UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143-08 PROBABILITY & STATISTICAL DATA ANALYSIS

SCHOOL OF COMPUTING
FACULTY OF ENGINEERING

SEMESTER 2 2020/2021

SECTION 08
PROJECT 2 REPORT

GROUP NAME: The Admirals

LIST OF MEMBERS

Name	Matric No.
1. CHONG TUNG HAN	A20EC0028
2. THAM CHUAN YEW	A20EC0166
3. LAI YEE JEN	A20EC0061
4. ZHAO XIN	A20EC4053

LECTURER: DR. SHARIN HAZLIN HUSPI

Introduction

The current age of technology is getting more and more modernized. The risk of giving birth is lower due to the medical technology is better. So, the fertility rate has increased while the death rate during birth is lower. Hence, the population for each of the country is increasing over time, as death rates are now lower. The growth rate of a country usually picked as a reference for their population distribution. However, there are many more factors which can be used to define the country population distribution. Thus, we should consider other factors that may result in the fluctuation of population. The purpose of this project is to investigate the relationship between life expectancy and mortality of the world population. By carrying out this project, we able to understand the relationship between the mentioned variables and how they influence each other.

Methodology

The dataset used in this research is a secondary data that is obtained from eLearning site of University of Technology Malaysia prepared by our lecturer Dr. Sharin Hazlin Huspi. The inference statistical analysis including hypothesis testing 1-sample, correlation, regression and chi-square test of independence is conducted by using RStudio for perform the calculations and obtain the results.

Hypothesis testing is used to test two different hypotheses: null hypothesis and alternative hypothesis regarding a population parameter. The analyst tests a statistical sample to provide evidence on the plausibility of the null hypothesis.

Correlation analysis is used to discover if there is a relationship between two variables/datasets, and how strong that relationship may be. A positive correlation result means that both variables increase in relation to each other, while a negative correlation means that as one variable decreases, the other increases.

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.

Chi-square test of independence is applied to determine whether there is a significant association between two categorical variables from a single population.

Dataset

The dataset used in this research is a secondary data that is obtained from eLearning site of University of Technology Malaysia prepared by our lecturer Dr. Sharin Hazlin Huspi. This dataset approach to the countries in the world. It contains variables such as populations, growth rate, percentage of population below age of 15, life expectancy and mortality. In the raw dataset, it contains 198 countries with 2 duplicate countries which are China and India. We proceed to data cleaning process by removing these 2 duplicate data which located at row 52 and row 114. There are total of 196 countries will be used in this project. The original dataset is made up of 5 variables which consists of population(country population), growth(growth rate of the country), under 15(percentage of population below age of 15), life expectancy(average life span of the population in that country) and mortality(death rate of the country). 2 variables are chosen to be used in this project which are life expectancy and mortality out of those 5 variables. The inference statistical analysis including hypothesis testing 1-sample, correlation, regression and chi-square test of independence is conducted by using RStudio for perform the calculations and obtain the results.

Data Analysis

Hypothesis Testing on 1 sample

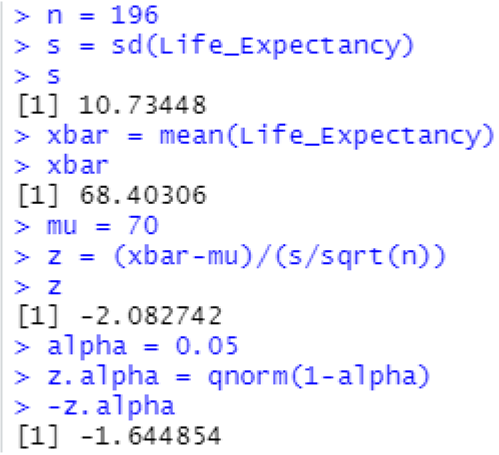
Hypothesis testing on 1 sample is conducted to test whether the mean life expectancy of 196 countries is lower than 70. This test is conducted at the level of significance of 0.05. The population variance is unknown and the sample size is 196.

The null hypothesis, H_0: The population mean of life expectancy is equal to 70.

The alternative hypothesis, H_1: The population mean of life expectancy is less than 70.

$$H_0: \mu = 70$$

$$H_1: \mu < 70$$



```
> n = 196
> s = sd(Life_Expectancy)
> s
[1] 10.73448
> xbar = mean(Life_Expectancy)
> xbar
[1] 68.40306
> mu = 70
> z = (xbar-mu)/(s/sqrt(n))
> z
[1] -2.082742
> alpha = 0.05
> z.alpha = qnorm(1-alpha)
> -z.alpha
[1] -1.644854
```

Figure 1: Calculations of test statistic value

From figure 1, the test statistic value that we obtained is z_{test} = -2.082742 and the critical value is $z_{0.05}$ = -1.644854. Since $z_{test} < z_{0.05}$, we reject the null hypothesis. There is sufficient evidence to conclude that the mean life expectancy of the population is less than 70. Personally, I think that this is because majority of the countries approaches some severe climate change and pandemic outbreak which ends up with lower life expectancy.

Correlation

In this test, we selected a sample of 196 countries and checked if there is any linear correlation between the life expectancy and mortality at 0.05 level of significance.

The null hypothesis, H_0: There is no linear correlation between life expectancy and mortality.

The alternative hypothesis, H_1: Linear correlation exists between life expectancy and mortality.

We reject null hypothesis, H_0 since t_0 = -30.376 < $t_{0.025,194}$ = -1.972268. Therefore, there is sufficient evidence of a linear relationship between life expectancy and mortality at the 5% level of significance.

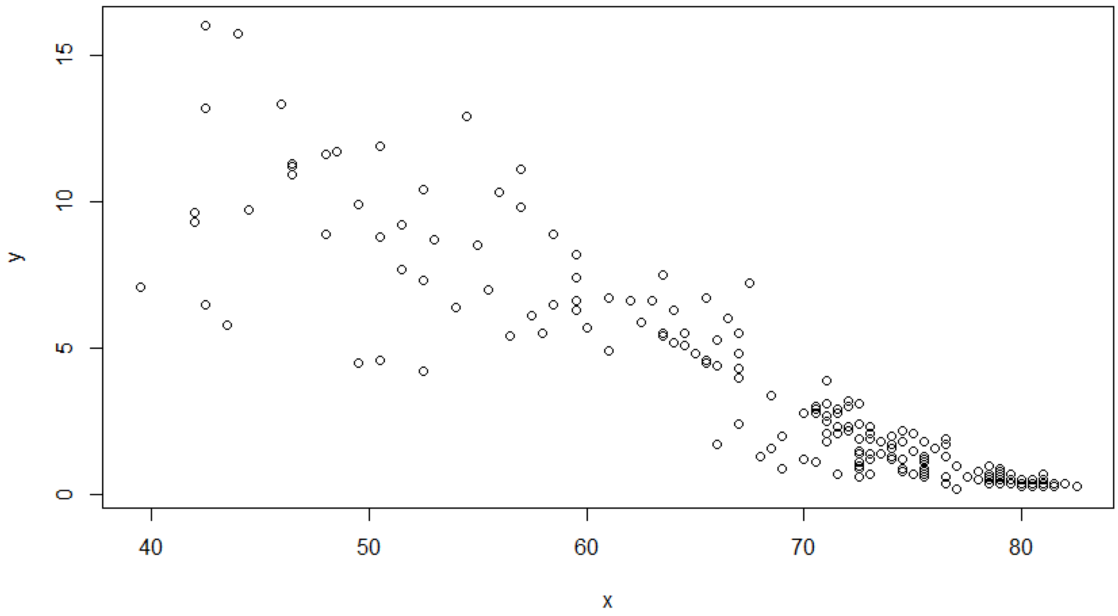


Figure 2 Scatter Plot of Mortality with Life Expectancy

From the scatter plot, it can be seen that mortality decreases as life expectancy increases. Correlation analysis of the data indicates that there is negative relationship between life expectancy and mortality. A value of r = -0.9089952 suggests that it is a strong relationship.

Regression

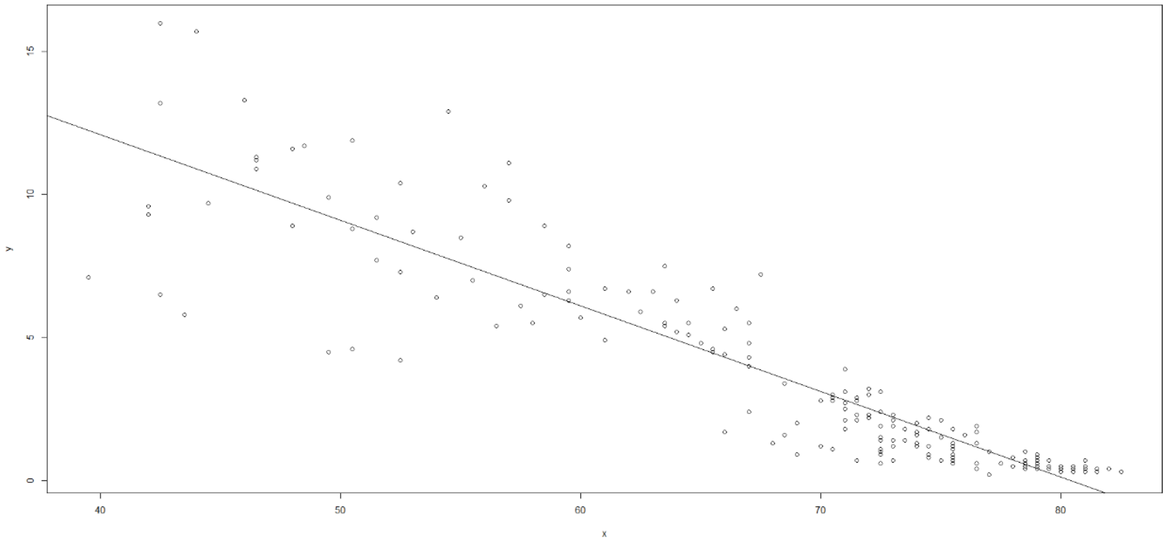


Figure 3 Average life expectancy against average percentage of mortality

In this test, a sample of 196 countries have been selected and we wanted to check whether if the average life expectancy can predict the average percentage of mortality at the significance level of 95%. The dependent variable (y) in this test is the average percentage of mortality whereas the independent variable (x) is the average life expectancy.

The null hypothesis, H_0: $\beta_1 = 0$

The alternative hypothesis, H_1: $\beta_1 \neq 0$

It can be seen from the graph that there is a negative linear relationship between the average life expectancy and the average percentage of mortality with the least squares regression equation of $y = 24.0699 - 0.2995x$.

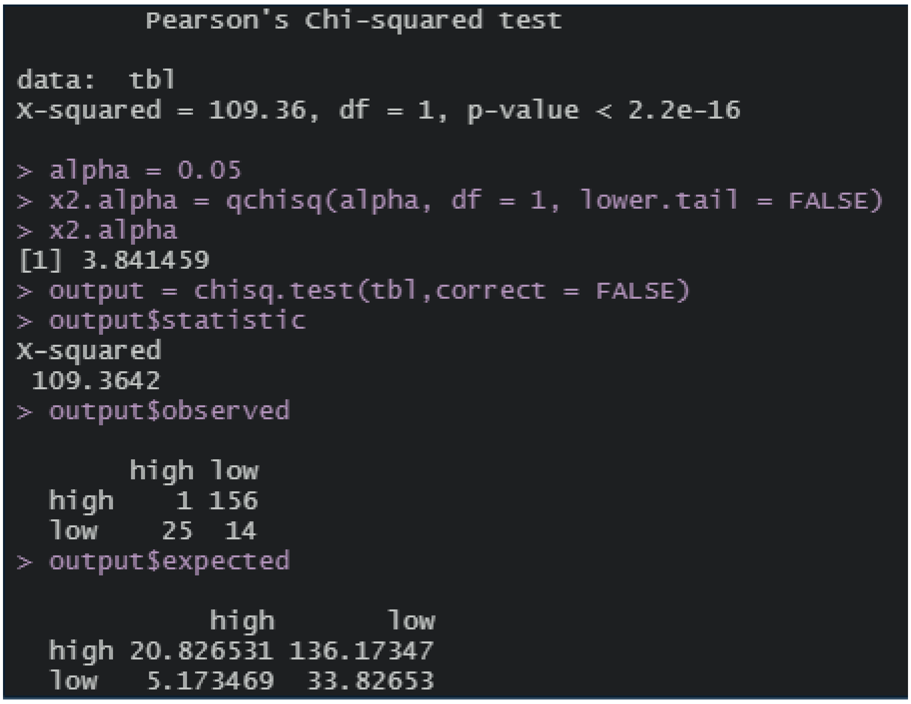
The value of estimate of regression intercept, β_0 is the estimated average value of y when the value of x is zero. In this test, $\beta_0 = 24.0699$ indicates that the estimated average percentage of mortality is 24.0699 without the influence of the average life expectancy. The value of estimate of the regression slope, β_1 is the estimated change in the average value of y as a result of a one-unit change in x. $\beta_1 = - 0.2995$ here shows that the value of y which is the average percentage of mortality will change by - 0.2995 times when there is one-unit change in x which is the average life expectancy. Hence, it can be concluded that the higher the average life expectancy, the lower the average percentage of mortality.

Chi Square Test of Independence

In this project, we conduct the chi square test of independence to investigate the relationship between two variables which are life expectancy and mortality. The life expectancy and mortality are categorized into 2 groups which are “low” and “high”. This test is conducted at the level of significance of 0.05.

The null hypothesis, H_0: The life expectancy is independent of mortality.

The alternative hypothesis, H_1: The life expectancy is not independent of mortality.



```
        Pearson's Chi-squared test

data:  tbl
X-squared = 109.36, df = 1, p-value < 2.2e-16

> alpha = 0.05
> x2.alpha = qchisq(alpha, df = 1, lower.tail = FALSE)
> x2.alpha
[1] 3.841459
> output = chisq.test(tbl,correct = FALSE)
> output$statistic
X-squared
 109.3642
> output$observed

       high low
  high    1 156
  low    25  14
> output$expected

            high       low
  high 20.826531 136.17347
  low   5.173469  33.82653
```

Figure 4: Calculations and two-way contingency table

From figure 4, we obtain the test statistic value is χ^2_{test} = 109.3642, degree of freedom = (2-1)(2-1) = 1 and the critical value $\chi^2_{0.05,1}$ = 3.841459. Since $\chi^2_{test} > \chi^2_{0.05,1}$, we reject the null hypothesis. There is sufficient evidence to conclude that the life expectancy is not independent of mortality. In other words, there is a relationship between life expectancy and mortality. For instance, a country with good healthcare supply is potentially having less casualties. Undoubtedly, the higher the life expectancy exists in a country reflects a lower mortality.

Conclusion

To conclude our project, we manage to choose a dataset and reprocess on it so that it fits our requirements for the analysis process. For the hypothesis testing on 1 sample, the mean life expectancy of the population is less than 70. For the correlation analysis, it is shown that life expectancy has a strong negative relationship with mortality. Meanwhile for the regression analysis, it can be seen that there is a negative linear relationship between the average life expectancy and the average percentage of mortality. It indicates that the country that have higher average life expectancy will have a lower average percentage of mortality. Finally, the chi square test of independence depicts that the life expectancy is not independent of the mortality. In short, the aim of this project is achieved as the relationship between the variables has been determined. Along with that, the discoveries of this project can be used as a guidance for a country to determine the factors affecting the life expectancy of their country.