DS501 HW4

What is Regression?

Regression models the relationship between a dependent variable (predicted outcome) and one or more independent variables (predictors). It explains how changes in predictors influence the outcome. These outcomes can either be continuous (predict values such as customer lifetime value) or discrete (predict probabilities such as loan default likelihood). Linear Regression models the relationship between one predictor and one response variable. It is a specific type of regression used to predict a continuous outcome as a linear function of predictors. The linear regression model represents the response variable Y, as a function of predictors X:

$$Y=\beta0+\beta1X+\epsilon$$

where $\beta0$ is the point where the regression line crosses the Y-axis, representing the value of Y when X = 0, $\beta1$ represents the change in Y for each unit increase in X, and ϵ is the error term accounting for variability not explained by X.

What data I collected?

The Life Expectancy dataset I used is publicly available on Kaggle. It uses data from the World Health Organization and the United Nations. It includes 193 countries' data from over 15 years and includes variables related to health and socioeconomic status and demographic trends. Specifically these variables include country, continent, year, life expectancy, and health metrics such as alcohol consumption and BMI.

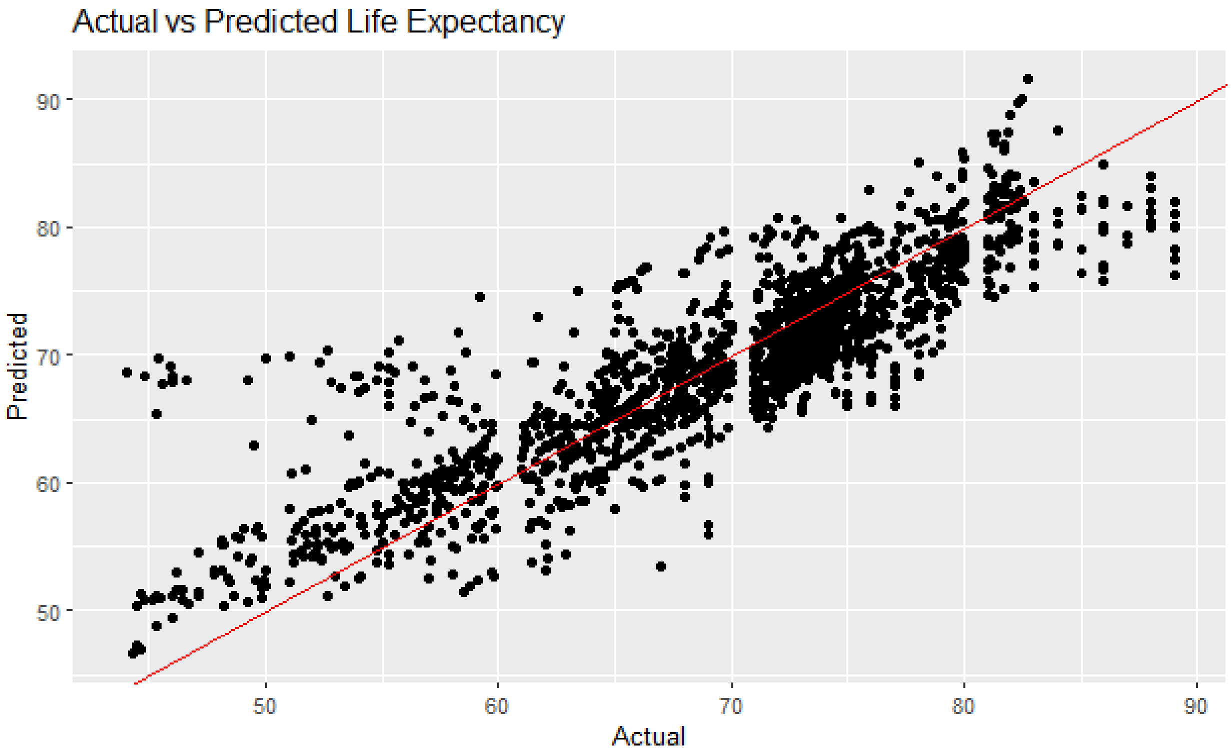
Motivation

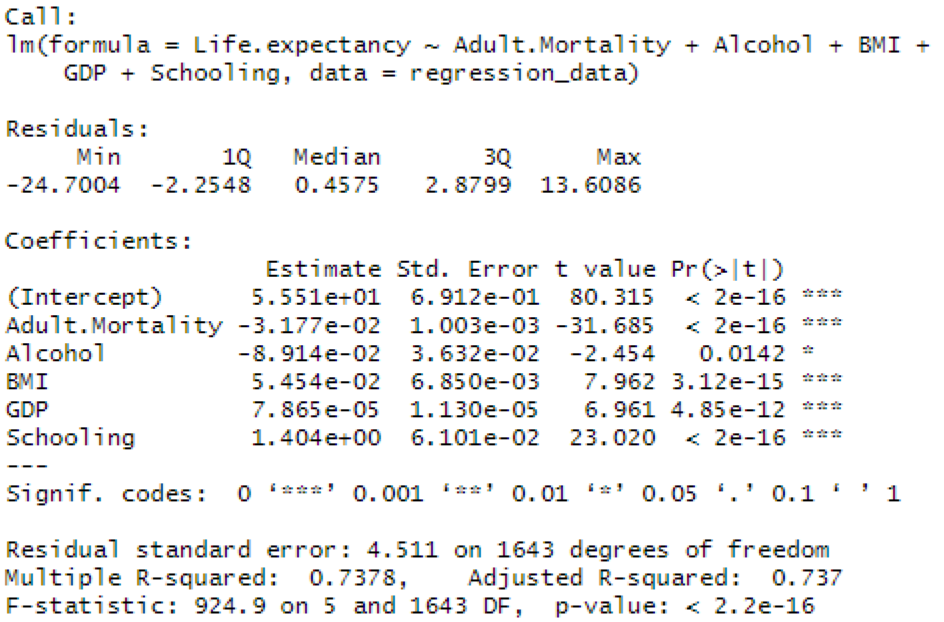
Life expectancy is an important statistic of a country's overall health, well-being and development. This project itself is important as it highlights disparities in life expectancy and overall health across countries, potentially helping to identify areas in need of policy interventions. A country can use this data and determine what aspects it needs to focus on and improve in order to make itself more healthy. For example if low schooling levels are leading to lower life expectancy, countries can use this data to spend more time and money to improve schooling.

How did you analyze the data?

For the purpose of data analysis I focused on key variables: Life expectancy, adult mortality, alcohol consumption, BMI, GDP, and schooling. I examined the correlation between life expectancy and other variables. I created a regression model using all the above variables, and utilized it to compare actual vs predicted life expectancy.

What did you find in the data?





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

Residuals:
     Min       1Q   Median       3Q      Max 
-24.7004  -2.2548   0.4575   2.8799  13.6086 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.551e+01  6.912e-01  80.315  < 2e-16 ***
Adult.Mortality -3.177e-02  1.003e-03 -31.685  < 2e-16 ***
Alcohol         -8.914e-02  3.632e-02  -2.454   0.0142 *  
BMI              5.454e-02  6.850e-03   7.962 3.12e-15 ***
GDP              7.865e-05  1.130e-05   6.961 4.85e-12 ***
Schooling        1.404e+00  6.101e-02  23.020  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.511 on 1643 degrees of freedom
Multiple R-squared:  0.7378,	Adjusted R-squared:  0.737 
F-statistic: 924.9 on 5 and 1643 DF,  p-value: < 2.2e-16
```

The model used is Life.expectancy ~ Adult.mortality + Alcohol + BMI + GDP + Schooling. The coefficients in the model tell the relationship between each predictor variable and the outcome. The intercept of the model is 55.51, meaning when all predictors are zero, life expectancy is 55.51 years. The coefficient for adult mortality is -0.03177, which means that for every unit increase in adult mortality, life expectancy decreases by 0.03177 years. The low p-value indicates that this variable is highly significant. The coefficient for alcohol is -0.08914, indicating that for every unit increase in alcohol consumption, life expectancy decreases by 0.08914 years. The p-value shows that alcohol is statistically significant. The coefficient for BMI is 0.05454 which means that for every unit increase in BMI, life expectancy increases by 0.05454 years. The very low p-value shows that BMI has a statistically significant positive effect on life expectancy. The coefficient of GDP is 0.00007865, indicating that for every unit increase in GDP, life expectancy increases by 0.00007865 years. The p-value shows that it is statistically significant. The coefficient for schooling is 1.404, meaning that for every additional year of schooling, life expectancy increases by 1.404. The low p-value shows that it is statistically significant. The median residual of 0.4575 shows that the model slightly overestimates life expectancy on average. The multiple r-squared value of 0.07378 means that 73.78% of the variability in life expectancy can be explained by the five predictors. The f-statistic of 924.9 with its given p-value of < 2.2e-16 tells that the predictors explain life expectancy in a meaningful way.