STA 141A Group 20 Project

Raghuram Palaniappan, Owen Levinthal, Yuhan Pu, Miranda Scheidner

11/20/2021

Contributions

Raghuram Palaniappan | rpalani@ucdavis.edu (mailto:rpalani@ucdavis.edu) : Quality Overseer, RMD Editor, Introduction, Data cleaning, Data Background, Key Questions, Model Set-Up and Selection, Formal Tests of Model, Statistical Analysis, Interpretation of Results, Conclusion

Yuhan Pu | ypu@ucdavis.edu (mailto:ypu@ucdavis.edu) : Formal Tests of Model, Data Visualization, Interpretation of Results

Owen Levinthal | Olevinthal@ucdavis.edu (mailto:Olevinthal@ucdavis.edu) : Country Life Expectancy Function, Statistical Analysis, Interpretation of Results

Miranda Scheidner | mschnier@ucdavis.edu (mailto:mschnier@ucdavis.edu) : Statistical Analysis, Interpretation of Results

Introduction

The Human Development Index uses life expectancy, GDP, and years of schooling as indicators from countries to assess the development of a country [1]. Our project isolates the life expectancy component of this index and questions what indicators influence it the greatest. The Australian Department of Health states that life expectancy is influenced by factors such as “socioeconomic status, including employment, income, education and economic wellbeing; the quality of the health system and the ability of people to access it; health behaviours such as tobacco and excessive alcohol consumption, poor nutrition and lack of exercise; social factors; genetic factors; and environmental factors including overcrowded housing, lack of clean drinking water and adequate sanitation [2].” With this knowledge, data related to these factors were extracted from the World Bank database.

Data Background

Using the Worldbank dataset, we extracted data for all countries from the year 2018 into a csv file. Using excel, all rows with missing values in data were removed, so only observations where data for all variables were considered. Originally, there were 217 countries of data but after filtering 167 countries remained which will be used to perform tests. Our response variable is Life Expectancy at Birth in years. The indicators chosen for explanatory variables are GDP in US$, Electricity is the Population Access to Electricity in percentage, CO2 is the CO2 emissions in metric tons, FTM is the Female to Male labor participation rate in %, Education is the average duration of primary school for individuals, Alcohol is the consumption in liters per capita, Health is the current expenditure in dollars per capita, Employment is the employment to population ratio, and Water is the population access to basic drinking water services.

```
data = read.csv("LifeExpectancyData.csv",header = T)
head(data)
```

```
##   CountryName       LE Electricity        CO2          GDP Alcohol    Water
## 1 Afghanistan 64.48600    98.71562  0.2001511 1.835388e+10    0.21 69.60193
## 2     Albania 78.45800   100.00000  1.9397316 1.514702e+10    7.17 94.43639
## 3     Algeria 76.69300    99.64192  3.5916574 1.754150e+11    0.95 94.03741
## 4      Angola 60.78200    45.29000  0.8873804 1.013530e+11    6.94 56.59175
## 5     Armenia 74.94500    99.90000  1.8802463 1.245794e+10    5.55 99.95192
## 6   Australia 82.74878   100.00000 15.4755165 1.432880e+12   10.51 99.96989
##        FTM     Health Education Employment
## 1 28.62503   49.84261         6      43.38
## 2 74.49190  274.91409         5      52.03
## 3 25.49049  255.86943         5      37.61
## 4 96.39149   87.61677         6      72.10
## 5 66.07909  422.28268         4      43.08
## 6 84.99437 5425.34033         7      62.17
```

Key Questions

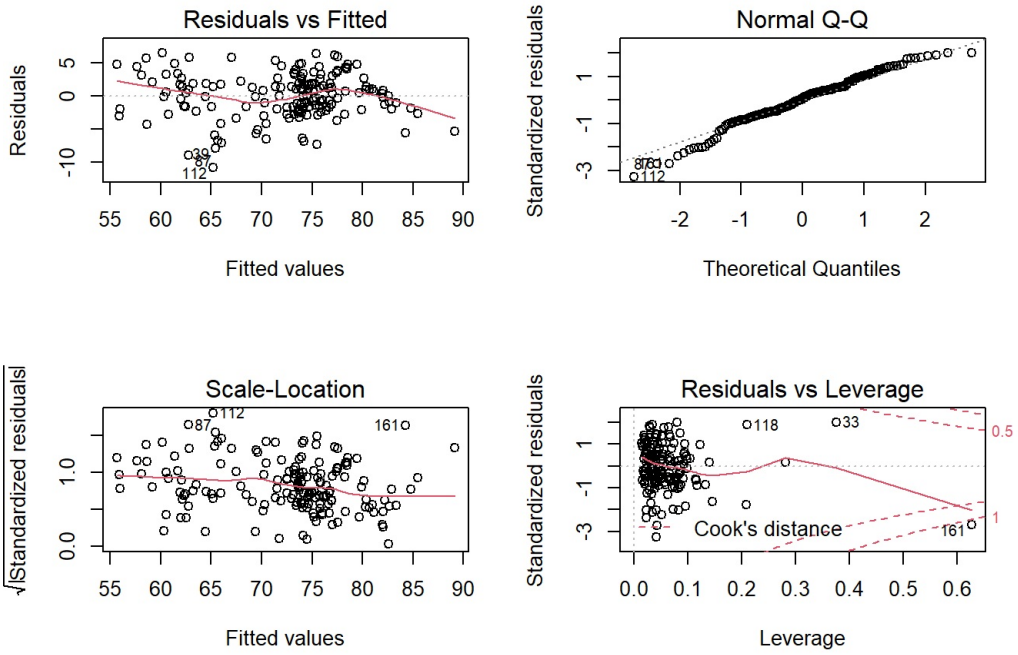
1. Does GDP, Electricity Access, CO2 Emissions, Female to Male Labor rate, Education years, Alcohol consumption, Health expenditure, Employment rate, Clean drinking water access in a country have a statistically significant influence on Life Expectancy?
2. Which of our explanatory variables explain our response variable of Life Expectancy the greatest?
3. When is life expectancy high in a country? What explanatory variables are influencing the life expectancy?

Model Setup

```
# Full Model
LE = lm(LE~GDP+Electricity+CO2+FTM+Education+Alcohol+Health+Water+Employment, data = data)
summary(LE)
```

```
##
## Call:
## lm(formula = LE ~ GDP + Electricity + CO2 + FTM + Education +
##     Alcohol + Health + Water + Employment, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.846  -1.975   0.276   1.920   6.496
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.536e+01  3.670e+00  12.360  < 2e-16 ***
## GDP         -2.570e-13  1.465e-13  -1.755 0.081210 .
## Electricity  1.231e-01  2.539e-02   4.848 2.98e-06 ***
## CO2         -4.561e-03  6.924e-02  -0.066 0.947569
## FTM         -4.411e-02  2.294e-02  -1.923 0.056279 .
## Education   -2.079e-01  3.443e-01  -0.604 0.546871
## Alcohol      1.539e-01  9.103e-02   1.691 0.092804 .
## Health       1.393e-03  1.974e-04   7.057 5.16e-11 ***
## Water        1.396e-01  4.226e-02   3.303 0.001185 **
## Employment   1.102e-01  3.067e-02   3.593 0.000437 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.399 on 157 degrees of freedom
## Multiple R-squared:  0.8083, Adjusted R-squared:  0.7973
## F-statistic: 73.53 on 9 and 157 DF,  p-value: < 2.2e-16
```

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

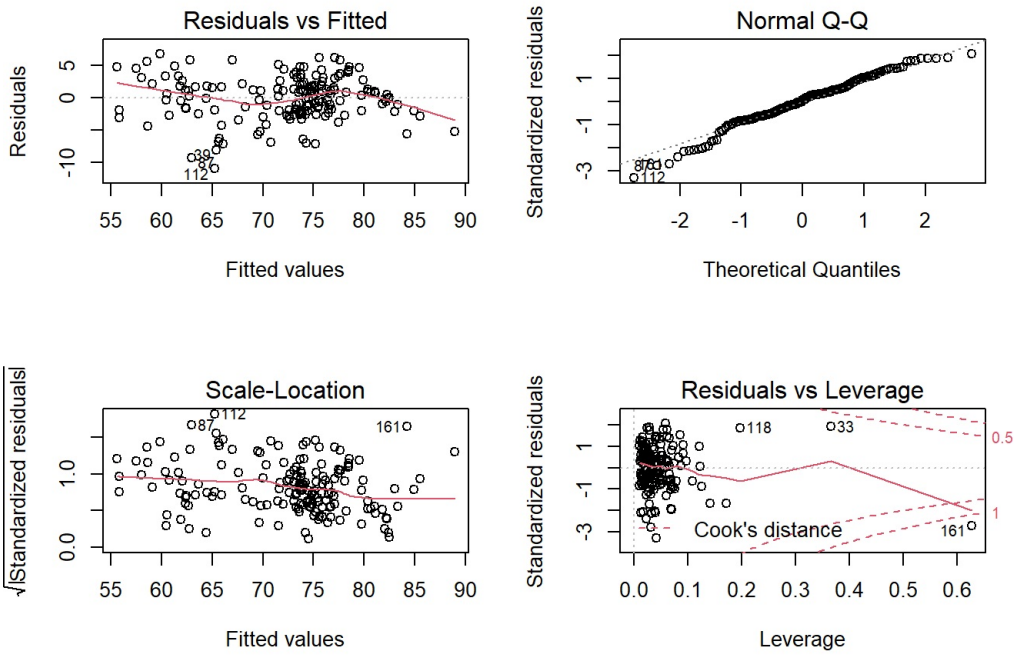


Certain variables such as CO2 and Education influence very little value towards the linear model. Along with this the QQ-plot does not show a normal distribution. To combat this the model will be reduced and the most optimal model will be selected. The forward stepwise regression will be used due to the large amount of variables to create a reduced model.

```
# Stepwise regression model
redmodel = stepAIC(LE, direction = "both", trace = FALSE)
summary(redmodel)
```

```
##
## Call:
## lm(formula = LE ~ GDP + Electricity + FTM + Alcohol + Health +
##     Water + Employment, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9071  -2.0373   0.2068   2.0203   6.7923
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.411e+01  2.914e+00  15.136  < 2e-16 ***
## GDP         -2.505e-13  1.450e-13  -1.728  0.08595 .
## Electricity  1.255e-01  2.490e-02   5.043 1.24e-06 ***
## FTM         -4.410e-02  2.255e-02  -1.956  0.05224 .
## Alcohol      1.603e-01  8.990e-02   1.783  0.07642 .
## Health       1.359e-03  1.805e-04   7.531 3.55e-12 ***
## Water        1.384e-01  4.141e-02   3.343  0.00103 **
## Employment   1.090e-01  2.893e-02   3.769  0.00023 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.381 on 159 degrees of freedom
## Multiple R-squared:  0.8078, Adjusted R-squared:  0.7993
## F-statistic: 95.47 on 7 and 159 DF,  p-value: < 2.2e-16
```

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



The CO2 and Education variable is removed via the stepwise regression reducing the model to 7 explanatory variables in Electricity, FTM, Alcohol, Health, Water, and Employment.CO2 and Education had smaller AIC values than the other variables.The Q-Q plot shows that the residuals are not normally distributed.

Formal Tests of Model

```
shapiro.test(resid(redmodel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(redmodel)
## W = 0.98175, p-value = 0.02698
```

```
bptest(redmodel)
```

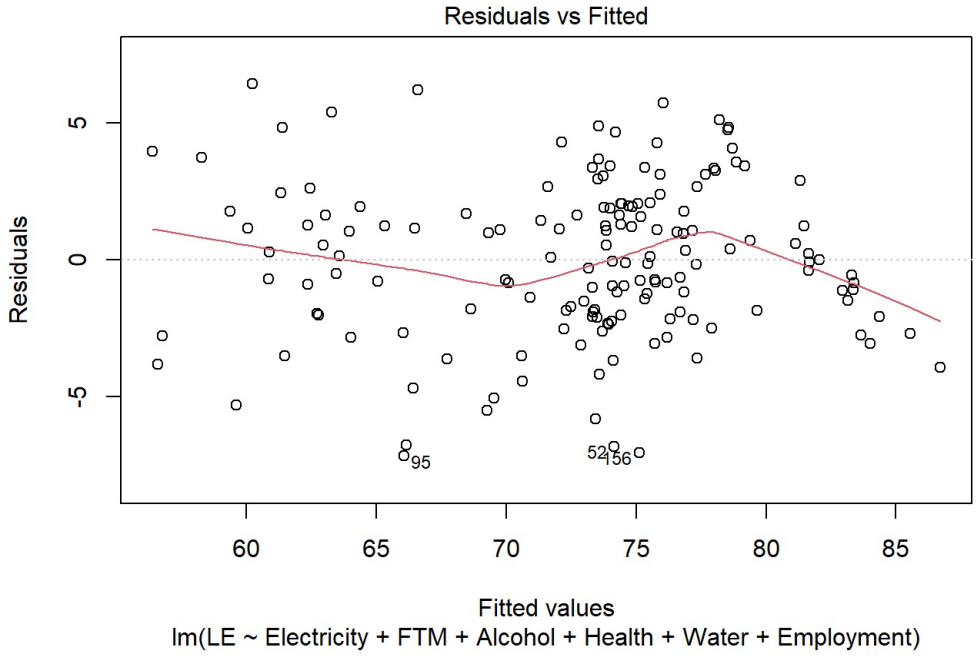
```
##
##  studentized Breusch-Pagan test
##
## data:  redmodel
## BP = 17.69, df = 7, p-value = 0.01345
```

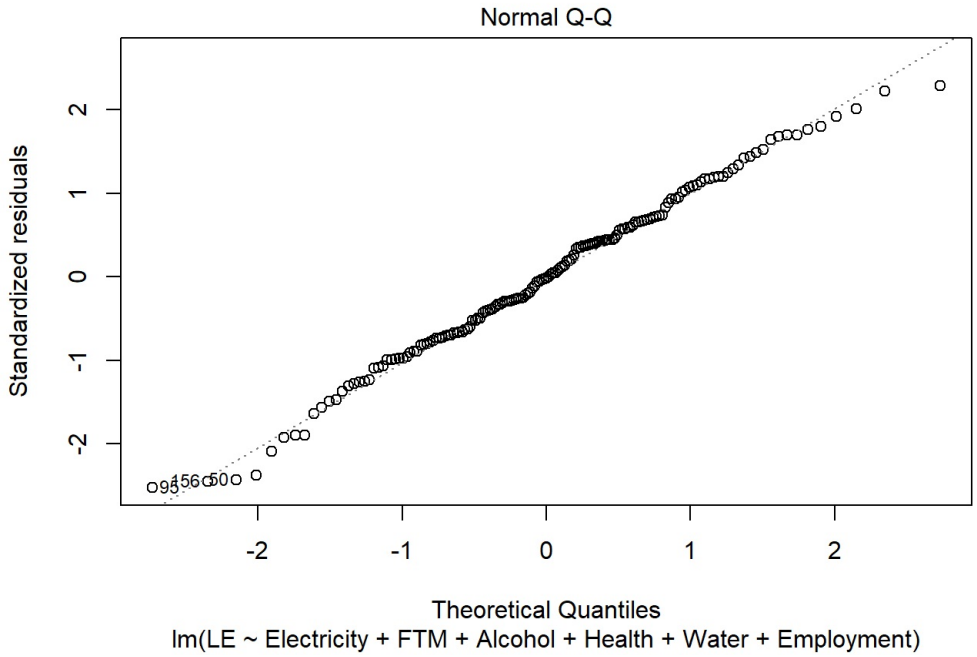
The Shapiro-Wilk test produces a p-value under 0.05 making us reject the null hypothesis that the residuals are normally distributed. The Breusch-Pagan test produces a p-value under 0.05 making us reject the null hypothesis that the errors are homoscedastic. To solve this issue the outliers shown in the plot will be removed from the data.

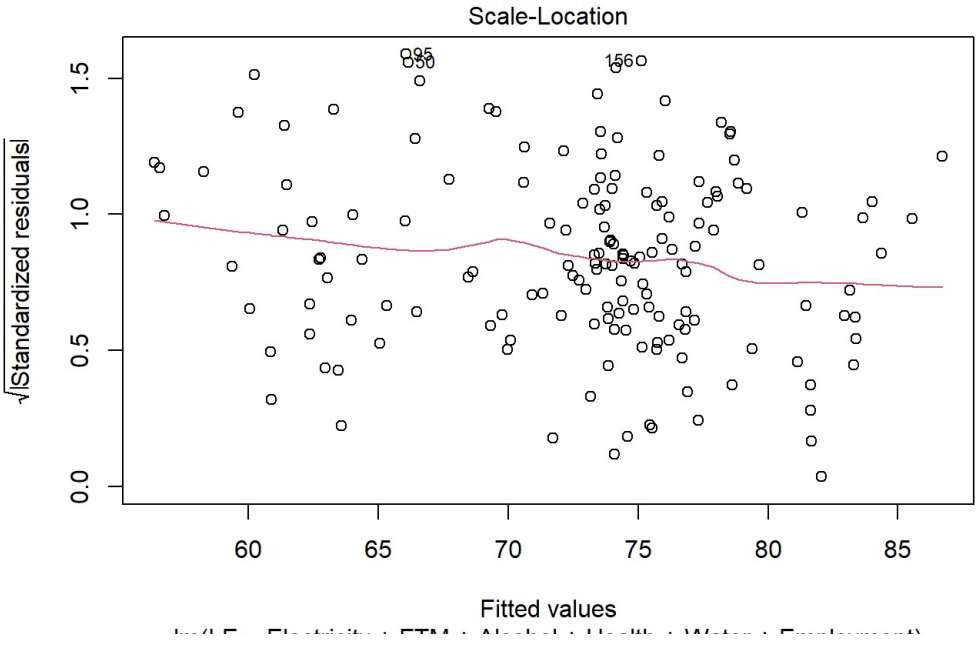
```
data = data[-c(28,39,87,112,118,138,146,161),]
LE = lm(LE~GDP+Electricity+CO2+FTM+Education+Alcohol+Health+Water+Employment, data = data)
finalmodel = stepAIC(LE, direction = "both", trace = FALSE)
summary(finalmodel)
```

```
##
## Call:
## lm(formula = LE ~ Electricity + FTM + Alcohol + Health + Water +
##     Employment, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1754 -1.9882 -0.0401  1.9281  6.4433
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.2958421  2.6406001  17.532  < 2e-16 ***
## Electricity  0.1035523  0.0228105   4.540 1.14e-05 ***
## FTM         -0.0499547  0.0200167  -2.496 0.013641 *
## Alcohol      0.1955819  0.0794118   2.463 0.014899 *
## Health       0.0015235  0.0001709   8.916 1.42e-15 ***
## Water        0.1474526  0.0385654   3.823 0.000192 ***
## Employment   0.0940358  0.0261657   3.594 0.000440 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.904 on 152 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8342
## F-statistic: 133.5 on 6 and 152 DF,  p-value: < 2.2e-16
```

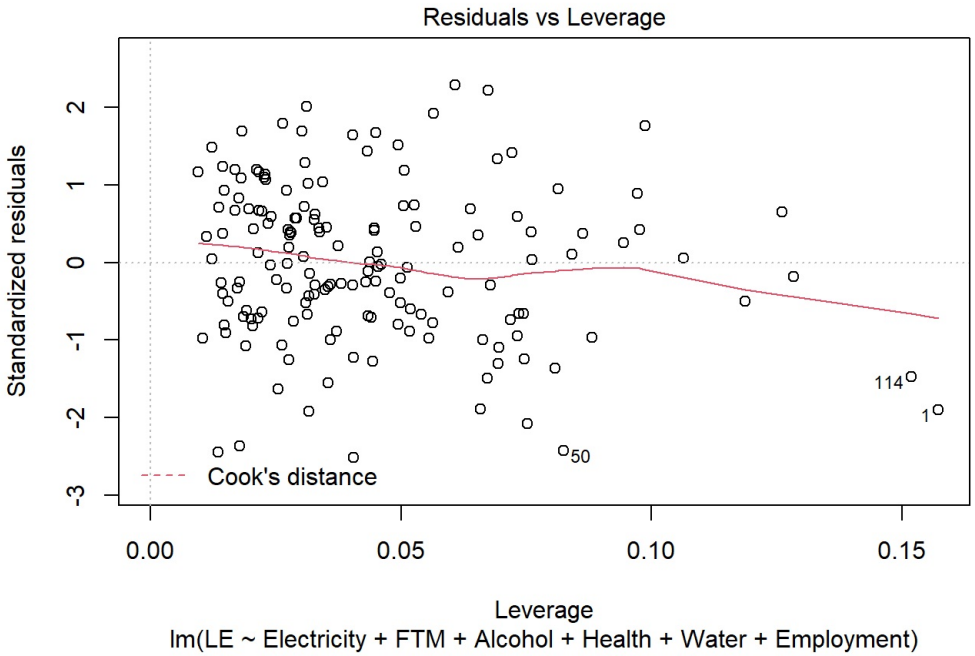
```
plot(finalmodel)
```


Residuals vs Fitted
Residuals
Fitted values
lm(LE ~ Electricity + FTM + Alcohol + Health + Water + Employment)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(LE ~ Electricity + FTM + Alcohol + Health + Water + Employment)


Scale-Location
√|Standardized residuals|
Fitted values

lm(LE ~ Electricity + FTM + Alcohol + Health + Water + Employment)




```
shapiro.test(resid(finalmodel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(finalmodel)
## W = 0.99226, p-value = 0.5499
```

```
bptest(finalmodel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  finalmodel
## BP = 11.355, df = 6, p-value = 0.07799
```

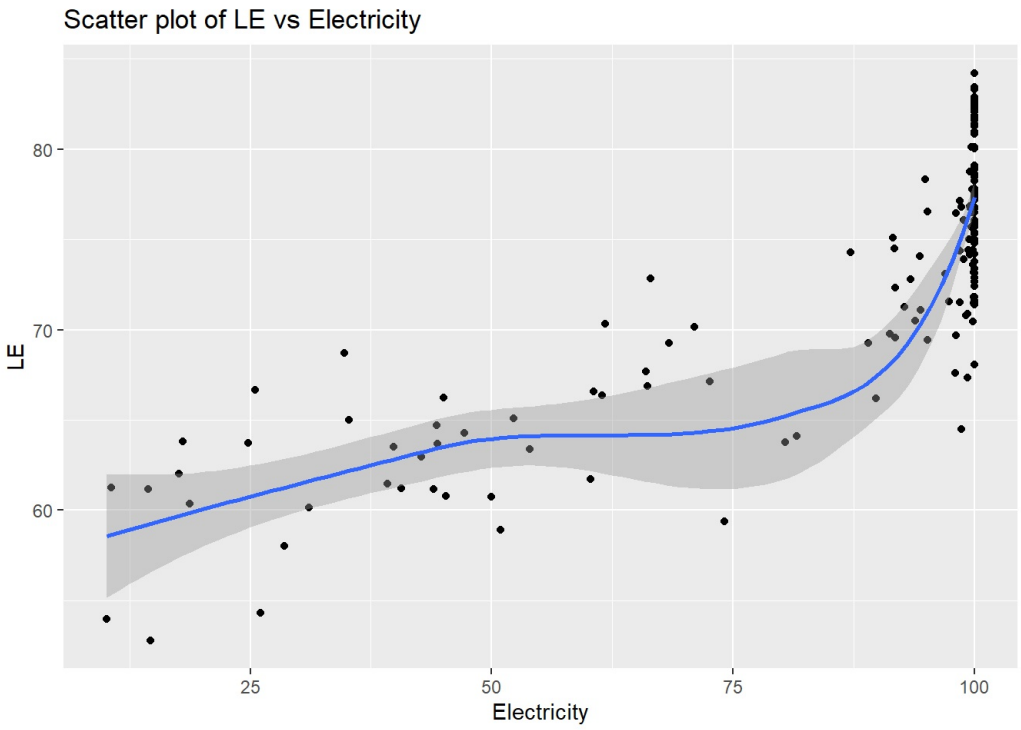
After removing the outliers, the new p-values for the Shapiro-Wilk test and the Breusch-Pagan test do not allow us to reject the null hypothesis.The linear model now has Normality and Homoscedasticity. Along with this the new stepwise regression without the outliers removes GDP as a predictor reducing our model to 6 variables.

Data Visualizaion

Every predictor against response, Life Expectancy

```
#Scatter plot and correlation (LE vs Electricity)#
ggplot(data=data, mapping=aes(x=Electricity,y=LE))+
  geom_point()+
  geom_smooth()+
  labs(title="Scatter plot of LE vs Electricity",x="Electricity",y="LE")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



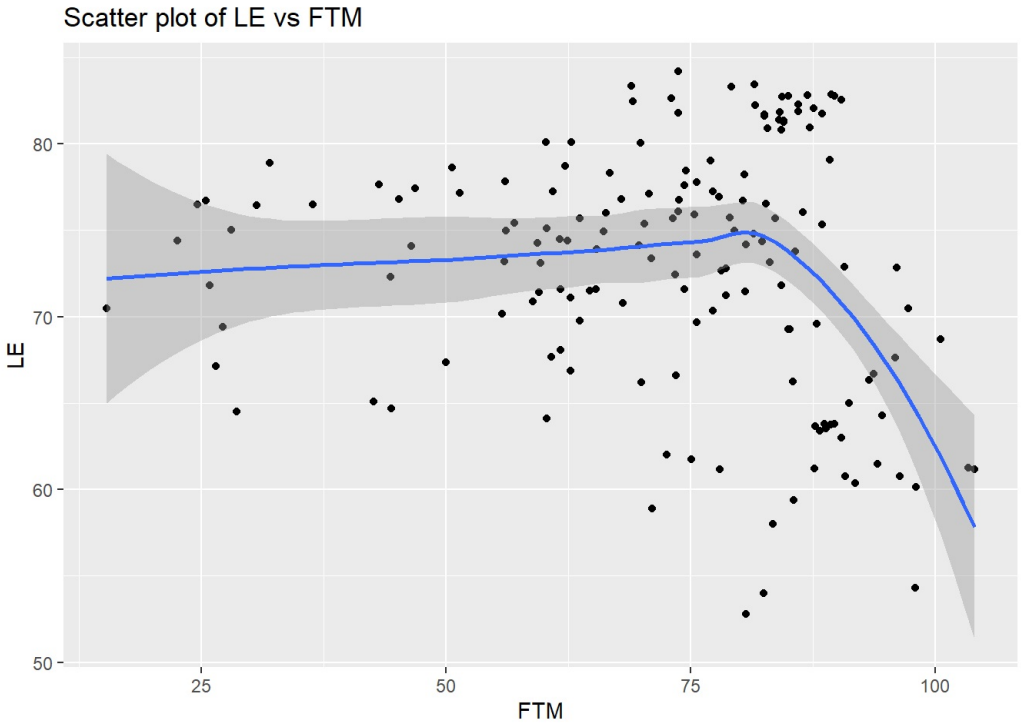

```
cor(data$LE,data$Electricity)
```

```
## [1] 0.8098451
```

Here r=0.8098451, which means that the electricity has highly positive linear relationship with LE.

```
#Scatter plot and correlation (LE vs FTM)#
ggplot(data=data, mapping=aes(x=FTM,y=LE))+
  geom_point()+
  geom_smooth()+
  labs(title="Scatter plot of LE vs FTM",x="FTM",y="LE")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



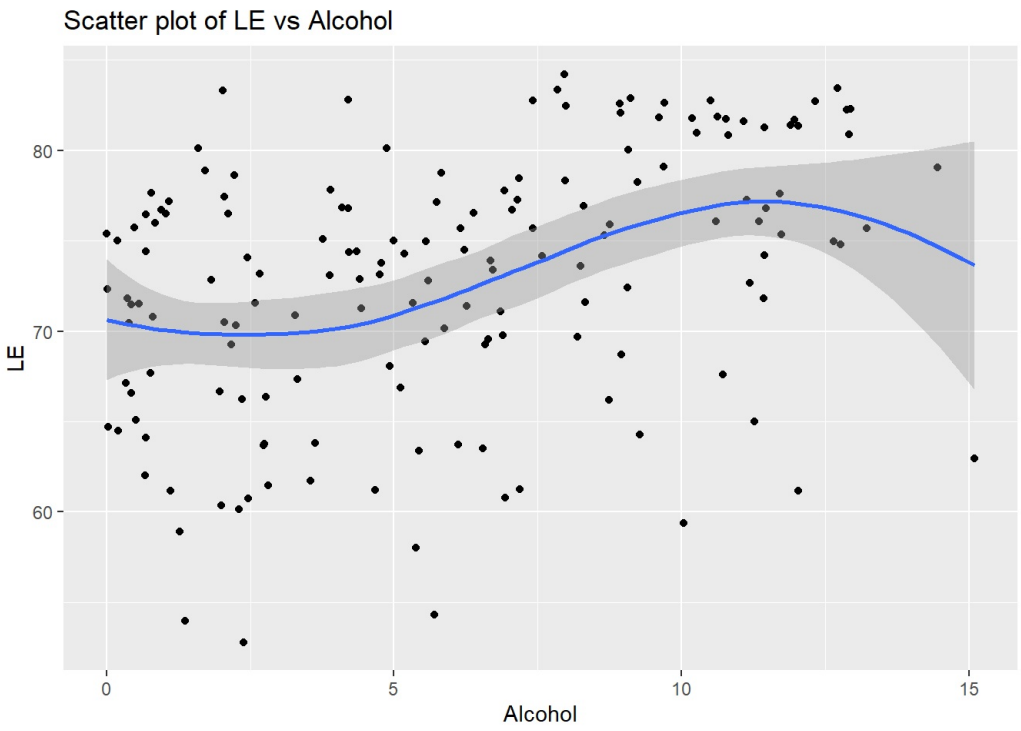

```
cor(data$LE,data$FTM)
```

```
## [1] -0.1403068
```

Here r=-0.1403068, which means FTM has no linear relationship with LE

```
#Scatter plot and correlation (LE vs Alcohol)#
ggplot(data=data, mapping=aes(x=Alcohol,y=LE))+
  geom_point()+
  geom_smooth()+
  labs(title="Scatter plot of LE vs Alcohol",x="Alcohol",y="LE")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



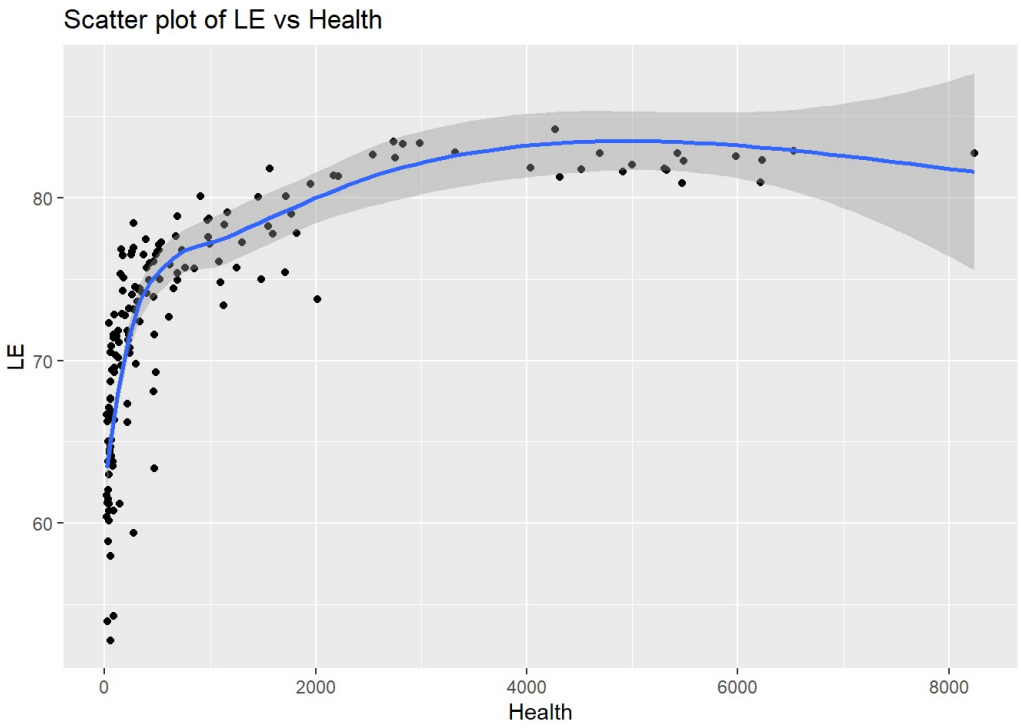

```
cor(data$LE,data$Alcohol)
```

```
## [1] 0.3808504
```

Here r=0.3808504, which means Alcohol has no linear relationship with LE

```
#Scatter plot and correlation (LE vs Health)#
ggplot(data=data, mapping=aes(x=Health,y=LE))+
  geom_point()+
  geom_smooth()+
  labs(title="Scatter plot of LE vs Health",x="Health",y="LE")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



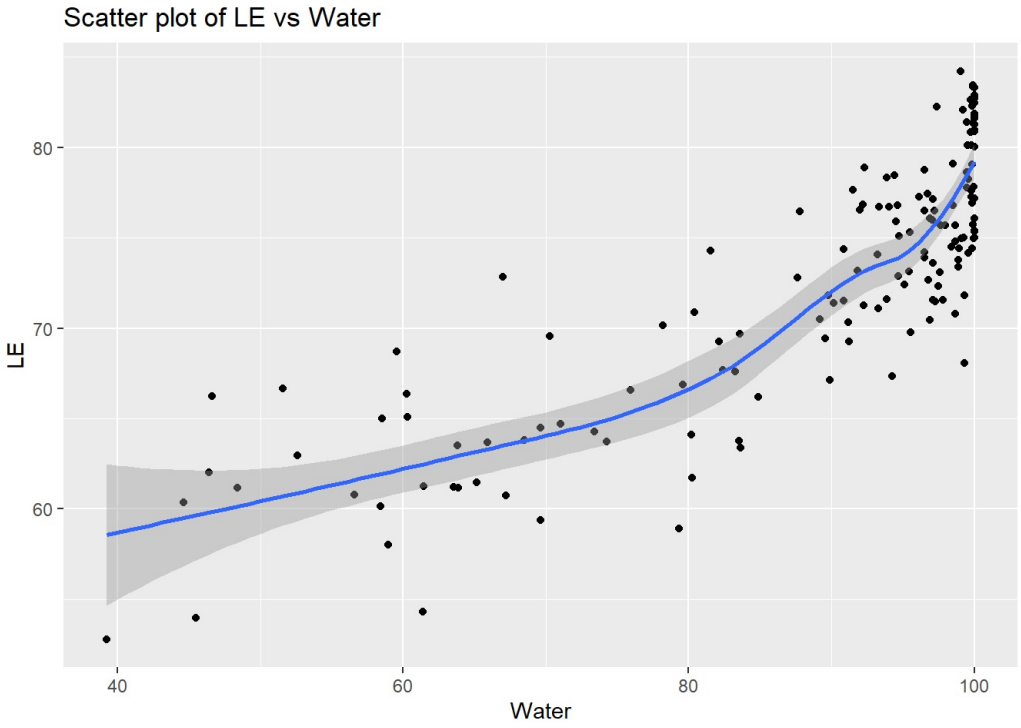

```
cor(data$LE,data$Health)
```

```
## [1] 0.6501304
```

Here r=0.6501304, which means Health has positive linear relationship with LE

```
#Scatter plot and correlation (LE vs Water)#
ggplot(data=data, mapping=aes(x=Water,y=LE))+
  geom_point()+
  geom_smooth()+
  labs(title="Scatter plot of LE vs Water",x="Water",y="LE")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



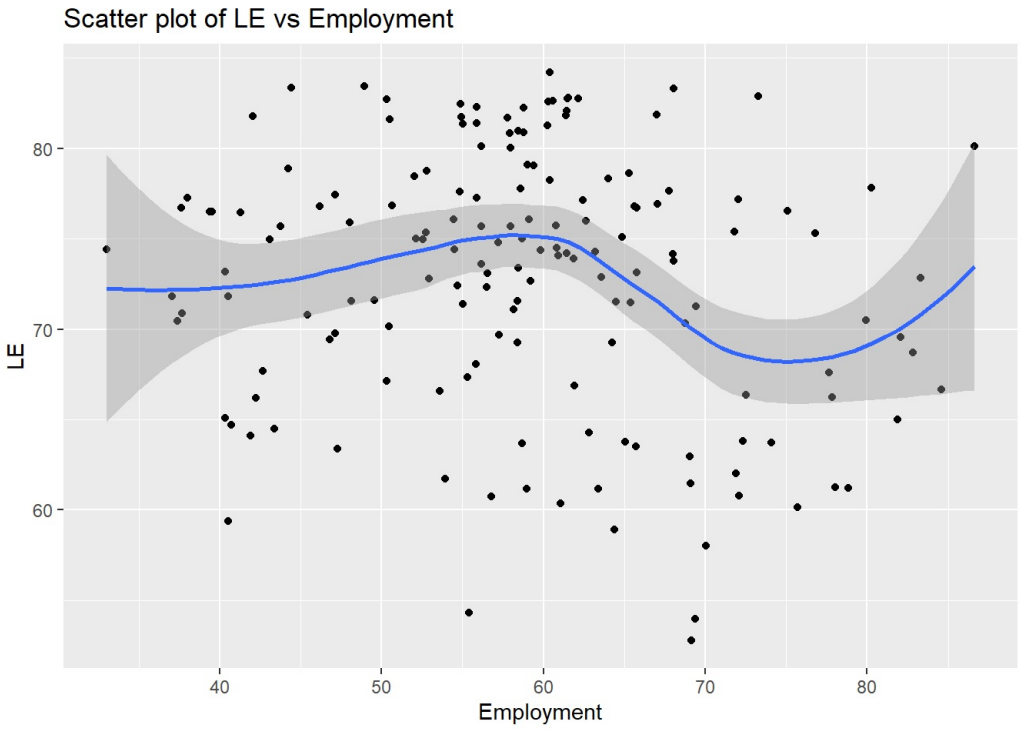

```
cor(data$LE,data$Water)
```

```
## [1] 0.8334759
```

Here r=0.8334759, which means Water has highly positive linear relationship with LE

```
#Scatter plot and correlation (LE vs Employment)#
ggplot(data=data, mapping=aes(x=Employment,y=LE))+
  geom_point()+
  geom_smooth()+
  labs(title="Scatter plot of LE vs Employment",x="Employment",y="LE")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
cor(data$LE,data$Employment)
```

```
## [1] -0.1586456
```

Here r=-0.1586456, which means Employment has no linear relationship with LE.

The two variables with the strongest correlation with our response variables are the Population Percentage Access to Electricity and Health Expenditure per Capita.

Statistical Analysis

F-Test on Regression

In order to finalize a conclusive statement regarding the fit of our data, we check the F statistic to check the fit of the model. To properly assess our results, the calculated p-value will be used to determine whether there is sufficient evidence to refute the null and accept our alternative hypothesis.

H_0 : A model with no independent variables fits the data as well as our created model

H_a : Our model fits the data better than a model with only B0

```
summary(finalmodel)
```

```
##
## Call:
## lm(formula = LE ~ Electricity + FTM + Alcohol + Health + Water +
##     Employment, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1754 -1.9882 -0.0401  1.9281  6.4433
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.2958421  2.6406001  17.532  < 2e-16 ***
## Electricity  0.1035523  0.0228105   4.540 1.14e-05 ***
## FTM         -0.0499547  0.0200167  -2.496 0.013641 *
## Alcohol      0.1955819  0.0794118   2.463 0.014899 *
## Health       0.0015235  0.0001709   8.916 1.42e-15 ***
## Water        0.1474526  0.0385654   3.823 0.000192 ***
## Employment   0.0940358  0.0261657   3.594 0.000440 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.904 on 152 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8342
## F-statistic: 133.5 on 6 and 152 DF,  p-value: < 2.2e-16
```

Our F value of 133.5 produces us a p-value of 2.2e-16 which is extremely small and strengthens our rejection of the null. This proves that our model including our explanatory variable is a better fit.

Confidence Intervals

```
confint(finalmodel)
```

```
##                    2.5 %       97.5 %
## (Intercept) 41.078824477 51.512859653
## Electricity  0.058485644  0.148618895
## FTM         -0.089501657 -0.010407831
## Alcohol      0.038688397  0.352475372
## Health       0.001185876  0.001861082
## Water        0.071259246  0.223646032
## Employment   0.042340364  0.145731160
```

The 95% confidence intervals for all our OLS estimators now all do not cross 0 showing that each OLS estimator is significant in our best model chosen. This shows that our model selection was accurate with a model with all of our explanatory variables significant.

T Test on Predictor Variables

```
coeftest(finalmodel)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) 46.29584206  2.64060010 17.5323 < 2.2e-16 ***
## Electricity  0.10355227  0.02281053  4.5397 1.139e-05 ***
## FTM         -0.04995474  0.02001672 -2.4957 0.0136406 *
## Alcohol      0.19558188  0.07941184  2.4629 0.0148987 *
## Health       0.00152348  0.00017088  8.9156 1.423e-15 ***
## Water        0.14745264  0.03856538  3.8234 0.0001915 ***
## Employment   0.09403576  0.02616569  3.5939 0.0004396 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The T-test creates the highest T-value for Health Expenditure per Capita at 8.9156 producing a p-value of 1.423e-15. The rest of the predictors in order from most significant to least significant in our model are Population Access to Electricity, Access to Clean Drinking Water, Employment Rate, Female to Male labor ratio, and Alcohol Consumption per Capita.

Interpretation of Results

Does GDP, Electricity Access, CO2 Emissions, Female to Male Labor rate, Education years, Alcohol consumption, Health expenditure, Employment rate, Clean drinking water access in a country have a statistically significant influence on Life Expectancy?

The original linear model had all these predictor variables in it. After the stepwise regression, Country GDP and Country CO2 emissions were removed as predictor variables. The F statistic guarantees the fit of the model and claims that none of the variables in our reduced model are

insignificant. This is further proven with our confidence intervals of predictor variables where none of the intervals go through 0 proving all predictors are significant.

Which of our explanatory variables explain our response variable of Life Expectancy the greatest?

The statistical T-test revealed that Health Expenditure per Capita had the most significant influence on the response variable of Life Expectancy. The test produced a T-value of 8.9156 which creates the extremely low p-value of 1.423e-15. The result reveals that countries with the highest spending on Health per capita are on average achieving higher life expectancies.

When is life expectancy high in a country? What explanatory variables are influencing the life expectancy?

From the plots of each predictor against the response, Health, Electricity, Water, and Employment had the highest correlation coefficients. These 4 correlation coefficients were all positive. This means for good life expectancy a country needs a high value in Health Expenditure per Capita, Population access to Electricity, Population access to clean drinking water, and high Employment rates.

Conclusion

Originally, the model started with 9 predictors, but after reducing our model via stepwise regression and the removing of the outliers we were left with 6 predictors. Our 6 predictors in the final model are Electricity is the Population Access to Electricity in percentage, FTM is the Female to Male labor participation rate in %, Alcohol is the consumption in liters per capita, Health is the current expenditure in dollars per capita, Employment is the employment to population ratio, and Water is the population access to basic drinking water services. We learned that CO2 emissions and Education weren't highly linked with Life expectancy. Unexpectedly, GDP was not a strong predictor towards life expectancy. However, health expenditure was extremely strong towards influencing life expectancy. It does not matter if a country has extremely high GDP if some of it is not used towards health expenditure. Only countries that allocated money towards health expenditure sustained higher values of life expectancy.

This raises the question of what types of Health Expediture lead to higher life expectancies. Since our strongest predictor is Health Expenditure per Capita, studies could be conducted on what types could influence our response. Could it be mental health expenditure, emergency health, vaccinations, preventative health or etc.? What type of health sectors require the most money allocated to achieve higher life expectancies in a country could be the next research question.

This study found that countries had the highest life expectancy when high values were found in Health Expenditure per Capita, Population access to Electricity, Population access to clean drinking water, and high Employment rates as supported by the T-values. If an individual has access to these they have a greater expected life in years compared to individuals who do not.

Country Life Expectancy Function

```
#Building prediction model
#Plug in new values into c() spots, NA where not available

LEpredict <- function(newdata) {
  temp <- predict(finalmodel, newdata = newdata)
  temp <- round(temp, digits = 1)
  temp1 <- predict(finalmodel, newdata = newdata, interval = "prediction", datatype = "list")
  temp1 <- round(temp1, digits = 2)
  percentile <- ecdf(data$LE)(temp)
  percentile <- round(percentile, digits = 1)

  print(paste("The average Life Expectancy for this country is", temp, "years"))
  print(paste0("This Life Expectancy places the country in the ", percentile*100, "th percentile"))
  print(paste("There is a 95% chance that the true Life Expectancy of the country falls within the range:"))
  print(temp1)
}

# Example of our Function
gdp <- c(1.835388e+10)
electricity <- c(98)
ftm <- c(30)
alcohol <- c(0.2)
health <- c(50)
water <- c(70)
employment <- c(43)

newdata <- data.frame("GDP" = gdp, "Electricity" = electricity, "FTM" = ftm, "Alcohol" = alcohol, "Health" = heal
th, "Water" = water, "Employment" = employment)

LEpredict(newdata)
```

```
## [1] "The average Life Expectancy for this country is 69.4 years"
## [1] "This Life Expectancy places the country in the 30th percentile"
## [1] "There is a 95% chance that the true Life Expectancy of the country falls within the range:"
##     fit   lwr   upr
## 1 69.43 63.28 75.58
```

References

1. Human development reports. Human Development Index (HDI) | Human Development Reports. (n.d.). Retrieved November 30, 2021, from http://hdr.undp.org/en/content/human-development-index-hdi (http://hdr.undp.org/en/content/human-development-index-hdi).
2. Tier 1-life expectancy and wellbeing-1.19 life expectancy at birth. Department of Health | Tier 1-Life expectancy and wellbeing-1.19 Life expectancy at birth. (n.d.). Retrieved December 10, 2021, from https://www1.health.gov.au/internet/publications/publishing.nsf/Content/oatsih-hpf-2012-toc~tier1~life-exp-wellb~119 (https://www1.health.gov.au/internet/publications/publishing.nsf/Content/oatsih-hpf-2012-toc~tier1~life-exp-wellb~119).
3. “STA141A - Project Effects on Ozone - Group 1”, Jun. 2021.