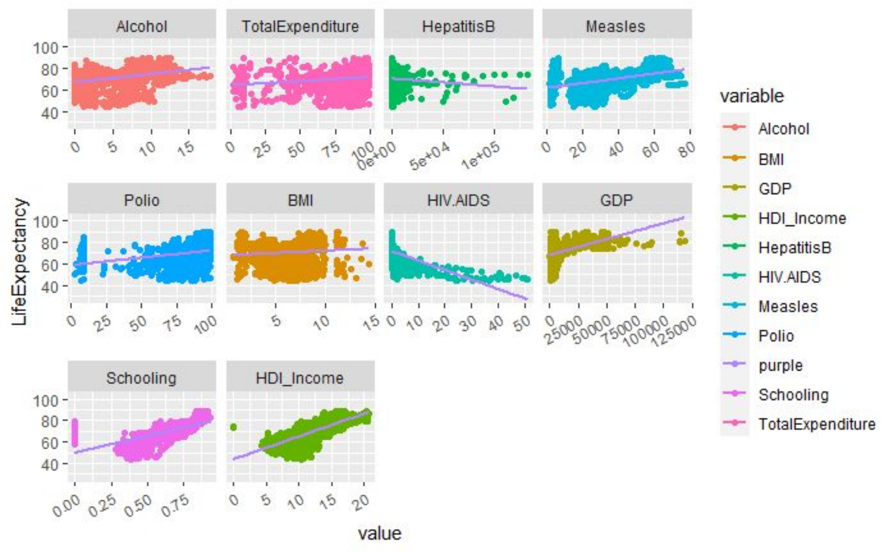
WHO Life Expectancy Data

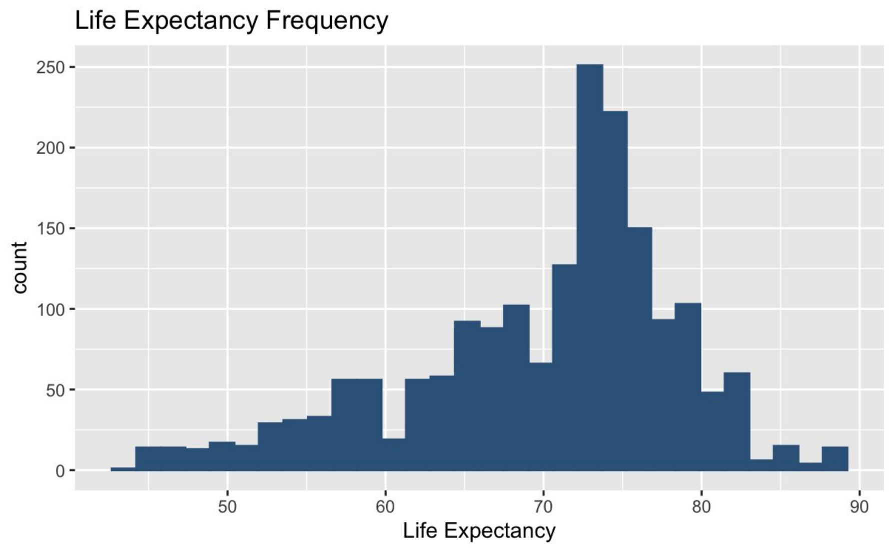
John K, Keerthi S, Kevin S
The Grim Reapers

Dataset Information / Cleaning

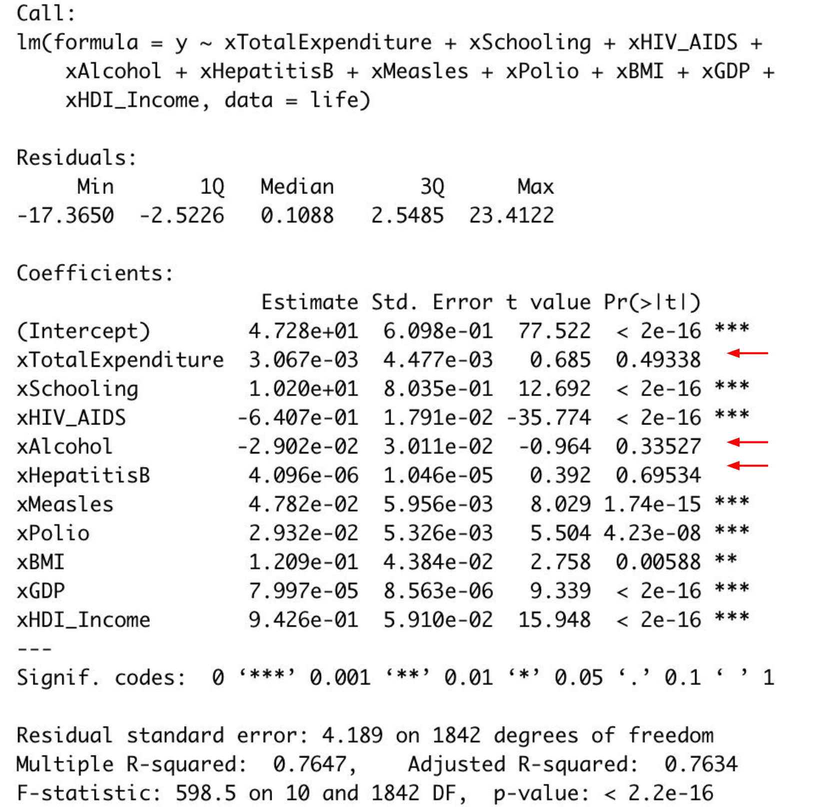
- Life Expectancy data from WHO found on Kaggle.
- Provides information on factors affecting life expectancy, such as various health conditions, income factors, and mortality rates, from 2000 to 2015 in 193 countries.
- Response Variable: Life Expectancy
- Predictor Variables: Alcohol, Health Expenditure, BMI, HIV/AIDS, Hepatitis B, Measles, Polio, Schooling, GDP per Capita, and Income Composition of Resources
- The initial dataset had around 2938 observations, but after importing certain variables and then cleaning the data to omit NA values, we resulted in 1853 observations.
- **ANALYSIS GOAL**: determine some of the population characteristics that affects a country’s overall life expectancy.

Initial Dataset Visualizations before Transformation





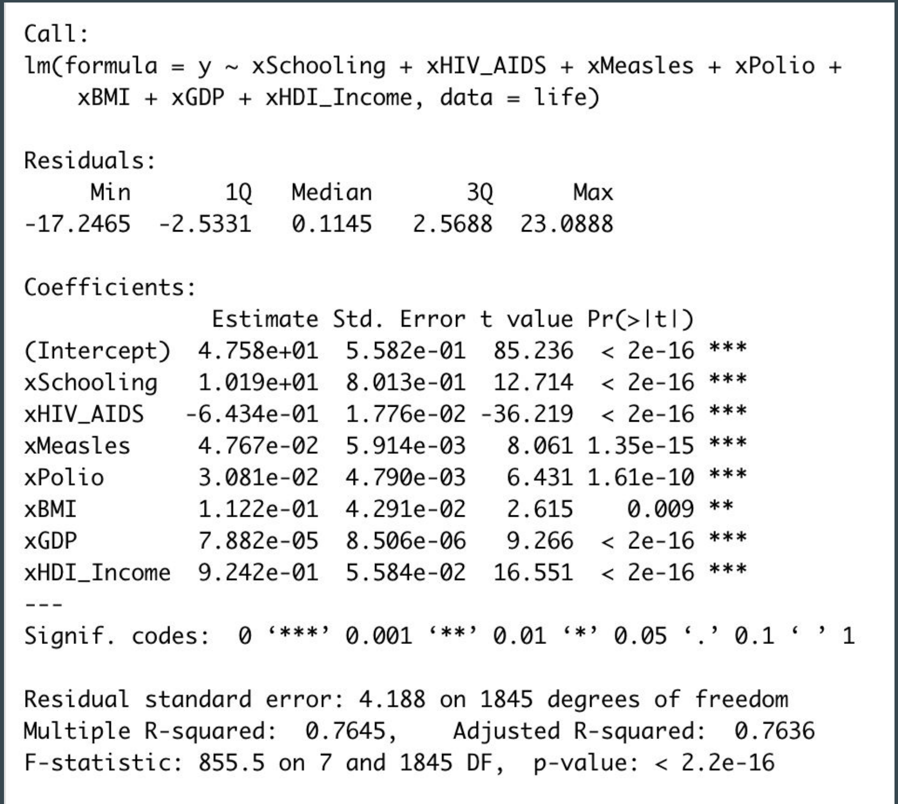
Choosing Predictors: Backward Approach



```
Call:
lm(formula = y ~ xTotalExpenditure + xSchooling + xHIV_AIDS +
    xAlcohol + xHepatitisB + xMeasles + xPolio + xBMI + xGDP +
    xHDI_Income, data = life)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3650  -2.5226   0.1088   2.5485  23.4122

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.728e+01  6.098e-01  77.522  < 2e-16 ***
xTotalExpenditure  3.067e-03  4.477e-03   0.685  0.49338    ←
xSchooling         1.020e+01  8.035e-01  12.692  < 2e-16 ***
xHIV_AIDS         -6.407e-01  1.791e-02 -35.774  < 2e-16 ***
xAlcohol          -2.902e-02  3.011e-02  -0.964  0.33527    ←
xHepatitisB        4.096e-06  1.046e-05   0.392  0.69534    ←
xMeasles           4.782e-02  5.956e-03   8.029 1.74e-15 ***
xPolio             2.932e-02  5.326e-03   5.504 4.23e-08 ***
xBMI               1.209e-01  4.384e-02   2.758  0.00588 **
xGDP               7.997e-05  8.563e-06   9.339  < 2e-16 ***
xHDI_Income        9.426e-01  5.910e-02  15.948  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.189 on 1842 degrees of freedom
Multiple R-squared:  0.7647,	Adjusted R-squared:  0.7634
F-statistic: 598.5 on 10 and 1842 DF,  p-value: < 2.2e-16
```

- **Variable Selection looking at p-values**
- Variables Total Expenditure, Alcohol, and Hepatitis B have a P-value of greater than 0.05.
- Adjusted R- Squared = 0.7634

Reduced Fit Model before Transformation

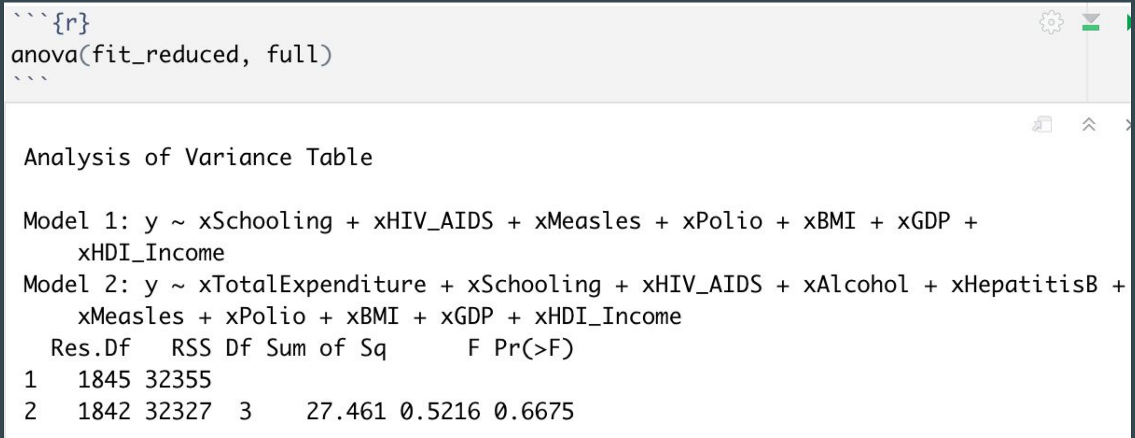


```
Call:
lm(formula = y ~ xSchooling + xHIV_AIDS + xMeasles + xPolio +
    xBMI + xGDP + xHDI_Income, data = life)

Residuals:
     Min       1Q   Median       3Q      Max
-17.2465  -2.5331   0.1145   2.5688  23.0888

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.758e+01  5.582e-01  85.236  < 2e-16 ***
xSchooling   1.019e+01  8.013e-01  12.714  < 2e-16 ***
xHIV_AIDS   -6.434e-01  1.776e-02 -36.219  < 2e-16 ***
xMeasles     4.767e-02  5.914e-03   8.061 1.35e-15 ***
xPolio       3.081e-02  4.790e-03   6.431 1.61e-10 ***
xBMI         1.122e-01  4.291e-02   2.615    0.009 **
xGDP         7.882e-05  8.506e-06   9.266  < 2e-16 ***
xHDI_Income  9.242e-01  5.584e-02  16.551  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.188 on 1845 degrees of freedom
Multiple R-squared:  0.7645,    Adjusted R-squared:  0.7636
F-statistic: 855.5 on 7 and 1845 DF,  p-value: < 2.2e-16
```

- After reducing, our linear regression model is:

  Y = **(1.019e+01)**Schooling + **(-6.434e-01)**HIV_AIDS + **(4.767e-02)**Measles + **(3.081e-02)**Polio + **(1.122e-01)**BMI + **(7.882e-05)**GDP + **(9.242e-01)**HDI_Income

Comparing Models



~~~
```{r}
anova(fit_reduced, full)
```
~~~

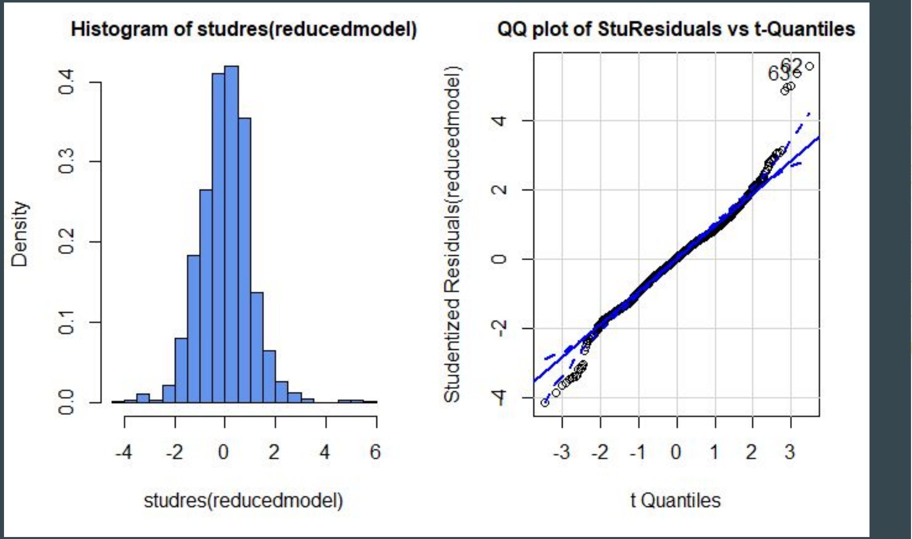
```
Analysis of Variance Table

Model 1: y ~ xSchooling + xHIV_AIDS + xMeasles + xPolio + xBMI + xGDP +
    xHDI_Income
Model 2: y ~ xTotalExpenditure + xSchooling + xHIV_AIDS + xAlcohol + xHepatitisB +
    xMeasles + xPolio + xBMI + xGDP + xHDI_Income
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1   1845 32355
2   1842 32327  3    27.461 0.5216 0.6675
```

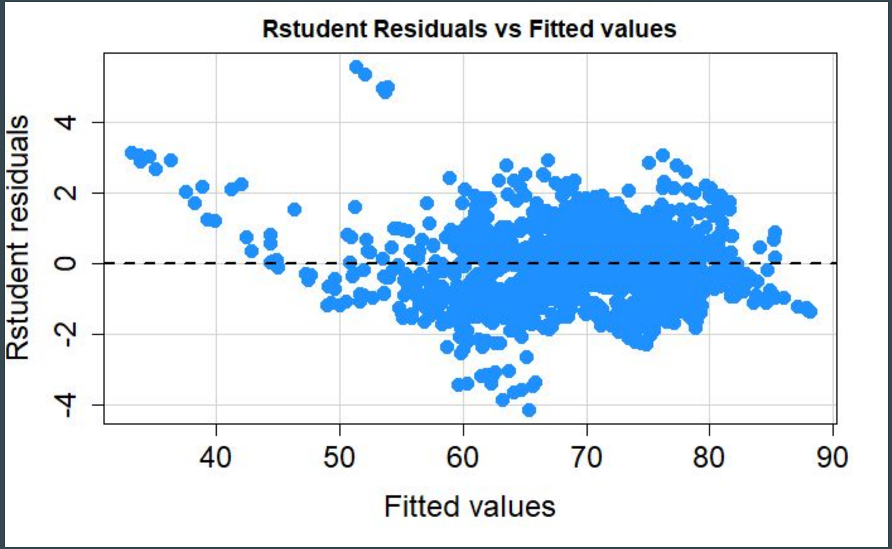
- Fit 1 = Model 1 (Reduced)

  Fit 2 = Model 2 (Full)
- Through ANOVA, we see that the p-value is sufficiently higher than 0.05, so we can conclude that the reduced model represents our dataset better.

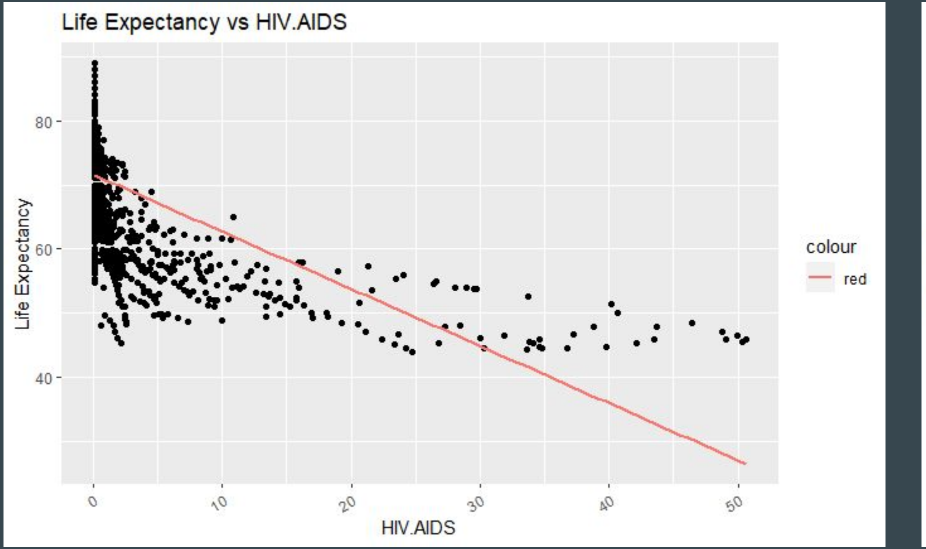
Residual Analysis Pre-transformation

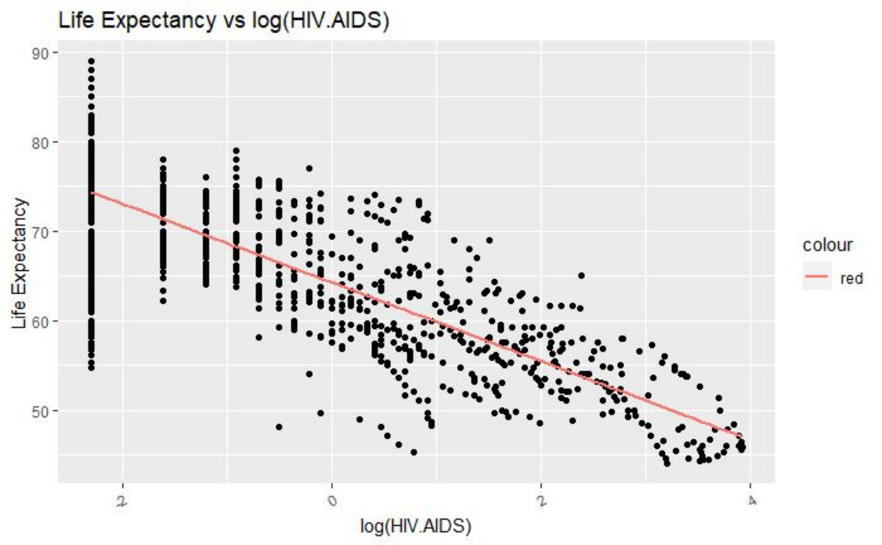
- We found that the Reduced Model is approximately normal with a few outliers seen
- We also saw that the Residuals were approximately evenly distributed vs. the fitted values.
- In the QQ-Plot the indices 62, 63 stand for the Country Antigua and Barbuda in the Year 2004 and the same Country Antigua and Barbuda in the year 2003.





Dataset Visualizations after a log Transformation of HIV/AIDS

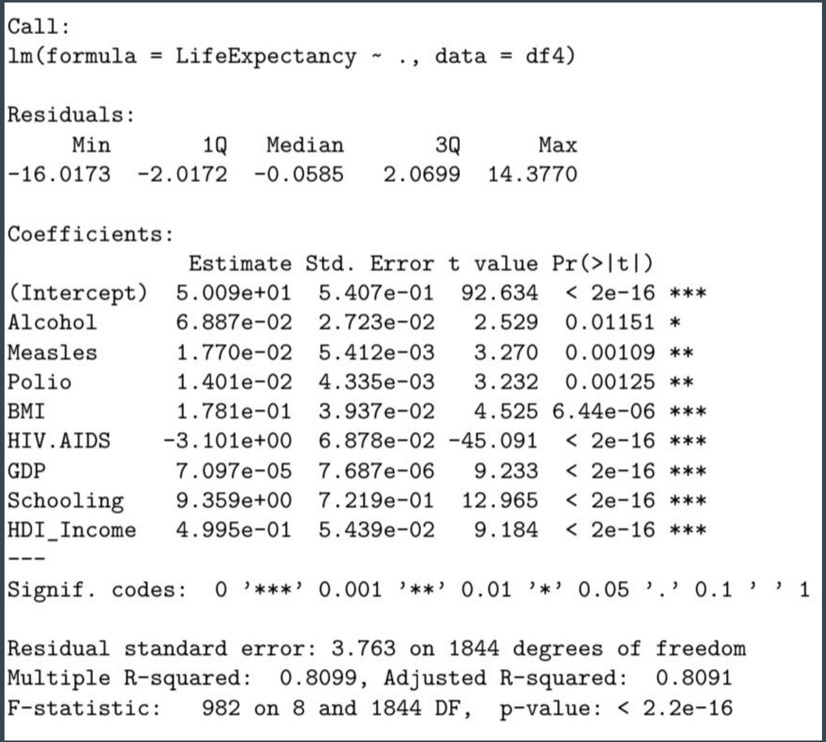




Before Transformation
R^2 = 0.3506

After Transformation
R^2 = 0.6429

Modifying Our Model Post Transformation



```
Call:
lm(formula = LifeExpectancy ~ ., data = df4)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.0173  -2.0172  -0.0585   2.0699  14.3770 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.009e+01  5.407e-01  92.634  < 2e-16 ***
Alcohol      6.887e-02  2.723e-02   2.529  0.01151 *  
Measles      1.770e-02  5.412e-03   3.270  0.00109 ** 
Polio        1.401e-02  4.335e-03   3.232  0.00125 ** 
BMI          1.781e-01  3.937e-02   4.525 6.44e-06 ***
HIV.AIDS    -3.101e+00  6.878e-02 -45.091  < 2e-16 ***
GDP          7.097e-05  7.687e-06   9.233  < 2e-16 ***
Schooling    9.359e+00  7.219e-01  12.965  < 2e-16 ***
HDI_Income   4.995e-01  5.439e-02   9.184  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.763 on 1844 degrees of freedom
Multiple R-squared:  0.8099, Adjusted R-squared:  0.8091 
F-statistic:   982 on 8 and 1844 DF,  p-value: < 2.2e-16
```

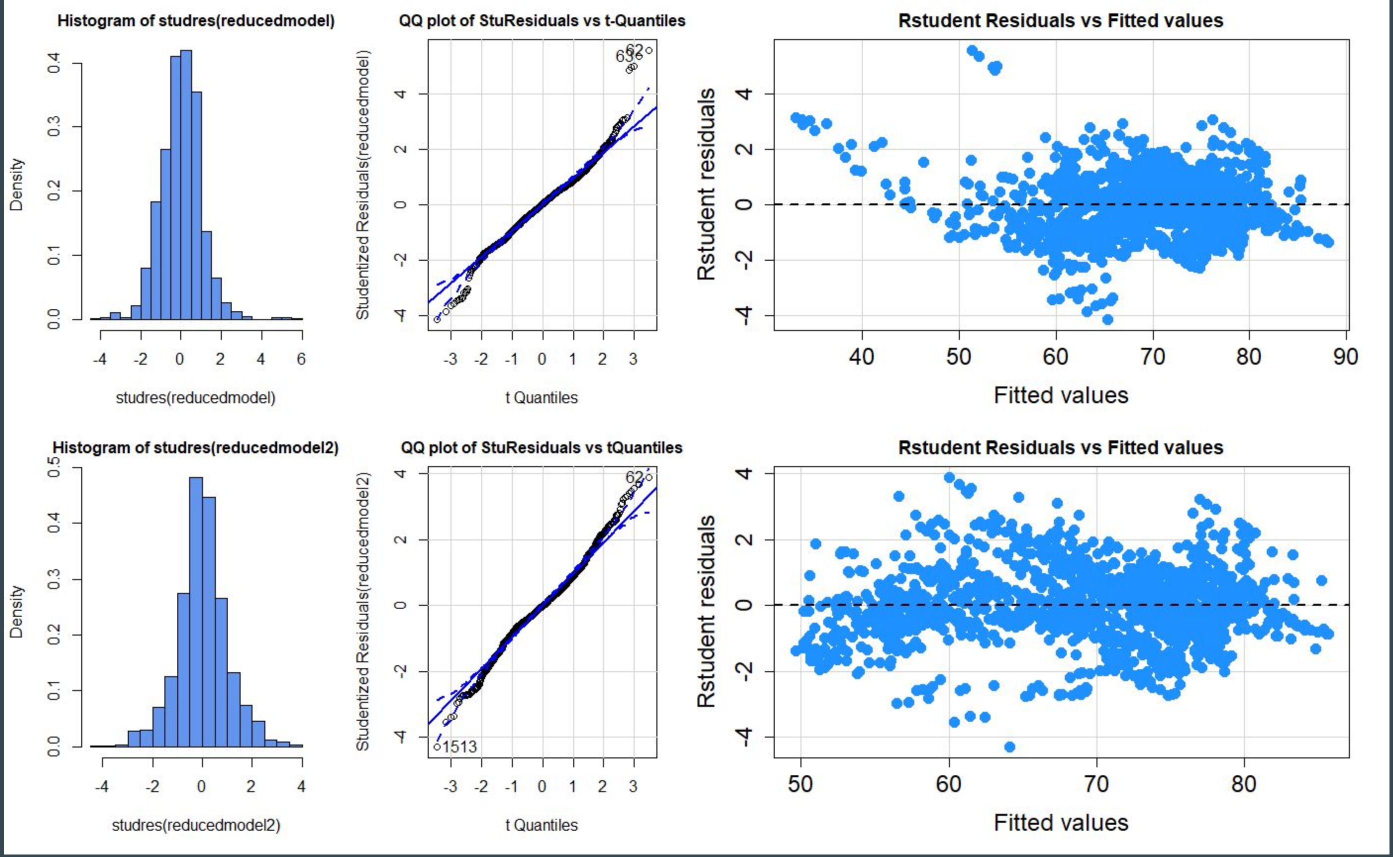
Reduced Model After Transformation.

- After transforming the HIV/AIDS predictor variable, we fit the full model to see what predictor variables best represented our data. This time we only had to omit Hepatitis B and Total Expenditure because their p-values were less than 0.05 (image not shown).
- Adjusted R^2 in reduced model Pre-Transformation: **0.7636**
- Adjusted R^2 in reduced model Post-Transformation: **0.8091**

Residual Analysis Pre-Transformation vs. Post-transformation

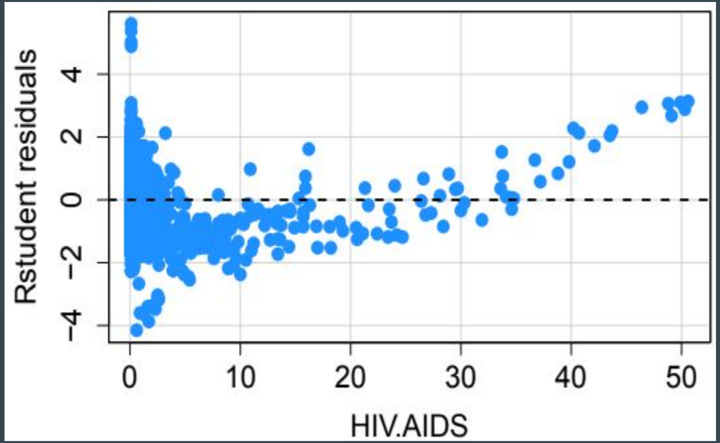
- In the QQ-Plot Pre-Transformation the indices 62, 63 stand for the Country Antigua and Barbuda in the Year 2004 and the same Country Antigua and Barbuda in the year 2003.

- In the QQ-Plot Post-Transformation the indices 62, 1513 stand for the Country Antigua and Barbuda in the Year 2004 and the Country Sierra Leone in the year 2014.

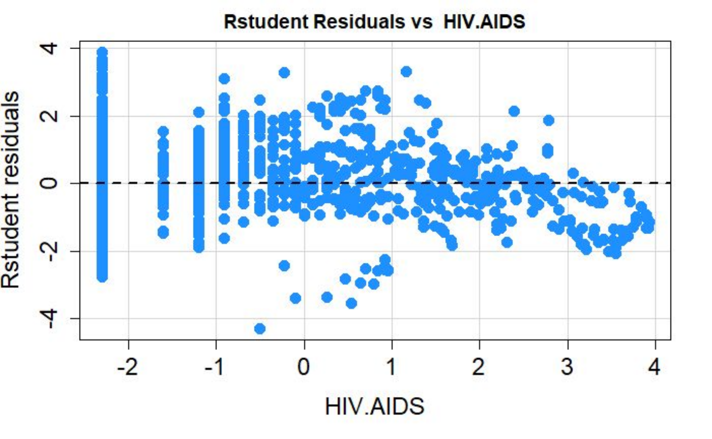


RStudent Residuals vs. HIV/AIDS Pre-transformation vs. Post-transformation

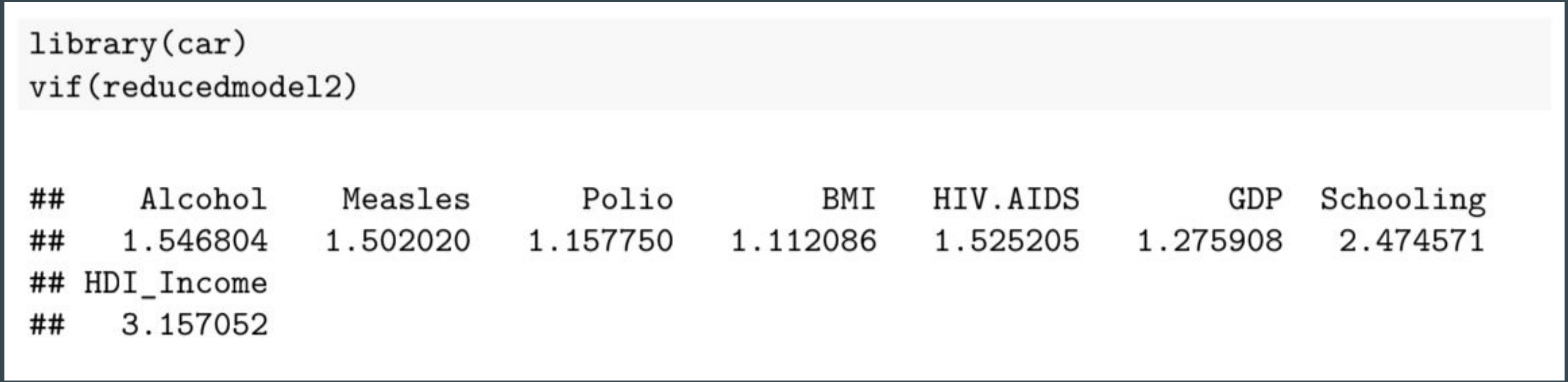
Before Transformation



After Transformation



VIF Post- Transformation



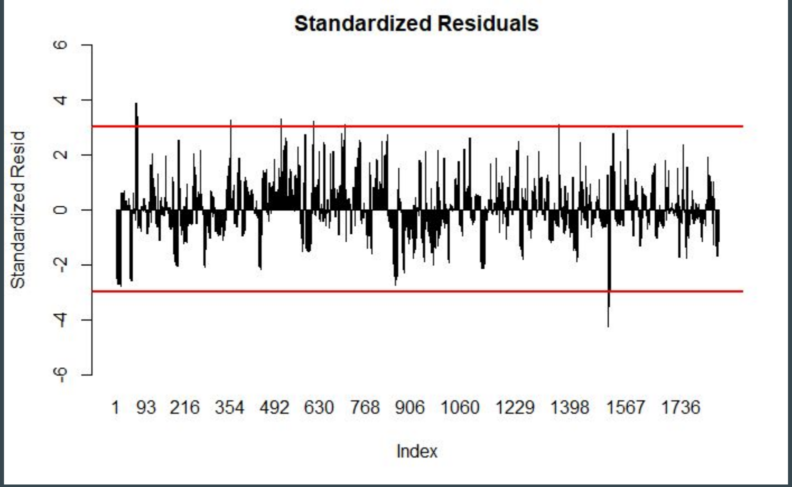
```
library(car)
vif(reducedmodel2)
```

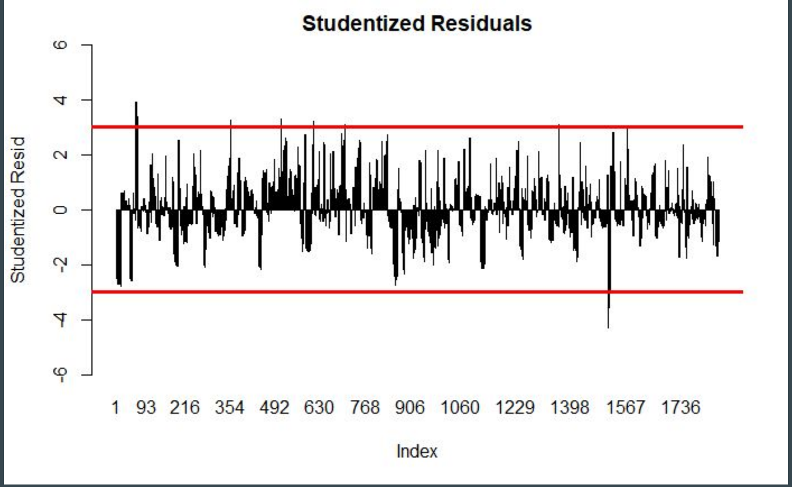
```
##    Alcohol    Measles      Polio        BMI   HIV.AIDS        GDP  Schooling
##   1.546804   1.502020   1.157750   1.112086   1.525205   1.275908   2.474571
## HDI_Income
##   3.157052
```

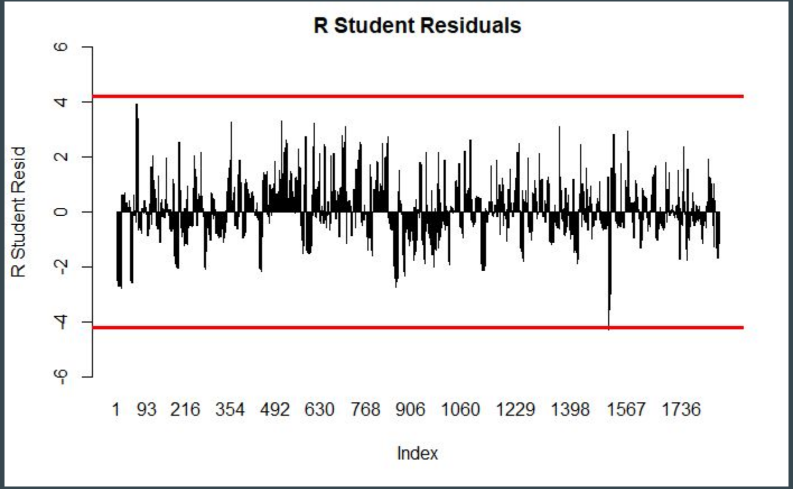
- No evidence of multicollinearity as all values are less than 10.
- Don’t need to remove any more variables

Influential Analysis Post-Transformation

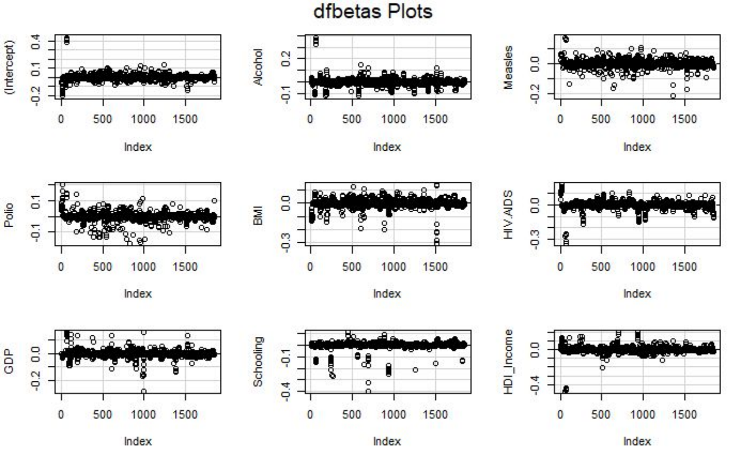
- We can see that there are a few data points that may be y-axis outliers and need further investigation.

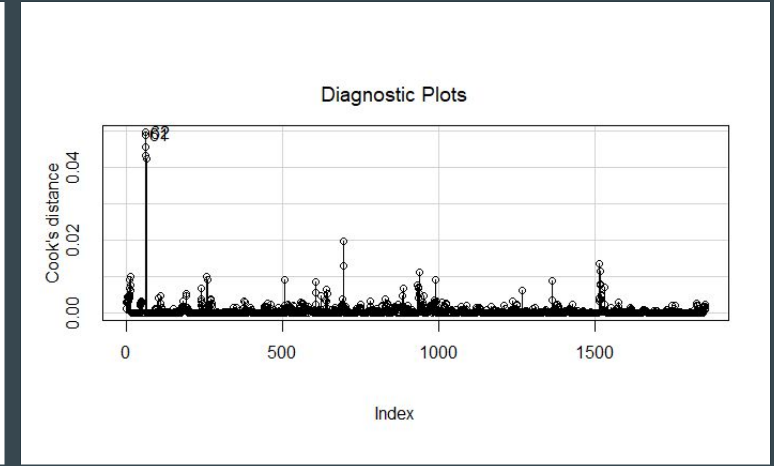


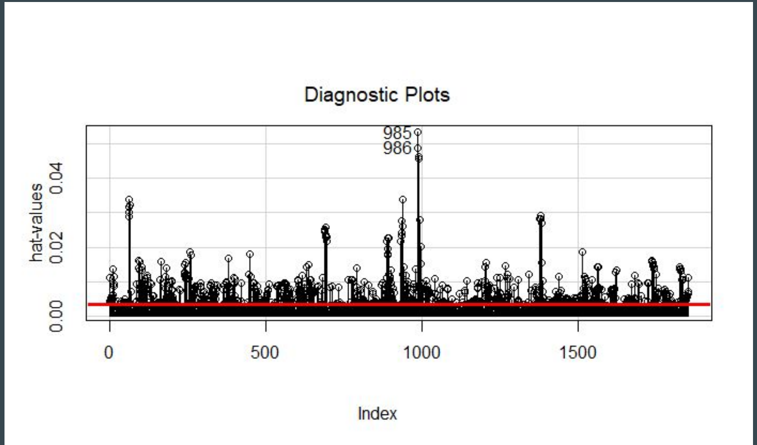




Influential Analysis Post- Transformation Continued



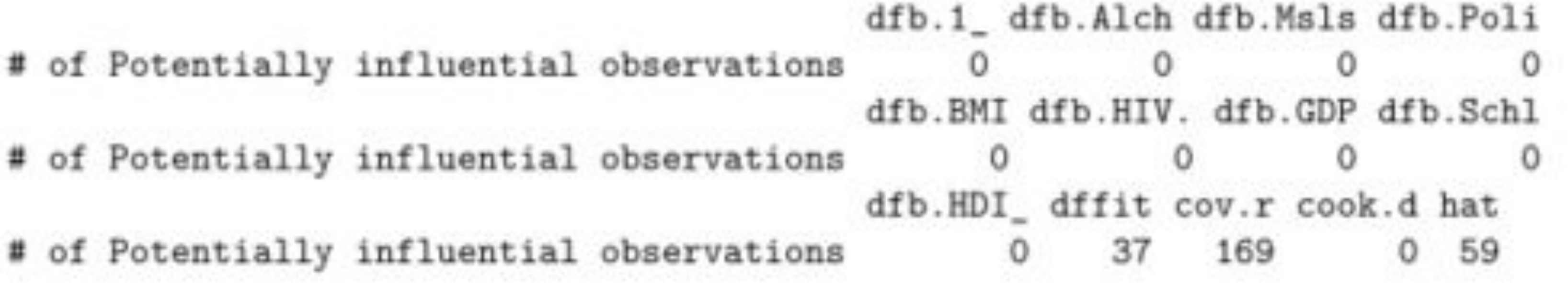




- The dfbetas Plots indicate no potentially influential observations.
- From the hat Diagnostic plot we can see that the leverage points that are above the red line that may be potentially influential.

- The indices 61, 62 stand for the Country Antigua and Barbuda in the Year 2005 and the same Country Antigua and Barbuda in the year 2004 and are potentially influential.
- The indices 985, 986 stand for the Country Luxembourg in the Year 2014 and the same Country Luxembourg in the year 2013 and are potentially influential and are leverage points.

Summary of the Influence Measures



```
                                          dfb.1_ dfb.Alch dfb.Msls dfb.Poli
# of Potentially influential observations      0        0        0        0
                                          dfb.BMI dfb.HIV. dfb.GDP dfb.Schl
# of Potentially influential observations       0        0       0        0
                                          dfb.HDI_ dffit cov.r cook.d hat
# of Potentially influential observations        0    37   169      0  59
```

- For our dffit measure we found a total number of 37 potentially influential observations of our deletion influence.
- For our measure of COVRATIO we found 169 instances of potentially influential points on our precision estimation.
- For our hat measure we found 59 potentially influential observations that leverage our model.
- We need to further investigate these observations that are potentially influential and be aware of their influence on our model.

Conclusion

- Through our analysis, we found that Alcohol, Measles, Polio, BMI, GDP, Schooling, HDI Income, and the log of HIV/AIDS are the best factors to predict the life Expectancy given data from a country and a specific year between 2000-2015.
- We hope to expand our work by analyzing data from more countries and past the year 2015.
- Life Expectancy = 5.009e+01 + (6.887e-02)Alcohol + (1.770e-02)Measles + (1.401e-02)Polio + (1.781e-01)BMI - (3.101)log(HIV/AIDS) + (7.097e-05)GDP + (9.359)Schooling + (4.995e-01)HDI_Income