Full Model

Live Laugh Life Expectancy

# Model Reduction

After cleaning the data, we are left with 15 explanatory variables for life expectancy, being Infant deaths, Under-five deaths, Adult mortality, Alcohol consumption, Hepatitis B, Measles, BMI, Polio, Diphtheria, Incidents of HIV, GDP per capita, Population (in millions), Thinness (ages 10–19 years), Thinness (ages 5–9 years), and Schooling.

# load data  
data <- read.csv("Life-Expectancy-Data-Updated.csv")  
  
# fit initial model  
model <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Under\_five\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Polio +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_ten\_nineteen\_years +  
 Thinness\_five\_nine\_years +  
 Schooling,  
 data = data  
)  
  
summary(model)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Under\_five\_deaths +   
## Adult\_mortality + Alcohol\_consumption + Hepatitis\_B + Measles +   
## BMI + Polio + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + Thinness\_ten\_nineteen\_years + Thinness\_five\_nine\_years +   
## Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.7675 -0.9412 -0.0560 0.8826 7.9078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.507e+01 6.114e-01 139.141 < 2e-16 \*\*\*  
## Infant\_deaths -5.152e-02 6.187e-03 -8.328 < 2e-16 \*\*\*  
## Under\_five\_deaths -5.295e-02 3.829e-03 -13.830 < 2e-16 \*\*\*  
## Adult\_mortality -4.893e-02 6.147e-04 -79.604 < 2e-16 \*\*\*  
## Alcohol\_consumption 9.023e-02 8.950e-03 10.081 < 2e-16 \*\*\*  
## Hepatitis\_B -9.262e-03 2.560e-03 -3.618 0.000302 \*\*\*  
## Measles 1.228e-03 1.728e-03 0.711 0.477264   
## BMI -1.674e-01 1.891e-02 -8.854 < 2e-16 \*\*\*  
## Polio 8.691e-04 5.876e-03 0.148 0.882421   
## Diphtheria 3.257e-03 5.926e-03 0.550 0.582629   
## Incidents\_HIV 1.005e-01 1.827e-02 5.502 4.08e-08 \*\*\*  
## GDP\_per\_capita 3.043e-05 2.146e-06 14.181 < 2e-16 \*\*\*  
## Population\_mln -1.814e-04 2.011e-04 -0.902 0.366939   
## Thinness\_ten\_nineteen\_years -3.313e-02 1.725e-02 -1.921 0.054867 .   
## Thinness\_five\_nine\_years -2.449e-03 1.690e-02 -0.145 0.884797   
## Schooling 1.119e-01 1.668e-02 6.709 2.35e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.369 on 2848 degrees of freedom  
## Multiple R-squared: 0.9789, Adjusted R-squared: 0.9788   
## F-statistic: 8817 on 15 and 2848 DF, p-value: < 2.2e-16

After fitting the full model, we have 9 highly statistically significant variables (Infant\_deaths, Under\_five\_deaths, Adult\_mortality, Alcohol\_consumption, Hepatitis\_B, BMI, Incidents\_HIV, GDP\_per\_capita, and Schooling). We have an overall p-value or 2.2e-16 Additionally, our R-squared is 0.9789 and Adjusted R-squared is 0.9788. This indicates that there might be overfitting issues within our model Next, we look at the variance inflation factors of each regressor variable to determine if there are any multicollinearity issues.

library(car)

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3

vif(model)

## Infant\_deaths Under\_five\_deaths   
## 44.329144 44.469991   
## Adult\_mortality Alcohol\_consumption   
## 7.618888 1.939630   
## Hepatitis\_B Measles   
## 2.560952 1.587422   
## BMI Polio   
## 2.627136 11.990081   
## Diphtheria Incidents\_HIV   
## 12.940367 2.890280   
## GDP\_per\_capita Population\_mln   
## 2.016715 1.150097   
## Thinness\_ten\_nineteen\_years Thinness\_five\_nine\_years   
## 8.949993 8.936053   
## Schooling   
## 4.273524

Multicollinearity issues, that is variables with a vif greater than 10 arise within the regressors Infant\_deaths (44.329144), Under\_five\_deaths (44.469991), Polio (11.990081), Diphtheria (12.940367). Additionally, Thinness\_ten\_nineteen\_years (8.949993) and Thinness\_five\_nine\_years (8.936053) also have rather high vif’s, but still bellow 10.

To combat this we will begin by removing Under\_five\_deaths.

model1 <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Polio +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_ten\_nineteen\_years +  
 Thinness\_five\_nine\_years +  
 Schooling,  
 data = data  
)  
  
summary(model1)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + Measles + BMI + Polio +   
## Diphtheria + Incidents\_HIV + GDP\_per\_capita + Population\_mln +   
## Thinness\_ten\_nineteen\_years + Thinness\_five\_nine\_years +   
## Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3062 -0.9366 -0.0298 0.9301 5.1858   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.398e+01 6.262e-01 134.110 < 2e-16 \*\*\*  
## Infant\_deaths -1.290e-01 2.715e-03 -47.501 < 2e-16 \*\*\*  
## Adult\_mortality -5.019e-02 6.279e-04 -79.934 < 2e-16 \*\*\*  
## Alcohol\_consumption 7.103e-02 9.132e-03 7.778 1.02e-14 \*\*\*  
## Hepatitis\_B -1.337e-02 2.626e-03 -5.090 3.82e-07 \*\*\*  
## Measles 2.159e-04 1.783e-03 0.121 0.9036   
## BMI -1.317e-01 1.934e-02 -6.806 1.22e-11 \*\*\*  
## Polio 1.491e-03 6.068e-03 0.246 0.8059   
## Diphtheria 1.166e-02 6.088e-03 1.915 0.0556 .   
## Incidents\_HIV 9.555e-02 1.886e-02 5.065 4.35e-07 \*\*\*  
## GDP\_per\_capita 2.312e-05 2.148e-06 10.762 < 2e-16 \*\*\*  
## Population\_mln -1.072e-04 2.076e-04 -0.516 0.6058   
## Thinness\_ten\_nineteen\_years -3.595e-02 1.781e-02 -2.018 0.0437 \*   
## Thinness\_five\_nine\_years 1.355e-02 1.742e-02 0.778 0.4366   
## Schooling 1.485e-01 1.701e-02 8.728 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.414 on 2849 degrees of freedom  
## Multiple R-squared: 0.9775, Adjusted R-squared: 0.9774   
## F-statistic: 8842 on 14 and 2849 DF, p-value: < 2.2e-16

vif(model1)

## Infant\_deaths Adult\_mortality   
## 8.003745 7.452012   
## Alcohol\_consumption Hepatitis\_B   
## 1.892972 2.526519   
## Measles BMI   
## 1.584574 2.578076   
## Polio Diphtheria   
## 11.989379 12.804362   
## Incidents\_HIV GDP\_per\_capita   
## 2.889161 1.894235   
## Population\_mln Thinness\_ten\_nineteen\_years   
## 1.149276 8.948743   
## Thinness\_five\_nine\_years Schooling   
## 8.894194 4.166226

After removing Under\_five\_deaths, we have the same highly statistically predictors, as well as Thinness\_ten\_nineteen\_years being a somewhat significant predictor. Furthermore, The vif for Infant\_deaths have improved. We still have a multicollinearity issue with Polio and Diphtheria, so it is in our best interest to remove one of them. We will remove polio because it has the highest p-value (0.8059) while having a large vif.

model2 <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_ten\_nineteen\_years +  
 Thinness\_five\_nine\_years +  
 Schooling,  
 data = data  
)  
  
summary(model2)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + Measles + BMI + Diphtheria +   
## Incidents\_HIV + GDP\_per\_capita + Population\_mln + Thinness\_ten\_nineteen\_years +   
## Thinness\_five\_nine\_years + Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3159 -0.9363 -0.0284 0.9320 5.1873   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.399e+01 6.225e-01 134.927 < 2e-16 \*\*\*  
## Infant\_deaths -1.291e-01 2.678e-03 -48.198 < 2e-16 \*\*\*  
## Adult\_mortality -5.018e-02 6.276e-04 -79.966 < 2e-16 \*\*\*  
## Alcohol\_consumption 7.097e-02 9.127e-03 7.775 1.04e-14 \*\*\*  
## Hepatitis\_B -1.338e-02 2.625e-03 -5.096 3.69e-07 \*\*\*  
## Measles 2.616e-04 1.773e-03 0.148 0.882688   
## BMI -1.314e-01 1.932e-02 -6.803 1.24e-11 \*\*\*  
## Diphtheria 1.291e-02 3.341e-03 3.864 0.000114 \*\*\*  
## Incidents\_HIV 9.567e-02 1.886e-02 5.074 4.15e-07 \*\*\*  
## GDP\_per\_capita 2.309e-05 2.146e-06 10.764 < 2e-16 \*\*\*  
## Population\_mln -1.082e-04 2.075e-04 -0.521 0.602240   
## Thinness\_ten\_nineteen\_years -3.580e-02 1.780e-02 -2.011 0.044412 \*   
## Thinness\_five\_nine\_years 1.348e-02 1.741e-02 0.774 0.438837   
## Schooling 1.484e-01 1.701e-02 8.728 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.414 on 2850 degrees of freedom  
## Multiple R-squared: 0.9775, Adjusted R-squared: 0.9774   
## F-statistic: 9526 on 13 and 2850 DF, p-value: < 2.2e-16

vif(model2)

## Infant\_deaths Adult\_mortality   
## 7.789556 7.447236   
## Alcohol\_consumption Hepatitis\_B   
## 1.891505 2.525566   
## Measles BMI   
## 1.567283 2.573423   
## Diphtheria Incidents\_HIV   
## 3.857871 2.887258   
## GDP\_per\_capita Population\_mln   
## 1.890494 1.148823   
## Thinness\_ten\_nineteen\_years Thinness\_five\_nine\_years   
## 8.937990 8.891916   
## Schooling   
## 4.166107

Another solution to this could possibly combining all of the vaccine percentage variables, Hepatitis\_B, Measles, Polio, and Diphtheria, into one variable, Vaccine\_percentage

data$Vaccine\_percentage <- rowMeans(data[, c("Hepatitis\_B", "Measles", "Polio", "Diphtheria")], na.rm = TRUE)  
  
model3 <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 BMI +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_ten\_nineteen\_years +  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Vaccine\_percentage,  
 data = data  
)  
  
summary(model3)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + BMI + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + Thinness\_ten\_nineteen\_years + Thinness\_five\_nine\_years +   
## Schooling + Vaccine\_percentage, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.2787 -0.9358 -0.0158 0.9075 5.3886   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.452e+01 6.139e-01 137.674 < 2e-16 \*\*\*  
## Infant\_deaths -1.313e-01 2.618e-03 -50.133 < 2e-16 \*\*\*  
## Adult\_mortality -5.023e-02 6.290e-04 -79.851 < 2e-16 \*\*\*  
## Alcohol\_consumption 7.140e-02 9.147e-03 7.805 8.26e-15 \*\*\*  
## BMI -1.420e-01 1.923e-02 -7.384 2.00e-13 \*\*\*  
## Incidents\_HIV 9.834e-02 1.893e-02 5.196 2.18e-07 \*\*\*  
## GDP\_per\_capita 2.354e-05 2.153e-06 10.932 < 2e-16 \*\*\*  
## Population\_mln -5.310e-05 2.080e-04 -0.255 0.7985   
## Thinness\_ten\_nineteen\_years -3.872e-02 1.781e-02 -2.174 0.0298 \*   
## Thinness\_five\_nine\_years 1.548e-02 1.739e-02 0.890 0.3736   
## Schooling 1.531e-01 1.693e-02 9.039 < 2e-16 \*\*\*  
## Vaccine\_percentage -2.668e-03 2.861e-03 -0.933 0.3511   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.42 on 2852 degrees of freedom  
## Multiple R-squared: 0.9773, Adjusted R-squared: 0.9772   
## F-statistic: 1.116e+04 on 11 and 2852 DF, p-value: < 2.2e-16

vif(model3)

## Infant\_deaths Adult\_mortality   
## 7.383893 7.418883   
## Alcohol\_consumption BMI   
## 1.883987 2.528624   
## Incidents\_HIV GDP\_per\_capita   
## 2.885007 1.887546   
## Population\_mln Thinness\_ten\_nineteen\_years   
## 1.144758 8.876975   
## Thinness\_five\_nine\_years Schooling   
## 8.797860 4.095898   
## Vaccine\_percentage   
## 2.236409

There does not seem to be a significant difference between these two models. If we chose to go with removing polio, we would have to remove each of the vaccines other than Diphteria as it is the only one that was statistically significant. This is likely because the vaccination somewhat consistent on each vaccine, if they had one they are likely to have the others.

Moving forward with removing Polio, our next highest p-value is measles, so we will remove that.

model4 <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_ten\_nineteen\_years +  
 Thinness\_five\_nine\_years +  
 Schooling,  
 data = data  
)  
  
summary(model4)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + BMI + Diphtheria + Incidents\_HIV +   
## GDP\_per\_capita + Population\_mln + Thinness\_ten\_nineteen\_years +   
## Thinness\_five\_nine\_years + Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3119 -0.9347 -0.0306 0.9271 5.1925   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.400e+01 6.194e-01 135.609 < 2e-16 \*\*\*  
## Infant\_deaths -1.291e-01 2.678e-03 -48.210 < 2e-16 \*\*\*  
## Adult\_mortality -5.019e-02 6.262e-04 -80.157 < 2e-16 \*\*\*  
## Alcohol\_consumption 7.104e-02 9.113e-03 7.795 8.96e-15 \*\*\*  
## Hepatitis\_B -1.333e-02 2.604e-03 -5.119 3.28e-07 \*\*\*  
## BMI -1.314e-01 1.931e-02 -6.803 1.24e-11 \*\*\*  
## Diphtheria 1.295e-02 3.327e-03 3.893 0.000101 \*\*\*  
## Incidents\_HIV 9.575e-02 1.884e-02 5.082 3.98e-07 \*\*\*  
## GDP\_per\_capita 2.309e-05 2.145e-06 10.765 < 2e-16 \*\*\*  
## Population\_mln -1.102e-04 2.070e-04 -0.532 0.594600   
## Thinness\_ten\_nineteen\_years -3.554e-02 1.771e-02 -2.007 0.044885 \*   
## Thinness\_five\_nine\_years 1.317e-02 1.728e-02 0.762 0.446030   
## Schooling 1.487e-01 1.688e-02 8.810 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.414 on 2851 degrees of freedom  
## Multiple R-squared: 0.9775, Adjusted R-squared: 0.9774   
## F-statistic: 1.032e+04 on 12 and 2851 DF, p-value: < 2.2e-16

vif(model4)

## Infant\_deaths Adult\_mortality   
## 7.788875 7.416150   
## Alcohol\_consumption Hepatitis\_B   
## 1.886438 2.486117   
## BMI Diphtheria   
## 2.570657 3.827148   
## Incidents\_HIV GDP\_per\_capita   
## 2.884409 1.890421   
## Population\_mln Thinness\_ten\_nineteen\_years   
## 1.143810 8.851921   
## Thinness\_five\_nine\_years Schooling   
## 8.762341 4.106941

Running the new model, we can see both Diphtheria and Hepatitis\_B are highly significant predictors so it is preferable to use this model rather than the average, Vaccine\_percentage. The next highest p-value is Population\_mlm, so we will move forward with removing that. This makes sense to not be a significatn predictor of life expectancy because there are plenty of high quality of life areas with low populations, low quality of life areas with high populations, and everything in between.

model5 <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Thinness\_five\_nine\_years +  
 Schooling,  
 data = data  
)  
  
summary(model5)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + BMI + Diphtheria + Incidents\_HIV +   
## GDP\_per\_capita + Thinness\_ten\_nineteen\_years + Thinness\_five\_nine\_years +   
## Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3152 -0.9332 -0.0287 0.9246 5.1923   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.397e+01 6.157e-01 136.384 < 2e-16 \*\*\*  
## Infant\_deaths -1.291e-01 2.677e-03 -48.213 < 2e-16 \*\*\*  
## Adult\_mortality -5.016e-02 6.228e-04 -80.537 < 2e-16 \*\*\*  
## Alcohol\_consumption 7.090e-02 9.109e-03 7.784 9.75e-15 \*\*\*  
## Hepatitis\_B -1.322e-02 2.596e-03 -5.093 3.75e-07 \*\*\*  
## BMI -1.302e-01 1.919e-02 -6.786 1.39e-11 \*\*\*  
## Diphtheria 1.293e-02 3.327e-03 3.887 0.000104 \*\*\*  
## Incidents\_HIV 9.540e-02 1.883e-02 5.067 4.31e-07 \*\*\*  
## GDP\_per\_capita 2.318e-05 2.139e-06 10.835 < 2e-16 \*\*\*  
## Thinness\_ten\_nineteen\_years -3.620e-02 1.767e-02 -2.049 0.040560 \*   
## Thinness\_five\_nine\_years 1.277e-02 1.726e-02 0.739 0.459688   
## Schooling 1.480e-01 1.683e-02 8.797 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.414 on 2852 degrees of freedom  
## Multiple R-squared: 0.9775, Adjusted R-squared: 0.9774   
## F-statistic: 1.126e+04 on 11 and 2852 DF, p-value: < 2.2e-16

vif(model5)

## Infant\_deaths Adult\_mortality   
## 7.788245 7.338100   
## Alcohol\_consumption Hepatitis\_B   
## 1.884960 2.470324   
## BMI Diphtheria   
## 2.539707 3.826591   
## Incidents\_HIV GDP\_per\_capita   
## 2.880807 1.880260   
## Thinness\_ten\_nineteen\_years Thinness\_five\_nine\_years   
## 8.808948 8.745244   
## Schooling   
## 4.080587

Thinness\_five\_nine\_years has the next highest p-value as well as rather high vif along with Thinness\_ten\_nineteen\_years, so it makes sense to remove Thinness\_five\_nine\_years.

model6 <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Schooling,  
 data = data  
)  
  
summary(model6)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + BMI + Diphtheria + Incidents\_HIV +   
## GDP\_per\_capita + Thinness\_ten\_nineteen\_years + Schooling,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3181 -0.9414 -0.0269 0.9250 5.1922   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.403e+01 6.089e-01 138.006 < 2e-16 \*\*\*  
## Infant\_deaths -1.291e-01 2.676e-03 -48.253 < 2e-16 \*\*\*  
## Adult\_mortality -5.017e-02 6.225e-04 -80.585 < 2e-16 \*\*\*  
## Alcohol\_consumption 7.055e-02 9.095e-03 7.756 1.21e-14 \*\*\*  
## Hepatitis\_B -1.331e-02 2.593e-03 -5.133 3.04e-07 \*\*\*  
## BMI -1.324e-01 1.896e-02 -6.982 3.60e-12 \*\*\*  
## Diphtheria 1.297e-02 3.326e-03 3.901 9.80e-05 \*\*\*  
## Incidents\_HIV 9.622e-02 1.879e-02 5.119 3.27e-07 \*\*\*  
## GDP\_per\_capita 2.304e-05 2.130e-06 10.812 < 2e-16 \*\*\*  
## Thinness\_ten\_nineteen\_years -2.461e-02 8.147e-03 -3.020 0.00255 \*\*   
## Schooling 1.485e-01 1.681e-02 8.831 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.413 on 2853 degrees of freedom  
## Multiple R-squared: 0.9775, Adjusted R-squared: 0.9774   
## F-statistic: 1.239e+04 on 10 and 2853 DF, p-value: < 2.2e-16

vif(model6)

## Infant\_deaths Adult\_mortality   
## 7.782878 7.333802   
## Alcohol\_consumption Hepatitis\_B   
## 1.879719 2.465090   
## BMI Diphtheria   
## 2.480304 3.825432   
## Incidents\_HIV GDP\_per\_capita   
## 2.870815 1.865489   
## Thinness\_ten\_nineteen\_years Schooling   
## 1.873529 4.074861

After this, all of the predictors are highly statistically significant. We still have a very high R-squared value of 0.9775. To combat this I think it may be a good idea to only have either Infant\_deaths or adult mortality, as they both measure their respective deaths per 1000 in the population.

model7 <- lm(Life\_expectancy ~   
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Schooling,  
 data = data  
)  
  
summary(model7)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Adult\_mortality + Alcohol\_consumption +   
## Hepatitis\_B + BMI + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Thinness\_ten\_nineteen\_years + Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.3883 -1.2058 -0.0079 1.2484 8.7670   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.098e+01 7.350e-01 96.567 < 2e-16 \*\*\*  
## Adult\_mortality -6.732e-02 6.886e-04 -97.767 < 2e-16 \*\*\*  
## Alcohol\_consumption 1.346e-01 1.212e-02 11.100 < 2e-16 \*\*\*  
## Hepatitis\_B -1.467e-02 3.493e-03 -4.200 2.75e-05 \*\*\*  
## BMI 4.859e-02 2.505e-02 1.940 0.0525 .   
## Diphtheria 7.365e-02 4.149e-03 17.751 < 2e-16 \*\*\*  
## Incidents\_HIV 3.077e-01 2.463e-02 12.496 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.137e-05 2.852e-06 3.987 6.86e-05 \*\*\*  
## Thinness\_ten\_nineteen\_years -1.667e-02 1.097e-02 -1.519 0.1289   
## Schooling 4.601e-01 2.092e-02 21.996 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.904 on 2854 degrees of freedom  
## Multiple R-squared: 0.9591, Adjusted R-squared: 0.959   
## F-statistic: 7442 on 9 and 2854 DF, p-value: < 2.2e-16

vif(model7)

## Adult\_mortality Alcohol\_consumption   
## 4.942269 1.839719   
## Hepatitis\_B BMI   
## 2.464798 2.383265   
## Diphtheria Incidents\_HIV   
## 3.278703 2.714672   
## GDP\_per\_capita Thinness\_ten\_nineteen\_years   
## 1.841474 1.872765   
## Schooling   
## 3.473744

After removing Infant\_deaths, we still have a high R-squared of 0.9591. Now we will try removing Adult\_mortality and keeping Infant\_deaths.

model8 <- lm(Life\_expectancy ~   
 Infant\_deaths +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Schooling,  
 data = data  
)  
  
summary(model8)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Alcohol\_consumption +   
## Hepatitis\_B + BMI + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Thinness\_ten\_nineteen\_years + Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.8589 -1.5134 0.1144 1.6817 11.4451   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.423e+01 1.080e+00 68.748 < 2e-16 \*\*\*  
## Infant\_deaths -2.523e-01 3.976e-03 -63.456 < 2e-16 \*\*\*  
## Alcohol\_consumption -6.113e-02 1.619e-02 -3.776 0.000163 \*\*\*  
## Hepatitis\_B -1.684e-02 4.692e-03 -3.589 0.000337 \*\*\*  
## BMI 8.984e-02 3.395e-02 2.646 0.008186 \*\*   
## Diphtheria 1.453e-02 6.019e-03 2.414 0.015829 \*   
## Incidents\_HIV -1.044e+00 2.240e-02 -46.587 < 2e-16 \*\*\*  
## GDP\_per\_capita 8.872e-05 3.562e-06 24.906 < 2e-16 \*\*\*  
## Thinness\_ten\_nineteen\_years 2.926e-02 1.469e-02 1.991 0.046562 \*   
## Schooling 3.488e-02 3.032e-02 1.150 0.250042   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.558 on 2854 degrees of freedom  
## Multiple R-squared: 0.9263, Adjusted R-squared: 0.926   
## F-statistic: 3984 on 9 and 2854 DF, p-value: < 2.2e-16

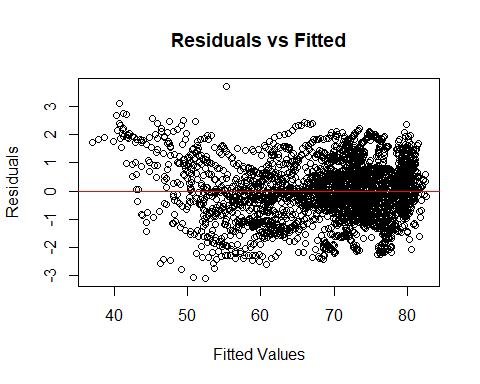
vif(model8)

## Infant\_deaths Alcohol\_consumption   
## 5.244902 1.819045   
## Hepatitis\_B BMI   
## 2.464386 2.427842   
## Diphtheria Incidents\_HIV   
## 3.825303 1.245089   
## GDP\_per\_capita Thinness\_ten\_nineteen\_years   
## 1.592441 1.860918   
## Schooling   
## 4.046216

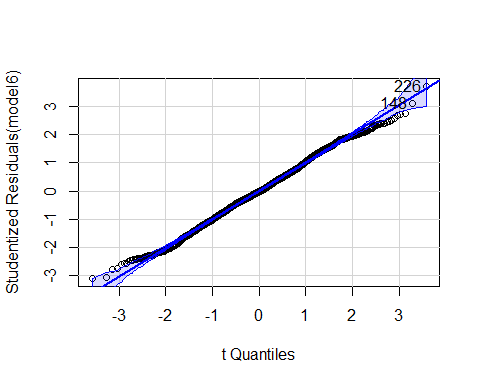
Once again, the R-Squared is still very high, with a value of 0.9263. I think the effort if trying to figure this out is not worth it, we will keep both variables and use model6. We opted for this model as it strikes a balance between high explanatory power and acceptable multicollinearity

# Residual Analysis

library(MASS)  
  
# residuals vs fitted values  
plot(fitted(model6), studres(model6),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

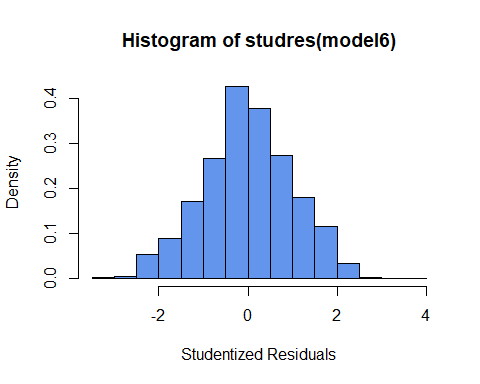


# QQ plot  
qqPlot(model6)



## [1] 148 226

# histogram of residuals  
hist(studres(model6), breaks = 15, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



The QQ plot values are very close to the line which is good. The histogram of the residuals also appears to follow a normal distribution. However, the Residuals vs. Fitted value plot indicates a transformation is needed.

# Assessment of model fit

Before we perform any transformations, we want to ensure that we have the best possible model fit. We can do this by using the ols\_step\_all\_possible, ols\_step\_forward\_p, and ols\_step\_backward\_p to see if those results match up with ours. Truth be told, we did not realize those functions existed before we began reducing the model manually.

# find all possible permutations of the full model (this is commented out because it took so long, it is saved as a csv file now)  
# MS\_ALL = ols\_step\_all\_possible(model)  
# write.csv(MS\_ALL$result, "MS\_ALL\_results.csv", row.names = FALSE)  
MS\_ALL <- read.csv("MS\_ALL\_results.csv")  
  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.2

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(olsrr)

## Warning: package 'olsrr' was built under R version 4.3.3

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:MASS':  
##   
## cement

## The following object is masked from 'package:datasets':  
##   
## rivers

# find model with highest adjusted r squared  
# adjr <- MS\_ALL$result$adjr  
# max\_adjr <- which.max(adjr)  
# best\_model\_adjr <- MS\_ALL$result[max\_adjr, ]  
# (commenting out original code and using new working with loaded data frame)  
  
adjr <- MS\_ALL$adjr  
max\_adjr <- which.max(adjr)  
best\_model\_adjr <- MS\_ALL[max\_adjr, ]  
  
best\_model\_adjr$predictors

## [1] "Infant\_deaths Under\_five\_deaths Adult\_mortality Alcohol\_consumption Hepatitis\_B BMI Diphtheria Incidents\_HIV GDP\_per\_capita Thinness\_ten\_nineteen\_years Schooling"

# find model with lowest AIC  
# aic <- MS\_ALL$result$aic  
# min\_aic <- which.min(aic)  
# best\_model\_aic <- MS\_ALL$result[min\_aic, ]  
  
aic <- MS\_ALL$aic  
min\_aic <- which.min(aic)  
best\_model\_aic <- MS\_ALL[min\_aic, ]  
  
best\_model\_aic$predictors

## [1] "Infant\_deaths Under\_five\_deaths Adult\_mortality Alcohol\_consumption Hepatitis\_B BMI Incidents\_HIV GDP\_per\_capita Thinness\_ten\_nineteen\_years Schooling"

# find model with lowest BIC  
# bic <- MS\_ALL$result$sbic  
# min\_bic <- which.min(bic)  
  
bic <- MS\_ALL$sbic  
min\_bic <- which.min(bic)  
min\_bic == min\_aic

## [1] TRUE

# note: did not run best model because the best bic has the same model as the best aic  
  
# stepwise selection with p-values  
ols\_stepwise\_forward <- ols\_step\_forward\_p(model)  
ols\_stepwise\_forward$metrics$variable

## [1] "Adult\_mortality" "Infant\_deaths"   
## [3] "GDP\_per\_capita" "Alcohol\_consumption"   
## [5] "Under\_five\_deaths" "Schooling"   
## [7] "BMI" "Incidents\_HIV"   
## [9] "Thinness\_ten\_nineteen\_years" "Hepatitis\_B"   
## [11] "Diphtheria"

ols\_stepwise\_backward <- ols\_step\_backward\_p(model)  
ols\_stepwise\_backward$metrics$variable

## [1] "Thinness\_five\_nine\_years" "Polio"   
## [3] "Measles" "Population\_mln"

# fit the best models that we found  
adjr\_model <- lm(Life\_expectancy ~ Infant\_deaths+Under\_five\_deaths+Adult\_mortality+Alcohol\_consumption+Hepatitis\_B+BMI+Diphtheria+Incidents\_HIV+GDP\_per\_capita+Thinness\_ten\_nineteen\_years+Schooling, data = data)  
summary(adjr\_model)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Under\_five\_deaths +   
## Adult\_mortality + Alcohol\_consumption + Hepatitis\_B + BMI +   
## Diphtheria + Incidents\_HIV + GDP\_per\_capita + Thinness\_ten\_nineteen\_years +   
## Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.7560 -0.9335 -0.0503 0.8733 7.9179   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.503e+01 5.940e-01 143.147 < 2e-16 \*\*\*  
## Infant\_deaths -5.200e-02 6.155e-03 -8.448 < 2e-16 \*\*\*  
## Under\_five\_deaths -5.265e-02 3.811e-03 -13.815 < 2e-16 \*\*\*  
## Adult\_mortality -4.890e-02 6.097e-04 -80.198 < 2e-16 \*\*\*  
## Alcohol\_consumption 9.031e-02 8.922e-03 10.122 < 2e-16 \*\*\*  
## Hepatitis\_B -8.838e-03 2.531e-03 -3.491 0.000488 \*\*\*  
## BMI -1.639e-01 1.850e-02 -8.857 < 2e-16 \*\*\*  
## Diphtheria 4.194e-03 3.283e-03 1.278 0.201516   
## Incidents\_HIV 1.001e-01 1.820e-02 5.498 4.18e-08 \*\*\*  
## GDP\_per\_capita 3.056e-05 2.134e-06 14.324 < 2e-16 \*\*\*  
## Thinness\_ten\_nineteen\_years -3.713e-02 7.940e-03 -4.676 3.06e-06 \*\*\*  
## Schooling 1.121e-01 1.649e-02 6.798 1.29e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.369 on 2852 degrees of freedom  
## Multiple R-squared: 0.9789, Adjusted R-squared: 0.9788   
## F-statistic: 1.203e+04 on 11 and 2852 DF, p-value: < 2.2e-16

vif(adjr\_model)

## Infant\_deaths Under\_five\_deaths   
## 43.911168 44.101047   
## Adult\_mortality Alcohol\_consumption   
## 7.503580 1.929304   
## Hepatitis\_B BMI   
## 2.506062 2.518500   
## Diphtheria Incidents\_HIV   
## 3.974430 2.871487   
## GDP\_per\_capita Thinness\_ten\_nineteen\_years   
## 1.995601 1.898273   
## Schooling   
## 4.181415

adjr\_model1 <- lm(Life\_expectancy ~ Infant\_deaths+Adult\_mortality+Alcohol\_consumption+Hepatitis\_B+BMI+Diphtheria+Incidents\_HIV+GDP\_per\_capita+Thinness\_ten\_nineteen\_years+Schooling, data = data)  
summary(adjr\_model1)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + BMI + Diphtheria + Incidents\_HIV +   
## GDP\_per\_capita + Thinness\_ten\_nineteen\_years + Schooling,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3181 -0.9414 -0.0269 0.9250 5.1922   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.403e+01 6.089e-01 138.006 < 2e-16 \*\*\*  
## Infant\_deaths -1.291e-01 2.676e-03 -48.253 < 2e-16 \*\*\*  
## Adult\_mortality -5.017e-02 6.225e-04 -80.585 < 2e-16 \*\*\*  
## Alcohol\_consumption 7.055e-02 9.095e-03 7.756 1.21e-14 \*\*\*  
## Hepatitis\_B -1.331e-02 2.593e-03 -5.133 3.04e-07 \*\*\*  
## BMI -1.324e-01 1.896e-02 -6.982 3.60e-12 \*\*\*  
## Diphtheria 1.297e-02 3.326e-03 3.901 9.80e-05 \*\*\*  
## Incidents\_HIV 9.622e-02 1.879e-02 5.119 3.27e-07 \*\*\*  
## GDP\_per\_capita 2.304e-05 2.130e-06 10.812 < 2e-16 \*\*\*  
## Thinness\_ten\_nineteen\_years -2.461e-02 8.147e-03 -3.020 0.00255 \*\*   
## Schooling 1.485e-01 1.681e-02 8.831 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.413 on 2853 degrees of freedom  
## Multiple R-squared: 0.9775, Adjusted R-squared: 0.9774   
## F-statistic: 1.239e+04 on 10 and 2853 DF, p-value: < 2.2e-16

aic\_bic\_model <- lm(Life\_expectancy ~ Infant\_deaths+Under\_five\_deaths+Adult\_mortality+Alcohol\_consumption+Hepatitis\_B+BMI+Incidents\_HIV+GDP\_per\_capita+Thinness\_ten\_nineteen\_years+Schooling, data = data)  
summary(aic\_bic\_model)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Under\_five\_deaths +   
## Adult\_mortality + Alcohol\_consumption + Hepatitis\_B + BMI +   
## Incidents\_HIV + GDP\_per\_capita + Thinness\_ten\_nineteen\_years +   
## Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.7541 -0.9321 -0.0426 0.8609 8.0225   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.534e+01 5.423e-01 157.367 < 2e-16 \*\*\*  
## Infant\_deaths -5.185e-02 6.154e-03 -8.424 < 2e-16 \*\*\*  
## Under\_five\_deaths -5.359e-02 3.739e-03 -14.332 < 2e-16 \*\*\*  
## Adult\_mortality -4.888e-02 6.096e-04 -80.182 < 2e-16 \*\*\*  
## Alcohol\_consumption 9.046e-02 8.922e-03 10.139 < 2e-16 \*\*\*  
## Hepatitis\_B -6.709e-03 1.906e-03 -3.520 0.000438 \*\*\*  
## BMI -1.677e-01 1.826e-02 -9.182 < 2e-16 \*\*\*  
## Incidents\_HIV 1.020e-01 1.814e-02 5.625 2.04e-08 \*\*\*  
## GDP\_per\_capita 3.067e-05 2.132e-06 14.382 < 2e-16 \*\*\*  
## Thinness\_ten\_nineteen\_years -3.744e-02 7.937e-03 -4.717 2.50e-06 \*\*\*  
## Schooling 1.118e-01 1.649e-02 6.782 1.44e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.369 on 2853 degrees of freedom  
## Multiple R-squared: 0.9789, Adjusted R-squared: 0.9788   
## F-statistic: 1.323e+04 on 10 and 2853 DF, p-value: < 2.2e-16

vif(aic\_bic\_model)

## Infant\_deaths Under\_five\_deaths   
## 43.894505 42.447727   
## Adult\_mortality Alcohol\_consumption   
## 7.499440 1.928949   
## Hepatitis\_B BMI   
## 1.420436 2.453165   
## Incidents\_HIV GDP\_per\_capita   
## 2.851115 1.992647   
## Thinness\_ten\_nineteen\_years Schooling   
## 1.896451 4.180747

aic\_bic\_model1 <- lm(Life\_expectancy ~ Infant\_deaths+Adult\_mortality+Alcohol\_consumption+Hepatitis\_B+BMI+Incidents\_HIV+GDP\_per\_capita+Thinness\_ten\_nineteen\_years+Schooling, data = data)  
summary(aic\_bic\_model1)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + BMI + Incidents\_HIV +   
## GDP\_per\_capita + Thinness\_ten\_nineteen\_years + Schooling,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.2869 -0.9260 -0.0151 0.9119 5.3714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.497e+01 5.608e-01 151.531 < 2e-16 \*\*\*  
## Infant\_deaths -1.331e-01 2.484e-03 -53.581 < 2e-16 \*\*\*  
## Adult\_mortality -5.018e-02 6.241e-04 -80.409 < 2e-16 \*\*\*  
## Alcohol\_consumption 6.991e-02 9.116e-03 7.668 2.37e-14 \*\*\*  
## Hepatitis\_B -6.725e-03 1.973e-03 -3.408 0.000662 \*\*\*  
## BMI -1.428e-01 1.882e-02 -7.589 4.34e-14 \*\*\*  
## Incidents\_HIV 1.023e-01 1.878e-02 5.448 5.53e-08 \*\*\*  
## GDP\_per\_capita 2.294e-05 2.136e-06 10.741 < 2e-16 \*\*\*  
## Thinness\_ten\_nineteen\_years -2.490e-02 8.167e-03 -3.048 0.002321 \*\*   
## Schooling 1.497e-01 1.685e-02 8.884 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.417 on 2854 degrees of freedom  
## Multiple R-squared: 0.9774, Adjusted R-squared: 0.9773   
## F-statistic: 1.37e+04 on 9 and 2854 DF, p-value: < 2.2e-16

vif(aic\_bic\_model1)

## Infant\_deaths Adult\_mortality   
## 6.670553 7.333555   
## Alcohol\_consumption Hepatitis\_B   
## 1.879108 1.420435   
## BMI Incidents\_HIV   
## 2.431035 2.851112   
## GDP\_per\_capita Thinness\_ten\_nineteen\_years   
## 1.865240 1.873373   
## Schooling   
## 4.073410

backward\_model <- lm(Life\_expectancy ~ Thinness\_five\_nine\_years+Polio+Measles+Population\_mln, data = data)  
summary(backward\_model)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Thinness\_five\_nine\_years + Polio +   
## Measles + Population\_mln, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.5921 -3.3228 0.8005 4.4436 17.6780   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.7593747 0.8457433 47.011 <2e-16 \*\*\*  
## Thinness\_five\_nine\_years -0.6028491 0.0302874 -19.904 <2e-16 \*\*\*  
## Polio 0.2990299 0.0095698 31.247 <2e-16 \*\*\*  
## Measles 0.0756454 0.0079028 9.572 <2e-16 \*\*\*  
## Population\_mln 0.0091009 0.0009285 9.802 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.539 on 2859 degrees of freedom  
## Multiple R-squared: 0.5173, Adjusted R-squared: 0.5167   
## F-statistic: 766.1 on 4 and 2859 DF, p-value: < 2.2e-16

The models that the functions from the olsrr library provided were quite similar our model6. It is important to note that these functions do not check for multicollinearity issues.

The model based on the highest adjusted R-squared value had predictors Infant\_deaths, Under\_five\_deaths, Adult\_mortality, Alcohol\_consumption, Hepatitis\_B, BMI, Diphtheria, Incidents\_HIV, GDP\_per\_capita, Thinness\_ten\_nineteen\_years, and Schooling. When looking at the vif’s we can see Infant\_deaths and Under\_five\_deaths are highly correlated, so it makes sense to remove Under\_five\_deaths as it has the higher vif. After doing so we can see that adjr\_model1 is the exact same model as our main model, model6.

The model based on the lowest AIC and BIC values (remember they are the same model) had predictors Infant\_deaths, Under\_five\_deaths, Adult\_mortality, Alcohol\_consumption, Hepatitis\_B, BMI, Incidents\_HIV, GDP\_per\_capita, Thinness\_ten\_nineteen\_years, and Schooling. Notice that this contains the same exact values as adj\_model other than the fact that it omits Diphtheria. After removing Under\_five\_deaths once more due to multicollinearity, we are left with the final model based on the AIC/BIC values, aic\_bic\_model1. Simillarly to our other models, this model’s adjusted R-squared was 0.9773

The model obtained by the forward stepwise method had the same predictors as the adjr\_model, meaning after cleaning it will be the same model as our main model, model6.

The model obtained by the backward stepwise method was the most different out of every model and had predictors Thinness\_five\_nine\_years, Polio, Measles, and Population\_mln. Interestingly enough, these are all variables that were removed from other models. This was the worst fit model, with an adjusted R-squared of 0.5167.

Taking all of this into consideration, the best choice of action is likely to go with aic\_bic\_model1, as the change caused by Diphtheria was rather insignificant.

# Transformations

# fit model taking log(Life\_expectancy)  
  
log\_aic\_bic\_model1 <- lm(log(Life\_expectancy) ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Schooling,  
 data = data  
)  
  
summary(log\_aic\_bic\_model1)

##   
## Call:  
## lm(formula = log(Life\_expectancy) ~ Infant\_deaths + Adult\_mortality +   
## Alcohol\_consumption + Hepatitis\_B + BMI + Incidents\_HIV +   
## GDP\_per\_capita + Thinness\_ten\_nineteen\_years + Schooling,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.07946 -0.01353 0.00153 0.01308 0.10058   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.466e+00 8.305e-03 537.745 < 2e-16 \*\*\*  
## Infant\_deaths -2.229e-03 3.678e-05 -60.606 < 2e-16 \*\*\*  
## Adult\_mortality -7.677e-04 9.243e-06 -83.060 < 2e-16 \*\*\*  
## Alcohol\_consumption 8.278e-04 1.350e-04 6.131 9.95e-10 \*\*\*  
## Hepatitis\_B -8.465e-05 2.922e-05 -2.897 0.00380 \*\*   
## BMI -1.585e-03 2.787e-04 -5.685 1.44e-08 \*\*\*  
## Incidents\_HIV -8.302e-04 2.781e-04 -2.985 0.00286 \*\*   
## GDP\_per\_capita 6.804e-09 3.163e-08 0.215 0.82970   
## Thinness\_ten\_nineteen\_years 8.982e-05 1.210e-04 0.743 0.45775   
## Schooling 1.847e-03 2.496e-04 7.402 1.76e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02099 on 2854 degrees of freedom  
## Multiple R-squared: 0.9797, Adjusted R-squared: 0.9797   
## F-statistic: 1.533e+04 on 9 and 2854 DF, p-value: < 2.2e-16

# WAIT

By including the predictors that relate directly to death we have essentially just found that when less adults/babies die, the life expectancy is longer. WOW!!! This is not exactly the information we wanted to capture, it is a pretty obvious assumption, so it makes sense to try to refit the model with out it. So yes, everything up til now was all in vain.

# New Initial Model Selection

new\_model <- lm(Life\_expectancy ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Polio +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_ten\_nineteen\_years +  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = data  
)  
  
summary(new\_model)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Polio + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + Thinness\_ten\_nineteen\_years + Thinness\_five\_nine\_years +   
## Schooling + Economy\_status\_Developed, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.844 -2.660 0.206 2.978 13.215   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.418e+01 1.206e+00 20.039 < 2e-16 \*\*\*  
## Alcohol\_consumption -9.599e-02 2.727e-02 -3.521 0.000437 \*\*\*  
## Hepatitis\_B -1.744e-02 7.199e-03 -2.423 0.015470 \*   
## Measles 1.678e-02 4.855e-03 3.457 0.000554 \*\*\*  
## BMI 8.963e-01 4.998e-02 17.933 < 2e-16 \*\*\*  
## Polio 1.160e-01 1.631e-02 7.115 1.41e-12 \*\*\*  
## Diphtheria 7.766e-02 1.658e-02 4.684 2.94e-06 \*\*\*  
## Incidents\_HIV -1.512e+00 3.194e-02 -47.349 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.012e-04 6.079e-06 16.639 < 2e-16 \*\*\*  
## Population\_mln 3.760e-03 5.617e-04 6.693 2.62e-11 \*\*\*  
## Thinness\_ten\_nineteen\_years -7.855e-02 4.863e-02 -1.615 0.106406   
## Thinness\_five\_nine\_years 1.725e-01 4.755e-02 3.627 0.000292 \*\*\*  
## Schooling 6.899e-01 4.318e-02 15.977 < 2e-16 \*\*\*  
## Economy\_status\_Developed 2.239e+00 3.003e-01 7.456 1.18e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.862 on 2850 degrees of freedom  
## Multiple R-squared: 0.8322, Adjusted R-squared: 0.8314   
## F-statistic: 1087 on 13 and 2850 DF, p-value: < 2.2e-16

vif(new\_model)

## Alcohol\_consumption Hepatitis\_B   
## 2.262824 2.545358   
## Measles BMI   
## 1.575468 2.308078   
## Polio Diphtheria   
## 11.609133 12.733138   
## Incidents\_HIV GDP\_per\_capita   
## 1.110464 2.034717   
## Population\_mln Thinness\_ten\_nineteen\_years   
## 1.128290 8.944351   
## Thinness\_five\_nine\_years Schooling   
## 8.886990 3.600979   
## Economy\_status\_Developed   
## 2.840275

# find all possible permutations of the new model   
# new\_MS\_ALL = ols\_step\_all\_possible(new\_model)  
# write.csv(new\_MS\_ALL$result, "new\_MS\_ALL\_results.csv", row.names = FALSE)  
new\_MS\_ALL <- read.csv("new\_MS\_ALL\_results.csv")  
  
new\_adjr <- new\_MS\_ALL$adjr  
new\_max\_adjr <- which.max(new\_adjr)  
new\_best\_model\_adjr <- new\_MS\_ALL[new\_max\_adjr, ]  
new\_best\_model\_adjr$predictors

## [1] "Alcohol\_consumption Hepatitis\_B Measles BMI Polio Diphtheria Incidents\_HIV GDP\_per\_capita Population\_mln Thinness\_ten\_nineteen\_years Thinness\_five\_nine\_years Schooling Economy\_status\_Developed"

new\_aic <- new\_MS\_ALL$aic  
new\_min\_aic <- which.min(new\_aic)  
new\_best\_model\_aic <- new\_MS\_ALL[new\_min\_aic, ]  
new\_best\_model\_aic$predictors

## [1] "Alcohol\_consumption Hepatitis\_B Measles BMI Polio Diphtheria Incidents\_HIV GDP\_per\_capita Population\_mln Thinness\_ten\_nineteen\_years Thinness\_five\_nine\_years Schooling Economy\_status\_Developed"

new\_max\_adjr == new\_min\_aic

## [1] TRUE

new\_bic <- new\_MS\_ALL$sbic  
new\_min\_bic <- which.min(new\_bic)  
new\_best\_model\_bic <- new\_MS\_ALL[new\_min\_bic, ]  
new\_best\_model\_bic$predictors

## [1] "Alcohol\_consumption Hepatitis\_B Measles BMI Polio Diphtheria Incidents\_HIV GDP\_per\_capita Population\_mln Thinness\_ten\_nineteen\_years Thinness\_five\_nine\_years Schooling Economy\_status\_Developed"

new\_min\_bic == new\_max\_adjr

## [1] TRUE

# stepwise selection with p-values  
new\_ols\_stepwise\_forward <- ols\_step\_forward\_p(new\_model)  
new\_ols\_stepwise\_forward$metrics$variable

## [1] "Schooling" "Incidents\_HIV"   
## [3] "Polio" "GDP\_per\_capita"   
## [5] "BMI" "Population\_mln"   
## [7] "Economy\_status\_Developed" "Thinness\_five\_nine\_years"   
## [9] "Diphtheria" "Alcohol\_consumption"   
## [11] "Measles" "Hepatitis\_B"   
## [13] "Thinness\_ten\_nineteen\_years"

new\_model1 <- lm(Life\_expectancy ~ Alcohol\_consumption+Hepatitis\_B+Measles+BMI+Polio+Diphtheria+Incidents\_HIV+GDP\_per\_capita+Population\_mln+Thinness\_ten\_nineteen\_years+Thinness\_five\_nine\_years+Schooling+Economy\_status\_Developed, data = data)  
  
summary(new\_model1)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Polio + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + Thinness\_ten\_nineteen\_years + Thinness\_five\_nine\_years +   
## Schooling + Economy\_status\_Developed, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.844 -2.660 0.206 2.978 13.215   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.418e+01 1.206e+00 20.039 < 2e-16 \*\*\*  
## Alcohol\_consumption -9.599e-02 2.727e-02 -3.521 0.000437 \*\*\*  
## Hepatitis\_B -1.744e-02 7.199e-03 -2.423 0.015470 \*   
## Measles 1.678e-02 4.855e-03 3.457 0.000554 \*\*\*  
## BMI 8.963e-01 4.998e-02 17.933 < 2e-16 \*\*\*  
## Polio 1.160e-01 1.631e-02 7.115 1.41e-12 \*\*\*  
## Diphtheria 7.766e-02 1.658e-02 4.684 2.94e-06 \*\*\*  
## Incidents\_HIV -1.512e+00 3.194e-02 -47.349 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.012e-04 6.079e-06 16.639 < 2e-16 \*\*\*  
## Population\_mln 3.760e-03 5.617e-04 6.693 2.62e-11 \*\*\*  
## Thinness\_ten\_nineteen\_years -7.855e-02 4.863e-02 -1.615 0.106406   
## Thinness\_five\_nine\_years 1.725e-01 4.755e-02 3.627 0.000292 \*\*\*  
## Schooling 6.899e-01 4.318e-02 15.977 < 2e-16 \*\*\*  
## Economy\_status\_Developed 2.239e+00 3.003e-01 7.456 1.18e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.862 on 2850 degrees of freedom  
## Multiple R-squared: 0.8322, Adjusted R-squared: 0.8314   
## F-statistic: 1087 on 13 and 2850 DF, p-value: < 2.2e-16

After removing the death variables, the best model, which has the highest adjusted R-squared and lowest AIC and BIC values is the new full model. Now we will begin reducing this model, Thinness\_ten\_nineteen\_years is not statistically significant so we will remove it.

new\_model1 <- lm(Life\_expectancy ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Polio +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln ++  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = data  
)  
  
summary(new\_model1)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Polio + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + +Thinness\_five\_nine\_years + Schooling +   
## Economy\_status\_Developed, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.858 -2.663 0.204 2.967 13.234   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.399e+01 1.202e+00 19.970 < 2e-16 \*\*\*  
## Alcohol\_consumption -9.344e-02 2.723e-02 -3.432 0.000608 \*\*\*  
## Hepatitis\_B -1.791e-02 7.195e-03 -2.489 0.012883 \*   
## Measles 1.605e-02 4.835e-03 3.320 0.000912 \*\*\*  
## BMI 9.010e-01 4.990e-02 18.055 < 2e-16 \*\*\*  
## Polio 1.152e-01 1.630e-02 7.066 2.00e-12 \*\*\*  
## Diphtheria 7.893e-02 1.656e-02 4.765 1.98e-06 \*\*\*  
## Incidents\_HIV -1.513e+00 3.194e-02 -47.376 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.008e-04 6.077e-06 16.586 < 2e-16 \*\*\*  
## Population\_mln 3.691e-03 5.602e-04 6.588 5.28e-11 \*\*\*  
## Thinness\_five\_nine\_years 1.047e-01 2.240e-02 4.674 3.09e-06 \*\*\*  
## Schooling 6.983e-01 4.288e-02 16.283 < 2e-16 \*\*\*  
## Economy\_status\_Developed 2.236e+00 3.004e-01 7.444 1.29e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.863 on 2851 degrees of freedom  
## Multiple R-squared: 0.832, Adjusted R-squared: 0.8313   
## F-statistic: 1177 on 12 and 2851 DF, p-value: < 2.2e-16

vif(new\_model1)

## Alcohol\_consumption Hepatitis\_B Measles   
## 2.255223 2.541284 1.561661   
## BMI Polio Diphtheria   
## 2.299923 11.597821 12.704416   
## Incidents\_HIV GDP\_per\_capita Population\_mln   
## 1.110047 2.031923 1.121850   
## Thinness\_five\_nine\_years Schooling Economy\_status\_Developed   
## 1.972073 3.549273 2.840155

Now, every explanatory variable is statistically significant, however, there is a multicollinearity issue between Polio and Diphtheria, so we will test models by removing each of them and see which one has a higher adjusted R-squared.

new\_model2 <- lm(Life\_expectancy ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = data  
)  
  
summary(new\_model2)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + Thinness\_five\_nine\_years + Schooling + Economy\_status\_Developed,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.6512 -2.7436 0.1714 2.9721 13.4598   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.393e+01 1.212e+00 19.746 < 2e-16 \*\*\*  
## Alcohol\_consumption -9.284e-02 2.746e-02 -3.381 0.000732 \*\*\*  
## Hepatitis\_B -1.917e-02 7.254e-03 -2.643 0.008255 \*\*   
## Measles 1.997e-02 4.844e-03 4.124 3.84e-05 \*\*\*  
## BMI 9.353e-01 5.009e-02 18.672 < 2e-16 \*\*\*  
## Diphtheria 1.809e-01 8.212e-03 22.022 < 2e-16 \*\*\*  
## Incidents\_HIV -1.513e+00 3.221e-02 -46.983 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.004e-04 6.128e-06 16.385 < 2e-16 \*\*\*  
## Population\_mln 3.669e-03 5.650e-04 6.494 9.85e-11 \*\*\*  
## Thinness\_five\_nine\_years 1.109e-01 2.258e-02 4.910 9.62e-07 \*\*\*  
## Schooling 7.211e-01 4.313e-02 16.720 < 2e-16 \*\*\*  
## Economy\_status\_Developed 2.154e+00 3.027e-01 7.114 1.42e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.896 on 2852 degrees of freedom  
## Multiple R-squared: 0.8291, Adjusted R-squared: 0.8284   
## F-statistic: 1258 on 11 and 2852 DF, p-value: < 2.2e-16

vif(new\_model2)

## Alcohol\_consumption Hepatitis\_B Measles   
## 2.255201 2.539698 1.541058   
## BMI Diphtheria Incidents\_HIV   
## 2.278211 3.070002 1.110045   
## GDP\_per\_capita Population\_mln Thinness\_five\_nine\_years   
## 2.031765 1.121815 1.969100   
## Schooling Economy\_status\_Developed   
## 3.529140 2.835865

new\_model3 <- lm(Life\_expectancy ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Polio +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln ++  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = data  
)  
  
summary(new\_model3)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Polio + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + +Thinness\_five\_nine\_years + Schooling +   
## Economy\_status\_Developed, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.3486 -2.6855 0.1679 2.9332 13.1950   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.429e+01 1.204e+00 20.165 < 2e-16 \*\*\*  
## Alcohol\_consumption -9.503e-02 2.733e-02 -3.477 0.000514 \*\*\*  
## Hepatitis\_B -5.415e-03 6.726e-03 -0.805 0.420811   
## Measles 1.516e-02 4.850e-03 3.125 0.001795 \*\*   
## BMI 8.851e-01 4.998e-02 17.709 < 2e-16 \*\*\*  
## Polio 1.828e-01 8.045e-03 22.728 < 2e-16 \*\*\*  
## Incidents\_HIV -1.516e+00 3.206e-02 -47.303 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.014e-04 6.099e-06 16.628 < 2e-16 \*\*\*  
## Population\_mln 3.785e-03 5.620e-04 6.734 1.99e-11 \*\*\*  
## Thinness\_five\_nine\_years 1.037e-01 2.249e-02 4.609 4.22e-06 \*\*\*  
## Schooling 7.068e-01 4.301e-02 16.434 < 2e-16 \*\*\*  
## Economy\_status\_Developed 2.322e+00 3.010e-01 7.715 1.66e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.878 on 2852 degrees of freedom  
## Multiple R-squared: 0.8307, Adjusted R-squared: 0.83   
## F-statistic: 1272 on 11 and 2852 DF, p-value: < 2.2e-16

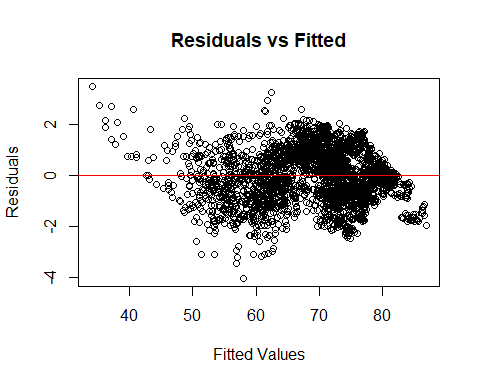
vif(new\_model3)

## Alcohol\_consumption Hepatitis\_B Measles   
## 2.254884 2.203988 1.559303   
## BMI Polio Incidents\_HIV   
## 2.289635 2.802595 1.109582   
## GDP\_per\_capita Population\_mln Thinness\_five\_nine\_years   
## 2.031008 1.120469 1.971879   
## Schooling Economy\_status\_Developed   
## 3.543072 2.829861

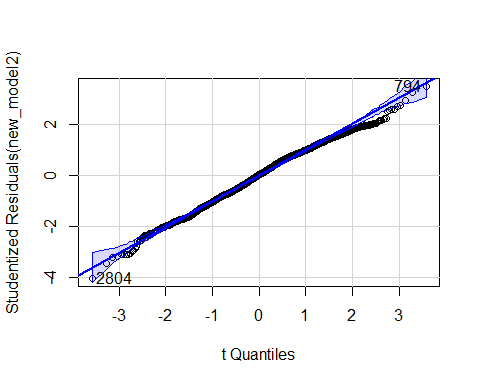
When dropping Diphtheria, Hepatitis-B becomes in significant, so we will drop Polio instead, so we will go with new\_model2 as our reduced model. It has an adjusted R-squared 0f 0.83 which is much more preferable to the previous 0.97, it indicated that these values have a strong relation to Life\_expectancy without over fitting the model.

# New Residual Analysis

# residuals vs fitted values  
plot(fitted(new\_model2), studres(new\_model2),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

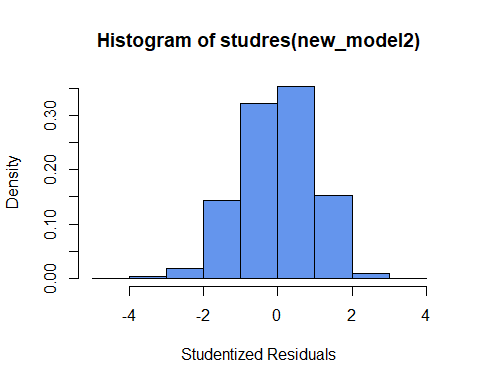


# QQ plot  
qqPlot(new\_model2)

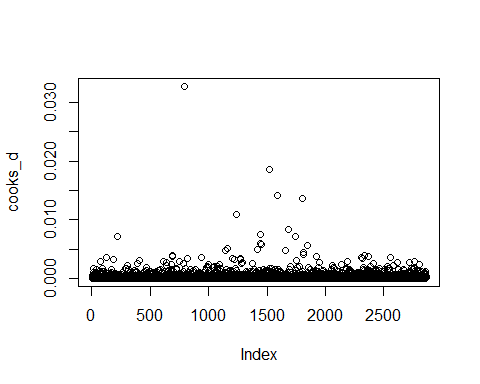


## [1] 794 2804

# histogram of residuals  
hist(studres(new\_model2), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



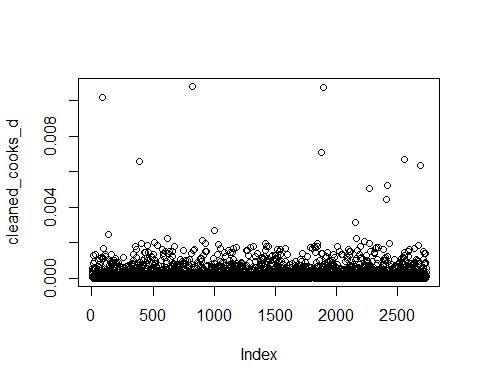
# cook's distance  
  
cooks\_d <- cooks.distance(new\_model2)  
plot(cooks\_d)



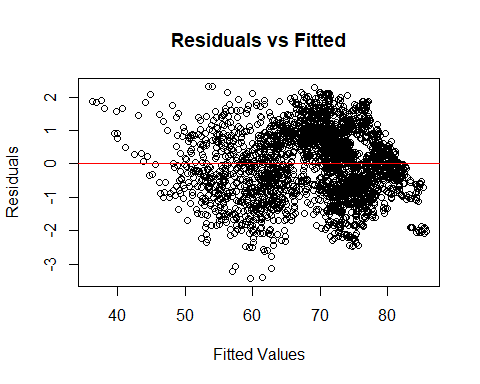
# remove influential points flagged by cook's distance  
  
n <- nrow(data)  
  
influential\_point <- which(cooks\_d > (4 / n))  
  
cleaned\_data <- data[-influential\_point,]  
  
cleaned\_new\_model2 <- lm(Life\_expectancy ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln ++  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = cleaned\_data  
)  
  
summary(cleaned\_new\_model2)

##   
## Call:  
## lm(formula = Life\_expectancy ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + +Thinness\_five\_nine\_years + Schooling +   
## Economy\_status\_Developed, data = cleaned\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.210 -2.614 0.057 2.821 8.274   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.446e+01 1.169e+00 20.927 < 2e-16 \*\*\*  
## Alcohol\_consumption -2.842e-02 2.682e-02 -1.059 0.289511   
## Hepatitis\_B -2.598e-02 6.982e-03 -3.721 0.000203 \*\*\*  
## Measles 1.827e-02 4.631e-03 3.945 8.19e-05 \*\*\*  
## BMI 9.450e-01 4.888e-02 19.332 < 2e-16 \*\*\*  
## Diphtheria 1.855e-01 8.030e-03 23.101 < 2e-16 \*\*\*  
## Incidents\_HIV -1.719e+00 3.651e-02 -47.075 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.170e-04 6.405e-06 18.264 < 2e-16 \*\*\*  
## Population\_mln 4.121e-03 5.461e-04 7.547 6.04e-14 \*\*\*  
## Thinness\_five\_nine\_years 9.544e-02 2.165e-02 4.408 1.08e-05 \*\*\*  
## Schooling 6.494e-01 4.116e-02 15.778 < 2e-16 \*\*\*  
## Economy\_status\_Developed 1.506e+00 2.886e-01 5.216 1.96e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.579 on 2722 degrees of freedom  
## Multiple R-squared: 0.8472, Adjusted R-squared: 0.8466   
## F-statistic: 1372 on 11 and 2722 DF, p-value: < 2.2e-16

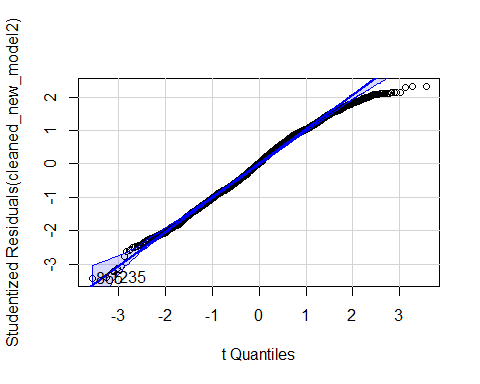
cleaned\_cooks\_d <- cooks.distance(cleaned\_new\_model2)  
plot(cleaned\_cooks\_d)



library(car)  
library(MASS)  
  
# residuals vs fitted values  
plot(fitted(cleaned\_new\_model2), studres(cleaned\_new\_model2),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

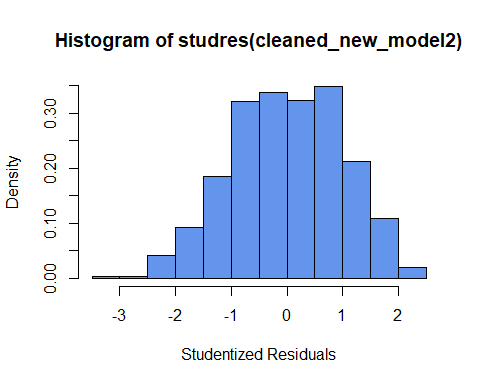


# QQ plot  
qqPlot(cleaned\_new\_model2)



## 865 1235   
## 830 1182

# histogram of residuals  
hist(studres(cleaned\_new\_model2), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")

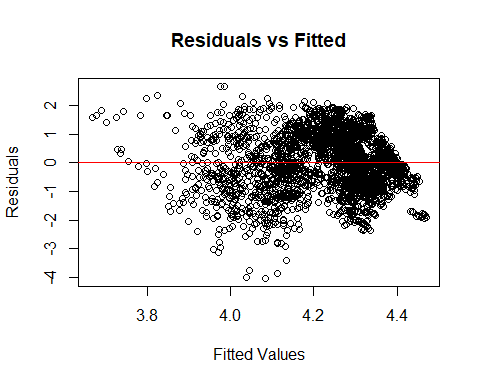


# New Transformations

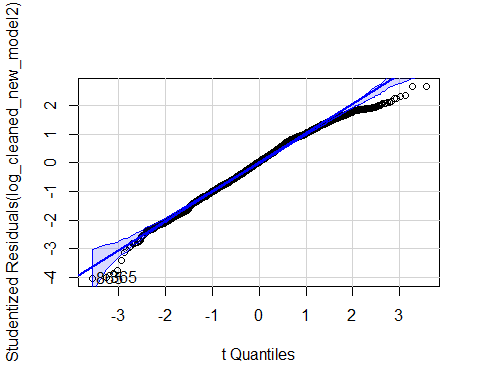
# log transformation  
  
log\_cleaned\_new\_model2 <- lm(log(Life\_expectancy) ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln ++  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = cleaned\_data  
)  
  
summary(log\_cleaned\_new\_model2)

##   
## Call:  
## lm(formula = log(Life\_expectancy) ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + +Thinness\_five\_nine\_years + Schooling +   
## Economy\_status\_Developed, data = cleaned\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.22248 -0.03907 0.00188 0.04390 0.14723   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.512e+00 1.806e-02 194.460 < 2e-16 \*\*\*  
## Alcohol\_consumption -1.830e-04 4.146e-04 -0.441 0.658951   
## Hepatitis\_B -4.203e-04 1.079e-04 -3.895 0.000101 \*\*\*  
## Measles 2.053e-04 7.158e-05 2.868 0.004165 \*\*   
## BMI 1.538e-02 7.555e-04 20.359 < 2e-16 \*\*\*  
## Diphtheria 3.028e-03 1.241e-04 24.396 < 2e-16 \*\*\*  
## Incidents\_HIV -2.951e-02 5.642e-04 -52.302 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.518e-06 9.899e-08 15.337 < 2e-16 \*\*\*  
## Population\_mln 6.270e-05 8.439e-06 7.430 1.45e-13 \*\*\*  
## Thinness\_five\_nine\_years 1.981e-03 3.346e-04 5.922 3.58e-09 \*\*\*  
## Schooling 1.025e-02 6.361e-04 16.112 < 2e-16 \*\*\*  
## Economy\_status\_Developed 1.484e-02 4.461e-03 3.327 0.000891 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05531 on 2722 degrees of freedom  
## Multiple R-squared: 0.8485, Adjusted R-squared: 0.8479   
## F-statistic: 1386 on 11 and 2722 DF, p-value: < 2.2e-16

# residuals vs fitted values  
plot(fitted(log\_cleaned\_new\_model2), studres(log\_cleaned\_new\_model2),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

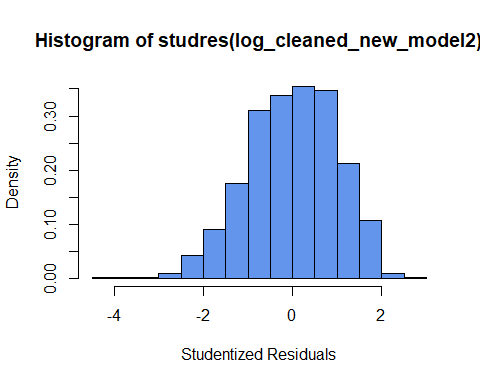


# QQ plot  
qqPlot(log\_cleaned\_new\_model2)



## 365 865   
## 353 830

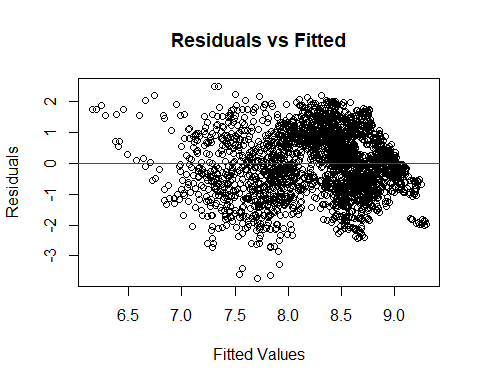
# histogram of residuals  
hist(studres(log\_cleaned\_new\_model2), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



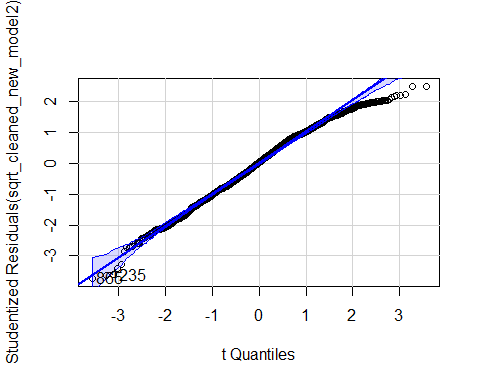
# square root transformation  
  
sqrt\_cleaned\_new\_model2 <- lm(sqrt(Life\_expectancy) ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln ++  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = cleaned\_data  
)  
  
summary(log\_cleaned\_new\_model2)

##   
## Call:  
## lm(formula = log(Life\_expectancy) ~ Alcohol\_consumption + Hepatitis\_B +   
## Measles + BMI + Diphtheria + Incidents\_HIV + GDP\_per\_capita +   
## Population\_mln + +Thinness\_five\_nine\_years + Schooling +   
## Economy\_status\_Developed, data = cleaned\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.22248 -0.03907 0.00188 0.04390 0.14723   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.512e+00 1.806e-02 194.460 < 2e-16 \*\*\*  
## Alcohol\_consumption -1.830e-04 4.146e-04 -0.441 0.658951   
## Hepatitis\_B -4.203e-04 1.079e-04 -3.895 0.000101 \*\*\*  
## Measles 2.053e-04 7.158e-05 2.868 0.004165 \*\*   
## BMI 1.538e-02 7.555e-04 20.359 < 2e-16 \*\*\*  
## Diphtheria 3.028e-03 1.241e-04 24.396 < 2e-16 \*\*\*  
## Incidents\_HIV -2.951e-02 5.642e-04 -52.302 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.518e-06 9.899e-08 15.337 < 2e-16 \*\*\*  
## Population\_mln 6.270e-05 8.439e-06 7.430 1.45e-13 \*\*\*  
## Thinness\_five\_nine\_years 1.981e-03 3.346e-04 5.922 3.58e-09 \*\*\*  
## Schooling 1.025e-02 6.361e-04 16.112 < 2e-16 \*\*\*  
## Economy\_status\_Developed 1.484e-02 4.461e-03 3.327 0.000891 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05531 on 2722 degrees of freedom  
## Multiple R-squared: 0.8485, Adjusted R-squared: 0.8479   
## F-statistic: 1386 on 11 and 2722 DF, p-value: < 2.2e-16

# residuals vs fitted values  
plot(fitted(sqrt\_cleaned\_new\_model2), studres(sqrt\_cleaned\_new\_model2),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

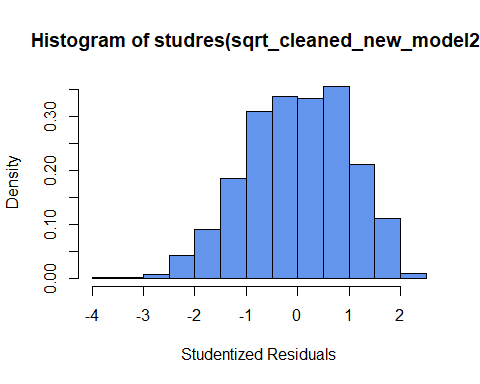


# QQ plot  
qqPlot(sqrt\_cleaned\_new\_model2)

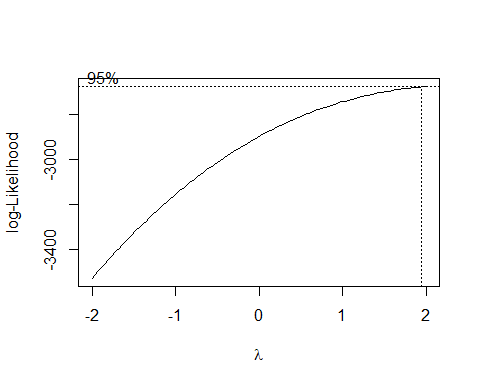


## 865 1235   
## 830 1182

# histogram of residuals  
hist(studres(sqrt\_cleaned\_new\_model2), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



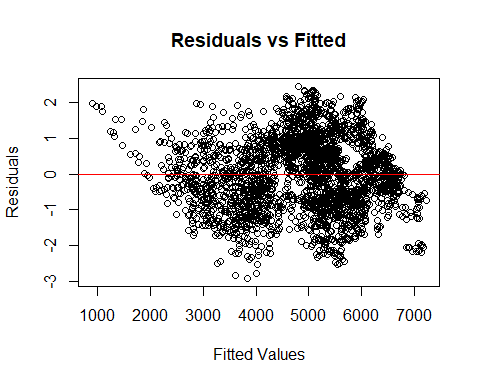
# box-cox transformation  
  
boxcox\_result <- boxcox(cleaned\_new\_model2)



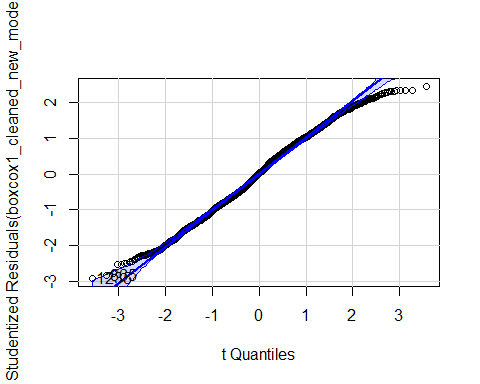
lambda <- boxcox\_result$x[which.max(boxcox\_result$y)]  
  
boxcox1\_cleaned\_new\_model2 <- lm((Life\_expectancy ^ lambda) ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln ++  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = cleaned\_data  
)  
  
summary(boxcox1\_cleaned\_new\_model2)

##   
## Call:  
## lm(formula = (Life\_expectancy^lambda) ~ Alcohol\_consumption +   
## Hepatitis\_B + Measles + BMI + Diphtheria + Incidents\_HIV +   
## GDP\_per\_capita + Population\_mln + +Thinness\_five\_nine\_years +   
## Schooling + Economy\_status\_Developed, data = cleaned\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1393.07 -348.71 4.71 357.31 1171.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.667e+02 1.561e+02 -4.912 9.54e-07 \*\*\*  
## Alcohol\_consumption -5.781e+00 3.582e+00 -1.614 0.106687   
## Hepatitis\_B -3.231e+00 9.324e-01 -3.465 0.000538 \*\*\*  
## Measles 2.942e+00 6.185e-01 4.757 2.06e-06 \*\*\*  
## BMI 1.164e+02 6.528e+00 17.831 < 2e-16 \*\*\*  
## Diphtheria 2.298e+01 1.072e+00 21.432 < 2e-16 \*\*\*  
## Incidents\_HIV -2.034e+02 4.876e+00 -41.726 < 2e-16 \*\*\*  
## GDP\_per\_capita 1.797e-02 8.553e-04 21.007 < 2e-16 \*\*\*  
## Population\_mln 5.433e-01 7.292e-02 7.450 1.25e-13 \*\*\*  
## Thinness\_five\_nine\_years 8.390e+00 2.891e+00 2.902 0.003739 \*\*   
## Schooling 8.322e+01 5.497e+00 15.139 < 2e-16 \*\*\*  
## Economy\_status\_Developed 2.712e+02 3.855e+01 7.034 2.53e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 477.9 on 2722 degrees of freedom  
## Multiple R-squared: 0.8429, Adjusted R-squared: 0.8423   
## F-statistic: 1328 on 11 and 2722 DF, p-value: < 2.2e-16

# residuals vs fitted values  
plot(fitted(boxcox1\_cleaned\_new\_model2), studres(boxcox1\_cleaned\_new\_model2),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

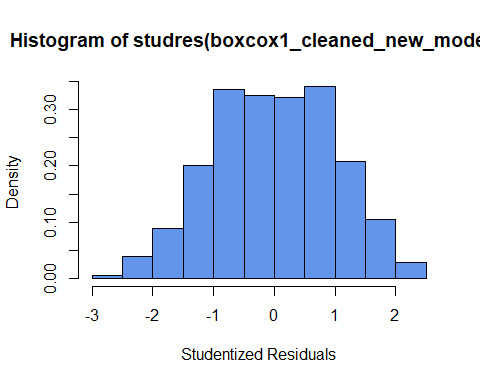


# QQ plot  
qqPlot(boxcox1\_cleaned\_new\_model2)



## 865 1235   
## 830 1182

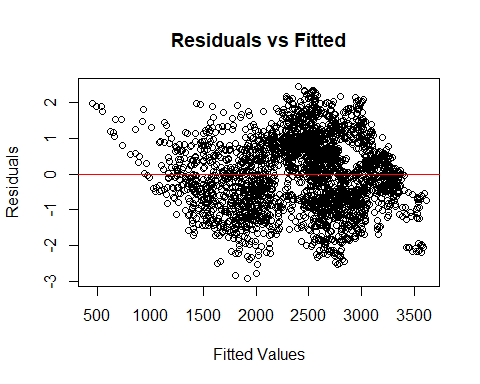
# histogram of residuals  
hist(studres(boxcox1\_cleaned\_new\_model2), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



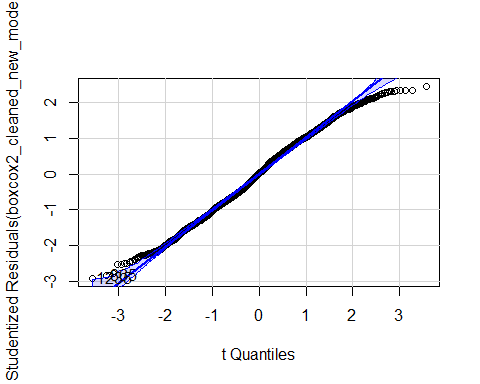
boxcox2\_cleaned\_new\_model2 <- lm((((Life\_expectancy ^ lambda) - 1) / 2) ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln ++  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = cleaned\_data  
)  
  
summary(boxcox2\_cleaned\_new\_model2)

##   
## Call:  
## lm(formula = (((Life\_expectancy^lambda) - 1)/2) ~ Alcohol\_consumption +   
## Hepatitis\_B + Measles + BMI + Diphtheria + Incidents\_HIV +   
## GDP\_per\_capita + Population\_mln + +Thinness\_five\_nine\_years +   
## Schooling + Economy\_status\_Developed, data = cleaned\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -696.54 -174.35 2.35 178.65 585.50   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.838e+02 7.804e+01 -4.919 9.24e-07 \*\*\*  
## Alcohol\_consumption -2.891e+00 1.791e+00 -1.614 0.106687   
## Hepatitis\_B -1.616e+00 4.662e-01 -3.465 0.000538 \*\*\*  
## Measles 1.471e+00 3.092e-01 4.757 2.06e-06 \*\*\*  
## BMI 5.820e+01 3.264e+00 17.831 < 2e-16 \*\*\*  
## Diphtheria 1.149e+01 5.362e-01 21.432 < 2e-16 \*\*\*  
## Incidents\_HIV -1.017e+02 2.438e+00 -41.726 < 2e-16 \*\*\*  
## GDP\_per\_capita 8.984e-03 4.277e-04 21.007 < 2e-16 \*\*\*  
## Population\_mln 2.716e-01 3.646e-02 7.450 1.25e-13 \*\*\*  
## Thinness\_five\_nine\_years 4.195e+00 1.446e+00 2.902 0.003739 \*\*   
## Schooling 4.161e+01 2.748e+00 15.139 < 2e-16 \*\*\*  
## Economy\_status\_Developed 1.356e+02 1.927e+01 7.034 2.53e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 239 on 2722 degrees of freedom  
## Multiple R-squared: 0.8429, Adjusted R-squared: 0.8423   
## F-statistic: 1328 on 11 and 2722 DF, p-value: < 2.2e-16

# residuals vs fitted values  
plot(fitted(boxcox2\_cleaned\_new\_model2), studres(boxcox2\_cleaned\_new\_model2),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

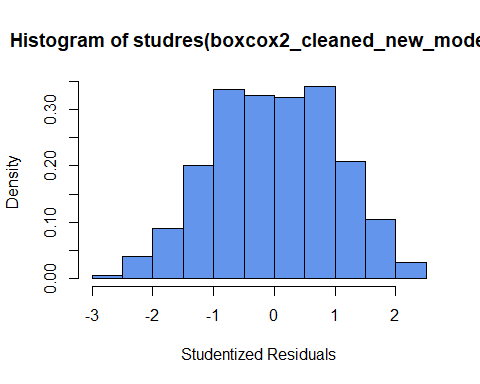


# QQ plot  
qqPlot(boxcox2\_cleaned\_new\_model2)



## 865 1235   
## 830 1182

# histogram of residuals  
hist(studres(boxcox2\_cleaned\_new\_model2), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



Now that we have transformed variables, we will compare their AIC values to determine which is best.

AIC(cleaned\_new\_model2, log\_cleaned\_new\_model2, sqrt\_cleaned\_new\_model2, boxcox1\_cleaned\_new\_model2, boxcox2\_cleaned\_new\_model2)

## df AIC  
## cleaned\_new\_model2 13 14744.3901  
## log\_cleaned\_new\_model2 13 -8056.1950  
## sqrt\_cleaned\_new\_model2 13 -469.7285  
## boxcox1\_cleaned\_new\_model2 13 41507.2209  
## boxcox2\_cleaned\_new\_model2 13 37717.0921

BIC(cleaned\_new\_model2, log\_cleaned\_new\_model2, sqrt\_cleaned\_new\_model2, boxcox1\_cleaned\_new\_model2, boxcox2\_cleaned\_new\_model2)

## df BIC  
## cleaned\_new\_model2 13 14821.2659  
## log\_cleaned\_new\_model2 13 -7979.3192  
## sqrt\_cleaned\_new\_model2 13 -392.8527  
## boxcox1\_cleaned\_new\_model2 13 41584.0967  
## boxcox2\_cleaned\_new\_model2 13 37793.9679

The log transformation has the lowest AIC and BIC values, indicating that it is the best model out of the original and these transformations. The tails of the QQ plot are not as good as the original mode but we think this is a valid trade off for the better AIC/BIC values.

# Final look at models

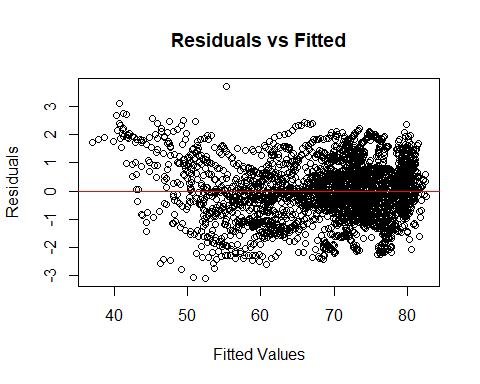
log\_model6 <- lm(log(Life\_expectancy) ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Schooling,  
 data = data  
)  
  
sqrt\_model6 <- lm(sqrt(Life\_expectancy) ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Schooling,  
 data = data  
)  
  
squared\_model6 <- lm((Life\_expectancy) ^ 2 ~   
 Infant\_deaths +  
 Adult\_mortality +  
 Alcohol\_consumption +  
 Hepatitis\_B +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Thinness\_ten\_nineteen\_years +  
 Schooling,  
 data = data  
)  
  
log\_new\_model2 <- lm(log(Life\_expectancy) ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = data  
)  
  
sqrt\_new\_model2 <- lm(sqrt(Life\_expectancy) ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = data  
)  
  
squared\_new\_model2 <- lm((Life\_expectancy) ^ 2 ~   
 Alcohol\_consumption +  
 Hepatitis\_B +  
 Measles +  
 BMI +  
 Diphtheria +  
 Incidents\_HIV +  
 GDP\_per\_capita +  
 Population\_mln +  
 Thinness\_five\_nine\_years +  
 Schooling +  
 Economy\_status\_Developed,  
 data = data  
)  
  
AIC(model6, log\_model6, sqrt\_model6, squared\_model6, new\_model2, log\_new\_model2, sqrt\_new\_model2, squared\_new\_model2)

## df AIC  
## model6 12 10122.60193  
## log\_model6 12 -14023.20345  
## sqrt\_model6 12 -6056.54814  
## squared\_model6 12 38965.56700  
## new\_model2 13 15931.27545  
## log\_new\_model2 13 -7863.11069  
## sqrt\_new\_model2 13 36.03168  
## squared\_new\_model2 13 43897.50726

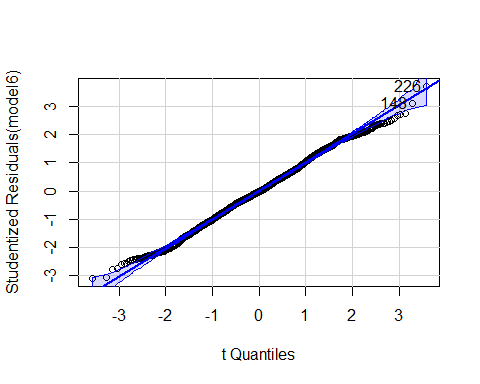
BIC(model6, log\_model6, sqrt\_model6, squared\_model6, new\_model2, log\_new\_model2, sqrt\_new\_model2, squared\_new\_model2)

## df BIC  
## model6 12 10194.1216  
## log\_model6 12 -13951.6838  
## sqrt\_model6 12 -5985.0284  
## squared\_model6 12 39037.0867  
## new\_model2 13 16008.7551  
## log\_new\_model2 13 -7785.6310  
## sqrt\_new\_model2 13 113.5114  
## squared\_new\_model2 13 43974.9869

# residuals vs fitted values  
plot(fitted(model6), studres(model6),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

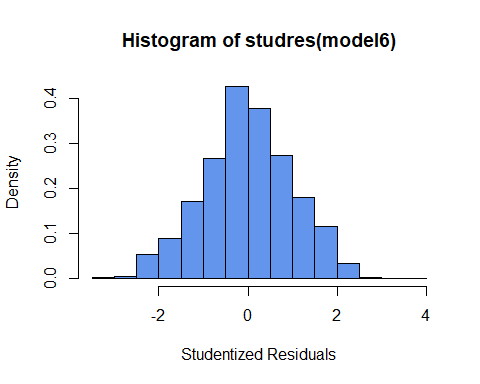


# QQ plot  
qqPlot(model6)

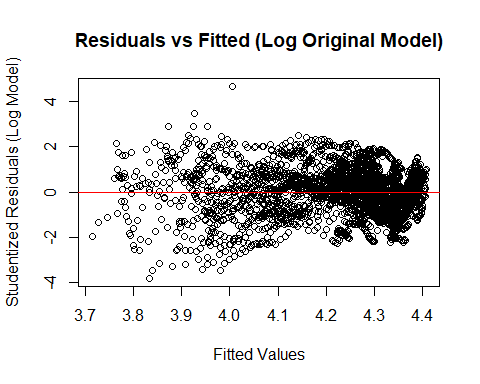


## [1] 148 226

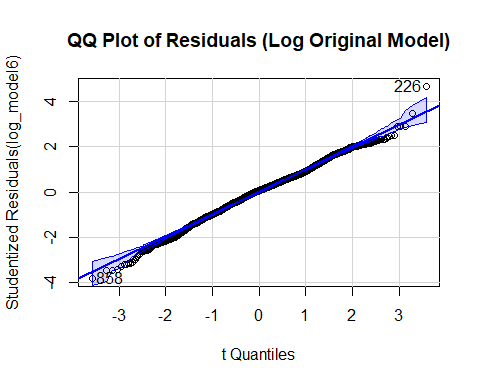
# histogram of residuals  
hist(studres(model6), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



# Residuals vs Fitted Values (Log Original Model)  
plot(fitted(log\_model6), studres(log\_model6),  
 main = "Residuals vs Fitted (Log Original Model)",  
 xlab = "Fitted Values", ylab = "Studentized Residuals (Log Model)")  
abline(h = 0, col = "red")

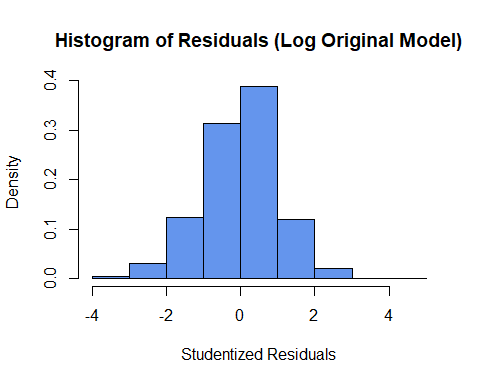


# QQ Plot (Log Original Model)  
qqPlot(log\_model6,   
 main = "QQ Plot of Residuals (Log Original Model)")

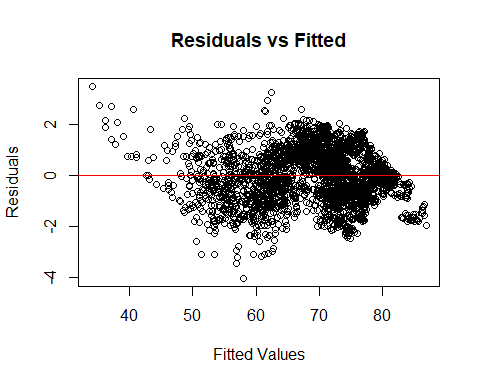


## [1] 226 858

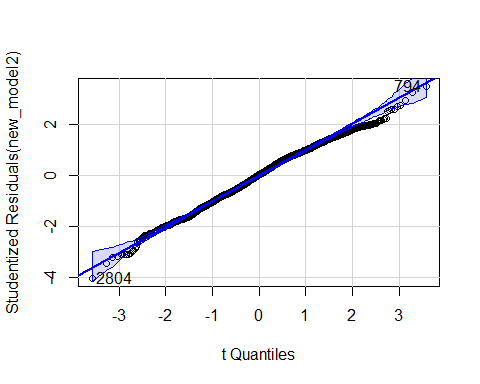
# Histogram of Residuals (Log Original Model)  
hist(studres(log\_model6), breaks = 10, freq = FALSE, col = "cornflowerblue",  
 xlab = "Studentized Residuals",   
 main = "Histogram of Residuals (Log Original Model)")



# residuals vs fitted values  
plot(fitted(new\_model2), studres(new\_model2),  
 main="Residuals vs Fitted",  
 xlab="Fitted Values", ylab="Residuals")  
abline(h = 0, col = "red")

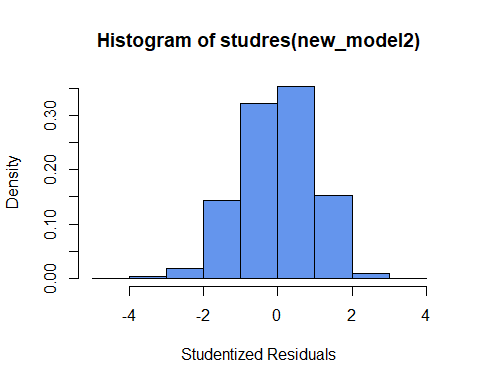


# QQ plot  
qqPlot(new\_model2)

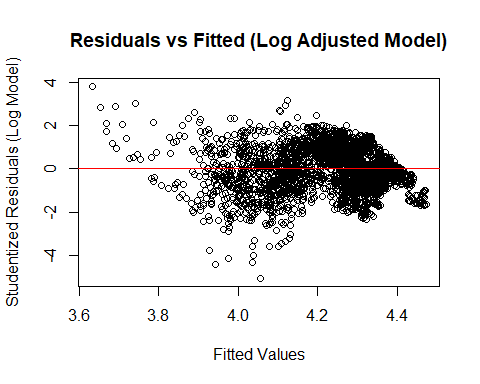


## [1] 794 2804

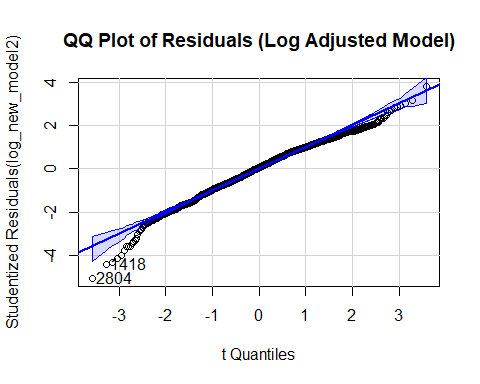
# histogram of residuals  
hist(studres(new\_model2), breaks = 10, freq = F, col="cornflowerblue",  
 xlab = "Studentized Residuals")



# Residuals vs Fitted Values Plot  
plot(fitted(log\_new\_model2), studres(log\_new\_model2),  
 main = "Residuals vs Fitted (Log Adjusted Model)",  
 xlab = "Fitted Values", ylab = "Studentized Residuals (Log Model)")  
abline(h = 0, col = "red")



# QQ Plot  
qqPlot(log\_new\_model2,   
 main = "QQ Plot of Residuals (Log Adjusted Model)")



## [1] 1418 2804

# Histogram of Residuals  
hist(studres(log\_new\_model2), breaks = 10, freq = FALSE, col = "cornflowerblue",  
 xlab = "Studentized Residuals",   
 main = "Histogram of Residuals (Log Adjusted Model)")

