STA238 Final Project

Name: Qiaoyu Wang & Zhanfei Gu

Due Date: April 19th

Introduction and Research Question

Nowadays, countries around the world pay attention to improving national health levels and average life expectancy, and the national healthcare system is receiving increasing attention. Over the past 15 years, great progress has been made in the health sector in all countries, with a decline in human mortality compared to the previous 30 years, especially in developing countries. The dataset we used comes from the Global Health Observatory (GHO) data repository at the World Health Organization (WHO), which tracks health in countries and many other relevant factors. The data include life expectancy and related factors affecting it for 193 countries from 2000 to 2015, which are recorded annually. Based on the dataset, we focus on the relationship between 193 countries' healthcare expenditure and the average lifespan.

Our study question is "is increased health expenditure(Total expenditure %) associated with increased lifespan (Life expectancy)". We will analyze and compare healthcare expenditures and people's average lifespan of different status countries (both developing and developed countries). Then, we will use three statistical methods, which are goodness of fit test, confidence interval and two group hypothesis tests to perform appropriate exploratory data analysis on our selected variables. The purpose of our study is to give some suggestions to all governments in the world about whether they should increase the expenditure in the healthcare field. The conclusion has practical implications for future national healthcare expenditure plans for all status countries.

Variable and Data Cleaning

The data was collected from WHO and the United Nations website with the help of Deeksha Russell and Duan Wang. In our analysis, the population is all countries over the world. We mainly use 5 variables which are **Country, Year, Status, Life expectancy, Total expenditure(%)**.

- ➔ **Country** means the 193 countries selected from the database.
- ➔ **Year** means the time of data recorded.
- ➔ **Status** means the development of selected countries (Developed or Developing status).
- ➔ **Life expectancy** means the life expectancy of countries in age.
- ➔ **Total expenditure** means the general government expenditure on health as a percentage of total government expenditure(%).

Since the dataset is from WHO, we found no obvious errors of data, but there is some missing data on life expectancy, total expenditure, and GDP for some years. In order to analyze data using R chunks, we need to do **data cleaning** first.

*(*See R chunks in Appendix A)*

We remove missing values on the variables Life Expectancy, Total expenditure and GDP, and only keep the 5 variables (the variables mentioned above) in the final dataset. Since too much data is lost in some years, we finally choose to only keep the data for 2014 and then only 145 countries are left.

Exploratory Data Analysis —— EDA

	Min	1st quantile	Median	Mean	3rd quantile	Max
Total expenditure(%)	1.21	4.38	5.79	6.072	7.56	13.73
Life expectancy (age)	48.1	65.5	73.3	71.07	76.4	89

Reason: The table above is numerical summaries *(*See R chunks in Appendix B-1)* of total expenditure(%) and life expectancy. We will use these statistics in three methods.

Conclusion: From the table, we find that the mean of total expenditure(%) and life expectancy are 6.072 billion dollars and 71.07 years old, respectively.

Histogram of Life Expectancy for 145 countries

Reason: Since our study question is about Total expenditure(%) and Life expectancy, we want to have an overview of life expectancy for 145 countries first. Hence, we plot a histogram below (Figure 1). *(*See R chunks in Appendix B-2.1)*

Conclusion: The histogram shows the distribution of life expectancy is left-skewed with a tail towards the left side. The spread of life expectancy is from 45 years old to 90 years old. The histogram attains a peak (mode) towards the right side at about 75 years old. It can be seen that the mean of life Expectancy is less than mode and it is about 70 years old.

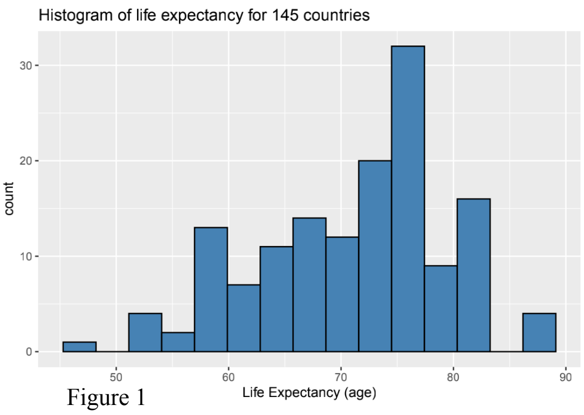


Figure 1

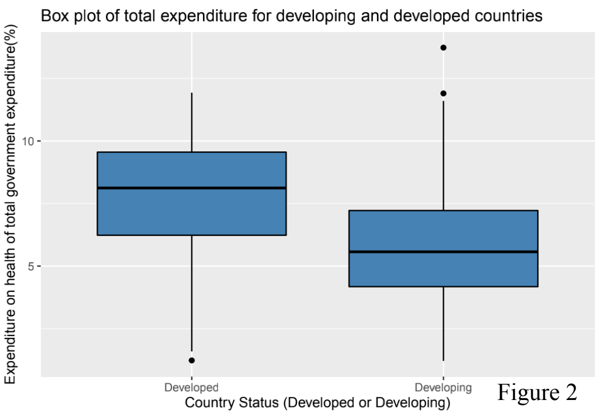


Figure 2

Side-by-side Boxplot of Total Expenditure for Developing and Developed Countries

Reason: Then we pay attention to the variable: Total Expenditure(%). We want to find out the difference between total expenditure(%) in developing countries and developed countries, so we plot a boxplot above (Figure 2). *(*See R chunks in Appendix B-2.2)*

Conclusion: The boxplot shows that developed countries are slightly right skewed, while developing countries are almost symmetrical shapes. Developed countries (about 8%) have a higher median of Total Expenditure(%) than developing countries (about 6%). The spread of health expenditure on developed countries and developing countries are almost the same. It's clear that developed countries have a high median even higher than the 3rd quantile of developing countries, which shows the developed countries have higher expenditure in healthcare compared with developing countries.

Scatterplot Between Total Expenditure and Life Expectancy

Reason: Since our study question is about Total expenditure(%) and Life expectancy, we draw a scatter plot to explore the association between them (Figure 3).

*(*See R chunks in Appendix B-2.3)*

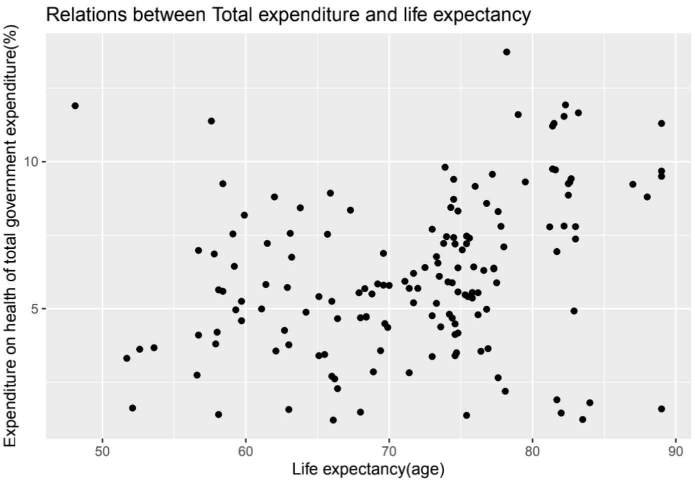


Figure 3

Conclusion: From the graph above, we find the strength is moderate and with a positive direction. The association is kind of linear, but there are still some outliers that are not on the line. Also from the graph we find it doesn't violate the constant variance assumption which implies that there is a linear relationship between life expectancy(age) and total expenditure(%).

Method 1 (Goodness of Fit Test) and Findings

Reason: The goodness of fit test for a multinomial distribution can be adapted to test generally whether a data set comes from a specific underlying distribution. In our analysis, we want to figure out whether the variable Total expenditure(%) follows Normal distribution, this is an assumption that must be met in our follow-up research.

Process: **See detailed steps for this method in R chunks in Appendix C*

Step 1: Set up the null and alternative hypotheses.

Null hypothesis (H_0):The variable Total Expenditure(%) follows Normal distribution.

Alternative hypothesis (H_a): The variable Total Expenditure(%) does not follow the Normal distribution.

Step 2: Calculate the sample mean and standard deviation of Total expenditure(%) . Standardized all the data points of variable Total expenditure (%) by using standardized formulas.

Step 3: Create 5 equal-probability intervals such that each interval represents 20 percentile of a normal distribution and find the boundary value for each interval above under normal distribution by using R function qnorm().

Step 4: According to the value of each data point of the Total variable, allocate them into correct intervals above and count the number of datapoints in each interval.

Step 5: Count the number of data points for variable Total expenditure(%) and spread them evenly into 5 different intervals such that there are equal numbers of data points in each interval.

Step 6: Calculate the test statistic Q by using its formula since Q follows chi-square distribution then we can now apply the R function pchisq and get the p-value.

Discuss finding:

After implementing the above steps, we get the p-value is 0.897166 which is close to 1. For the p-value > 0.1, we can make a conclusion that there is no evidence against the null hypotheses(H_0). That means that **the variable Total expenditure(%) does follow the Normal distribution.**

Contribution to our study question:

This method doesn't have a direct contribution to our study question, but it proves the assumption that method 2 needs to hold.

Method 2 (Confidence Interval) and Findings

Reason: We plan to use the confidence interval because we want to study the average interval of total expenditure(%) for all countries on the dataset.

Process: **See detailed steps for this method in R chunks in Appendix D*

Step 1: Calculate the sample mean, standard deviation and sample size of the variable Total expenditure(%) by using R.

Step 2: Calculate the critical value by using R function qt().

Step 3: With the assumption proved by method 1, we now can apply the t/z approach directly and calculate the confidence interval.

Step 4: From step 3, we found the confidence interval is (5.77427, 6.679562). Next we separate our dataset into two groups. First group consists of data with a value of Total expenditure(%) that is less or equal to 5.77427. Second group consists of data with a value of Total expenditure(%) that is bigger or equal to 6.679562.

Step 5: Calculate the mean of life expectancy for each group above and compare.

Discuss finding:

We finally build a 98% confidence interval and its value is (5.722427, 6.679562), which interprets that we are 98% confident that the mean of the general government expenditure on health for all countries in the world as a percentage is between 5.722427 and 6.679562 .

Besides, we find the average life expectancy for the first group is 68.87 years old and average life expectancy for the second group is 74.83 years old. Something interesting here is that countries with high general expenditure on health as a percentage corresponds to high life expectancy. This indicates there exists a positive association between the variable Total expenditure(%) and Life expectancy(age).

Contribution to our study question:

The confidence interval helps us separate the data set and find the association between the Total expenditure(%) and Life expectancy(age).

Method 3 (Two group hypothesis tests) and Findings

Reason: Since we want to know the difference of Total expenditure(%) between developed and developing countries, we use the two-group hypothesis test.

Process: **See detailed steps for this method in R chunks in Appendix E-1*

Step 1: Set up the null and alternative hypotheses.

Null hypothesis (H_0): $\mu_1 = \mu_2$

Alternative hypothesis (H_a): $\mu_1 > \mu_2$

μ_1 represents the mean of Total Expenditure(%) for developing countries.

μ_2 represents the mean of Total Expenditure(%) for developing countries.

Step 2: Separate our dataset into two groups. Group 1 only contains data for developing countries. Group 2 only contains data for developed countries.

Step 3: Calculate the sample mean standard deviation, and sample size of Total expenditure(%) and Life expectancy(age) for each group using R.

Step 4: Calculate the test statistic and degrees of freedom by using formulas.

Step 5: Calculate the p-value using the statistic from step 4.

Discuss finding:

We find the p-value is 0.0006228666, which is close to 0. Since p-value is less than 0.001, we have strong evidence against the null hypothesis, which means the mean of Total Expenditure(%) for developed countries is bigger than the mean of Total Expenditure(%) for developing countries.

Contribution to our study question:

**See detailed steps for this method in R chunks in Appendix E-2.1 & E-2.2*

Also we find for developed countries, their average life expectancy is 81.14 years old. For developing countries, their average life expectancy is 69.58 years old. The interesting fact is that high Total Expenditure(%) corresponds to a relatively high Life expectancy(age) for developed countries. For developing countries, low Total Expenditure(%) corresponds to a relatively low Life expectancy. This implies there exists a positive association between the variable Total expenditure(%) and Life expectancy(age).

Linear Regression Model

Reason: Linear regression uses linear equations to model the relationship between one or more independent and dependent variables. Since our study question is about Total expenditure(%) and Life expectancy, we will use linear regression to verify whether there is a linear relationship between two variables.

Process: **See detailed steps for this method in R chunks in Appendix F-1*

Step 1: Use variable Total expenditure and Life expectancy to model the fitted regression line equation and find the Coefficient of R^2 .

Step 2: Set up a hypothesis test for $\widehat{\beta_1}$ (which represents the slope of the fitted regression line equation).

Null hypothesis (H_0): $\widehat{\beta_1} = 0$; Alternative hypothesis (H_a): $\widehat{\beta_1} \neq 0$

Step 3: Build the confidence interval for $\widehat{\beta_0}$ (which represents the intercept of the fitted regression line equation) and $\widehat{\beta_1}$.

Discuss finding:

**See detailed steps for this method in R chunks in Appendix F-2 & F-3*

We find the fitted regression line equation is: $\widehat{y_i}$**= 69.5846+0.9845**x_i , which means $\widehat{\beta_0}$ = 69.5846 and $\widehat{\beta_1}$ = 0.9845. Here $\widehat{y_i}$ represents the estimated Life expectancy and x_i represents the Total expenditure(%). When the Total expenditure is 0 the estimated life expectancy is 69.5846 years old. (However the situation that the Total expenditure = 0 is unusual thus this estimated life expectancy is not important)

Also when the Total expenditure(%) increases by 1 percent then the life expectancy will increase 0.9845 years, so higher Total expenditure likely generate high life expectancy.

We get R^2=0.08901, which is quite low. It interprets that only 8.9% of Life expectancy can be explained by Total expenditure(%).

We get the p-value is 0.000171 which is close to 0. Since p-value is less than 0.001, we have strong evidence against the null hypothesis, which means $\widehat{\beta_1}$ ≠ 0. Since $\widehat{\beta_1}$ ≠ 0, there's a linear relationship between Total expenditure(%) and Life expectancy.

We get two 98% confidence interval of $\widehat{\beta_0}$ (-85954.32, -34462.68) and $\widehat{\beta_1}$ (624.5984, 1338.4016) , which mean $\widehat{\beta_0}$ will float up and down between -85954.32 and -34462.68 and $\widehat{\beta_1}$ will float up and down between 624.5984 and 1338.4016. We can easily see that both $\widehat{\beta_1}$ and $\widehat{\beta_0}$ fluctuate which means they are not stable and will easily change when we change the sample.

Conclusion

From the method of goodness of fit test, we conclude that the variable fits the Normal distribution well. Then we build a 98% confidence interval and conclude that we are 98% confident that the mean of the general government expenditure on health for all countries in the world as a percentage is between 5.722427 and 6.679562. We also find countries with Total expenditure bigger than 6.679562 have a longer Life expectancy compared with countries with Total expenditure less than 5.722427. Therefore, we make a conclusion that there exists a positive association between variables Total expenditure and Life expectancy.

In the last two group hypothesis test methods, we found the mean of Total expenditure for developed countries is bigger than the Total expenditure for developing countries. We also find the mean of Life expectancy for developed countries is longer than Life expectancy for developing countries. That again implies there exists a positive association between variable Total expenditure and Life expectancy.

Finally, in our fitted regression line equation we found that when the Total expenditure(%) increases by 1 percent then the life expectancy will increase 0.9845 years, which means higher Total expenditure is likely to generate higher life expectancy. Overall, we can answer our study question by concluding increased health expenditure will be associated with increased lifespan.

Limitation

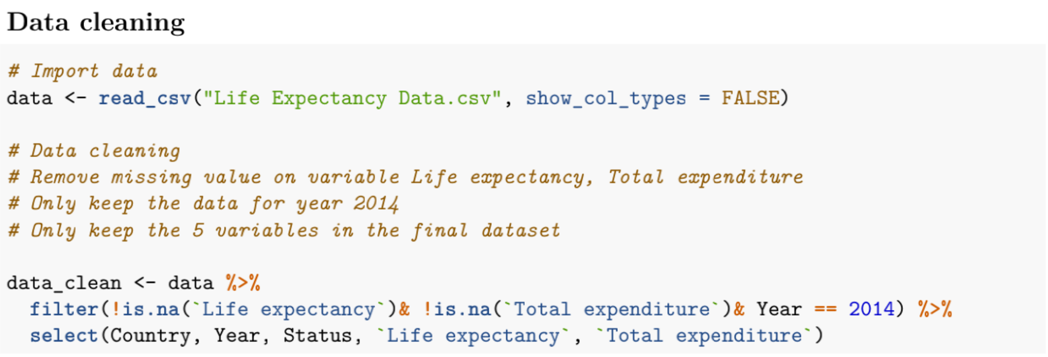
Although our analysis circumvents many potential errors, there are still some unavoidable limitations, and we have reflected on our analysis to get some summaries below.

The sample size is not large enough. After data cleaning, our sample size is only 145 countries, but our populations are analyzed for all countries in the world. Countries that are not in the database didn’t have the similar total healthcare expenditure(%) of the 145 countries, so the samples are not enough to represent all countries in the world. Therefore, the conclusions we obtained are limited and cannot be generalized to a wider range.

Besides, the coefficient of determination R^2 is extremely low. It means the variable Total expenditure can not predict Life expectancy well, so the validity of linear regression model is doubtful.

Appendix

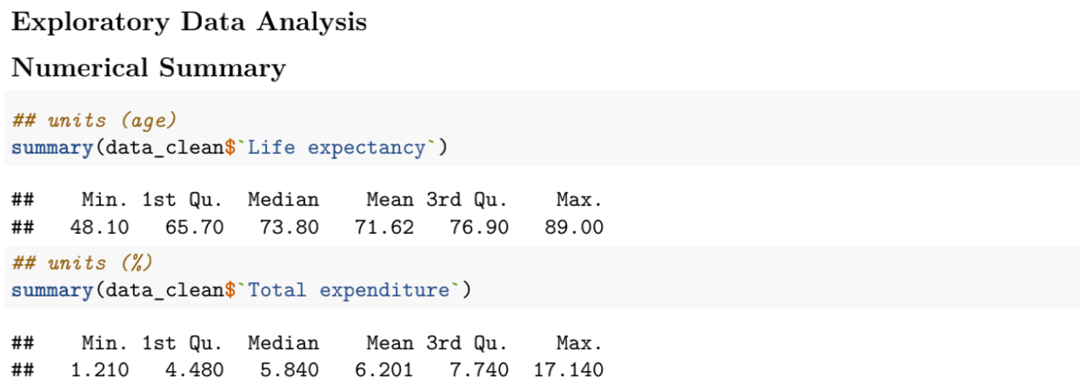
Appendix A:



Data cleaning

```
# Import data
data <- read_csv("Life Expectancy Data.csv", show_col_types = FALSE)

# Data cleaning
# Remove missing value on variable Life expectancy, Total expenditure
# Only keep the data for year 2014
# Only keep the 5 variables in the final dataset

data_clean <- data %>%
  filter(!is.na(`Life expectancy`)& !is.na(`Total expenditure`)& Year == 2014) %>%
  select(Country, Year, Status, `Life expectancy`, `Total expenditure`)
```

Appendix B-1:

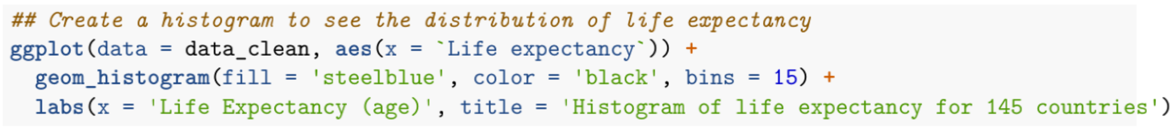


Exploratory Data Analysis

Numerical Summary

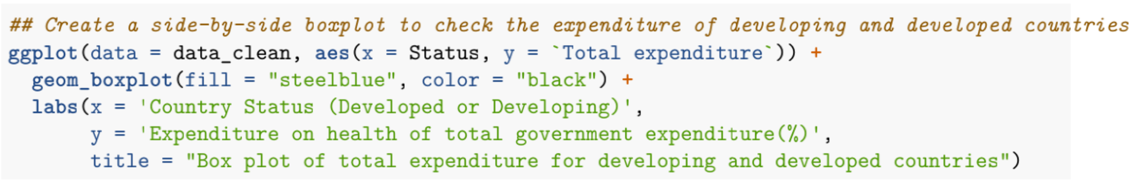
```
## units (age)
summary(data_clean$`Life expectancy`)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.10   65.70   73.80   71.62   76.90   89.00

## units (%)
summary(data_clean$`Total expenditure`)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.210   4.480   5.840   6.201   7.740  17.140
```

Appendix B-2.1:



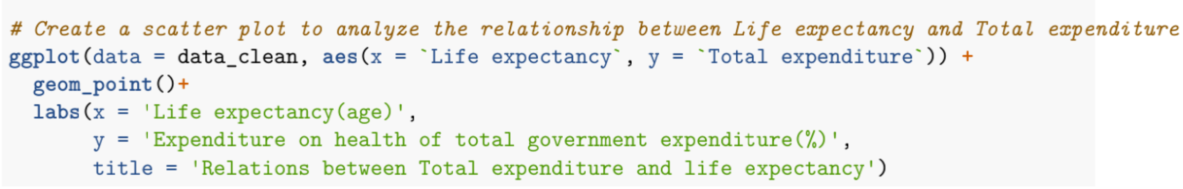
```
## Create a histogram to see the distribution of life expectancy
ggplot(data = data_clean, aes(x = `Life expectancy`)) +
  geom_histogram(fill = 'steelblue', color = 'black', bins = 15) +
  labs(x = 'Life Expectancy (age)', title = 'Histogram of life expectancy for 145 countries')
```

Appendix B-2.2:



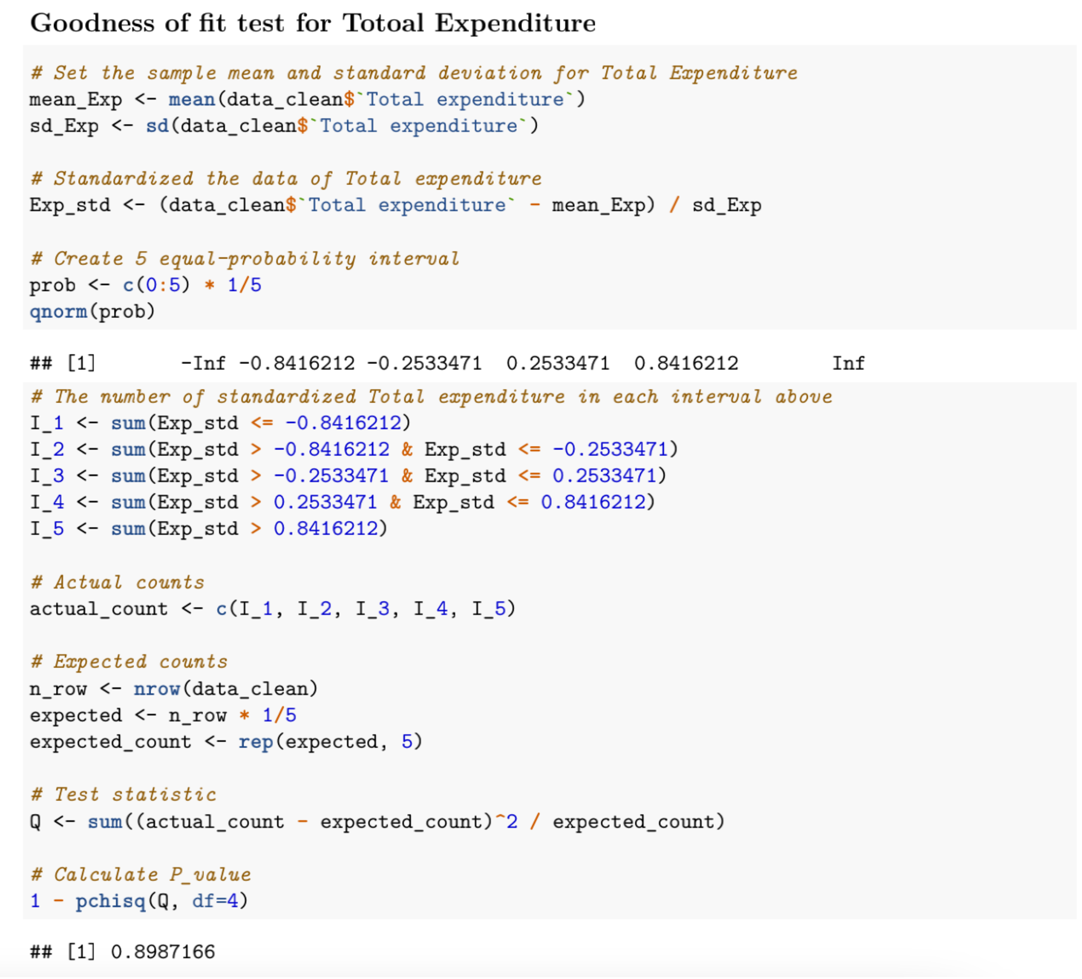
```
## Create a side-by-side boxplot to check the expenditure of developing and developed countries
ggplot(data = data_clean, aes(x = Status, y = `Total expenditure`)) +
  geom_boxplot(fill = "steelblue", color = "black") +
  labs(x = 'Country Status (Developed or Developing)',
       y = 'Expenditure on health of total government expenditure(%)',
       title = "Box plot of total expenditure for developing and developed countries")
```

Appendix B-2.3:



```
# Create a scatter plot to analyze the relationship between Life expectancy and Total expenditure
ggplot(data = data_clean, aes(x = `Life expectancy`, y = `Total expenditure`)) +
  geom_point()+
  labs(x = 'Life expectancy(age)',
       y = 'Expenditure on health of total government expenditure(%)',
       title = 'Relations between Total expenditure and life expectancy')
```

Appendix C:



Goodness of fit test for Totoal Expenditure

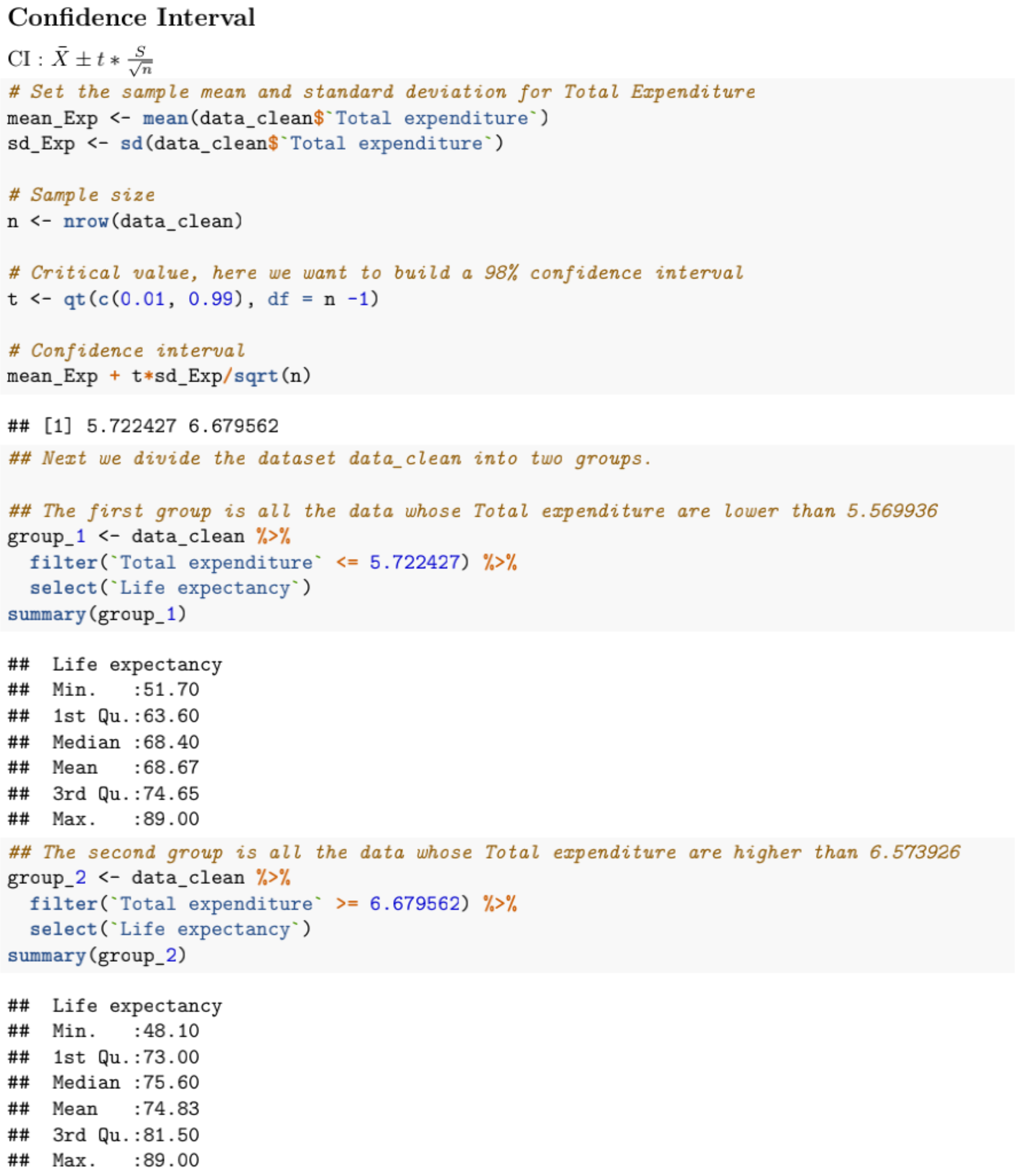
```
# Set the sample mean and standard deviation for Total Expenditure
mean_Exp <- mean(data_clean$`Total expenditure`)
sd_Exp <- sd(data_clean$`Total expenditure`)

# Standardized the data of Total expenditure
Exp_std <- (data_clean$`Total expenditure` - mean_Exp) / sd_Exp

# Create 5 equal-probability interval
prob <- c(0:5) * 1/5
qnorm(prob)
```

```
## [1]       -Inf -0.8416212 -0.2533471  0.2533471  0.8416212        Inf
```

```
# The number of standardized Total expenditure in each interval above
I_1 <- sum(Exp_std <= -0.8416212)
I_2 <- sum(Exp_std > -0.8416212 & Exp_std <= -0.2533471)
I_3 <- sum(Exp_std > -0.2533471 & Exp_std <= 0.2533471)
I_4 <- sum(Exp_std > 0.2533471 & Exp_std <= 0.8416212)
I_5 <- sum(Exp_std > 0.8416212)

# Actual counts
actual_count <- c(I_1, I_2, I_3, I_4, I_5)

# Expected counts
n_row <- nrow(data_clean)
expected <- n_row * 1/5
expected_count <- rep(expected, 5)

# Test statistic
Q <- sum((actual_count - expected_count)^2 / expected_count)

# Calculate P_value
1 - pchisq(Q, df=4)
```

```
## [1] 0.8987166
```

Appendix D:



Confidence Interval

CI : $\bar{X} \pm t * \frac{S}{\sqrt{n}}$

```
# Set the sample mean and standard deviation for Total Expenditure
mean_Exp <- mean(data_clean$`Total expenditure`)
sd_Exp <- sd(data_clean$`Total expenditure`)

# Sample size
n <- nrow(data_clean)

# Critical value, here we want to build a 98% confidence interval
t <- qt(c(0.01, 0.99), df = n -1)

# Confidence interval
mean_Exp + t*sd_Exp/sqrt(n)
```

```
## [1] 5.722427 6.679562
```

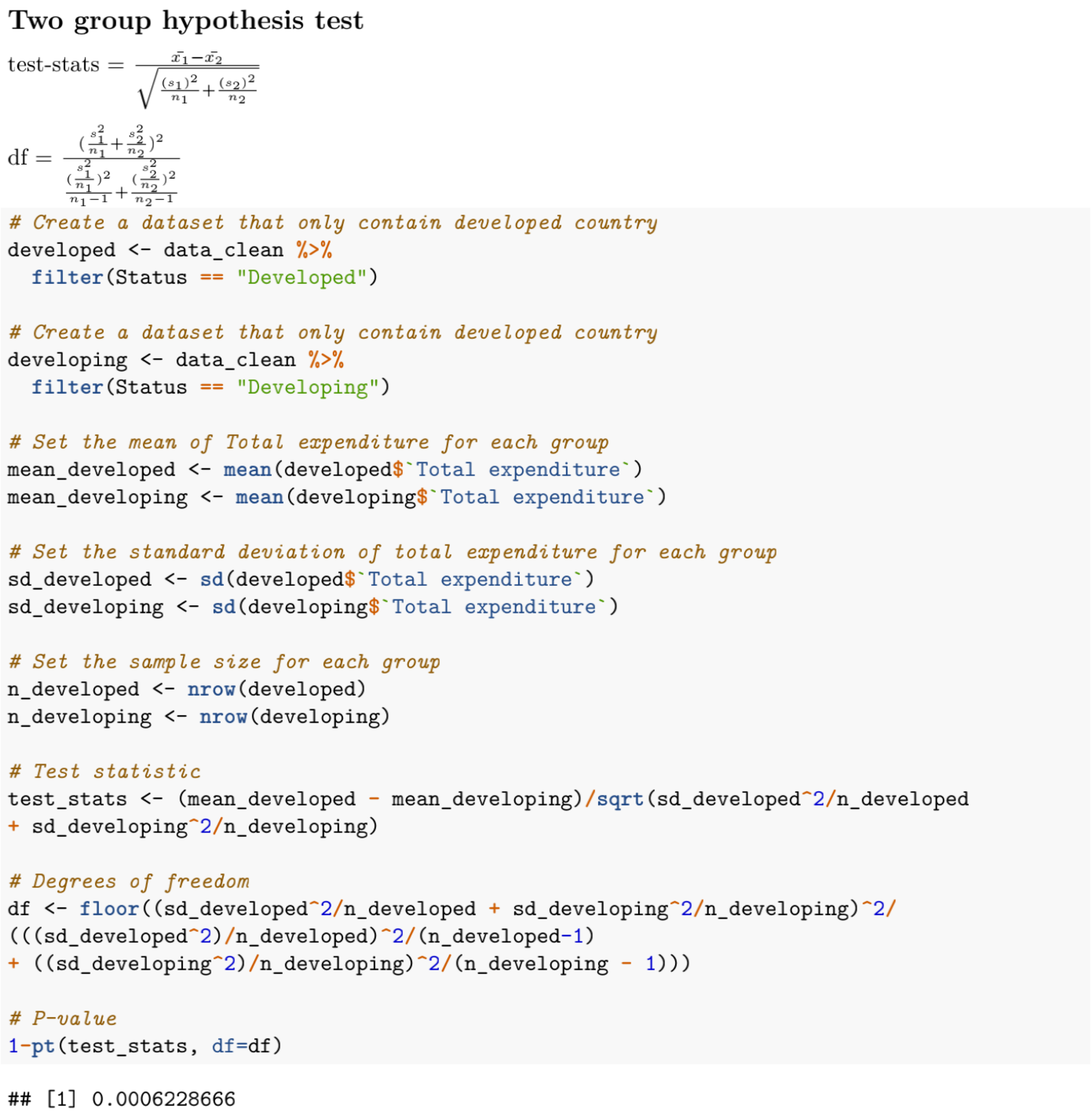
```
## Next we divide the dataset data_clean into two groups.

## The first group is all the data whose Total expenditure are lower than 5.569936
group_1 <- data_clean %>%
  filter(`Total expenditure` <= 5.722427) %>%
  select(`Life expectancy`)
summary(group_1)
```

```
##  Life expectancy
##  Min.   :51.70
##  1st Qu.:63.60
##  Median :68.40
##  Mean   :68.67
##  3rd Qu.:74.65
##  Max.   :89.00
```

```
## The second group is all the data whose Total expenditure are higher than 6.573926
group_2 <- data_clean %>%
  filter(`Total expenditure` >= 6.679562) %>%
  select(`Life expectancy`)
summary(group_2)
```

```
##  Life expectancy
##  Min.   :48.10
##  1st Qu.:73.00
##  Median :75.60
##  Mean   :74.83
##  3rd Qu.:81.50
##  Max.   :89.00
```

Appendix E-1:



Two group hypothesis test

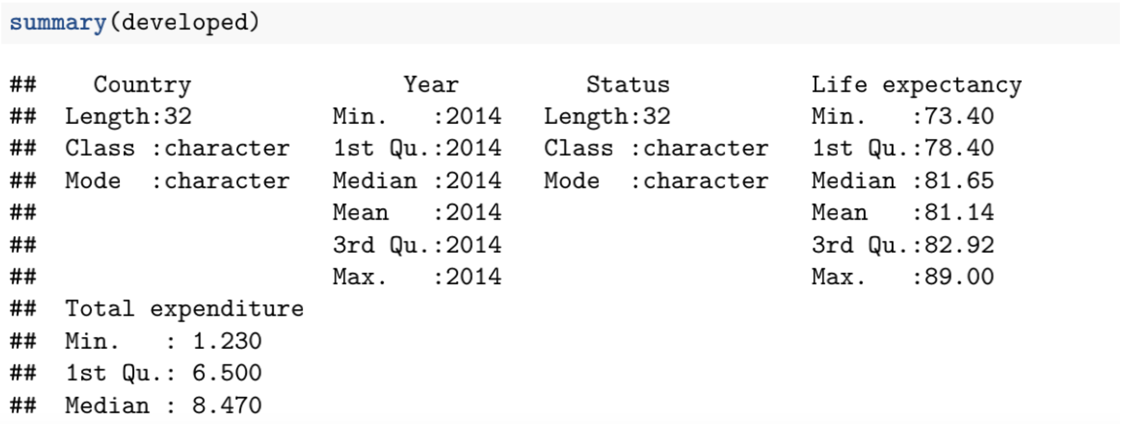
test-stats $= \frac{\bar{x_1}-\bar{x_2}}{\sqrt{\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}}}$

df $= \frac{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1}+\frac{(\frac{s_2^2}{n_2})^2}{n_2-1}}$

```
# Create a dataset that only contain developed country
developed <- data_clean %>%
  filter(Status == "Developed")

# Create a dataset that only contain developed country
developing <- data_clean %>%
  filter(Status == "Developing")

# Set the mean of Total expenditure for each group
mean_developed <- mean(developed$`Total expenditure`)
mean_developing <- mean(developing$`Total expenditure`)

# Set the standard deviation of total expenditure for each group
sd_developed <- sd(developed$`Total expenditure`)
sd_developing <- sd(developing$`Total expenditure`)

# Set the sample size for each group
n_developed <- nrow(developed)
n_developing <- nrow(developing)

# Test statistic
test_stats <- (mean_developed - mean_developing)/sqrt(sd_developed^2/n_developed
+ sd_developing^2/n_developing)

# Degrees of freedom
df <- floor((sd_developed^2/n_developed + sd_developing^2/n_developing)^2/
(((sd_developed^2)/n_developed)^2/(n_developed-1)
+ ((sd_developing^2)/n_developing)^2/(n_developing - 1)))

# P-value
1-pt(test_stats, df=df)
```

```
## [1] 0.0006228666
```

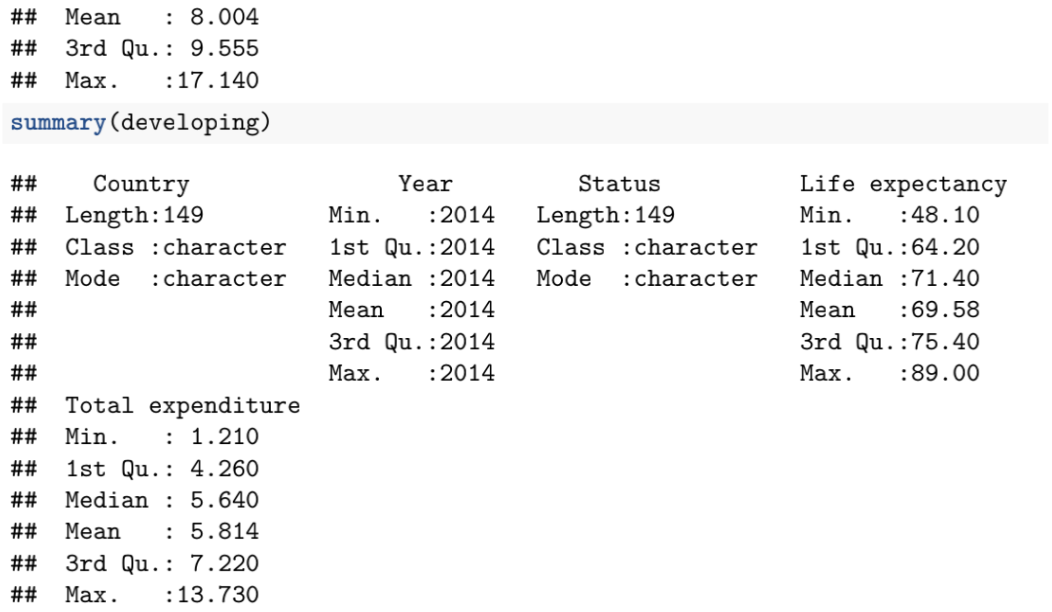
Appendix E-2.1:



```
summary(developed)
```

```
##    Country               Year         Status             Life expectancy
##  Length:32          Min.   :2014   Length:32          Min.   :73.40
##  Class :character   1st Qu.:2014   Class :character   1st Qu.:78.40
##  Mode  :character   Median :2014   Mode  :character   Median :81.65
##                     Mean   :2014                      Mean   :81.14
##                     3rd Qu.:2014                      3rd Qu.:82.92
##                     Max.   :2014                      Max.   :89.00
##  Total expenditure
##  Min.   : 1.230
##  1st Qu.: 6.500
##  Median : 8.470
```

Appendix E-2.2:

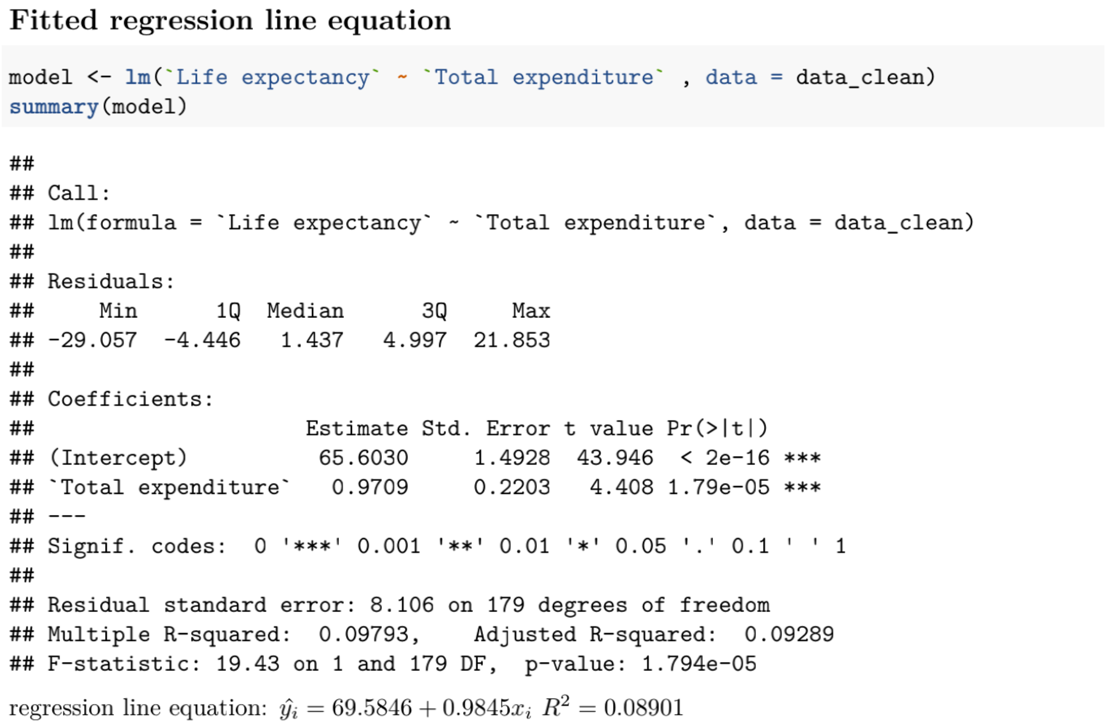


```
##  Mean   : 8.004
##  3rd Qu.: 9.555
##  Max.   :17.140
```

```r
summary(developing)
```

```
##    Country               Year          Status           Life expectancy
##  Length:149         Min.   :2014   Length:149         Min.   :48.10
##  Class :character   1st Qu.:2014   Class :character   1st Qu.:64.20
##  Mode  :character   Median :2014   Mode  :character   Median :71.40
##                     Mean   :2014                      Mean   :69.58
##                     3rd Qu.:2014                      3rd Qu.:75.40
##                     Max.   :2014                      Max.   :89.00
##  Total expenditure
##  Min.   : 1.210
##  1st Qu.: 4.260
##  Median : 5.640
##  Mean   : 5.814
##  3rd Qu.: 7.220
##  Max.   :13.730
```

Appendix F-1:



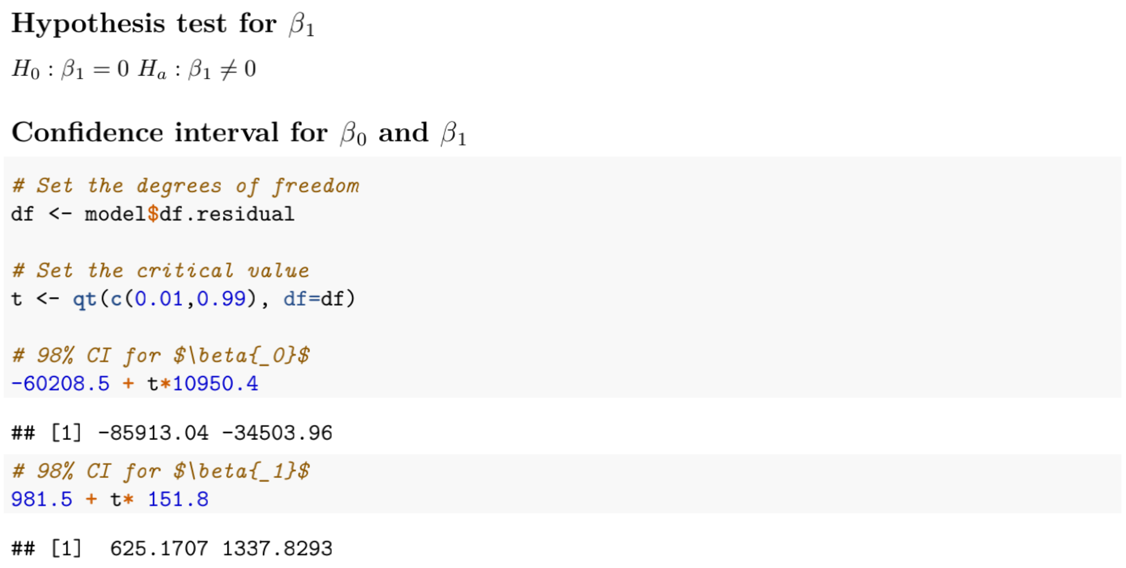
Fitted regression line equation

```r
model <- lm(`Life expectancy` ~ `Total expenditure` , data = data_clean)
summary(model)
```

```
##
## Call:
## lm(formula = `Life expectancy` ~ `Total expenditure`, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.057  -4.446   1.437   4.997  21.853
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          65.6030     1.4928  43.946  < 2e-16 ***
## `Total expenditure`   0.9709     0.2203   4.408 1.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.106 on 179 degrees of freedom
## Multiple R-squared:  0.09793,    Adjusted R-squared:  0.09289
## F-statistic: 19.43 on 1 and 179 DF,  p-value: 1.794e-05
```

regression line equation: $\hat{y}_i = 69.5846 + 0.9845x_i$ $R^2 = 0.08901$

Appendix F-2:



Hypothesis test for β_1

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$

Confidence interval for β_0 and β_1

```
# Set the degrees of freedom
df <- model$df.residual

# Set the critical value
t <- qt(c(0.01,0.99), df=df)

# 98% CI for $\beta{_0}$
-60208.5 + t*10950.4
```

```
## [1] -85913.04 -34503.96
```

```
# 98% CI for $\beta{_1}$
981.5 + t* 151.8
```

```
## [1]  625.1707 1337.8293
```

Citations

KumarRajarshi. (2018, February 10). Life expectancy (WHO). Kaggle. Retrieved March 21, 2022, from https://www.kaggle.com/kumarajarshi/life-expectancy-who