SECI 1143-01
Probability & Statistical Data Analysis

Project 2

Lecturer:

Dr Sharin Hazlin Binti Huspi

Group 1

Name	Matric number
Choh Jing Yi	A23CS0296
Liow Jia Feng	A23CS0302
Chua Shang Yeet	A23CS0297
Tay Ching Xian	A23CS0307
Muhammad Hilmi Hijazi bin Jamal	A23CS0303

Table of Contents

Introduction

Heart diseases are also known as cardiovascular diseases (CVDs). Heart diseases are a group of disorders of the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, and valvular heart disease. Heart diseases is the most causes of death in the world. Regarding to World Health Organization (WHO), approximately 17.9 million of people dead in 2019 which also representing almost 32% of all global death. Among these deaths, 85% of the data collected was killed by heart attack and strokes.

In this case, we have curiosity about the causes of heart diseases those lead a person to death. Therefore, the aim of this project is to research the relation between heart diseases and the potential factors that may cause heart diseases. The approach of statistical analysis such as two sample tests, correlation tests, regression tests and Chi-square test of independence is used to test the correlation between the heart diseases and the potential factors.

The data set in this project is provided by Lapp, D. (2019, June 6). *Heart disease dataset*. Kaggle. The data is collected from Cleveland, Hungary, Switzerland, and Long Beach V with the contains the data:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. old peak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. Thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

Dataset

2.1 Description of Data

The dataset that we have chosen for our project is Heart Disease Data. This dataset contains heart disease patient in Cleveland, Hungary, Switzerland, and Long Beach V. The variables in the dataset are shown in Table 1 and the summary of the selected variables and the test with their respective description is demonstrated in Table 2.

Data Description:

Variables (Description)	Type of Variable	Measurement Level
ID (patients from 1 to 1025)	Quantitative	Nominal
Age (age of patients)	Quantitative	Ratio
Sex (gender of patients)	Qualitative	Nominal
Chest pain type (4 values)	Qualitative	Ordinal
Resting blood pressure	Quantitative	Ratio
Serum cholesterol in mg/dl	Quantitative	Ratio
Fasting blood sugar > 120 mg/dl	Qualitative	Nominal
Resting electrocardiographic results (values 0,1,2)	Qualitative	Ordinal
Maximum heart rate achieved	Quantitative	Ratio
Exercise induced angina	Qualitative	Nominal
Oldpeak (ST depression induced by exercise relative to rest)	Quantitative	Ratio
The slope of the peak exercise ST segment	Qualitative	Ordinal
Number of major vessels (0-3) colored by flouroscopy	Quantitative	Ratio
Thal (0 = normal; 1 = fixed defect; 2 = reversable defect)	Qualitative	Ordinal

Table 1: Variables in Health Disease Data

2.2 Statistical Test Analysis

Selected Variable(s)	Test	Description
Maximum heart rate achieved, Sex	Hypothesis testing (two sample test)	Explanation: The variable is used to test whether the mean maximum heart rate achieved of male patient different from mean maximum heart rate achieved of female patient Possible outcome: The mean maximum heart rate achieved of male patient same from mean maximum heart rate achieved of female patient
Age, resting blood pressure	Hypothesis testing (correlation test)	Explanation: The variables are selected to test whether the resting blood pressure can be predicted based on the age using linear correlation analysis at a significant level of 0.05. Possible outcome: There is a linear relation between age and resting blood pressure achieved to allow the prediction of resting blood pressure based on age.
Age, maximum heart rate achieved	Hypothesis testing (Regression test)	Explanation: The variables are selected to test whether the maximum heart rate achieved can be predicted based on the age using linear regression analysis at a significant level of 0.05. Possible outcome: There is a linear relation between age and maximum heart rate achieved to allow the prediction of maximum heart rate achieved based on age.
Sex and Fasting Blood Sugar	Hypothesis testing (Chi Square of Independence)	Explanation: Investigating the relationship between sex (gender) and fasting blood sugar (FBS) levels is important in medical research because it helps understand if there are gender-based differences in the risk or prevalence of high blood sugar levels, which can indicate diabetes or pre-diabetes. Possible Outcome:

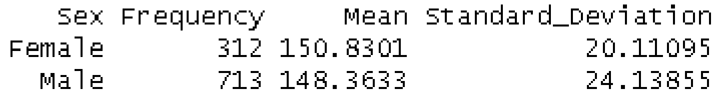
		No Association (Fail to Reject H0). The test statistic is less than the critical value, and the p-value is greater than the significance level (0.05).

Table 2: Summary of Selected Variables and Test with their Respective Description

3.0 Data Analysis

3.1 Two Sample Hypothesis Test

In this analysis, we will use variables maximum heart rate achieved and sex, where we will test whether the mean maximum heart rate achieved of male patient different from mean maximum heart rate achieved of female patient or not at 95% confidence level, assuming unequal variances. From the data, frequency (n), mean (x), standard deviations (s) are calculated.



```
    Sex Frequency     Mean Standard_Deviation
 Female       312 150.8301           20.11095
   Male       713 148.3633           24.13855
```

From the calculation, we know that:

	Female	**Male**
Frequency	$n_1 = 312$	$n_2 = 713$
Mean	$\overline{x_1} = 150.8301$	$\overline{x_2} = 148.3633$
Standard Deviation	$s_1 = 20.11095$	$s_2 = 24.13855$

Hypothesis Statement:

H_0: μ1 = μ2

H_1: μ1 ≠ μ2

2. Given 95% confidence level, α = 0.05. The test statistics, to can be calculated using the formula:

$$t_0 = \frac{\bar{x}_1 - \overline{x_2} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

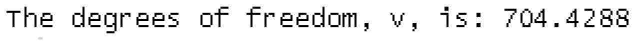
By using RStudio, the test statistics, t_0 = -1.6969

```
  t0                        -1.69685021199889
```

3. Calculate the degree of freedom using the formula:

$$v = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\frac{\left({s_1}^2/_{n_1}\right)^2}{n_1 - 1} + \frac{\left({s_2}^2/_{n_2}\right)^2}{n_2 - 1}}$$

By using RStudio, degree of freedom, v = 704.4288 ≈ 704



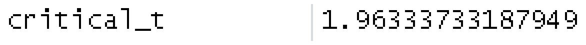
```
The degrees of freedom, v, is: 704.4288
```

Therefore, using $\boldsymbol{\alpha} = 0.05$, we reject H_0

if $t_0 > t_{0.025,\ 704} = 1.96$

or $t_0 < -t_{0.025,\ 704} = -1.96$

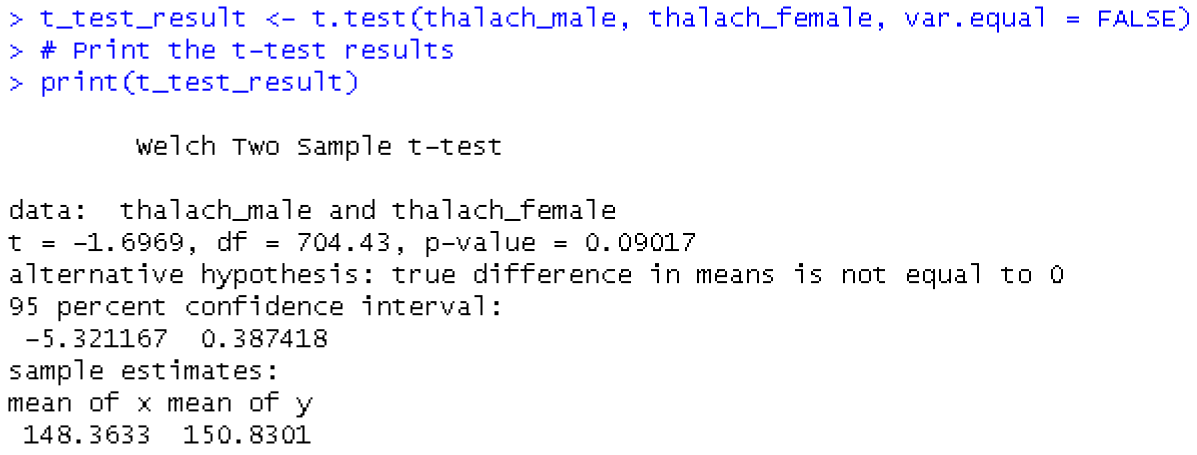
∴ Critical value $t_{-0.025,\ 704} = -1.96$, $t_{0.025,\ 704} = 1.96$



```
critical_t            1.96333733187949
```

4. Conclusion

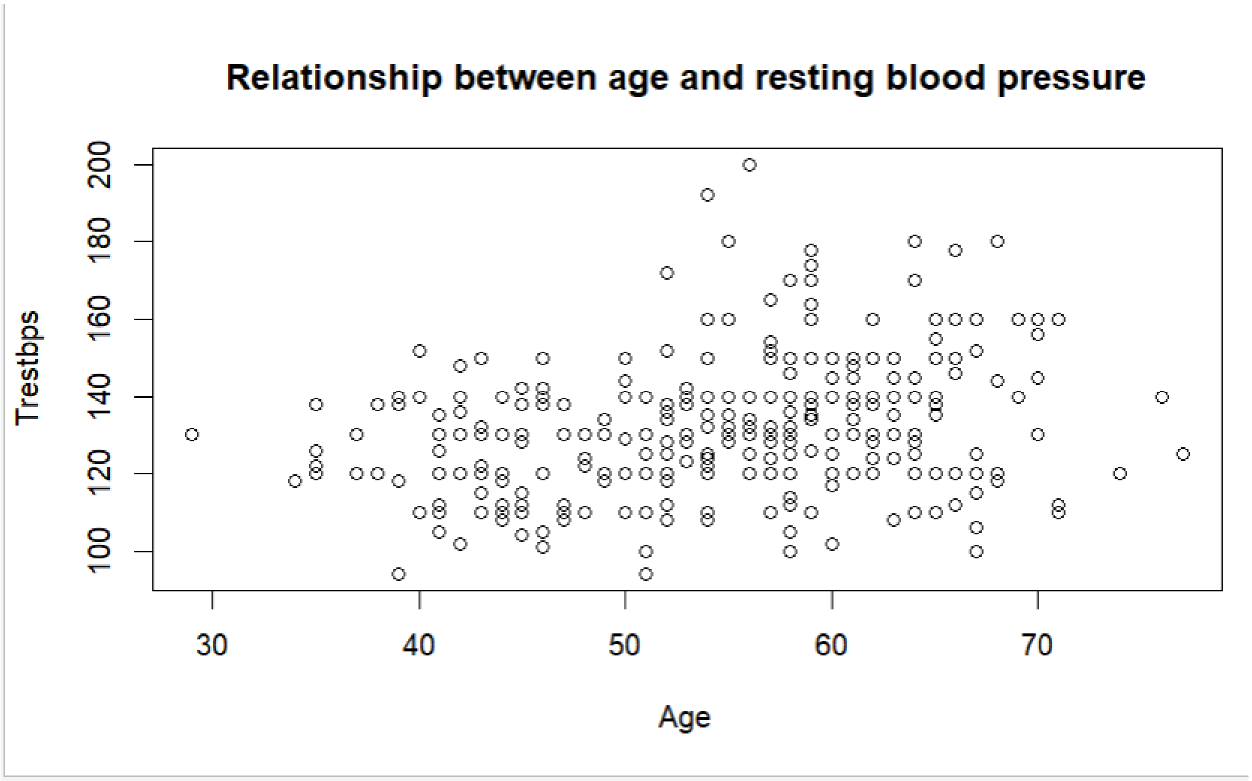
Since t0 = -1.6969 > critical value t-0.025, 704 = -1.96, we fail to reject the null hypothesis. Hence, there is not sufficient evidence to conclude that the maximum heart rate achieved of male patient is different with the mean maximum heart rate achieved of female patient.



```
> t_test_result <- t.test(thalach_male, thalach_female, var.equal = FALSE)
> # Print the t-test results
> print(t_test_result)

        Welch Two Sample t-test

data:  thalach_male and thalach_female
t = -1.6969, df = 704.43, p-value = 0.09017
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.321167  0.387418
sample estimates:
mean of x mean of y 
 148.3633  150.8301 
```

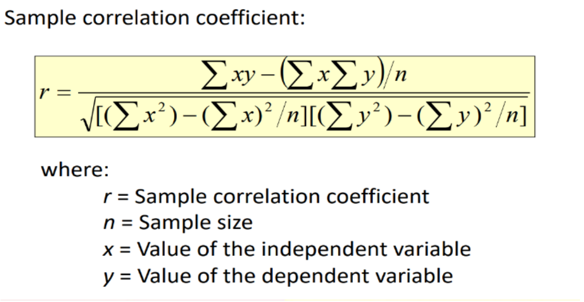
3.2 Correlation Test

In this analysis, we will use the variables age and trestbps to determine whether there is a linear relationship between age and resting blood pressure at a 95% confidence level using Pearson's Product-Moment Correlation Coefficient. The strength of association, which is the linear relationship between two variables, is measured using correlation analysis.



According to the scatter plot above, there will be a weaker positive correlation relationship between the patients' age and resting blood pressure.

1. Calculate the sample correlation coefficient using Pearson's method by:

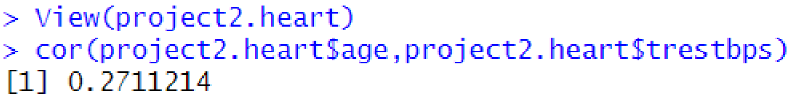


Sample correlation coefficient:

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

where:

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

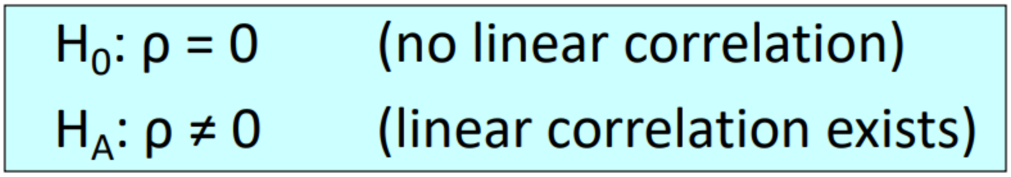


```
> View(project2.heart)
> cor(project2.heart$age,project2.heart$trestbps)
[1] 0.2711214
```

By using RStudio, we can see that the sample correlation coefficient, r is 0.2711214, which indicates that there is a relatively weak positive linear correlation between the age and resting blood pressure.

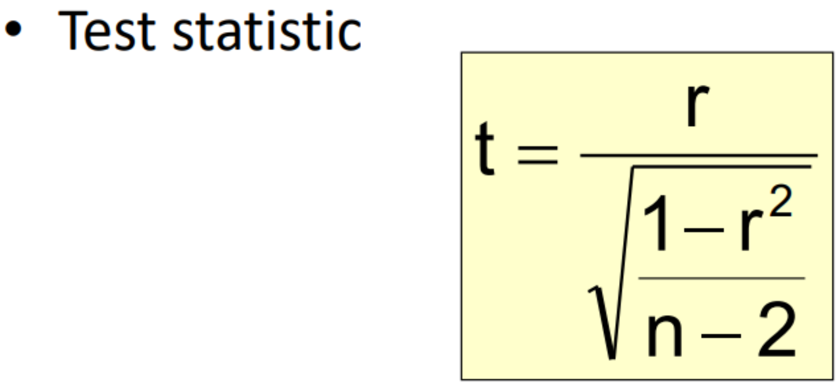
2. Significance Test for Correlation

Hypothesis:



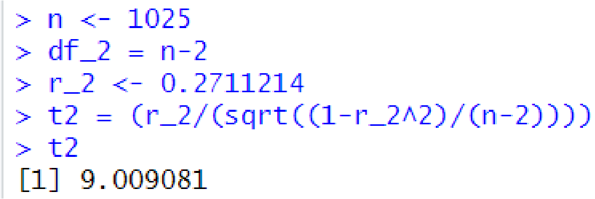
$H_0: \rho = 0$ (no linear correlation)

$H_A: \rho \neq 0$ (linear correlation exists)



- Test statistic

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



```
> n <- 1025
> df_2 = n-2
> r_2 <- 0.2711214
> t2 = (r_2/(sqrt((1-r_2^2)/(n-2))))
> t2
[1] 9.009081
```

By using RStudio, we get the test statistic, t is 9.009081.

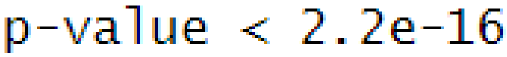
Find critical value, using α = 0.05, df = n-2 = 1023

From t-table, since this is a two-tailed test, there are two critical values:

Lower tail critical value = -1.962346

Higher tail critical value = 1.962346

From RStudio, we also get p-value = 1.962346.



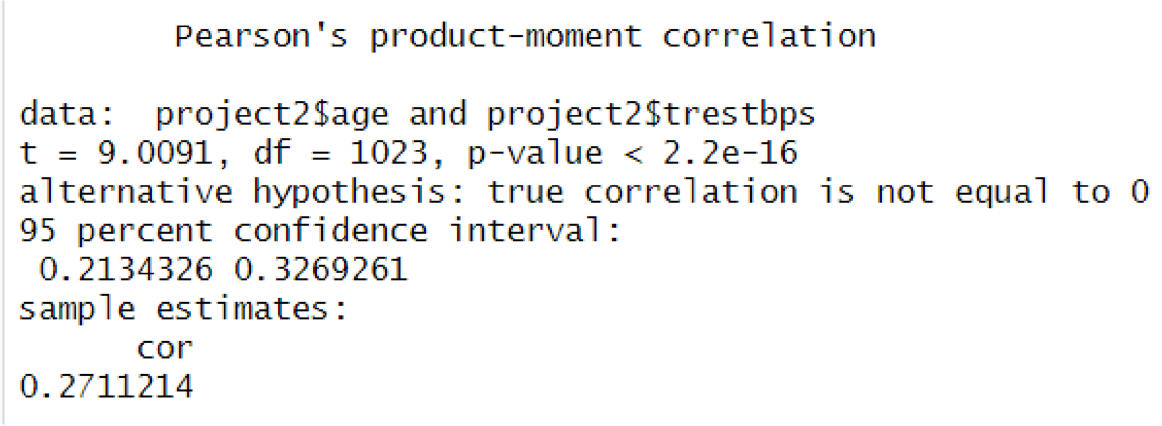
```
p-value < 2.2e-16
```

From R-studio

Hence, if test statistic > 1.962346 or test statistic < -1.962346, we will reject null hypothesis.

Decision

Since the test statistic, t = 9.009081 > upper tail critical value which is 1.962346, we will reject the null hypothesis. In conclusion, there is sufficient evidence to conclude that there is a linear relationship between the age and resting blood pressure at α = 0.05.



```
        Pearson's product-moment correlation

data:  project2$age and project2$trestbps
t = 9.0091, df = 1023, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2134326 0.3269261
sample estimates:
      cor 
0.2711214 
```

Perform Pearson's product-moment correlation by R-studio

3.3 Regression Test

In this regression analysis, we aim to determine whether a student's age can predict their maximum heart rate achieved. The hypothesis is that age negatively correlates with maximum heart rate achieved due to the natural decline in heart rate with aging. Here, age is considered the independent variable (x), and the maximum heart rate achieved (thalach) is the dependent variable (y). The population regression line is estimated using the sample regression line derived from the data.

$$\underset{\substack{\uparrow\\ \text{Dependent}\\ \text{Variable}}}{Y_i} = \overset{\substack{\text{Constant/Intercept}\\ \downarrow}}{\beta_0} + \underset{\substack{\uparrow\\ \text{Slope/Coefficient}}}{\beta_1} \overset{\substack{\text{Independent}\\ \text{Variable}\\ \downarrow}}{X_i}$$

From equation above:

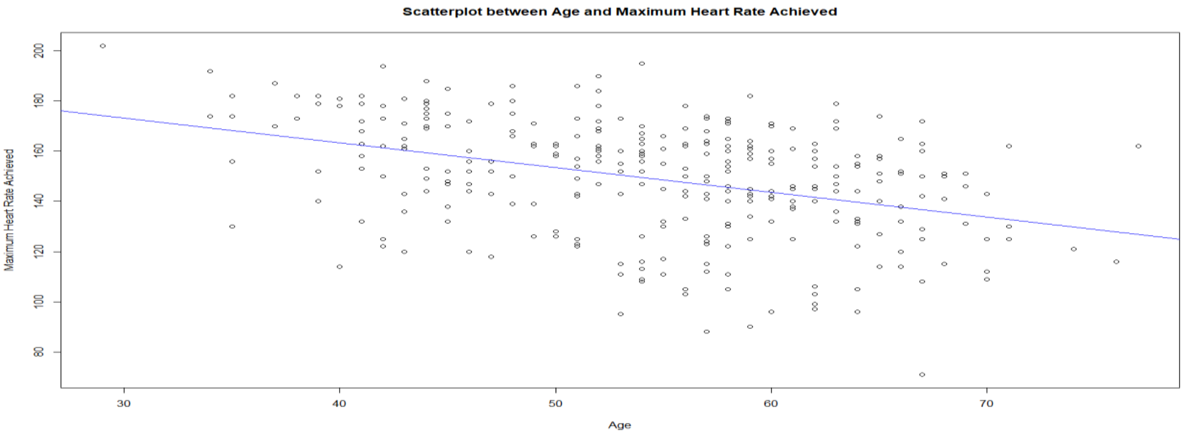
y = Estimated (or predicted) Y value (thalach)

b0 = Estimate of the regression intercept

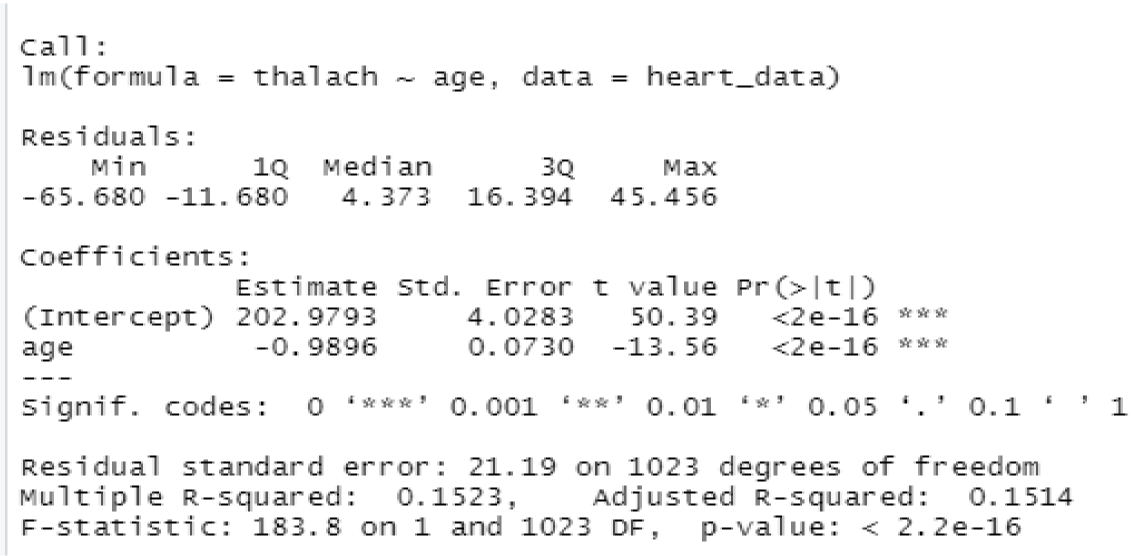
b1 = Estimate of the regression slope

x = Independent variable (age)

Using the dataset, we calculate the correlation between age and maximum heart rate achieved: Correlation (r)=−0.39. A linear regression model is constructed to predict the maximum heart rate achieved based on age. In the visualization, the data points will be represented by the black circle and the regression line will be represented in blue line.



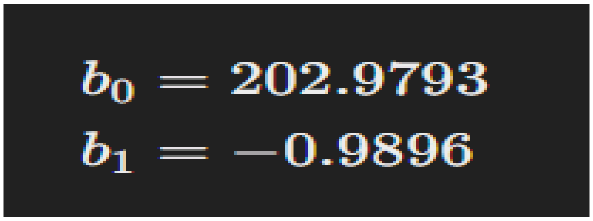
Scatterplot between age and maximum heart rate



```
Call:
lm(formula = thalach ~ age, data = heart_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.680 -11.680   4.373  16.394  45.456 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 202.9793     4.0283   50.39   <2e-16 ***
age          -0.9896     0.0730  -13.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.19 on 1023 degrees of freedom
Multiple R-squared:  0.1523,	Adjusted R-squared:  0.1514 
F-statistic: 183.8 on 1 and 1023 DF,  p-value: < 2.2e-16
```

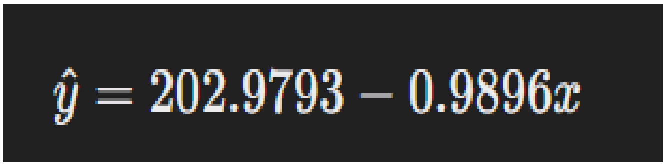
Summary of scatter plot graph from RStudio

By using the RStudio, we can get the b0 and b1:



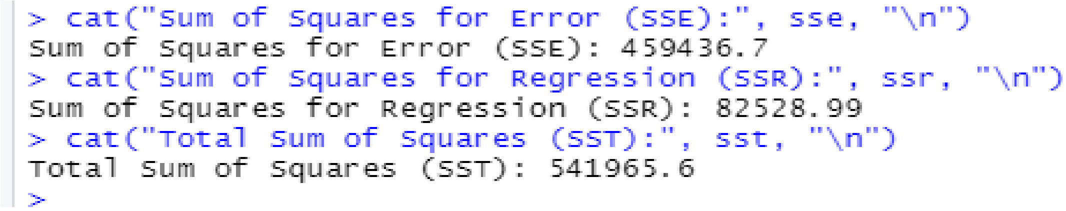
$$b_0 = 202.9793$$
$$b_1 = -0.9896$$

From the summary, the formula for the estimated regression model is:



$$\hat{y} = 202.9793 - 0.9896x$$

Above the equation, we can interpret the data of intersection coefficient b0 and slope coefficient b1. In the data, b0 =202.9793 indicates that for patients with an Age of 0, the Maximum Heart Rate Achieved would become 202.9793. b1 =−0.9896 indicates that the Maximum Heart Rate Achieved decreases by 0.9896 on average for each additional one year of Age.



```
> cat("Sum of Squares for Error (SSE):", sse, "\n")
Sum of Squares for Error (SSE): 459436.7 
> cat("Sum of Squares for Regression (SSR):", ssr, "\n")
Sum of Squares for Regression (SSR): 82528.99 
> cat("Total Sum of Squares (SST):", sst, "\n")
Total Sum of Squares (SST): 541965.6 
> 
```

Value for SSE, SSR and SST

Coefficient of Determination: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Sum of Squares Total: $SST = \sum (y - \bar{y})^2$

Sum of Regression Total: $SSR = \sum (y' - \bar{y'})^2$

Sum of Errors Total: $SSE = \sum (y - y')^2$

Where:
y` (y hat) is the estimate of y
$\bar{y}$ (y bar) is the mean of y

Where:

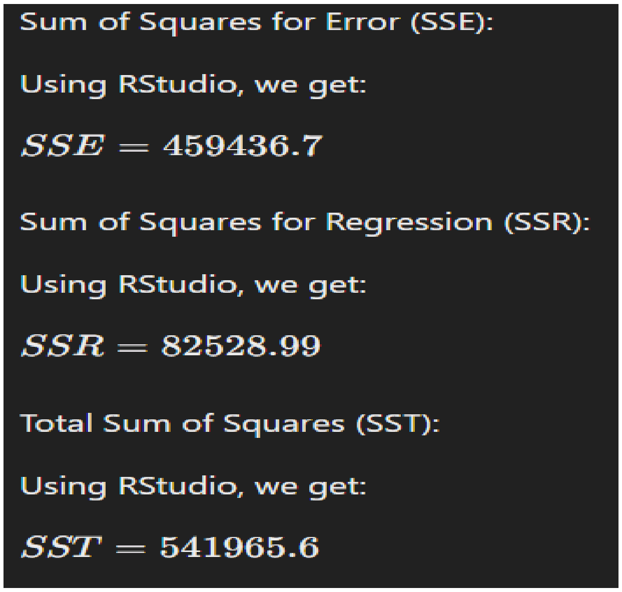
R2 = coefficient of determination

SSR = sum of squares for regression

SSE = sum of squares for error

SST= total sum of squares

By using the RStudio, we can get the value for SSE, SSR and SST



Sum of Squares for Error (SSE):

Using RStudio, we get:

$SSE = 459436.7$

Sum of Squares for Regression (SSR):

Using RStudio, we get:

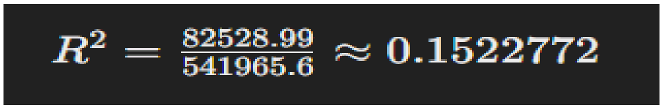
$SSR = 82528.99$

Total Sum of Squares (SST):

Using RStudio, we get:

$SST = 541965.6$

Thus, the R2 value can be calculated as follows:



$$R^2 = \frac{82528.99}{541965.6} \approx 0.1522772$$

This indicates that 15.23% of the variation in Maximum Heart Rate Achieved is explained by the variation in Age. Since the value of R2 being 0.1522772, so it also indicates that the linear relationship between Age and Maximum Heart Rate Achieved is relatively weak.

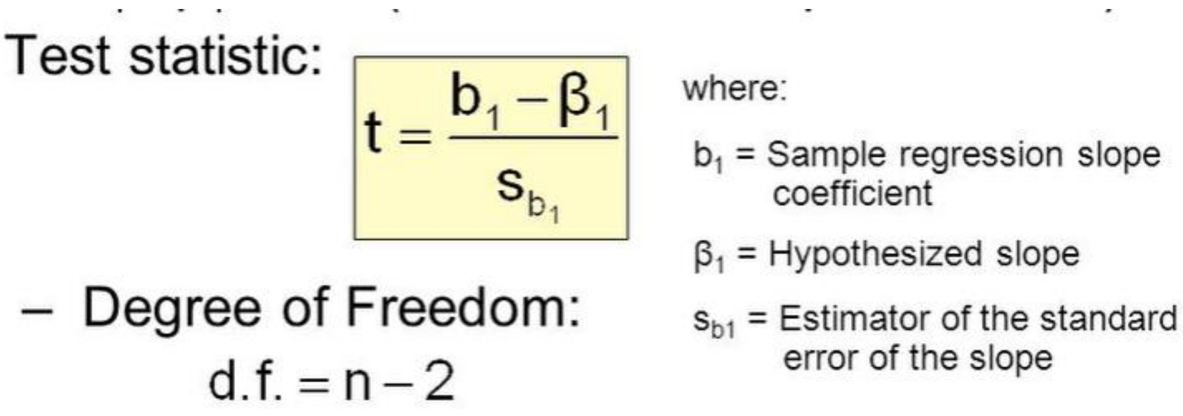
Inference for Regression T-test

Hypothesis testing:

Null hypothesis, H0:B1 =0(no linear relationship)

Alternative hypothesis, H1:B1 ≠0(linear relationship does exist)

Formula for test statistics:



Test statistic:

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

– Degree of Freedom:

$$d.f. = n - 2$$

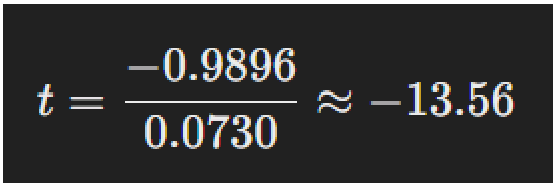
where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

By using the RStudio, we can get the test statistic is –13.56



$$t = \frac{-0.9896}{0.0730} \approx -13.56$$

Conclusion:

Since the t value for age in absolute value is greater than the critical value of t-test with the significant level of 0.05, we reject the null hypothesis. Thus, there is sufficient evidence to indicate that age has a statistically significant negative effect on maximum heart rate.

3.4 Chi-square Test of Independence

In this analysis, we will use the variables sex and fasting blood sugar to assess the independence of sex and fasting blood sugar status using a Two-Way Contingency Table at a 95% confidence level. As a result, we employ the Chi-Square Test of Independence in conjunction with a two-way contingency table.

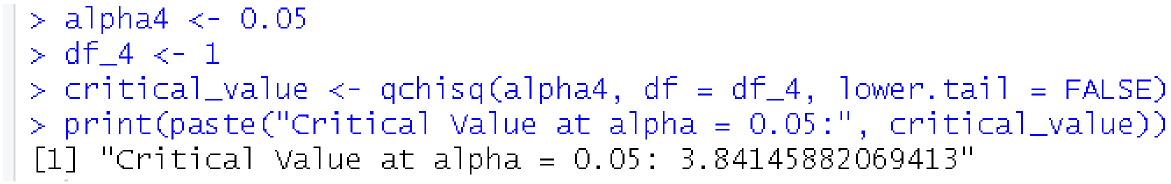
Sex/Gender	**Fasting Blood Sugar ≤ 120 mg/dl (fbs=0)**	**Fasting Blood Sugar > 120 mg/dl (fbs=1)**	**Total**
Male (1)	602	111	713
Female (0)	270	42	312
Total	872	153	1025

1. State the Hypotheses

- Null Hypothesis (H0): There is no association between gender and fasting blood sugar levels > 120 mg/dl. The variables are independent.
- Alternative Hypothesis (H1): There is an association between gender and fasting blood sugar levels > 120 mg/dl. The variables are dependent.

2. Find the critical value:

- The critical value for the Chi-Square test is determined by the significance level (α) and the degrees of freedom (df). For a typical significance level of 0.05 and df = 1:
- Using the Chi-Square distribution table, we look up the critical value for α=0.05 and df = 1:
- Critical value x^2 = 3.841



```
> alpha4 <- 0.05
> df_4 <- 1
> critical_value <- qchisq(alpha4, df = df_4, lower.tail = FALSE)
> print(paste("Critical Value at alpha = 0.05:", critical_value))
[1] "Critical Value at alpha = 0.05: 3.84145882069413"
```

- By using RStudio, the value of the critical value, x^2 is 3.84145882069413.

3. Calculate the expected counts

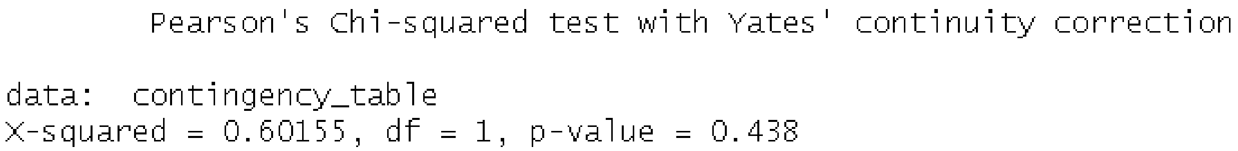
Sex/Gender	**Fasting Blood Sugar ≤ 120 mg/dl (fbs=0)**		**Fasting Blood Sugar > 120 mg/dl (fbs=1)**		**Total**
	Observed	Expected	Observed	Expected	
Male (1)	602	$\frac{872 \times 713}{1025} = 606.57$	111	$\frac{153 \times 713}{1025} = 106.43$	713
Female (0)	270	$\frac{872 \times 312}{1025} = 265.43$	42	$\frac{153 \times 312}{1025} = 46.57$	312
Total	872	872	153	153	1025

4. Calculate the statistic value

Cell, ij	**Observed Count, o_{ij}**	**Expected Count, e_{ij}**	$\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$
1,1	602	606.57	0.0344
1,2	270	265.43	0.0787
2,1	111	106.43	0.1962
2,2	42	46.57	0.4485
		x^2	0.7578

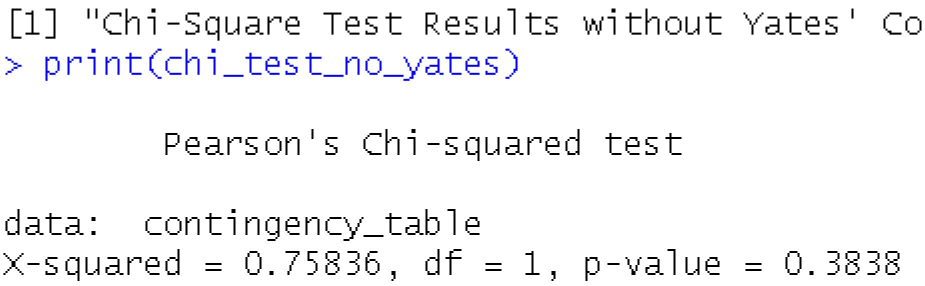
When we calculate test statistic manually, we get test statistic, x^2= 0.7578

- Using Rstudio with Yates correction



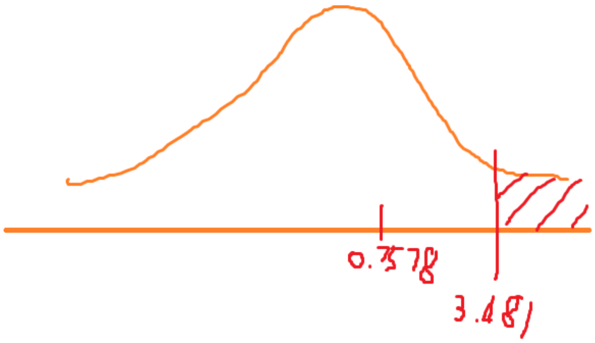
```
	Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 0.60155, df = 1, p-value = 0.438
```

- Using Rstudio with Yates correction



```
[1] "Chi-Square Test Results without Yates' Correction:"
> print(chi_test_no_yates)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 0.75836, df = 1, p-value = 0.3838
```

5. Decision



Since the test statistic value x^2= 0.7578 < critical value ($x^2_{k=1,\ \alpha = 0.05 = 3.841}$), it does not fall within the critical region. Thus, we fail to reject the null hypothesis, H0. Therefore, there is not enough evidence to conclude that there is a relationship between the variables sex and fasting blood sugar, at α = 0.05.

Reflection

1. Choh Jing Yi

I think it was an enlightening experience to work on the heart disease dataset project. I gained knowledge on how to use statistical methods on real data, including regression, chi-square testing, hypothesis testing, and correlation. Since every group member brought something different to the table, the collaborative element helped me strengthen my communication and cooperation abilities. My statistical analysis was enhanced, and my understanding of statistical concepts was reinforced by using RStudio for statistical analysis. Overall, this project has greatly increased my understanding of data analysis and its practical applications in the health sciences.

2. Muhammad Hilmi Hijazi bin Jamal

Working collaboratively on this project has been a delightful experience that has highlighted the value of teamwork and diverse perspectives. Each team member contributed distinct talents and ideas, resulting in a more complete and rigorous study. Our main hurdle was correctly handling file locations and data formats, which we overcame via excellent communication and problem-solving. The process of integrating R for statistical analysis helped us gain a better knowledge of data manipulation and inference. This project demonstrated the value of thorough data processing and the collaborative effort necessary to diagnose and solve technological challenges. As a group, we created a coherent process, improved our technical skills, and got a better understanding of the complexities of data analysis in a real-world setting.

3. Liow Jia Feng

In my opinion, I think this project can enhance my understanding of statistical analysis and its practical applications in health sciences. Working with real data using various statistical methods, including regression, chi-square testing, hypothesis testing and correlation,thus it will allow me to see the real-world implications of these techniques. Besides, the collaborative aspect of the project was equally valuable as each team member brought unique skills and perspectives, fostering effective communication and cooperation. Utilizing RStudio for our analysis will improve my technical skills so I can use this skill in my future. In conclusion, this project has been instrumental in developing both my analytical skills and my appreciation for the complexity and importance of data analysis in medical research.

4. Chua Shang Yeet

It is an interesting project since I am a Bioinformatic student, I am enabled to learn about heart disease and data analysis. This is a fresh experience to face a big data set and do statistical tests with real-world data. There are also some factors that will increase the risk of heart disease, such as smoking, alcohol consumption, eating habits and many more. Public health campaigns focusing on lifestyle modifications could have a substantial impact on reducing the burden of heart diseases. Lastly, my teammates are very cooperative and done their part within the time given.

5.Tay Ching Xian

From this project, I’ve learned a lot of skills, particularly in problem-solving and critical thinking. To be honest, I’m more favored in manual mathematical calculations by using paper and pencils. However, as our data is expanded to 1250 data points, manual calculation becomes impractical. Thus, it prompts me to learn using RStudio to do data analysis. The process of discovering and mastering RStudio was incredibly enjoyable for me and it really enhanced my problem-solving skill. It is really a good experience for me for personal growth and skill development. I'm looking forward to developing more skills in the future.

Conclusion

From the two sample hypothesis tests that have been done, we fail to reject the average of maximum heart rate of men and women are equal. Hence, there is not sufficient evidence to conclude that the maximum heart rate achieved of male patient is different from the mean maximum heart rate achieved of female patient at 95% of confidence level.

From correlation test, we will reject hypothesis that state that there is no linear relationship between the age and the resting blood pressure. Hence, there is sufficient evidence to conclude that there is a linear relationship between the age and resting blood pressure at significant level at 0.05.

For the regression test, we reject the statement that states there is no relation between age and maximum heart rate. Thus, there is sufficient evidence to indicate that age has a statistically significant negative effect on maximum heart rate at significant level of 0.05.

Lastly, the Chi-square of independence test, we fail to reject the null hypothesis, H0 that states that is no relationship between sex and fasting blood sugar. Therefore, there is not enough evidence to conclude that there is a relationship between the variables sex and fasting blood sugar, at α = 0.05.

In conclusion, the factor that increases the heart disease risk is the age of the person. When a human being grows, the risk of the person suffering from heart diseases is higher. All the tests are done by using R-programming and R-studio, and with the knowledges taught by our dear lecturer， Dr. Sharin.

Appendix

Lapp, D. (2019, June 6). *Heart disease dataset*. Kaggle.

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data

Cardiovascular diseases (CVDs)

https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)?gad_source=1&gclid=CjwKCAjwp4m0BhBAEiwAsdc4aIHb_kn6VZnaVftsMgd-SLC_zfrGGOrWWc4okrN5cH0-Rrvew_H4rRoCGeEQAvD_BwE