PROJECT 2

LIFE EXPECTANCY

GROUP 4 MEMBERS

CHANG WEN XUEN	(A23CS5012)
LIM CHEN XI	(A23CS0103)
FOO MING KUANG	(A23CS5026)
KAREN VOON XIU WEN	(A23CS0229)

INTRODUCTION

The primary objective of this study is to identify and analyze the key determinants of life expectancy using a dataset from the World Health Organization (WHO) and supplementary economic data from the United Nations. By integrating various health, economic, social, and immunization-related factors, we aim to develop a predictive model that can help countries understand and improve their life expectancy rates. This comprehensive approach allows for a deeper insight into how different variables interact and contribute to life expectancy across diverse global contexts.

DATA DESCRIPTION

Variables (Description)	**Type of Variable**	**Measurement Level**
Status	Qualitative	Nominal
Life expectancy	Quantitative	Ratio
Adult Mortality	Quantitative	Ratio
Hepatitis B	Quantitative	Ratio
BMI	Quantitative	Ratio
HIV/AIDS	Quantitative	Ratio
GDP	Quantitative	Ratio

STATISTICAL TEST ANALYSIS

Selected Variables	Objectives	Test Analysis
Life expectancy	To test whether the average life expectancy in the dataset is significantly different from a global benchmark	Hypothesis Test: One Sample t-Test
Life expectancy, GDP	To determine the strength and direction of the relationship between life expectancy and GDP.	Correlation Test: Pearson Correlation
Life expectancy, Adult Mortality, BMI, GDP, HIV/AIDS	To predict life expectancy based on adult mortality, BMI, GDP, and HIV/AIDS.	Regression Test: Multiple Linear Regression
Life expectancy, Status	To test if there are significant differences in life expectancy between developing and developed countries.	ANOVA test developing and developed countries.
Status, Hepatitis B immunization coverage	To determine if there is an association between a country's development status and its Hepatitis B immunization coverage.	Chi-Square Test of Independence

HYPOTHESIS TESTING: ONE SAMPLE T-TEST

To determine whether the average life expectancy in the dataset is significantly different from a global benchmark.

Hypothesis statement:

$H_0: \mu = 70$

$H_1: \mu \neq 70$

VARIANCE UNKNOWN, T-TEST FORMULA:

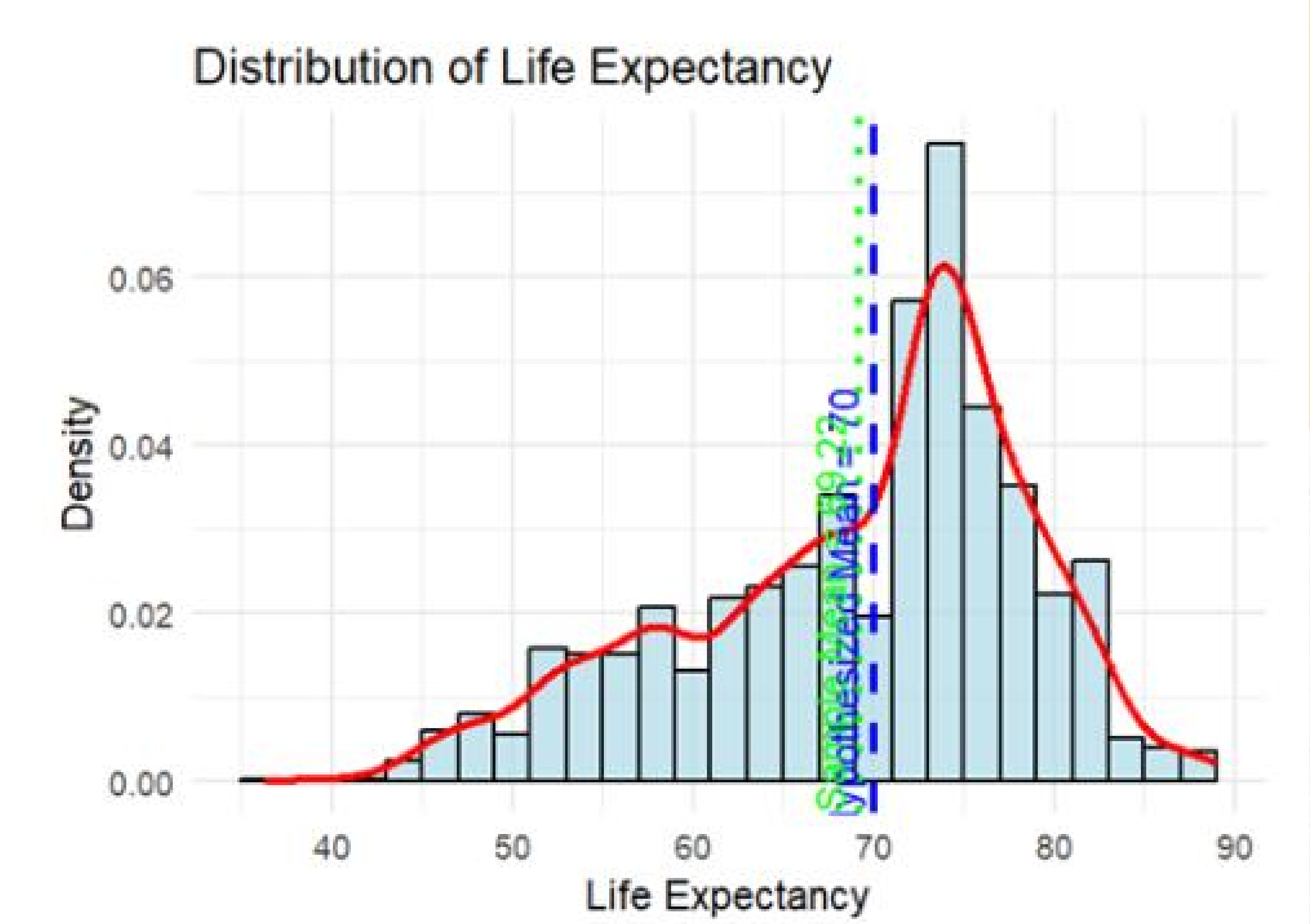
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

p-value= twice the area to the left of the test statistic

HYPOTHESIS TESTING: ONE SAMPLE T-TEST

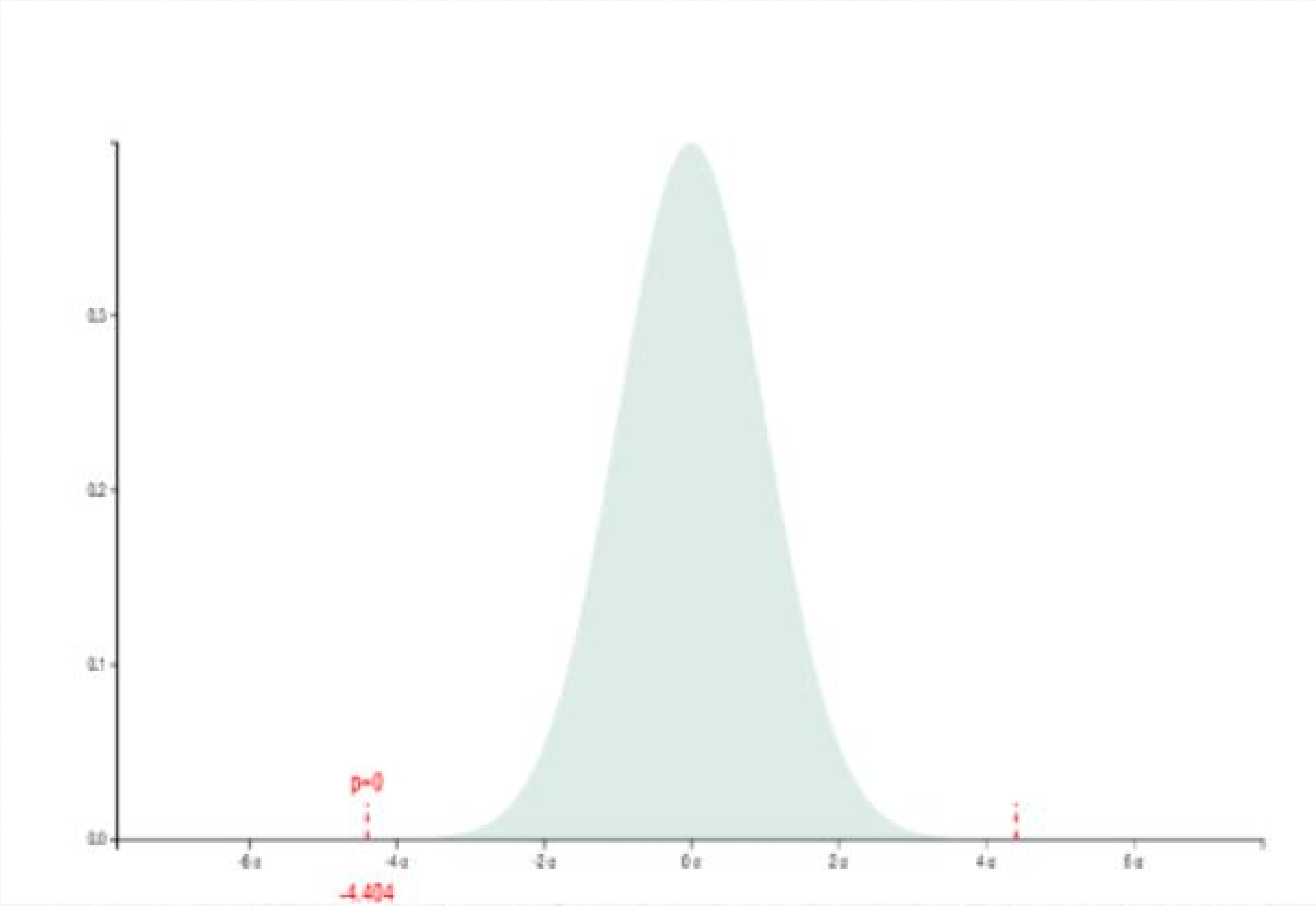
```r
# Perform the t-test
t_test_result <- t.test(data$'Life expectancy' , mu = 70)
print(t_test_result)
```

```
> print(t_test_result)

        One Sample t-test

data:  data$"Life expectancy"
t = -4.4036, df = 2927, p-value = 1.103e-05
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 68.87982 69.57004
sample estimates:
mean of x 
 69.22493 

> print(paste("p-value:", p_value))
[1] "p-value: 1.10263601464004e-05"
```



test statistic t= -4.4036,
p-value=1.1×10-5
sample
mean=69.22

HYPOTHESIS TESTING: ONE SAMPLE T-TEST



Distribution Graph when test statistic = -4.4036 and P-value=1.1×10^-5

Since P-value 1.1×10^-5< 0.05, we reject the null hypothesis, there is sufficient evidence to conclude that the average life expectancy is significantly different from 70 years.

CORRELATION TEST

To determine the strength and direction of the relationship (linear relationship) between life expectancy and GDP

Hypothesis statement:
H_0: $\rho = 0$ (no linear correlation)
H_1: $\rho \neq 0$ (linear correlation exists)

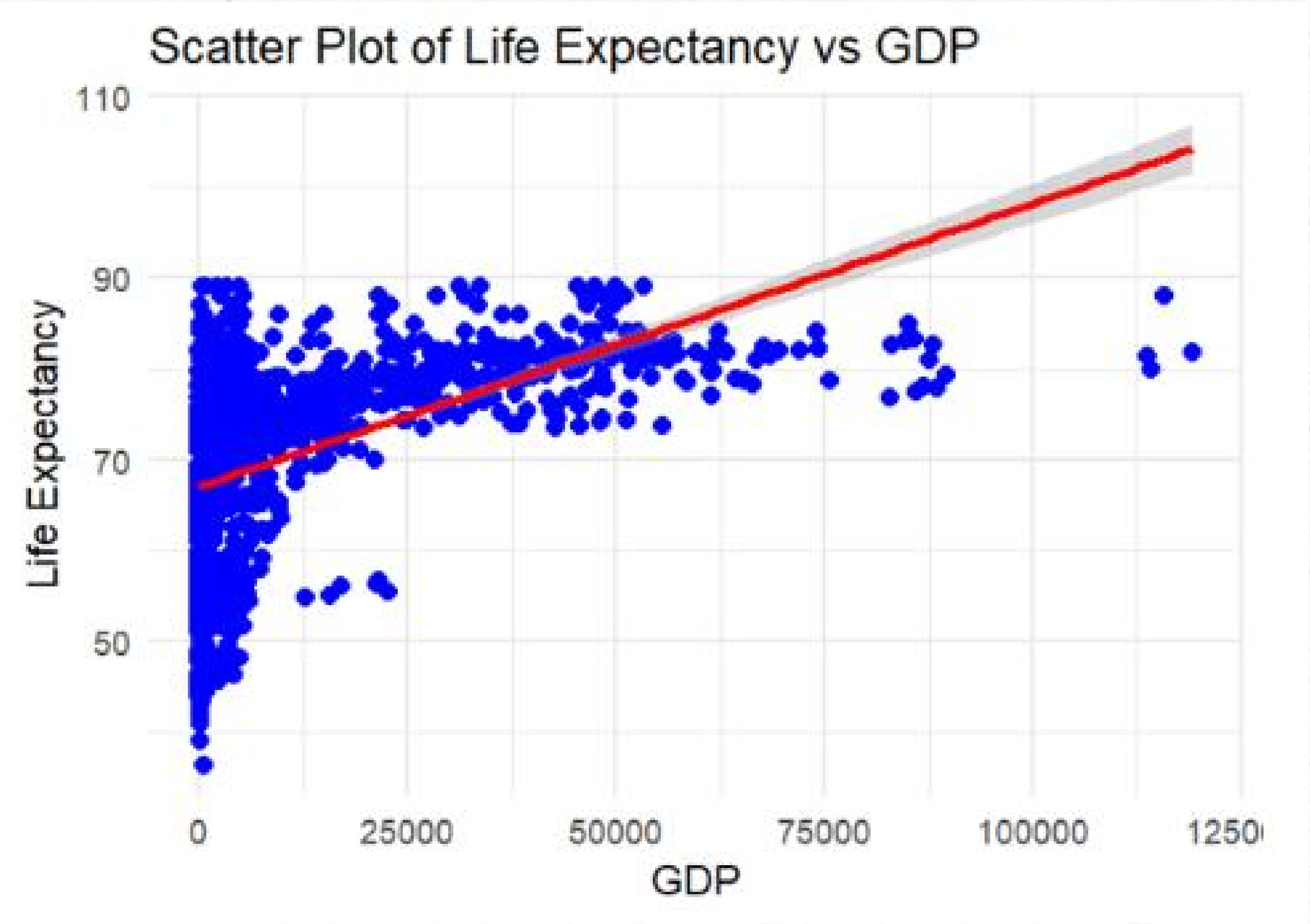
Pearson's Product-Moment Correlation test formula:

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

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

Degree of freedom =n-2

CORRELATION TEST




```
Correlation coefficient (r): 0.461455192620738"
```

From the scatter plot , we can see that the points slope slightly upward, it indicates that there is a positive correlation between life expectancy and GDP, that is the higher the life expectancy, the higher the GDP.

By using RStudio, we also get a sample correlation coefficient, r = 0.461455, which indicates that there is a moderate positive linear correlation between life expectancy and GDP.

CORRELATION TEST

```
        Pearson's product-moment correlation

data:  data$"Life expectancy" and data$GDP
t = 25.919, df = 2483, p-value < 2.2e-16
a [1] "Critical value: 1.9607723063617"
95 percent confidence interval:
 0.4299354 0.4918515
sample estimates:
      cor
0.4614552
```

```
[1] "Critical value: 1.9607723063617"
```

```
> df <- nrow(data) - 2
> print(paste("Degrees of freedom:", df))
[1] "Degrees of freedom: 2936"
```

From R studio:

test statistic = 25.919

critical value = 1.9608

df = 2936

Conclusion:

Since the test statistic t=25.919 > upper tail critical value = 1.9608, we reject the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between life expectancy and GDP.

REGRESSION TEST

Regression Test - Multiple Linear Regression

To examine the relationship between life expectancy (dependent variable) and multiple predictors (independent variables: adult mortality, BMI, GDP, HIV/AIDS).

$$\text{Life Expectancy} = \beta_0 + \beta_1(\text{Adult Mortality}) + \beta_2(\text{BMI}) + \beta_3(\text{GDP}) + \beta_4(\text{HIV/AIDS}) + \epsilon$$

- Dependent Variable: Life Expectancy
- Independent Variables: Adult Mortality, BMI, GDP, HIV/AIDS

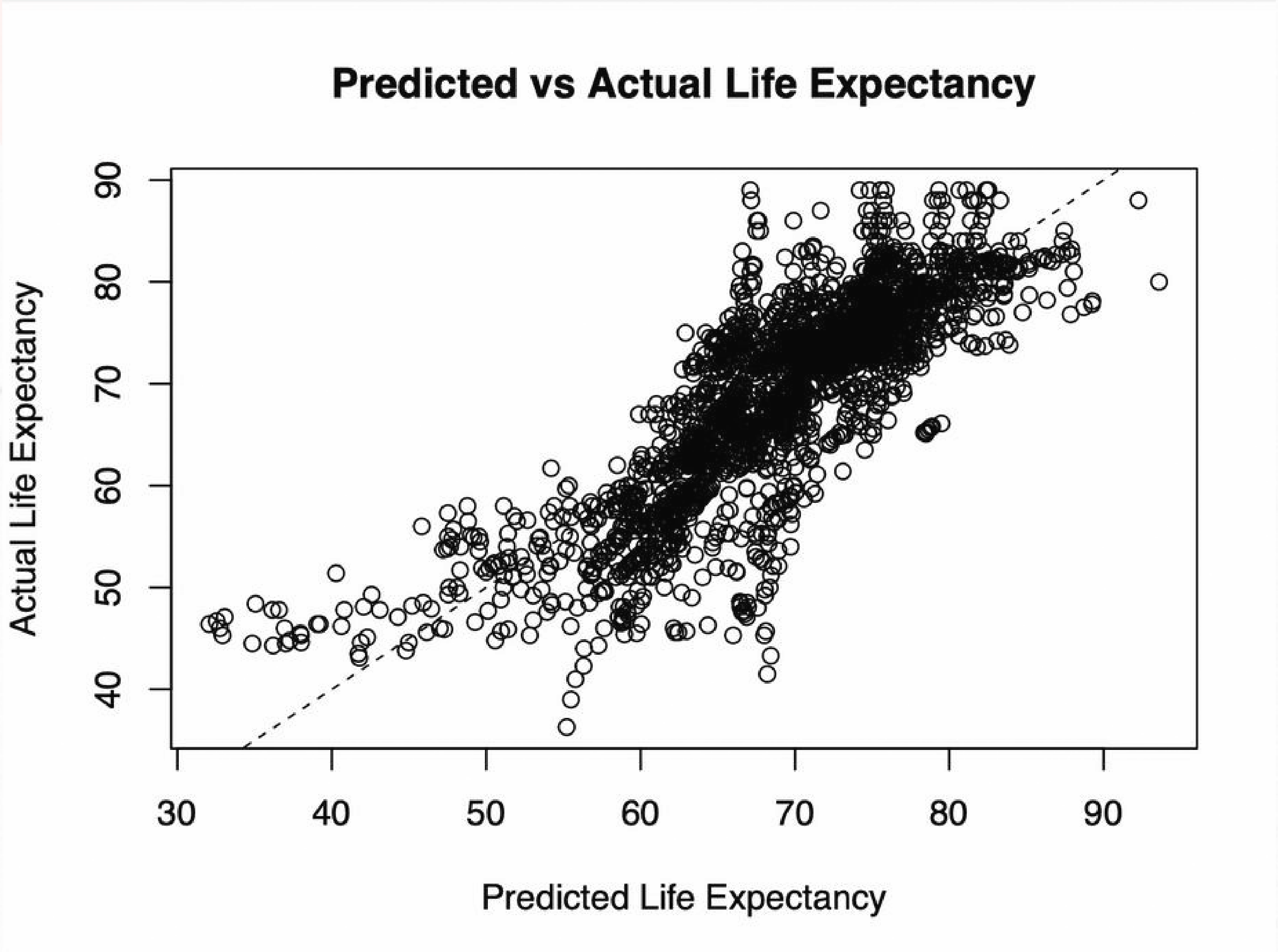
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.789e+01  3.429e-01  198.00   <2e-16 ***
AdultMortality -2.702e-02  1.110e-03  -24.34   <2e-16 ***
BMI             1.487e-01  6.140e-03   24.21   <2e-16 ***
GDP             1.513e-04  8.228e-06   18.38   <2e-16 ***
HIVAIDS        -4.836e-01  2.396e-02  -20.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Intercept ($\beta 0$): 67.89
- Adult Mortality ($\beta 1$): -0.027
- BMI ($\beta 2$): 0.15
- GDP ($\beta 3$): 0.00015
- HIV/AIDS ($\beta 4$): -4.84

ŷ = 67.89 - 0.027 (AdultMortality) + 0.15(BMI) + 0.00015(GDP) - 4.84(HIV AIDS)

GOODNESS-OF-FIT
COEFFICIENT OF DETERMINATION (R^2)

$$R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = 0.877$$



Key findings:

- Higher adult mortality rates are associated with lower life expectancy.
- Higher BMI and GDP are associated with higher life expectancy.
- Higher HIV/AIDS prevalence is associated with significantly lower life expectancy.

✓ strong linear relationship, indicating a good model fit.

ANOVA (ANALYSIS OF VARIANCE) TEST

To test if there are significant differences in life expectancy between developing and developed countries.

H_0: μDeveloped = μDeveloping
H_1: μDeveloped ≠μDeveloping

Status	n	$\bar{x}$	s
Developed	2426	79.19785	3.930942
Developing	512	67.11147	9.006092

F-TEST STATISTIC FORMULA:

$$MS_{between} = \frac{SS_{between}}{df_{between}} \qquad MS_{within} = \frac{SS_{within}}{df_{within}}$$

$$F = \frac{MS_{between}}{MS_{within}}$$

```
> cat("Sum of Squares Between Groups (SSB): ", ss_between, "\n")
Sum of Squares Between Groups (SSB):  61714.72
> cat("Degrees of Freedom Between Groups (df_between): ", df_between, "\n")
Degrees of Freedom Between Groups (df_between):  1
> cat("Mean Square Between Groups (MSB): ", ms_between, "\n\n")
Mean Square Between Groups (MSB):  61714.72

> cat("Sum of Squares Within Groups (SSW): ", ss_within, "\n")
Sum of Squares Within Groups (SSW):  203776
> cat("Degrees of Freedom Within Groups (df_within): ", df_within, "\n")
Degrees of Freedom Within Groups (df_within):  2926
> cat("Mean Square Within Groups (MSW): ", ms_within, "\n\n")
Mean Square Within Groups (MSW):  69.64321
```

Results for MSB and MSW

F-TEST STATISTIC FORMULA:

$$F = \frac{MS_{between}}{MS_{within}}$$

```
> cat("F-test statistic: ", f_statistic, "\n")
F-test statistic:  886.1556
> cat("F-critical value at alpha =", alpha, "is:", f_critical, "\n")
F-critical value at alpha = 0.05 is: 3.844639
>
```

results for F-test statistic and F-critical value

F-test statistic > F-critical value (886.1556 > 3.844639),
we reject the null hypothesis, H_0, as there is sufficient evidence to claim that there is a significant difference in the means of life expectancy between Developed and Developing countries.

CHI-SQUARE TEST OF INDEPENDENCE

To determine if there is an association between a country's development status (developing/developed) and Hepatitis B immunization coverage (low, medium, high).

- Low: Hepatitis B < 60
- Medium: 60 ≤ Hepatitis B < 80
- High: Hepatitis B ≥ 80

H_0: Hepatitis B immunization coverage is independent of a country's development status.
H_1: Hepatitis B immunization coverage is dependent on a country's development status.

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

Oij = Observed count at row i column j
Eij = Expected count at row i column j

$$df = (2 - 1)(3 - 1)$$

```
> alpha <- 0.05
> df <- (2 - 1) * (3 - 1)
> critical_value <- qchisq(1 - alpha, df)
> critical_value
[1] 5.991465
```

OBSERVED FREQUENCIES

```
[1] "Observed Frequencies:"
> print(observed)

           High  Low Medium
Developed   306   24      9
Developing 1413  298    335
```

EXPECTED FREQUENCIES

```
[1] "Expected Frequencies:"
> print(expected)

                High       Low   Medium
Developed   244.3358  45.76855  48.8956
Developing 1474.6642 276.23145 295.1044
```

CALCULATING TEST STATISTICS

```
> # Print chi-square test results
> print(chi_square_test)


        Pearson's Chi-squared test

data:  contingency_table
X-squared = 68.156, df = 2, p-value = 1.585e-15
```

Since χ^2 = 68.156 > 5.991, we reject the null hypothesis.

- There is sufficient evidence to conclude that there is an association between a country's development status and its Hepatitis B immunization coverage.
- Implications: Hepatitis B immunization coverage varies significantly between developing and developed countries.

THANK YOU VERY MUCH!