Module 5 | Practice 5: Life Expectancy Dataset

Trang Tran

CPS, Northeastern University

ALY6010 | Probability Theory and Introductory Statistics

Patrick McQuillan

Mar 25, 2023

Introduction

The dataset related to life expectancy consists of 22 columns and 2938 rows which have 20 numeric variables. I dropped the character variables of "Country" and "Status" out of the dataset to run a whole correlation table. After getting the initial overview, I decided to choose these variables to make further correlation and regression analysis: "Life.expectancy" (Life Expectancy in age), "Alcohol" (recorded per capita (15+) consumption (in liters of pure alcohol)), "BMI" (Average Body Mass Index of the entire population), "GDP" (Gross Domestic Product per capita (in USD), and "Schooling" (Number of years of Schooling(years)).

Correlation Table and Chart

```
> corr <- cor(df[, c("Life.expectancy", "Alcohol", "BMI", "GDP", "Schooling")], method =
"pearson")
> corr
                Life.expectancy   Alcohol       BMI       GDP Schooling
Life.expectancy       1.0000000 0.4027183 0.5420416 0.4413218 0.7276300
Alcohol               0.4027183 1.0000000 0.3533962 0.4434328 0.6169748
BMI                   0.5420416 0.3533962 1.0000000 0.2661140 0.5548439
GDP                   0.4413218 0.4434328 0.2661140 1.0000000 0.4679470
Schooling             0.7276300 0.6169748 0.5548439 0.4679470 1.0000000
```

Table 1: Correlation Table

This correlation table shows the strength and direction of relationships between 5 variables. While we see a strong positive linear relationship between Schooling and Life expectancy (>0.7), most of the other relationship remains moderately positive status with correlation coefficients ranging around 0.4 to 0.6. There is no negative linear relationship in this matrix. The correlation plot presents a visualization of the correlation relationships.

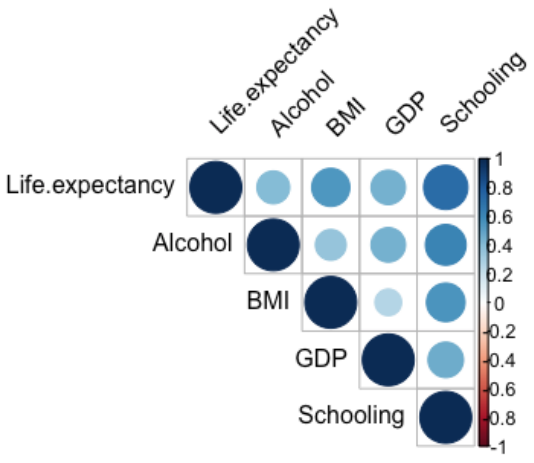


Figure 1: Correlation Plot

Linear Regression Analysis

Firstly, we take a look at the distribution of the targeted dependent variable – Life expectancy by using histogram and density functions. Figure 2 illustrates a negatively skewed normal distribution with a blue dash line of the mean value.

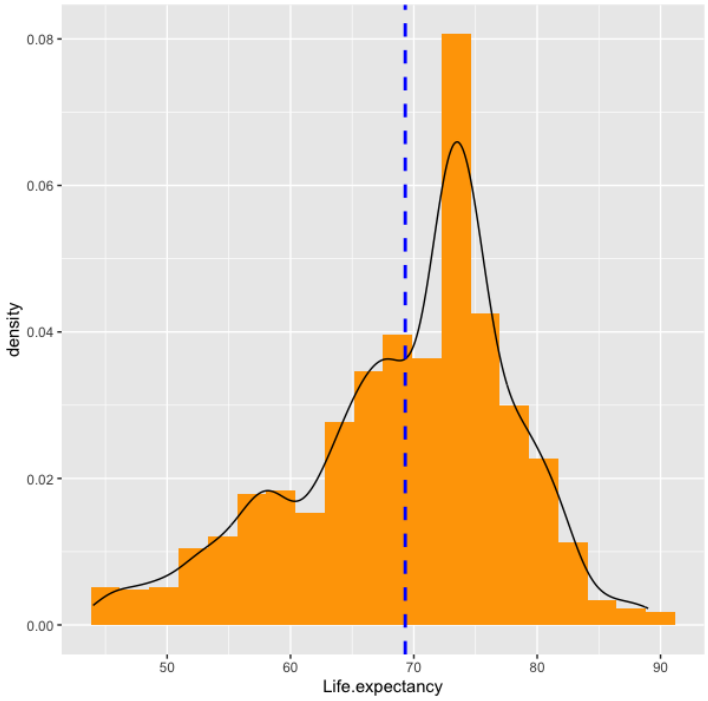


Figure 2: Life Expectancy Distribution

Secondly, we check for outliers of all 5 chosen variables by using the boxplot charts. Only the GDP variable boxplot shows a lot of interfering outliers that could have a significant impact on the outcome of the regression model (Figure 3). Therefore, I use the logarithmic scale to mutate the GDP column (Figure 4).

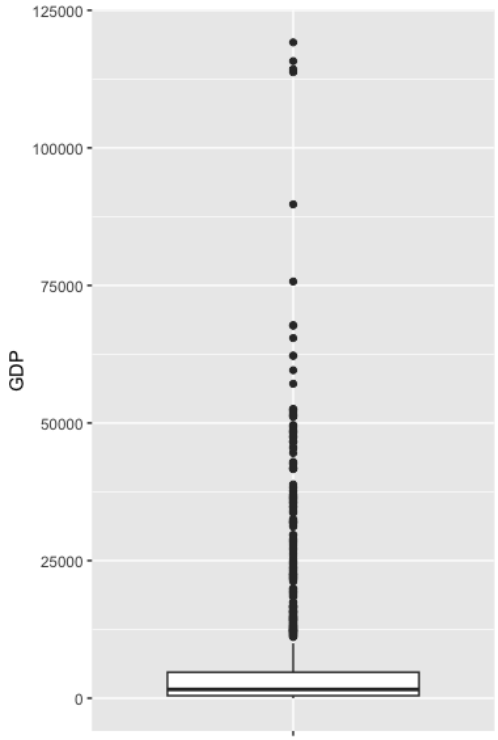


Figure 3: GDP boxplot (raw data)

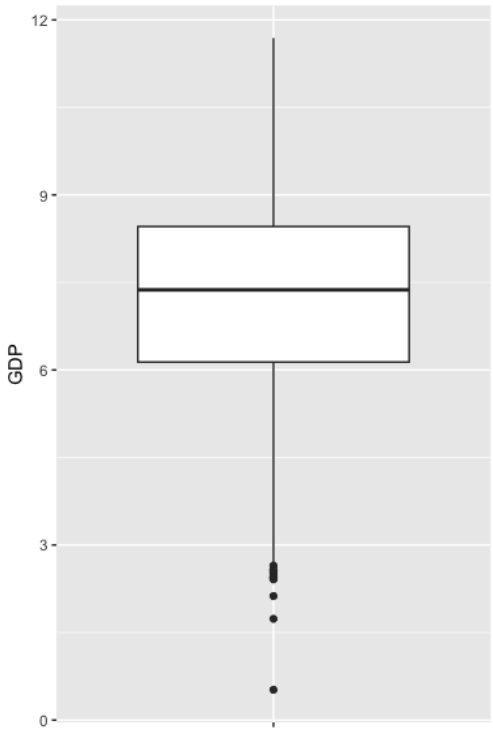


Figure 4: GDP boxplot (log data)

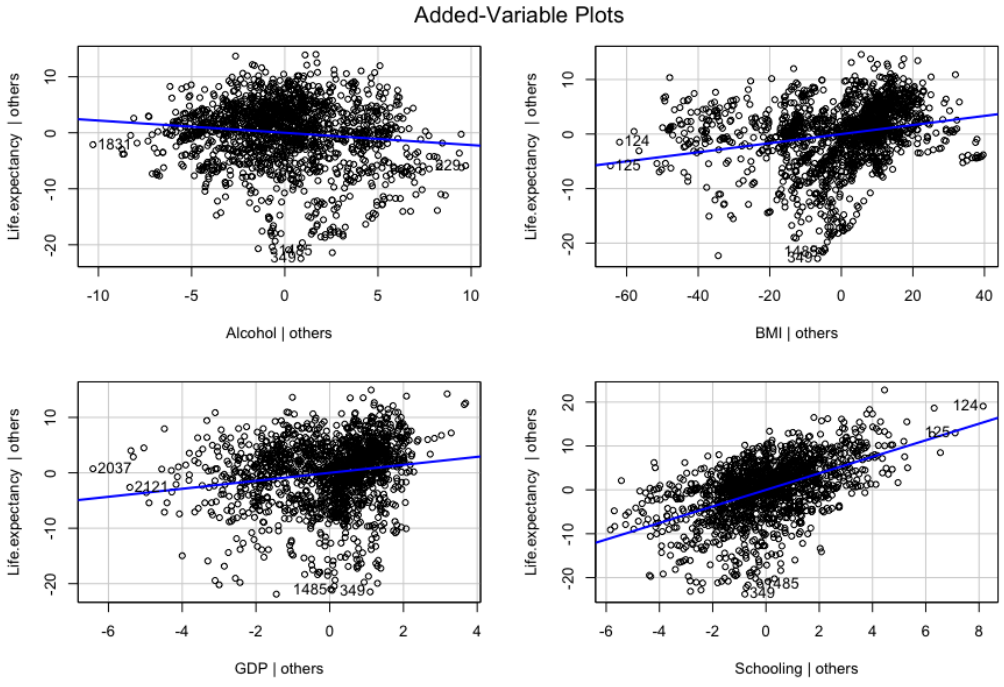
Finally, we run the OLS regression model with the outcome variable of Life expectancy, and four predictor variables: Alcohol, BMI, GDP, and Schooling. The regression table and the added-variable plots are as follows. While correlation analysis measures the strength and direction of the relationship between two variables, regression analysis attempts to model the relationship between a dependent variable (the outcome variable) and one or more independent variables. Regression analysis can also be used to make predictions about the value of the dependent variable based on the values of the independent variables. Additionally, regression analysis can be used to identify which predictor variables have a significant impact on the dependent variable and to quantify the strength and direction of these relationships. We will interpret this regression output in Module 6.

```
Call:
lm(formula = Life.expectancy ~ Alcohol + BMI + GDP + Schooling,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-22.2732  -2.9292   0.8249   3.8336  14.3524

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 38.987186   0.769978  50.634 < 0.0000000000000002 ***
Alcohol     -0.219382   0.045225  -4.851      0.00000134555896 ***
BMI          0.083453   0.008666   9.630 < 0.0000000000000002 ***
GDP          0.720527   0.103047   6.992      0.00000000000392 ***
Schooling    1.885823   0.078118  24.141 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.752 on 1644 degrees of freedom
Multiple R-squared:  0.5736,    Adjusted R-squared:  0.5725
F-statistic: 552.8 on 4 and 1644 DF,  p-value: < 0.00000000000000022
```



References

1. Zach (2020, Dec 23). *How to Plot Multiple Linear Regression Results in R*. Statology. Retrieved March 25, 2023. https://www.statology.org/plot-multiple-linear-regression-in-r/
2. Kaggle. Life Expectancy (WHO). Retrieved March 25, 2023. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?datasetId=12603
3. R Coder. Correlation plot in R. Retrieved March 25, 2023. https://r-coder.com/correlation-plot-r/