Analysis of Life Expectancy

Rameen Usman, Mercedes De La Garza, Faaiz Nadeem

STAT 4355.001

May 2021

Table of Contents Page

Introduction:

The purpose of this project was to utilize R and use the skills we learned in class to analyze a set of data, and make predictions through the use of regression analysis. Our group used a dataset from Kaggle compiled by Kumar Rajarshi under the name “Life Expectancy WHO”. The dataset looked into factors that affected life expectancy including some that were not previously considered in other studies. These factors include immunization factors, mortality factors, economic factors, social factors, as well as other health factors, based on different countries around the world. The dataset itself has 2938 total observations from 193 countries around the world with 22 separate variables. The predictor variables include:

- Country
- Year
- Status - Developed or Developing (Countries)
- Adult Mortality - Probability of dying from age 15-60 per 1000 population
- Infant Deaths - Infant deaths per 1000 population
- Alcohol - Consumption recorded per capita
- Percentage Expenditure - Expenditure on health as a percentage of Gross Domestic Product per Capita %
- Hepatitis B - Immunization coverage among one-year-olds (%)
- Measles - Number of reported cases per 1000 population
- BMI - Average BMI of entire population
- Under five deaths - Number of deaths under age five per 1000 population
- Polio - Pol3 immunization coverage among one-year olds (%)
- Total expenditure - General government expenditure on health as percentage of total government expenditure
- Diphtheria - DTP3 immunization coverage among one-year-olds (%)
- HIV/AIDS - Deaths per 1000 live births HIV/AIDS (0-4 years)
- GDP
- Population
- Thinness 1-19 years - Prevalence of thinness among children and adolescents for age 10-19 (%)
- Thinness 5-9 years - Prevalence of thinness among children and adolescents for age 5-9 (%)
- Income Composition of Resources - Human development index in terms of income composition of resources
- Schooling - Number of years of schooling

Since we wanted to analyze the different factors above and their effect on life expectancy, life expectancy was our response variable for this project. This led to our question that we wanted to answer throughout the duration of this project: “What factors are statistically the most significant when calculating life expectancy globally?”.

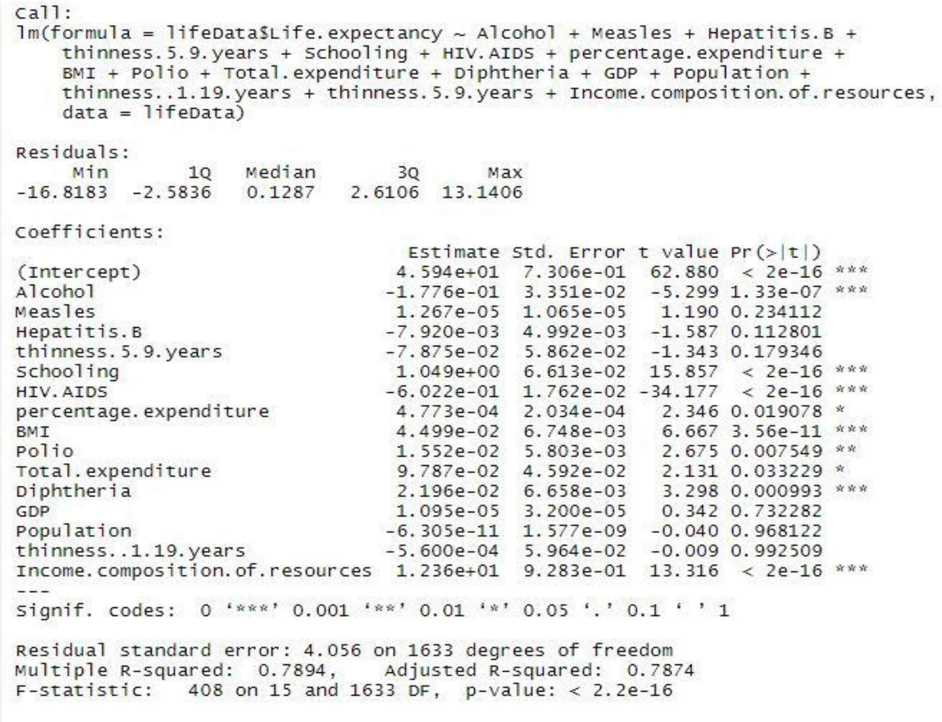
Variable Selection:

When building our full model we included numerous variables that were provided in the dataset. However, we excluded variables such as adult mortality, infant deaths, and other predictor variables that also factor into the calculation of life expectancy. First, we wanted to remove any variable in which there were a lot of null values. We removed variables such as GDP and Population size as we did not have consistent data and we deemed them as insignificant to our model.

We used a backwards selection approach with our remaining variables. We began by creating a model with our remaining variables and fit a regression line. We were able to create a satisfactory regression line however we wanted to be more precise in our variable selection and lower our p-value to a significance level of under 0.05.

Also, we had variables which could be considered too similar to examine individually such as Thinness of 5 to 9 years and Thinness of 1 to 19 years (there are years that overlap in these variables and analyze the same concept). For the sake of diversity in the variables that we would be selecting for our reduced model, we only examined one of the two variables in our full model.

The figure below represents a summary of our regression line which includes all variables from our full model.



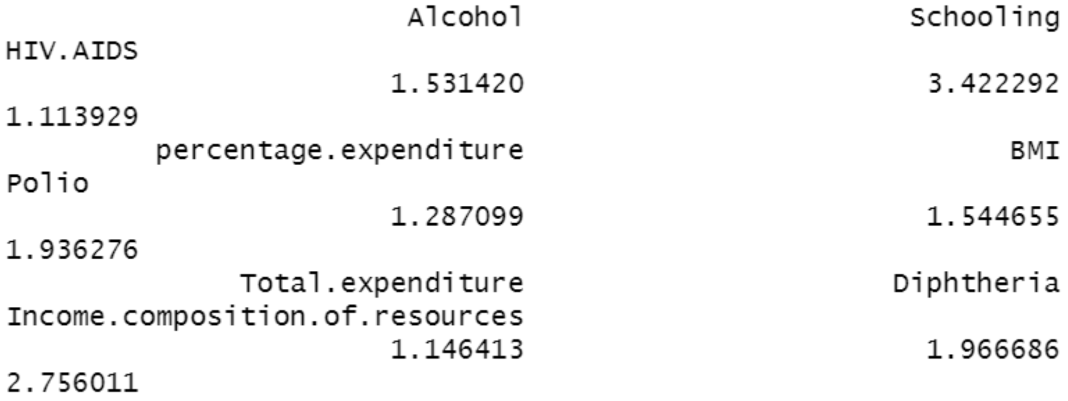
```
Call:
lm(formula = lifeData$Life.expectancy ~ Alcohol + Measles + Hepatitis.B + 
    thinness.5.9.years + Schooling + HIV.AIDS + percentage.expenditure + 
    BMI + Polio + Total.expenditure + Diphtheria + GDP + Population + 
    thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources, 
    data = lifeData)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.8183  -2.5836   0.1287   2.6106  13.1406 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     4.594e+01  7.306e-01  62.880  < 2e-16 ***
Alcohol                        -1.776e-01  3.351e-02  -5.299 1.33e-07 ***
Measles                         1.267e-05  1.065e-05   1.190 0.234112    
Hepatitis.B                    -7.920e-03  4.992e-03  -1.587 0.112801    
thinness.5.9.years             -7.875e-02  5.862e-02  -1.343 0.179346    
Schooling                       1.049e+00  6.613e-02  15.857  < 2e-16 ***
HIV.AIDS                       -6.022e-01  1.762e-02 -34.177  < 2e-16 ***
percentage.expenditure          4.773e-04  2.034e-04   2.346 0.019078 *  
BMI                             4.499e-02  6.748e-03   6.667 3.56e-11 ***
Polio                           1.552e-02  5.803e-03   2.675 0.007549 ** 
Total.expenditure               9.787e-02  4.592e-02   2.131 0.033229 *  
Diphtheria                      2.196e-02  6.658e-03   3.298 0.000993 ***
GDP                             1.095e-05  3.200e-05   0.342 0.732282    
Population                     -6.305e-11  1.577e-09  -0.040 0.968122    
thinness..1.19.years           -5.600e-04  5.964e-02  -0.009 0.992509    
Income.composition.of.resources  1.236e+01  9.283e-01  13.316  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.056 on 1633 degrees of freedom
Multiple R-squared:  0.7894,	Adjusted R-squared:  0.7874 
F-statistic:   408 on 15 and 1633 DF,  p-value: < 2.2e-16
```

Figure 1: Linear model summary

As we were using a backwards selection approach, we removed insignificant variables one at a time until we were left with only significant variables in which we could create our reduced model. After thorough examination and research we concluded that Alcohol, Schooling, HIV.AIDS, Percent Expenditure, BMI, Polio, Total Expenditure, Diphtheria, and Income Composition of Resources would be the best predictor variables to include in our reduced model. These variables were picked based on their individual p-values. The significance level that was used remained 0.05. Hence, the variables that were selected were all under 0.05 and the lowers p-values of the full model.

After verifying that these variables are significant and are sufficient to use in our reduced model we wanted to check for any mulit-collinearity issues.

The figure above represents the variance inflation measures summary for our reduced model. There is no variable in which its inflation measure is greater than 10 so it does not imply serious problems with multicollinearity so we do not need to remove some violating predictors from the model. We can keep all of these variables in our model.



```
                       Alcohol                       Schooling
HIV.AIDS
                      1.531420                        3.422292
1.113929
        percentage.expenditure                             BMI
Polio
                      1.287099                        1.544655
1.936276
             Total.expenditure                      Diphtheria
Income.composition.of.resources
                      1.146413                        1.966686
2.756011
```

Figure 2: variance inflation measure summary for reduced model

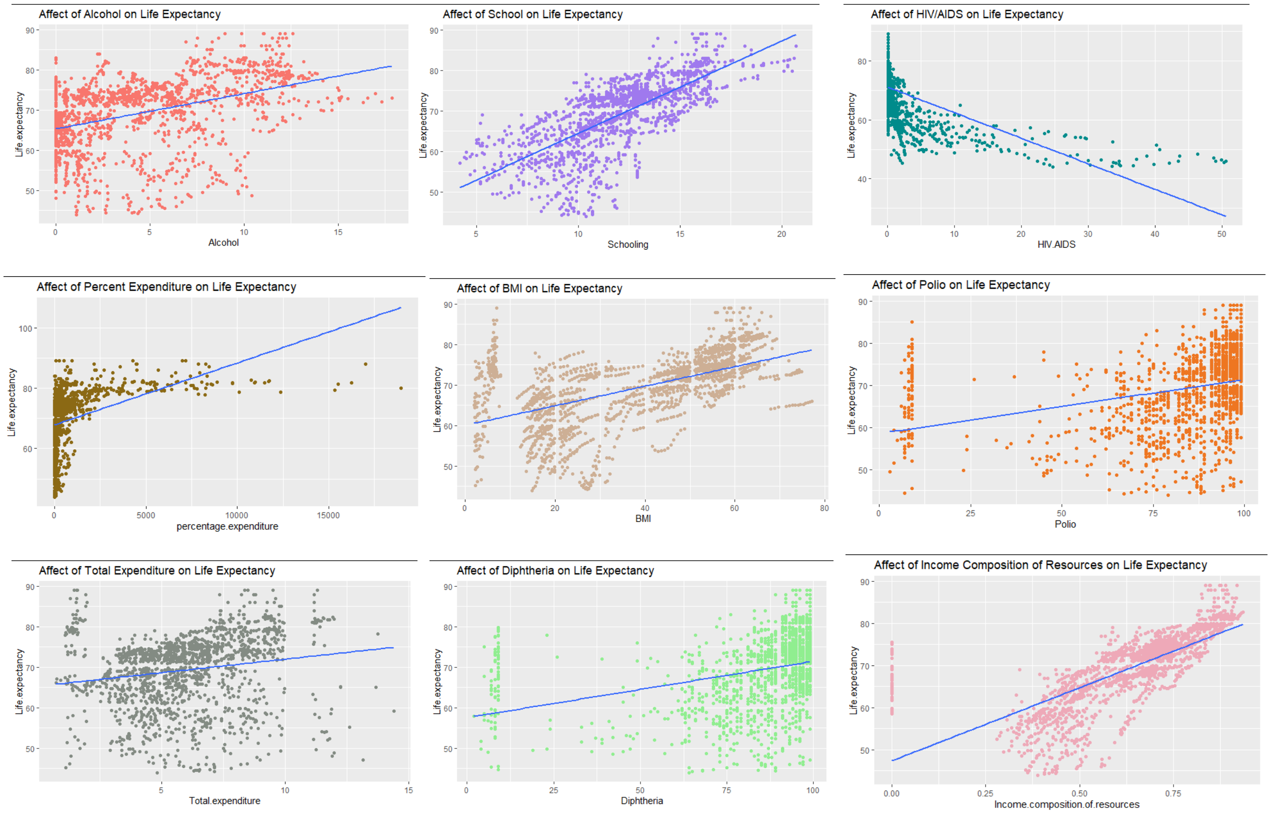
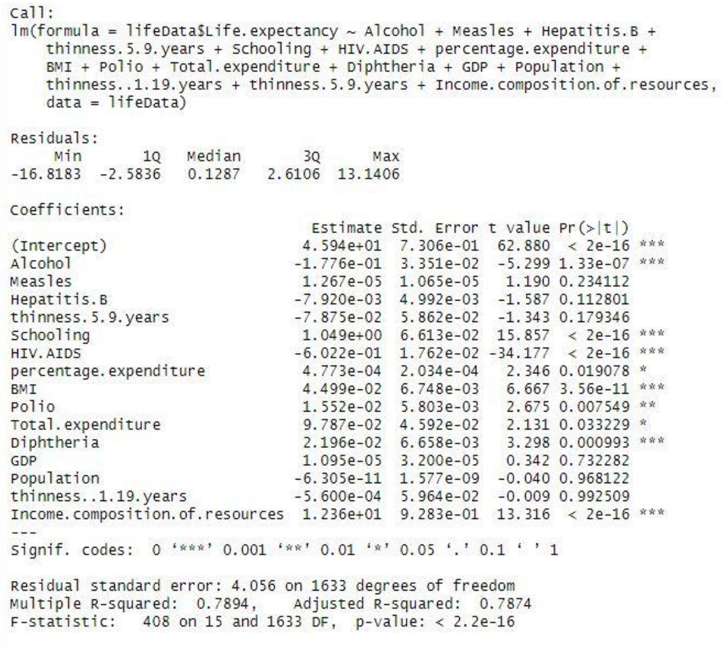


Figure 3: Each variable of reduced model vs. life expectancy scatterplots with regression line

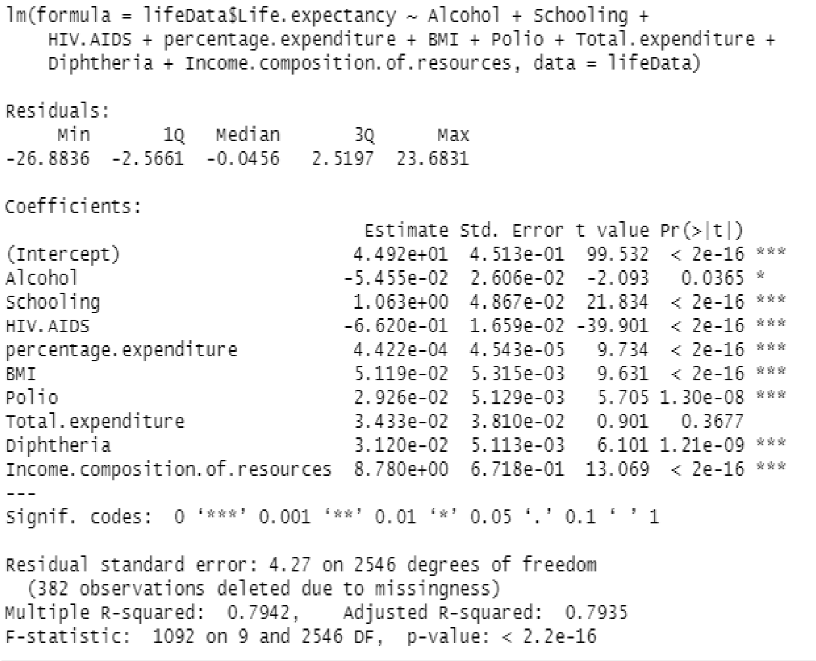
The figure above represents each variable that was selected for the reduced model. Here we are comparing each variable to life expectancy and adding a linear regression line. Many of the graphs that are seen above are quite linear however some variables indicate that the model may need a transformation. Variables such as HIV/AIDS, Percent Expenditure, Polio, Total Expenditure, and Diphtheria are models that can be improved with a transformation however are still sufficient to use in our reduced models evaluation.

Lastly, in order to confirm that these variables are sufficient we can compare the R squared values of the full and reduced models. In the *figures 4* and *5*, we can see the summary outputs of both the full and reduced models.



```
Call:
lm(formula = lifeData$Life.expectancy ~ Alcohol + Measles + Hepatitis.B +
    thinness.5.9.years + Schooling + HIV.AIDS + percentage.expenditure +
    BMI + Polio + Total.expenditure + Diphtheria + GDP + Population +
    thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources,
    data = lifeData)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8183  -2.5836   0.1287   2.6106  13.1406

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.594e+01  7.306e-01  62.880  < 2e-16 ***
Alcohol                         -1.776e-01  3.351e-02  -5.299 1.33e-07 ***
Measles                          1.267e-05  1.065e-05   1.190 0.234112
Hepatitis.B                     -7.920e-03  4.992e-03  -1.587 0.112801
thinness.5.9.years              -7.875e-02  5.862e-02  -1.343 0.179346
Schooling                        1.049e+00  6.613e-02  15.857  < 2e-16 ***
HIV.AIDS                        -6.022e-01  1.762e-02 -34.177  < 2e-16 ***
percentage.expenditure           4.773e-04  2.034e-04   2.346 0.019078 *
BMI                              4.499e-02  6.748e-03   6.667 3.56e-11 ***
Polio                            1.552e-02  5.803e-03   2.675 0.007549 **
Total.expenditure                9.787e-02  4.592e-02   2.131 0.033229 *
Diphtheria                       2.196e-02  6.658e-03   3.298 0.000993 ***
GDP                              1.095e-05  3.200e-05   0.342 0.732282
Population                      -6.305e-11  1.577e-09  -0.040 0.968122
thinness..1.19.years            -5.600e-04  5.964e-02  -0.009 0.992509
Income.composition.of.resources  1.236e+01  9.283e-01  13.316  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.056 on 1633 degrees of freedom
Multiple R-squared:  0.7894,	Adjusted R-squared:  0.7874
F-statistic:   408 on 15 and 1633 DF,  p-value: < 2.2e-16
```

Figure 4: Summary output for full model



```
lm(formula = lifeData$Life.expectancy ~ Alcohol + Schooling +
    HIV.AIDS + percentage.expenditure + BMI + Polio + Total.expenditure +
    Diphtheria + Income.composition.of.resources, data = lifeData)

Residuals:
     Min       1Q   Median       3Q      Max
-26.8836  -2.5661  -0.0456   2.5197  23.6831

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.492e+01  4.513e-01  99.532  < 2e-16 ***
Alcohol                         -5.455e-02  2.606e-02  -2.093   0.0365 *
Schooling                        1.063e+00  4.867e-02  21.834  < 2e-16 ***
HIV.AIDS                        -6.620e-01  1.659e-02 -39.901  < 2e-16 ***
percentage.expenditure           4.422e-04  4.543e-05   9.734  < 2e-16 ***
BMI                              5.119e-02  5.315e-03   9.631  < 2e-16 ***
Polio                            2.926e-02  5.129e-03   5.705 1.30e-08 ***
Total.expenditure                3.433e-02  3.810e-02   0.901   0.3677
Diphtheria                       3.120e-02  5.113e-03   6.101 1.21e-09 ***
Income.composition.of.resources  8.780e+00  6.718e-01  13.069  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.27 on 2546 degrees of freedom
  (382 observations deleted due to missingness)
Multiple R-squared:  0.7942,	Adjusted R-squared:  0.7935
F-statistic:  1092 on 9 and 2546 DF,  p-value: < 2.2e-16
```

Figure 5: Summary output of reduced model

Although our p-value is unchanged it is quite small at 2e-16. We can see that the models we have chosen are both significant. However, our R squared value did increase slightly from our reduced model (R^2 = .7935) from our reduced model (R^2 = .7841). This increase shows that there is a stronger correlation, however more research and analysis would need to be done before making this claim.

Reduced Model Fitting:

After reducing the model by eliminating the variables that had the least amount of influence. We refit the model with the variables Alcohol, Schooling, HIV/AIDS, Percent Expenditure, BMI, Polio, Total Expenditure, Diphtheria, and Income Composition of Resources. Then we plotted the R-student residuals vs. fitted values for the model, which can be seen in *figure 6*. There is more of a concentration of points towards the middle of the plot right around 60 to 80 and this gave the appearance of a double bow in our data. We then generated the QQ plot which can be seen in *figure 7*.There were several data points that were not within the lines, but were not marked by R. There were two specific data points marked by R, Antigua and Haiti, that were not within the acceptable range. To get further information we output the studentized residuals histogram which is *figure 8*. This did show a relatively nice bell curve. However, due to the two other plots we decided to apply a transformation. To correct the nonconstant variance we applied the method of weighted least squares to the reduced model.

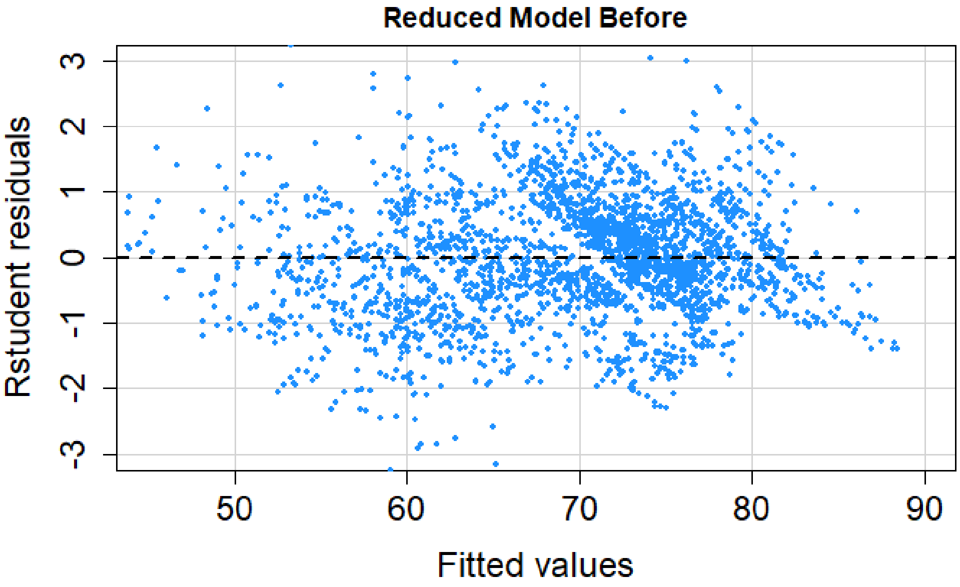


Figure 6: Residuals vs. Fitted plot for reduced model

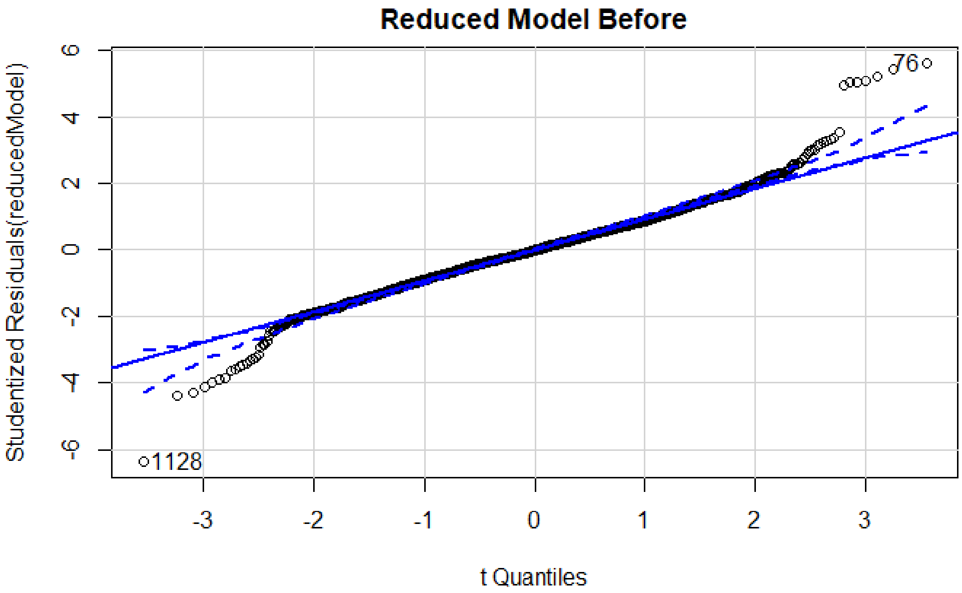


Figure 7: QQ plot for reduced model

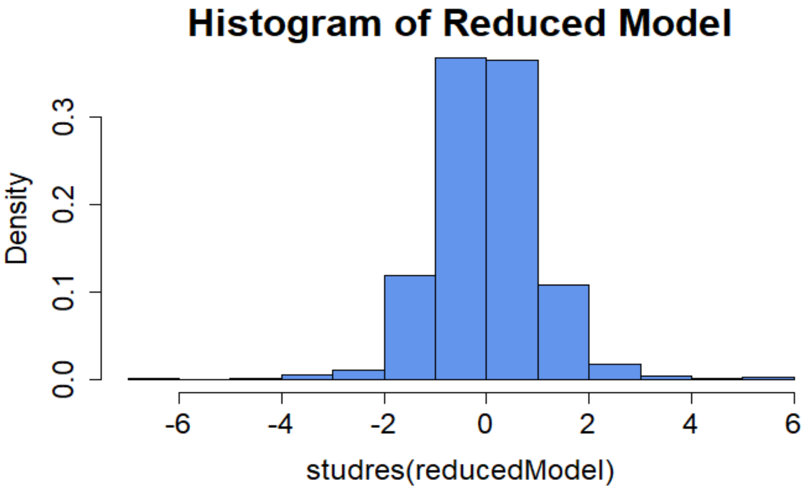


Figure 8: Histogram for reduced model

Explanation of Data Transformations:

The method of weighted least squares is often employed to correct nonconstant variance. So initially we created a linear model that contains all the variables that we chose to keep. From there we fit another linear model with the value of the residuals of the initial model. We inserted the residual values of the initial model as the y variable and all the x variables remained the same. Then we took the absolute value of the new linear model and extracted the fitted values. The final step was to divide 1 by the squared fitted values of the new model, which gave us the weights. The formula for weight can be seen below. From here we created a brand-new linear model just like the first, but we applied the value of the weights to it.

$$weights = \frac{1}{(fitted)^2}$$

Transformed Reduced Model Fitting:

After applying the method of weighted least squares, the R-student residuals vs. fitted values plot looked more scattered which is in *figure 9*. There was less of a concentration of points in the 60 to 80 range and over all the variance seemed more constant. The next plot was the QQ plot which is in *figure 10*. All the points that were previously not within an acceptable range are now within the blue lines indicating that they are no longer being shown as outliers here. There are still two countries that were marked on the QQ plot, Jordan and Chad, but they are within the boundary of acceptance. Then we output the studentized residuals histogram which is *figure 11*. This shows that there is still a bell curve for the data. After generating these three plots we decided that the transformation was adequate.

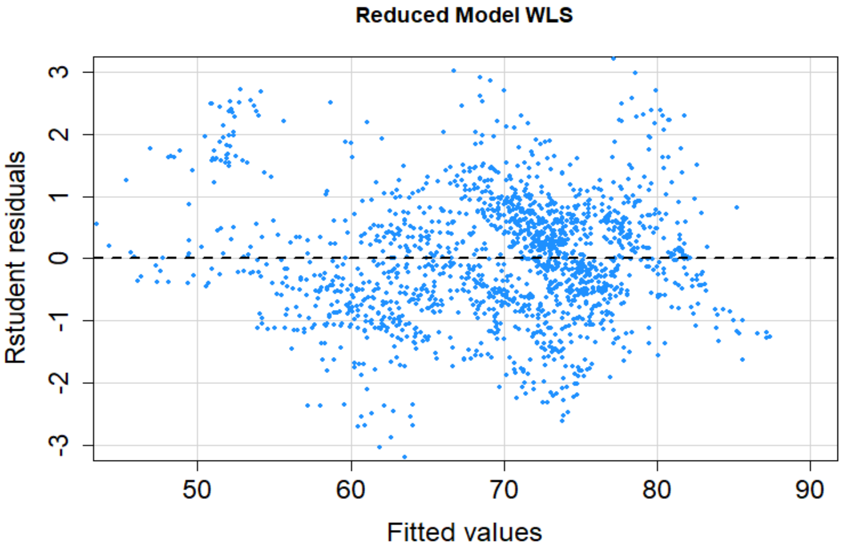


Figure 9: Residuals vs. Fitted plot for the reduced model with variable transformations

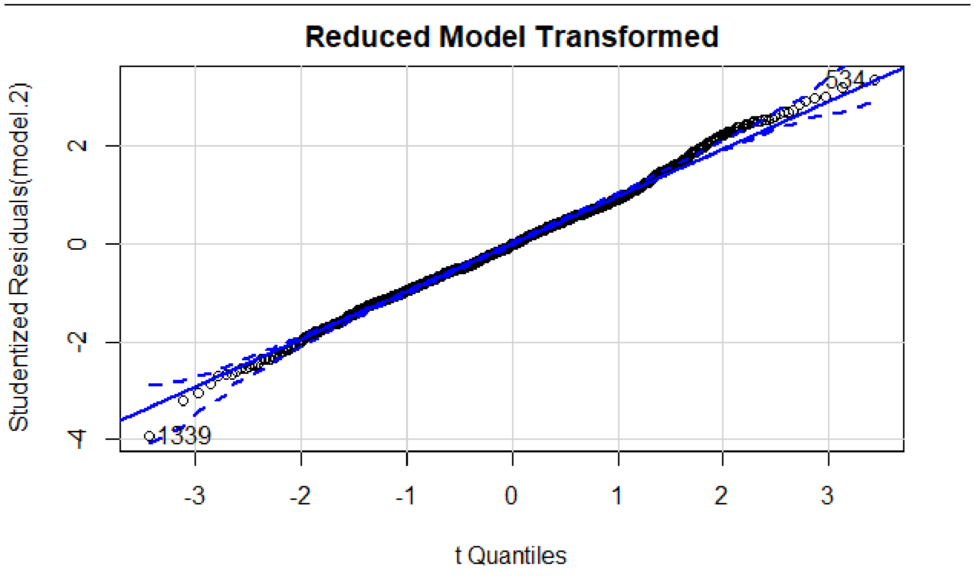


Figure 10: QQ plot for the reduced model with variable transformations

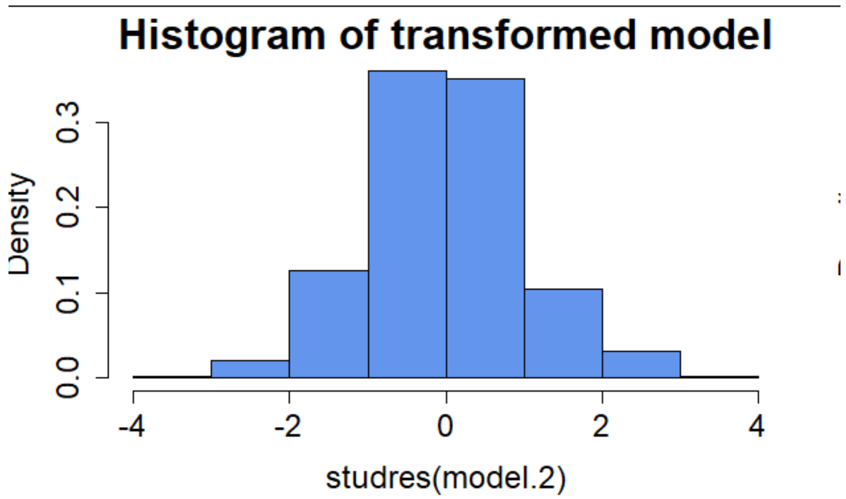


Figure 11: Histogram for the reduced model with variable transformations

Residual Analysis:

Now, we want to assess the validity of our linear regression model by defining residuals and creating residual graphs. We will look at multiple perspectives in order to find outliers and any inconsistencies in our reduced models.

Firstly, we can look at our standardized, studentized, and Rstudent residual graphs. The figures below are barplots of these respective residual graphs with a line of significance at 3 for standardized and studentized and 3.5 line of significance for Rstudent residual graphs. If any of the observations cross this line we can consider it an outlier.

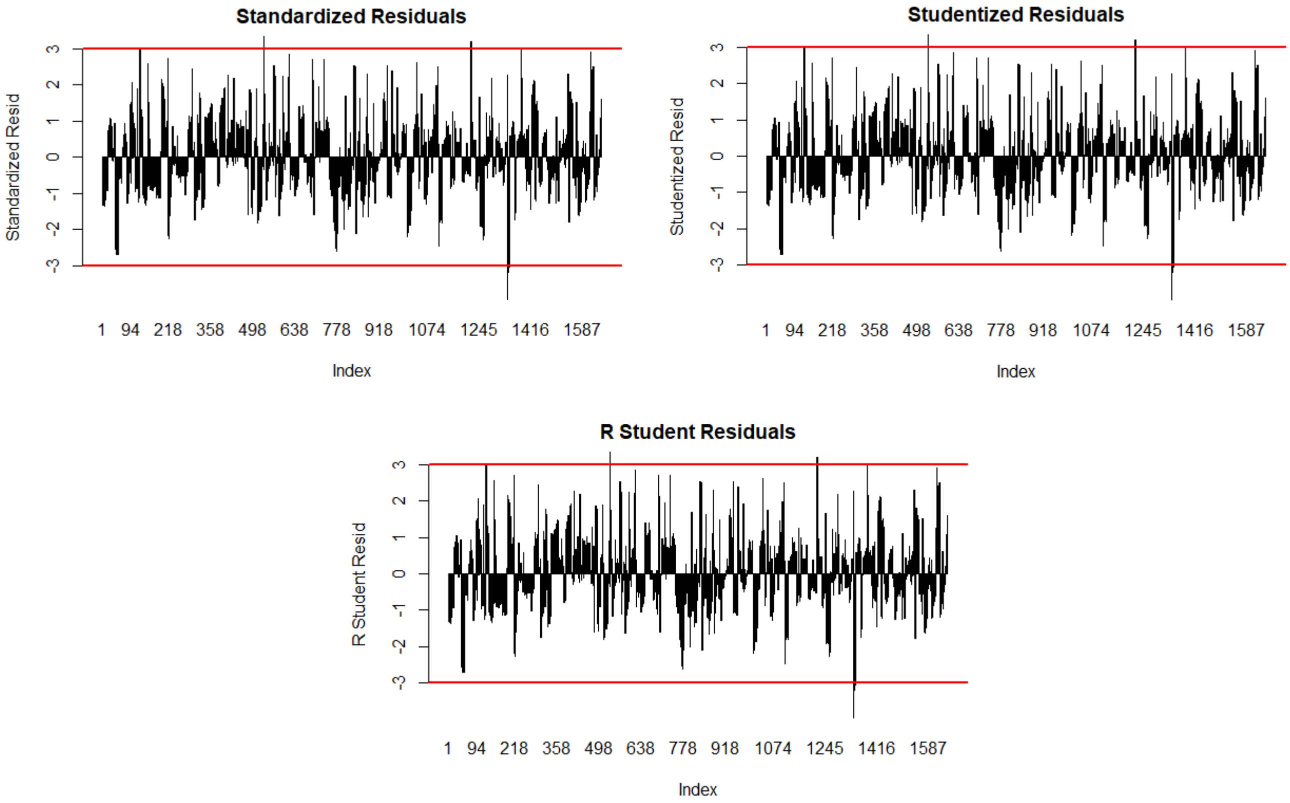


Figure 12: Bar plots of standardized, studentied, and Rstudent residual graphs

Standardized residual plots estimate error of a particular point which accounts for the leverage/influence of the point. Studentized residuals for a given point calculate a model fit to every other point except for the observation that it is checking. We can see here that we do have a few outliers in our dataset through this bar plot. The R-student residual re-normalized residuals to a unit variance to measure the error similar to the methodology that the other two graphs use.

Observations between 1245 and 1416, 498 and 638, are seen to be outliers for our reduced regression model. However, since most of our observations lie within the normal range we do not see this as a problem in our model. We can consider removing these observations in order to create a more precise linear model. These countries could have had other factors at the time which influenced these observations to be different from the rest of the dataset and will need further research and analysis to better understand the reasoning.

Next, we wanted to analyze our outliers further so we created DFBETAS plots. We are able to see the effect of deleting each observation on the estimate of regression coefficients. Now, we can see if we have any outliers and how that will affect each variable individually.

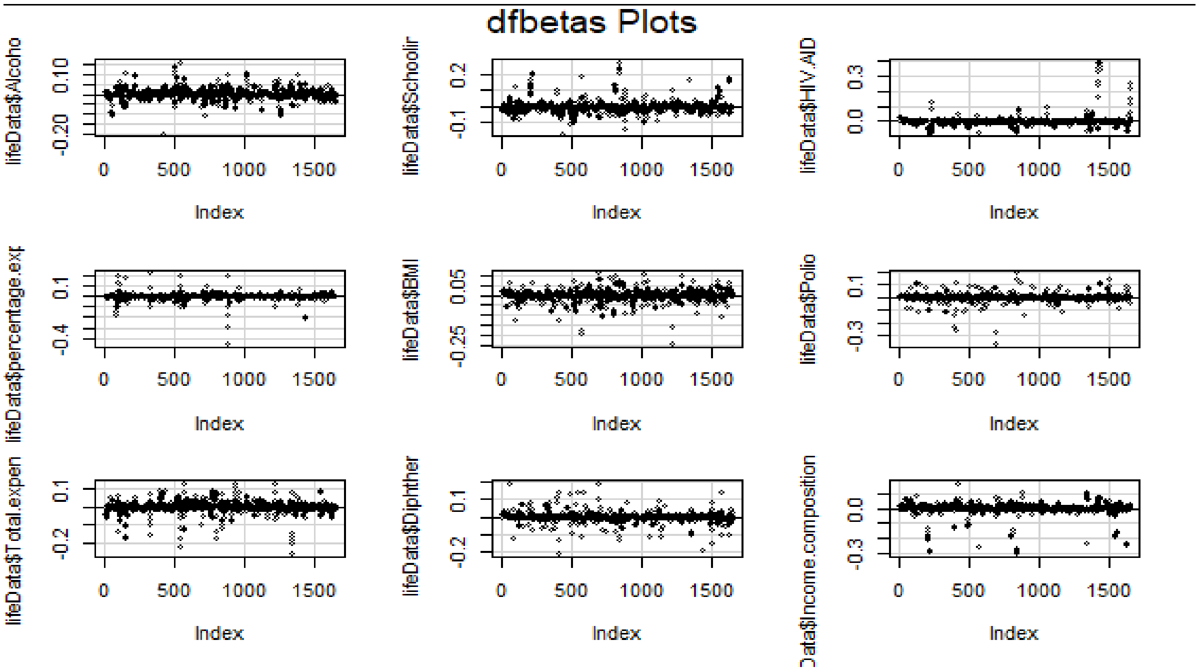


Figure 13: DFBETAS plot for each regression coefficient

Looking at the DFBETAS plots, we are able to find outliers more accurately. In the graphs above we can see that the range of these points is relatively small with the large range being from 0.50 (percent expenditure range is -0.4-0.1). This implies that there are very few (if any) outliers. We can see that most of the plots are concentrated around 0 which implies that these fit the models adequately. HIV.AIDS seems to imply the most irregularity as it has the largest range and most points that are outside of the normal range.

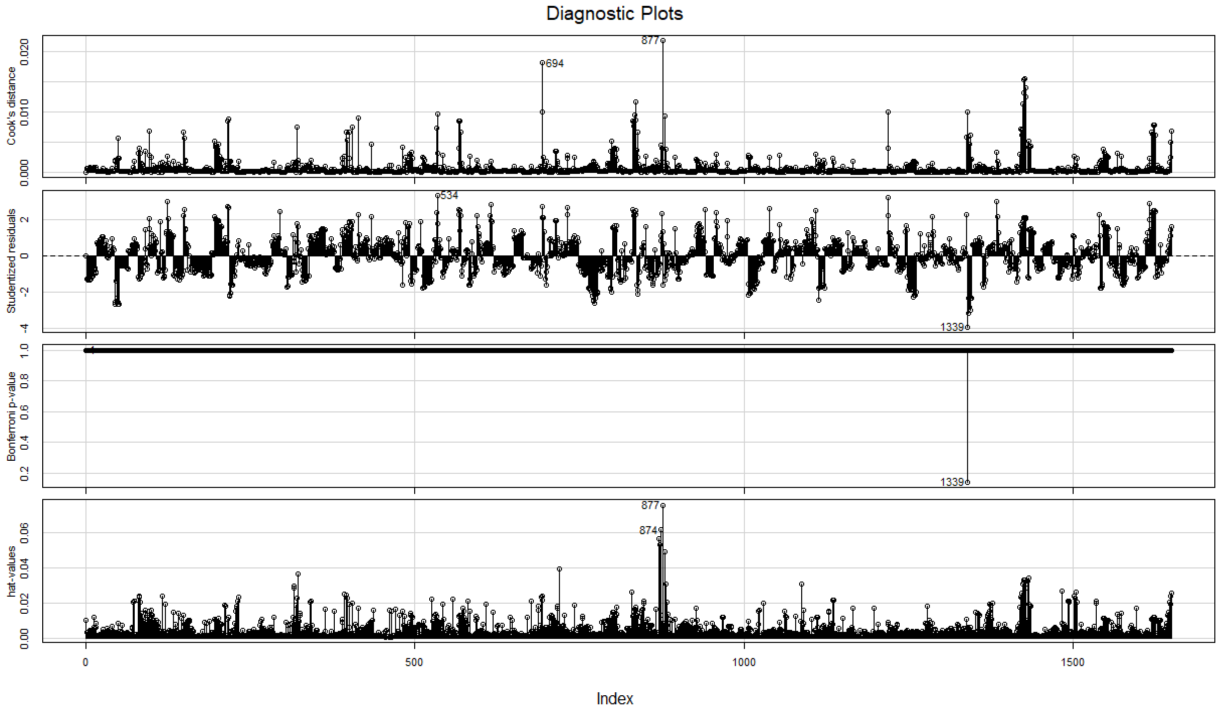


Figure 14: Diagnostic plot for each regression coefficient

Our diagnostic plots show that we have some common points specifically on the Cook's and Hat's plots such as 877 which represent Estonia. Observation numbers 1339, 874, 534, and 694 represent Jordan, Estonia, Chad, and Czechia respectively, showing us that they are possible prediction errors. However since we do not have any points that are the same across Cooks, Hats, and Studentized, it is safe to say that we do not have any points that we have to remove/test for further analysis. We can see from our Cook's plots all of our points are less than one so it does not warrant examination.

Conclusion:

After a thorough analysis of the dataset, we have come to the conclusion that the variables presented from our reduced model are sufficient in predicting life expectancy of individuals on a global scale. With our reduced model we saw that across the world life expectancy has overall a positive trend with alcohol consumption, percentage expenditure, BMI, Polio immunizations, total expenditure, Diphtheria immunizations, and income composition of resources, but a negative trend with HIV/AIDS which was expected. When performing more thorough analysis of residual plots and dfbetas we discovered that although we saw some outliers, the majority were found within the normal range or error and did need to be examined further.

With data analysis we also answered other questions we had regarding the effects of alcohol and schooling on life expectancy. For schooling our discovery was what we expected. As the years of schooling that a person takes increases, their overall life expectancy also increases.
This could be due to many factors such as a possible low literacy rate in less developed countries which have poor health conditions. Our residual plot shows that there is also a normal distribution between these two variables

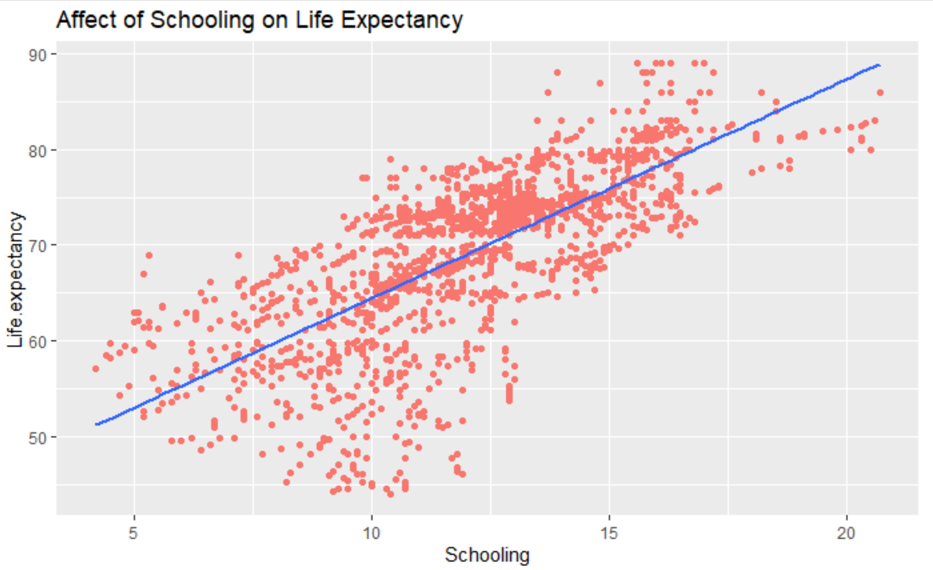


Figure 15: plot for life expectancy vs schooling

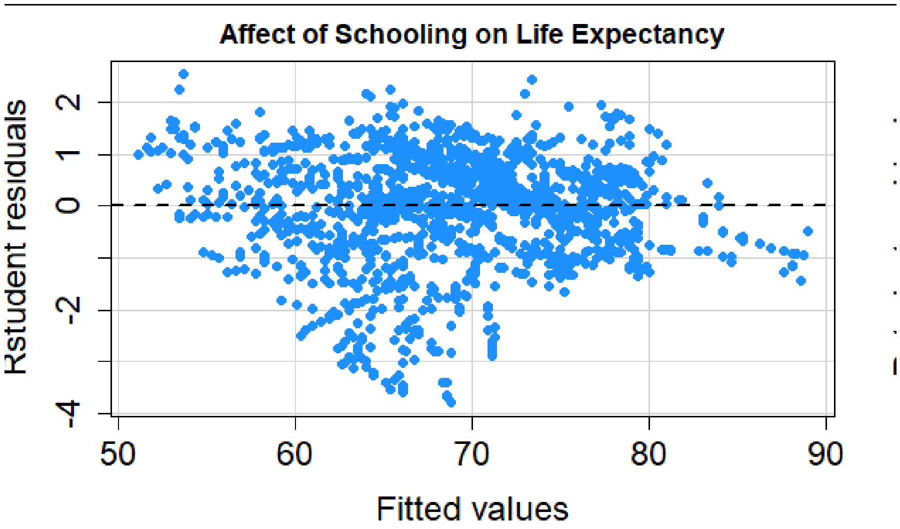


Figure 16: Residuals vs. Fitted plot for schooling

However for alcohol, our expectations did not match the results. As the liters of alcohol a person drank increased, their life expectancy also seemed to increase.

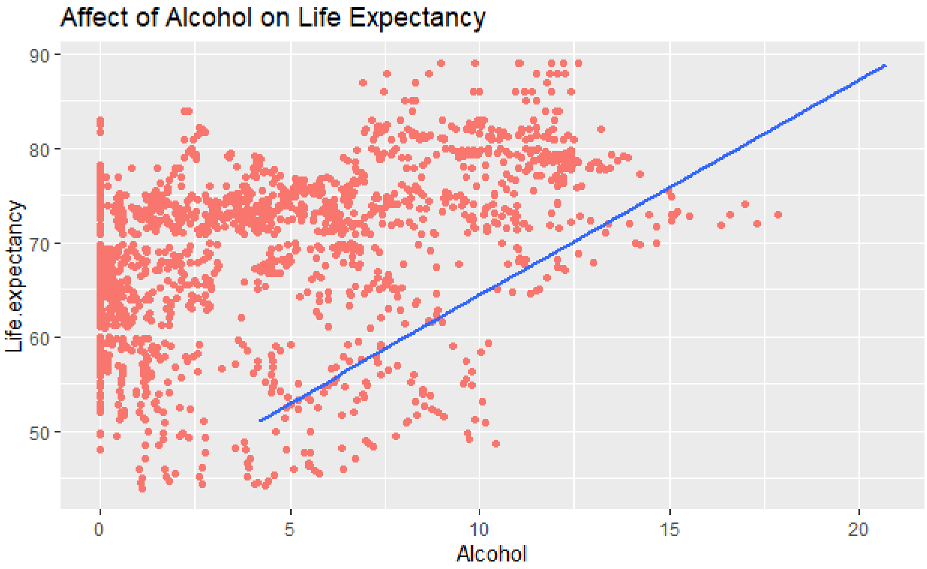


Figure 17: plot for life expectancy vs alcohol

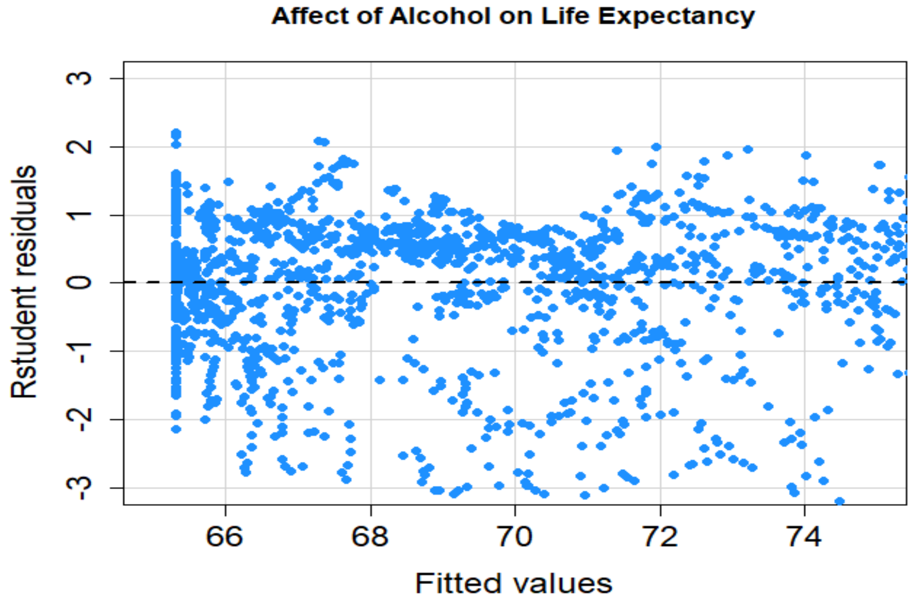


Figure 16: Residuals vs. Fitted plot for alcohol

We believe this could be due to a number of reasons as alcohol is more available and affordable in more developed countries which have a higher life expectancy. Also we can see that the points tend to scatter more after 5 liters of alcohol.

Overall, we still believe that we can analyze more with this dataset and are looking forward to doing so in the future.

Reflection:

Variable selection and analysis with our reduced model was one of our major successes for this project because it allowed us to come to our conclusions for this dataset. The only major difficulty encountered was making transformations on the reduced model to create a more linear outcome. Additional ideas we have for future analysis includes analyzing the difference in life expectancy between developing and developed countries as our dataset also has a variable indicating whether the observation is from a developing or developed country. We could also determine the difference in life expectancy based on the continent. Alongside with examining recent years to determine if life expectancy has increased in recent years since the dataset cut off in 2015.

Appendix A:

Team Member Roles:

- Rameen Usman: Residual Analysis and Variable Selection
- Mercedes De La Garza: Model Fitting and Transformations
- Faaiz Nadeem: Dataset Information, Reflection, and Conclusion

Code:

Reading the file:

```
library(readr)
library(tidyverse)
library(ggplot2)
library(reshape2)
library(MASS)
library(car)
library(olsrr)

lifeData <- read.csv("Life Expectancy Data.csv")
summary(lifeData)

lifeData <- na.omit(lifeData)

full.model <-
lm(lifeData$Life.expectancy~Alcohol+Measles+Hepatitis.B+Schooling+

HIV.AIDS+percentage.expenditure+BMI+Polio+Total.expenditure+
                   Diphtheria+GDP+Population+thinness..1.19.years+

thinness.5.9.years+Income.composition.of.resources,
                 data = lifeData)

reducedModel <-
lm(lifeData$Life.expectancy~Alcohol+Schooling+HIV.AIDS+percentage.exp
enditure+BMI+Polio+Total.expenditure+Diphtheria+Income.composition.of
.resources, data = lifeData)
```

Figure A.1:

```
summary(full.model)
```

Figure A.2:

```
vif(reducedModel)
```

Figure A.3:

```
alcoholPlot <- ggplot(lifeData) + geom_point(aes(Alcohol,
Life.expectancy, colour = "red")) +
  geom_smooth(aes(Alcohol, Life.expectancy), method = lm, se = FALSE)
+
  ggtitle("Affect of Alcohol on Life Expectancy") +
  theme(legend.position = "none")

alcoholPlot

schoolPlot <- ggplot(lifeData) + geom_point(aes(Schooling,
Life.expectancy), color = "mediumpurple2") +
  geom_smooth(aes(Schooling, Life.expectancy), method = lm, se =
FALSE) +
  ggtitle("Affect of School on Life Expectancy") +
  theme(legend.position = "none")

schoolPlot

hivPlot <- ggplot(lifeData) + geom_point(aes(HIV.AIDS,
Life.expectancy), color = "darkcyan") +
  geom_smooth(aes(HIV.AIDS, Life.expectancy), method = lm, se =
FALSE) +
  ggtitle("Affect of HIV/AIDS on Life Expectancy") +
  theme(legend.position = "none")

hivPlot

percentPlot <- ggplot(lifeData) +
geom_point(aes(percentage.expenditure, Life.expectancy), color =
"goldenrod4") +
  geom_smooth(aes(percentage.expenditure, Life.expectancy), method =
lm, se = FALSE) +
  ggtitle("Affect of Percent Expenditure on Life Expectancy") +
  theme(legend.position = "none")

percentPlot

bmiPlot <- ggplot(lifeData) + geom_point(aes(BMI, Life.expectancy),
color = "peachpuff3") +
  geom_smooth(aes(BMI, Life.expectancy), method = lm, se = FALSE) +
  ggtitle("Affect of BMI on Life Expectancy") +
  theme(legend.position = "none")

bmiPlot

polioPlot <- ggplot(lifeData) + geom_point(aes(Polio,
Life.expectancy), color = "chocolate2") +
  geom_smooth(aes(Polio, Life.expectancy), method = lm, se = FALSE) +
  ggtitle("Affect of Polio on Life Expectancy") +
  theme(legend.position = "none")

polioPlot

totalPlot <- ggplot(lifeData) + geom_point(aes(Total.expenditure,
Life.expectancy), color = "honeydew4") +
  geom_smooth(aes(Total.expenditure, Life.expectancy), method = lm,
se = FALSE) +
  ggtitle("Affect of Total Expenditure on Life Expectancy") +
  theme(legend.position = "none")

totalPlot

diphPlot <- ggplot(lifeData) + geom_point(aes(Diphtheria,
Life.expectancy), color = "palegreen2") +
  geom_smooth(aes(Diphtheria, Life.expectancy), method = lm, se =
FALSE) +
  ggtitle("Affect of Diphtheria on Life Expectancy") +
  theme(legend.position = "none")

diphPlot

incomePlot <- ggplot(lifeData) +
geom_point(aes(Income.composition.of.resources, Life.expectancy),
color = "pink2") +
  geom_smooth(aes(Income.composition.of.resources, Life.expectancy),
method = lm, se = FALSE) +
  ggtitle("Affect of Income Composition of Resources on Life
Expectancy") +
  theme(legend.position = "none")

incomePlot
```

Figure A.4:

```
summary(full.model)
```

Figure A.5:

```
summary(reducedModel)
```

Figure A.6:

```
residualPlot(reducedModel, type="rstudent", quadratic=F, col =
"dodgerblue", pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5, ylim=c(-3,
3), main = "Reduced Model Before")
```

Figure A.7:

```
qqPlot(reducedModel, main = "Reduced Model Before")
```

Figure A.8:

```
hist(studres(reducedModel), breaks = 10, freq = F,
col="cornflowerblue",
     cex.axis = 1.5, cex.lab=1.5, cex.main = 2, main = "Histogram of
reduced model")
```

Figure A.9:

```
wts <-
1/fitted(lm(abs(residuals(reducedModel))~lifeData$Alcohol+lifeData$Sc
hooling+lifeData$HIV.AIDS+lifeData$percentage.expenditure+lifeData$BM
I+lifeData$Polio+lifeData$Total.expenditure+lifeData$Diphtheria+lifeD
ata$Income.composition.of.resources))^2

model.2 <- lm(lifeData$Life.expectancy ~
lifeData$Alcohol+lifeData$Schooling+lifeData$HIV.AIDS+lifeData$percen
tage.expenditure+lifeData$BMI+lifeData$Polio+lifeData$Total.expenditu
re+lifeData$Diphtheria+lifeData$Income.composition.of.resources,
weights=wts)

residualPlot(model.2, type="rstudent", quadratic=F,
col = "dodgerblue", pch=16, cex=0.5,
cex.axis=1.5, cex.lab=1.5, ylim=c(-3, 3), main = "Reduced Model WTS")
```

Figure A.10:

```
qqPlot(model.2, main = "Reduced Model Transformed")
```

Figure A.11:

```
hist(studres(model.2), breaks = 10, freq = F, col="cornflowerblue",
     cex.axis = 1.5, cex.lab=1.5, cex.main = 2, main = "Histogram of
transformed model")
```

Figure A.12:

```
barplot(height = stdres(model.2), main = "General Standardized
Residuals",
xlab = "Index", ylab = "Standardized Resid", ylim= c(-5,5))
abline(h = 3, col = "Red", lwd = 2)
abline(h = -3, col = "Red", lwd = 2)

barplot(height = studres(model.2), main = "Studentized Residuals",
xlab = "Index",
ylab = "Studentized Resid", ylim=c(-5,5))
# Add cutoff values.
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)


RStudent <- rstudent(model.2)
RStudent
#Find the range of the values.
range(RStudent)
#Set the range of y axis with argument ylim.
barplot(height = RStudent,
main = "R Student Residuals", xlab = "Index",
ylab = "R Student Resid", ylim=c(-5,5))
cor.level <- 0.05/(2*28)
cor.qt <- qt(cor.level, 24, lower.tail=F)
cor.qt
RStudent> cor.qt
abline(h=cor.qt , col = "Red", lwd=3)
abline(h=-cor.qt , col = "Red", lwd=3)
```

Figure A.13:

```
dfbetasPlots(model.2,intercept=F)
```

Figure A.14:

```
influenceIndexPlot(model.2)
```

Figure A.15:

```
schoolPlot <- ggplot(lifeData) + geom_point(aes(Schooling,
Life.expectancy), color = "mediumpurple2") +
  geom_smooth(aes(Schooling, Life.expectancy), method = lm, se =
FALSE) +
  ggtitle("Affect of School on Life Expectancy") +
  theme(legend.position = "none")

schoolPlot
```

Figure A.16:

```
genSchool <- lm(lifeData$Life.expectancy~Schooling, data = lifeData)

residualPlot(genSchool, quadratic=F, col = "dodgerblue",
             pch = 16, cex = 1, cex.axis=1.5, cex.lab=1.5,
             main = "Affect of Schooling on Life Expectancy")
```

Figure A.17:

```
alcoholPlot <- ggplot(lifeData) + geom_point(aes(Alcohol,
Life.expectancy, colour = "red")) +
  geom_smooth(aes(Alcohol, Life.expectancy), method = lm, se = FALSE)
+
  ggtitle("Affect of Alcohol on Life Expectancy") +
  theme(legend.position = "none")

alcoholPlot
```

Figure A.18:

```
genAlc <- lm(lifeData$Life.expectancy~Alcohol, data = lifeData)

residualPlot(genAlc, quadratic=F, col = "dodgerblue",
             pch = 16, cex = 1, cex.axis=1.5, cex.lab=1.5,
             main = "Affect of Alcohol on Life Expectancy")
```

References:

Rajarshi ,Kumar.(2018, February). Life Expectancy (WHO): Statistical Analysis on factors influencing Life Expectancy, Version 1. Retrieved March 29, 2021 from https://www.kaggle.com/kumarajarshi/life-expectancy-who/metadata.