STAT 420 Final Project - National Life Expectancy prediction

xinyupi2

2019年7月22日

Introduction

About our Data

We are using a dataset called Life Expectancy (WHO). The observations of the dataset is based on different countries. And for each country, the observations contain information of life expectancy values (which will undoubtedly be used as the response variable), immunization factors, mortality factors, economic factors, social factors and other health related factors. Overally, the dimension of our data is (2939,22).

We found the dataset from @Kaggle. The website where we get this dataset is https://www.kaggle.com/kumarajarshi/life-expectancy-who (https://www.kaggle.com/kumarajarshi/life-expectancy-who). According to the author, the data was collected from WHO and United Nations website.

We all have some personal interest in biology and health science, and how long a person can live is certianly one of the most important and mysterious questions in the two fields. Thinking about this question purely from the perspective of biology theory can make it too complicated. But this dataset brings us another point of view: thinking about this question from the perspective of a statistician. So we are really interested in exploring this dataset, and we hope to build a model of life expectancy using other factors with high accuracy.

A view of the data(Explortory Data Analysis)

```
life_data = read.csv("Life Expectancy Data.csv")
# View(life_data)
```

```
## an overview of data
summary(life_data)
```

```
##                 Country          Year             Status
##  Afghanistan        :  16   Min.   :2000   Developed : 512
##  Albania            :  16   1st Qu.:2004   Developing:2426
##  Algeria            :  16   Median :2008
##  Angola             :  16   Mean   :2008
##  Antigua and Barbuda:  16   3rd Qu.:2012
##  Argentina          :  16   Max.   :2015
##  (Other)            :2842
##  Life.expectancy Adult.Mortality infant.deaths       Alcohol
##  Min.   :36.30   Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100
##  1st Qu.:63.10   1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775
##  Median :72.10   Median :144.0   Median :   3.0   Median : 3.7550
##  Mean   :69.22   Mean   :164.8   Mean   :  30.3   Mean   : 4.6029
##  3rd Qu.:75.70   3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025
##  Max.   :89.00   Max.   :723.0   Max.   :1800.0   Max.   :17.8700
##  NA's   :10      NA's   :10                       NA's   :194
##  percentage.expenditure  Hepatitis.B       Measles              BMI
##  Min.   :    0.000      Min.   : 1.00   Min.   :     0.0   Min.   : 1.00
##  1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30
##  Median :   64.913      Median :92.00   Median :    17.0   Median :43.50
##  Mean   :  738.251      Mean   :80.94   Mean   :  2419.6   Mean   :38.32
##  3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20
##  Max.   :19479.912      Max.   :99.00   Max.   :212183.0   Max.   :87.30
##                         NA's   :553                        NA's   :34
##  under.five.deaths     Polio       Total.expenditure   Diphtheria
##  Min.   :   0.00   Min.   : 3.00   Min.   : 0.370    Min.   : 2.00
##  1st Qu.:   0.00   1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00
##  Median :   4.00   Median :93.00   Median : 5.755    Median :93.00
##  Mean   :  42.04   Mean   :82.55   Mean   : 5.938    Mean   :82.32
##  3rd Qu.:  28.00   3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00
##  Max.   :2500.00   Max.   :99.00   Max.   :17.600    Max.   :99.00
##                    NA's   :19      NA's   :226       NA's   :19
##     HIV.AIDS           GDP              Population
##  Min.   : 0.100   Min.   :     1.68   Min.   :3.400e+01
##  1st Qu.: 0.100   1st Qu.:   463.94   1st Qu.:1.958e+05
##  Median : 0.100   Median :  1766.95   Median :1.387e+06
##  Mean   : 1.742   Mean   :  7483.16   Mean   :1.275e+07
##  3rd Qu.: 0.800   3rd Qu.:  5910.81   3rd Qu.:7.420e+06
##  Max.   :50.600   Max.   :119172.74   Max.   :1.294e+09
##                   NA's   :448         NA's   :652
##  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##  Min.   : 0.10        Min.   : 0.10      Min.   :0.0000
##  1st Qu.: 1.60        1st Qu.: 1.50      1st Qu.:0.4930
##  Median : 3.30        Median : 3.30      Median :0.6770
##  Mean   : 4.84        Mean   : 4.87      Mean   :0.6276
##  3rd Qu.: 7.20        3rd Qu.: 7.20      3rd Qu.:0.7790
##  Max.   :27.70        Max.   :28.60      Max.   :0.9480
##  NA's   :34           NA's   :34         NA's   :167
##    Schooling
##  Min.   : 0.00
##  1st Qu.:10.10
##  Median :12.30
##  Mean   :11.99
##  3rd Qu.:14.30
##  Max.   :20.70
##  NA's   :163
```

```
#drop the country column and this is not useful for predicting
life_data = subset(life_data, select=-c(Country))

#data overview
names(life_data)
```

```
##  [1] "Year"                            "Status"                         
##  [3] "Life.expectancy"                 "Adult.Mortality"                
##  [5] "infant.deaths"                   "Alcohol"                        
##  [7] "percentage.expenditure"          "Hepatitis.B"                    
##  [9] "Measles"                         "BMI"                            
## [11] "under.five.deaths"               "Polio"                          
## [13] "Total.expenditure"               "Diphtheria"                     
## [15] "HIV.AIDS"                        "GDP"                            
## [17] "Population"                      "thinness..1.19.years"           
## [19] "thinness.5.9.years"              "Income.composition.of.resources"
## [21] "Schooling"
```

```
head(life_data,10)
```

```
##    Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol
## 1  2015 Developing            65.0             263            62    0.01
## 2  2014 Developing            59.9             271            64    0.01
## 3  2013 Developing            59.9             268            66    0.01
## 4  2012 Developing            59.5             272            69    0.01
## 5  2011 Developing            59.2             275            71    0.01
## 6  2010 Developing            58.8             279            74    0.01
## 7  2009 Developing            58.6             281            77    0.01
## 8  2008 Developing            58.1             287            80    0.03
## 9  2007 Developing            57.5             295            82    0.02
## 10 2006 Developing            57.3             295            84    0.03
##    percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio
## 1               71.279624          65    1154 19.1                83     6
## 2               73.523582          62     492 18.6                86    58
## 3               73.219243          64     430 18.1                89    62
## 4               78.184215          67    2787 17.6                93    67
## 5                7.097109          68    3013 17.2                97    68
## 6               79.679367          66    1989 16.7               102    66
## 7               56.762217          63    2861 16.2               106    63
## 8               25.873925          64    1599 15.7               110    64
## 9               10.910156          63    1141 15.2               113    63
## 10              17.171518          64    1990 14.7               116    58
##    Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1               8.16         65      0.1 584.25921   33736494
## 2               8.18         62      0.1 612.69651     327582
## 3               8.13         64      0.1 631.74498   31731688
## 4               8.52         67      0.1 669.95900    3696958
## 5               7.87         68      0.1  63.53723    2978599
## 6               9.20         66      0.1 553.32894    2883167
## 7               9.42         63      0.1 445.89330     284331
## 8               8.33         64      0.1 373.36112    2729431
## 9               6.73         63      0.1 369.83580   26616792
## 10              7.43         58      0.1 272.56377    2589345
##    thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                  17.2               17.3                           0.479
## 2                  17.5               17.5                           0.476
## 3                  17.7               17.7                           0.470
## 4                  17.9               18.0                           0.463
## 5                  18.2               18.2                           0.454
## 6                  18.4               18.4                           0.448
## 7                  18.6               18.7                           0.434
## 8                  18.8               18.9                           0.433
## 9                  19.0               19.1                           0.415
## 10                 19.2               19.3                           0.405
##    Schooling
## 1       10.1
## 2       10.0
## 3        9.9
## 4        9.8
## 5        9.5
## 6        9.2
## 7        8.9
## 8        8.7
## 9        8.4
## 10       8.1
```

```
cat("The data has", nrow(life_data), "rows")
```

```
## The data has 2938 rows
```

```
cat("and", ncol(life_data), "columns")
```

```
## and 21 columns
```

```
##See the data types and levels
str(life_data)
```

```
## 'data.frame':    2938 obs. of  21 variables:
##  $ 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 ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  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                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ 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              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  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                     : int  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..1.19.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 ...
```

```
levels(life_data$Status)
```

```
## [1] "Developed"  "Developing"
```

We found only “Status” is a factor class. Our data has 22 features and 2938 observations.

```
# check which columns has NA values.
anyNA(life_data)
```

```
## [1] TRUE
```

```
cols_has_na = names(life_data)[colSums(is.na(life_data)) != 0]
cols_has_na
```

```
##  [1] "Life.expectancy"                 "Adult.Mortality"
##  [3] "Alcohol"                         "Hepatitis.B"
##  [5] "BMI"                             "Polio"
##  [7] "Total.expenditure"               "Diphtheria"
##  [9] "GDP"                             "Population"
## [11] "thinness..1.19.years"            "thinness.5.9.years"
## [13] "Income.composition.of.resources" "Schooling"
```

```
cat(length(cols_has_na), "columns has na values. Status, the categorical variable, does not contain any na.")
```

```
## 14 columns has na values. Status, the categorical variable, does not contain any na.
```

Methods

A general overview

- Handling missing data

we will be using pakcage “mice” to impute all the missing data with random forest(5th time) to make the prediction each better.

- Modeling

Linear: we will fit a additive full model, a raw AIC model, and aa model selected based on AIC. Non-linear: We will fit a random forest regression model and a KNN(both with and without scaling) model and compare these non-linear models to MLR. The results from random forest would be used to interpret the importances of features.

- Feature Selection

Generalization indicator: LOOCV RMSE for linear model, test RMSE for non-linear model. This is because training of random forest is too expensive.

Diagnostic: Ajusted R square, Influential points, QQ plot, residual vs. fitted, shapiro test, bp test will be used for selecting the features for MLR. Non-linear models is not the big focus of this project so feature engineering would be limited.

Pacakges will be used

```
library(caret)
library(lmtest)
library(faraway)
library(mice)
library(randomForest)
library(lmtest)
library(knitr)
```

Handling missing data and highly correlated data

```
### we drop the rows whose reponse is NA
life_data = life_data[!is.na(life_data$Life.expectancy),]
cat("now the data has",nrow(life_data),"observations and",ncol(life_data),"features")
```

```
## now the data has 2928 observations and 21 features
```

```
##
##  iter imp variable
##   1   1  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   1   2  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   1   3  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   1   4  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   1   5  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   2   1  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   2   2  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   2   3  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   2   4  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   2   5  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   3   1  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   3   2  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   3   3  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   3   4  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   3   5  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   4   1  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   4   2  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   4   3  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   4   4  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   4   5  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   5   1  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   5   2  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   5   3  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   5   4  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
##   5   5  Alcohol  Hepatitis.B  BMI  Polio  Total.expenditure  Diphtheria  GDP  Population  thinness..1.19.years
thinness.5.9.years  Income.composition.of.resources  Schooling
```

```
# check all the NA's have been removed and see the new data
life_data = read.csv("life_data.csv")
life_data = subset(life_data, select = -c(1))
summary(life_data)
```

```
##       Year             Status     Life.expectancy Adult.Mortality
##  Min.   :2000   Developed : 512   Min.   :36.30   Min.   :  1.0  
##  1st Qu.:2004   Developing:2416   1st Qu.:63.10   1st Qu.: 74.0  
##  Median :2008                     Median :72.10   Median :144.0  
##  Mean   :2008                     Mean   :69.22   Mean   :164.8  
##  3rd Qu.:2011                     3rd Qu.:75.70   3rd Qu.:228.0  
##  Max.   :2015                     Max.   :89.00   Max.   :723.0  
##  infant.deaths        Alcohol       percentage.expenditure  Hepatitis.B
##  Min.   :   0.00   Min.   : 0.010   Min.   :    0.000      Min.   : 1.00
##  1st Qu.:   0.00   1st Qu.: 0.590   1st Qu.:    4.854      1st Qu.:73.00
##  Median :   3.00   Median : 3.520   Median :   65.611      Median :91.00
##  Mean   :  30.41   Mean   : 4.458   Mean   :  740.321      Mean   :78.91
##  3rd Qu.:  22.00   3rd Qu.: 7.550   3rd Qu.:  442.614      3rd Qu.:96.00
##  Max.   :1800.00   Max.   :17.870   Max.   :19479.912      Max.   :99.00
##     Measles              BMI        under.five.deaths     Polio      
##  Min.   :     0.0   Min.   : 1.00   Min.   :   0.00   Min.   : 3.00  
##  1st Qu.:     0.0   1st Qu.:19.10   1st Qu.:   0.00   1st Qu.:78.00  
##  Median :    17.0   Median :43.00   Median :   4.00   Median :93.00  
##  Mean   :  2427.9   Mean   :37.99   Mean   :  42.18   Mean   :82.57  
##  3rd Qu.:   362.2   3rd Qu.:56.10   3rd Qu.:  28.00   3rd Qu.:97.00  
##  Max.   :212183.0   Max.   :77.60   Max.   :2500.00   Max.   :99.00  
##  Total.expenditure   Diphtheria       HIV.AIDS           GDP           
##  Min.   : 0.370    Min.   : 2.00   Min.   : 0.100   Min.   :     1.68  
##  1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   392.82  
##  Median : 5.725    Median :93.00   Median : 0.100   Median :  1439.39  
##  Mean   : 5.927    Mean   :82.34   Mean   : 1.748   Mean   :  6778.67  
##  3rd Qu.: 7.470    3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5335.76  
##  Max.   :17.600    Max.   :99.00   Max.   :50.600   Max.   :119172.74  
##    Population        thinness..1.19.years thinness.5.9.years
##  Min.   :3.400e+01   Min.   : 0.100       Min.   : 0.100    
##  1st Qu.:1.907e+05   1st Qu.: 1.600       1st Qu.: 1.600    
##  Median :1.351e+06   Median : 3.400       Median : 3.400    
##  Mean   :1.210e+07   Mean   : 4.878       Mean   : 4.908    
##  3rd Qu.:7.463e+06   3rd Qu.: 7.200       3rd Qu.: 7.300    
##  Max.   :1.294e+09   Max.   :27.700       Max.   :28.600    
##  Income.composition.of.resources   Schooling    
##  Min.   :0.0000                  Min.   : 0.00  
##  1st Qu.:0.4880                  1st Qu.:10.10  
##  Median :0.6755                  Median :12.30  
##  Mean   :0.6229                  Mean   :11.98  
##  3rd Qu.:0.7802                  3rd Qu.:14.30  
##  Max.   :0.9480                  Max.   :20.70
```

```
cor_mat = cor(subset(life_data, select=-c(2)))
cor_mat
```

```
##                                        Year Life.expectancy
## Year                             1.00000000      0.17003302
## Life.expectancy                  0.17003302      1.00000000
## Adult.Mortality                 -0.07905159     -0.69635931
## infant.deaths                   -0.03646405     -0.19655718
## Alcohol                         -0.10240924      0.38681485
## percentage.expenditure           0.03272257      0.38186350
## Hepatitis.B                      0.16482515      0.34939521
## Measles                         -0.08184033     -0.15758580
## BMI                              0.10410611      0.57334044
## under.five.deaths               -0.04197985     -0.22252912
## Polio                            0.09351400      0.46254453
## Total.expenditure                0.08379779      0.21924793
## Diphtheria                       0.13282994      0.47628734
## HIV.AIDS                        -0.13878854     -0.55655625
## GDP                              0.09485729      0.44300460
## Population                       0.01829211     -0.03552482
## thinness..1.19.years            -0.04239973     -0.47411907
## thinness.5.9.years              -0.04636332     -0.46836129
## Income.composition.of.resources  0.23422808      0.71063151
## Schooling                        0.20561002      0.74723863
##                                 Adult.Mortality infant.deaths     Alcohol
## Year                              -0.0790515894   -0.03646405 -0.10240924
## Life.expectancy                   -0.6963593138   -0.19655718  0.38681485
## Adult.Mortality                    1.0000000000    0.07875601 -0.19034231
## infant.deaths                      0.0787560117    1.00000000 -0.11167407
## Alcohol                           -0.1903423135   -0.11167407  1.00000000
## percentage.expenditure            -0.2428595283   -0.08590584  0.34092695
## Hepatitis.B                       -0.2069469484   -0.21944382  0.12949216
## Measles                            0.0311764119    0.50103772 -0.04793560
## BMI                               -0.3942273941   -0.22712775  0.32206422
## under.five.deaths                  0.0941461272    0.99662815 -0.10854760
## Polio                             -0.2743173607   -0.17095205  0.21744974
## Total.expenditure                 -0.1175450172   -0.12913785  0.29570304
## Diphtheria                        -0.2748298071   -0.17537304  0.21295734
## HIV.AIDS                           0.5238205079    0.02495467 -0.04232854
## GDP                               -0.2913972211   -0.10415209  0.32527981
## Population                        -0.0001027216    0.55289214 -0.03713300
## thinness..1.19.years               0.3015995498    0.46295685 -0.41534054
## thinness.5.9.years                 0.3084199130    0.46870946 -0.40497616
## Income.composition.of.resources   -0.4643517551   -0.15284023  0.41293724
## Schooling                         -0.4643946036   -0.20511670  0.51163229
##                                 percentage.expenditure Hepatitis.B
## Year                                        0.03272257  0.16482515
## Life.expectancy                             0.38186350  0.34939521
## Adult.Mortality                            -0.24285953 -0.20694695
## infant.deaths                              -0.08590584 -0.21944382
## Alcohol                                     0.34092695  0.12949216
## percentage.expenditure                      1.00000000  0.07072946
## Hepatitis.B                                 0.07072946  1.00000000
## Measles                                    -0.05683054 -0.13870163
## BMI                                         0.23344585  0.24162778
## under.five.deaths                          -0.08815223 -0.23094044
## Polio                                       0.14692778  0.51632864
## Total.expenditure                           0.16902702  0.12693009
## Diphtheria                                  0.14337558  0.61453572
## HIV.AIDS                                   -0.09822981 -0.14351898
## GDP                                         0.87321854  0.10922547
## Population                                 -0.02670560 -0.10300983
## thinness..1.19.years                       -0.25200985 -0.19150233
## thinness.5.9.years                         -0.25362669 -0.19686348
## Income.composition.of.resources             0.36572597  0.26057766
## Schooling                                   0.37565388  0.31020657
##                                     Measles         BMI under.five.deaths
## Year                            -0.08184033  0.10410611       -0.04197985
## Life.expectancy                 -0.15758580  0.57334044       -0.22252912
## Adult.Mortality                  0.03117641 -0.39422739        0.09414613
## infant.deaths                    0.50103772 -0.22712775        0.99662815
## Alcohol                         -0.04793560  0.32206422       -0.10854760
## percentage.expenditure          -0.05683054  0.23344585       -0.08815223
## Hepatitis.B                     -0.13870163  0.24162778       -0.23094044
## Measles                          1.00000000 -0.17371142        0.50771799
## BMI                             -0.17371142  1.00000000       -0.23793022
## under.five.deaths                0.50771799 -0.23793022        1.00000000
## Polio                           -0.13622949  0.28948530       -0.18901215
## Total.expenditure               -0.10514678  0.24148852       -0.13071951
## Diphtheria                      -0.14188765  0.28824952       -0.19587913
## HIV.AIDS                         0.03067341 -0.24261210        0.03778323
## GDP                             -0.07255652  0.28925820       -0.10770734
## Population                       0.26730334 -0.06984789        0.54042050
## thinness..1.19.years             0.22314454 -0.52802044        0.46521690
## thinness.5.9.years               0.21884812 -0.53475504        0.46982624
## Income.composition.of.resources -0.15127734  0.51228009       -0.17102076
## Schooling                       -0.15782322  0.55693962       -0.22119142
##                                       Polio Total.expenditure  Diphtheria
## Year                             0.09351400      0.0837977869  0.13282994
## Life.expectancy                  0.46254453      0.2192479289  0.47628734
## Adult.Mortality                 -0.27431736     -0.1175450172 -0.27482981
## infant.deaths                   -0.17095205     -0.1291378471 -0.17537304
## Alcohol                          0.21744974      0.2957030410  0.21295734
## percentage.expenditure           0.14692778      0.1690270163  0.14337558
## Hepatitis.B                      0.51632864      0.1269300859  0.61453572
## Measles                         -0.13622949     -0.1051467847 -0.14188765
## BMI                              0.28948530      0.2414885222  0.28824952
## under.five.deaths               -0.18901215     -0.1307195128 -0.19587913
## Polio                            1.00000000      0.1485139903  0.67134553
## Total.expenditure                0.14851399      1.0000000000  0.16056505
## Diphtheria                       0.67134553      0.1605650510  1.00000000
## HIV.AIDS                        -0.15998656      0.0002208433 -0.16546908
## GDP                              0.18960224      0.1526468576  0.19107441
## Population                      -0.05086496     -0.0519543691 -0.03790595
## thinness..1.19.years            -0.21716604     -0.2732040534 -0.22493807
## thinness.5.9.years              -0.22024147     -0.2820275068 -0.22114229
## Income.composition.of.resources  0.36553625      0.1841025559  0.39629940
## Schooling                        0.40040506      0.2734388128  0.41759877
##                                      HIV.AIDS         GDP    Population
## Year                            -0.1387885438  0.09485729  0.0182921141
## Life.expectancy                 -0.5565562534  0.44300460 -0.0355248207
## Adult.Mortality                  0.5238205079 -0.29139722 -0.0001027216
## infant.deaths                    0.0249546750 -0.10415209  0.5528921366
## Alcohol                         -0.0423285426  0.32527981 -0.0371329963
## percentage.expenditure          -0.0982298117  0.87321854 -0.0267056022
## Hepatitis.B                     -0.1435189813  0.10922547 -0.1030098299
## Measles                          0.0306734076 -0.07255652  0.2673033377
## BMI                             -0.2426121040  0.28925820 -0.0698478900
## under.five.deaths                0.0377832289 -0.10770734  0.5404204974
## Polio                           -0.1599865586  0.18960224 -0.0508649602
## Total.expenditure                0.0002208433  0.15264686 -0.0519543691
## Diphtheria                      -0.1654690761  0.19107441 -0.0379059475
## HIV.AIDS                         1.0000000000 -0.12553067 -0.0219858188
## GDP                             -0.1255306748  1.00000000 -0.0326330743
## Population                      -0.0219858188 -0.03263307  1.0000000000
## thinness..1.19.years             0.2006633479 -0.27650202  0.2346754894
## thinness.5.9.years               0.2040340747 -0.28045266  0.2348746264
## Income.composition.of.resources -0.2417315313  0.43516738 -0.0165344492
## Schooling                       -0.2197712747  0.43498892 -0.0396778149
##                                 thinness..1.19.years thinness.5.9.years
## Year                                     -0.04239973        -0.04636332
## Life.expectancy                          -0.47411907        -0.46836129
## Adult.Mortality                           0.30159955         0.30841991
## infant.deaths                             0.46295685         0.46870946
## Alcohol                                  -0.41534054        -0.40497616
## percentage.expenditure                   -0.25200985        -0.25362669
## Hepatitis.B                              -0.19150233        -0.19686348
## Measles                                   0.22314454         0.21884812
## BMI                                      -0.52802044        -0.53475504
## under.five.deaths                         0.46521690         0.46982624
## Polio                                    -0.21716604        -0.22024147
## Total.expenditure                        -0.27320405        -0.28202751
## Diphtheria                               -0.22493807        -0.22114229
## HIV.AIDS                                  0.20066335         0.20403407
## GDP                                      -0.27650202        -0.28045266
## Population                                0.23467549         0.23487463
## thinness..1.19.years                      1.00000000         0.93527661
## thinness.5.9.years                        0.93527661         1.00000000
## Income.composition.of.resources          -0.41091658        -0.40014430
## Schooling                                -0.47011917        -0.45915272
##                                 Income.composition.of.resources
## Year                                                 0.23422808
## Life.expectancy                                      0.71063151
## Adult.Mortality                                     -0.46435176
## infant.deaths                                       -0.15284023
## Alcohol                                              0.41293724
## percentage.expenditure                               0.36572597
## Hepatitis.B                                          0.26057766
## Measles                                             -0.15127734
## BMI                                                  0.51228009
## under.five.deaths                                   -0.17102076
## Polio                                                0.36553625
## Total.expenditure                                    0.18410256
## Diphtheria                                           0.39629940
## HIV.AIDS                                            -0.24173153
## GDP                                                  0.43516738
## Population                                          -0.01653445
## thinness..1.19.years                                -0.41091658
## thinness.5.9.years                                  -0.40014430
## Income.composition.of.resources                      1.00000000
## Schooling                                            0.79092723
##                                   Schooling
## Year                             0.20561002
## Life.expectancy                  0.74723863
## Adult.Mortality                 -0.46439460
## infant.deaths                   -0.20511670
## Alcohol                          0.51163229
## percentage.expenditure           0.37565388
## Hepatitis.B                      0.31020657
## Measles                         -0.15782322
## BMI                              0.55693962
## under.five.deaths               -0.22119142
## Polio                            0.40040506
## Total.expenditure                0.27343881
## Diphtheria                       0.41759877
## HIV.AIDS                        -0.21977127
## GDP                              0.43498892
## Population                      -0.03967781
## thinness..1.19.years            -0.47011917
## thinness.5.9.years              -0.45915272
## Income.composition.of.resources  0.79092723
## Schooling                        1.00000000
```

```
drop_index= which(cor(subset(life_data, select=-c(2)))  > 0.9 &
                    cor(subset(life_data, select=-c(2))) < 1)
cor_mat[drop_index]
```

```
## [1] 0.9966281 0.9966281 0.9352766 0.9352766
```

```
drop_index
```

```
## [1]  70 184 338 357
```

We discover `infant.deaths ~ under.five.deaths` has correlation 0.9966281 and `thinness.5.9.years ~ thinness..1.19.years` has correlation 0.9363631 So we decided to only keep `infant.deaths` and `thinness..1.19.years` (`thinness..1.19.years` appears to be less corrlated with other features)

```
## drop the columns
life_data = subset(life_data, select=-c(under.five.deaths, thinness.5.9.years))
```

```
### do the train_test_split
set.seed(1969)
split = rbinom(0.66, size = 1, n = nrow(life_data))
sum(split == 0)
```

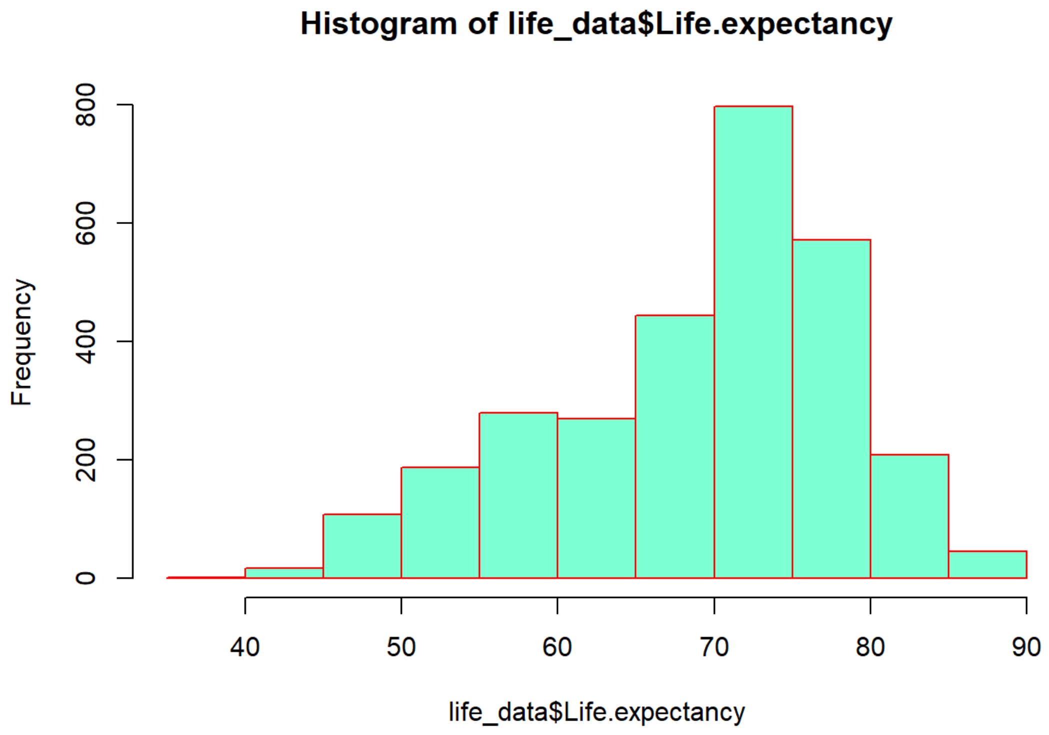
```
## [1] 1016
```

```
train_data = life_data[split == 1,]
test_data = life_data[!split == 0,]
```

Separate original data into train and test set (66% vs. 33%)

Data Visuailzation

```
hist(life_data$Life.expectancy, col = "aquamarine", border = "red")
```



Feature Engineering

```
### change the year to year_i - min(year)
life_data$Year = life_data$Year - min(life_data$Year)
```

It’s a commonly used trick to make scales the years to max(year) - min(year), or current - year to make the years have higher contrast from each other.

Helper functions

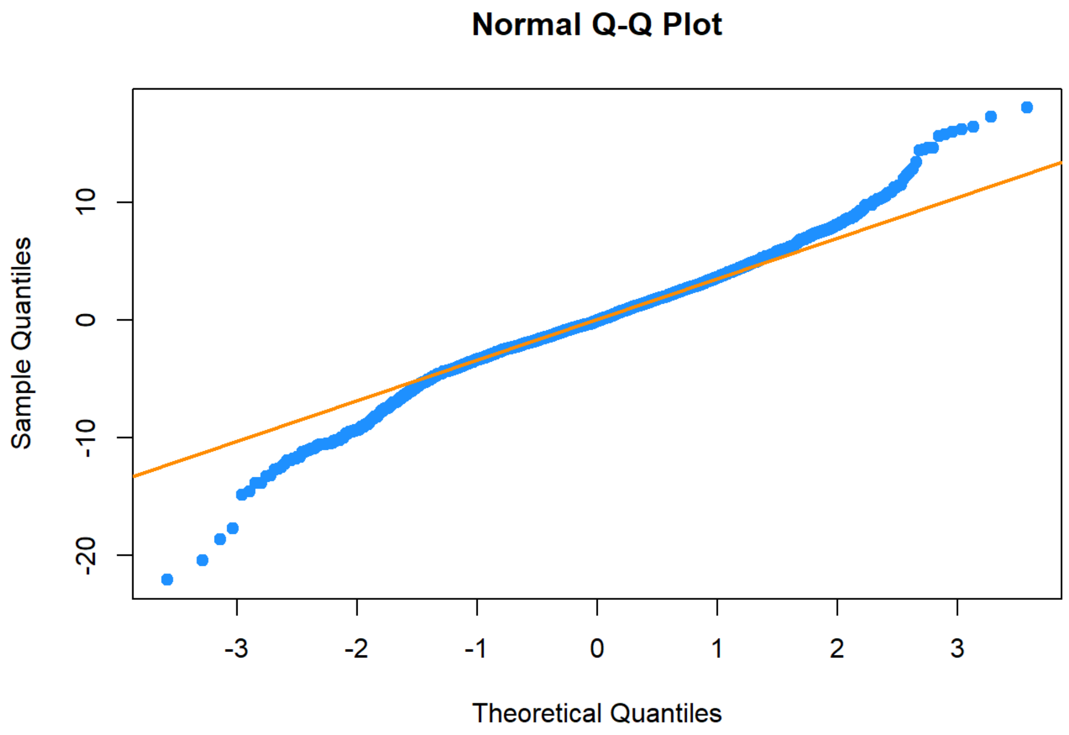
```
##### Scoring tool kits

##RMSE
calc_loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}

calc_rmse = function(predicted, actual) {
  return(sqrt(mean((predicted - actual)^2)))
}

## calculate test performace in one step
test_rmse = function(model_name, model) {
  pred = predict(model, test_data)
  cat("Model", model_name, "has test rmse:",
      calc_rmse(predicted=pred, actual=test_data$Life.expectancy))
}
```

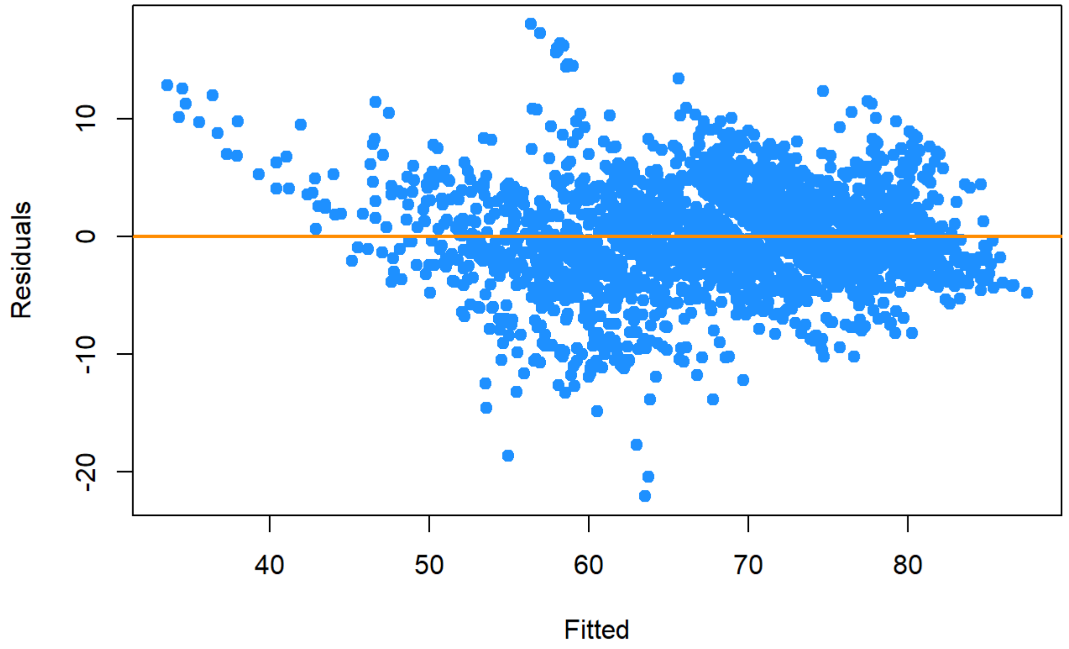
Full additive model

```
full_additive = lm(data = life_data, Life.expectancy ~ .)
cat("\nAdjusted R^2:", summary(full_additive)$adj)
```

```
##
## Adjusted R^2: 0.8194621
```

```
check_assumptions(full_additive)
```





```
## shapiro test p-value : 1.614711e-19
##  the normality assumption was violated
##  bptest p-value : 4.295118e-83
##  the equal variance assumption was violated
```

```
test_rmse("\nfull additive", full_additive)
```

```
## Model
## full additive has test rmse: 61.64631
```

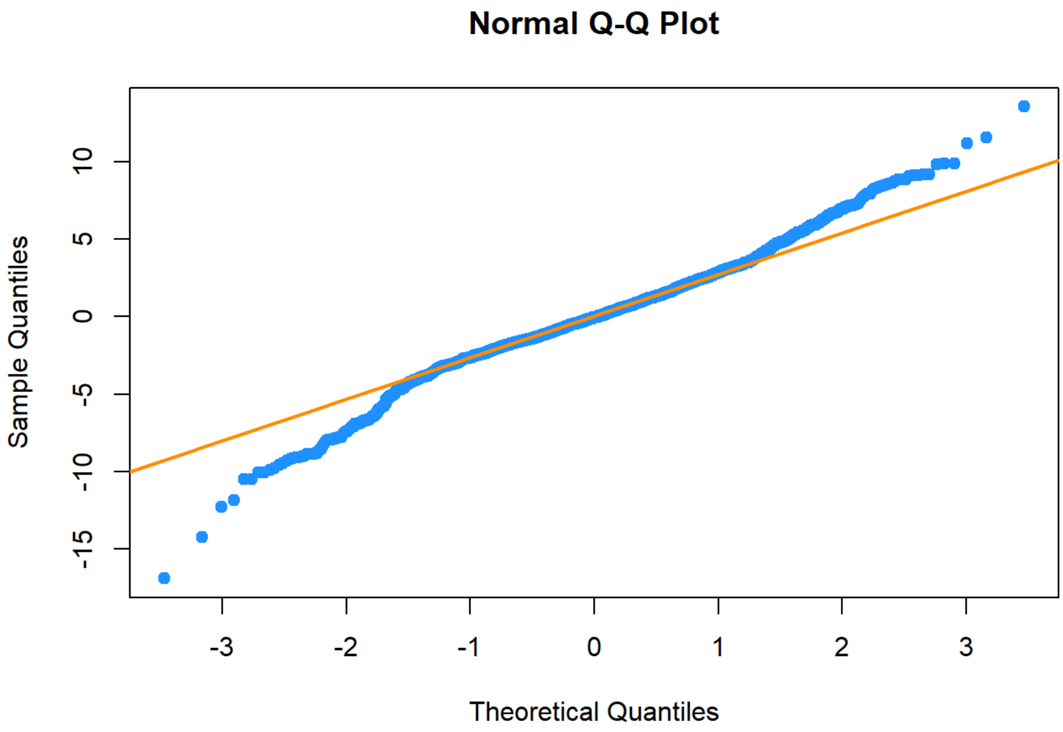
```
cat("\nLOOCV rmse:",calc_loocv_rmse(full_additive))
```

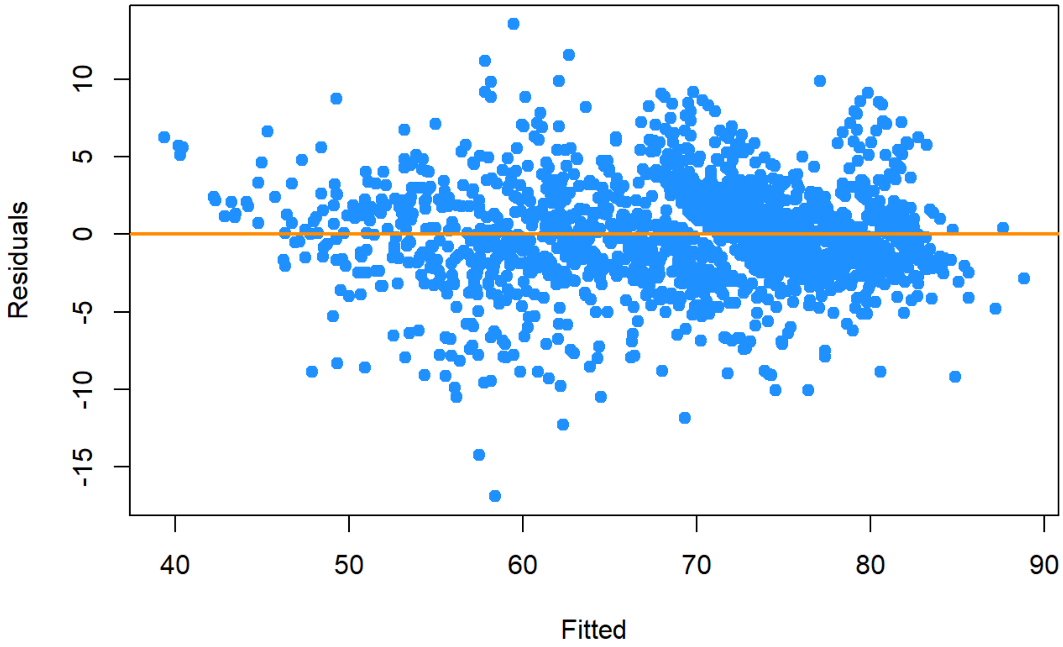
```
##
## LOOCV rmse: 4.068498
```

This is not good enough. All assumptions are violated and test rmse is really high because of unusual obersvations

AIC selection Model

```
##############
############## This is the model chosen by aic initially. (begin with . ^ 2)
aic_ini = lm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
    infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B +
    Measles + BMI + Polio + Total.expenditure + Diphtheria +
    HIV.AIDS + GDP + Population + thinness..1.19.years + Income.composition.of.resources +
    Schooling + Year:infant.deaths + Year:Alcohol + Year:Measles +
    Year:HIV.AIDS + Year:GDP + Year:thinness..1.19.years + Year:Income.composition.of.resources +
    Year:Schooling + Status:Adult.Mortality + Status:Alcohol +
    Status:Hepatitis.B + Status:BMI + Status:Total.expenditure +
    Status:Population + Status:thinness..1.19.years + Status:Schooling +
    Adult.Mortality:Alcohol + Adult.Mortality:percentage.expenditure +
    Adult.Mortality:BMI + Adult.Mortality:Total.expenditure +
    Adult.Mortality:Diphtheria + Adult.Mortality:HIV.AIDS + Adult.Mortality:GDP +
    Adult.Mortality:thinness..1.19.years + Adult.Mortality:Schooling +
    infant.deaths:Measles + infant.deaths:BMI + infant.deaths:Polio +
    infant.deaths:Total.expenditure + infant.deaths:Diphtheria +
    infant.deaths:Population + infant.deaths:Income.composition.of.resources +
    infant.deaths:Schooling + Alcohol:Hepatitis.B + Alcohol:Measles +
    Alcohol:Polio + Alcohol:Total.expenditure + Alcohol:HIV.AIDS +
    Alcohol:Population + Alcohol:thinness..1.19.years + Alcohol:Income.composition.of.resources +
    Alcohol:Schooling + percentage.expenditure:BMI + percentage.expenditure:thinness..1.19.years +
    Hepatitis.B:Measles + Hepatitis.B:Polio + Hepatitis.B:Total.expenditure +
    Hepatitis.B:Diphtheria + Hepatitis.B:Income.composition.of.resources +
    Measles:Polio + Measles:Diphtheria + BMI:Diphtheria + BMI:thinness..1.19.years +
    BMI:Income.composition.of.resources + BMI:Schooling + Polio:Diphtheria +
    Polio:GDP + Polio:Schooling + Total.expenditure:Diphtheria +
    Total.expenditure:HIV.AIDS + Total.expenditure:GDP + Total.expenditure:Population +
    Total.expenditure:thinness..1.19.years + Total.expenditure:Schooling +
    HIV.AIDS:GDP + HIV.AIDS:Population + HIV.AIDS:Income.composition.of.resources +
    HIV.AIDS:Schooling + GDP:thinness..1.19.years + GDP:Income.composition.of.resources +
    GDP:Schooling + Population:Income.composition.of.resources +
    thinness..1.19.years:Income.composition.of.resources + thinness..1.19.years:Schooling +
    Income.composition.of.resources:Schooling, data = train_data)
check_assumptions(aic_ini)
```





```
## shapiro test p-value : 2.115756e-15
##  the normality assumption was violated
##  bptest p-value : 2.510088e-21
##  the equal variance assumption was violated
```

```
cat("\nmodel adj.R^2 :", summary(aic_ini)$adj, "\n")
```

```
##
## model adj.R^2 : 0.8796093
```

```
test_rmse("AIC linear", aic_ini)
```

```
## Model AIC linear has test rmse: 3.20321
```

```
cat("\nLoocv is:", calc_loocv_rmse(aic_ini), "\n")
```

```
##
## Loocv is: 3.450851
```

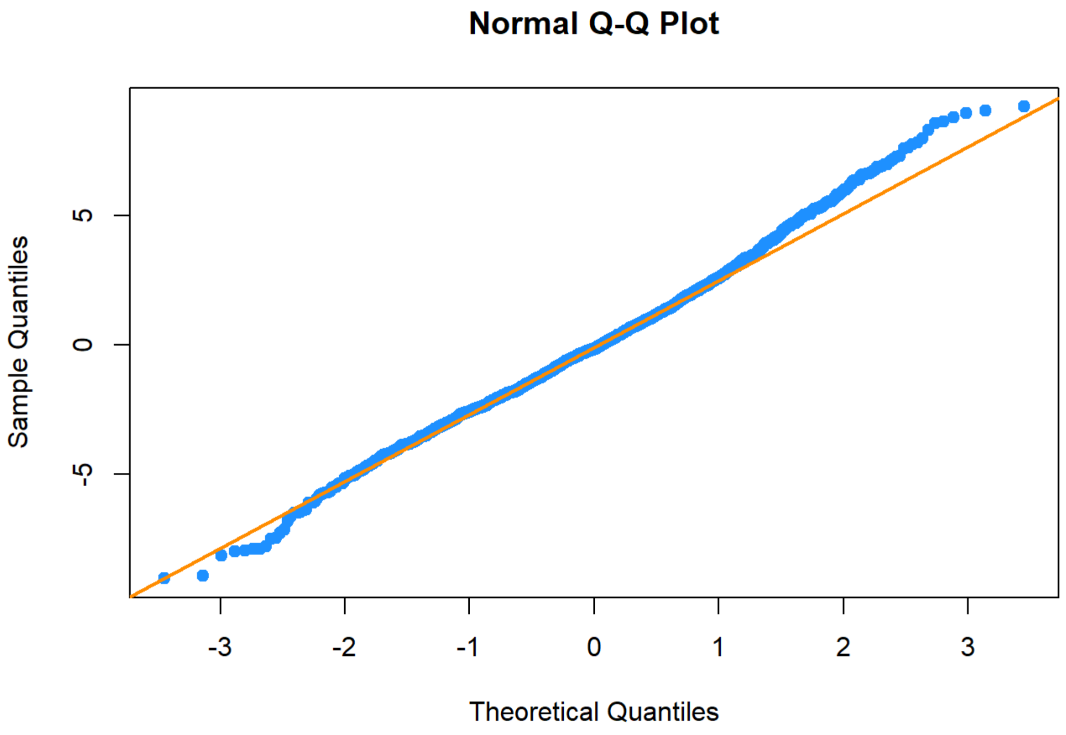
If we choose BIC, the penalty is too large so that the full model was surprisingly reduced to only one parameter(the mean of all life expectancy). So we choose AIC instead. There’s so much parameters so we might had better give up the three way interactions.

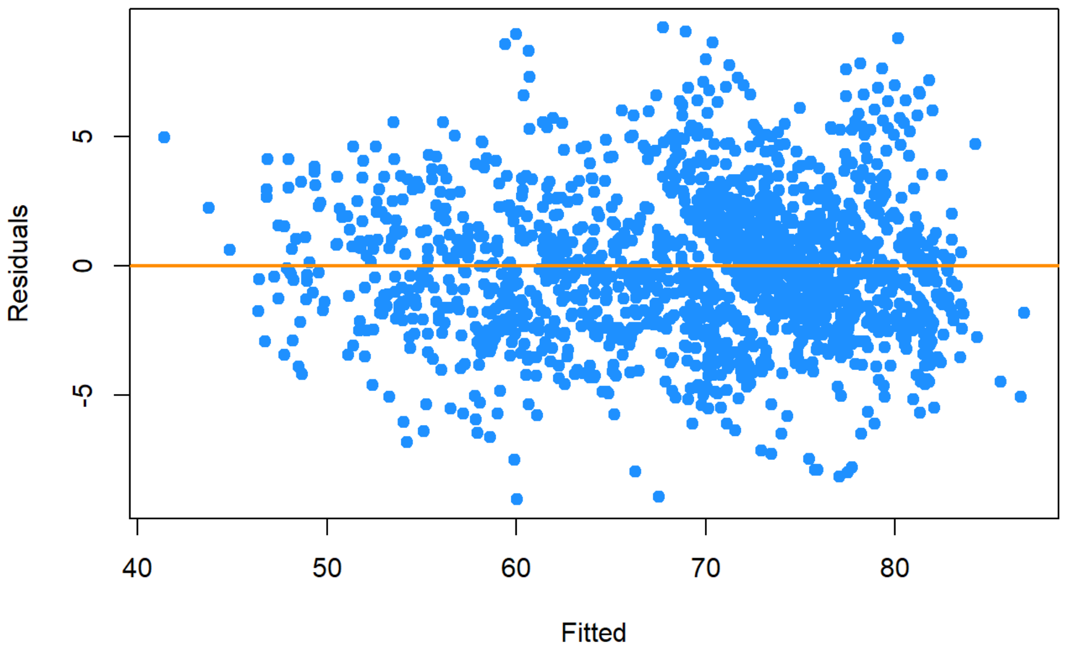
This model makes a lot more sense, however, both assumptions are violated. We will try to make some improvement on this.

Our best linear model after manual selection

```
###### use AIC to choose feature, takes super long to run!!
# full_model = lm(data = train_data, Life.expectancy ~ . ^ 2)
# best_model = step(full_model, direction="backward",trace=0)
```

```
### This is the best multi-linear regression model we can make for this problem
best_model = lm(formula = Life.expectancy ~ Year +
    poly(infant.deaths,2) + poly(percentage.expenditure,2) + poly(Hepatitis.B, 2) +
    I(Measles ^ 2) + poly(BMI,3) + poly(Polio,2) +  poly(Diphtheria,3) +
    poly(HIV.AIDS, 3) + GDP + poly(Population,2) +
    poly(Schooling,2) +Income.composition.of.resources +
    Status:Adult.Mortality  +
    Status:Hepatitis.B  + Status:thinness..1.19.years +
    Adult.Mortality:Alcohol  +
    Adult.Mortality:Diphtheria + Adult.Mortality:HIV.AIDS + Adult.Mortality:GDP + Adult.Mortality:Schooling +
    infant.deaths:Measles + infant.deaths:BMI +
    infant.deaths:Total.expenditure + Alcohol:Measles +
    Alcohol:Polio + Alcohol:HIV.AIDS +
    Alcohol:Population  + Alcohol:Income.composition.of.resources +
     percentage.expenditure:BMI +
    Hepatitis.B:Diphtheria + Hepatitis.B:Income.composition.of.resources  + BMI:Schooling + Polio:Diphtheria +
    Total.expenditure:HIV.AIDS +
     Total.expenditure:Schooling +
    GDP:Schooling + thinness..1.19.years:Schooling, data = train_data)
non_influential_idx = cooks.distance(best_model) <= 4 / length(cooks.distance(best_model))
train_wo_influential = train_data[non_influential_idx, ]

best_model_wo_influential = lm(formula = Life.expectancy ~ Year +
    poly(infant.deaths,2) + poly(percentage.expenditure,2) + poly(Hepatitis.B, 2) +
    I(Measles ^ 2) + poly(BMI,3) + poly(Polio,2) +  poly(Diphtheria,3) +
    poly(HIV.AIDS, 3) + GDP + poly(Population,2) +
    poly(Schooling,2) +Income.composition.of.resources +
    Status:Adult.Mortality  +
    Status:Hepatitis.B  + Status:thinness..1.19.years +
    Adult.Mortality:Alcohol  +
    Adult.Mortality:Diphtheria + Adult.Mortality:HIV.AIDS + Adult.Mortality:GDP + Adult.Mortality:Schooling +
    infant.deaths:Measles + infant.deaths:BMI +
    infant.deaths:Total.expenditure + Alcohol:Measles +
    Alcohol:Polio + Alcohol:HIV.AIDS +
    Alcohol:Population  + Alcohol:Income.composition.of.resources +
     percentage.expenditure:BMI +
    Hepatitis.B:Diphtheria + Hepatitis.B:Income.composition.of.resources  + BMI:Schooling + Polio:Diphtheria +
    Total.expenditure:HIV.AIDS +
     Total.expenditure:Schooling +
    GDP:Schooling + Population:Income.composition.of.resources  + thinness..1.19.years:Schooling, data = train_wo_in
fluential)
check_assumptions(best_model_wo_influential)
```





```
## shapiro test p-value : 7.362058e-06
##  the normality assumption was violated
##  bptest p-value : 0.08418819
##  the equal variance assumption was not violated
```

```
cat("\nmodel adj.R^2 :", summary(best_model_wo_influential)$adj, "\n")
```

```
##
## model adj.R^2 : 0.9024191
```

```
test_rmse("AIC linear", best_model_wo_influential)
```

```
## Model AIC linear has test rmse: 3.403834
```

```
cat("\nLoocv is:", calc_loocv_rmse(best_model_wo_influential), "\n")
```

```
##
## Loocv is: 2.8191
```

```
vif(best_model_wo_influential)
```

```
##                                         Year
##                                 1.338897e+00
##                      poly(infant.deaths, 2)1
##                                 4.884539e+01
##                      poly(infant.deaths, 2)2
##                                 7.226052e+00
##             poly(percentage.expenditure, 2)1
##                                 2.535296e+01
##             poly(percentage.expenditure, 2)2
##                                 1.283999e+00
##                        poly(Hepatitis.B, 2)1
##                                 2.469494e+01
##                        poly(Hepatitis.B, 2)2
##                                 2.522269e+00
##                                 I(Measles^2)
##                                 2.701284e+00
##                                poly(BMI, 3)1
##                                 6.598683e+01
##                                poly(BMI, 3)2
##                                 1.488535e+00
##                                poly(BMI, 3)3
##                                 1.388269e+00
##                              poly(Polio, 2)1
##                                 1.538570e+01
##                              poly(Polio, 2)2
##                                 4.725789e+00
##                         poly(Diphtheria, 3)1
##                                 2.326166e+01
##                         poly(Diphtheria, 3)2
##                                 5.445606e+00
##                         poly(Diphtheria, 3)3
##                                 1.322417e+00
##                           poly(HIV.AIDS, 3)1
##                                 4.867696e+01
##                           poly(HIV.AIDS, 3)2
##                                 1.969458e+00
##                           poly(HIV.AIDS, 3)3
##                                 1.821509e+00
##                                          GDP
##                                 1.210936e+02
##                         poly(Population, 2)1
##                                 3.622341e+01
##                         poly(Population, 2)2
##                                 2.215512e+00
##                          poly(Schooling, 2)1
##                                 3.151408e+01
##                          poly(Schooling, 2)2
##                                 3.341624e+00
##              Income.composition.of.resources
##                                 1.882433e+01
##              StatusDeveloped:Adult.Mortality
##                                 1.078743e+01
##             StatusDeveloping:Adult.Mortality
##                                 4.212955e+01
##                  StatusDeveloped:Hepatitis.B
##                                 8.196251e+00
##                 StatusDeveloping:Hepatitis.B
##                                 3.224130e+04
##         StatusDeveloped:thinness..1.19.years
##                                 4.454081e-01
##        StatusDeveloping:thinness..1.19.years
##                                 2.467662e-04
##                      Adult.Mortality:Alcohol
##                                 1.105126e-01
##                  Adult.Mortality:Diphtheria 
##                                2.115073e+02 
##                    Adult.Mortality:HIV.AIDS 
##                                3.131802e-05 
##                         GDP:Adult.Mortality 
##                                2.754126e+07 
##                   Adult.Mortality:Schooling 
##                                2.597590e-06 
##                       infant.deaths:Measles 
##                                7.394855e+07 
##                           infant.deaths:BMI 
##                                5.097523e+02 
##             infant.deaths:Total.expenditure 
##                                4.109923e-04 
##                             Alcohol:Measles 
##                                1.118259e+05 
##                               Alcohol:Polio 
##                                2.843543e+03 
##                            Alcohol:HIV.AIDS 
##                                8.214245e-14 
##                          Alcohol:Population 
##                                4.014434e+16 
##     Income.composition.of.resources:Alcohol 
##                                2.063964e-08 
##                  BMI:percentage.expenditure 
##                                4.407175e+04 
##                      Hepatitis.B:Diphtheria 
##                                4.505473e+05 
## Income.composition.of.resources:Hepatitis.B 
##                                6.506612e-01 
##                               Schooling:BMI 
##                                4.069883e-01 
##                            Diphtheria:Polio 
##                                2.576285e+05 
##                  HIV.AIDS:Total.expenditure 
##                                1.246521e+00 
##                 Schooling:Total.expenditure 
##                                4.991431e-06 
##                               GDP:Schooling 
##                                2.917815e-03 
##  Income.composition.of.resources:Population 
##                                8.161897e+12 
##              thinness..1.19.years:Schooling 
##                                1.142766e+02
```

We can see that after dropping the influential points, we fail to reject null hypothesis of BP test, which means that the model does not violate contant variance assumption. However, p-value of the shapiro test still indicates that the normality assumption is violated. Also, through investigating the variance inflation factor, we can see that there is a huge multicollinearity issue as many of the predictors have a VIF greater than 5. The failure of normality assumption and the multicollinearity issue may due to the loss of some features in the data. These features may provide critical information for the prediction model. Without these information, we may not able to find a linear model which fulfills both assumptions and has little multicollinearity issue.

```
# lm_drop_outliers
sum(cooks.distance(best_model_wo_influential) > 4 / length(nrow(train_data)))
```

```
## [1] 0
```

This is to make sure there’s no influential points in our model.

Random Forest Regression

```
#### Random Forest
rf = train(data = train_data, Life.expectancy ~ ., method="rf")
```

```
test_rmse("Randomforest", rf)
```

```
## Model Randomforest has test rmse: 0.8247877
```

Random Forest works the best!!!!!!!!!!!! The testing RMSe is soooooo small However, this takes quite long to run(over 10 mins) to we might had better give up cross validation on this.

K Nearest Neighbours

```
### KNN without scaling
knn_without_scale = train(data = train_data, Life.expectancy ~ ., method="knn")
test_rmse("KNN_without_scale", knn_without_scale)
```

```
## Model KNN_without_scale has test rmse: 8.017768
```

```
### KNN with scaling
knn_with_scale = train(data = train_data, Life.expectancy ~ ., method="knn", preProcess=c("center", "scale"))
test_rmse("KNN with scale", knn_with_scale)
```

```
## Model KNN with scale has test rmse: 2.439092
```

We found KNN without scaling works poorly, but the KNN with scaling works quite well.

```
##### This chunk is Left for stacking
```

We originally wanted to do some stacking to even improve our accuracy. However, with the limitation of computation power, it takes unacceptably long to do cross validation. So we finally give up on this idea. Random forest has already gave good prediction.

Results

```
** To conclude which model we would prefer, we should discuss separately, as we've employed models from different family. **
```

Inside the Linear family

- Inside the linear family:

```
results = data.frame(full_additive=c(0.8206541,4.054004,5.593882e-20,1.919205e-80),
                     raw_aic = c(0.8823423,3.392533,8.960947e-15,7.925938e-17),
                     manual_selected_aic=c(0.9030391,2.81044,4.216292e-06,0.03813797))
rownames(results) = c("Adj.R2", "LOOCV test rmse", "Normality p_val", "Homo-variance p_val")
kable(results, format = "markdown")
```

	full_additive	raw_aic	manual_selected_aic
Adj.R2	0.8206541	0.8823423	0.9030391
LOOCV test rmse	4.0540040	3.3925330	2.8104400
Normality p_val	0.0000000	0.0000000	0.0000042
Homo-variance p_val	0.0000000	0.0000000	0.0381380

LOOCV is a better indicator of generallization ability. So we will use this instead of test rmse.

We have fitted three models inside linear family: full additive, raw aic, and manual selected model based on AIC. We can easily conclude that the third column, our manually selected model based on AIC performs the best in all of the indicators. This model has the highest Ajusted R square, lowest LOOCV rmse(the best gerneralization ability), and lowest evidence of violating our LINE assumptions. Though the only pity of this model is we cannot satisfy the normality assuption because of potential lack of information.

Considering more families

```
results = data.frame(test_rmse = c(81.09287,3.166643,3.370076,0.8169445,2.418823),
                     training_selection_time = c("less than 1 min", "around 10 mins","over 5 hours", "around 15 mins"
, "less than 1 min"))
rownames(results) = c("MLR(full additive)","MLR(raw AIC)","MLR(Manual selected AIC)", "Random Forest", "KNN(with sca
ling)")
kable(results, format = "markdown")
```

	test_rmse	training_selection_time
MLR(full additive)	81.0928700	less than 1 min
MLR(raw AIC)	3.1666430	around 10 mins
MLR(Manual selected AIC)	3.3700760	over 5 hours
Random Forest	0.8169445	around 15 mins
KNN(with scaling)	2.4188230	less than 1 min

The method we choose to measure the generalization ability is test rmse in this scenerio. Otherwise the waiting time would be uncceptably long

We can see from the table that our random forest model beat other two models with **HUGE** advatange. The test rmse is really low so we do not worry about overfitting. We can also see that the non-linear family, either RF or KNN outperforms our linear model.

Summing up

Inside the linear family, we will choose our manually selected model based on AIC.(the precise model is not shown because of the heavy parameters)

Beyond this, we will prefer the Random Forest model.

Discussion

Final Linear Model

```
mean = mean(life_data$Life.expectancy)
loocv = 2.8104400
cat("error percent:", loocv / mean)
```

```
## error percent: 0.04059867
```

```
cat("\nthat is about", (loocv / mean) * mean, "years' error")
```

```
##
## that is about 2.81044 years' error
```

In the linear family scope, our best selection would have estimated error percentage of about 4%.(calculated by loocv RMSE / mean response)
This is super satifying. The generalization ability is guranteed, and the low error rate would make sure our model is useful in prediction. We can see from the previous section that our adjusted R square is also over 90%. With these limited information, I think this is already an awesome linear model. Given enough information, our model would be confident to give you a prediction with error around 2.81044 years.

Final Overall Model

```
rf_rmse = 0.8169445
cat("error percent:", rf_rmse / mean)
```

```
## error percent: 0.0118013
```

```
cat("\nthat is about", (rf_rmse/mean)*mean,"years' error")
```

```
##
## that is about 0.8169445 years' error
```

When not limited to linear family, we would prefer the Random Forest model. The error rate is around 1%, and that is about 0.8169 years' error! This is really amazing.

Put it into a more general case

We would NEVER choose a linear model for prototyping a complex modeling task. In this simple study(which is not even comparable to the simplist Titanic Kaggle competition), we spent over **FIVE** hours trying to manually select useful interactive & polynomial features while obtaining not satisfying prediction. The training of KNN only takes few seconds(while giving better results than the best MLR), while the training of random forest takes about 15 mins and give us super good predictions(only 1% error!). Random forest is really the best model for prototyping.

Appendix

Some interpretation

```
importance(rf$finalModel)
```

```
##                                  IncNodePurity
## Year                                1252.6310
## StatusDeveloping                     774.7990
## Adult.Mortality                    27079.7573
## infant.deaths                       2418.7429
## Alcohol                             1433.0314
## percentage.expenditure               661.3213
## Hepatitis.B                          512.5422
## Measles                              640.5772
## BMI                                 4340.9375
## Polio                               1596.7390
## Total.expenditure                   1305.7754
## Diphtheria                          1167.4476
## HIV.AIDS                           58111.9335
## GDP                                  825.6973
## Population                           627.5607
## thinness..1.19.years                2815.0294
## Income.composition.of.resources    53017.1438
## Schooling                          11816.5101
```

We can easily retrieve the feature importance with the help of random forest. **We then need to create a simpler model to interpret the feature effects. Our best model for prediction has too many parameters and is impossible to predict.**

```
simple_lm = lm(data = life_data, Life.expectancy ~ Adult.Mortality + HIV.AIDS + Income.composition.of.resources + Sc
hooling + BMI)
test_rmse("simple lm",simple_lm)
```

```
## Model simple lm has test rmse: 4.294979
```

```
cat("\nthe Adjust R2 of simple lm:",summary(simple_lm)$adj)
```

```
## 
## the Adjust R2 of simple lm: 0.7936622
```

```
cat("\nloocv of simple lm:",calc_loocv_rmse(simple_lm))
```

```
## 
## loocv of simple lm: 4.339406
```

We take the top five important features into consideration and use this to fit a much simpler linear model. We see that **Over 80% of the overall effect of model is accredited to only the top 20% features.**

```
summary(simple_lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + HIV.AIDS + Income.composition.of.resources +
##     Schooling + BMI, data = life_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.6999  -2.1886  -0.0867   2.2801  22.8986
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     54.2443168  0.4049418  133.96   <2e-16 ***
## Adult.Mortality                 -0.0197702  0.0008464  -23.36   <2e-16 ***
## HIV.AIDS                        -0.5072988  0.0185156  -27.40   <2e-16 ***
## Income.composition.of.resources  7.9532617  0.6128265   12.98   <2e-16 ***
## Schooling                        1.0153309  0.0406038   25.01   <2e-16 ***
## BMI                              0.0528240  0.0049375   10.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.326 on 2922 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.7937
## F-statistic:  2253 on 5 and 2922 DF,  p-value: < 2.2e-16
```

- We observe that all of these important features has extremely low p_values.
- When all of the features are `0`, the life expectancy is predicted to be **54.2298933**.
- For each unit increment of `Adult Mortality`, the life expectancy is predicted to decrese **0.0195968**.
- For each unit increment of `HIV.AIDS`, the life expectancy is predicted to decrese **0.5015563**.
- For each unit increment of `Income.composition.of.resources`, the life expectancy is predicted to increase **8.0312173**.
- For each unit increment of `Schooling`, the life expectancy is predicted to increase **1.0152195**.
- For each unit increment of `BMI`, the life expectancy is predicted to increase **0.0514571**.

These are the effects of important features on Life expectancy.

Other conclusions

- The random forest model has the lowest test RMSE(Though it’s completely uninterpretable.). Followed by KNN. (Also not quite interpretable) The linear model might be easier to intepret, its accuracy is really bad. Accuracy and interpretablility has is negatively related to each other.

- Try to use random forest for prototyping instead of linear model. The underlying relationship between repsonse and features might be complicated, and can not be easily found out through feature selection process(this may cost you hours and still give bad results.) The lack of information might never make you able to satisfy LINE assumptions and find good model, so try non-linear models to make life easier instead.

- For model with very large amount of parameters(even raw additive model), the BIC selection might be so much that it simply the model to a null model(always predict with average).

- The **twenty-eighty rule** really robust. This rule would enable you to interpret the model with just a few important features.

- It’s really important to find good teammates to avoid do it all.