UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI 2143-07
Probability & Statistical Data Analysis

SEMESTER II, SESSION 2021/2022

Project 2

(A Study on The Features That Increase The Chances of Having Cardiovascular Disease)

Lecturer: Dr. Nor Azizah Ali

Group Name: Little Group

Name	Student ID
TIA SIAW XUEN	A21EC0233
CHUA XIN LIN	A21EC0020
MA ZE JUN	A21EC4009
KEE LE WEI	A21EC0189

Presentation Video Link:
https://drive.google.com/file/d/1OJt5XakDHH-S1K2pkARSCIKMrpOM4Ana/view?usp=sharing

Introduction

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. According to the World Health Organization (WHO), cardiovascular diseases are the leading cause of death globally. As proof, 32% of all deaths (an estimated 17.9 million people) were caused by cardiovascular diseases in 2019. Despite that, people can still avoid cardiovascular diseases by maintaining a healthy lifestyle and doing regular body check-ups. Once patients are diagnosed with cardiovascular disease, they are advised to receive treatment immediately to avoid stroke or heart attack. Hence, cardiovascular diseases are always a hot topic to discuss and deserve more attention.

Therefore, we decided to do a study on patients with and without cardiovascular disease. The dataset chosen for the study was collected during the medical examination. The dataset consists of 70 000 observations and 12 variables. However, we only use 500 observations out of 70 000 with cardiovascular diseases to improve the accuracy of data. The variables taken into account for this study are age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol level, and glucose level. Besides, behaviour risk factors such as smoking habits and alcohol consumption are considered in the data. Lastly, variables such as physical activity and the presence of cardiovascular disease are also included in the dataset.

The patient's age, height and weight are recorded in days, cm, and kg, respectively. Meanwhile, the cholesterol and glucose levels are represented in 3 levels, 1 represents normal, 2 represents above normal, and 3 represents well above normal. Furthermore, variables such as smoking habit, alcohol intake, physical activity and presence of cardiovascular disease are documented in binary, which is 0 means no while 1 means yes.

We hope this study will increase public awareness of cardiovascular diseases. Everyone should avoid cardiovascular diseases because they are very dangerous and might cause sudden death. For example, stroke and heart attack. The elders and youngsters must pay close attention to cardiovascular disease, especially those with hypertension. This is because hypertension can be a major contributing factor to cardiovascular disease. Therefore, everyone must maintain a healthy lifestyle and avoid bad behaviour such as smoking, drinking alcohol, and staying up late. All of us must have the idea to prevent ourselves from suffering from cardiovascular diseases before it is too late. As the saying goes, prevention is better than cure.

Dataset

To increase the accuracy for inference statistical analysis, We use five hundred observations out of seventy thousand observations with people that presence of cardiovascular disease.The dataset chosen contains the medical examination on patients with and without cardiovascular disease. The medical examination has a total of 12 variables which contains 4 objective features, 4 examination features, 3 subjective features, and 1 target variable. From the dataset, there are continuous and discontinuous variables that we have identified. The continuous variables in this dataset are age, height, weight, systolic blood pressure and diastolic blood pressure. Systolic blood pressure and diastolic blood pressure are used to measure blood pressure. Meanwhile, gender, cholesterol, glucose, smoking, alcohol intake, physical activity and presence or absence of cardiovascular disease are discontinuous variables. Cholesterol and glucose represent the patients' cholesterol and glucose level whether it is normal, above normal or well above normal. On the other hand, smoking, alcohol intake and physical activity are from

one category which is used to represent the social history of patients.

Variables	**Suggested Answer**	**Data Type**	**Level of Measurement**
Age	Metric value	Quantitative	Ratio
Height	Metric value	Quantitative	Ratio
Gender	Male/Female	Qualitative	Nominal
Systolic blood pressure	Metric value	Quantitative	Ratio
Diastolic blood pressure	Metric value	Quantitative	Ratio
Cholesterol	Normal/Above normal/ Well above normal	Qualitative	Interval
Glucose	Normal/Above normal/ Well above normal	Qualitative	Interval
Smoking	Yes/No	Qualitative	Nominal
Alcohol intake	Yes/No	Qualitative	Nominal
Physical Activity	Active/Inactive	Qualitative	Nominal
Presence or Absence of cardiovascular disease	Presence/Absence	Qualitative	Nominal

In our project, variable systolic blood pressure is chosen to conduct one-sample hypothesis testing whether to test whether there is sufficient evidence to support the claim that the population of patients with cardiovascular disease have a systolic blood pressure more than 120 mmHg. Other than that, we have chosen variables such as age and diastolic blood pressure to perform a regression test to test whether the age does affect the diastolic blood pressure. Moreover, we also chose age of patient and systolic blood pressure to conduct a correlation test to determine if there is a relationship between age of patient and systolic blood pressure. Lastly, variables smoking and physical activity are chosen to conduct Chi-Square Test of Independence to determine if there is a relationship between the cardiovascular disease patients are smokers and whether the cardiovascular disease patients are active in physical activity.

Hypothesis Testing (One Sample Test)

Since systolic blood pressure is one of the examination features of cardiovascular disease and there is a study that states that a normal systolic blood pressure is less than 120 mmHg. Thus, we will conduct a one sample hypothesis testing to test whether there is sufficient evidence to support the claim that the population of patients with cardiovascular disease have a systolic blood pressure more than 120 mmHg using a significance level, $\alpha = 0.05$. A sample of 500 data is used in this hypothesis testing and the population variance is unknown.

In this case, hypothesis statement is:

H_0: $\mu = 120$

H_1: $\mu > 120$

Then, we will calculate the test statistics for population mean by apply the following formula:

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

```
> # Hypothesis Testing (One Sample Test) for population mean with unknown population variance (n>30)
> systolic_bp <- Cardiovascular_dataset$ap_hi
> n = 500 # Sample size
> alpha = 0.05 # Significance level
> u = 120 # Claimed population mean
> # Calculate and display sample mean of systolic_bp
> x <-  mean(systolic_bp)
> x
[1] 135.636
> # Calculate and display sample standard deviation of systolic_bp
> s <- sd(systolic_bp)
> s
[1] 18.69102
> # Calculate and display z, test statistics
> z <- (x - u)/(s / sqrt(n))
> z
[1] 18.70586
> # Calculate and display critival value for right-tail test
> z.alpha <- qnorm(1-alpha)
> z.alpha
[1] 1.644854
> # Calculate and display p-value for z, test statistics
> pval <- pnorm(z, lower.tail = FALSE)
> pval
[1] 2.217613e-78
```

Figure 1

The code in Figure 1 shows the steps to conduct the one sample hypothesis test in Rstudio. From the calculations, the sample mean of systolic blood pressure is 135.636 and the sample standard deviation of systolic blood pressure is 18.6910. By applying the formula above, we calculated the test statistics, $z = 18.7059$. The critical value with significance level of 0.05, $z_{0.05} = 1.6449$. In this case, we will only reject the null hypothesis, H_0 if the $z > z_{0.05}$ since it is a right-tailed test.

Since $z = 18.7059 > z_{0.05} = 1.6449$, hence we reject the null hypothesis, H_0 at the significance level of 0.05. There is sufficient evidence to support the claim that the population of patients with cardiovascular disease have a systolic blood pressure more than 120 mmHg.

Correlation

In this correlation test, we decided to find out if there is any linear relationship between age of patient and their systolic blood pressure with a random sample of 500. Since both variables are ratio-type data, we will use Pearson's product-moment correlation coefficient method. The significant level, α used is 0.05. Therefore, the critical value, $t_{(\pm 0.025,\ 498)} = \pm 1.960$.

The hypothesis statement for the correlation test is as below:

H_0: $\rho = 0$ (No linear correlation)

H_1: $\rho \neq 0$ (Linear correlation exist)

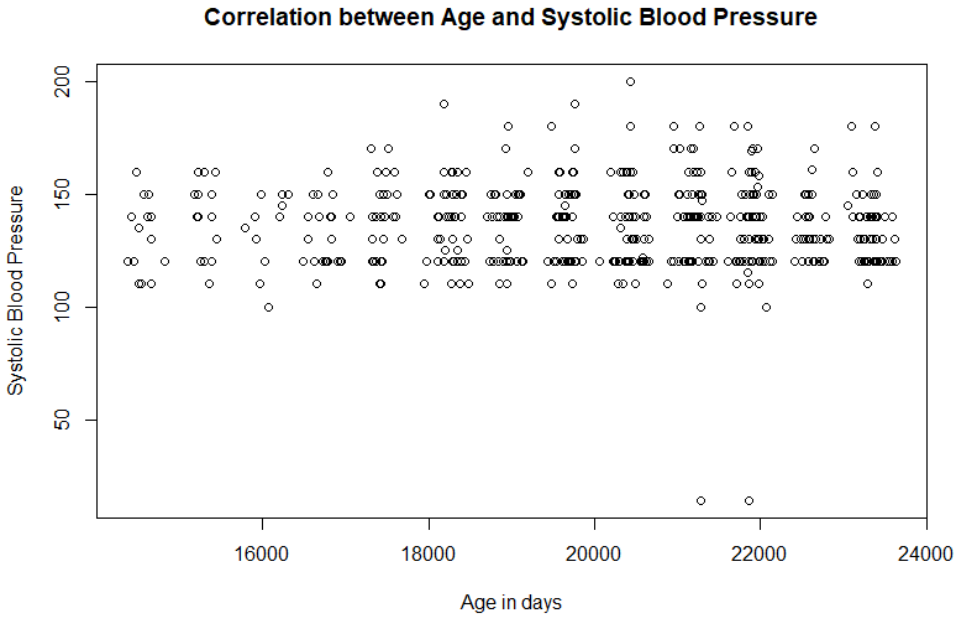
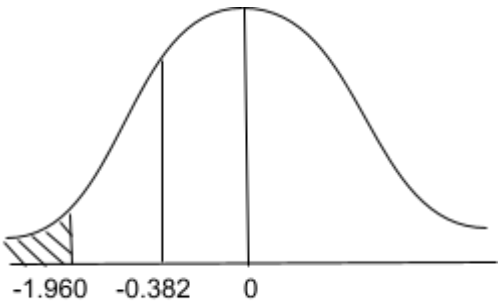


Figure2

```
> # Correlation test
> x <- Cardiovascular_dataset$age
> # Correlation test
> x <- Cardiovascular_dataset$age
> # Correlation coefficient, r and test statistic, t value
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = -0.38207, df = 498, p-value = 0.7026
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.10465210  0.07067807
sample estimates:
        cor
-0.01711861
```



Based on the value from R programming, t = -0.38207 does not fall at the critical region. Hence, we fail to reject H_0. There is insufficient evidence to prove that there exists a linear relationship between age and systolic blood pressure. As proof, the r value is very close to 0.

Regression

Regression analysis is used to:

1. Predict the value of a dependent variable based on the value of at least one independent variable.
2.Explain the impact of changes in an independent variable on the dependent variable.

formula:

$\hat{y}i = b0 + b1x$

bo and b1 can get value from Rstudio

Regression analysis predicts the value of a dependent variable through the value of at least one independent variable, and can also explain the influence of the change of independent variable on the dependent variable. Our project is about the characteristics of cardiovascular disease. The examination of cardiovascular disease can be divided into many kinds. Here, I want to study the relationship between age and vasoconstriction

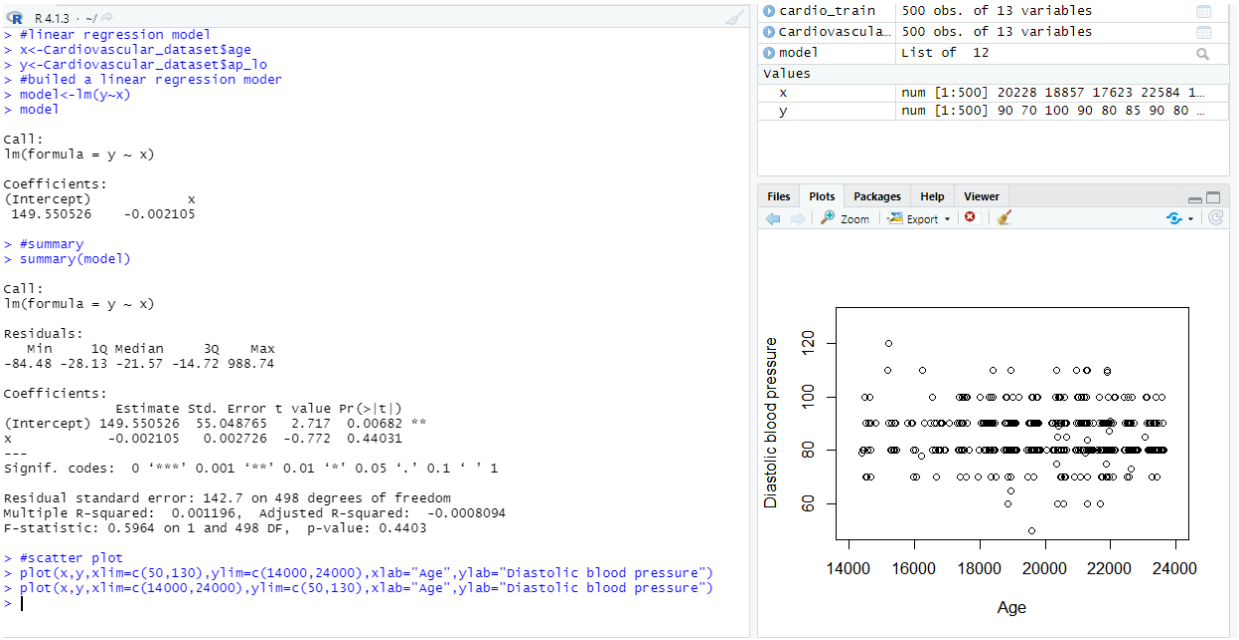


Figure 4

According to figure 4, we get the estimated regression equation for our sample

$\hat{y}i = 149.550526 - 0.002105x$ from Rstudio

so the intercept b0= *149.550526,* and regression slope b1=*-0.002105*

then we know the diastolic blood pressure at age 0 was *149.550526*mmHg, and b1 it tells us that diastolic blood pressure increases by an average of *-0.002105* mmHg for each additional year

The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable

R^2=SSR/SST so, we also can get coefficient of determination value R^2= *-0.002105*

$0 \leq R^2 \leq 1$ this indicates that the linear relationship between age and diastolic blood pressure is weak, only *0.2105*% of the change in diastolic blood pressure could be explained by changes in patient age

Regression slope test: t test

We will also perform a t-test on the population slope to determine whether there is a linear relationship between age and diastolic blood pressure. The assumptions are stated as follows:

H_0 : $\beta 1 = 0$ (Age and Diastolic blood pressure no linear relationship)
H_1 : $\beta 1 \neq 0$ (Age and Diastolic blood pressure linear relationship does exist)

$t = (b1 - \beta 1)/sb1$ d.f.=n-2 df=500-2=498
where:
$b1$ = Sample regression slope coefficient
$\beta 1$ = Hypothesized slope
$sb1$ = Estimator of the standard error of the slope

According to the figure4, we can get value for
b1= -0.002105 Sb1=0.002726 t= -0.772
it is a two-tailed test, so a=0.05, so a/2=0.025
Critical value, $t_{(\pm 0.025,\ 498)} = \pm 1.960$, since -1.960<t<1.960, so fail to reject H0,
there is no sufficient evidence that age does affect diastolic blood pressure.

Chi-Square Test of Independence

The sample size of this project is 500. The two variables considered are whether the cardiovascular disease patients are smokers and whether the patients are active in physical activity. We will test for evidence of the relationship between these variables at the significance level of 0.05.

Hypothesis statement:

H_0**:** No relationship between whether the cardiovascular disease patients are smokers and whether the cardiovascular disease patients are active in physical activity.

H_1**:** There is a relationship between the cardiovascular disease patients are smokers and whether the cardiovascular disease patients are active in physical activity.

```
> #Create two-way contingency table
> table <- table(Cardiovascular_dataset$smoke,Cardiovascular_dataset$active)
> table

     0   1
  0 104 351
  1   9  36
> #Perform Chi-Square Test on data table
> res <- chisq.test(table,correct = FALSE)
> res

        Pearson's Chi-squared test

data:  table
X-squared = 0.1911, df = 1, p-value = 0.662

> #Calculate expected frequency
> res$expected

         0      1
  0 102.83 352.17
  1  10.17  34.83
> #Critical value
> alpha <- 0.05
> x2.alpha <- qchisq(alpha,df=1,lower.tail = FALSE)
> x2.alpha
[1] 3.841459
```

Figure 5

Figure 5 shows how we perform Chi-Square Test of Independence using RStudio and the results of Chi-Square Test of Independence for whether the cardiovascular disease patients are smoker and whether the cardiovascular disease patients are active in physical activity. We have created the two-way contingency table that is obtained from the result:

	Active in Physical Activity				
	No		Yes		
Smoker	Observed count	Expected Count	Observed count	Expected Count	Total
No	104	102.83	351	352.17	455
Yes	9	10.17	36	34.83	45
Total	113	113	387	387	500

Function chisq.test() in R is used to find test statistic value, χ 2 = 0.1911. Based on the formula df = (r-1) (c-1), we can obtain a degree of freedom, df = (2-1) (2-1) =1. The function qchisq(alpha,df,lower.tail = FALSE) is used to compute the critical value for the significance level of 0.05. Therefore, we know that the critical value for test statistic value, χ 2 = 0.1911, degree of freedom, df = 1, is 3.841459.

Since test statistic value (0.1911) < critical value (3.841459), the test result does not fall within the critical region, we do not reject the null hypothesis, H_0 at the 0.05 significance level. There is insufficient evidence to show that there is a relationship between the cardiovascular disease patients who are smokers and whether the cardiovascular disease patients are active in physical activity.

Conclusion

Our group project is to analyze the datasets of cardiovascular disease, in the selected datasets stage, our group at first confused, do not know from which aspect to consider, because there are a lot of choice, then we began to look for suitable datasets, in selecting some consider the datasets, we started to carry on the analysis, but in the process of analysis, We found that there are problems for some of the data accuracy, and some data is incomplete, finally we decided to choose cardiovascular disease dataset as the research target of our group project, our project by different indicators to calculate the problem of cardiovascular disease, hope that citizens can improve the cognition to the cardiovascular disease, In daily life we can do more exercise, healthy eating, go to hospital making inspection regularly to reduce cardiovascular disease as it is harmful to people. In this project, we have four team members with a clear division of responsibilities, everyone finishes their part seriously, then discusses and analyzes together, we've learned how to use Excel to analyze process data, How to use RStudio to complete the test and make various charts. Through the test, we can get some informations, such as there is enough evidence to support the statement that the systolic blood pressure of patients with cardiovascular diseases exceeds 120 mmHg, and the relationship between diastolic blood pressure and age etc., so that we can expand the knowledge from textbooks to the Internet. From the Internet to our daily life, I think this project allows us to learn more than what we have learned from books. Finally, our group would like to thank Dr Nor Azizah Ali again. She is patient in solving problems for us and asking questions in class, so that everyone can concentrate on the class.

Appendix

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?select=cardio_train.csv