Final Project

Mingchen Wang

2023-03-24

Introduction

Background

An essential measure of a country's health and prosperity is its life expectancy. Many would argue that this is influenced by a number of variables, including lifestyle choices, social influences, genetics, and access to healthcare. There are still a lot of differences in life expectancy between nations and areas, with some populations having substantially lower life expectancy than others. But why? This project intends to investigate how various variables in different nations affect one's life expectancy. Or, in a broad sense, what makes you live longer?

The values used in this project is sourced from the data set **Life Expectancy** (hyperlink included) from Kaggle. This data set contains information about life expectancy for people, CO2 emissions and Health expenditure, Obesity rate, and people using at least basic drinking water services across the world (contains 1904 unique values)

Modeling and Structure

A linear model will be made based on Life Expectancy and other factors, where the average life expectancy in a country will be our response variable. Our model will be adjusted to fit, we will determine which variables effects one's life expectancy, and make adjustments based on our calculations.

This report will be divided into five sections:

- The current introduction section
- Data Description, some basic summary and cleaning process of the data.
- Modeling, analyzation, and interpretation of the data
- A summary of results.
- A discussion of my findings

Data Description.

Data cleaning:

In order to see the variables we wanted in one data set, let's first load our data, then remove all the terms that is not necessary for this study.

```
Life_expectancy_old <- read.csv(file = "Life_Expectancy_00_15.csv", head = TRUE, sep=";")


LifeExp <- Life_expectancy_old %>%
  filter (Year == "2015") %>%  # for the purpose of this study,
  # let's only look at one specific year.
  select("Life.Expectancy", "CO2.emissions",
         "Health.expenditure", "Electric.power.consumption",
         "People.using.at.least.basic.drinking.water.services",
         "Beer.consumption.per.capita",
         "Obesity.among.adults") %>%
  rename("DrinkableWater"
         = "People.using.at.least.basic.drinking.water.services",
         "ElectricPower" = "Electric.power.consumption",
          "Beer" = "Beer.consumption.per.capita",
          "Obesity" = "Obesity.among.adults")
  #for the sake of making our summary table shorter, let's rename our column values

head(LifeExp)
```

```
##   Life.Expectancy CO2.emissions Health.expenditure ElectricPower DrinkableWater
## 1          78.025      1.603775           4.896312     2309.3665       93.39433
## 2          76.090      3.933496           6.978489     1362.8719       93.40956
## 3          59.398      1.135044           2.605795      312.2289       54.31693
## 4          76.068      4.301914          10.229339     3074.7021       98.96659
## 5          74.467      1.825292          10.117634     1961.6104       99.55257
## 6          82.400     15.863288           9.327589    10071.3990       99.97001
##   Beer Obesity
## 1 1.48    21.7
## 2 0.31    25.7
## 3 3.25     6.5
## 4 3.39    28.0
## 5 0.49    20.2
## 6 3.76    29.8
```

Summary

Now we have our data set in hand, let's check out some summary statistics.

```
summary(LifeExp)
```

```
##  Life.Expectancy CO2.emissions      Health.expenditure ElectricPower
##  Min.   :53.11   Min.   : 0.06968   Min.   : 1.973     Min.   :   51.2
##  1st Qu.:70.03   1st Qu.: 1.10117   1st Qu.: 4.869     1st Qu.:  773.5
##  Median :74.90   Median : 3.72989   Median : 6.551     Median : 2584.4
##  Mean   :73.31   Mean   : 5.05221   Mean   : 6.576     Mean   : 3937.7
##  3rd Qu.:77.90   3rd Qu.: 6.70010   3rd Qu.: 8.297     3rd Qu.: 5246.5
##  Max.   :82.90   Max.   :33.04351   Max.   :16.524     Max.   :22999.9
##  DrinkableWater        Beer          Obesity
##  Min.   : 42.07   Min.   :0.000   Min.   : 3.10
##  1st Qu.: 88.44   1st Qu.:0.570   1st Qu.:11.95
##  Median : 96.53   Median :2.000   Median :21.90
##  Mean   : 89.65   Mean   :2.251   Mean   :19.58
##  3rd Qu.: 99.62   3rd Qu.:3.470   3rd Qu.:26.15
##  Max.   :100.00   Max.   :6.870   Max.   :36.70
```

It seems like the lowest average Life Expectancy in a country is about 53 years, while the highest is almost 83 years. This data is consisitent with our currernt understanding of life expectancy. We are also able to see some summary statistics about other factors such as CO2 emission rates and electric power consumption in the chart above as well.

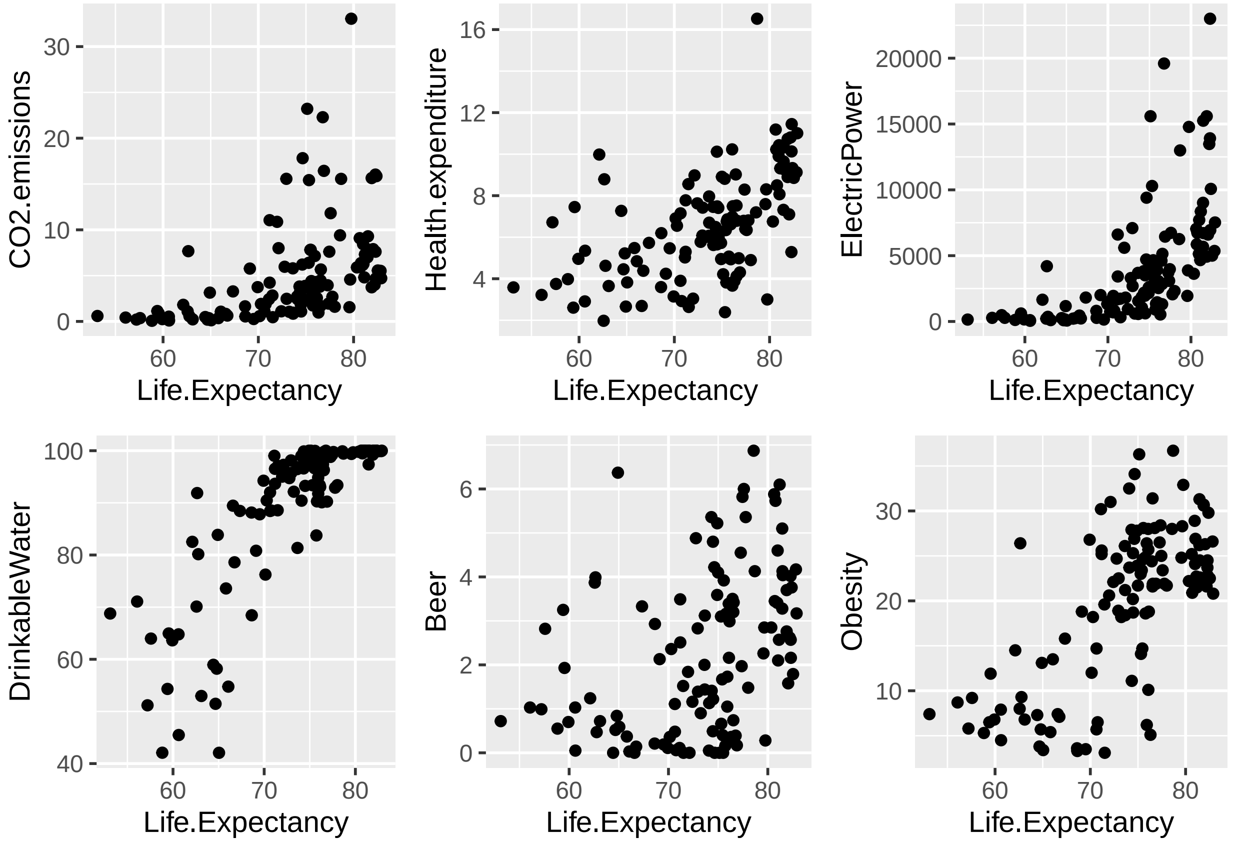
Modeling and Analyzation

Modeling

We are interested in developing model for Life Expectancy (Life.Expectancy) based on our predictors. (CO2.emissions, Health.expenditure, ElectricPower, DrinkableWater, Beer, Obesity)

Let's first take a look a the scatter plots between Life Expectancy and possible predictor variables listed.

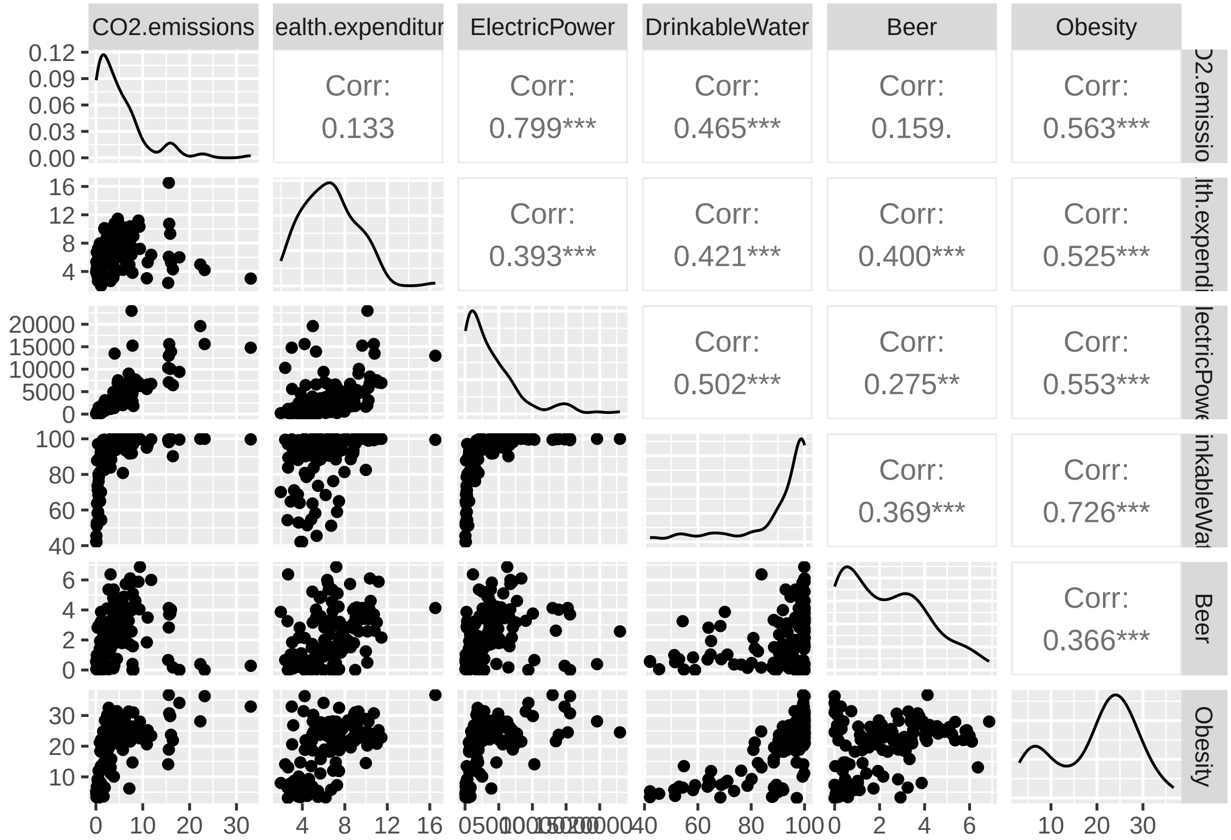
```
plot1 <- ggplot(LifeExp, aes(x=Life.Expectancy, y=CO2.emissions)) +
    geom_point()
plot2 <- ggplot(LifeExp, aes(x=Life.Expectancy, y=Health.expenditure)) + geom_point()
plot3 <- ggplot(LifeExp, aes(x=Life.Expectancy, y=ElectricPower)) +
    geom_point()
plot4 <- ggplot(LifeExp, aes(x=Life.Expectancy, y=DrinkableWater)) +
    geom_point()
plot5 <- ggplot(LifeExp, aes(x=Life.Expectancy, y=Beer)) +
    geom_point()
plot6 <- ggplot(LifeExp, aes(x=Life.Expectancy, y=Obesity)) +
    geom_point()

grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, ncol=3)
```



All the predictor variables seems to have at least some amount of correlation to Life Expectancy. There are no graphs that seems significantly concerning, or have absolutely no correlation here.

We can also create a scatter plot matrix between the variables to see the relationship between each of the variables, and to check if there are any predictors that are correlated to each other.

```
ggpairs(LifeExp[,2:7])
```



As seen in the plot, there is a high correlation between electric power consumption and CO2 emission. The correlation value of 0.799 further proves this point, the closer this value is to one, the higher the correlation is. This values makes sense in a real world context as well, since we know that electric power is often tied to energy production, and is seen as one of the source of greenhouse gas emission.(Here is a study by United States Environmental Protection Agency, where it is found that about 25% of Greenhouse Gas Emissions is caused by electric power consumption: [https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions])

People using at least basic drinking water services and Adult Obesity rate as seem to have a high correlation, with a correlation value of 0.726. Putting this into real world context, it is true that drinking unfiltered water could cause many diseases, some of which could have the effect of weight gaining. (Here is a study about contamination in water and obesity for reference: [https://www.sciencedirect.com/science/article/pii/S0160412020322571])

The separate relationships between the response variable and each of the predictor variables do not appear to be linear. Let's try to build a linear model based on these predictor variables first, and see if this correlation might effect our model.

```
model <- lm(Life.Expectancy ~ CO2.emissions + Health.expenditure +
              ElectricPower + DrinkableWater + Beer + Obesity,
            data = LifeExp)
summary(model)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ CO2.emissions + Health.expenditure +
##     ElectricPower + DrinkableWater + Beer + Obesity, data = LifeExp)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6880  -1.8985   0.2351   2.4360   8.6153
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        42.2730167  2.3995925  17.617  < 2e-16 ***
## CO2.emissions      -0.0709610  0.1181315  -0.601  0.54926
## Health.expenditure  0.4764516  0.1798435   2.649  0.00923 **
## ElectricPower       0.0003710  0.0001451   2.557  0.01190 *
## DrinkableWater      0.2885587  0.0330126   8.741 2.62e-14 ***
## Beer                0.1030593  0.2059288   0.500  0.61773
## Obesity             0.0355923  0.0645253   0.552  0.58232
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.577 on 112 degrees of freedom
## Multiple R-squared:  0.7517, Adjusted R-squared:  0.7384
## F-statistic: 56.52 on 6 and 112 DF,  p-value: < 2.2e-16
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Life.Expectancy
##                     Df  Sum Sq Mean Sq  F value    Pr(>F)
## CO2.emissions        1 1262.93 1262.93  98.7288 < 2.2e-16 ***
## Health.expenditure   1 1389.37 1389.37 108.6132 < 2.2e-16 ***
## ElectricPower        1  132.52  132.52  10.3596  0.001685 **
## DrinkableWater       1 1545.74 1545.74 120.8378 < 2.2e-16 ***
## Beer                 1    3.72    3.72   0.2906  0.590917
## Obesity              1    3.89    3.89   0.3043  0.582319
## Residuals          112 1432.69   12.79
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From our summary table, we see that we have an adjusted R-squared value of 0.7384. This means gives the 73% of life expectancy can be explained by our predictors. However, from before, we know that our model might be impacted by a Multicollinearity issue, where our predictor factors are highly correlated to each other. This can also be seen in our p-value difference, especially in CO2 emission, in our ANOVA vs linear model.

To further expand this point, we can calculate the variance inflation factors of the linear model, as seen here.

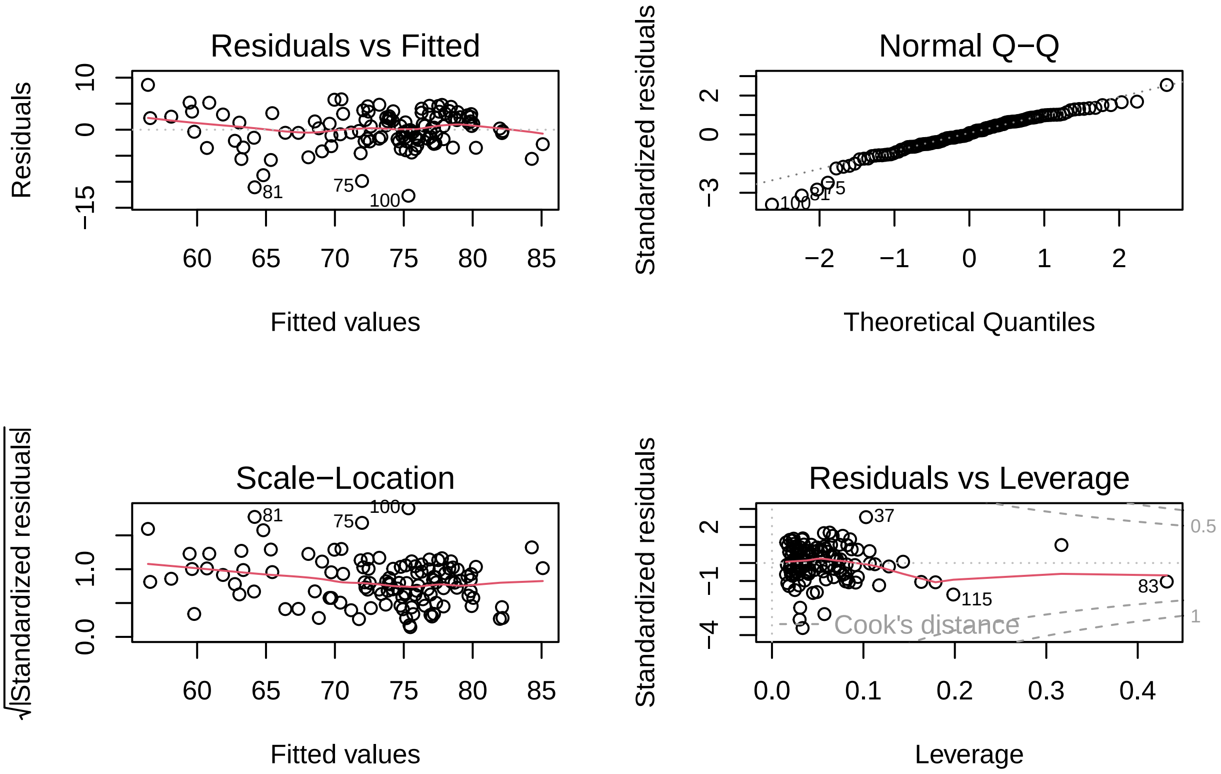
```
vif(model)
```

```
##      CO2.emissions Health.expenditure      ElectricPower     DrinkableWater
##           3.825798           1.898445           3.695907           2.230427
##               Beer            Obesity
##           1.276212           3.013537
```

A variance inflation higher than 5 is considered extremely high, while between 2~5 can be considered as moderately correlated. There are no extremely high correlation between any of our variables, however, many variables are moderately correlated.

Let's also take a loot at the summary plots of our model and see if we can apply a transformation to adjust model.

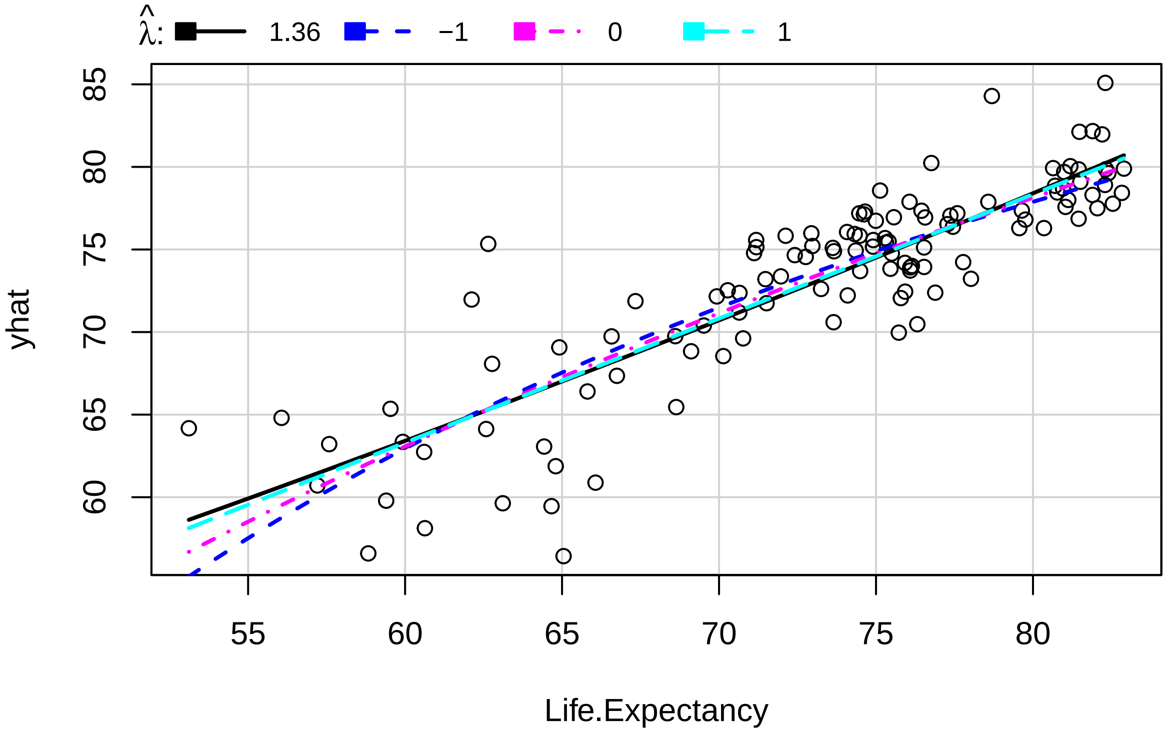
```
par(mfrow = c(2, 2))
plot(model)
```



From our QQ plot, it seems like a transformation could be done due to the normality of the errors. A good QQ plot should have all of its pointer in a straight line, while our QQ plot seems to have heavy tails on its end.

Perhaps the transformation will fix the Multicollinearity as well. Let's try to do a Inverse response plot to test for which kind of transformation is the best.

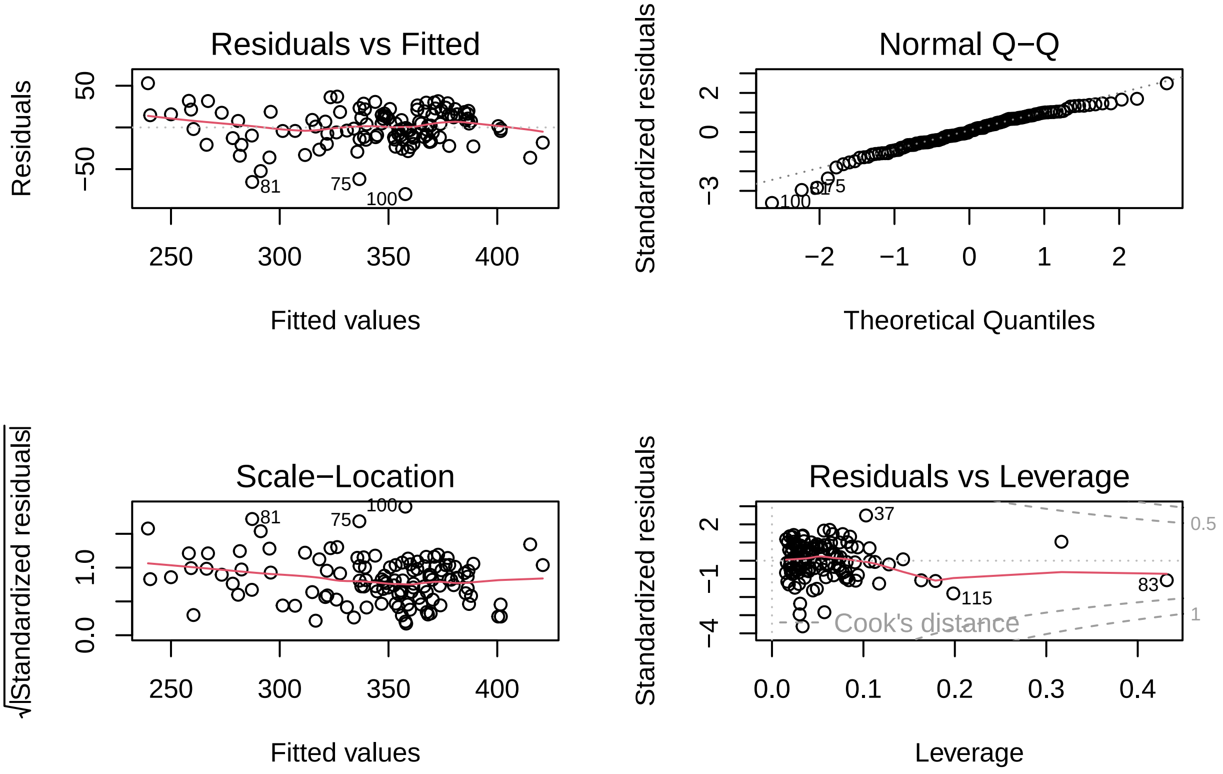
```
inverseResponsePlot(model, key=TRUE)
```



```
##      lambda      RSS
## 1  1.364495 1075.286
## 2 -1.000000 1151.225
## 3  0.000000 1100.062
## 4  1.000000 1077.005
```

Our lowest RSS value is at lambda = 1.36, meaning that we can apply the transformation of Y^1.36.

```
model2 <- lm(Life.Expectancy^(1.36) ~ CO2.emissions + Health.expenditure
             + ElectricPower + DrinkableWater + Beer + Obesity, data
             = LifeExp)

par(mfrow = c(2, 2))
plot(model2)
```



```
summary(model2)
```

```
##
## Call:
## lm(formula = Life.Expectancy^(1.36) ~ CO2.emissions + Health.expenditure +
##     ElectricPower + DrinkableWater + Beer + Obesity, data = LifeExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.913 -12.247   1.022  15.797  53.004
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        151.184090  15.081859  10.024  < 2e-16 ***
## CO2.emissions       -0.474893   0.742477  -0.640  0.52373
## Health.expenditure   3.070030   1.130348   2.716  0.00766 **
## ElectricPower        0.002445   0.000912   2.681  0.00845 **
## DrinkableWater       1.787442   0.207490   8.615 5.09e-14 ***
## Beer                 0.742558   1.294299   0.574  0.56731
## Obesity              0.218909   0.405552   0.540  0.59042
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22.48 on 112 degrees of freedom
## Multiple R-squared:  0.7523, Adjusted R-squared:  0.739
## F-statistic: 56.68 on 6 and 112 DF,  p-value: < 2.2e-16
```

There doesn't seem to be too much changes in our model after the transformation, this makes sense since our values are only raised to the 1.36th power. It seems like more steps needs to be taken.

Let's remove variables that has a high correlation and high p-values for a reduced model. This will help with our Multicollinearity issues. We can try a step-wise reduction, with remove the predictor with the highest p-value first.

```
model3 <- lm(Life.Expectancy^(1.36) ~  CO2.emissions + Health.expenditure +
              ElectricPower + DrinkableWater + Obesity, data = LifeExp)
summary(model3)
```

```
##
## Call:
## lm(formula = Life.Expectancy^(1.36) ~ CO2.emissions + Health.expenditure +
##     ElectricPower + DrinkableWater + Obesity, data = LifeExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.976 -12.732   0.739  15.462  53.177
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.502e+02  1.494e+01  10.053  < 2e-16 ***
## CO2.emissions      -5.038e-01  7.386e-01  -0.682  0.49654
## Health.expenditure  3.194e+00  1.106e+00   2.888  0.00465 **
## ElectricPower       2.484e-03  9.069e-04   2.739  0.00717 **
## DrinkableWater      1.804e+00  2.048e-01   8.811  1.7e-14 ***
## Obesity             2.347e-01  4.034e-01   0.582  0.56193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.41 on 113 degrees of freedom
## Multiple R-squared:  0.7515, Adjusted R-squared:  0.7405
## F-statistic: 68.36 on 5 and 113 DF,  p-value: < 2.2e-16
```

```
vif(model3)
```

```
##      CO2.emissions Health.expenditure      ElectricPower     DrinkableWater
##           3.808165           1.828956           3.675978           2.185501
##            Obesity
##           2.999714
```

The issue with Multicollinearity still exists and there seems to be values with high p-values, let's try to remove the next predictor with highest p-value, the obesity rate.

```
model4 <- lm(Life.Expectancy^(1.36) ~  CO2.emissions + Health.expenditure +
              ElectricPower + DrinkableWater, data = LifeExp)
summary(model4)
```

```
##
## Call:
## lm(formula = Life.Expectancy^(1.36) ~ CO2.emissions + Health.expenditure +
##     ElectricPower + DrinkableWater, data = LifeExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.522 -12.487   0.832  15.774  53.736
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.467e+02  1.360e+01  10.782  < 2e-16 ***
## CO2.emissions      -3.440e-01  6.836e-01  -0.503 0.615737
## Health.expenditure  3.466e+00  9.993e-01   3.469 0.000739 ***
## ElectricPower       2.415e-03  8.965e-04   2.694 0.008137 **
## DrinkableWater      1.869e+00  1.713e-01  10.913  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22.35 on 114 degrees of freedom
## Multiple R-squared:  0.7508, Adjusted R-squared:  0.742
## F-statistic: 85.86 on 4 and 114 DF,  p-value: < 2.2e-16
```

```r
vif(model4)
```

```
##      CO2.emissions Health.expenditure      ElectricPower     DrinkableWater
##           3.281501           1.501389           3.613139           1.537958
```

The issue with Multicollinearity still exists and there seems to be values with high p-values, let’s try to remove the next predictor with highest p-value, the CO2 emission rate.

```r
model5 <- lm(Life.Expectancy^(1.36) ~ Health.expenditure +
              ElectricPower + DrinkableWater, data = LifeExp)
summary(model5)
```

```
##
## Call:
## lm(formula = Life.Expectancy^(1.36) ~ Health.expenditure + ElectricPower +
##     DrinkableWater, data = LifeExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.709 -12.409   1.599  15.380  53.624
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.469e+02  1.355e+01  10.839  < 2e-16 ***
## Health.expenditure 3.658e+00  9.213e-01   3.970 0.000126 ***
## ElectricPower      2.063e-03  5.588e-04   3.691 0.000343 ***
## DrinkableWater     1.849e+00  1.659e-01  11.145  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22.27 on 115 degrees of freedom
## Multiple R-squared:  0.7502, Adjusted R-squared:  0.7437
## F-statistic: 115.1 on 3 and 115 DF,  p-value: < 2.2e-16
```

```
vif(model5)
```

```
## Health.expenditure       ElectricPower     DrinkableWater
##            1.284438            1.413215           1.451892
```

Our reduced model has a R-squared value of 0.7434, which is higher than our original model. All of the predictors in our model also appears to be significant with a p-value less than 0.05. There are no Multicollinearity issues anymore between the predictor variables as their variance inflation are all less than 2. This appears to be a good model to predict our data.

Results

From our modelling, it is suggested that our best model is model5. This model suggests that the most highly related factors to one's life expectancy in a country is their health expenditure, their electric power consumption, and the amount of people that is using basic water services. We can predict one's life expectancy from a country with the equation:

$$Predicted.LifeExpectancy = 1.469 \times 10^2 + 3.658 \times 10^1 \times HealthExpenditure + \times 10^{-3} \times ElectricPower + 1.849 \times 10^1 \times DrinkableWater$$

Discussion

Overall, through our study, we find that a country is their health expenditure, their electric power consumption, and the amount of people that is using basic water services are the most influential to one's life expectancy.

This result makes sense in the context of our study, we know that filtered water is a very essential component to live a healthy life. Many studies had proved that good hydration is linked with longevity, this research conducted by National Heart, Lung, and Blood Institute had stated found that good hydration will reduce risks for heart failure: [https://www.nhlbi.nih.gov/news/2022/good-hydration-may-reduce-long-term-risks-heart-failure]

To continue, what our country can do for our physical health is definitely another very influential factor. With proper healthcare, vaccines, and annual checkups, we can prevent ourselves from certain diseases, or get treatment when it is still early. Lastly, the usage of electric powers allows one to live a better lifestyle, thus increasing one's life expectancy.

One limitation to our model is that we only has the chance to study the data from the year 2015. Many things would've changed in the world by now, especially after the global pandemic. If possible, it would be a good idea to further expand our study through out the past few years to get a better and more accurate prediction based on the recent data.

Another limitation is that there are still many factors we can consider in our study that haven't been covered, such as smoking, diet on certain type of foods, physical activity level, etc. Many other factors can also significantly contribute to our life expectancy, and unfortunately many of which are not quantitative. We could collect new data by surveying people around the world to rate themselves on a scale of 1~10, but that is too big of a project on a personal level, and would have to be saved for another time.

All in all, we should be appreciative of the advancements we have made in healthcare and other fields. It's crucial to remember that there are still many differences in life expectancy among countries. By making investments and donations in services like healthcare, water filters, and electric power systems, we can help to create a society where everyone has the chance to lead long and happy lives.