DATA ANALYSIS METHOD-FINAL PROJECT REPORT

Regression on life expectancy data

Group: Karm veer, Anirudh Sharma , Prashant Sahu

Data source: https://www.kaggle.com/kumarajarshi/life-expectancy-who

Intra-group evaluation:

Sharma, Anirudh 30% (evaluator)

Veer, Karm 35%

Sahu, Prashant 35%

Summary:

We chose a dataset for life expectancy prediction as our model for this project we first cleaned the data by removing null values and correcting some obvious errors. We then did data exploration and visualization to get a better feel of the data and spot some relations. We consequently prepared the best model for our data using forward selection method. We checked the summary of the model and then changed the spotted the strong relationships and the positive and negative relations. We then did the model adequacy check and did any of the necessary changes required using transformation we did spot the outliers in the same sections. Then we did the model residual analysis and did the log transformation for two of the covariates. Later we checked for multicollinearity and did ridge regression for correction at last we did model prediction and model validation.

**1. Data Exploration &amp; Data Cleaning –**

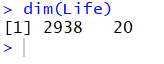
**1.1. Loading &amp; Basic Data Exploration —**

**Goal** :Here we will load the data into R from the csv file we have downloaded and we will also check if there are null values in the data set.

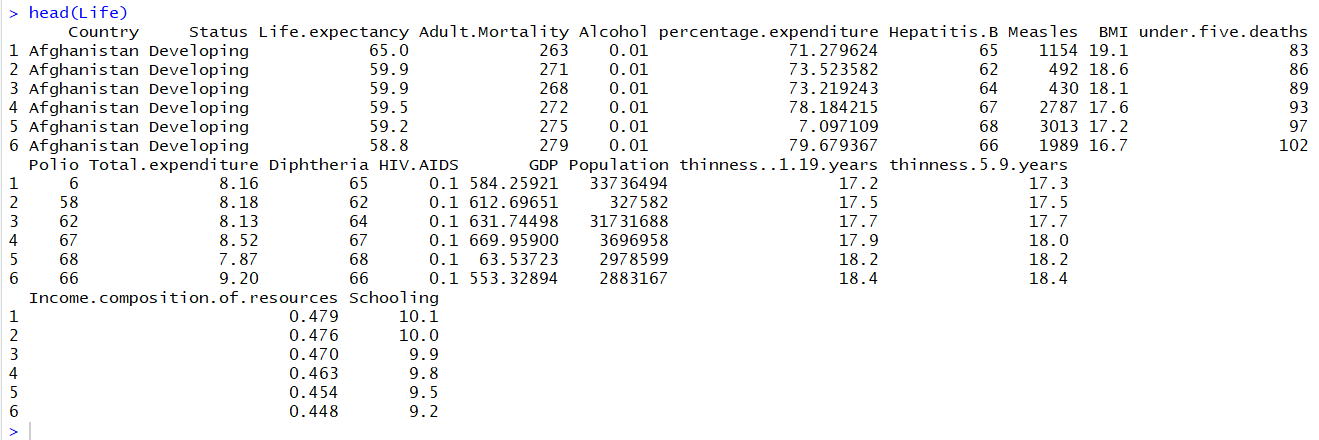
**Rcode and Output:**

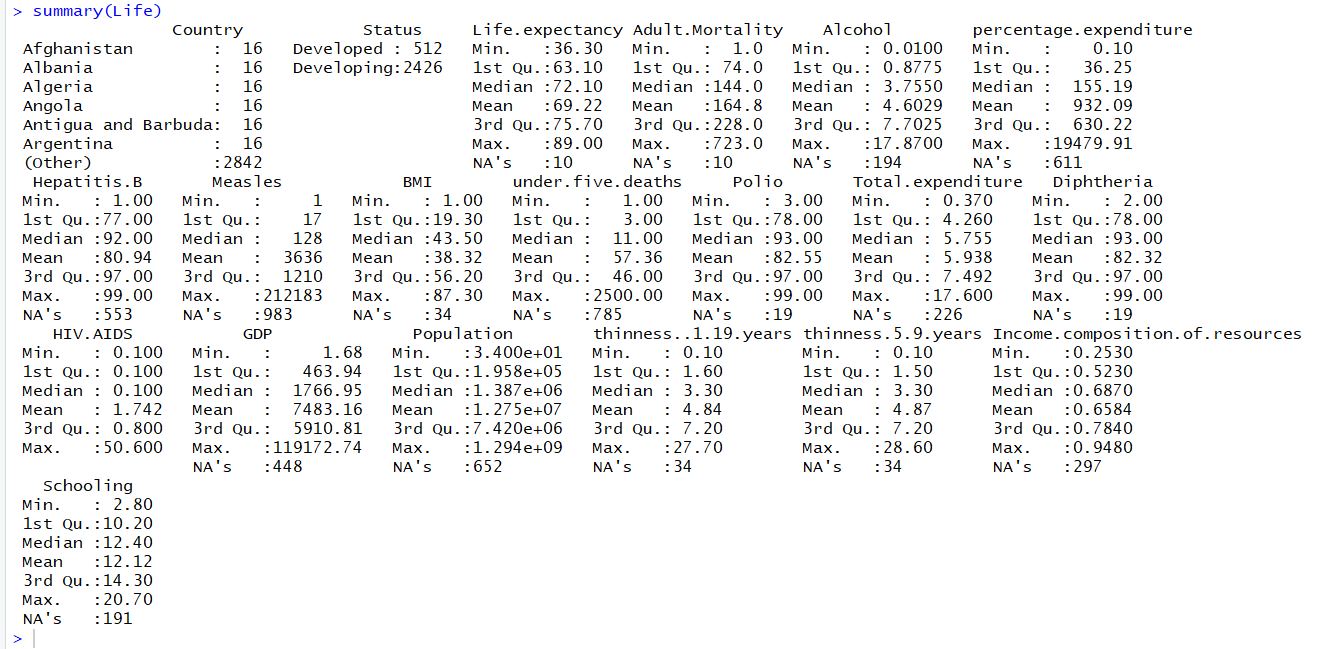
Loading the data into the system-





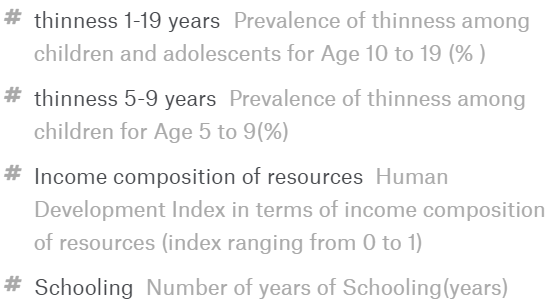
Checking the





**Observation:**

We have 2938 rows and 20 columns in our dataset. We further investigated on the website to get more details on the about the data set here is the details of the each columns and what it meant

**2. Data cleansing :**

* + We Removed null values from our data set as it had a lot of null values
  + Our data set had a lot of zero values instead of the missing of NA we were not able to remove those values using R so we took only those cells with values not equal to zero in our dataset
  + We Took the mean values to replace the null values for some of the columns with critical data (avoiding deleting large number of rows)
  + We did change dew of the obvious values manually from our data set since some of the countries had a dubious data so we turned them null and then took mean for some of the values

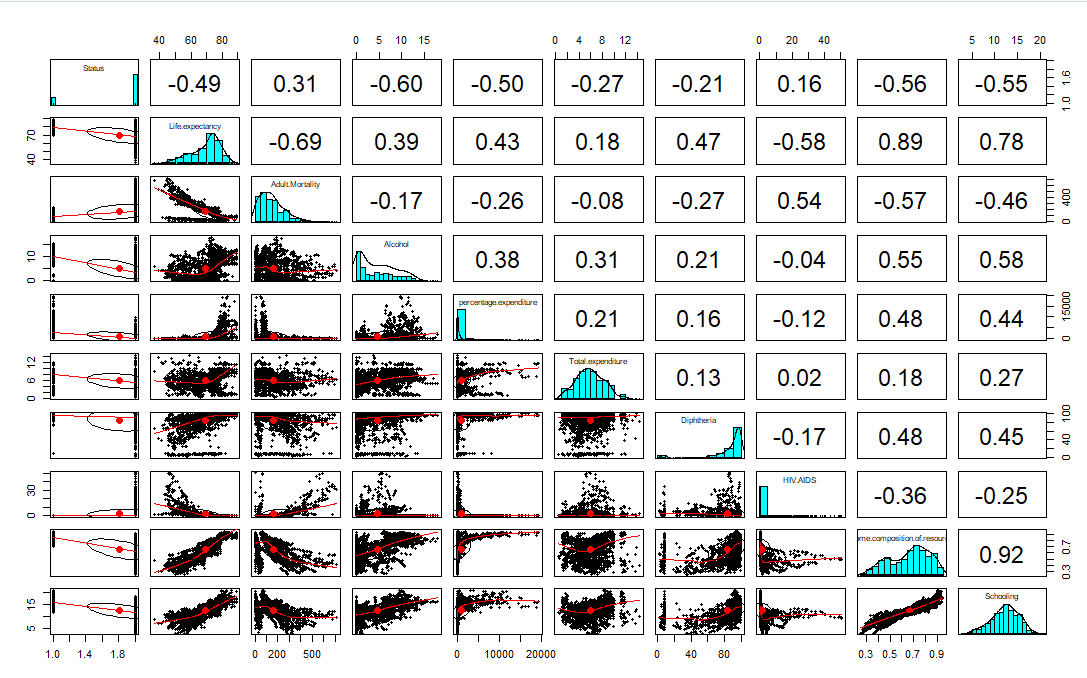


**Data Visualization :**

We are using pair.panels() function of psych library to plot all the numerical variables against each other.







**Observation**:

Here we can see that there is a strong positive linear relationship of Life expectancy with Schooling and Income composition of resources.

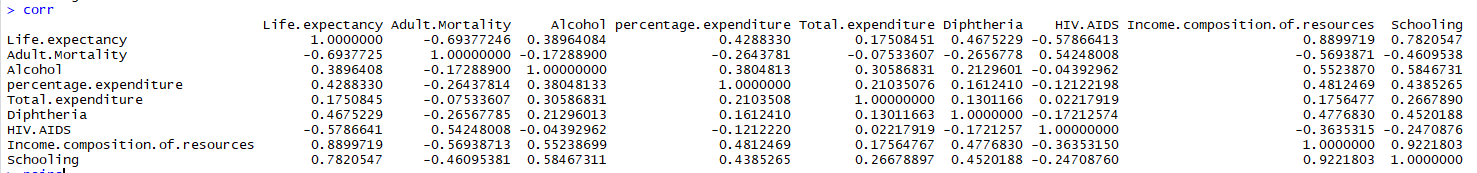
Also we can see that Life expectancy has a negative linear relationship with Adult mortality and HIV AIDS.

**Correlation**:

We are using cor() and corrplot() functions of library corrplot to plot the correlations between the numerical regressors in our dataset.

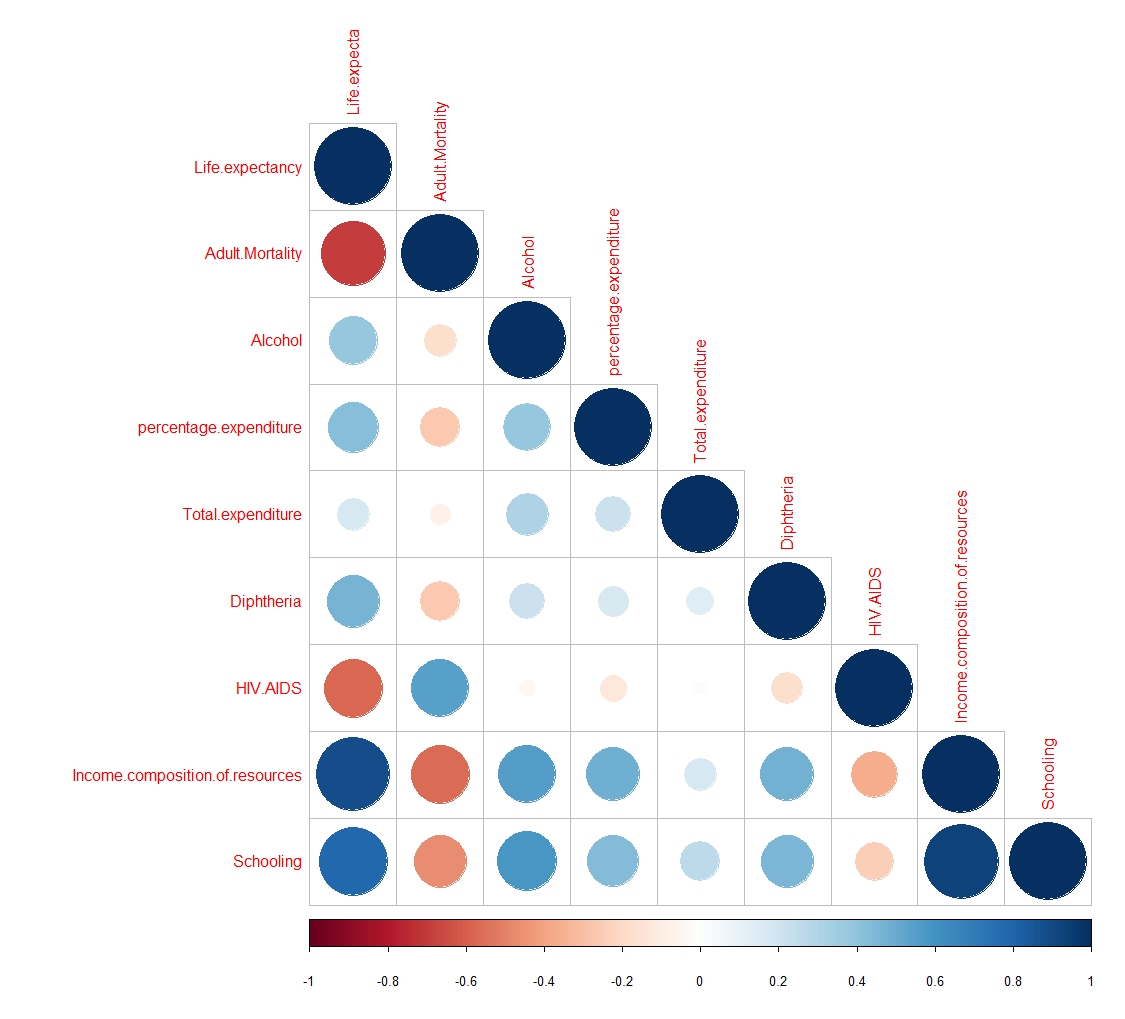






**Observation:**

As seen in above table Life expectancy have a high positive correlation of 0.889 and 0782 for Income composition and Schooling respectively. Other than that life expectancy has a high correlation with the Schooling Income composition resources . It has a negative correlation with adult mortality and HIV aids.



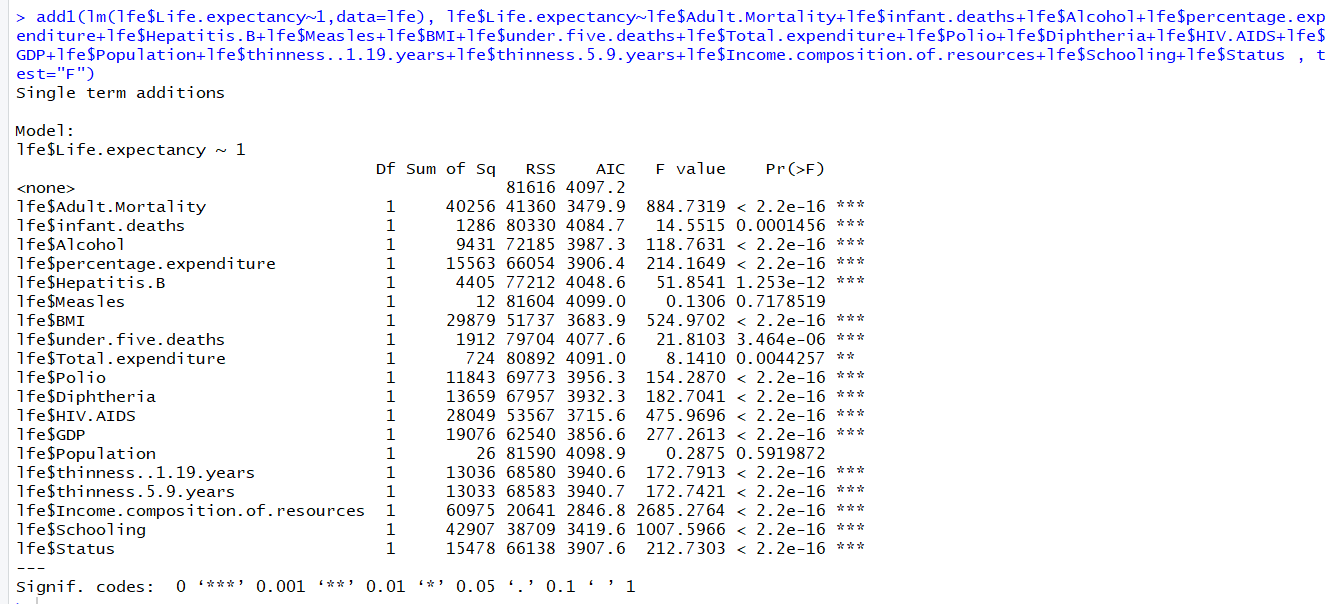
**Observations**:

In the above plot , the correlation increases as the circle becomes bigger and darker blue or darker brown. Based on this, we can see that there is a strong positive correlation of Life expectancy with Schooling and Income composition of the resources and there is a negative correlation of Life expectancy with Adult mortality and HIV AIDS. Bigger circles signifies stronger relationship and orange relation signifies negative relationship. We get the same observational insight as we got from corr function in previous example

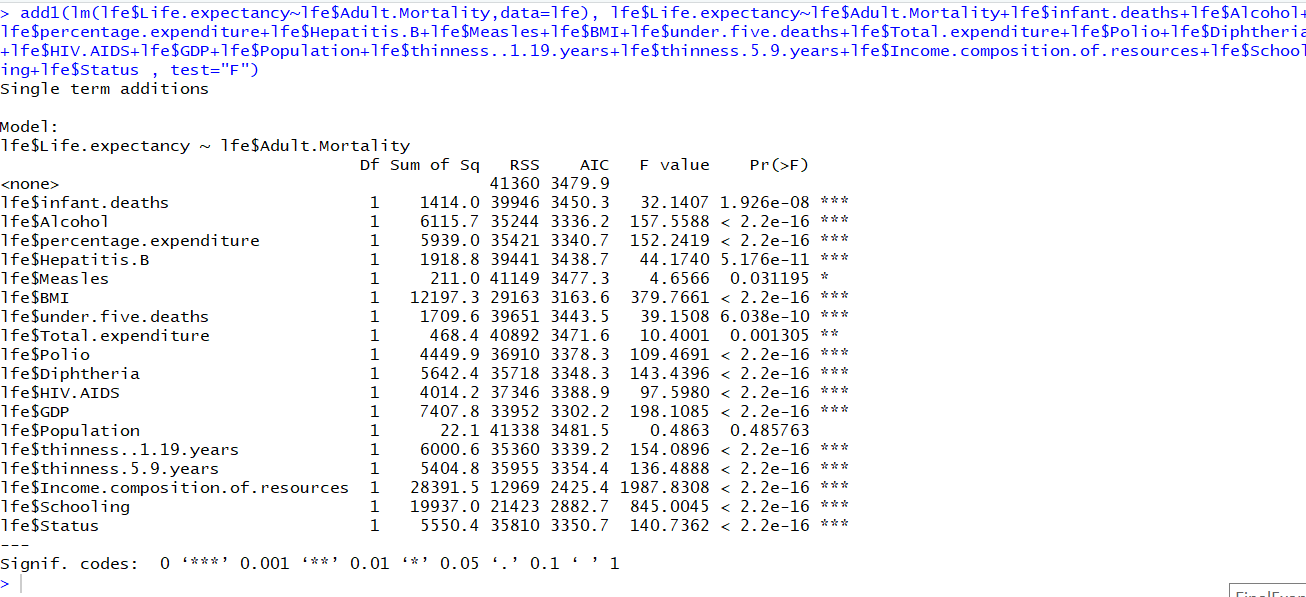
**Choosing best possible covariates**

**Checking the model with forward selection-**

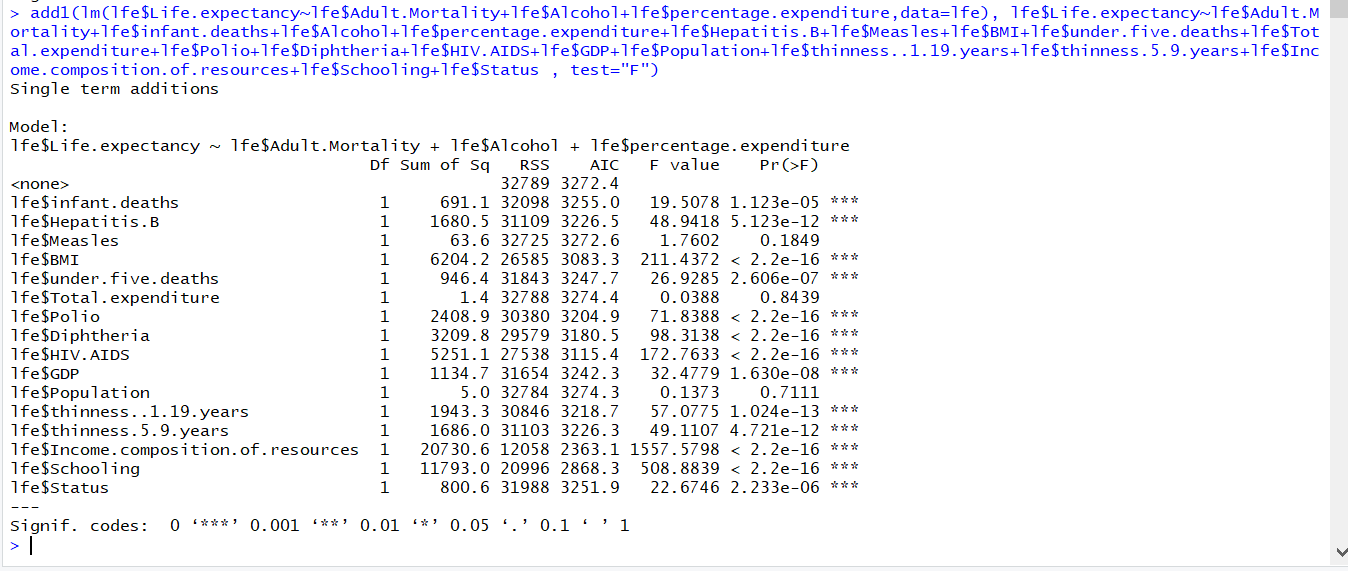
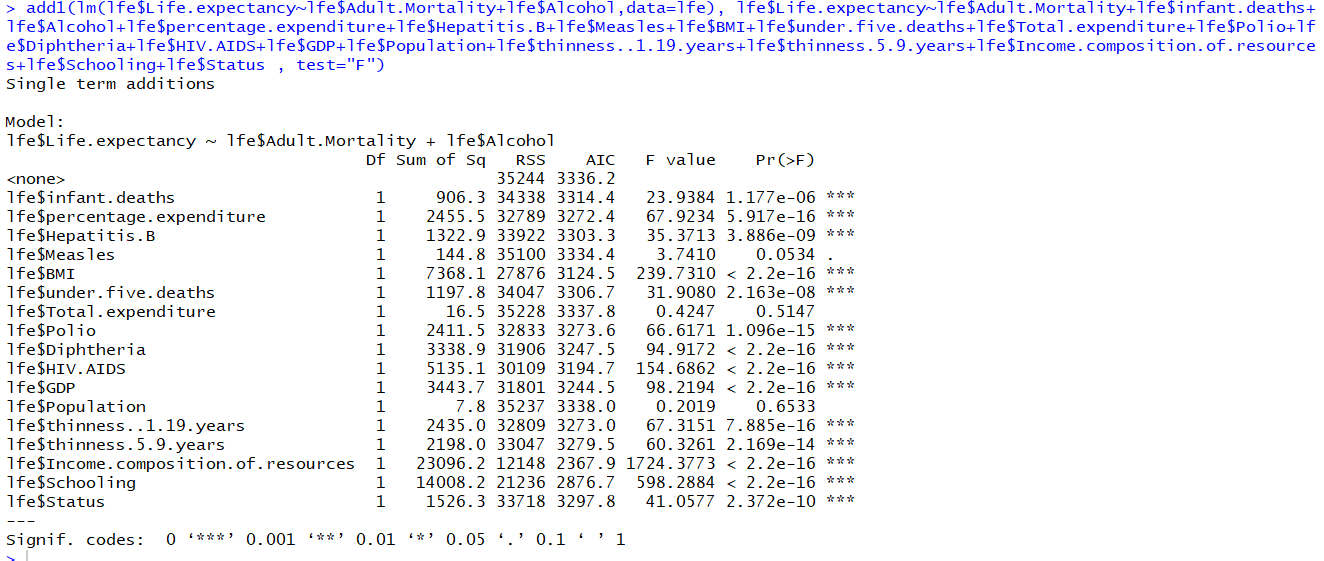
Forward select is a technique for choosing the best possible covariates for the regression in a multiple covariate model where we start with zero covariates and then keep adding the covariates in the model and keep checking the relative p values of the rest of the variables and then adding the most relevant covariate in the next round until we all the other covariates insignificant in the model



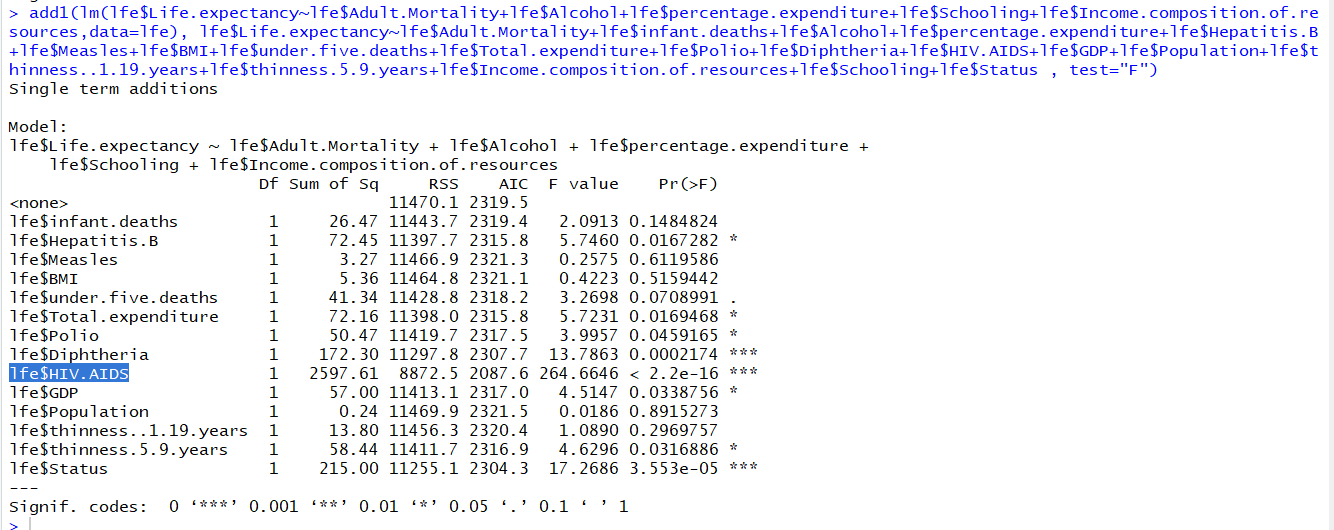
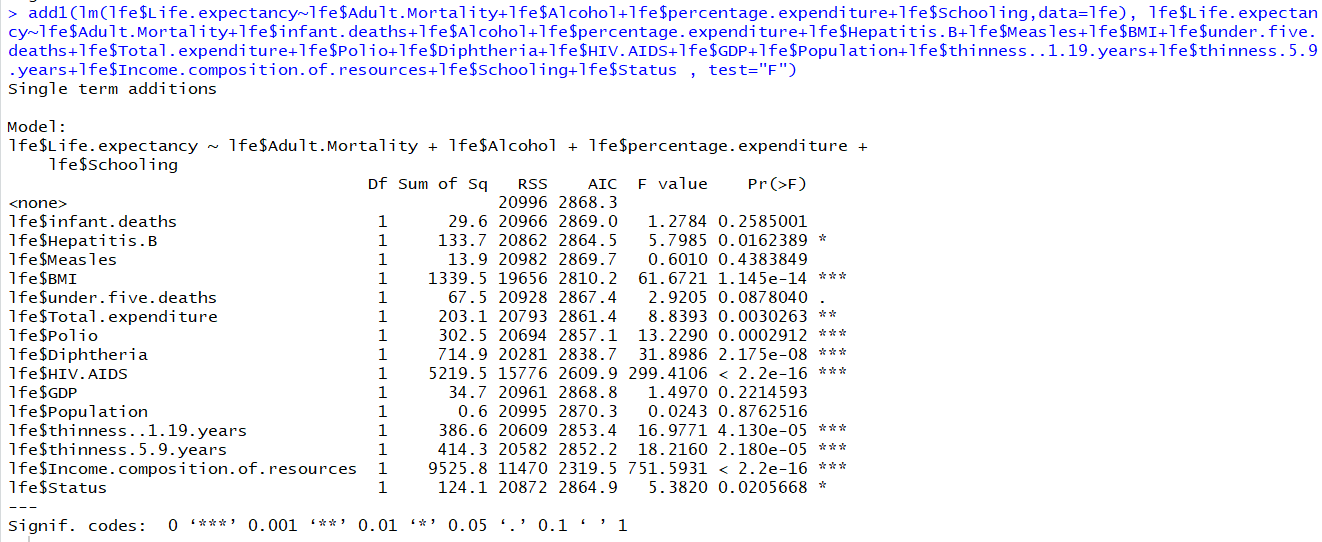
**Adding lfe$adult.mortality** We can see from the results here adultmortality is the most significant variable because of the p value similarly there are few others with the same p values but we will choose adult mortality for now . We will now observe the values for rest of the variables and



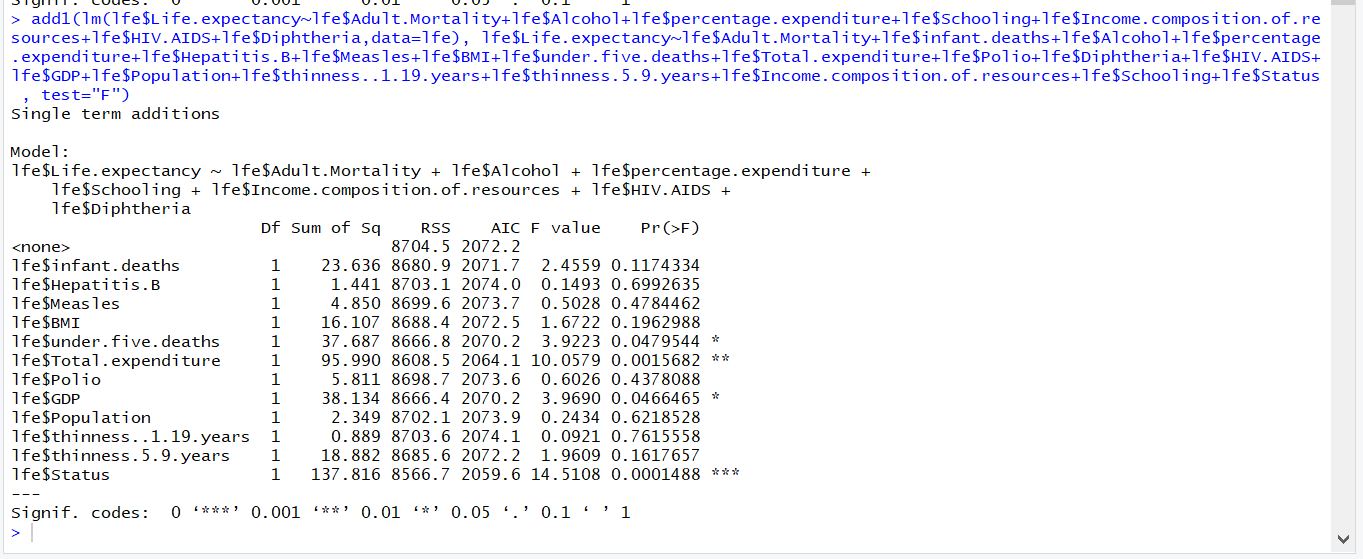
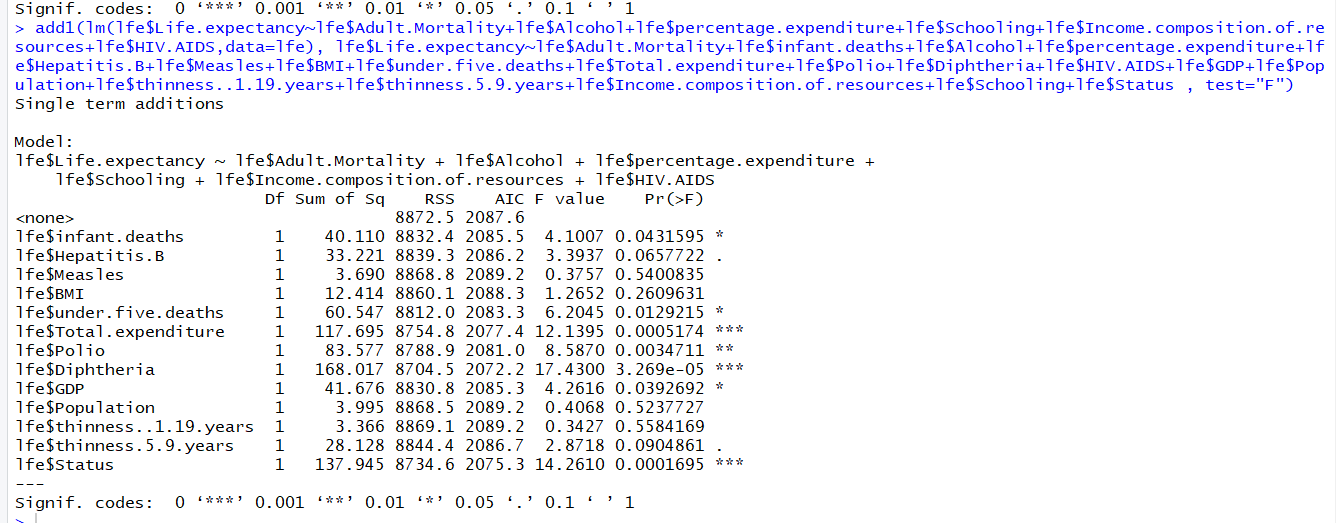
**Adding** lfe$Alcohol lfe$percentage.expenditure



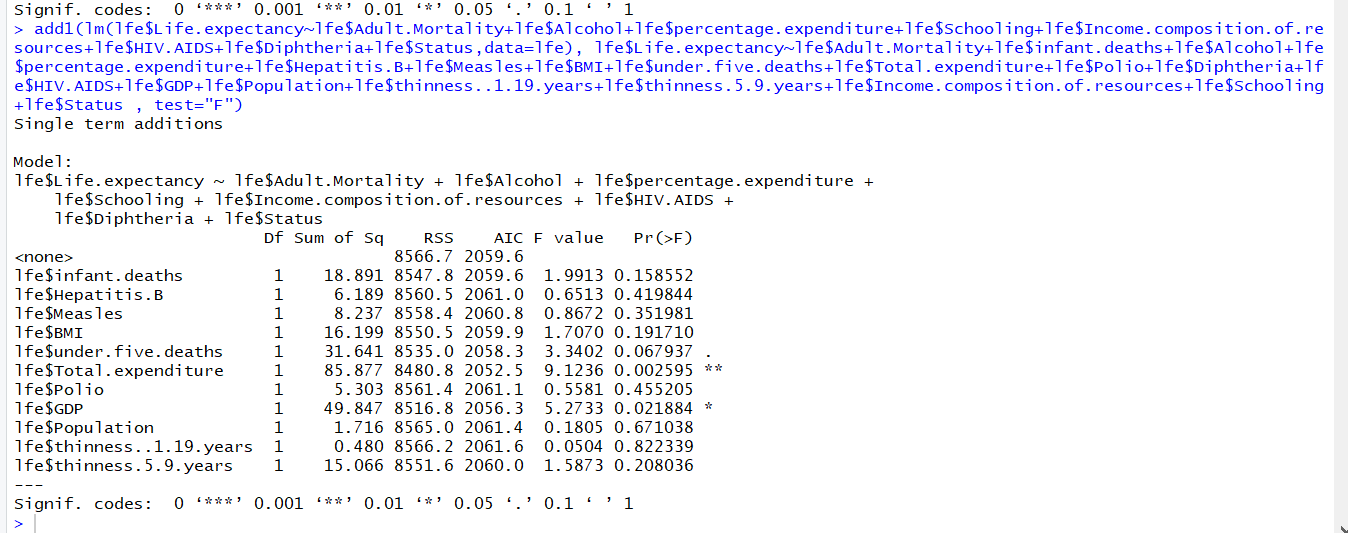
lfe$Schooling lfe$Income.composition.of.resources



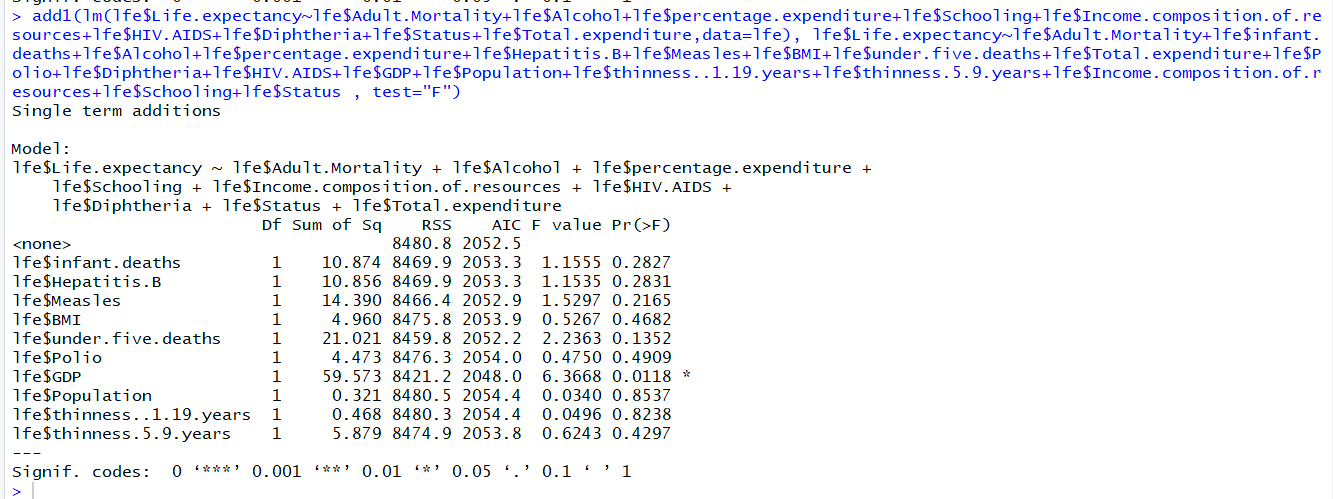
lfe$HIV.AIDS lfe$Diphtheria



lfe$Status



lfe$Total.expenditure



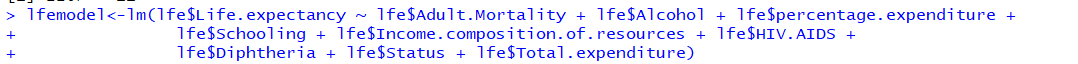
We removed all the unnecessary columns from our dataset and we have total 9 covariates remaining

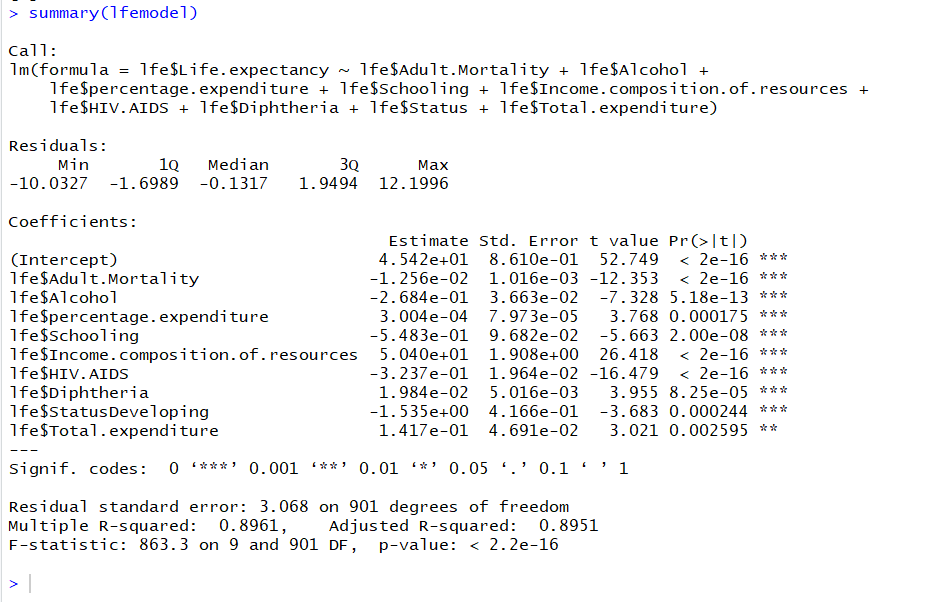
After the step wise elimination we have these columns

lfe$Adult.Mortality,lfe$Alcohol,lfe$Schooling,lfe$Income.composition.of.resources ,lfe$HIV.AIDS+lfe$Status, lfe$Diphtheria, lfe$percentage.expenditure , lfe$Total.expenditure, HivAids

**Creating a linear model for data:**

**Rcode:**





The intercept in our example is the expected Life expectancy if the value of our covariates was zero.

. The coefficient standard errors tell us the average variation of the estimated coefficients from the actual average of our response variable.

**t value:**

This is a test statistic that measures how many standard deviations the estimated coefficient

is from zero.

**Pr(>|t|):**

This number is the p-value, defined as the probability of observing any value equal or larger

than t if H0 is true. The larger the t statistic, the smaller the p-value. Generally, we use 0.05 as

the cutoff for significance; when p-values are smaller than 0.05, we reject H0. We can reject the

null hypothesis in favor of believing there to be a relationship between price and all covariates.

How well does the model fit the data?

**Residuals:**

This section of the output provides us with a summary of the residuals (recall that these

are the distances between our observation and the model), which tells us something about how

well our model fit our data. The residuals should have a pretty symmetrical distribution around

zero. Generally, we’re looking for the residuals to be normally distributed around zero (i.e. a

bell curve distribution), but the important thing is that there’s no visually obvious pattern to

them, which would indicate that a linear model is not appropriate for the data. Our residuals

look pretty symmetrical around 0, suggesting that our model fits the data well.

**Residual standard error:**

This term represents the average amount that our response variable measurements deviate

from the fitted linear model (the model error term).

**Degrees of freedom (DoF):**

It is the number of independent pieces of information that were used to calculate an es-

timate. DoF are related to, but not the same as, the number of measurements.

**Multiple R-squared:**

The R2 value is a measure of how close our data are to the linear regression model. R2

values are always between 0 and 1; numbers closer to 1 represent well-fitting models. R2 always 10

increases as more variables are included in the model, and so adjusted R2 is included to account

for the number of independent variables used to make the model.

F statistic:

This test statistic tells us if there is a relationship between the dependent and independent

variables we are testing. Generally, a large F indicates a stronger relationship.

**p-value:**

This p-value is associated with the F statistic, and is used to interpret the significance for

the whole model fit to our data.

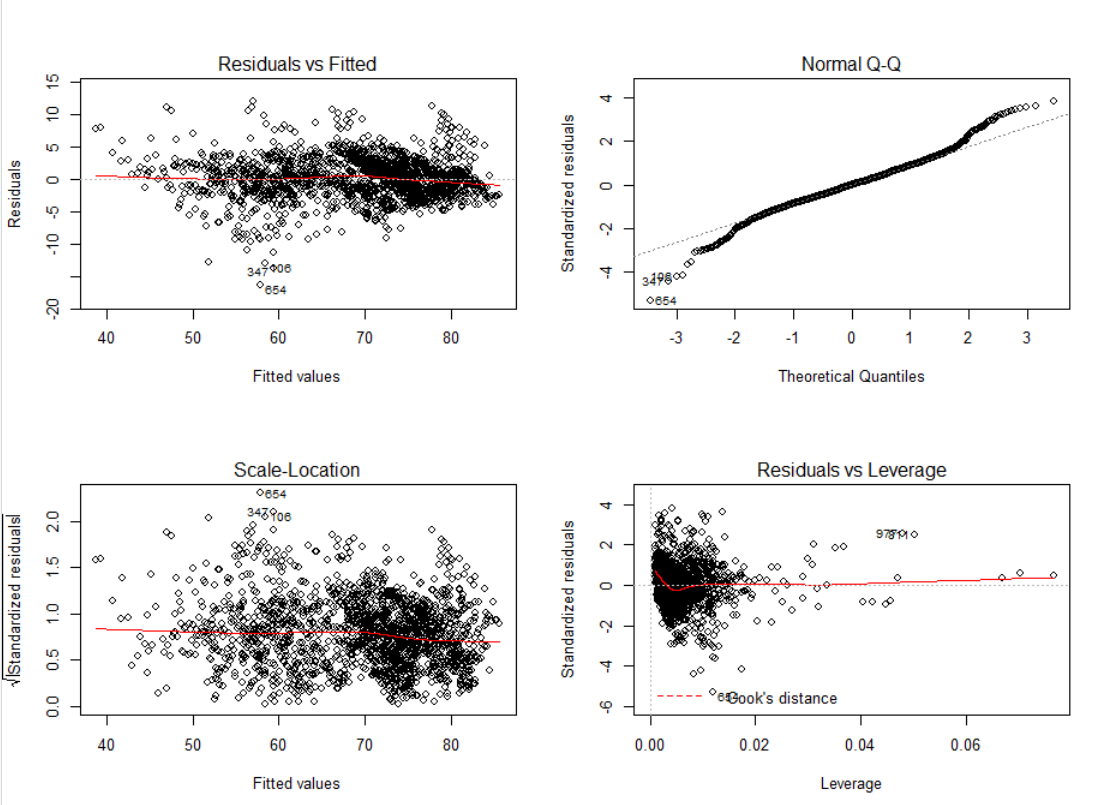
**Observation:**

We may observe our p values for individual model is quite low so we are which suggests all these attributes are significant in deciding the life expectancy. Our r-square is quite on the lower side because this is a social data and a lower r square is expected for these models.one might suspect we have a lot of covariates resulting in increased R square but our adjusted r square is close to our R square suggesting that is not the case. Our F values is also very low suggesting overall model is highly relevant.

**Model Adequacy check :**

Under this we will check different parameters of our finalized model and try to increase the model adequacy even further using transformation or elimination of other variables



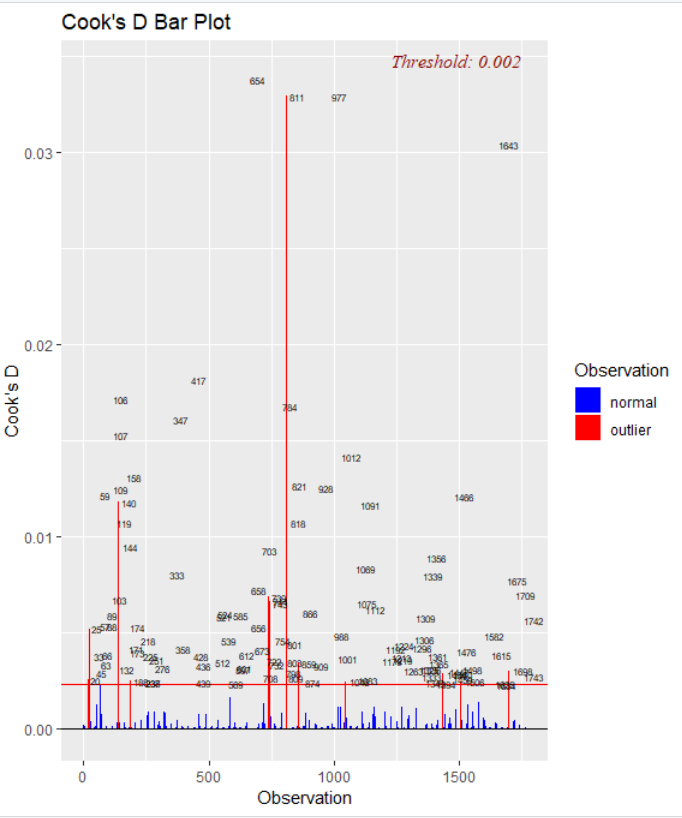


QQ plot shows us a lightly tailed graph, Residual vs fitted values are some what satisfactory as most of the values are centered around the zero residual and plot looks evenly distributed we will do the detailed residual analysis for individual graphs later when we do the transformation.

Outlier analysis:

We may notice the residuals vs leverage graph of ours doesn’t show the curve for cook distance according to our research this is bas sign this only means our model doesn’t has a point with large enough leverage to appear on the graph. From the above graph we can not see any such outliers in residual vs leverage graph





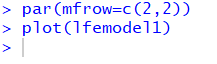
Cook’s D bar graph gives us the outliers with their observation number

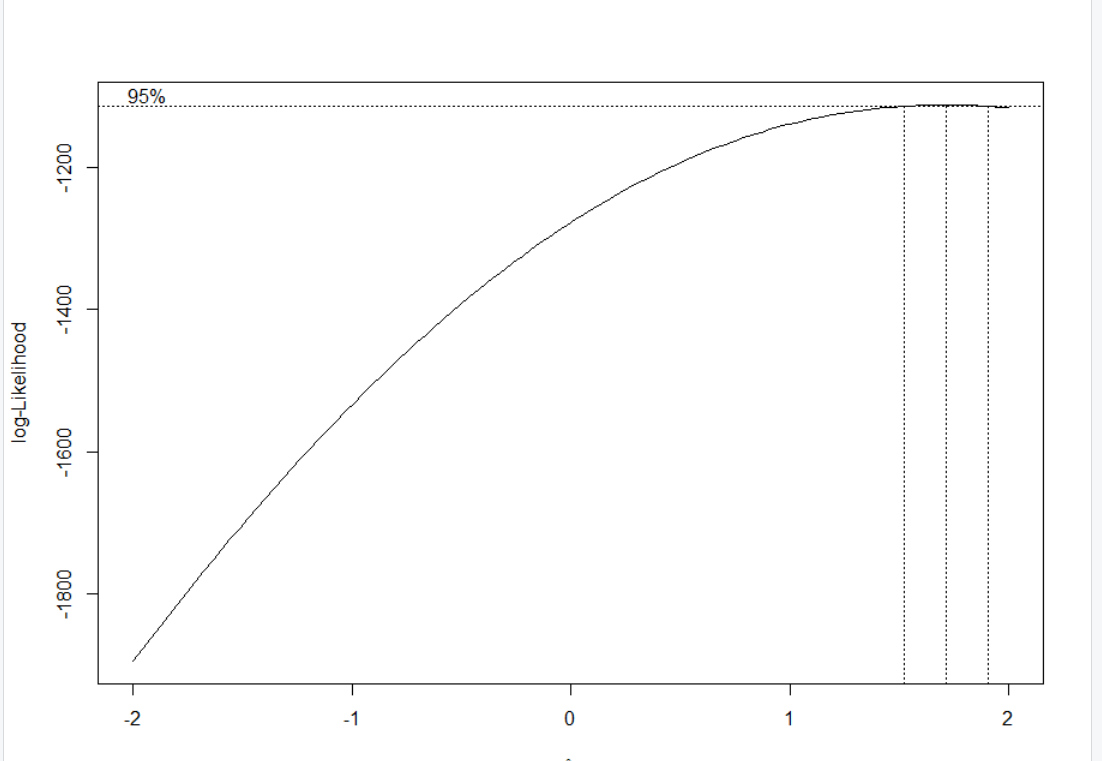
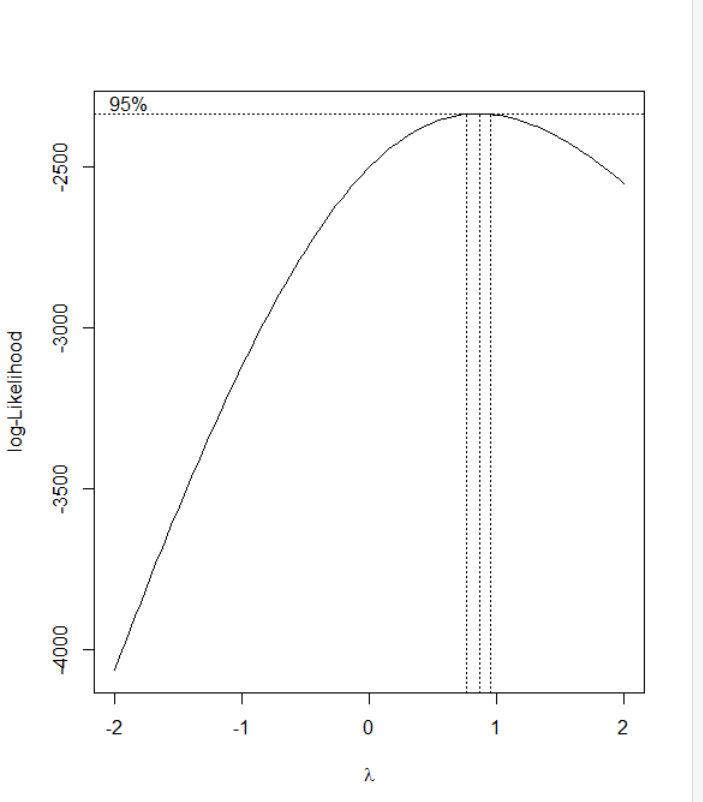
**Observation:**

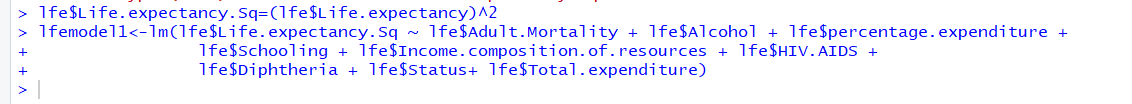
From the cook distance leverage graph and the Cook’s D bar graph gives us the outliers with their observation number

**BOXCOX:**

We will check for the boxcox plot of our model if we need to do any transformation in x or y to make our model even more adequate



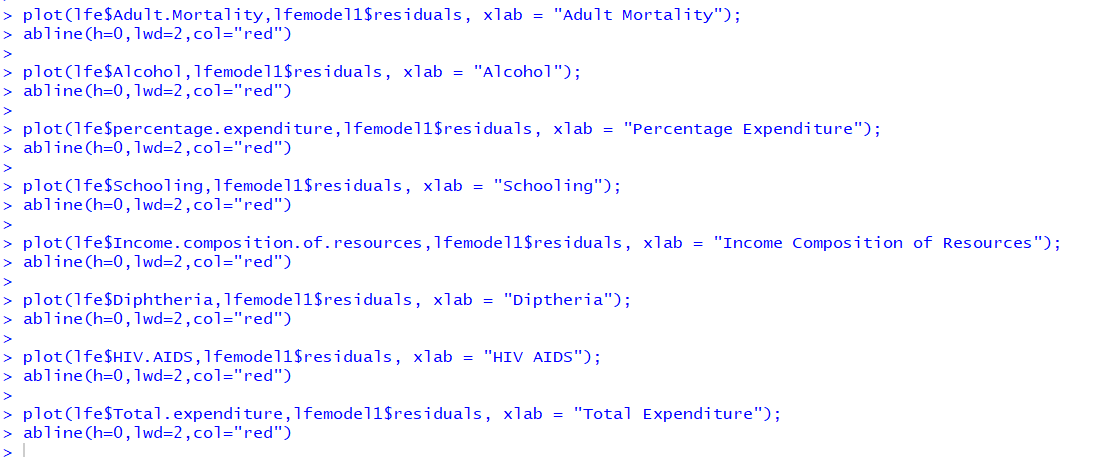
**Observation:**

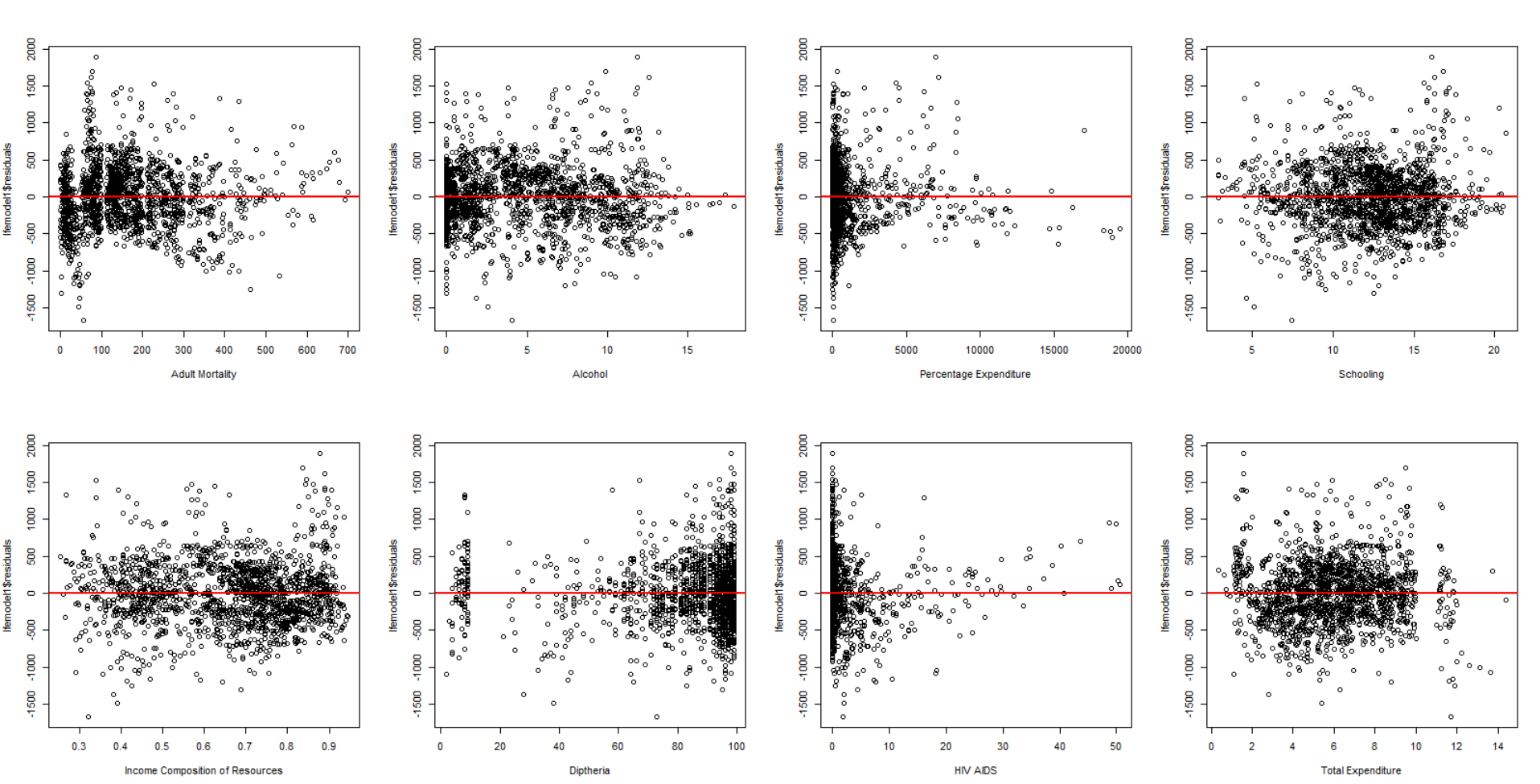
We can see the box plot for our data indicates we have a lamda close to 2 in this case we can do the square transformation of y to resolve the issue, on doing square transformation we will get lamda close to 1 which is a satisfactory value so we will go ahead with our model

**Residual Analysis:**

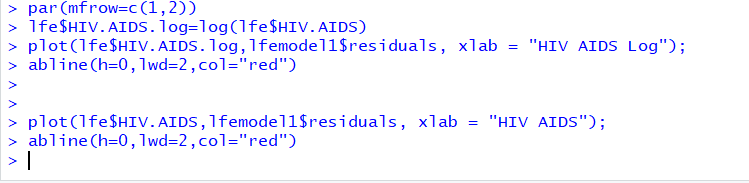
Goal here is to check for the residual values and remove any nonlinearity, non-normality or any unequal variance we can have from our figure

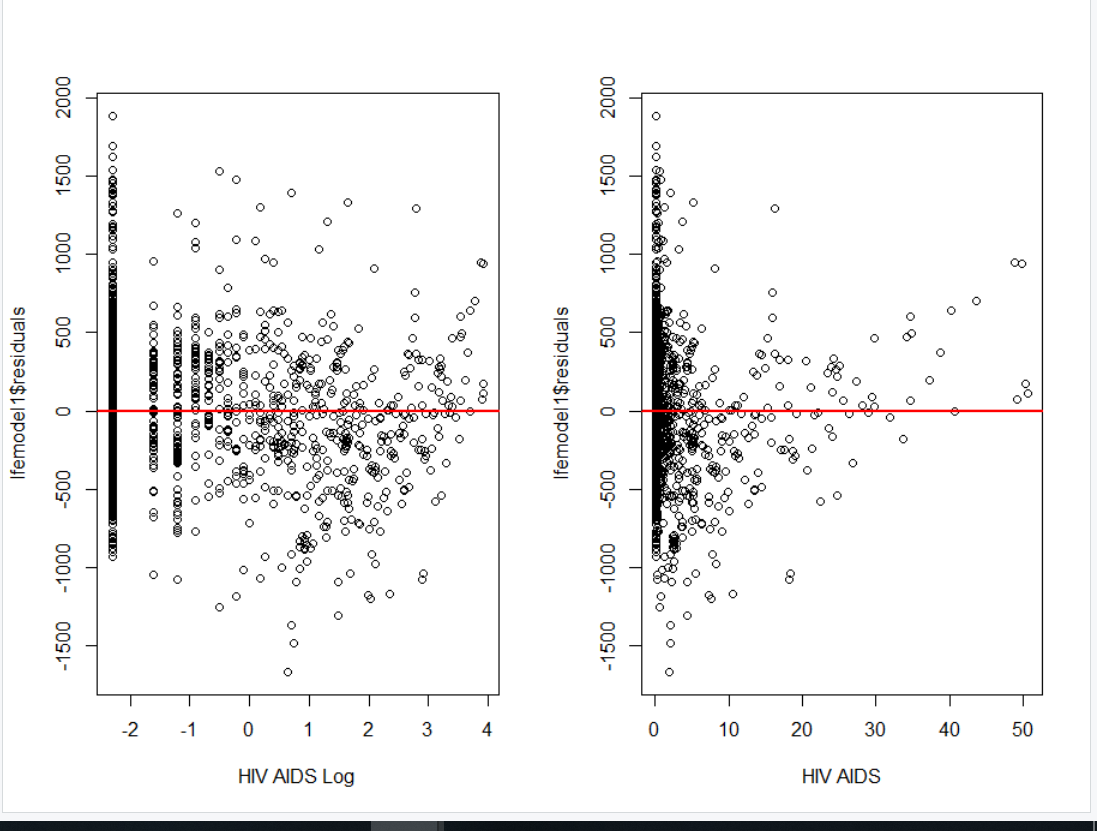
**R-code and results:**

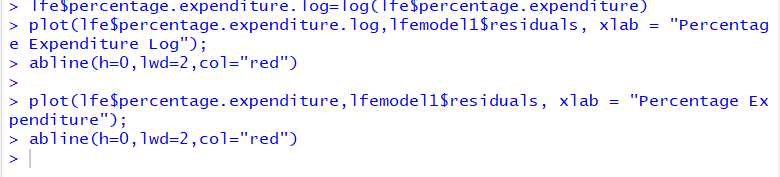


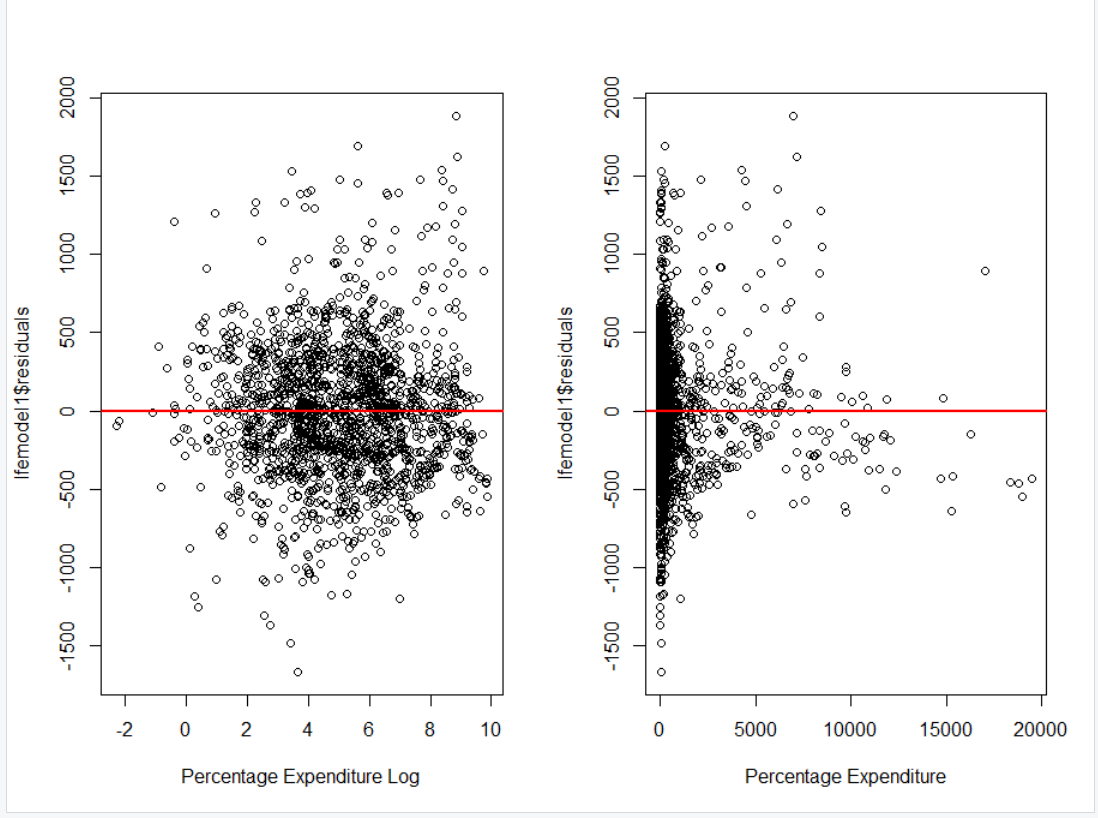


**Observation**: We can see for HIV aids and for percentage expenditure is left skewed and hence we will do a log transformation for these covariates.





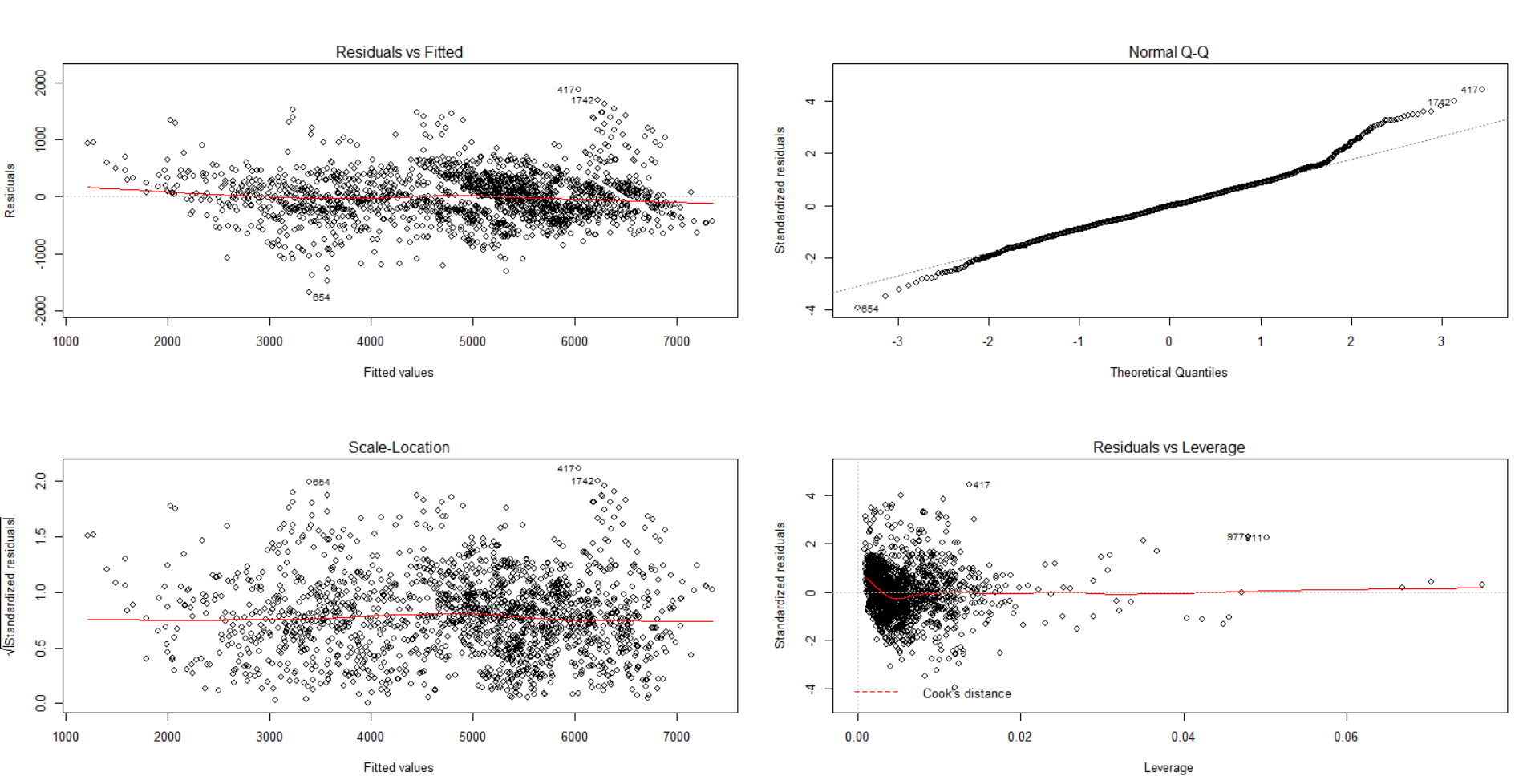


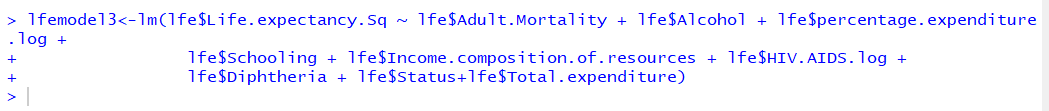


**Observation:**

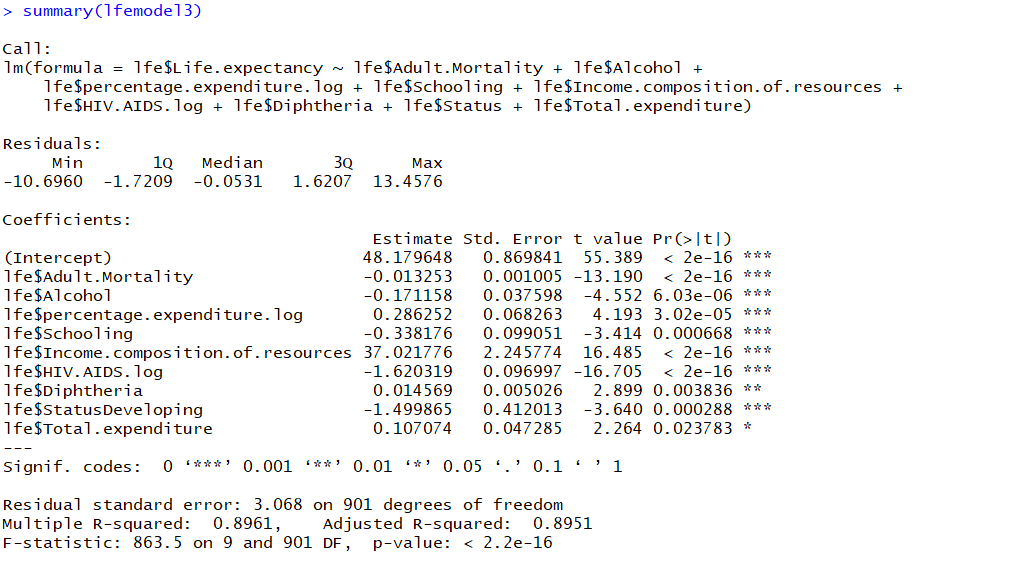
We have kept the results of transformation side by side just to compare the results and we may notice there is a significant improvement In the residual plot after the transformation so we have decided to go ahead with this model and take a log transformation of the respective variable

**Plot(model1):**





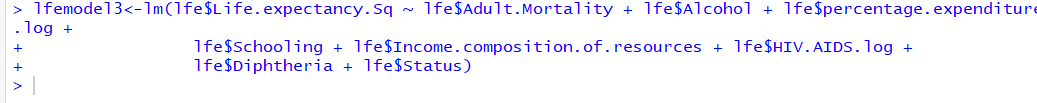
We will make a new model with all these transformed values



**Observation:**

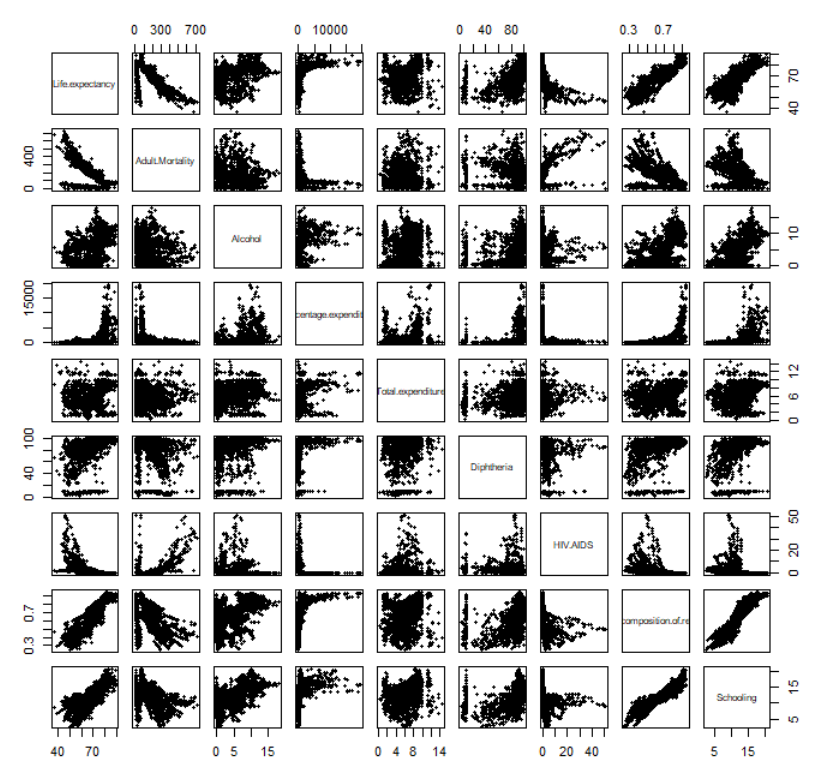
Thought the r square has increase our p value for the Total expenditure has become insignificant after the transformation this covariate has a high p value anyways so we have decided to remove this covariate just to make sure we maintain our model adequacy.

New model:



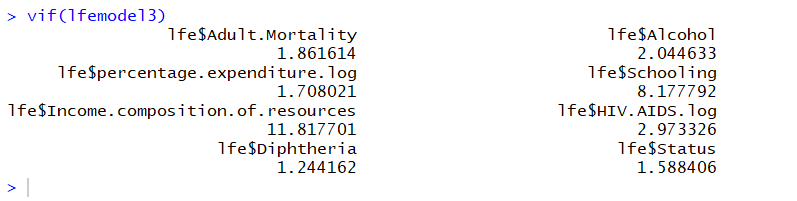
**Checking for multicollinearity of the model:**

**Goal:**  Our c



From the graph it looks like there is a multicollinearity between schooling Income composition as the graphs look very linear.

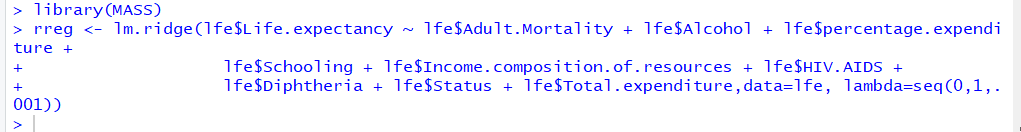
We will check the vif value to confirm the multicollinearity that we observed in the graph

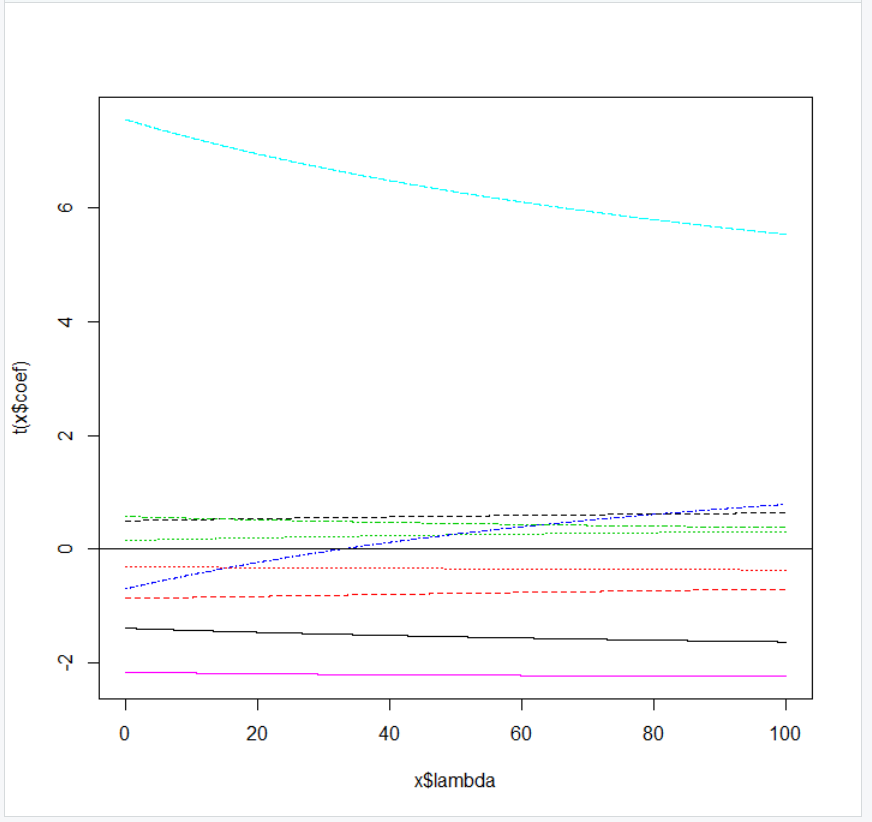


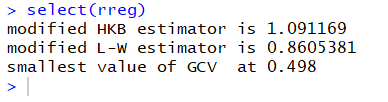
The variance inflation factor confirms there is a multicollinearity in schooling and income composition resources is very high as both have a very high vif

**Removing multicollinearity:**

We will use ridge regression to remove multicollinearity

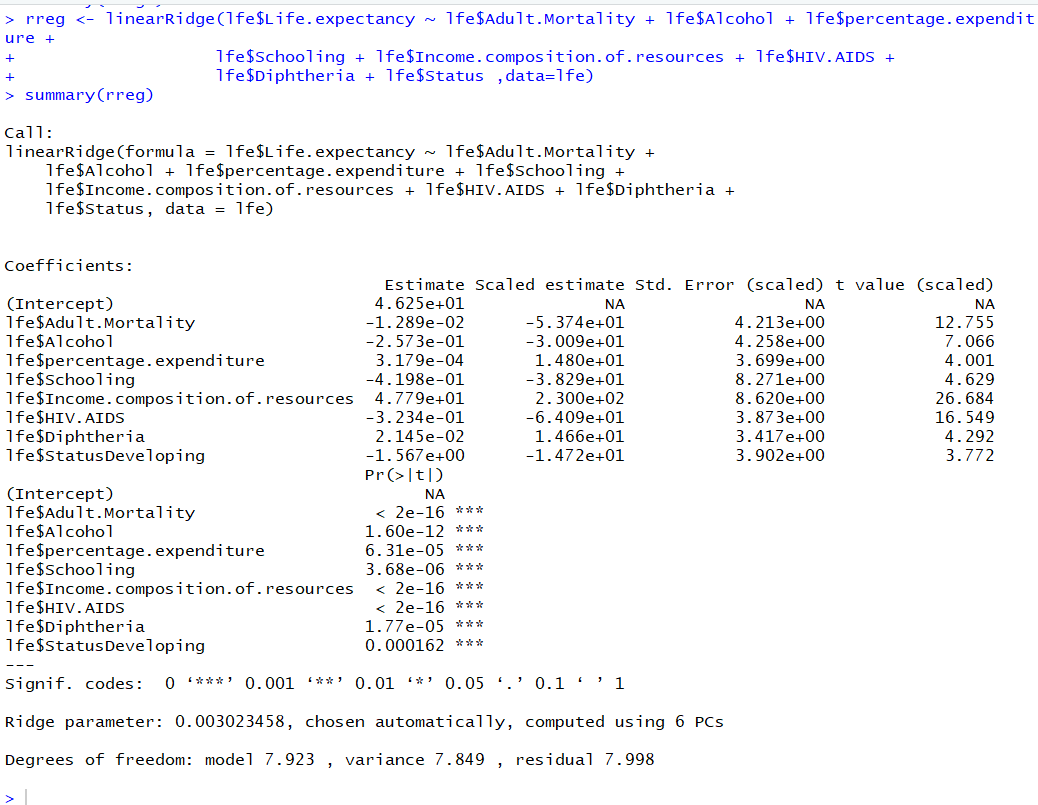






**Observation :**

We observe that VIF value is significant i.e. greater than 5 for 2 of the variables schooling and income composition suggesting the two variables are correlated. To deal with multicollinearity we did the ridge regression but even for high Lamda we are not able to see any significant change in the



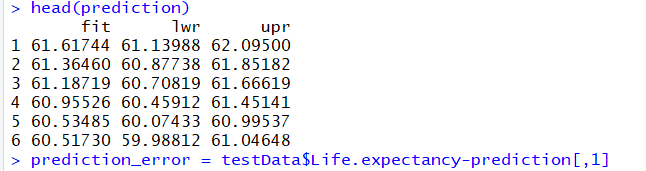
We had another try on ridge regression here is the solution we tried to find the VIF Values in order to check

**Prediction and Validation Model:**

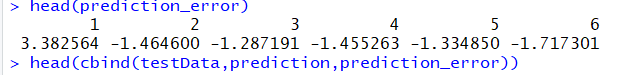
We will make a prediction using predict function using the test data we kept earlier aside for this time only all the other modelling we have done is on lfe data which was 80 percent of our data set.



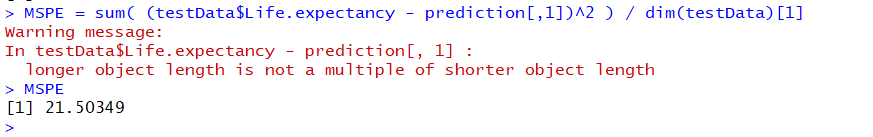
Next we check for the predicted value we use the head function to see the first 6 values of the dataset



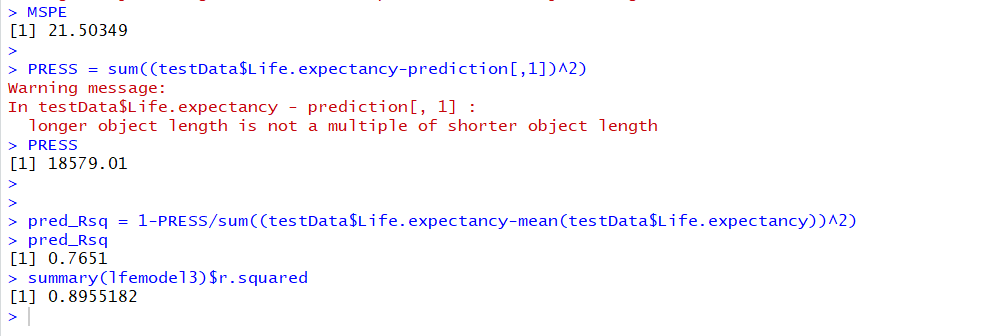
Next we calculate the prediction error by substracting the test value from the predicted values. Head will again give us first 6 values.



Next we calculated the MSPE value using the formula.



Next we calculated the press values and predicted R square value



**Observation :**

Through our prediction analysis we got the predicted r square as 0.7651 which is less as compared to r square value of the original model but this is maybe because our dataset had some unreliable values in the middle this might be the sole reason we don’t have a very r square value altogether .Since we have a social data at hand this kind of anomalies are expected In fact this value is quite high when you compare it with other social models this must be because of the data cleaning we did with removing all the run values

**Conclusion:**

As per our analysis we can say that Life expectancy= -0.0133\*Adult Mortality -0.165Alcohol -284 \*Schooling + 35.9Income.composition.of.resources -1.61\*HIV.AIDS -1.54\*Status+ 0.015\*Diphtheria+0.309\*percentage expenditure , We observe that the life expectancy has a positive relationship between income diphtheria and percentage expenditure a negative relationship with adult mortality , HIV deaths, status, Alcohol