**An Analysis and Application of Linear Regression**

John Ayres, Sam Malnati, Sebastian Staskiewicz

Statistics 490: Design Of Experiments

Rutgers University

v. 4-2022

**An Analysis and Application of Linear Regression**

John Ayres, Sam Malnati, Sebastian Staskiewicz

Stat 490: Rutgers University

24 April 2022

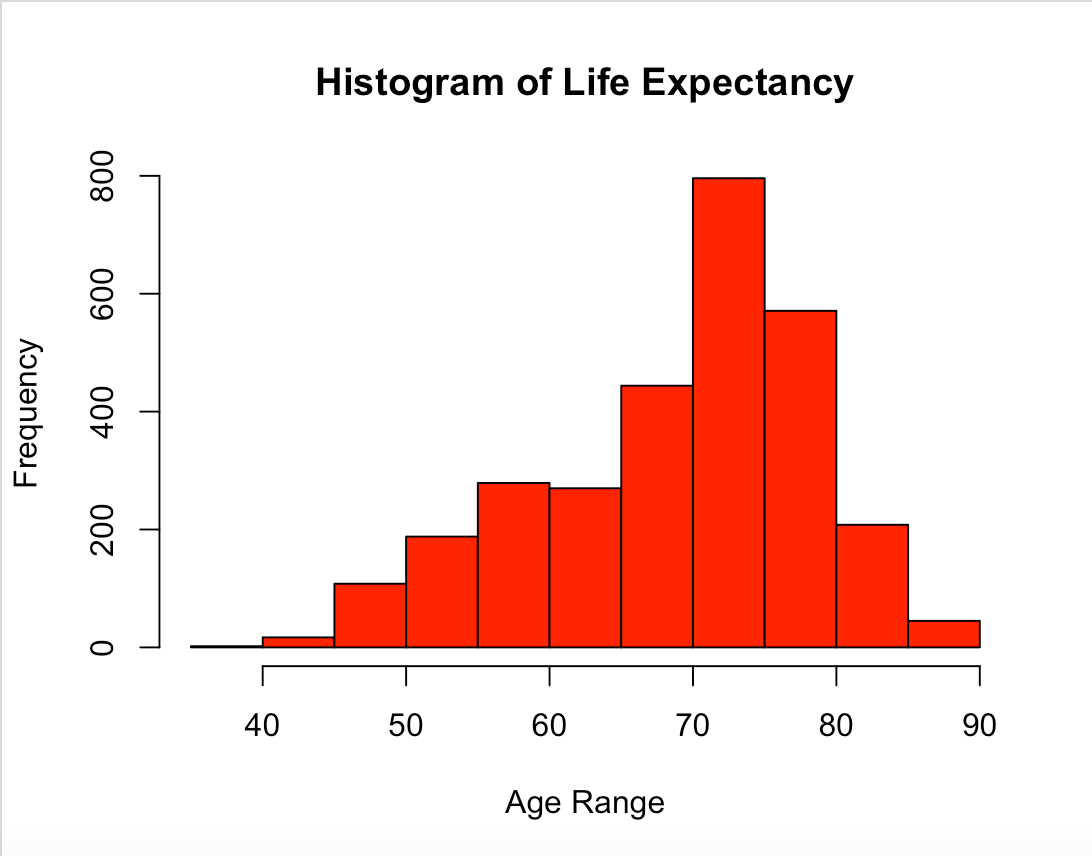
**ABSTRACT**

We chose to further study and apply linear regression to real world applications. We chose a data set from the WHO (World Health Organization) which dealt with predicting life expectancy. The data set was large with many abnormalities, our challenge was to create a linear regression model which would be able to best predict the life expectancy dependent upon many variables. We tackled this by employing a multiple linear regression model and analyzing the residuals to produce the best possible model which we could obtain.

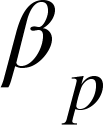
**INTRODUCTION**

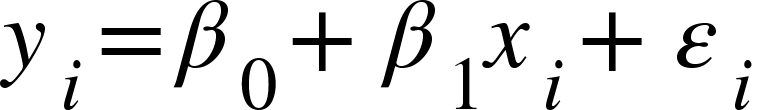
A topic which interested all of us was linear regression since it is something that is heavily relied on and applied to many aspects in the real world. We specifically chose to focus on a data set from the WHO (World Health Organization) which allows us to predict life expectancy using linear regression. This would allow us to model and see how independent variables, such as alcohol consumption or income level could affect life expectancy. Furthermore, you are able to interpret the different coefficients associated with a variable so you know how a singular change in that x-value affects the dependent variable (life expectancy). While making our linear regression model we can also see which terms are interaction terms and develop a more sophisticated model to predict better results, and then by ultimately putting in our own information from the data set we would be able to predict average life expectancy.

The problem is we have a data set which contains many independent variables all trying to predict the life expectancy of an individual. Given all of this valuable information we have to create a model which is able to best predict the average life expectancy of an individual dependent on multiple factors. We must also see which terms would be considered statistically significant and possibly interact with each other. After creating our model we would then have to make sure no assumptions are violated and be able to visualize the model. Once we finalize our model with which variables we would like to use we then have to be able to interpret the coefficients corresponding to the variables to understand how a specific variable is affecting the life expectancy. In addition to this another problem we have to be able to combat is the fact that there are many missing data points for certain variables as the data could not be collected or was difficult to obtain by the original owner of the data. This also leads us to believe that there may be outliers in the data which could have been caused by a previous miscalculation or typing error. Assuming this we would attempt to eliminate the outliers to get a more accurate model.

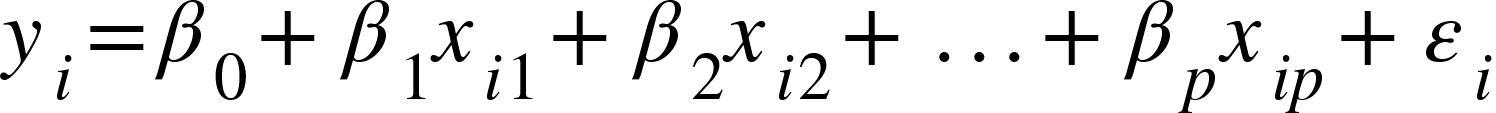


**METHODS**

Our method of analysis is, as suggested prior, linear regression. Represented in the equations outlined below, our predictor value’s(  ) relationship to the response variable is explained by our regression coefficient .Additionally, included in the model is an error term().



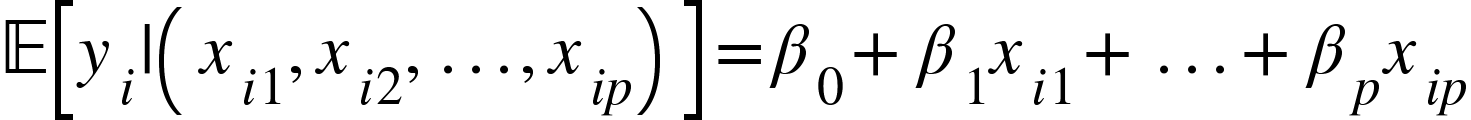
In this form the equation is very clear and easy to understand, however it is not particularly useful, Therefore we may extend our simple linear regression(SLR) equation to multiple linear regression(MLR).



The MLR model extends our linearly related equation, allowing for the use of predictors.

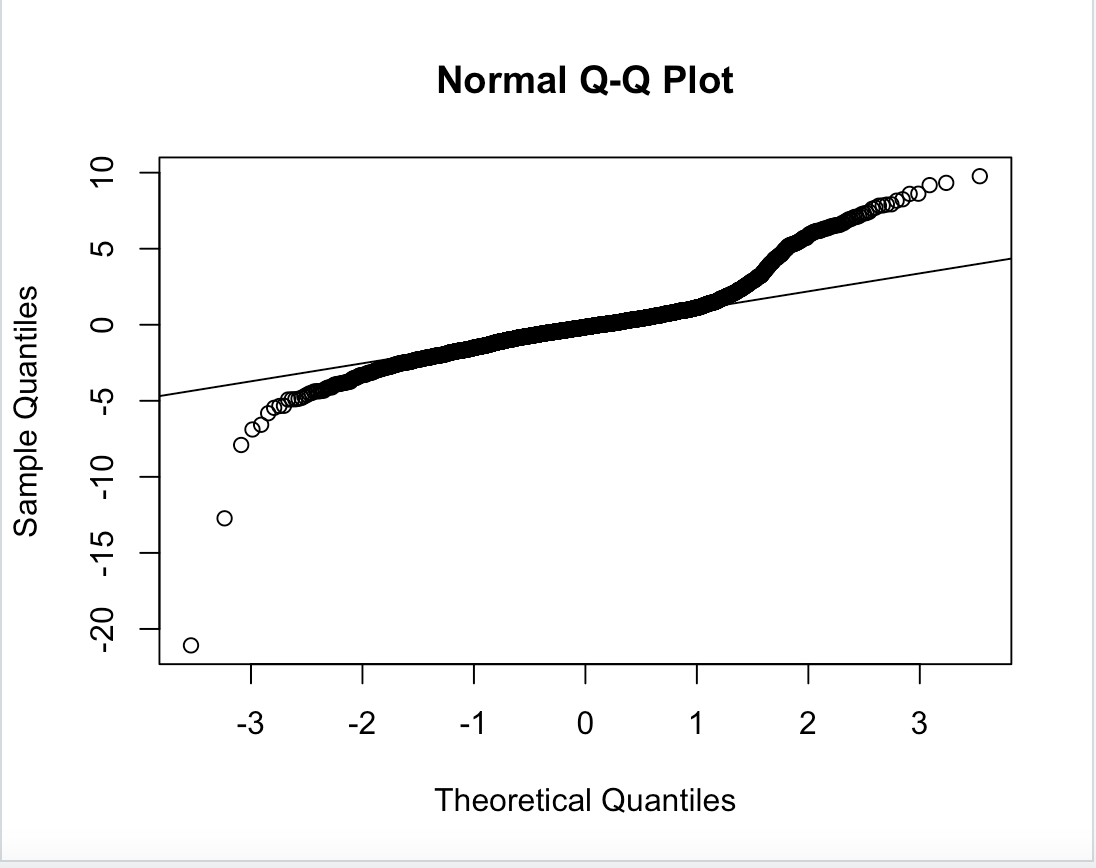
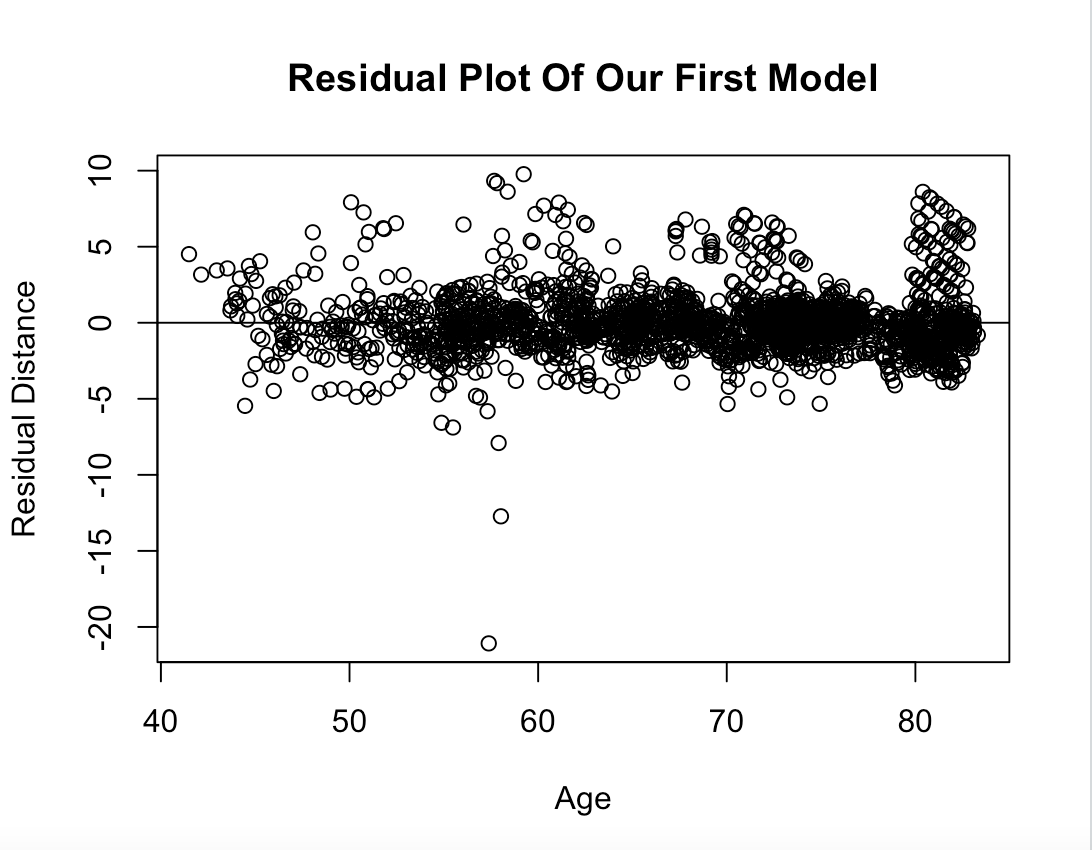
***Assumptions of model***

A linear regression model makes several assumptions for the model to be adequate. First, it is assumed the errors are independent of predictor variables. Second, it is assumed observations are independent. Third, it is assumed that the error terms are normally distributed. Last, Linear regression makes the assumption that the mean of the response  is linearly related to the predictors. That is,

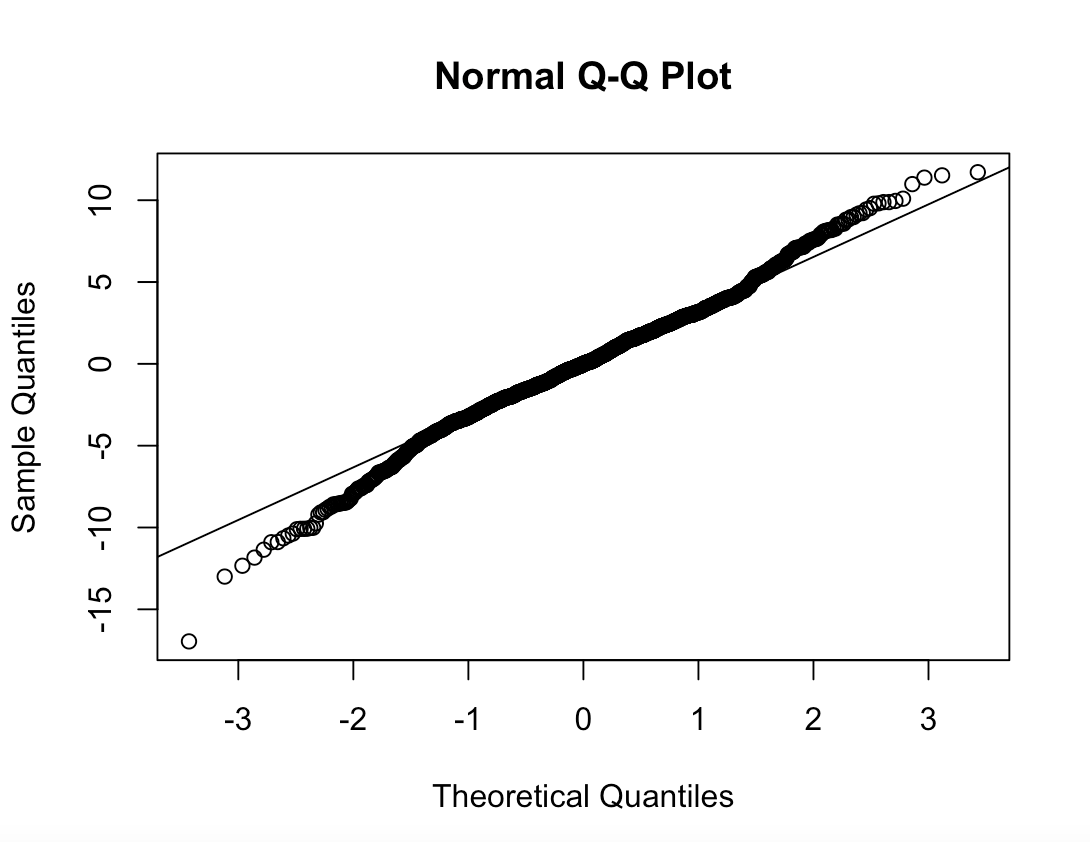


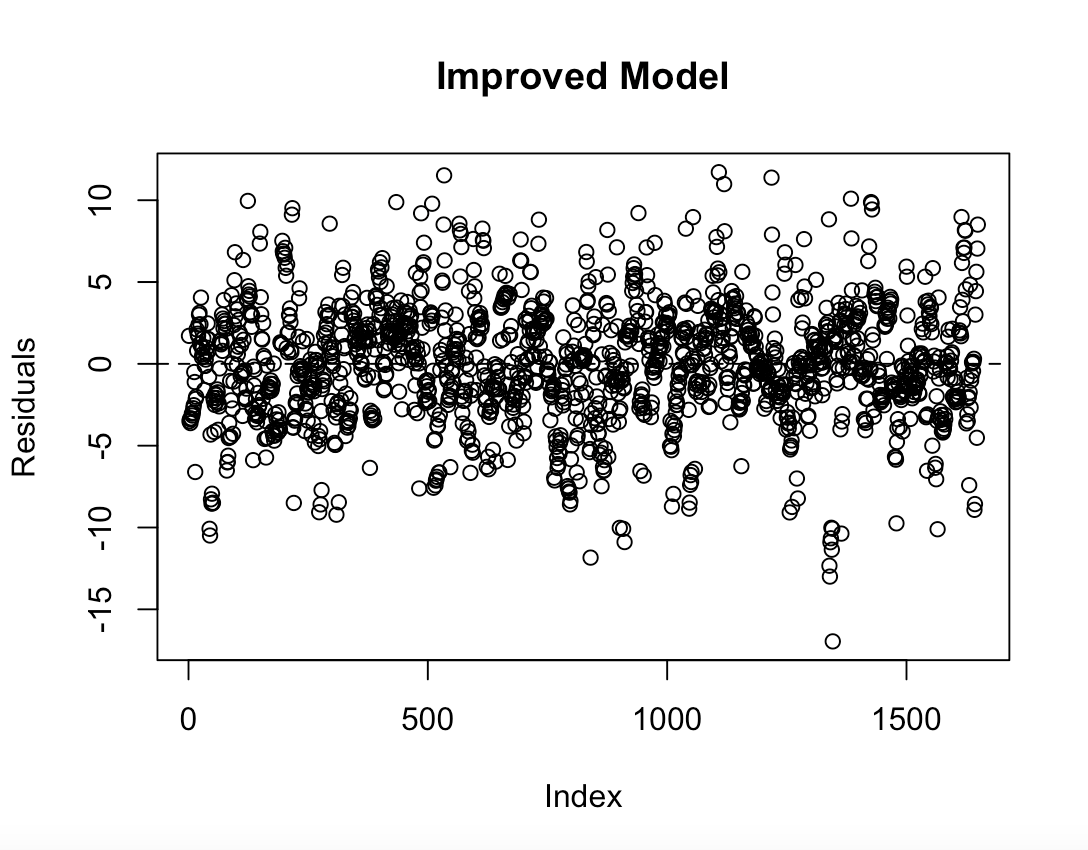
**RESULTS**

Running our very first simple linear regression, we get an extremely high adjusted R^2 value of about 0.9575. This would at first glance indicate a good result and one may assume this an extremely effective model at predicting life expectancy. Although, after further investigation we see this is not the case. The data is still filled with outliers, has too many predictor variables, and the residuals do not seem to be normally distributed.

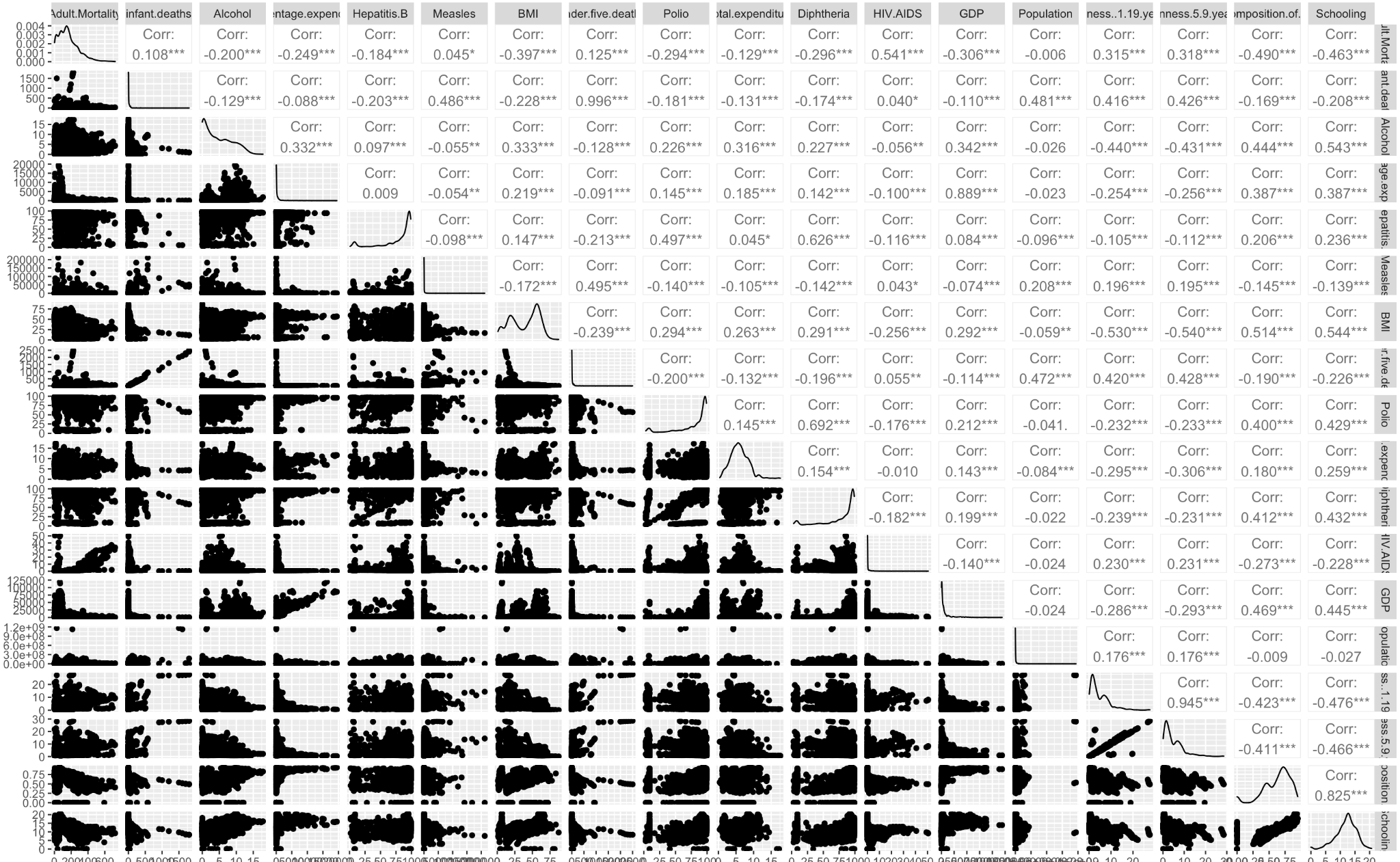


It is evident that there are many outliers still present in the data set and the residuals corresponding to the older individuals all seem to be positive which means that our model is giving too low of a prediction value for the more elderly people within the population. This is evident in the normal Q-Q plot and we see how the plotted values increase much above our Q-Q line. We made a new attempt of making a model by eliminating the insignificant variables and by creating all of the interaction terms and eliminating those which were deemed insignificant hoping this would lead to a more accurate model. After running these models we say they were no better than our first and we would have to tackle this problem with new methods. We now had to approach this problem differently by eliminating the country variable and focusing more on the status of a country's development. Our new model was much simpler to interpret and allowed us to better analyze where our problems were coming from.

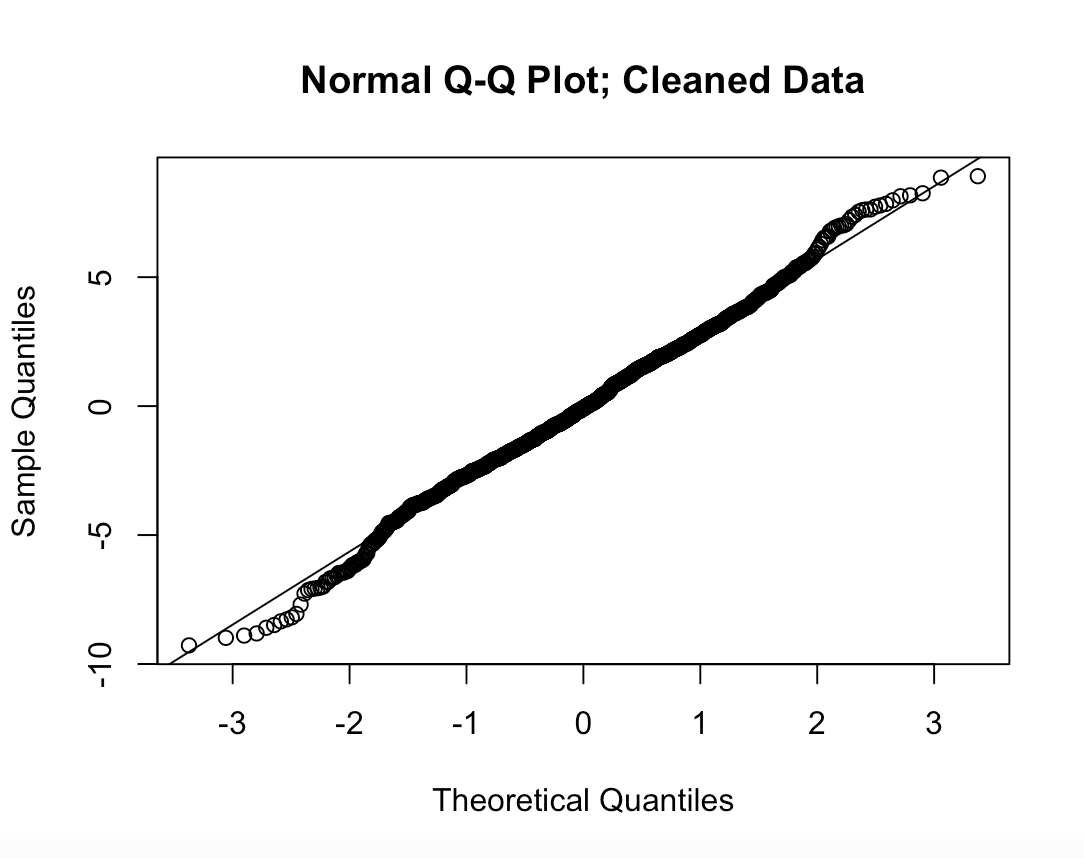


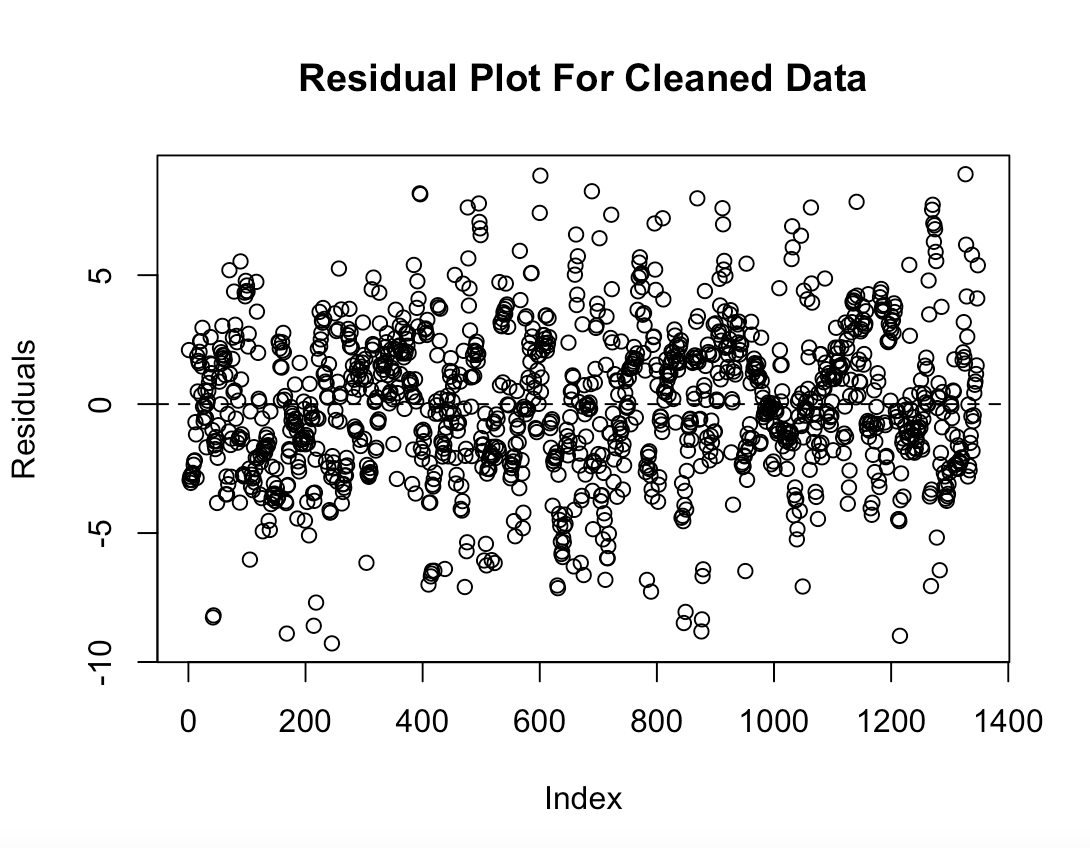


Our new model decreases our adjusted R^2 value to 0.8336, which was expected since the number of predictor variables had drastically decreased. Our standard error is now 3.588, which is an increase compared to our previous 1.776. We see by creating a simpler model with only 20 variables the distribution of the residuals is more normal and by using the Q-Q Plot we can see how our residuals are closer to the actual values then before. Although a problem in this model still shows that the residuals are quite large in absolute terms, meaning there must be outliers within the dataset. While analyzing the outliers we also determined that there is a multicollinearity problem within the data set.



From the correlation matrix our assumption that there is multicollinearity present within the dataset is correct and we need to adjust our model in order to minimize this problem. There is a high positive correlation between adult mortality rate and BMI. This is expected since a higher BMI is usually an indicator of a person being overweight, which would lead to a higher mortality rate. Our biggest correlation problem we observed was their correlation between rate of deaths for infants and children under 5. The problem occurs since infants are included in children under 5, so we see an overlap of the same information twice. This suggests to us to eliminate one of the variables. We see a large positive correlation between diphtheria and hepatitis B. After further research we noticed both diseases are easily preventable by vaccination. We assume the two variables are correlated because in countries with less access to vaccines for these diseases individuals are much more susceptible to them. It may happen in a less developed country that an individual is likely to get diphtheria and hepatitis B because of the lack of vaccines available for prevention. This idea also supports the high correlation noticed between diphtheria and polio, which is also very easily preventable through vaccination. Our last extremely high positive correlation observed in the dataset is expenditure and GDP. This is since the higher the GDP the more a country is able to buy and invest in healthcare. Furthermore, we next decide to implement a Cook’s distance procedure in order to eliminate possible outliers and have a more accurate representation of the data. We now create a model which eliminates the influential points and are able to bring improvement to the fit.





By eliminating the influential points we are able to decrease our standard error from 3.588 to 2.91 and increase our adjusted R^2 value from 0.8336 to 0.8757 all while making the residuals more normally distributed. We see this in our residual plot as the variance across the model tends to stay the same and random.

The ordinary least squares model is motivated by minimizing the squared residuals. More formally, it minimizes 𝝐2 of the equation y = XB + 𝝐. X is the matrix of data and we solve for B, a vector of coefficients. The model assumptions were previously mentioned and apply to 𝝐 in this equation.

Running a simple linear regression using lm function in R

model1 <- lm(Life.expectancy ~ Country + Status + Adult.Mortality +

infant.deaths + Alcohol + Hepatitis.B + Measles

+ BMI + under.five.deaths + Polio + Diphtheria +

HIV.AIDS + GDP + Schooling, data = Life.Expectancy.Data)

summary(model1)

new\_model1 <- lm(Life.expectancy ~ Country + Status + Adult.Mortality +

infant.deaths + Alcohol + Hepatitis.B + Measles

+ under.five.deaths + HIV.AIDS + GDP +

Schooling, data = Life.Expectancy.Data)

In this model we have life expectancy as the dependent variable, and from the summary we can see that the adjusted R^2 value is extremely high, (0.975), which one might believe to be a very sign in the prediction of life expectancy. We analyze the different independent variables and check for statistical significance at the 5% level. We throw away the variables which were deemed insignificant. The variables no longer used are BMI, Polio, and Diphtheria. We were surprised to see that BMI was thrown away as we thought it would be a crucial value to estimate the health of someone. We were not surprised that Polio and Diphtheria were thrown away as those cases are extremely uncommon and are usually dealt with very efficiently and effectively due to modern knowledge and medicine. Although after further investigation into the residuals we see through the qq norm plot that there is more to this dataset than a simple linear regression model and we see through the shapiro-wilk test that our test turns out to be significant. This means that the data is not normally distributed and the current model should not be used to predict life expectancy and many of the values of the coefficients could be wrong and should not be interpreted.

We ran a couple more models with interaction terms and standardized variables in order to be able to interpret the results better. After investigating the data much more, we saw there are many problems with the data. There are many mistakes in the data such as the measles cases per 1000 sometimes being reported as the total number of measles for a country in that specific year. Furthermore we see a problem within the GDP variable. Often there are cases where the decimal is misplaced in the GDP value leading to an incorrect calculation. This leads us to working with a data set that is inaccurate which could lead us to having inaccurate models. We continue to see a problem as well in the multicollinearity of the dataset. We inspected this through a correlation matrix in R, which is able to obtain the correlation associated amongst each of the independent variables and say this problem. Also, many factors have missing values. This will require us to do more advanced methods to be able to better predict the life expectancy of an individual.

Our new method was to not make a country a variable as we then would have too many variables and instead deemed the status of whether a country was developing or developed would be a much better predictor. We now once again factor status and treat the year and country as part of the null. We then input the independent variables from the dataset into our linear regression model and attempt to run a new analysis. We then analyze the model which has been given to us and using Cook's distance attempt to eliminate the influential points in order to make sure variables are not being influenced heavily by a smaller number of points.

***Stepwise regression***

> full.model <- lm(Life.expectancy ~Status+Adult.Mortality+infant.deaths+Alcohol+

+ percentage.expenditure+BMI+ under.five.deaths+HIV.AIDS+thinness..1.19.years+thinness.5.9.years+

+ Income.composition.of.resources+Schooling, data = Data)

> step.model <- stepAIC(full.model, direction = "both",

+ trace = FALSE)

> summary(step.model)

Call:

lm(formula = Life.expectancy ~ Status + Adult.Mortality + infant.deaths +

Alcohol + percentage.expenditure + BMI + under.five.deaths +

HIV.AIDS + thinness..1.19.years + Income.composition.of.resources,

data = Data)

Residuals:

Min 1Q Median 3Q Max

-22.6383 -2.4317 0.0294 2.4442 21.1259

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.286e+01 5.789e-01 108.579 < 2e-16 \*\*\*

StatusDeveloping -1.785e+00 3.125e-01 -5.712 1.25e-08 \*\*\*

Adult.Mortality -1.888e-02 9.005e-04 -20.968 < 2e-16 \*\*\*

infant.deaths 1.266e-01 9.055e-03 13.978 < 2e-16 \*\*\*

Alcohol 1.290e-01 2.843e-02 4.539 5.92e-06 \*\*\*

percentage.expenditure 3.817e-04 4.813e-05 7.930 3.24e-15 \*\*\*

BMI 6.630e-02 5.590e-03 11.861 < 2e-16 \*\*\*

under.five.deaths -9.595e-02 6.656e-03 -14.417 < 2e-16 \*\*\*

HIV.AIDS -4.825e-01 1.904e-02 -25.346 < 2e-16 \*\*\*

thinness..1.19.years -5.977e-02 2.651e-02 -2.255 0.0242 \*

Income.composition.of.resources 1.426e+01 5.609e-01 25.427 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.326 on 2558 degrees of freedom

(369 observations deleted due to missingness)

Multiple R-squared: 0.7886, Adjusted R-squared: 0.7878

F-statistic: 954.5 on 10 and 2558 DF, p-value: < 2.2e-16

The above stepwise regression confirms some of our observations with the t-test of coefficients, but also differs. Polio and Diphtheria are still deemed insignificant, but Heptatitis.B was also deemed insignificant. Although, stepwise regression deemed BMI significant even though the coefficient value is small. With the loss of some factors the R2 has lowered to .7886 but the adj-R2 is very close to the R2 at .7878.

***Full factor model with data set cleaning***

#Author: John Ayres ##########################################################

library(car)

library(dplyr)

Data = read.csv(file = "Life Expectancy Data.csv",header = TRUE)

Data

as.factor(Data$Status)

model.lm = lm(Life.expectancy~Status + Adult.Mortality + infant.deaths + Alcohol

+ percentage.expenditure + Hepatitis.B + Measles + BMI

+ under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS

+ GDP + Population + thinness..1.19.years + thinness.5.9.years

+ Income.composition.of.resources + Schooling, data = Data)

summary(model.lm)

model.lm$coefficients

# Testing for normality among residuals ######################################

qqnorm(residuals(model.lm))

# Suggest normality

plot(residuals(model.lm))

abline(a=0,b=0,lty=2)

# Suggests normality

shapiro.test(residuals(model.lm))

#Suggests the residuals are not normal. OLR and T-test under large N and CLT

#nullify this argument, see references section

# Cooks distance to identify influential points

cookd.lm<-cooks.distance(model.lm)

influential<-cookd.lm[(cookd.lm>3\*mean(cookd.lm,na.rm=TRUE))]

influential

# We see that are multiple influential points or outliers, therefore we eliminate such points form the data set

names\_influential<-names(influential)

outliers<-Data[names\_influential,]

Data\_no\_outliers<-Data %>% anti\_join(outliers)

model2.lm = lm(Life.expectancy~Status + Adult.Mortality + infant.deaths + Alcohol

+ percentage.expenditure + Hepatitis.B + Measles + BMI

+ under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS

+ GDP + Population + thinness..1.19.years + thinness.5.9.years

+ Income.composition.of.resources + Schooling, data = Data\_no\_outliers)

summary(model2.lm)

model2.lm$coefficients

# Testing for normality among residuals ######################################

qqnorm(residuals(model2.lm),main='Normal Q-Q Plot; Cleaned Data')

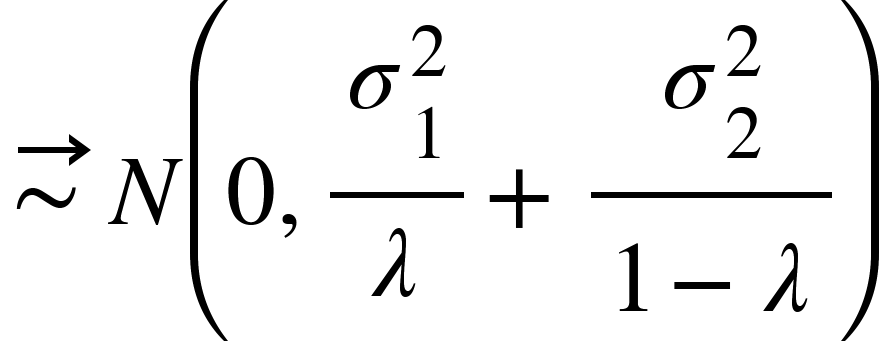
# Suggest normality

plot(residuals(model2.lm))

abline(a=0,b=0,lty=2)

# Suggests normality

shapiro.test(residuals(model2.lm))

Running this model we are able to see a good R^2 adjusted value of 0.8336, meaning 83.36% of the variation in life expectancy is explained by all of the variables listed in this model. Although after further investigation we see some of the variables are statistically insignificant, which means we can most likely eliminate them as long as they are not included in interaction terms. Most importantly, we see the residuals are not normally distributed. The non-normal distribution of the residuals raises some interesting concerns. Firstly, can the model still be used? Yes, but with some consideration. The normality of the residuals is not an assumption of the model, but the residuals non-normality suggests that there is possible a better fitting model. One major concern of non-normal residuals is the potential for Type II error(failing to reject null hypothesis when it is false). The parametric estimation of our beta terms comes from the t-test, hence the column pr>|t| in R output. This is a probability of the values of the beta being within the absolute value of the t-test. This is important because the t-test is our major computational method for checking our beta value obtained from our least squares optimization. Professor Snijders of Oxford university provides guidance to the robustness of the t-test against type II error. The law of large numbers implies that for large n, the standard error converges to a normal distribution,.

Therefore, t has an asymptotic normal distribution under equal means, which is an assumption of the model. This demonstrates the robustness of the T-test to type II error, showing that our non-normal residuals(Snijders, 2011). Our non-normal residuals in this case simply give more information as to the nature of the data

**DISCUSSION**

For the discussion we will be analyzing the results from the reduced factor model with and adj-R2 of .9575. One of the most important factors was the country coefficients. The country variable was one-hot encoded so that each unique country in the dataset has its own coefficient. Depending on the country, these coefficients could be in the single digits, or even as large as in the tens. This means that the country a person lives in affects their life expectancy by years or even decades. Another prominent factor was alcohol consumption. We would also like to emphasize one of the main factors out of our control which influenced our project significantly. Most importantly, were the outliers or missing data points in many of the variables. Although we tried our best to eliminate outliers and missing data points, it is still highly probable that errors may have not been caught. A reduced data set also reduces our ability to better predict life expectancy.

This data set from WHO showcases some of the flexibility and use of linear regression. Through this model we are able to take a full factorial model, and run an initial regression, use other tools such as cooks distance to calculate influential points and remove them from the data set, effectively helping to clean the data and remove outliers. We also see the application of linear regression in the event where residuals are not normally distributed. Effectively, these different iterations give us more and more information about the data, the variables relations, and interactions. In another iteration a reduced model was used, eliminating the factors whose influence was insignificant(by t-test). This reduced model boasted better fit than the previous model as exemplified by its higher R-squared coefficient. Another iteration of the reduction finalized an adjusted R-Squared value of 0.9575. While we see our specific method and use of regression allows us to build a better fitting model with each iteration, it shows that the linear regression model in terms of experimentation, not only gives us the ability to make inferences from the data in terms of prediction/confidence intervals, but we can better understand the specific influence the factors have, in order to better improve life expectancy in certain regions. This shows further that linear regression in combination with other test methodologies, is a flexible, robust system to understand influence from your factors in a designed experiment.

**ACKNOWLEDGMENTS**

Special thanks to user Alexis Dinno(Sc.D Oregon Health and Science University) on stack exchange, for providing insight and advising on non-normal residuals in Linear regression.

**LITERATURE CITED**

Dataset: <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>

J. Kaczmarski, “Explained Gauss-Markov proof: Ordinary least squares and B.L.U,” *University of Mexico, Department of Economics*, 2021. [Online]. Available: https://www.unm.edu/~jikaczmarski/working\_papers/gm\_proof.pdf. [Accessed: 24-Apr-2022]

Snijders, Tom A.B. “Statistical Methods: Robustness - University of Oxford.” *Statistical Methods: Robustness*, University of Oxford, 13 Nov. 2011. [Online]. Available: http://www.stats.ox.ac.uk/~snijders/SM\_robustness.pdf.