**Life Expectancy**

DS6372: Project 1

[Introduction](#_bjntdd108i34) **1**

[Data Description](#_x0wxmladl51) 1

[Exploratory Data Analysis](#_2glkgg7yw9k) 1

[Objective 1](#_iv4c8jwkrjuq) **3**

[Problem](#_xrgx36ecyol) 3

[Approach](#_g35amjantwyf) 3

[Model variable selection](#_44qgl49x2dim) 4

[Forward, Backward and Stepwise](#_xtshph2ab1a2) 4

[LASSO](#_dibuuff4jqxj) 4

[Model 1 (Interpretable)](#_8xpn5ymvtptm) 4

[Assumptions](#_beuhqx3ihlsn) 4

[Parameter Interpretation](#_2pgkwouci335) 5

[Model 2 (Predictive)](#_erixlk668mt7) 6

[Hypothesis Test](#_yldz8seow2nf) 6

[Compare Competing Models](#_v272tgdgoyfv) 6

[Objective 2](#_2fiakmjzqdgl) **7**

[Problem](#_8bgii91qphbx) 7

[Approach](#_berrr9du7vpp) 7

[Compare Competing Models](#_ln45y85ky9ii) 8

[Conclusion](#_nwvlv0t3ppjk) **8**

[Appendix](#_qf41vgd5rb61) **9**

[EDA](#_oug9htc3llx8) 9

[Variable selection method analysis](#_q23k4z6vxv7f) 14

## 

## Introduction

What factors impact life expectancy? Can life expectancy be predicted? The World Health Organization (WHO) maintains a database with life expectancy information for all countries by year along with other variables that could contribute to life expectancy. The data falls into four main areas: health, economic, social, and immunization. This project makes use of the WHO data to better understand life expectancy and how to predict it. In this paper, three different models will be explored:

1) A regression model to identify key relationships between life expectancy and factors related to health, economic, social, and immunization.

2) A parametric regression model to predict life expectancy.

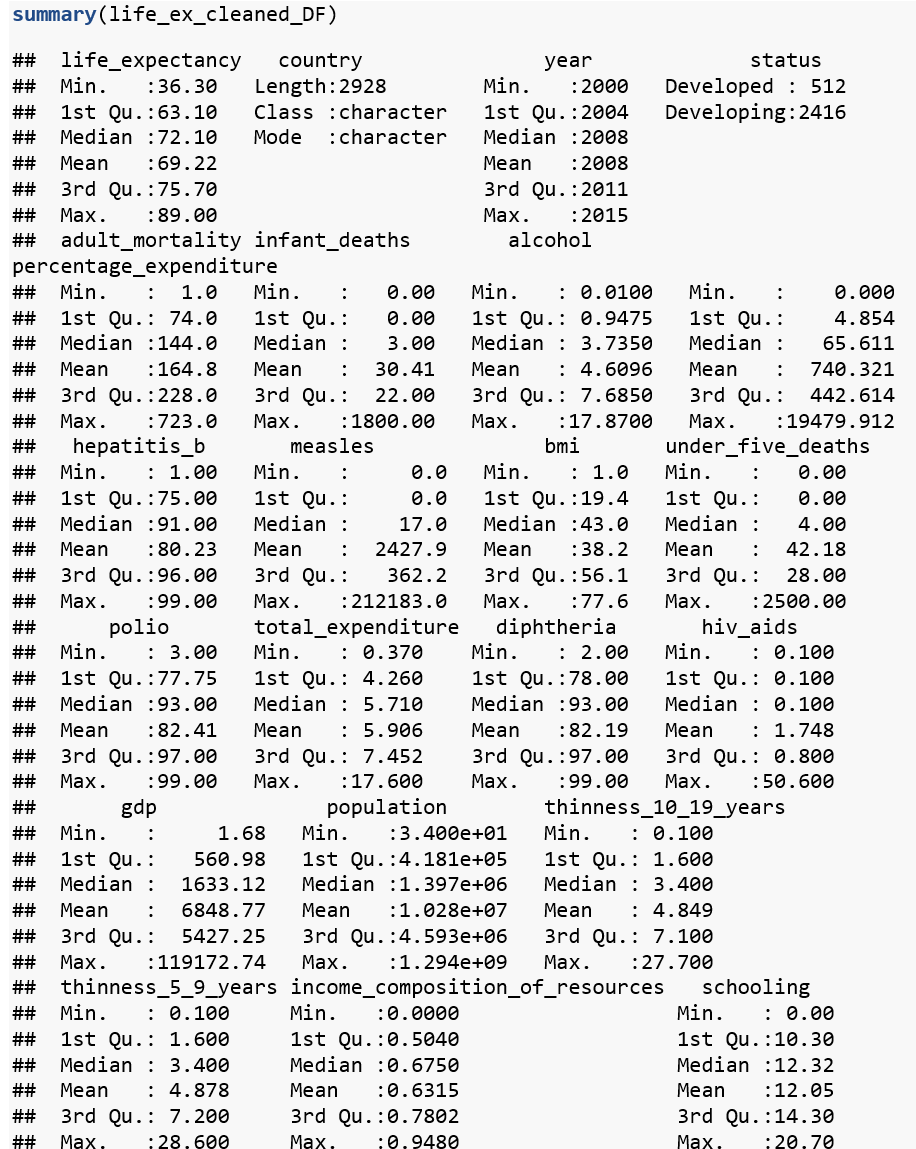
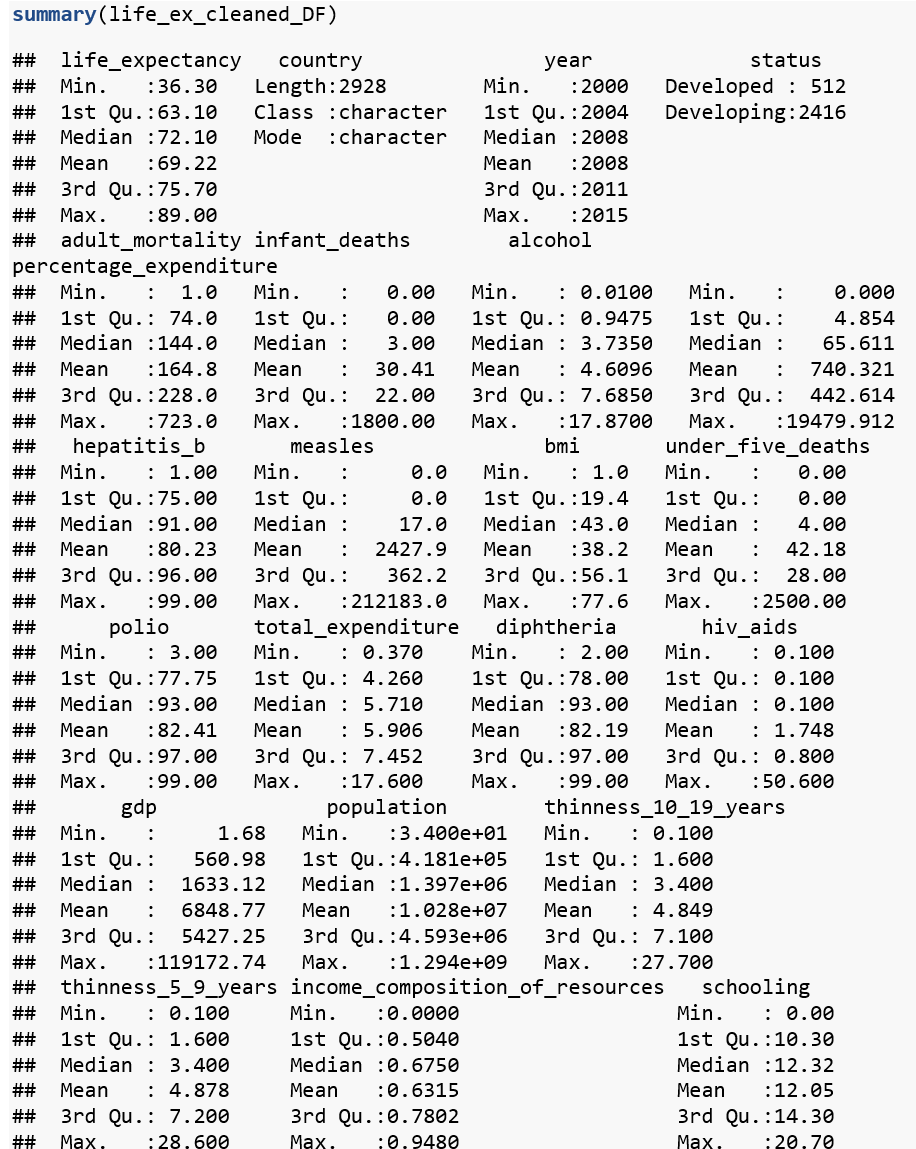
3) A non-parametric regression model to predict life expectancy.

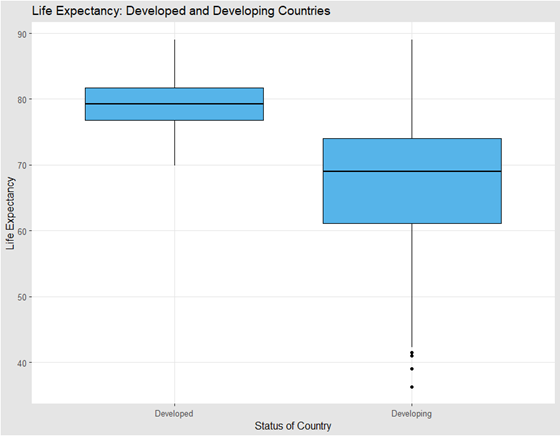
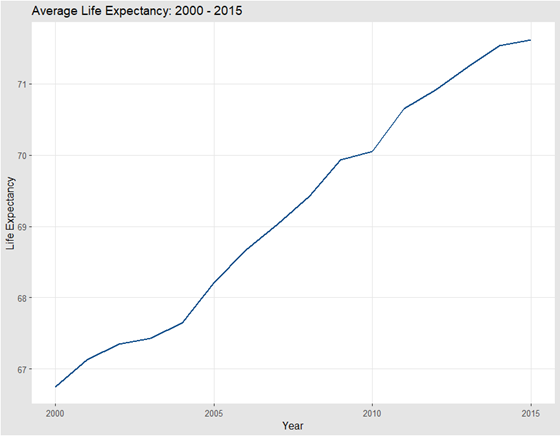
### Data Description

The Life Expectancy dataset used in this analysis was collected from the WHO and made available at Kaggle: <https://www.kaggle.com/kumarajarshi/life-expectancy-who>. It includes data from 193 countries for the years 2000 – 2015. The 2938 rows contained in the dataset include average life expectancy per country for each year. In addition to the average life expectancy, there are 21 columns of data related to health, economic, social, and immunization.

### Exploratory Data Analysis

Of the 22 columns of data available, two columns are factors (country and status) and the rest are numeric. Life expectancy values in this dataset range from 36.30 years to 89 years. BMI has some very weird mins and maxes. Will explore further, but this data may later show correlations that aren’t actually there because the min and max of BMI are not practical or possible.





Life Expectancy is linearly related to year, and year should be treated as continuous instead of as a factor in this model. We will drop BMI for having unrealistic range of data, and no way to correct for it.

**Testing the life\_expectancy mean difference by status**

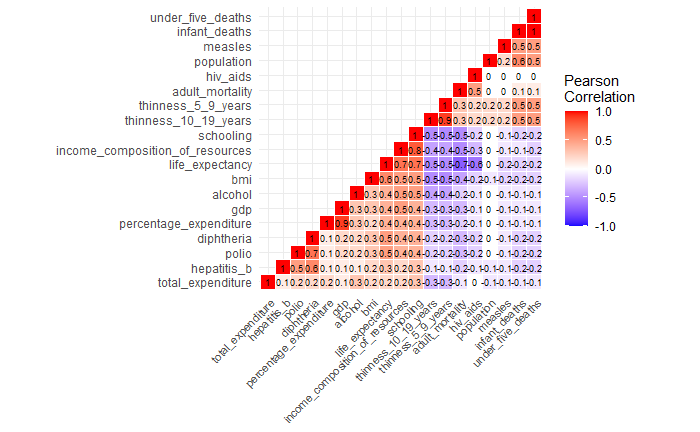
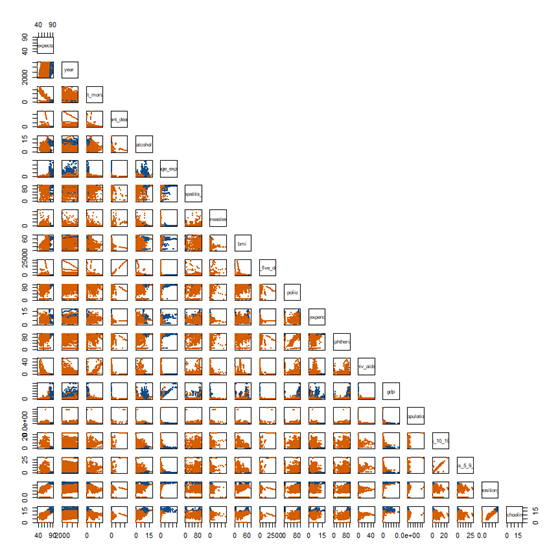
H0: μdeveloped = μdeveloping Ha: μdeveloped ≠ μdeveloping

**There is sufficient evidence to suggest that the means are equal**

**Reject H0**

**(p-value <2e-16)**

The following four variables are removed based on the multicollinearity among certain variables and having too many null values: Thiness.5.9.years, GDP, under.five.deaths, Population



See Appendix for additional EDA

## Objective 1

### Problem

1. Build a highly interpretable model that identifies key relationships associated with life expectancy.
2. Build a highly predictive model for life expectancy.

### Approach

1. Explored and cleaned the data
   1. Removed the population variable because there were too many missing values.
   2. Removed BMI because of extreme values that did not seem practically possible.
   3. Imputed missing values for the other variables.
2. Used VIFs to evaluate inflation factors and eliminate variables with multicollinearity.
3. With the remaining data, used variable selection techniques to identify the best variables for the model. Techniques used: forward, backward, stepwise selection and LASSO.
4. Added quadratics for hiv\_aids variable and percentage\_expenditure to the predictive model because a non-linear relationship was observed.
5. Added interactions for status and percentage\_expenditure because of visual evidence of an interaction. This was confirmed with a test for significance.
6. Hypothesis tests:
   1. Extra sum of squares: Difference in means for status of Developing vs. Developed countries.
   2. Checking the significance of the factorial variables - country and status. Country was found to not be significant and was removed.
   3. Lack of fit: Lasso and Backward
   4. Lack of fit: Stepwise and Backward
   5. Lack of fit: Interpretable and Predictive models (Stepwise)
7. Checked assumptions for full, interpretable, and predictive models.
8. Model diagnostics and comparisons
   1. ASE
   2. AIC
   3. BIC
   4. Adjusted R squared
   5. RMSE
9. Model selection based on fit test significance and model diagnostics metrics.

### Model variable selection

|  |  |  |
| --- | --- | --- |
| **Forward and Backward** | **Stepwise** | **LASSO** |
|  |  |  |
| # of vars: 12 | # of vars: 12 | # of vars: 13 |

##### Forward, Backward and Stepwise

* Forward and Backward methods had the same result for the variable selection.
* The optimal number of variables is 12 (11 explanatory variables plus 1 intercept) based on the lowest ASE associated with the number of the variables as shown below.

##### LASSO

* Two vertical dotted lines represent the following:

1. The one is at the minimum(the one pointing at 17 on top)

2. The other vertical line is within one standard error of the minimum(the one at 13). We have this line because cross validation error is measured with some variance. The result chose the second line of 13 variables as seen in the plot above.

See Appendix for additional variable selection analysis in depth.

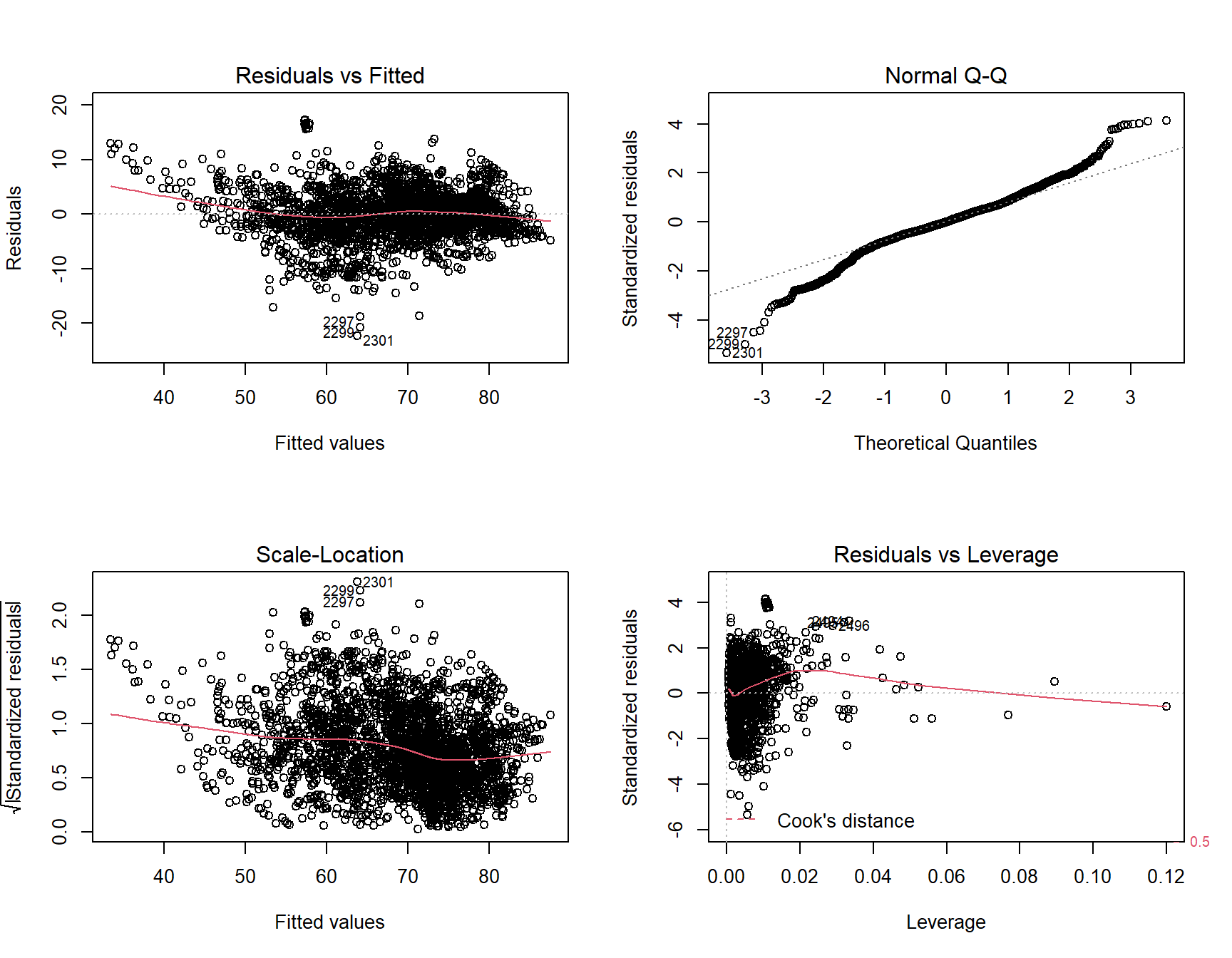
### Model 1 (Interpretable)

life\_expectancy ~ status + adult\_mortality + percentage\_expenditure + measles + polio + total\_expenditure + diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources + schooling

### Assumptions

**Constant Variance:**  The evidence suggests that the variance is mostly constant.

**Normality:** There isn’t enough evidence to suggest that the normality assumption is not met.  
**Independence:**  The observations are from 193 countries. Measurements were taken for each country over a 15-year time span, potentially causing a violation of independence. We will note the potential lack of independence and proceed with caution.



### Parameter Interpretation

The data suggest that for every additional year of schooling, the estimated mean life expectancy increases 0.796 years (p-value < 0.0001) holding all other variables constant.

The mean life expectancy for people in Developing countries is estimated to be -1.01 years (p-value < 0.0001) than the mean life expectancy of people in Developed countries holding all other variables constant.

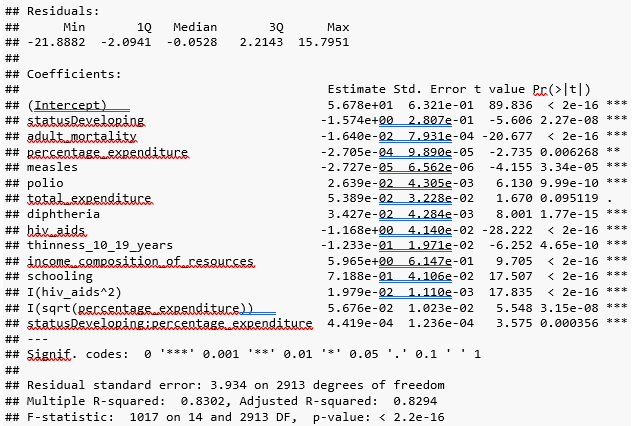
A 95% confidence interval for the mean increase in life expectancy with each additional year of schooling is (0.71, 0.88). A 95% confidence interval for mean change in life expectancy for people in developing countries is (-1.52, -0.50).

In addition to statistical significance, we believe there is practical significance in these predictors as well.

|  |  |
| --- | --- |
|  |  |

### Model 2 (Predictive)

life\_expectancy ~ status + adult\_mortality + percentage\_expenditure + measles + polio + total\_expenditure + diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources + schooling + status:percentage\_expenditure + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure))



### Hypothesis Test

We wish to test the assumption that a full model (predictive model) is not the same fit as a reduced model (interpretable model).

Ho: Predictive Model = Interpretable Model Ha: Predictive Model != Interpretable Model

**There is sufficient evidence to suggest that these models do not fit the data the same.**

**Reject H0**

**(p-value: < 0.001)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Res.Df**  <dbl> | **RSS**  <dbl> | **Df**  <dbl> | **Sum of Sq**  <dbl> | **F**  <dbl> | **Pr(>F)**  <dbl> |
| **1** | **2913** | **45082.67** | ***NA*** | ***NA*** | ***NA*** | ***NA*** |
| **2** | **2916** | **51073.09** | **-3** | **-5990.429** | **129.0231** | **1.790762e-78** |

### Compare Competing Models

We see that there isn’t enough evidence to suggest that the fit between stepwise and backward is different. We do see that both of them are indicated to be better fitting than the lasso model. Additional evidence suggests that both the stepwise and backward models returned lower diagnostic fit metrics in RMSE, BIC, and AIC, and higher Adjusted R Squared compared to the lasso model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **ASE** | **RMSE** | **Adj R^2** | **AIC** | **BIC** | **DF** |
| Lasso | 18.4412 | 4.2943 | 0.7959915 | 16865.22 | 16931.02 | 11 |
| Interpretable | 17.443 | 4.176481 | 0.8069019 | 16706.28 | 16784.04 | 13 |
| Backward | 15.41182 | 3.925789 | 0.8292704 | 16347.78 | 16437.51 | 15 |
| Stepwise | 15.39709 | 3.923912 | 0.8293751 | 16346.98 | 16442.69 | 16 |

We will proceed with the predictive stepwise model as it has a slightly lower RMSE, higher Adjusted R Squared, lower AIC, and an additional degree of freedom; even though we see lower lower BIC from the backward model.

Based on the hypothesis test and the diagnostic metrics, the predictive model is preferred over the interpretable one.

## Objective 2

### Problem

Build a highly non-parametric model that identifies what can predict life expectancy.

### Approach

1. Build a KNN model

* **Model Phase 1:**

Our first model utilized the KNN function to predict values of life expectancy based on the neighboring data points. We initially used the same variables identified previously as highly influential on the outcome and removed those that were considered to have little impact. This type of model required a categorical comparison for the predictions, therefore the life expectancy predictions were grouped into two categories for those within or outside two standard deviations from the prediction mean. This model predicted values within the deviation range with an accuracy of 88.4%. However the standard deviation for this data set was quite large at 9.63 years making this model a poor predictor with values at ±19.3 years from our test set being considered accurate.

* **Model Phase 2:**

The second model utilized the KNN.reg function to calculate the predictions. The same predictor variables were utilized as done in previous models. To select a K value a range was tested to gather values of mean square and absolute error. These values were then analyzed for their minimum with respect to each other and the K value. The best fit was found at K = 3 based on the resulting lowest MSE value. The MAE values were found to typically increase with each increase in K value and therefore a low K value was prioritized.

1. Build a Random Forest Model

* This model utilized the Random Forest function to calculate predictions. The training data is split into multiple regions or trees to create a mean prediction. The predictions are then aggregated into a single prediction.
* **Pros:** It can determine the importance of each variable in the dataset, it predicts continuous or categorical responses, and it handles missing data.
* **Cons:**It’s conceptually explainable, but the inner workings of the model are not, it does not predict outside the range of the data, and it can overfit with noisy data.
* **Model Parameters:**  After running tuneRF and observing the Out of Bag errors, 5 was chosen to avoid overfitting (16/3, rounded down). The OOB Error was 0.514. The number of trees variable was set to the default value. Importance and Proximity were set to True.

### Compare Competing Models

|  |  |
| --- | --- |
| **Predictive Models** | **ASE** |
| KNN | 8.44 |
| Random Forest | 5.24 |

|  |  |
| --- | --- |
| KNN | Random Forest |
|  |  |

## Conclusion

Based on this analysis, if prediction is the only goal and we wanted to predict within the life expectancy range of the dataset, our recommendation would be the Random Forest model because it has the lowest ASE (5.24), but it could overfit due to the high percentage of variation explained (94.7).

The KNN model also produced a low ASE in comparison to the interpretable and predictive MLR models which would also make this model a viable choice, however it does little to explain why or how the relationships exist and are interacting.

## Appendix

## Importing the necessary libraries

library(ggplot2)  
library(ggthemes)  
library(gridExtra)  
library(tidyverse)  
library(plyr)  
library(dplyr)   
library(naniar) #keep  
library(zoo)  
  
  
# create not in operator  
`%notin%` <- Negate(`%in%`)

## Introduction

What factors impact life expectancy? Can life expectancy be predicted? The World Health Organization (WHO) maintains a database with life expectancy information for all countries by year along with other variables that could contribute to life expectancy. The data falls into four main areas: health, economic, social, and immunization. This project makes use of the WHO data to better understand life expectancy and how to predict it. In this paper, three different models will be explored: 1) A regression model to identify key relationships between life expectancy and factors related to health, economic, social, and immunization. 2) A parametric regression model to predict life expectancy. 3) A non-parametric regression model to predict life expectancy.

## Data Description

The Life Expectancy dataset used in this analysis was collected from the WHO and made available at Kaggle: <https://www.kaggle.com/kumarajarshi/life-expectancy-who>. It includes data from 193 countries for the years 2000 – 2015. The 2938 rows contained in the dataset include average life expectancy per country for each year. In addition to the average life expectancy, there are 21 columns of data related to health, economic, social, and immunization.

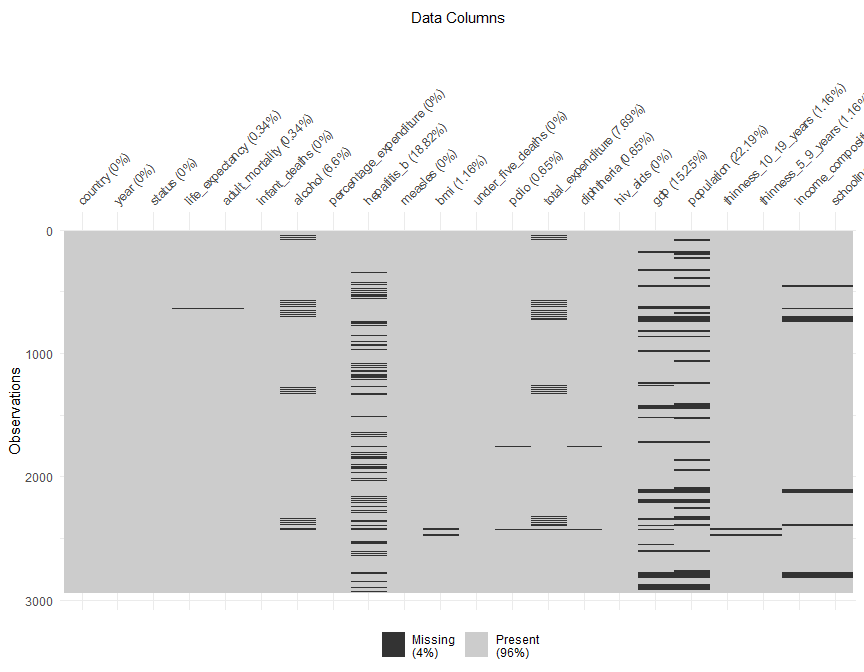
## Exploratory Data Analysis

## Importing and Cleaning the Data

########### Loading the data  
life\_ex\_DF <- read.csv("https://raw.githubusercontent.com/kristxh/Applied\_Stats\_Project1/master/datasets\_12603\_17232\_Life\_Expectancy\_Data.csv")  
  
# cleaning up column names  
colnames(life\_ex\_DF) <- tolower(colnames(life\_ex\_DF))  
colnames(life\_ex\_DF) <- gsub("\\.\\.",".",colnames(life\_ex\_DF))  
colnames(life\_ex\_DF) <- gsub("\\.","\_",colnames(life\_ex\_DF))  
names(life\_ex\_DF)[19] <- "thinness\_10\_19\_years"  
  
# converting status and year to factor columns  
life\_ex\_DF$status <- as.factor(life\_ex\_DF$status)

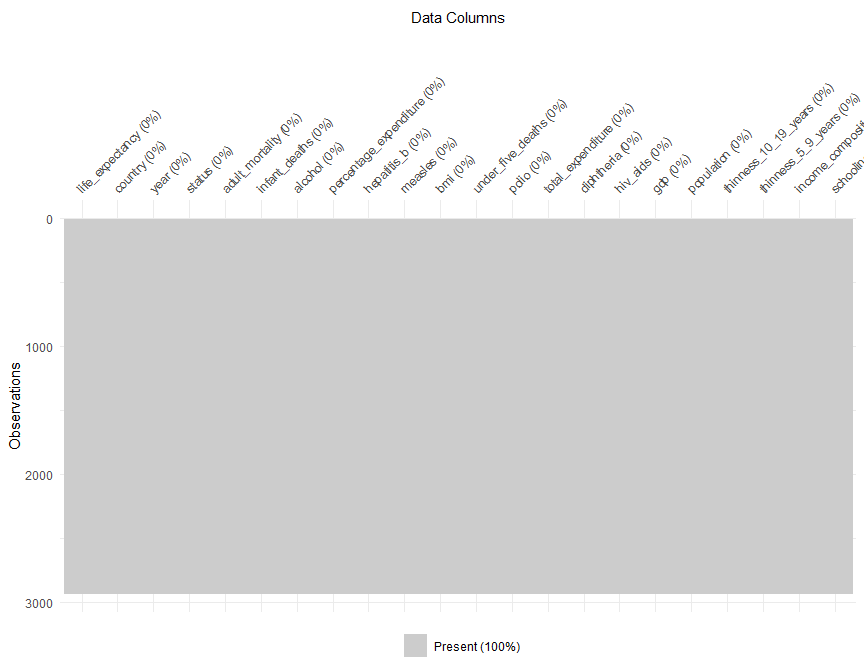
## Checking for missing data

# checking for missing values in the data  
vis\_miss(life\_ex\_DF) + xlab("Data Columns")



## Fix missing data

# sorting the columns so that life\_expectancy column is first  
life\_ex\_DF <- life\_ex\_DF %>% select("life\_expectancy", everything())  
  
factorcols <- c("status","country","year")  
allcols <- colnames(life\_ex\_DF)  
  
# selecting only numerical columns for correlation  
corvars <- allcols[allcols %notin% factorcols]  
  
  
# impute with median missing values per country  
life\_ex\_median\_DF <- as.data.frame(life\_ex\_DF %>%  
 group\_by(country) %>%  
 mutate\_at(corvars, na.aggregate, FUN=median))  
  
  
# dropping rows where the country has a single entry/record in the dataset and their records are mostly incomplete  
life\_ex\_median\_DF <- life\_ex\_median\_DF[!is.na(life\_ex\_median\_DF$life\_expectancy), ]  
  
  
# imputing the median by country still left a few nulls throughout. We will then impute the rest of the missing values as the mean of status (developing, developed, by year)  
life\_ex\_cleaned\_DF <- as.data.frame(life\_ex\_median\_DF %>%  
 group\_by(status,year) %>%  
 mutate\_at(corvars, na.aggregate, FUN=median))  
  
# checking for missing values in the data  
vis\_miss(life\_ex\_cleaned\_DF) + xlab("Data Columns")



## Summary of cleaned up data

summary(life\_ex\_cleaned\_DF)

## life\_expectancy country year status   
## Min. :36.30 Length:2928 Min. :2000 Developed : 512   
## 1st Qu.:63.10 Class :character 1st Qu.:2004 Developing:2416   
## Median :72.10 Mode :character Median :2008   
## Mean :69.22 Mean :2008   
## 3rd Qu.:75.70 3rd Qu.:2011   
## Max. :89.00 Max. :2015   
## adult\_mortality infant\_deaths alcohol percentage\_expenditure  
## Min. : 1.0 Min. : 0.00 Min. : 0.0100 Min. : 0.000   
## 1st Qu.: 74.0 1st Qu.: 0.00 1st Qu.: 0.9475 1st Qu.: 4.854   
## Median :144.0 Median : 3.00 Median : 3.7350 Median : 65.611   
## Mean :164.8 Mean : 30.41 Mean : 4.6096 Mean : 740.321   
## 3rd Qu.:228.0 3rd Qu.: 22.00 3rd Qu.: 7.6850 3rd Qu.: 442.614   
## Max. :723.0 Max. :1800.00 Max. :17.8700 Max. :19479.912   
## hepatitis\_b measles bmi under\_five\_deaths  
## Min. : 1.00 Min. : 0.0 Min. : 1.0 Min. : 0.00   
## 1st Qu.:75.00 1st Qu.: 0.0 1st Qu.:19.4 1st Qu.: 0.00   
## Median :91.00 Median : 17.0 Median :43.0 Median : 4.00   
## Mean :80.23 Mean : 2427.9 Mean :38.2 Mean : 42.18   
## 3rd Qu.:96.00 3rd Qu.: 362.2 3rd Qu.:56.1 3rd Qu.: 28.00   
## Max. :99.00 Max. :212183.0 Max. :77.6 Max. :2500.00   
## polio total\_expenditure diphtheria hiv\_aids   
## Min. : 3.00 Min. : 0.370 Min. : 2.00 Min. : 0.100   
## 1st Qu.:77.75 1st Qu.: 4.260 1st Qu.:78.00 1st Qu.: 0.100   
## Median :93.00 Median : 5.710 Median :93.00 Median : 0.100   
## Mean :82.41 Mean : 5.906 Mean :82.19 Mean : 1.748   
## 3rd Qu.:97.00 3rd Qu.: 7.452 3rd Qu.:97.00 3rd Qu.: 0.800   
## Max. :99.00 Max. :17.600 Max. :99.00 Max. :50.600   
## gdp population thinness\_10\_19\_years  
## Min. : 1.68 Min. :3.400e+01 Min. : 0.100   
## 1st Qu.: 560.98 1st Qu.:4.181e+05 1st Qu.: 1.600   
## Median : 1633.12 Median :1.397e+06 Median : 3.400   
## Mean : 6848.77 Mean :1.028e+07 Mean : 4.849   
## 3rd Qu.: 5427.25 3rd Qu.:4.593e+06 3rd Qu.: 7.100   
## Max. :119172.74 Max. :1.294e+09 Max. :27.700   
## thinness\_5\_9\_years income\_composition\_of\_resources schooling   
## Min. : 0.100 Min. :0.0000 Min. : 0.00   
## 1st Qu.: 1.600 1st Qu.:0.5040 1st Qu.:10.30   
## Median : 3.400 Median :0.6750 Median :12.32   
## Mean : 4.878 Mean :0.6315 Mean :12.05   
## 3rd Qu.: 7.200 3rd Qu.:0.7802 3rd Qu.:14.30   
## Max. :28.600 Max. :0.9480 Max. :20.70

str(life\_ex\_cleaned\_DF)

## 'data.frame': 2928 obs. of 22 variables:  
## $ life\_expectancy : num 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...  
## $ country : chr "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...  
## $ year : int 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...  
## $ status : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2 2 2 ...  
## $ adult\_mortality : num 263 271 268 272 275 279 281 287 295 295 ...  
## $ infant\_deaths : num 62 64 66 69 71 74 77 80 82 84 ...  
## $ alcohol : num 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...  
## $ percentage\_expenditure : num 71.3 73.5 73.2 78.2 7.1 ...  
## $ hepatitis\_b : num 65 62 64 67 68 66 63 64 63 64 ...  
## $ measles : num 1154 492 430 2787 3013 ...  
## $ bmi : num 19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...  
## $ under\_five\_deaths : num 83 86 89 93 97 102 106 110 113 116 ...  
## $ polio : num 6 58 62 67 68 66 63 64 63 58 ...  
## $ total\_expenditure : num 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...  
## $ diphtheria : num 65 62 64 67 68 66 63 64 63 58 ...  
## $ hiv\_aids : num 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...  
## $ gdp : num 584.3 612.7 631.7 670 63.5 ...  
## $ population : num 33736494 327582 31731688 3696958 2978599 ...  
## $ thinness\_10\_19\_years : num 17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...  
## $ thinness\_5\_9\_years : num 17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...  
## $ income\_composition\_of\_resources: num 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...  
## $ schooling : num 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...

t(aggregate(life\_expectancy~status,data=life\_ex\_cleaned\_DF,summary))

## [,1] [,2]   
## status "Developed" "Developing"  
## life\_expectancy.Min. "69.90000" "36.30000"   
## life\_expectancy.1st Qu. "76.80000" "61.10000"   
## life\_expectancy.Median "79.25000" "69.00000"   
## life\_expectancy.Mean "79.19785" "67.11147"   
## life\_expectancy.3rd Qu. "81.70000" "74.00000"   
## life\_expectancy.Max. "89.00000" "89.00000"

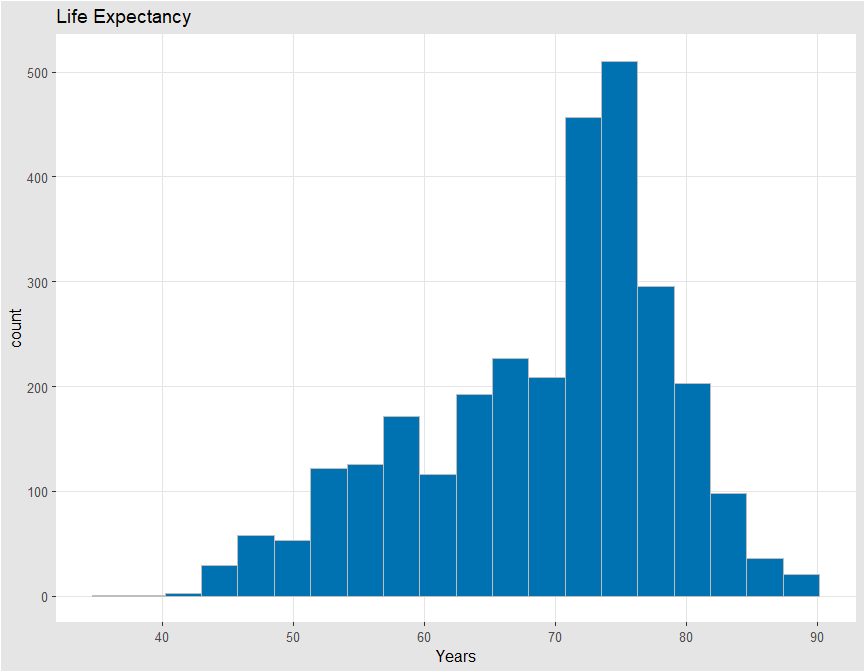
t(aggregate(life\_expectancy~status,data=life\_ex\_cleaned\_DF,sd))

## [,1] [,2]   
## status "Developed" "Developing"  
## life\_expectancy "3.930942" "9.006092"

BMI has some very weird mins and maxes. Will explore further, but this data may later show correlations that aren’t actually there because the min and max of BMI are not practical or possible.

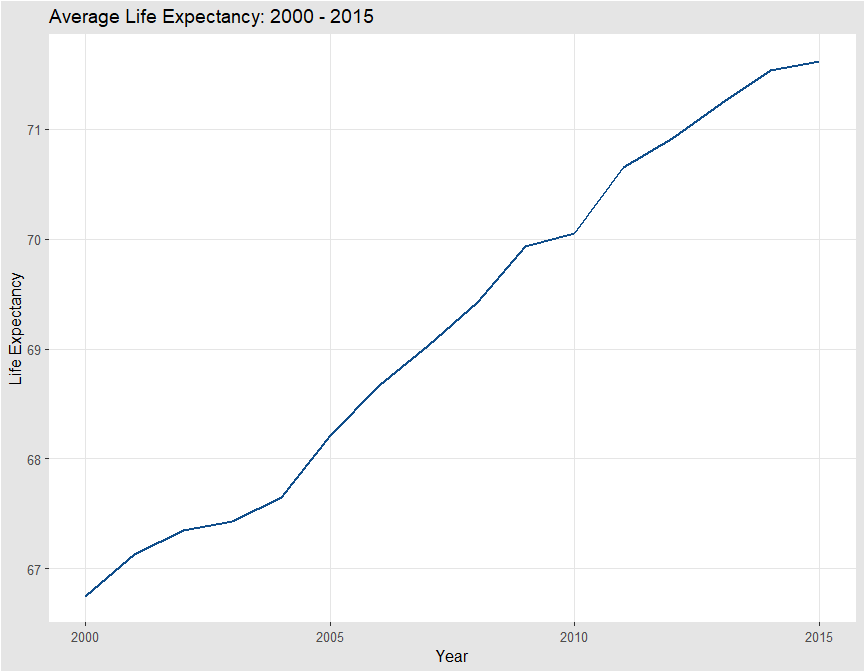
## Histogram of life\_expectancy

theme\_set(theme\_igray())  
  
le\_hist <- geom\_histogram(  
 mapping = aes(life\_expectancy),  
 data = life\_ex\_cleaned\_DF,  
 bins = 20,  
 colour = "gray",  
 fill = "#0072B2"  
)  
  
ggplot()+ le\_hist + xlab("Years") + ggtitle("Life Expectancy")



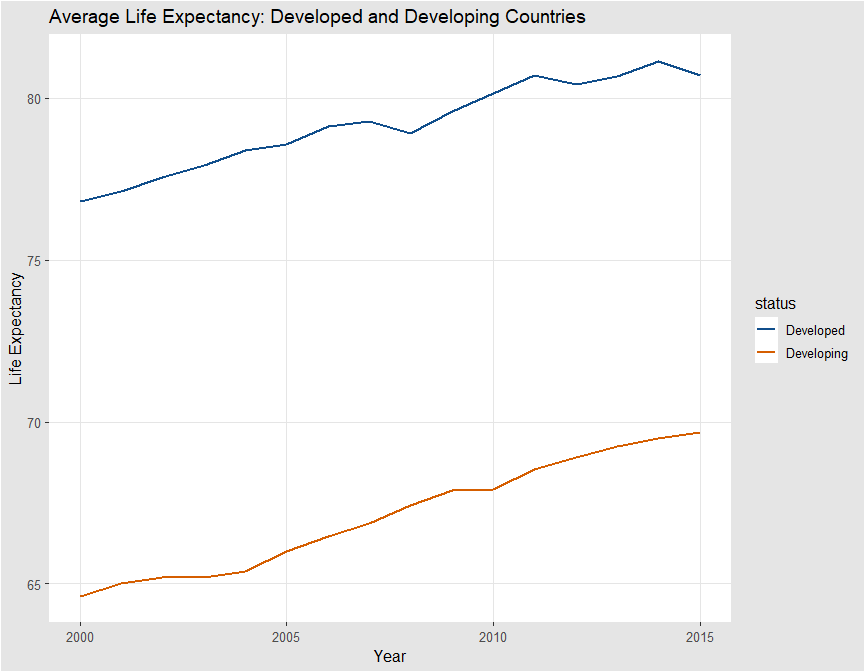
## Overall Life Expectancy

le\_by\_yr <- ddply(life\_ex\_cleaned\_DF, .(year), summarize, avg\_le=mean(life\_expectancy))  
  
le\_line <- geom\_line(  
 mapping = aes(x = year, y = avg\_le),  
 data = le\_by\_yr,  
 size = 1,  
 colour = "dodgerblue4"  
)  
  
ggplot() + le\_line + xlab("Year") + ylab("Life Expectancy") + ggtitle("Average Life Expectancy: 2000 - 2015")

 Life Expectancy is linearly related to year, and year should be treated as continuous instead of as a factor in this model. We will drop BMI for having unrealistic range of data, and no way to correct for it.

## Developed/Developing Life Expectancy

le\_by\_status <- ddply(life\_ex\_cleaned\_DF, .(year,status), summarize, avg\_le=mean(life\_expectancy))  
line\_colors <- c("dodgerblue4", "#D55E00")  
  
le\_line2 <- geom\_line(  
 mapping = aes(x = year, y = avg\_le, group=status, color=status),  
 data = le\_by\_status,  
 size = 1  
)  
  
ggplot() + le\_line2 + xlab("Year") + ylab("Life Expectancy") + ggtitle("Average Life Expectancy: Developed and Developing Countries") + scale\_colour\_manual(values=line\_colors)

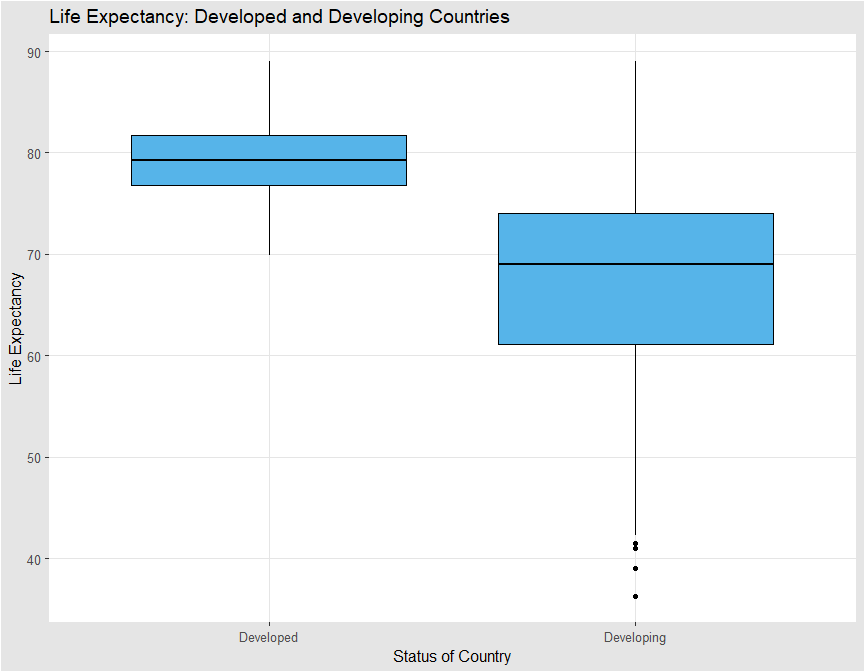


## Boxplot by status (developing, developed)

# Plot boxplot  
  
dev\_box <- geom\_boxplot(  
 mapping = aes(x=status, y=life\_expectancy),  
 data = life\_ex\_cleaned\_DF,  
 colour = "black",  
 fill = "#56B4E9",  
 groupColors=c('#999999','#E69F00')  
)

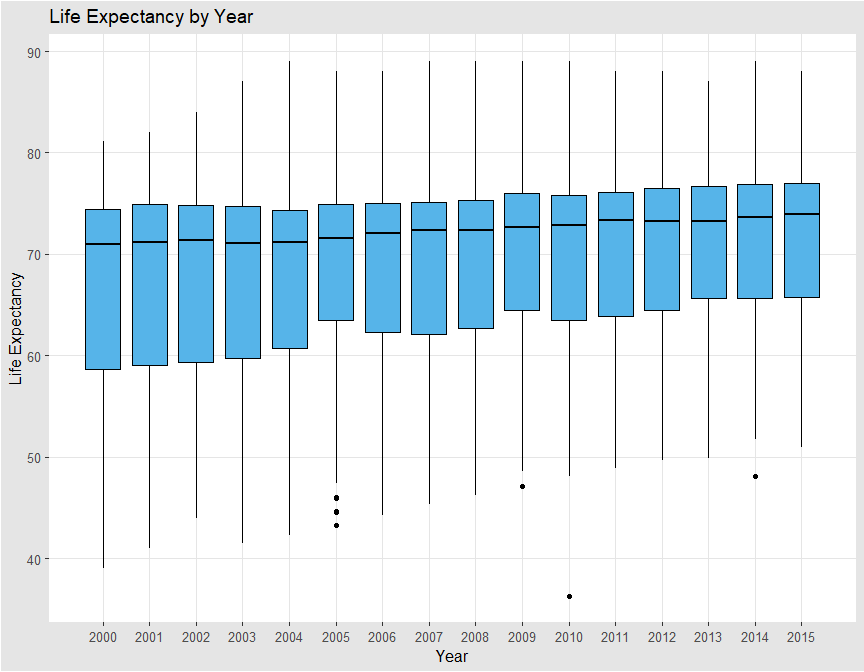
## Warning: Ignoring unknown parameters: groupColors

ggplot()+ dev\_box + xlab("Status of Country") + ylab("Life Expectancy") + ggtitle("Life Expectancy: Developed and Developing Countries ")

 There appears to be enough of a difference between life\_expectancy mean by developed and developing status. We will perform an anova and analyze signifance in the difference in means.

## Boxplots by status by year

# Plot boxplot  
  
yr\_box <- stat\_boxplot(  
 mapping = aes(x=year, y=life\_expectancy, group=year),  
 data = life\_ex\_cleaned\_DF,  
 colour = "black",  
 fill = "#56B4E9"  
 )   
   
  
ggplot()+ yr\_box + xlab("Year") + ylab("Life Expectancy") + scale\_x\_continuous(breaks=c(2000:2015)) + ggtitle("Life Expectancy by Year")



## Checking the significance of the factorial variables using ANOVA

# convert categorical variables to factors  
names <- c("country", "status")  
  
life\_ex\_cleaned\_factor\_DF <- life\_ex\_cleaned\_DF  
  
# set the categorical columns to factor  
for (name in names)  
{  
 life\_ex\_cleaned\_factor\_DF[,name]<-factor(life\_ex\_cleaned\_DF[,name])  
}  
  
# convert factors to numeric for factor analysis  
life\_ex\_cleaned\_asnum\_DF <- life\_ex\_cleaned\_factor\_DF[,names] %>% mutate\_all(as.numeric)  
  
# adding suffix ASNUM to numerical representation of factors columns  
colnames(life\_ex\_cleaned\_asnum\_DF) <- paste(colnames(life\_ex\_cleaned\_asnum\_DF), "ASNUM", sep = "\_")  
  
# adding numerical factor columns to base Df in order to perform stepwise for variable selection  
life\_ex\_cleaned\_withnum\_DF <- cbind(life\_ex\_cleaned\_factor\_DF, life\_ex\_cleaned\_asnum\_DF)  
  
# dropping factor columns  
life\_ex\_cleaned\_withnum\_DF <- life\_ex\_cleaned\_withnum\_DF[, sapply(life\_ex\_cleaned\_withnum\_DF, class) != "factor"]  
  
# columns for anova factor analysis  
anovacols <- c(colnames(life\_ex\_cleaned\_asnum\_DF))  
  
# creating the formula to pass into aov function  
anovafmla <- as.formula(paste("life\_expectancy ~ ", paste(anovacols, collapse= "+")))  
  
# passing my auto generated formula to analyze  
anova\_fit <- aov(anovafmla, data=life\_ex\_cleaned\_withnum\_DF)  
summary(anova\_fit)

## Df Sum Sq Mean Sq F value Pr(>F)   
## country\_ASNUM 1 57 57 0.816 0.367   
## status\_ASNUM 1 61910 61910 889.749 <2e-16 \*\*\*  
## Residuals 2925 203524 70   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Testing the life\_expectancy mean difference by status

Equal means model significance here to test significance of mean difference of life\_expectancy between status.

H0: μdeveloped = μdeveloping  
Ha: μdeveloped ≠ μdeveloping

Reject H0 that the means are equal

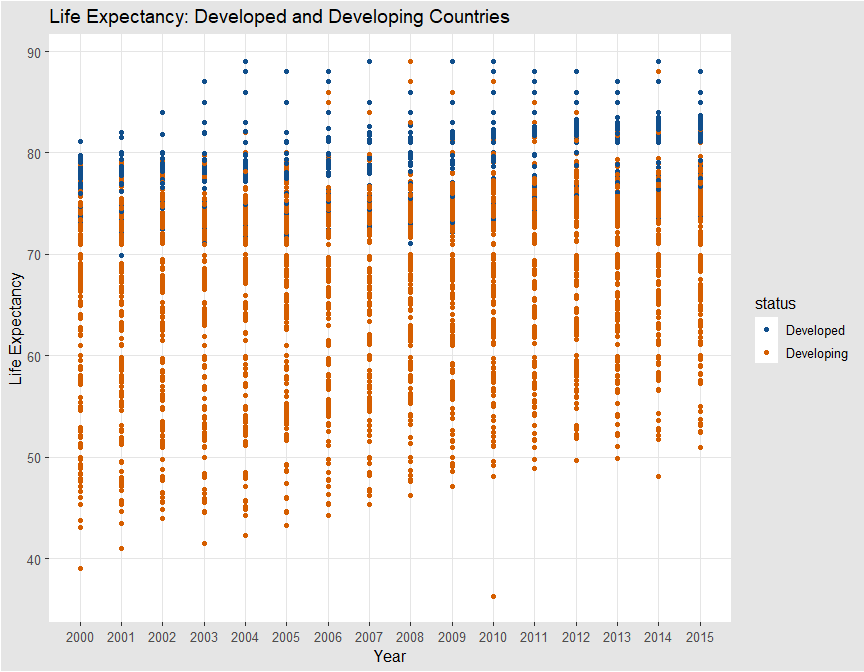
This is statistically significant (p-value <2e-16), and we will include status as a factor to consider in the model.

### Both year and status are statistically significant as categorical predictors.

Later will remove country from the data set for variable selection (p-value 0.367).

## Scatterplot boxplot

le\_scatter <- geom\_point(  
 mapping = aes(x=year, y=life\_expectancy, color=status),  
 data = life\_ex\_cleaned\_DF  
 )   
   
  
ggplot()+ le\_scatter + xlab("Year") + ylab("Life Expectancy") + scale\_x\_continuous(breaks=c(2000:2015)) + ggtitle("Life Expectancy: Developed and Developing Countries") + scale\_colour\_manual(values=line\_colors)



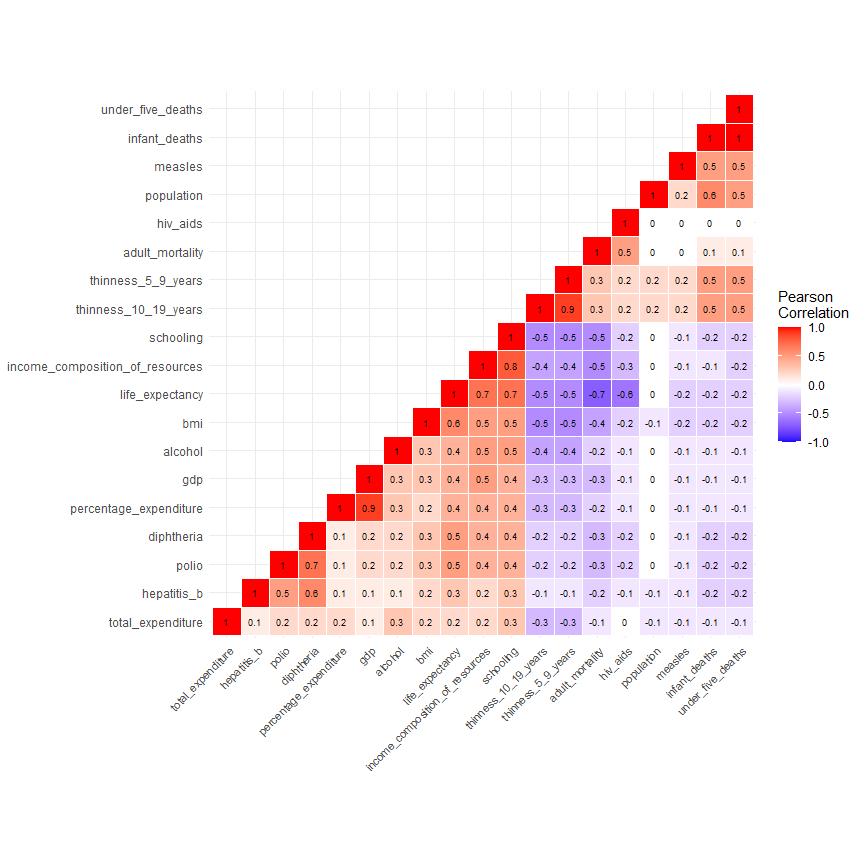
## Correlation matrix with pearson correlation coefficients

library(reshape2)

##   
## Attaching package: 'reshape2'

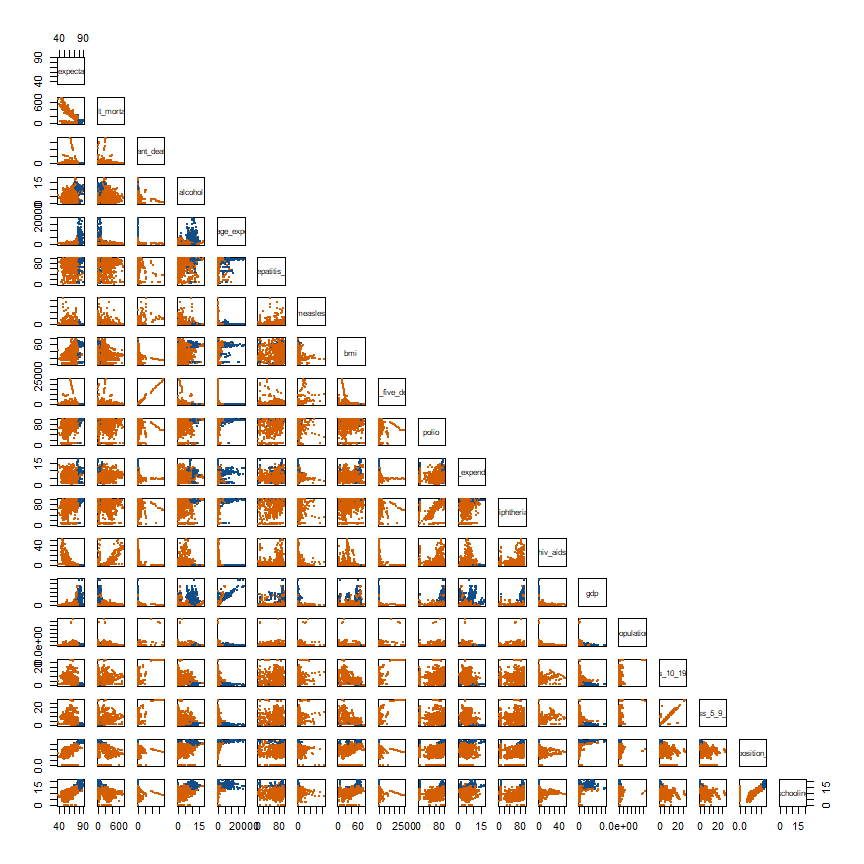
## The following object is masked from 'package:tidyr':  
##   
## smiths

corr <- round(cor(life\_ex\_cleaned\_DF[,corvars]),1)  
melted\_cormat <- melt(corr)  
  
  
# Get lower triangle of the correlation matrix  
 get\_lower\_tri<-function(cormat){  
 cormat[lower.tri(cormat)] <- NA  
 return(cormat)  
 }  
   
  
# reorder the correlation matrix  
reorder\_cormat <- function(cormat){  
# Use correlation between variables as distance  
dd <- as.dist((1-cormat)/2)  
hc <- hclust(dd)  
cormat <-cormat[hc$order, hc$order]  
}  
  
  
# reorder the correlation matrix  
corr <- reorder\_cormat(corr)  
  
# get lower half  
lower\_half <- get\_lower\_tri(corr)  
  
# reshape the correlation matrix  
melted\_cormat <- melt(lower\_half, na.rm = TRUE)  
  
# plot it  
ggcorrmap <- ggplot(melted\_cormat, aes(Var2, Var1, fill = value))+  
 geom\_tile(color = "white")+  
 scale\_fill\_gradient2(low = "blue", high = "red", mid = "white",   
 midpoint = 0, limit = c(-1,1), space = "Lab",   
 name="Pearson\nCorrelation") +  
 theme\_minimal()+  
 theme(axis.text.x = element\_text(angle = 45, vjust = 1,   
 size = 8, hjust = 1), axis.title.x = element\_blank(), axis.title.y = element\_blank())+  
 coord\_fixed()  
  
  
ggcorrmap + geom\_text(aes(Var2, Var1, label = value), color = "black", size = 2.5)

 It appears there is a high degree of correlation between the following:  
- thinnes\_5\_9\_years & thinnes\_10\_19\_years: 0.9  
- percentage\_expenditure & gdp: 0.9  
- under\_5\_deaths & infant\_deaths: 1.0  
- schooling & income\_composition\_of\_resources: 0.8  
- population has 0 correlation with the response variable and will be removed as a predictor and is incomplete throughout the data set

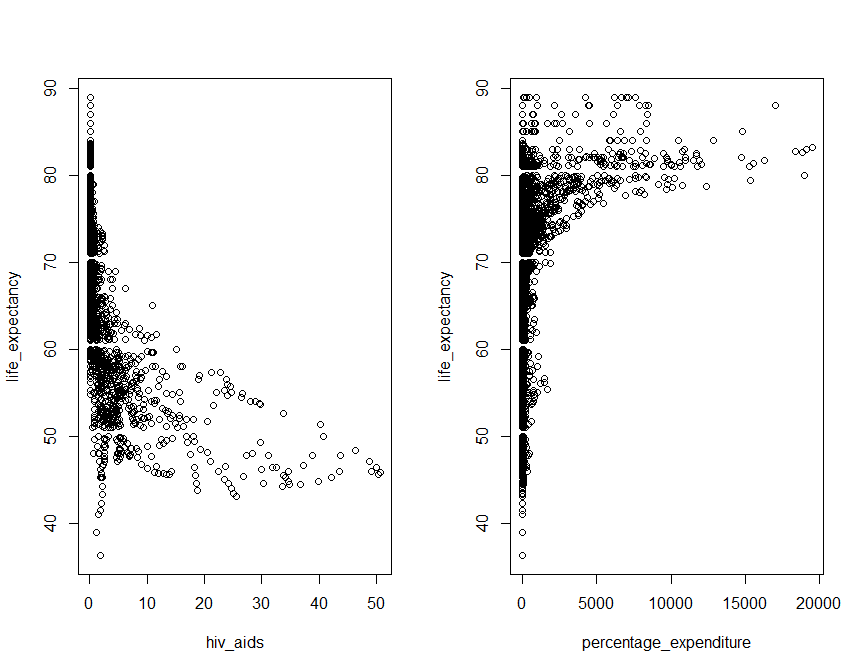
## Further examining the issue of multicolinearity

pairs(life\_ex\_cleaned\_DF[,corvars], pch = 19, cex = 0.5,  
 col = line\_colors[life\_ex\_cleaned\_DF$status],  
 upper.panel=NULL)



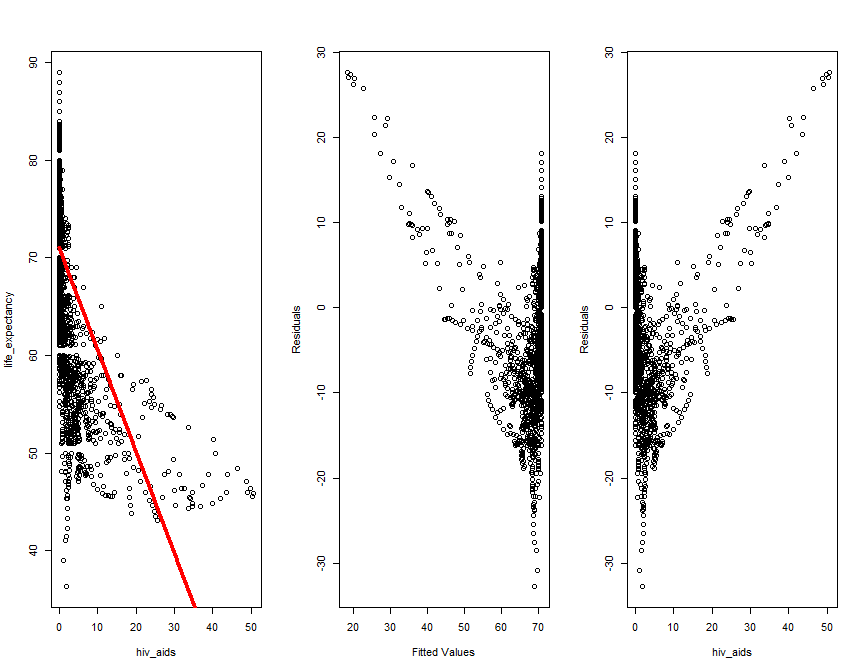
## Further exploration of hiv\_ads and percentage\_expenditure

par(mfrow=c(1,2))  
  
plot(life\_ex\_cleaned\_DF$hiv\_aids,life\_ex\_cleaned\_DF$life\_expectancy, xlab="hiv\_aids",ylab="life\_expectancy")  
  
  
plot(life\_ex\_cleaned\_DF$percentage\_expenditure,life\_ex\_cleaned\_DF$life\_expectancy, xlab="percentage\_expenditure",ylab="life\_expectancy")

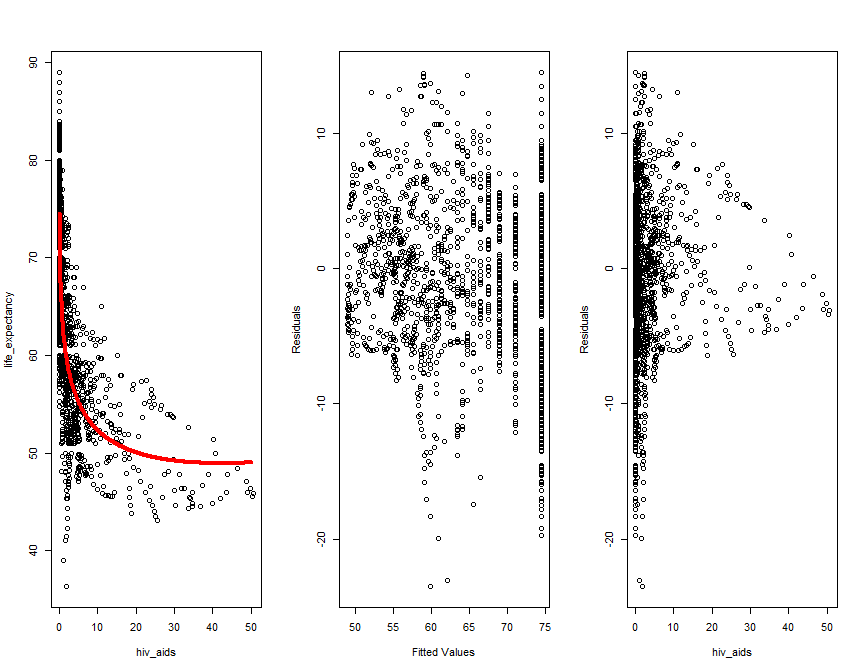
 It appears that both hiv\_aids and percentage\_expenditure have nonlinear relationships with the response variable. We will further investigate.

There also appears to be some interaction between status and percentage\_expenditure, and status and schooling. Further exploration needed.

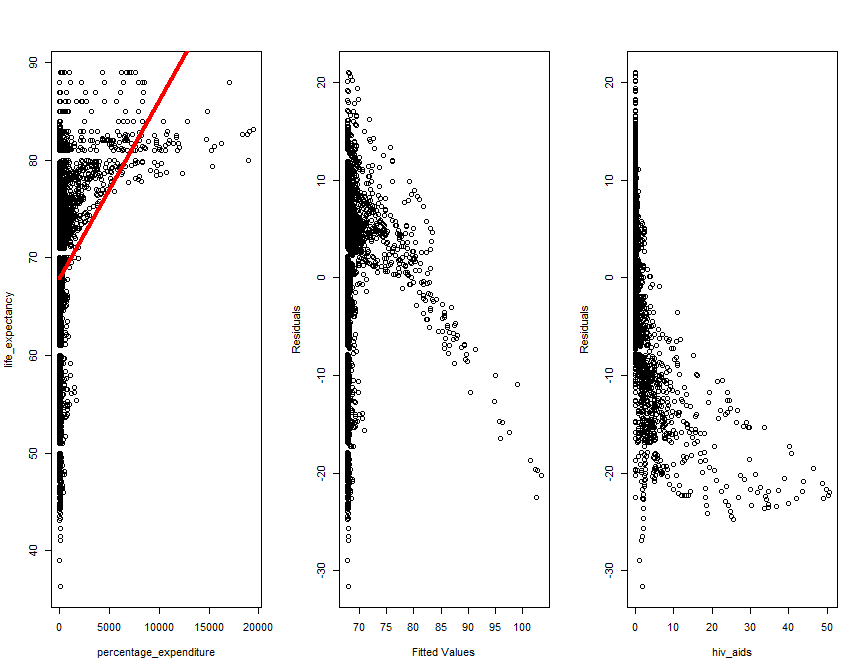
life\_expectancy <- life\_ex\_cleaned\_DF$life\_expectancy  
hiv\_aids <- life\_ex\_cleaned\_DF$hiv\_aids  
  
hiv\_aids.model <- lm(life\_expectancy~hiv\_aids)  
  
# visually inspecting the reltationship between life\_expectancy and hiv\_aids before adding complexity  
par(mfrow=c(1,3))  
plot(hiv\_aids,life\_expectancy, xlab="hiv\_aids",ylab="life\_expectancy")  
new<-data.frame(hiv\_aids=seq(0,50,.1))  
lines(seq(0,50,.1),predict(hiv\_aids.model,newdata=new),col="red",lwd=4)  
plot(hiv\_aids.model$fitted.values,hiv\_aids.model$residuals,xlab="Fitted Values",ylab="Residuals")  
plot(life\_ex\_cleaned\_DF$hiv\_aids,hiv\_aids.model$residuals,xlab="hiv\_aids",ylab="Residuals")



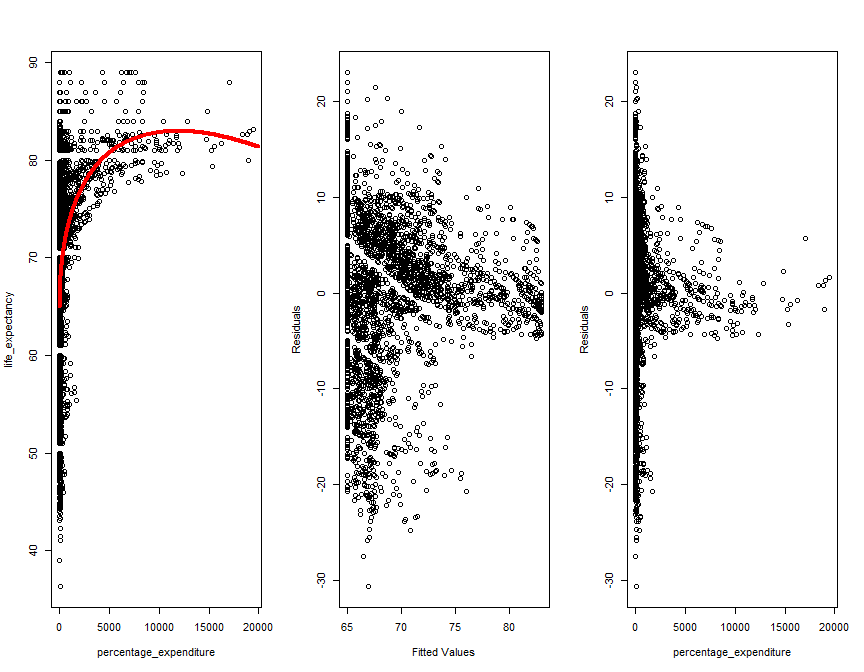
# adding complexity to relationship between life\_expectancy and hiv\_aids  
hiv\_aids.model2 <- lm(life\_expectancy~hiv\_aids+I(log(hiv\_aids)))  
  
  
# visually inspecting the reltationship between life\_expectancy and hiv\_aids after transformations  
par(mfrow=c(1,3))  
plot(hiv\_aids,life\_expectancy, xlab="hiv\_aids",ylab="life\_expectancy")  
new<-data.frame(hiv\_aids=seq(0,50,.1))  
lines(seq(0,50,.1),predict(hiv\_aids.model2,newdata=new),col="red",lwd=4)  
plot(hiv\_aids.model2$fitted.values,hiv\_aids.model2$residuals,xlab="Fitted Values",ylab="Residuals")  
plot(life\_ex\_cleaned\_DF$hiv\_aids,hiv\_aids.model2$residuals,xlab="hiv\_aids",ylab="Residuals")



percentage\_expenditure <- life\_ex\_cleaned\_DF$percentage\_expenditure  
  
perc\_exp.model <- lm(life\_expectancy~percentage\_expenditure)  
  
# visually inspecting the reltationship between life\_expectancy and percentage\_expenditure before adding complexity  
par(mfrow=c(1,3))  
plot(percentage\_expenditure,life\_expectancy, xlab="percentage\_expenditure",ylab="life\_expectancy")  
new2<-data.frame(percentage\_expenditure=seq(0,20000,100))  
lines(seq(0,20000,100),predict(perc\_exp.model,newdata=new2),col="red",lwd=4)  
plot(perc\_exp.model$fitted.values,perc\_exp.model$residuals,xlab="Fitted Values",ylab="Residuals")  
plot(life\_ex\_cleaned\_DF$hiv\_aids,perc\_exp.model$residuals,xlab="hiv\_aids",ylab="Residuals")



# adding complexity to relationship between life\_expectancy and percentage\_expenditure  
perc\_exp.model2 <- lm(life\_expectancy~percentage\_expenditure+I(sqrt(percentage\_expenditure)))  
  
  
# visually inspecting the reltationship between life\_expectancy and percentage\_expenditure after transformations  
par(mfrow=c(1,3))  
plot(percentage\_expenditure,life\_expectancy, xlab="percentage\_expenditure",ylab="life\_expectancy")  
new2<-data.frame(percentage\_expenditure=seq(0,20000,100))  
lines(seq(0,20000,100),predict(perc\_exp.model2,newdata=new2),col="red",lwd=4)  
plot(perc\_exp.model2$fitted.values,perc\_exp.model2$residuals,xlab="Fitted Values",ylab="Residuals")  
plot(life\_ex\_cleaned\_DF$percentage\_expenditure,perc\_exp.model2$residuals,xlab="percentage\_expenditure",ylab="Residuals")



## Checking colinearity with VIF and removing redundant columns

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':  
##   
## some

full.model <- lm(life\_expectancy~. ,data=life\_ex\_cleaned\_DF[,corvars])  
  
vif(full.model, data = life\_ex\_cleaned\_DF)

## adult\_mortality infant\_deaths   
## 1.740193 177.362727   
## alcohol percentage\_expenditure   
## 1.667368 4.927272   
## hepatitis\_b measles   
## 1.612879 1.376254   
## bmi under\_five\_deaths   
## 1.753643 176.345484   
## polio total\_expenditure   
## 1.980492 1.194978   
## diphtheria hiv\_aids   
## 2.310742 1.423468   
## gdp population   
## 5.333755 1.484830   
## thinness\_10\_19\_years thinness\_5\_9\_years   
## 8.752647 8.845432   
## income\_composition\_of\_resources schooling   
## 3.121290 3.642254

# removing columns and re-evaluating  
colstodelete <- c("under\_five\_deaths", "thinness\_5\_9\_years", "gdp", "population", "bmi")  
  
full.model1 <- lm(life\_expectancy~. ,data=life\_ex\_cleaned\_DF[,corvars] %>% select(-all\_of(colstodelete)))  
  
vif(full.model1, data = life\_ex\_cleaned\_DF[,corvars] %>% select(-all\_of(colstodelete)))

## adult\_mortality infant\_deaths   
## 1.709290 1.703851   
## alcohol percentage\_expenditure   
## 1.622942 1.232169   
## hepatitis\_b measles   
## 1.605898 1.352350   
## polio total\_expenditure   
## 1.969422 1.177501   
## diphtheria hiv\_aids   
## 2.272748 1.416032   
## thinness\_10\_19\_years income\_composition\_of\_resources   
## 1.787287 3.034277   
## schooling   
## 3.535584

* infant\_deaths and under\_five\_deaths have a VIF of 177.36 and 176.35 respectively, indicating a very high degree of multicolinearity between the two. We will remove under\_five\_deaths and re-evaluate.
* thinness\_5\_9\_years and thinness\_10\_19\_years have a VIF of 8.82 and 8.74 respectively, indicating a very high degree of multicolinearity between the two. We will remove thinness\_5\_9\_years and re-evaluate.
* percentage\_expenditure and gdp have a VIF of 4.92 and 5.32 respectively, indicating a very high degree of multicolinearity between the two. We will remove gdp and re-evaluate.
* schooling and income\_composition\_of\_resources are correlated to each other with a high Pearson Correlation Coefficient of .8, but have VIFs of 3.62 and 3.05 respectively. While these are highly correlated, their VIFs are relatively low. We will leave them in them in for variable selection.

## Removing columns with high colinearity or low significance

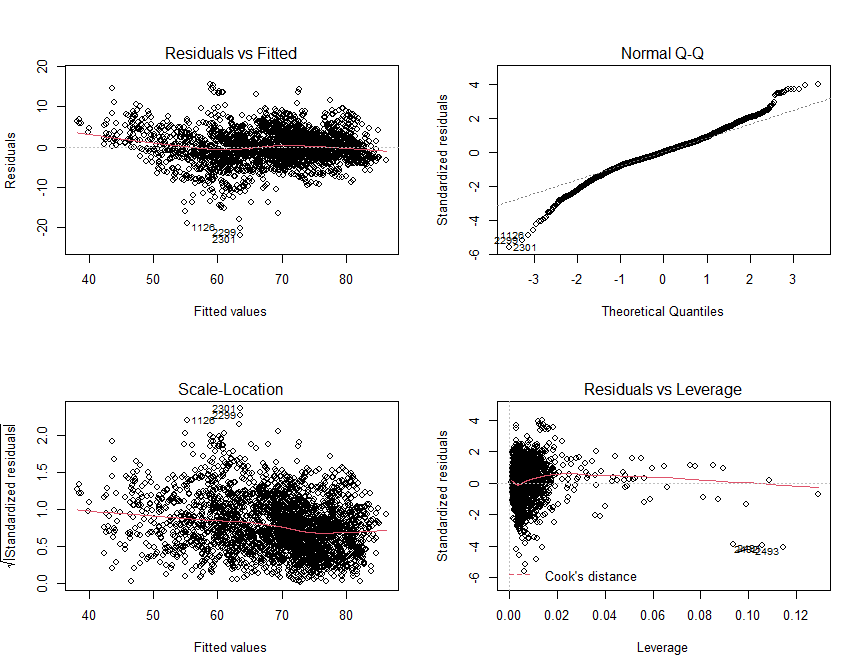
“under\_five\_deaths”, “thinness\_5\_9\_years”, “gdp”, “population” and “country” Removing “bmi” as a result of unrealistic data

# adding "country" to the list of columns to delete after its poor performance in ANOVA analysis  
colstodelete <- append(colstodelete, "country")   
Life <- life\_ex\_cleaned\_factor\_DF %>% select(-all\_of(colstodelete))

## Checking residuals and q-q plot of the full model

Including complexity for the terms hiv\_aids and percentage\_expenditure.

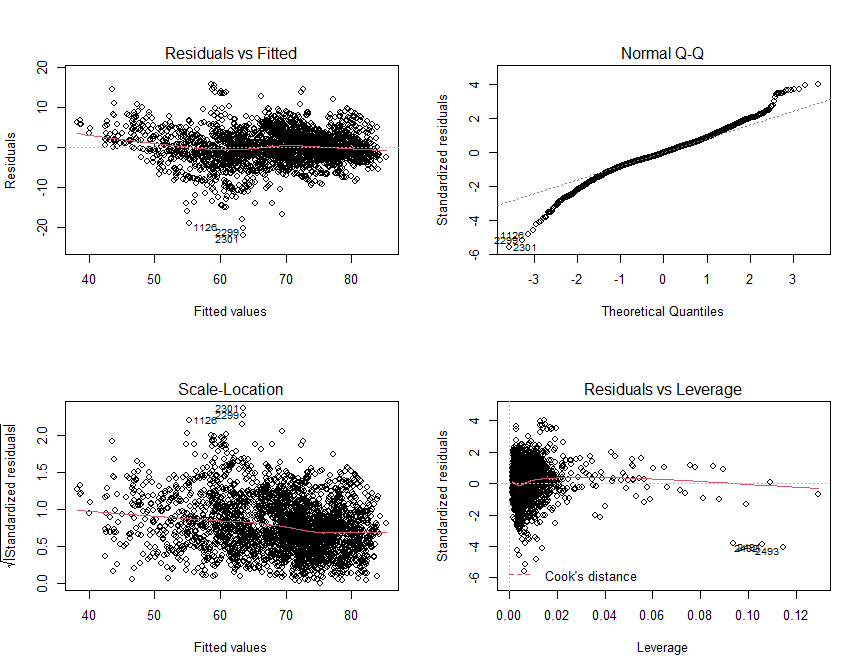
full.model <- lm(life\_expectancy~. + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) ,data=Life)  
par(mfrow=c(2,2))  
plot(full.model)

 ##################### ADD COMMENTARY HERE ABOUT CHECKING THE RESIDUALS OF THE MODEL #####################

## Adding interaction and checking significance

Visual assessment of pair plots showed potential interactions between the following:  
- status >> percentage\_expenditure  
- status >> schooling

full.modelwinter <- lm(life\_expectancy~. + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) + status:percentage\_expenditure + status:schooling, data=Life)  
  
# Re-assessing the residual plots and assumptions  
par(mfrow=c(2,2))  
plot(full.modelwinter)



summary(full.modelwinter)

##   
## Call:  
## lm(formula = life\_expectancy ~ . + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) +   
## status:percentage\_expenditure + status:schooling, data = Life)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.9650 -2.1015 -0.0345 2.1999 15.7021   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.514e+01 3.408e+01 1.618 0.105777  
## year 1.850e-03 1.708e-02 0.108 0.913750  
## statusDeveloping -3.684e+00 1.769e+00 -2.083 0.037356  
## adult\_mortality -1.649e-02 7.971e-04 -20.684 < 2e-16  
## infant\_deaths -4.656e-04 8.048e-04 -0.578 0.562971  
## alcohol 3.344e-02 2.546e-02 1.313 0.189165  
## percentage\_expenditure -2.638e-04 9.975e-05 -2.644 0.008227  
## hepatitis\_b -1.005e-03 3.789e-03 -0.265 0.790764  
## measles -2.606e-05 7.389e-06 -3.527 0.000427  
## polio 2.607e-02 4.365e-03 5.972 2.62e-09  
## total\_expenditure 5.396e-02 3.262e-02 1.654 0.098189  
## diphtheria 3.420e-02 4.651e-03 7.353 2.51e-13  
## hiv\_aids -1.174e+00 4.169e-02 -28.154 < 2e-16  
## thinness\_10\_19\_years -1.114e-01 2.218e-02 -5.025 5.35e-07  
## income\_composition\_of\_resources 5.871e+00 6.306e-01 9.311 < 2e-16  
## schooling 5.766e-01 1.079e-01 5.345 9.75e-08  
## I(hiv\_aids^2) 1.988e-02 1.115e-03 17.824 < 2e-16  
## I(sqrt(percentage\_expenditure)) 5.752e-02 1.033e-02 5.568 2.81e-08  
## statusDeveloping:percentage\_expenditure 3.991e-04 1.263e-04 3.161 0.001588  
## statusDeveloping:schooling 1.477e-01 1.131e-01 1.306 0.191736  
##   
## (Intercept)   
## year   
## statusDeveloping \*   
## adult\_mortality \*\*\*  
## infant\_deaths   
## alcohol   
## percentage\_expenditure \*\*   
## hepatitis\_b   
## measles \*\*\*  
## polio \*\*\*  
## total\_expenditure .   
## diphtheria \*\*\*  
## hiv\_aids \*\*\*  
## thinness\_10\_19\_years \*\*\*  
## income\_composition\_of\_resources \*\*\*  
## schooling \*\*\*  
## I(hiv\_aids^2) \*\*\*  
## I(sqrt(percentage\_expenditure)) \*\*\*  
## statusDeveloping:percentage\_expenditure \*\*   
## statusDeveloping:schooling   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.935 on 2908 degrees of freedom  
## Multiple R-squared: 0.8304, Adjusted R-squared: 0.8293   
## F-statistic: 749.5 on 19 and 2908 DF, p-value: < 2.2e-16

Upon further examination, we find that the following are insignificant and will be fruther analyzesd for removal from the model:  
- year - infant\_deaths - alcohol - hepatitis\_b - total\_expenditure - status:schooling interaction

## Splitting the data into test and training sets for cross validation

# Cross Validation on Train vs Test   
# Splitting 2939 observations of the data into Training and Test set in the ratio of 70:30  
set.seed(1234)  
trainIndices = sample(seq(1:dim(Life)[1]),round(.7\*dim(Life)[1]))  
train = Life[trainIndices,]  
test = Life[-trainIndices,]  
dim(train)

## [1] 2050 16

dim(test)

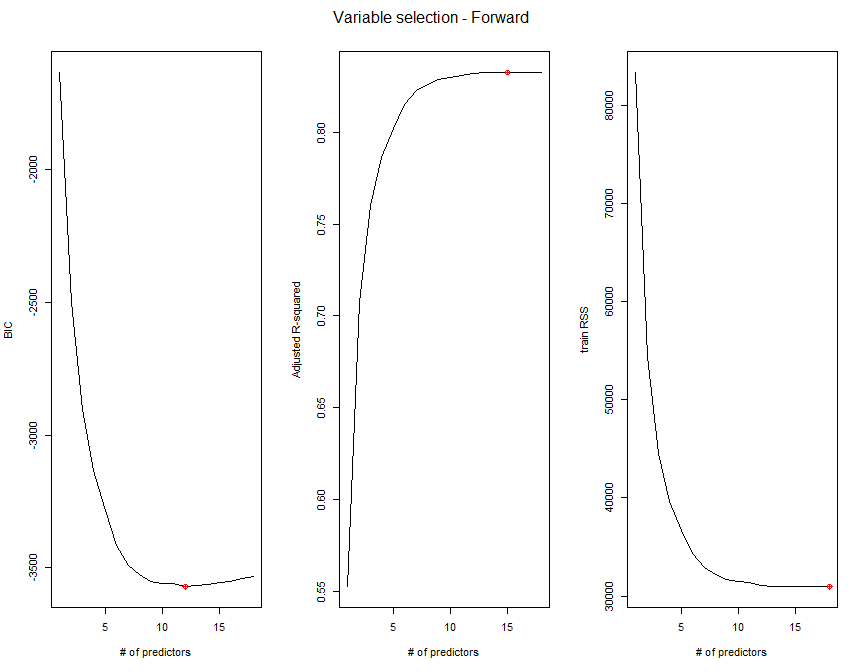
## [1] 878 16

## Creating a prediction function

predict.regsubsets = function (object , newdata ,id ,...){  
 form=as.formula (object$call [[2]])  
 mat=model.matrix(form ,newdata )  
 coefi=coef(object ,id=id)  
 xvars=names(coefi)  
 mat[,xvars]%\*%coefi  
}

## Variable selection - Forward

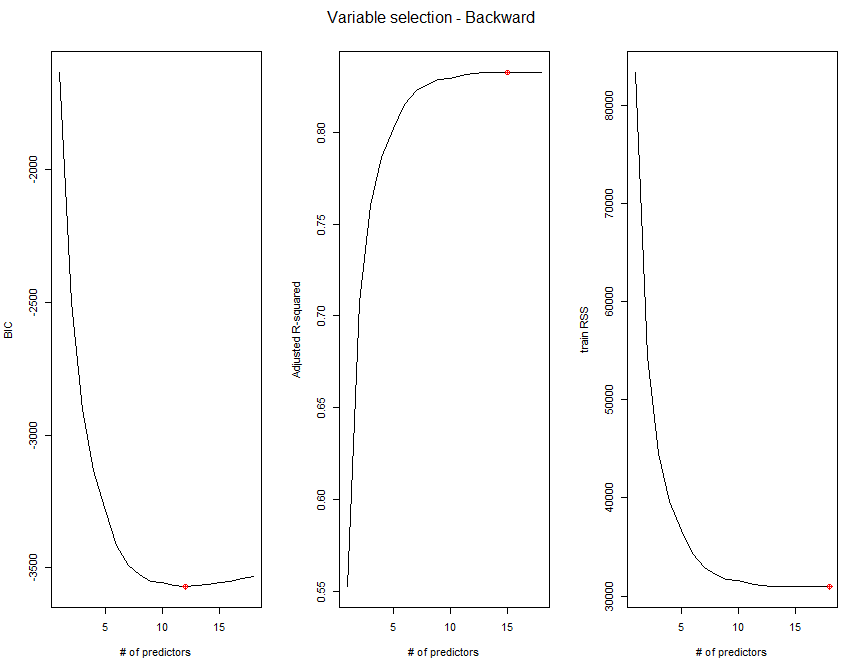
# Just the training set  
library(leaps)  
  
reg.fwd=regsubsets(life\_expectancy~. + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) + status:percentage\_expenditure, data=train, method="forward", nvmax=18)  
  
par(mfrow=c(1,3))  
  
# BIC  
bics<-summary(reg.fwd)$bic  
plot(1:18,bics,type="l",ylab="BIC",xlab="# of predictors")  
index<-which(bics==min(bics))  
points(index,bics[index],col="red",pch=10)  
  
# Adjusted R2  
adjr2<-summary(reg.fwd)$adjr2  
plot(1:18,adjr2,type="l",ylab="Adjusted R-squared",xlab="# of predictors")  
index<-which(adjr2==max(adjr2))  
points(index,adjr2[index],col="red",pch=10)  
  
# RSS  
rss<-summary(reg.fwd)$rss  
plot(1:18,rss,type="l",ylab="train RSS",xlab="# of predictors")  
index<-which(rss==min(rss))  
points(index,rss[index],col="red",pch=10)  
mtext("Variable selection - Forward", side = 3, line = -2, outer = TRUE)



# -> 12 predictors

## Variable selection - Backward

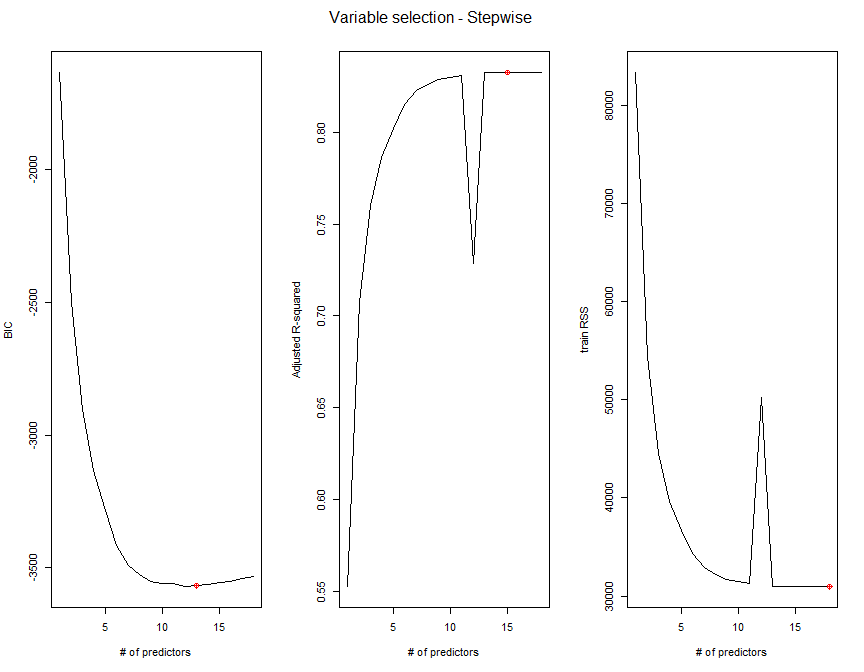
# Just the training set  
reg.bwd=regsubsets(life\_expectancy~. + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) + status:percentage\_expenditure, data=train, method="backward", nvmax=18)  
  
par(mfrow=c(1,3))  
  
# BIC  
bics2<-summary(reg.bwd)$bic  
plot(1:18,bics2,type="l",ylab="BIC",xlab="# of predictors")  
index2<-which(bics2==min(bics2))  
points(index2,bics2[index2],col="red",pch=10)  
  
# Adjusted R2  
adjr2<-summary(reg.bwd)$adjr2  
plot(1:18,adjr2,type="l",ylab="Adjusted R-squared",xlab="# of predictors")  
index.r2<-which(adjr2==max(adjr2))  
points(index.r2,adjr2[index.r2],col="red",pch=10)  
  
# RSS  
rss2<-summary(reg.bwd)$rss  
plot(1:18,rss2,type="l",ylab="train RSS",xlab="# of predictors")  
index.rss2<-which(rss2==min(rss2))  
points(index.rss2,rss[index.rss2],col="red",pch=10)  
mtext("Variable selection - Backward", side = 3, line = -2, outer = TRUE)



# -> 12 predictors

## Variable selection - Stepwise

# Just the training set  
reg.stp=regsubsets(life\_expectancy~. + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) + status:percentage\_expenditure, data=train, method="seqrep", nvmax=18)  
  
par(mfrow=c(1,3))  
  
# BIC  
bics3<-summary(reg.stp)$bic  
plot(1:18,bics,type="l",ylab="BIC",xlab="# of predictors")  
index3<-which(bics3==min(bics3))  
points(index3,bics[index3],col="red",pch=10)  
  
# Adjusted R2  
adjr3<-summary(reg.stp)$adjr  
plot(1:18,adjr3,type="l",ylab="Adjusted R-squared",xlab="# of predictors")  
index3<-which(adjr3==max(adjr3))  
points(index3,adjr3[index3],col="red",pch=10)  
  
# RSS  
rss3<-summary(reg.stp)$rss  
plot(1:18,rss3,type="l",ylab="train RSS",xlab="# of predictors")  
index3<-which(rss3==min(rss3))  
points(index3,rss[index3],col="red",pch=10)  
mtext("Variable selection - Stepwise", side = 3, line = -2, outer = TRUE)



# -> 13 predictors

## Variable selection - LASSO

library(glmnet)

## Loading required package: Matrix

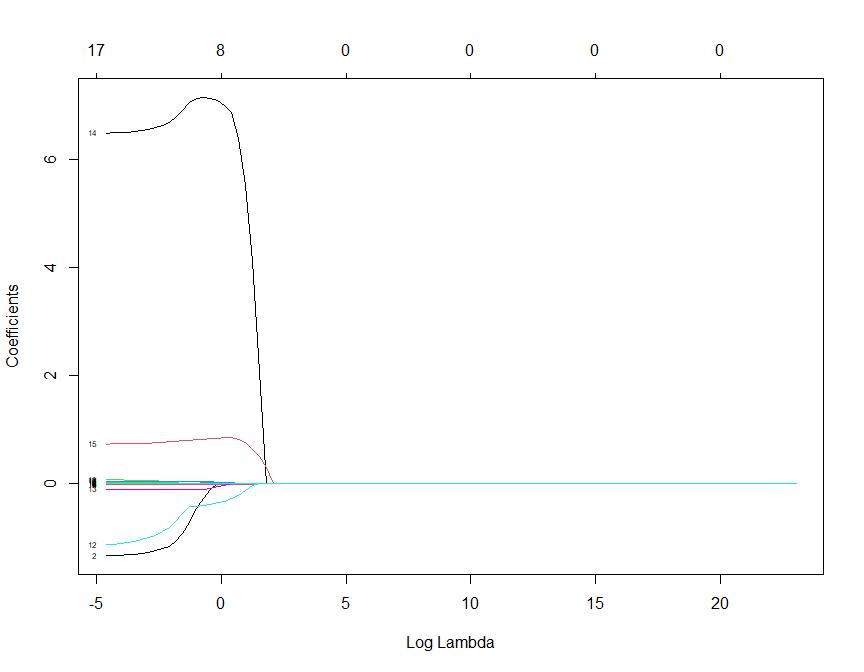
##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.0-2

x=model.matrix(life\_expectancy~. + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) + status:percentage\_expenditure, train)[,-1]  
y=train$life\_expectancy  
  
xtest<-model.matrix(life\_expectancy~. + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)) + status:percentage\_expenditure, test)[,-1]  
ytest<-test$life\_expectancy  
  
  
grid=10^seq(10,-2, length =100)  
lasso.mod=glmnet(x,y,alpha=1, lambda =grid)  
  
  
# LASSO regularization and rsquare plot  
plot(lasso.mod,xvar="lambda",label=TRUE) # shows shrinkage and variable selection

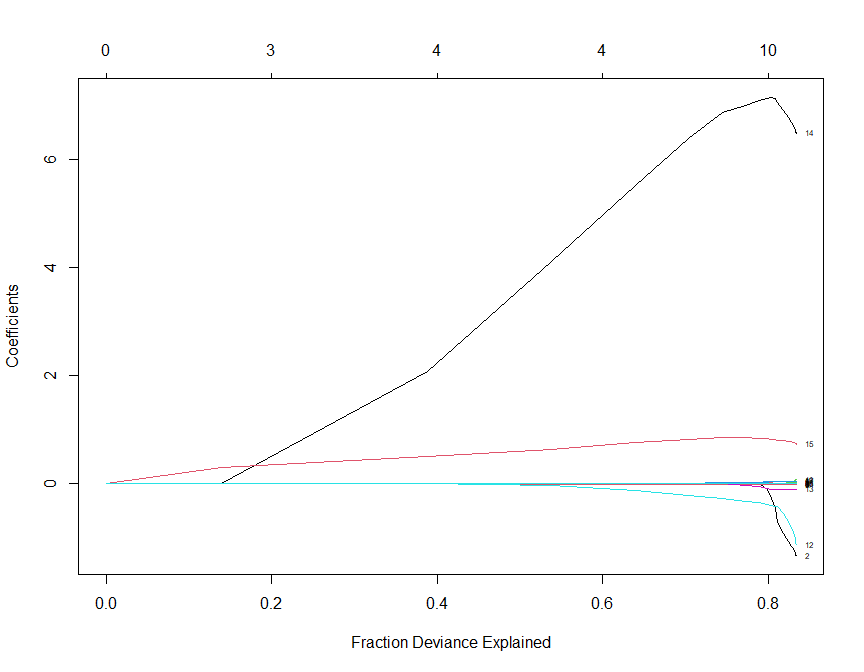
# Plot Analysis : Here in this lambda graph, each colored line represents the value taken by a different coefficient in the model. LASSO minimizes the residual sum of squares plus a shrinkage penalty of lambda multiplied by the sum of absolute values of the coefficients. This model performs variable selection while some of the coefficients are nearing exactly zero. The plot shows how many non-zero variables are in the model at the top. So at a logLambda of 0, the model has 8 variables. The earlier the variables enter the model (going from right to left), the more important and significant those variables are in this model.



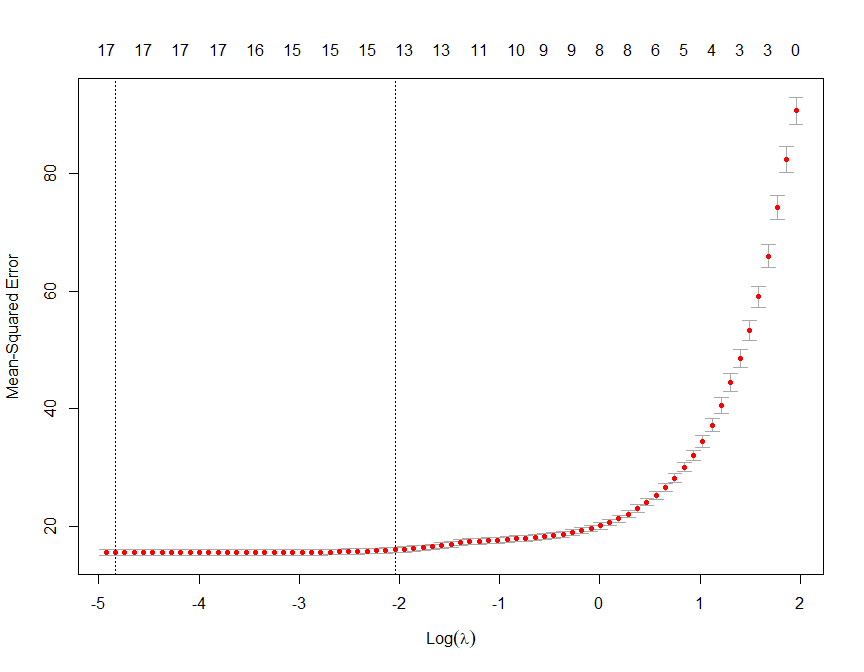
plot(lasso.mod,xvar="dev",label=TRUE) # shows rsquare change as the number of variable changes

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):  
## collapsing to unique 'x' values

## Plot analysis : There is a steady increase in coefficients as r-squared value increases from 0.2 to 0.8 as the coefficients grow. This indicates that the end of the path is slowly getting overfit. This plot indicates that about 40% of the deviance is explained by four variables whereas the full model explains about 80% of the variance.



# CV plot for LASSO  
cv.out=cv.glmnet(x,y,alpha=1) #alpha=1 performs LASSO  
plot(cv.out) #The output has 11 non-zero coefficients which shows that the function has chosen the second vertical second line on the cross-validation plot (within one standard error of the minimum) because cross validation error is measured with some variance.



# Two vertical lines.  
# 1. The one is at the minimum(the one at 17)  
# 2. The other vertical line is within one standard error of the minimum(the one at 13). We have this line because cross validation error is measured with some variance.   
  
# => The result chose the second line of 13 variables as seen in the plot(cv.out)  
  
bestlambda<-cv.out$lambda.min #Optimal penalty parameter. You can make this call visually.  
lasso.pred=predict (lasso.mod ,s=bestlambda ,newx=xtest)  
  
testMSE\_LASSO<-mean((ytest-lasso.pred)^2)  
testMSE\_LASSO

## [1] 16.22354

#Coefficients of the lasso model - Shows 13 variables (12 explanatory variables + 1 intercept), here's a coefficient function extractor that works on a cross validation object and pick the coefficient vector corresponding to the best model  
coef(lasso.mod,s=bestlambda)

## Coefficient Analysis : From the coefficient table, it is observed that the variables with negative coefficients get removed from the model as theirs have been shrunk to zero. A final model using this result is populated as below and ASE metrics was applied to be compared with other models.

## 19 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 5.564403e+01  
## year .   
## statusDeveloping -1.339117e+00  
## adult\_mortality -1.555710e-02  
## infant\_deaths -5.487916e-04  
## alcohol 3.453298e-02  
## percentage\_expenditure -1.238898e-04  
## hepatitis\_b 4.314213e-04  
## measles -2.706404e-05  
## polio 2.874677e-02  
## total\_expenditure 6.914683e-02  
## diphtheria 3.251946e-02  
## hiv\_aids -1.133162e+00  
## thinness\_10\_19\_years -1.025565e-01  
## income\_composition\_of\_resources 6.496306e+00  
## schooling 7.315119e-01  
## I(hiv\_aids^2) 1.813152e-02  
## I(sqrt(percentage\_expenditure)) 4.282561e-02  
## statusDeveloping:percentage\_expenditure 5.282712e-04

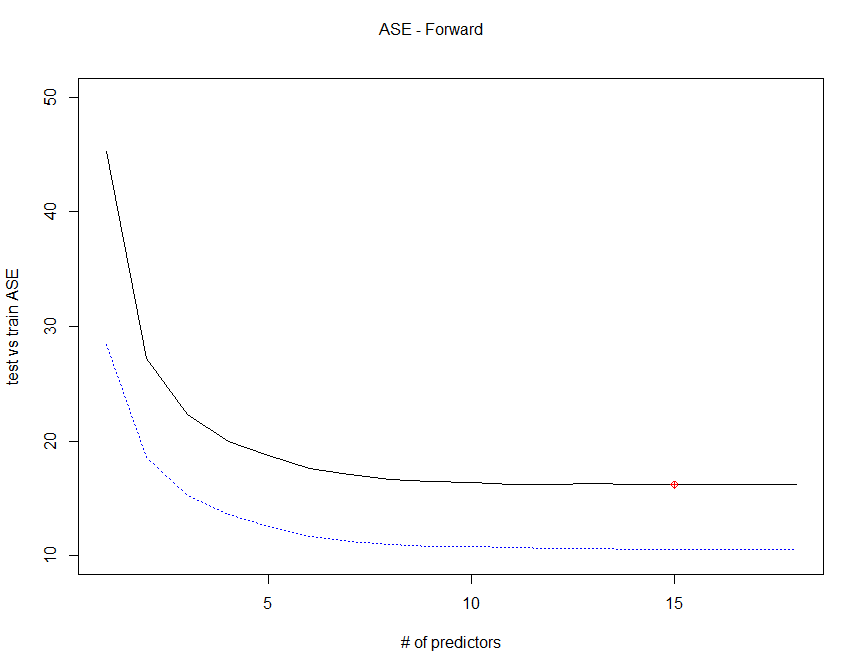
# Recreating the final LASSO model based on variable selection  
# reg.lso <- lm(life\_expectancy ~ adult\_mortality + polio + diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources + schooling + I(sqrt(percentage\_expenditure)), data=Life)  
  
reg.lso <- lm(life\_expectancy ~ adult\_mortality + measles + polio + diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources + schooling + I(sqrt(percentage\_expenditure)), data=Life)

## Cross validating the models on the test data

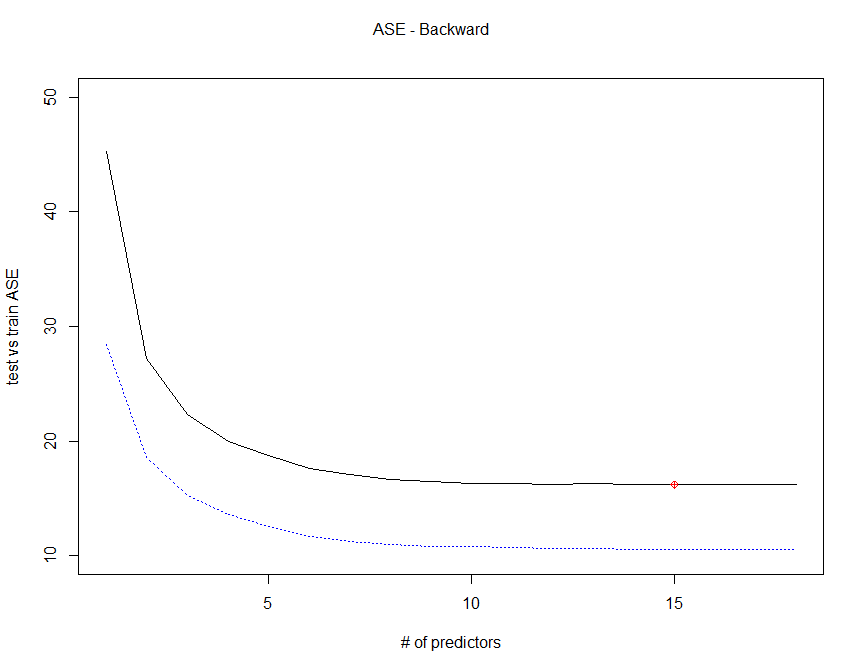
testASEfwd<-c()  
testASEbwd<-c()  
testASEstp<-c()  
  
for (i in 1:18){  
 # predicting the forward model on test set  
 predictionsfwd<-predict.regsubsets(object=reg.fwd,newdata=test,id=i)  
 testASEfwd[i]<-mean((test$life\_expectancy-predictionsfwd)^2)  
   
 # predicting the backward model on test set  
 predictionsbwd<-predict.regsubsets(object=reg.bwd,newdata=test,id=i)  
 testASEbwd[i]<-mean((test$life\_expectancy-predictionsbwd)^2)  
   
 # predicting stepwise model on test set  
 predictionsstp<-predict.regsubsets(object=reg.stp,newdata=test,id=i)  
 testASEstp[i]<-mean((test$life\_expectancy-predictionsstp)^2)  
   
}

## Plots for ASE

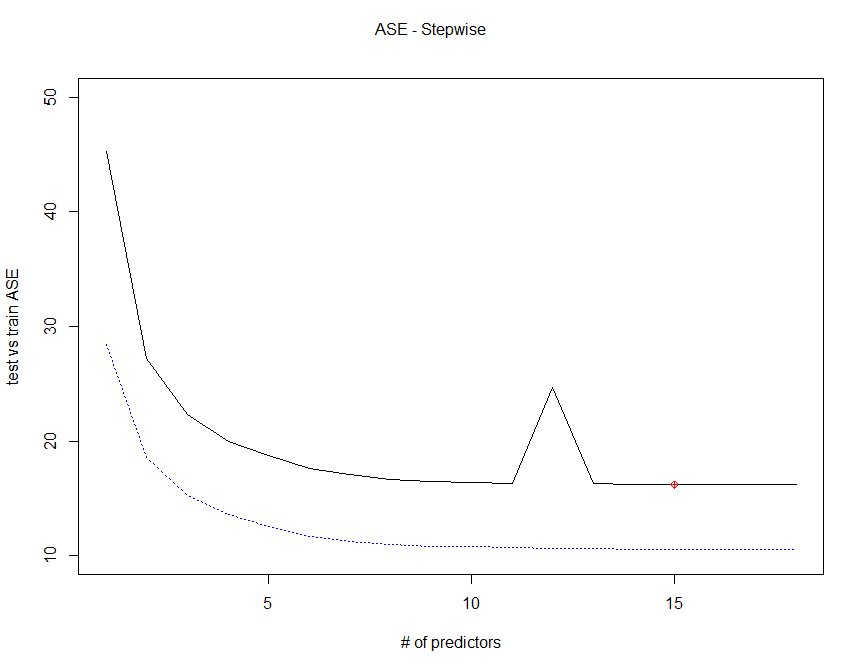
par(mfrow=c(1,1))  
plot(1:18,testASEfwd,type="l",xlab="# of predictors",ylab="test vs train ASE",ylim=c(10,50))  
index<-which(testASEfwd==min(testASEfwd))  
points(index,testASEfwd[index],col="red",pch=10)  
rss<-summary(reg.fwd)$rss  
lines(1:18,rss/2928,lty=3,col="blue") #Dividing by 2928 since ASE=RSS/sample size  
mtext("ASE - Forward", side = 3, line = -2, outer = TRUE)



par(mfrow=c(1,1))  
plot(1:18,testASEbwd,type="l",xlab="# of predictors",ylab="test vs train ASE",ylim=c(10,50))  
index<-which(testASEbwd==min(testASEbwd))  
points(index,testASEbwd[index],col="red",pch=10)  
rss<-summary(reg.bwd)$rss  
lines(1:18,rss/2928,lty=3,col="blue") #Dividing by 2928 since ASE=RSS/sample size  
mtext("ASE - Backward", side = 3, line = -2, outer = TRUE)



par(mfrow=c(1,1))  
plot(1:18,testASEstp,type="l",xlab="# of predictors",ylab="test vs train ASE",ylim=c(10,50))  
index<-which(testASEstp==min(testASEstp))  
points(index,testASEstp[index],col="red",pch=10)  
rss3<-summary(reg.stp)$rss3  
lines(1:18,rss/2928,lty=3,col="blue") #Dividing by 2928 since ASE=RSS/sample size  
mtext("ASE - Stepwise", side = 3, line = -2, outer = TRUE)



## Using forward backward, and stepwise to select max number of variables

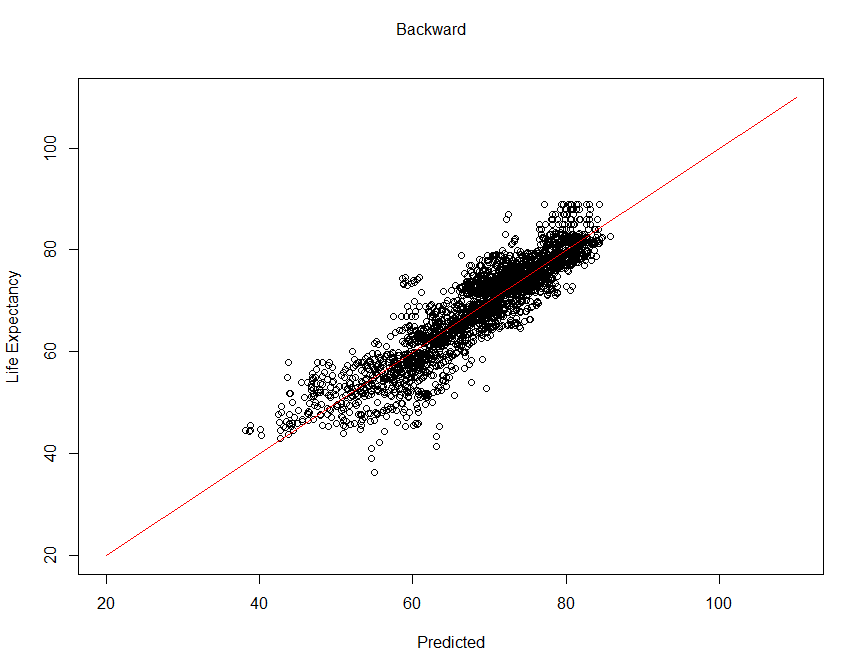
# the forward and backward models are exactly the same, we will proceed in analyzing only the bakward model  
  
  
# final variable selection for backward variable selection - the forward and backwards method selected the same model in this case  
reg.finalbwd=regsubsets(life\_expectancy~. ,data=Life,method="backward",nvmax=12)  
coef(reg.finalbwd,10)

## (Intercept) statusDeveloping   
## 5.502614e+01 -1.049242e+00   
## adult\_mortality percentage\_expenditure   
## -2.054205e-02 3.201944e-04   
## measles polio   
## -3.469699e-05 3.131628e-02   
## diphtheria hiv\_aids   
## 4.042362e-02 -4.856260e-01   
## thinness\_10\_19\_years income\_composition\_of\_resources   
## -1.458279e-01 6.818905e+00   
## schooling   
## 8.002560e-01

final.modelbwd <- lm(life\_expectancy ~ status + adult\_mortality + percentage\_expenditure + measles + polio + diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources + schooling + status:percentage\_expenditure + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)), data=Life)  
  
summary(final.modelbwd)

##   
## Call:  
## lm(formula = life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + diphtheria + hiv\_aids + thinness\_10\_19\_years +   
## income\_composition\_of\_resources + schooling + status:percentage\_expenditure +   
## I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)), data = Life)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.5297 -2.1364 -0.0472 2.2408 15.7545   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.711e+01 6.022e-01 94.834 < 2e-16  
## statusDeveloping -1.646e+00 2.775e-01 -5.933 3.32e-09  
## adult\_mortality -1.642e-02 7.932e-04 -20.702 < 2e-16  
## percentage\_expenditure -2.620e-04 9.880e-05 -2.652 0.008035  
## measles -2.771e-05 6.558e-06 -4.225 2.46e-05  
## polio 2.648e-02 4.306e-03 6.149 8.84e-10  
## diphtheria 3.461e-02 4.280e-03 8.087 8.91e-16  
## hiv\_aids -1.164e+00 4.134e-02 -28.164 < 2e-16  
## thinness\_10\_19\_years -1.281e-01 1.951e-02 -6.565 6.16e-11  
## income\_composition\_of\_resources 5.871e+00 6.123e-01 9.590 < 2e-16  
## schooling 7.279e-01 4.071e-02 17.881 < 2e-16  
## I(hiv\_aids^2) 1.974e-02 1.110e-03 17.791 < 2e-16  
## I(sqrt(percentage\_expenditure)) 5.588e-02 1.022e-02 5.468 4.94e-08  
## statusDeveloping:percentage\_expenditure 4.458e-04 1.236e-04 3.606 0.000317  
##   
## (Intercept) \*\*\*  
## statusDeveloping \*\*\*  
## adult\_mortality \*\*\*  
## percentage\_expenditure \*\*   
## measles \*\*\*  
## polio \*\*\*  
## diphtheria \*\*\*  
## hiv\_aids \*\*\*  
## thinness\_10\_19\_years \*\*\*  
## income\_composition\_of\_resources \*\*\*  
## schooling \*\*\*  
## I(hiv\_aids^2) \*\*\*  
## I(sqrt(percentage\_expenditure)) \*\*\*  
## statusDeveloping:percentage\_expenditure \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.935 on 2914 degrees of freedom  
## Multiple R-squared: 0.83, Adjusted R-squared: 0.8293   
## F-statistic: 1095 on 13 and 2914 DF, p-value: < 2.2e-16

plot(final.modelbwd$fitted.values,Life$life\_expectancy,xlab="Predicted",ylab="Life Expectancy",xlim=c(20,110),ylim=c(20,110))  
lines(c(20,110),c(20,110),col="red")  
mtext("Backward", side = 3, line = -2, outer = TRUE)



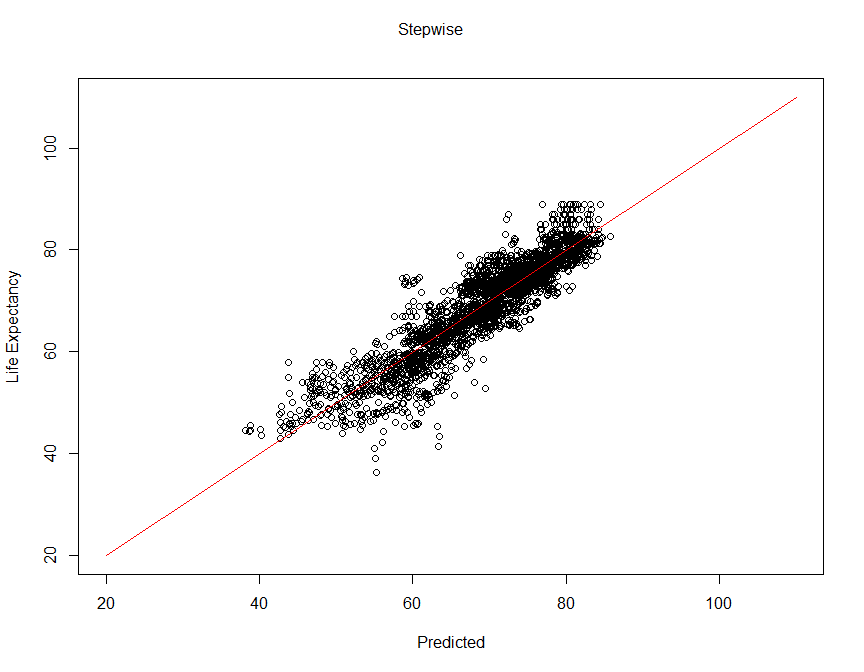
# final variable selection for stepwise variable selection  
reg.finalstp=regsubsets(life\_expectancy~.,data=Life,method="seqrep",nvmax=13)  
coef(reg.finalstp,10)

## (Intercept) statusDeveloping   
## 5.502614e+01 -1.049242e+00   
## adult\_mortality percentage\_expenditure   
## -2.054205e-02 3.201944e-04   
## measles polio   
## -3.469699e-05 3.131628e-02   
## diphtheria hiv\_aids   
## 4.042362e-02 -4.856260e-01   
## thinness\_10\_19\_years income\_composition\_of\_resources   
## -1.458279e-01 6.818905e+00   
## schooling   
## 8.002560e-01

final.modelstp <- lm(life\_expectancy ~ status + adult\_mortality + percentage\_expenditure + measles + polio + total\_expenditure + diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources + schooling + status:percentage\_expenditure + I(hiv\_aids^2) + I(sqrt(percentage\_expenditure)), data=Life)  
  
summary(final.modelstp)

##   
## Call:  
## lm(formula = life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + total\_expenditure + diphtheria + hiv\_aids +   
## thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling + status:percentage\_expenditure + I(hiv\_aids^2) +   
## I(sqrt(percentage\_expenditure)), data = Life)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.8882 -2.0941 -0.0528 2.2143 15.7951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.678e+01 6.321e-01 89.836 < 2e-16  
## statusDeveloping -1.574e+00 2.807e-01 -5.606 2.27e-08  
## adult\_mortality -1.640e-02 7.931e-04 -20.677 < 2e-16  
## percentage\_expenditure -2.705e-04 9.890e-05 -2.735 0.006268  
## measles -2.727e-05 6.562e-06 -4.155 3.34e-05  
## polio 2.639e-02 4.305e-03 6.130 9.99e-10  
## total\_expenditure 5.389e-02 3.228e-02 1.670 0.095119  
## diphtheria 3.427e-02 4.284e-03 8.001 1.77e-15  
## hiv\_aids -1.168e+00 4.140e-02 -28.222 < 2e-16  
## thinness\_10\_19\_years -1.233e-01 1.971e-02 -6.252 4.65e-10  
## income\_composition\_of\_resources 5.965e+00 6.147e-01 9.705 < 2e-16  
## schooling 7.188e-01 4.106e-02 17.507 < 2e-16  
## I(hiv\_aids^2) 1.979e-02 1.110e-03 17.835 < 2e-16  
## I(sqrt(percentage\_expenditure)) 5.676e-02 1.023e-02 5.548 3.15e-08  
## statusDeveloping:percentage\_expenditure 4.419e-04 1.236e-04 3.575 0.000356  
##   
## (Intercept) \*\*\*  
## statusDeveloping \*\*\*  
## adult\_mortality \*\*\*  
## percentage\_expenditure \*\*   
## measles \*\*\*  
## polio \*\*\*  
## total\_expenditure .   
## diphtheria \*\*\*  
## hiv\_aids \*\*\*  
## thinness\_10\_19\_years \*\*\*  
## income\_composition\_of\_resources \*\*\*  
## schooling \*\*\*  
## I(hiv\_aids^2) \*\*\*  
## I(sqrt(percentage\_expenditure)) \*\*\*  
## statusDeveloping:percentage\_expenditure \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.934 on 2913 degrees of freedom  
## Multiple R-squared: 0.8302, Adjusted R-squared: 0.8294   
## F-statistic: 1017 on 14 and 2913 DF, p-value: < 2.2e-16

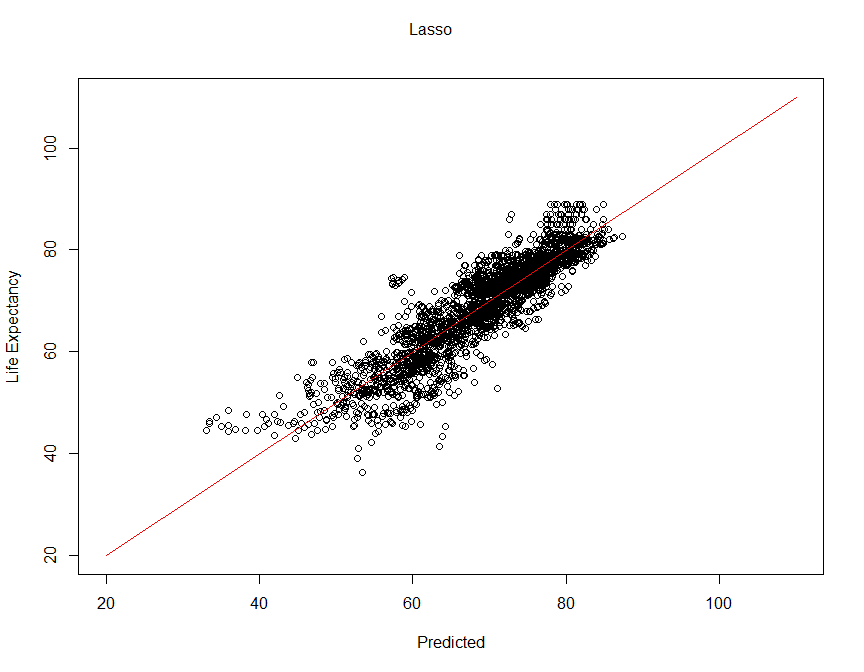
plot(final.modelstp$fitted.values,Life$life\_expectancy,xlab="Predicted",ylab="Life Expectancy",xlim=c(20,110),ylim=c(20,110) )  
lines(c(20,110),c(20,110),col="red")  
mtext("Stepwise", side = 3, line = -2, outer = TRUE)



# final variable selection for lasso variable selection  
final.modellso <- reg.lso  
  
summary(final.modellso)

##   
## Call:  
## lm(formula = life\_expectancy ~ adult\_mortality + measles + polio +   
## diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling + I(sqrt(percentage\_expenditure)), data = Life)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.0137 -2.1757 -0.0672 2.3400 17.1769   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.370e+01 5.072e-01 105.882 < 2e-16 \*\*\*  
## adult\_mortality -2.033e-02 8.103e-04 -25.085 < 2e-16 \*\*\*  
## measles -3.312e-05 6.931e-06 -4.779 1.85e-06 \*\*\*  
## polio 3.120e-02 4.548e-03 6.860 8.35e-12 \*\*\*  
## diphtheria 4.005e-02 4.518e-03 8.864 < 2e-16 \*\*\*  
## hiv\_aids -4.901e-01 1.787e-02 -27.418 < 2e-16 \*\*\*  
## thinness\_10\_19\_years -1.502e-01 2.044e-02 -7.346 2.64e-13 \*\*\*  
## income\_composition\_of\_resources 6.846e+00 6.421e-01 10.662 < 2e-16 \*\*\*  
## schooling 7.991e-01 4.223e-02 18.920 < 2e-16 \*\*\*  
## I(sqrt(percentage\_expenditure)) 4.465e-02 4.064e-03 10.986 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.166 on 2918 degrees of freedom  
## Multiple R-squared: 0.8093, Adjusted R-squared: 0.8087   
## F-statistic: 1376 on 9 and 2918 DF, p-value: < 2.2e-16

plot(final.modellso$fitted.values,Life$life\_expectancy,xlab="Predicted",ylab="Life Expectancy",xlim=c(20,110),ylim=c(20,110))  
lines(c(20,110),c(20,110),col="red")  
mtext("Lasso", side = 3, line = -2, outer = TRUE)



## ANOVA lack of fit test between backward model and lasso

anova(final.modelbwd, final.modellso)

## Analysis of Variance Table  
##   
## Model 1: life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + diphtheria + hiv\_aids + thinness\_10\_19\_years +   
## income\_composition\_of\_resources + schooling + status:percentage\_expenditure +   
## I(hiv\_aids^2) + I(sqrt(percentage\_expenditure))  
## Model 2: life\_expectancy ~ adult\_mortality + measles + polio + diphtheria +   
## hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling + I(sqrt(percentage\_expenditure))  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2914 45126   
## 2 2918 50637 -4 -5510.9 88.966 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

H0: Backward Model Fit = Lasso Model Fit  
Ha: Backward Model Fit != Lasso Model Fit

There is sufficient evidence to suggest that these models do not fit the data the same.

Reject H0

p-value: < 2.2e-16

## ANOVA lack of fit test between stepwise and backward model

anova(final.modelstp, final.modelbwd)

## Analysis of Variance Table  
##   
## Model 1: life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + total\_expenditure + diphtheria + hiv\_aids +   
## thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling + status:percentage\_expenditure + I(hiv\_aids^2) +   
## I(sqrt(percentage\_expenditure))  
## Model 2: life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + diphtheria + hiv\_aids + thinness\_10\_19\_years +   
## income\_composition\_of\_resources + schooling + status:percentage\_expenditure +   
## I(hiv\_aids^2) + I(sqrt(percentage\_expenditure))  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2913 45083   
## 2 2914 45126 -1 -43.138 2.7873 0.09512 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

H0: Stepwise Model Fit = Backward Model Fit  
Ha: Stepwise Model Fit != Backward Model Fit

There is not enough evidence to suggest that these models fit the data differently.

Fail to reject H0

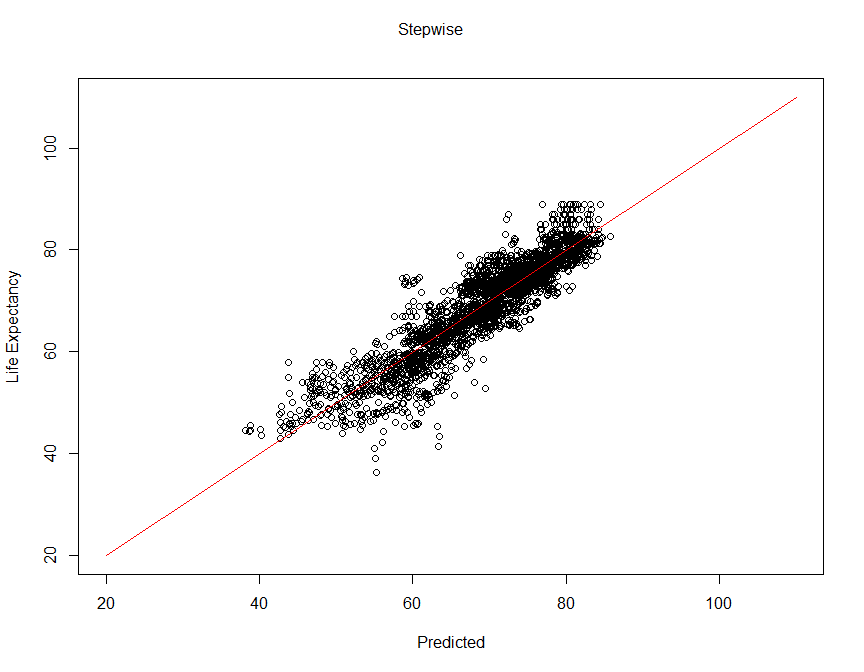
p-value: 0.095

## Using forward backward, and stepwise to select max number of variables

# final predictive model  
final.predictive <- final.modelstp  
  
  
summary(final.predictive)

##   
## Call:  
## lm(formula = life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + total\_expenditure + diphtheria + hiv\_aids +   
## thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling + status:percentage\_expenditure + I(hiv\_aids^2) +   
## I(sqrt(percentage\_expenditure)), data = Life)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.8882 -2.0941 -0.0528 2.2143 15.7951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.678e+01 6.321e-01 89.836 < 2e-16  
## statusDeveloping -1.574e+00 2.807e-01 -5.606 2.27e-08  
## adult\_mortality -1.640e-02 7.931e-04 -20.677 < 2e-16  
## percentage\_expenditure -2.705e-04 9.890e-05 -2.735 0.006268  
## measles -2.727e-05 6.562e-06 -4.155 3.34e-05  
## polio 2.639e-02 4.305e-03 6.130 9.99e-10  
## total\_expenditure 5.389e-02 3.228e-02 1.670 0.095119  
## diphtheria 3.427e-02 4.284e-03 8.001 1.77e-15  
## hiv\_aids -1.168e+00 4.140e-02 -28.222 < 2e-16  
## thinness\_10\_19\_years -1.233e-01 1.971e-02 -6.252 4.65e-10  
## income\_composition\_of\_resources 5.965e+00 6.147e-01 9.705 < 2e-16  
## schooling 7.188e-01 4.106e-02 17.507 < 2e-16  
## I(hiv\_aids^2) 1.979e-02 1.110e-03 17.835 < 2e-16  
## I(sqrt(percentage\_expenditure)) 5.676e-02 1.023e-02 5.548 3.15e-08  
## statusDeveloping:percentage\_expenditure 4.419e-04 1.236e-04 3.575 0.000356  
##   
## (Intercept) \*\*\*  
## statusDeveloping \*\*\*  
## adult\_mortality \*\*\*  
## percentage\_expenditure \*\*   
## measles \*\*\*  
## polio \*\*\*  
## total\_expenditure .   
## diphtheria \*\*\*  
## hiv\_aids \*\*\*  
## thinness\_10\_19\_years \*\*\*  
## income\_composition\_of\_resources \*\*\*  
## schooling \*\*\*  
## I(hiv\_aids^2) \*\*\*  
## I(sqrt(percentage\_expenditure)) \*\*\*  
## statusDeveloping:percentage\_expenditure \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.934 on 2913 degrees of freedom  
## Multiple R-squared: 0.8302, Adjusted R-squared: 0.8294   
## F-statistic: 1017 on 14 and 2913 DF, p-value: < 2.2e-16

plot(final.predictive$fitted.values,Life$life\_expectancy,xlab="Predicted",ylab="Life Expectancy",xlim=c(20,110),ylim=c(20,110) )  
lines(c(20,110),c(20,110),col="red")  
mtext("Stepwise", side = 3, line = -2, outer = TRUE)



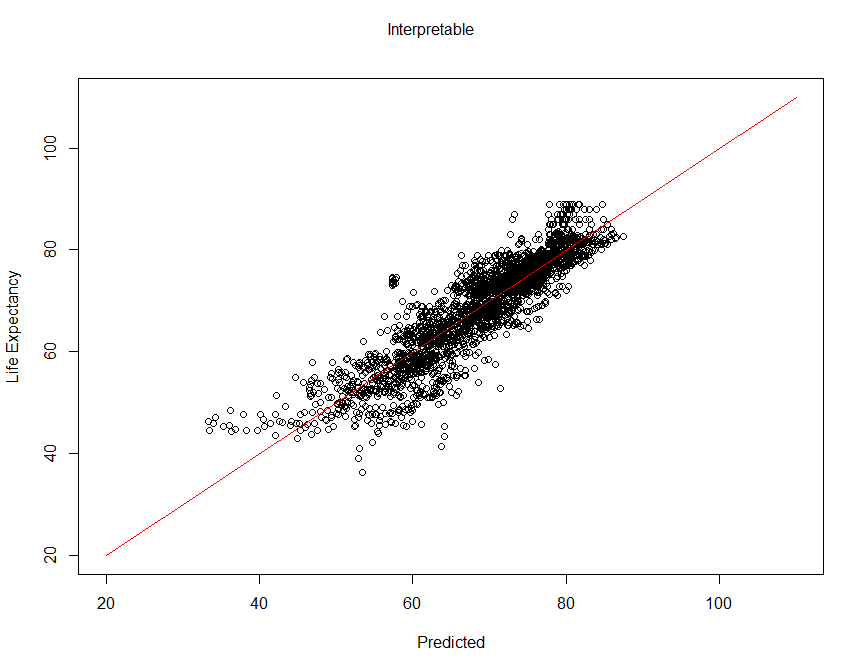
# final interpretable model  
final.interpretable <- lm(life\_expectancy ~ status + adult\_mortality + percentage\_expenditure + measles + polio + total\_expenditure + diphtheria + hiv\_aids + thinness\_10\_19\_years + income\_composition\_of\_resources + schooling, data=Life)  
  
  
summary(final.interpretable)

##   
## Call:  
## lm(formula = life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + total\_expenditure + diphtheria + hiv\_aids +   
## thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling, data = Life)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.2898 -2.0852 -0.0477 2.3129 17.2392   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.485e+01 6.388e-01 85.867 < 2e-16 \*\*\*  
## statusDeveloping -1.012e+00 2.613e-01 -3.874 0.000109 \*\*\*  
## adult\_mortality -2.054e-02 8.131e-04 -25.257 < 2e-16 \*\*\*  
## percentage\_expenditure 3.197e-04 4.463e-05 7.164 9.88e-13 \*\*\*  
## measles -3.447e-05 6.968e-06 -4.947 7.96e-07 \*\*\*  
## polio 3.128e-02 4.570e-03 6.845 9.31e-12 \*\*\*  
## total\_expenditure 2.900e-02 3.428e-02 0.846 0.397633   
## diphtheria 4.025e-02 4.545e-03 8.856 < 2e-16 \*\*\*  
## hiv\_aids -4.869e-01 1.800e-02 -27.049 < 2e-16 \*\*\*  
## thinness\_10\_19\_years -1.433e-01 2.091e-02 -6.851 8.91e-12 \*\*\*  
## income\_composition\_of\_resources 6.870e+00 6.497e-01 10.574 < 2e-16 \*\*\*  
## schooling 7.956e-01 4.319e-02 18.419 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.185 on 2916 degrees of freedom  
## Multiple R-squared: 0.8076, Adjusted R-squared: 0.8069   
## F-statistic: 1113 on 11 and 2916 DF, p-value: < 2.2e-16

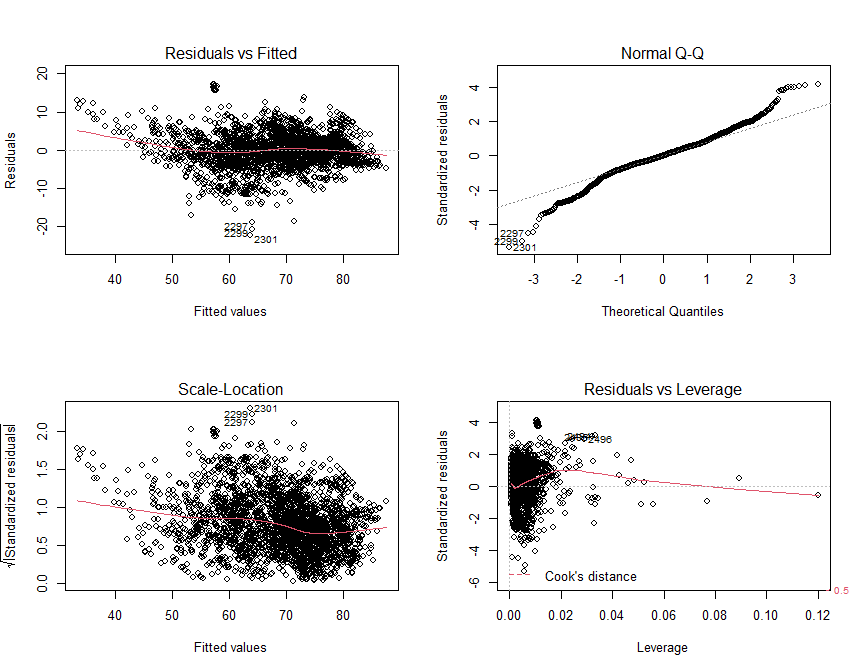
confint(final.interpretable)

## 2.5 % 97.5 %  
## (Intercept) 5.360161e+01 5.610681e+01  
## statusDeveloping -1.524759e+00 -4.999658e-01  
## adult\_mortality -2.213172e-02 -1.894300e-02  
## percentage\_expenditure 2.322167e-04 4.072333e-04  
## measles -4.813747e-05 -2.081051e-05  
## polio 2.231698e-02 4.023690e-02  
## total\_expenditure -3.821435e-02 9.621340e-02  
## diphtheria 3.133698e-02 4.916035e-02  
## hiv\_aids -5.221528e-01 -4.515681e-01  
## thinness\_10\_19\_years -1.842852e-01 -1.022717e-01  
## income\_composition\_of\_resources 5.596393e+00 8.144404e+00  
## schooling 7.109263e-01 8.803179e-01

plot(final.interpretable$fitted.values,Life$life\_expectancy,xlab="Predicted",ylab="Life Expectancy",xlim=c(20,110),ylim=c(20,110) )  
lines(c(20,110),c(20,110),col="red")  
mtext("Interpretable", side = 3, line = -2, outer = TRUE)



par(mfrow=c(2,2))  
plot(final.interpretable)



## ANOVA lack of fit test between final predictive model and interpretable model

anova(final.predictive, final.interpretable)

## Analysis of Variance Table  
##   
## Model 1: life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + total\_expenditure + diphtheria + hiv\_aids +   
## thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling + status:percentage\_expenditure + I(hiv\_aids^2) +   
## I(sqrt(percentage\_expenditure))  
## Model 2: life\_expectancy ~ status + adult\_mortality + percentage\_expenditure +   
## measles + polio + total\_expenditure + diphtheria + hiv\_aids +   
## thinness\_10\_19\_years + income\_composition\_of\_resources +   
## schooling  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2913 45083   
## 2 2916 51073 -3 -5990.4 129.02 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

H0: Predictive Model Fit = Interpretable Model Fit  
Ha: Predictive Model Fit != Interpretable Model Fit

There is sufficient evidence to suggest that these models do not fit the data the same.

Reject H0

p-value: < 2.2e-16

## Comparing the models

# Function for Root Mean Squared Error  
RMSE <- function(error) { sqrt(mean(error^2)) }  
ASE <- function(error) { mean(error^2) }  
  
RMSE(final.modelbwd$residuals)

## [1] 3.925789

RMSE(final.modelstp$residuals)

## [1] 3.923912

RMSE(final.modellso$residuals)

## [1] 4.1586

RMSE(final.interpretable$residuals)

## [1] 4.176481

ASE(final.modelbwd$residuals)

## [1] 15.41182

ASE(final.modelstp$residuals)

## [1] 15.39709

ASE(final.modellso$residuals)

## [1] 17.29395

ASE(final.interpretable$residuals)

## [1] 17.443

summary(final.modelbwd)$adj.r.squared

## [1] 0.8292704

summary(final.predictive)$adj.r.squared

## [1] 0.8293751

summary(final.modellso)$adj.r.squared

## [1] 0.8086831

summary(final.interpretable)$adj.r.squared

## [1] 0.8069019

AIC(final.modelbwd, final.predictive, final.modellso, final.interpretable)

## df AIC  
## final.modelbwd 15 16347.78  
## final.predictive 16 16346.98  
## final.modellso 11 16677.15  
## final.interpretable 13 16706.28

BIC(final.modelbwd, final.modelstp, final.modellso, final.interpretable)

## df BIC  
## final.modelbwd 15 16437.51  
## final.modelstp 16 16442.69  
## final.modellso 11 16742.95  
## final.interpretable 13 16784.04

## Model Summary

We see that isn’t enough evidence to suggest that the fit between stepwise and backward is different. We do see that both of them are indicated to be better fitting than the lasso model. Additional evidence suggests that both the stepwise and backward models returned lower diagnostic fit metrics in RMSE, BIC, and AIC, and higher Adjusted R Squared compared to the lasso model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | ASE | RMSE | Adj R^2 | AIC | BIC | DF |
| Lasso | 17.29395 | 4.1586 | 0.8086831 | 16677.15 | 16742.95 | 11 |
| Interpretable | 17.443 | 4.176481 | 0.8069019 | 16706.28 | 16784.04 | 13 |
| Backward | 15.41182 | 3.925789 | 0.8292704 | 16347.78 | 16437.51 | 15 |
| Stepwise | 15.39709 | 3.923912 | 0.8293751 | 16346.98 | 16442.69 | 16 |

We will proceed with the stepwise model as it has a slightly lower RMSE, higher Adjusted R Squared, lower AIC, and an additional degree of freedom; even though we see lower lower BIC from the backward model.

###KNN Model1 (categorical)  
#changing variable name to what I was using in my code from Neil's  
LED = Life  
  
##factorizing status into 0-1  
x = LED$status  
x = factor(x)  
LED$status <- sapply(as.integer(x), switch, "Developing" = 0, "Developed" = 1)  
  
#Data Normalization  
normalize <- function(x) {  
 return ((x - min(x)) / (max(x) - min(x))) }  
  
LED.df <- as.data.frame(lapply(LED[,-1], normalize)) #normalizes data and removes life expectancy as response variable to be predicted  
  
#splicing into test/train set  
set.seed(123)  
dat.d <- sample(1:nrow(LED.df), size=nrow(LED.df)\*0.7, replace = FALSE) #random selection for 70% of data  
  
train.LED <- LED[dat.d,] # 70% of data  
test.LED <- LED[-dat.d,] # 30% of data  
  
#creates separate data frame for 'life expectancy' feature  
train.LED\_labels <- LED[dat.d,1] #1 for life expectency  
test.LED\_labels <- LED[-dat.d,1]  
  
library(class)  
#NROW(train.LED\_labels) # number of observations indicates knn(38) sqrt of 1458  
  
ACC.model = data.frame(matrix(ncol = 1, nrow = 0))  
  
for(i in 1:10){  
 knn.model <- knn(train=train.LED, test=test.LED, cl=train.LED\_labels, k=i) #gave errors due to non-numeric variables, attempting to remove non-numeric values (country & status)  
   
 #accuracy test  
 ACC.model[i,] = 100\* sum(test.LED\_labels == knn.model)/NROW(test.LED\_labels)  
}  
  
max(ACC.model) #retrieves maximum accuracy obtained

## [1] 2.389078

which(ACC.model == max(ACC.model)) #finds the k value that gives the maximum accuracy

## [1] 1

#convert knn and test set to dataframes for analysis  
library(taRifx)

##   
## Attaching package: 'taRifx'

## The following objects are masked from 'package:dplyr':  
##   
## between, distinct, first, last

## The following object is masked from 'package:purrr':  
##   
## rep\_along

knn.df <- as.data.frame(knn.model)  
knn.df <- remove.factors(knn.df) #removes factor levels to make use of comparitive operations  
  
test.LED\_labels.df <- as.data.frame(test.LED\_labels)  
  
#find standard deviation for test set to use as range for predictions  
test.LED.sd <- sd(test.LED\_labels.df$test.LED\_labels, na.rm = TRUE)  
  
#finding if predictions is within 2 sd of test set  
  
acc\_count = data.frame(matrix(ncol = 1, nrow = 0))  
  
knn.df[1,] <= test.LED\_labels.df[1,] + (test.LED.sd/4)

## [1] TRUE

for(i in 1:nrow(knn.df)){  
if (knn.df[i,] <= test.LED\_labels.df[i,] + (2\*test.LED.sd) && knn.df[i,] >= test.LED\_labels.df[i,] - (2\*test.LED.sd)){  
 acc\_count[i,] = 1  
} else {  
 acc\_count[i,] = 0  
}}  
  
accuracy = 100\* sum(acc\_count == 1)/NROW(acc\_count)  
accuracy

## [1] 88.16837

###knn.reg model  
  
LED\_outcome <- LED %>% select(life\_expectancy)  
LED.df <- as.data.frame(lapply(LED[,-1], normalize)) #normalizes data and removes life expectancy as response variable to be predicted  
  
str(LED\_outcome)

## 'data.frame': 2928 obs. of 1 variable:  
## $ life\_expectancy: num 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...

str(LED.df)

## 'data.frame': 2928 obs. of 15 variables:  
## $ year : num 1 0.933 0.867 0.8 0.733 ...  
## $ status : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ adult\_mortality : num 0.363 0.374 0.37 0.375 0.38 ...  
## $ infant\_deaths : num 0.0344 0.0356 0.0367 0.0383 0.0394 ...  
## $ alcohol : num 0 0 0 0 0 ...  
## $ percentage\_expenditure : num 0.003659 0.003774 0.003759 0.004014 0.000364 ...  
## $ hepatitis\_b : num 0.653 0.622 0.643 0.673 0.684 ...  
## $ measles : num 0.00544 0.00232 0.00203 0.01313 0.0142 ...  
## $ polio : num 0.0312 0.5729 0.6146 0.6667 0.6771 ...  
## $ total\_expenditure : num 0.452 0.453 0.45 0.473 0.435 ...  
## $ diphtheria : num 0.649 0.619 0.639 0.67 0.68 ...  
## $ hiv\_aids : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ thinness\_10\_19\_years : num 0.62 0.63 0.638 0.645 0.656 ...  
## $ income\_composition\_of\_resources: num 0.505 0.502 0.496 0.488 0.479 ...  
## $ schooling : num 0.488 0.483 0.478 0.473 0.459 ...

library(FNN)

##   
## Attaching package: 'FNN'

## The following objects are masked from 'package:class':  
##   
## knn, knn.cv

set.seed(1234) # set the seed to make the partition reproducible  
  
# 75% of the sample size  
smp\_size <- floor(0.75 \* nrow(LED.df))  
  
train\_ind <- sample(seq\_len(nrow(LED.df)), size = smp\_size)  
  
# creating test and training sets that contain all of the predictors  
reg\_pred\_train <- LED.df[train\_ind, ]  
reg\_pred\_test <- LED.df[-train\_ind, ]  
  
LED\_outcome\_train <- LED\_outcome[train\_ind, ]  
LED\_outcome\_test <- LED\_outcome[-train\_ind, ]  
  
model\_error = data.frame(matrix(ncol = 2, nrow = 100))  
colnames(model\_error) <- c('MSE','MAE')  
  
#cycles through k = 1 to 100 storing all mean error values  
for(i in 1:100)  
{  
   
#knn predictions  
reg\_results <- knn.reg(reg\_pred\_train, reg\_pred\_test, LED\_outcome\_train, k = i)  
  
#mean square prediction error  
model\_error[i,1] = mean((LED\_outcome\_test - reg\_results$pred) ^ 2)  
  
#mean absolute error  
model\_error[i,2] = mean(abs(LED\_outcome\_test - reg\_results$pred))  
}  
  
min(model\_error$MSE) #retrieves minimum errorobtained

## [1] 8.082158

MSE\_val <- which(model\_error$MSE == min(model\_error$MSE)) #finds the k value that gives the min MSE  
min(model\_error$MAE) #retrieves minimum errorobtained

## [1] 1.731148

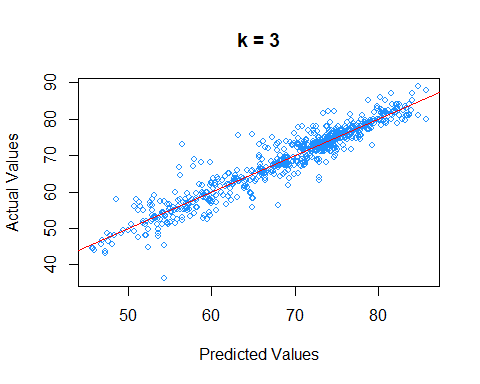
MAE\_val <- which(model\_error$MAE == min(model\_error$MAE)) #finds the k value that gives the min MAE  
  
model\_error[MSE\_val,]

## MSE MAE  
## 2 8.082158 1.787022

model\_error[MAE\_val,]

## MSE MAE  
## 1 9.448443 1.731148

#Results of K that produced lowest MSE  
reg\_results <- knn.reg(reg\_pred\_train, reg\_pred\_test, LED\_outcome\_train, k = 3)  
  
plot(LED\_outcome\_test~reg\_results$pred, cex = .8, col = "dodgerblue", main = "k = 3", xlab="Predicted Values", ylab="Actual Values")  
reg\_line = lm(LED\_outcome\_test~reg\_results$pred)  
abline(reg\_line, col = "red")

 #### Random Forest Model

# Random Forest Train/Test split  
library(MASS)

##   
## Attaching package: 'MASS'

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

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

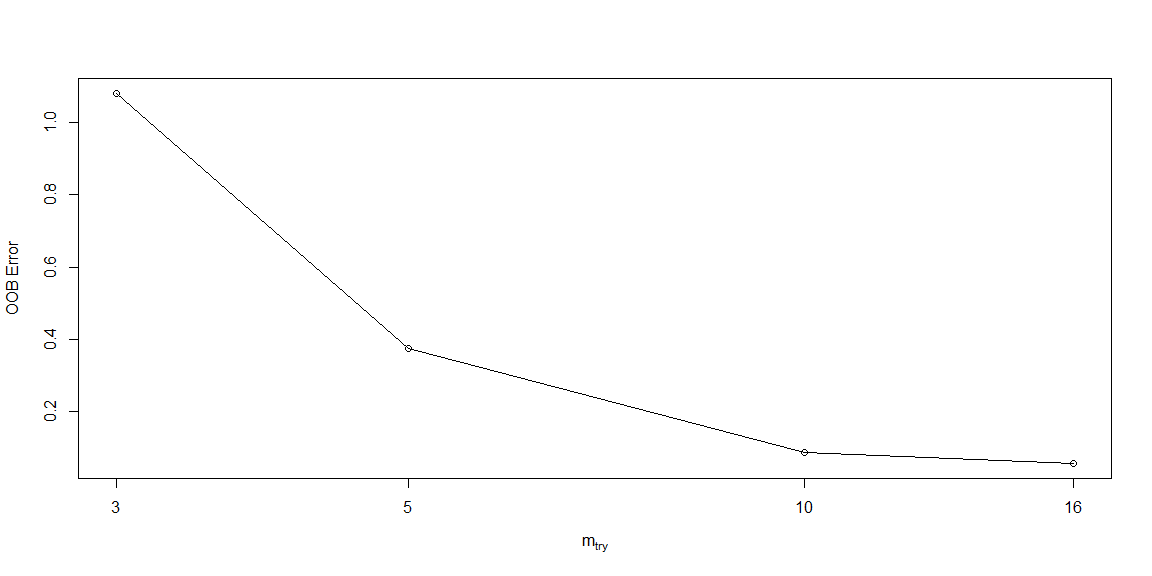
## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:gridExtra':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

set.seed(128)  
trainIndices = sample(1:nrow(Life), nrow(Life)/2)  
tree.test = Life[-trainIndices,"life\_expectancy"]  
  
# Tuning to find mtry value  
tuneRF(Life, life\_expectancy)

## mtry = 5 OOB error = 0.3751084   
## Searching left ...  
## mtry = 3 OOB error = 1.079876   
## -1.878837 0.05   
## Searching right ...  
## mtry = 10 OOB error = 0.085842   
## 0.7711542 0.05   
## mtry = 16 OOB error = 0.05721444   
## 0.3334913 0.05

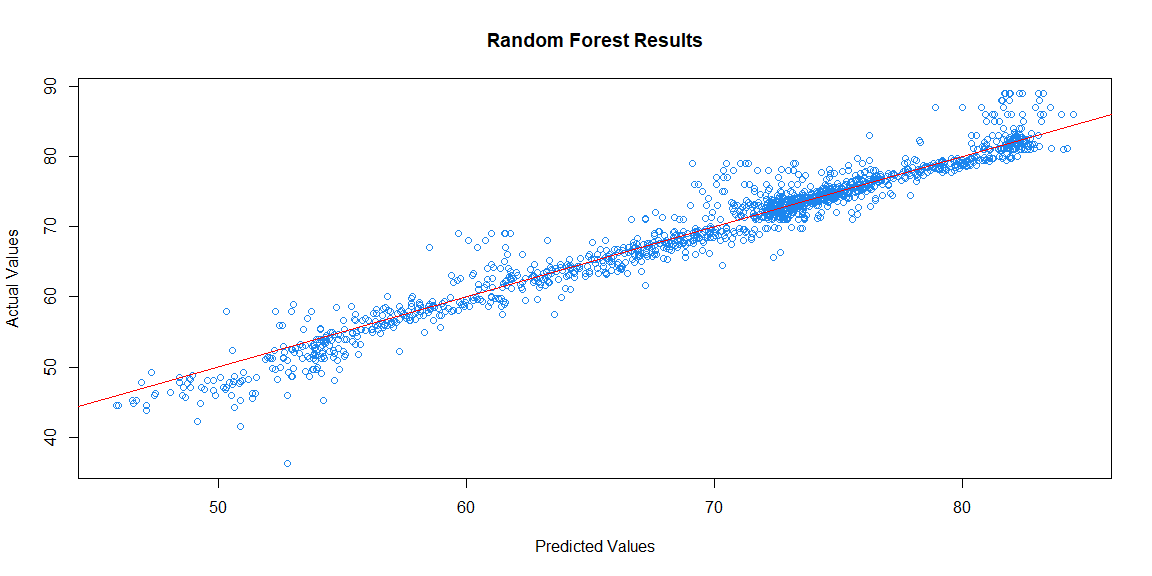


## mtry OOBError  
## 3 3 1.07987598  
## 5 5 0.37510844  
## 10 10 0.08584200  
## 16 16 0.05721444

# Prediction  
rf.le = randomForest(life\_expectancy~.,data=Life,subset=trainIndices, mtry=5, importance=TRUE, proximity=TRUE)  
yhat.rf = predict(rf.le, newdata = Life[-trainIndices,])  
  
# ASE  
mean((yhat.rf-tree.test)^2)

## [1] 4.583365

# Plot  
plot(yhat.rf,tree.test, main="Random Forest Results", xlab="Predicted Values", ylab="Actual Values", col="dodgerblue2")  
abline(0,1, col="red")



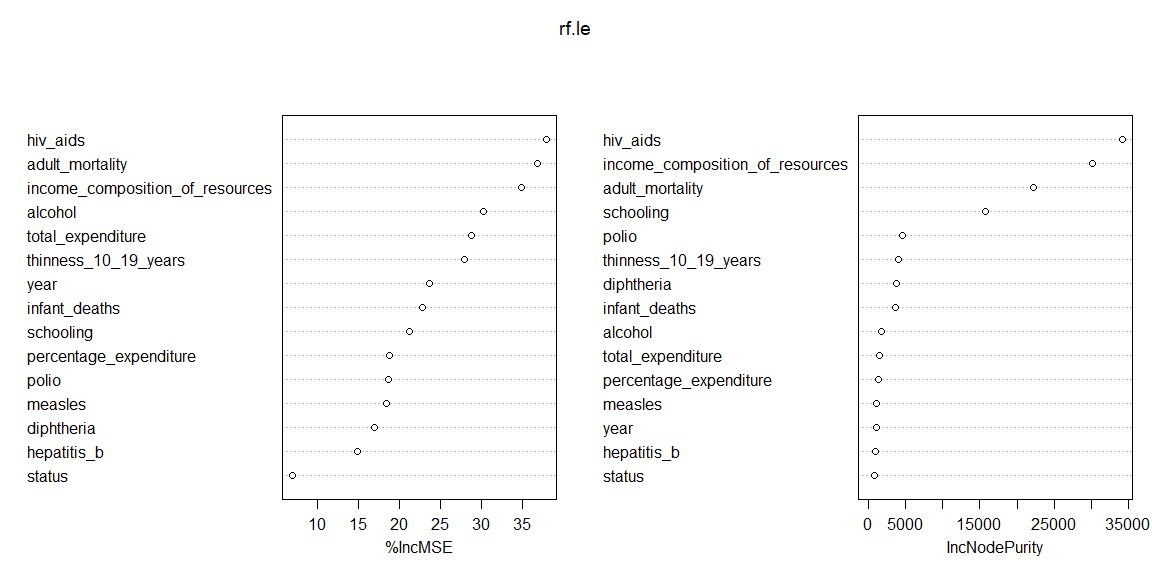
mean((yhat.rf-tree.test)^2)

## [1] 4.583365

# Importance of variables  
importance(rf.le)

## %IncMSE IncNodePurity  
## year 23.706145 1084.8799  
## status 6.980001 890.5310  
## adult\_mortality 36.767670 22215.0890  
## infant\_deaths 22.855102 3679.1394  
## alcohol 30.263816 1794.9587  
## percentage\_expenditure 18.819069 1406.3031  
## hepatitis\_b 14.878400 991.1083  
## measles 18.363278 1126.6754  
## polio 18.648251 4585.6970  
## total\_expenditure 28.806982 1442.3710  
## diphtheria 17.014871 3827.2812  
## hiv\_aids 37.886529 34194.3226  
## thinness\_10\_19\_years 27.865090 4059.6647  
## income\_composition\_of\_resources 34.785340 30103.3069  
## schooling 21.186849 15749.2618

varImpPlot(rf.le)



# Print  
print(rf.le)

##   
## Call:  
## randomForest(formula = life\_expectancy ~ ., data = Life, mtry = 5, importance = TRUE, proximity = TRUE, subset = trainIndices)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 5  
##   
## Mean of squared residuals: 4.037818  
## % Var explained: 95.38

summary(Life)

## life\_expectancy year status adult\_mortality  
## Min. :36.30 Min. :2000 Developed : 512 Min. : 1.0   
## 1st Qu.:63.10 1st Qu.:2004 Developing:2416 1st Qu.: 74.0   
## Median :72.10 Median :2008 Median :144.0   
## Mean :69.22 Mean :2008 Mean :164.8   
## 3rd Qu.:75.70 3rd Qu.:2011 3rd Qu.:228.0   
## Max. :89.00 Max. :2015 Max. :723.0   
## infant\_deaths alcohol percentage\_expenditure hepatitis\_b   
## Min. : 0.00 Min. : 0.0100 Min. : 0.000 Min. : 1.00   
## 1st Qu.: 0.00 1st Qu.: 0.9475 1st Qu.: 4.854 1st Qu.:75.00   
## Median : 3.00 Median : 3.7350 Median : 65.611 Median :91.00   
## Mean : 30.41 Mean : 4.6096 Mean : 740.321 Mean :80.23   
## 3rd Qu.: 22.00 3rd Qu.: 7.6850 3rd Qu.: 442.614 3rd Qu.:96.00   
## Max. :1800.00 Max. :17.8700 Max. :19479.912 Max. :99.00   
## measles polio total\_expenditure diphtheria   
## Min. : 0.0 Min. : 3.00 Min. : 0.370 Min. : 2.00   
## 1st Qu.: 0.0 1st Qu.:77.75 1st Qu.: 4.260 1st Qu.:78.00   
## Median : 17.0 Median :93.00 Median : 5.710 Median :93.00   
## Mean : 2427.9 Mean :82.41 Mean : 5.906 Mean :82.19   
## 3rd Qu.: 362.2 3rd Qu.:97.00 3rd Qu.: 7.452 3rd Qu.:97.00   
## Max. :212183.0 Max. :99.00 Max. :17.600 Max. :99.00   
## hiv\_aids thinness\_10\_19\_years income\_composition\_of\_resources  
## Min. : 0.100 Min. : 0.100 Min. :0.0000   
## 1st Qu.: 0.100 1st Qu.: 1.600 1st Qu.:0.5040   
## Median : 0.100 Median : 3.400 Median :0.6750   
## Mean : 1.748 Mean : 4.849 Mean :0.6315   
## 3rd Qu.: 0.800 3rd Qu.: 7.100 3rd Qu.:0.7802   
## Max. :50.600 Max. :27.700 Max. :0.9480   
## schooling   
## Min. : 0.00   
## 1st Qu.:10.30   
## Median :12.32   
## Mean :12.05   
## 3rd Qu.:14.30   
## Max. :20.70

## 