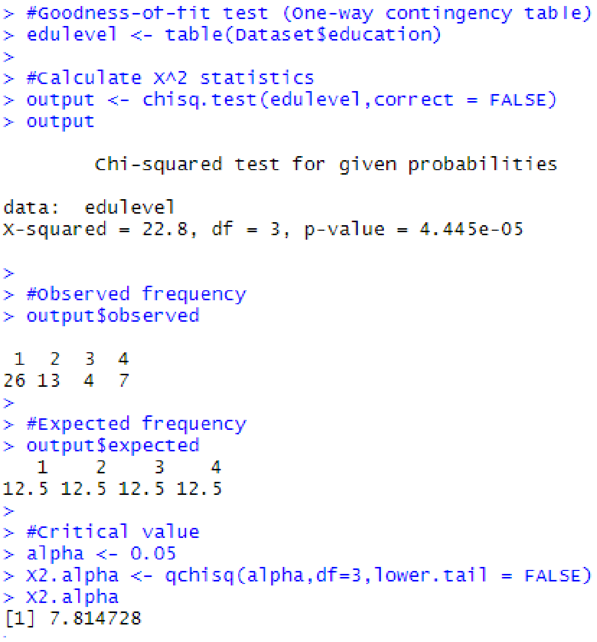
Next, we will find the critical value by referring to the Chi-square value from table by using k-1 degrees of freedom where k is number of categories. Lastly, we will determine whether χ^2 calculated falls within the critical region, if the value falls within critical region, the null hypothesis, H_0 will be rejected, otherwise fail to reject the null hypothesis, H_0.

In our project, we will choose the variable which is education level and it is considered as one of the demographic risk factors that increases the chance of developing coronary heart disease. Goodness-of-fit test is conducted to test whether there are equal proportions for four education levels of patients at the significance level, α = 0.05 and the sample size, n is 50. Also, we will carry out the test in RStudio using certain functions.

Since the claim is that all patients in this project have equal proportions for four education levels, hence we can state our hypothesis statement like below:

$H_0 : p_1 = p_2 = p_3 = p_4$
H_1 : At least 1 of the 4 proportions is different from the others

Next, the code in the next page has shown us that how we can use functions in RStudio to conduct the goodness-of-fit test.



```
> #Goodness-of-fit test (One-way contingency table)
> edulevel <- table(Dataset$education)
>
> #Calculate X^2 statistics
> output <- chisq.test(edulevel,correct = FALSE)
> output

	Chi-squared test for given probabilities

data:  edulevel
X-squared = 22.8, df = 3, p-value = 4.445e-05

>
> #Observed frequency
> output$observed

 1  2  3  4 
26 13  4  7 
>
> #Expected frequency
> output$expected
   1    2    3    4 
12.5 12.5 12.5 12.5 
>
> #Critical value
> alpha <- 0.05
> X2.alpha <- qchisq(alpha,df=3,lower.tail = FALSE)
> X2.alpha
[1] 7.814728
```

>Figure 7

The code in Figure 7 has shown that how we conduct goodness-of-fit test in RStudio. First, we can get the test statistic value, $\chi^2 = 22.8$ based on the formula mentioned before using chisq.test() function in R. The value of degrees of freedom, df = k-1 = 4-1 = 3 is obtained as shown above. Also, we obtain the information of the one-way contingency table by inspecting the observed frequencies and the expected frequencies. For the expected frequencies, it is calculated in RStudio based on the formula which is E = n / k. The one-way contingency table is shown below:

Education Level	1	2	3	4
Observed Frequency, O	26	13	4	7
Expected Frequency, E	12.5	12.5	12.5	12.5

Next, the critical value for the significance level of 0.05 is calculated using the function which is qchisq(alpha,df,lower.tail = FALSE). Hence, the critical value for the test, χ^2 for significance level, α = 0.05 and degrees of freedom, df = 3 is 7.814728.

Therefore, since the $\chi^2 = 22.8 > 7.814728$, that is the test statistics falls within the critical region, thus we reject the null hypothesis, H_0 at the significance level of 0.05. There is sufficient evidence to conclude that there are different proportions of patients for four education levels.

5) Chi-square Test of Independence

Chi-square test of independence is used to test if a relationship exists between two qualitative variables and two-way contingency table is used.

In chi-square test of independence, the expected count is calculated using the formula:

$$e_{ij} = \frac{(i^{th} Row\ total)(j^{th} Column\ total)}{Total\ sample\ size}$$

Also, the test statistic value, χ^2 will be calculated using the formula:

$$\chi^2 = \sum \frac{[o_{ij} - e_{ij}]^2}{e_{ij}}$$

Next, the critical value at the significance level is found by referring to the chi-square table where the degree of freedom, df = (r-1) (c-1). Lastly, if the test statistic, χ^2 > critical value, which falls within critical region, then the null hypothesis, H_0 is rejected and there is evidence that there is relationship between two variables, otherwise if χ^2 < critical value, in which it does not fall within the critical region, then fail to reject the null hypothesis, H_0 and there is insufficient evidence that there is relationship between two variables which means that the no relationship between two variables that they are independent.

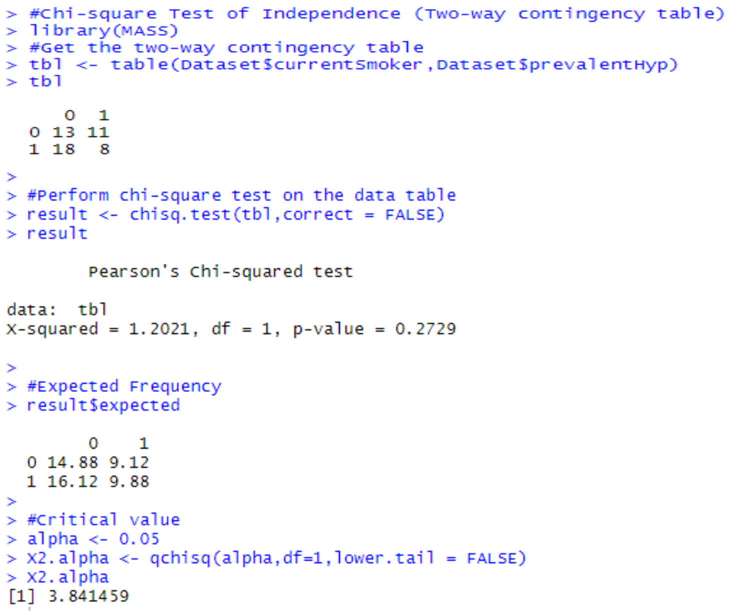
a) In the project, we have 50 data as the sample. The two variables chosen are whether the patients are smoker and whether the patients have hypertension. We need to test if there is evidence of the relationship between them at the significance level of 0.05.

The hypothesis statement is shown below:

H_0 : No relationship between whether the patients are smoker and whether the patients have hypertension.

H_1 : Relationship exists between whether the patients are smoker and whether the patients have hypertension.

We have shown that how do we implement chi-square test of independence using RS tudio in the next page.



```
> #Chi-square Test of Independence (Two-way contingency table)
> library(MASS)
> #Get the two-way contingency table
> tbl <- table(Dataset$currentSmoker,Dataset$prevalentHyp)
> tbl

     0  1
  0 13 11
  1 18  8

> 
> #Perform chi-square test on the data table
> result <- chisq.test(tbl,correct = FALSE)
> result

	Pearson's Chi-squared test

data:  tbl
X-squared = 1.2021, df = 1, p-value = 0.2729

> 
> #Expected Frequency
> result$expected

        0    1
  0 14.88 9.12
  1 16.12 9.88
> 
> #Critical value
> alpha <- 0.05
> X2.alpha <- qchisq(alpha,df=1,lower.tail = FALSE)
> X2.alpha
[1] 3.841459
```

>Figure 8

The code in Figure 8 has shown that the results of chi-square test of independence for whether the patients are smoker and whether the patients have hypertension. The two-way contingency table that is obtained from the result which includes the observed count and the expected count is shown below:

	Hypertension				
	No		Yes		
Smoker	Obs.	Exp.	Obs.	Exp.	Total
No	13	14.88	11	9.12	24
Yes	18	16.12	8	9.88	26
Total	31	31	19	19	50

Not only that, the test statistic value, $\chi^2 = 1.2021$ is obtained using the function chisq.test() in RStudio based on the formula mentioned before. The degree of freedom, df = 1 is obtained as well. if based on the formula, df = (r-1) (c-1) = (2-1) (2-1) = 1. Also, the critical value for the significance level of 0.05 is calculated using the function which is qchisq(alpha,df,lower.tail = FALSE). Hence, the critical value for the test, χ^2 for significance level, $\alpha = 0.05$ and degree of freedom, df = 1 is 3.841459.

Therefore, since $\chi^2 = 1.2021 < 3.841459$, that is the test statistic does not fall within the critical region, thus we fail to reject the null hypothesis, H_0 at the significance level of 0.05. There is insufficient evidence of the relationship between whether the patients are smoker and whether the patients have hypertension.

3. 0 Discussion and Conclusion

To recapitulate, based on the result of the one-sample hypothesis testing for population proportion, we can conclude that there is sufficient evidence to support the claim that the proportion of patients who are overweight or obese is different from 80%. Also, after conducting the one-sample hypothesis testing for population mean, we have found that there is insufficient evidence to support the claim that the population of patients with coronary heart disease age more than 65 years old. Hence, we can say that the mean age of patients with coronary heart disease is no more than 65 years old. However, it is undeniable that there is still more likely for elderly people to develop coronary heart disease. Furthermore, the correlation test shows that there is relatively weak positive linear relationship between number of cigarettes smoked per day and heart rate in bpm with correlation coefficient, r = 0.336344. Hence, the heart rate of patients increases as the number of cigarettes smoked by them per day increases. Not only that, we have also shown that there is sufficient evidence of a linear relationship between number of cigarettes smoked per day and heart rate in bpm through the significance test for correlation.

Moreover, we have obtained the estimated regression equation for age and systolic blood pressure in mmHg which is $\hat{y}$ = 84.898 + 1.036x. It can greatly help us to predict the value of the dependent variable which is systolic blood pressure in mmHg based on the value of the independent variable which is the age of patients and also enable us to explain impact of changes in the age of patients on the systolic blood pressure in mmHg. Since the estimate of regression intercept, b_0 = 84.898, it indicates that systolic blood pressure is 84.898 mmHg when the age of the patients is 0 years old. Also, b_1 = 1.036 tells us that the average value of systolic blood pressure increases by 0.3609 mmHg, on average, for each additional 1 year old of age. Also, we can conclude that there is sufficient evidence that the age does affect the systolic blood pressure in mmHg through the t test for population slope. Next, goodness-of-fit test has shown that there is sufficient evidence to conclude that there are different proportions of patients for four education levels. Last but not least, Chi-square test of independence has told us that there is insufficient evidence of the relationship between whether the patients are smoker and whether the patients have hypertension.

Although we have come to a conclusion that the mean age of patients with coronary heart disease is no more than 65 years old, however we know that the age of patients will affect the systolic blood pressure of patients. Since the higher the age of patients, the higher the systolic blood pressure of patients. Higher blood pressure will cause the higher risk of having hypertension, which in turn causes the higher risk of coronary heart disease in the elderly and we can conclude that the elderly are the high-risk populations for coronary heart disease. Thus, some risk factors of coronary heart disease can be seen as interrelated. Also, even though smoking habit is known as the risk factor for coronary heart disease, but the association between smoking and hypertension is still being determined. Nevertheless, in the chi-square test of independence, we conclude that there is no relationship between the current smoking status and hypertension. However, it is still undeniable that smoking will increase blood pressure and heart rate of people, hence it is the major risk factors for coronary heart disease.

All in all, we can conclude that we have successfully done the statistical analysis using hypothesis testing, correlation, regression, goodness-of-fit test and chi-square test of independence. The relationship between the risk factors of coronary heart disease have been investigated and we have drawn the conclusions for the analysis. This project is accomplished in order to increase people's awareness about the importance of taking preventive actions as they can understand and identify all types of risk factors of coronary heart disease, so that they will change their lifestyles and lead a healthy life. It is because the more risk factors they have and the greater the degree of each risk factor, the higher their chances of developing coronary heart disease. Thus, a healthy lifestyle will certainly go a long way towards living a long and normal life, hence people will not suffer from coronary heart disease by modifying, treating and controlling some of the risk factors of coronary heart disease.

4.0 References

1) Aman Ajmera. (2017). *Framingham Heart Study*. [Dataset]. Retrieved from https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset

2) Cardiovascular diseases (CVDs). (2017, 17 May). Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

3) Ades, P. A., & Savage, P. D. (2017). Obesity in coronary heart disease: An unaddressed behavioral risk factor. *Preventive medicine*, *104*, 117–119. https://doi.org/10.1016/j.ypmed.2017.04.013.