# Life Expectancy Data

Tianxing Li, Yaodong Wang, Yao Xiao, Donglai Xu

# Overview

Country

Year

Status（developing/developed)

Life expectancy

Adult Mortality

Infant deaths

Alcohol

Percentage.expenditure

Hepatitis B

Measles

BMI

under-five deaths

Polio

Total expenditure

Diphtheria

HIV/AIDS

GDP

Population

thinness 1-19 years

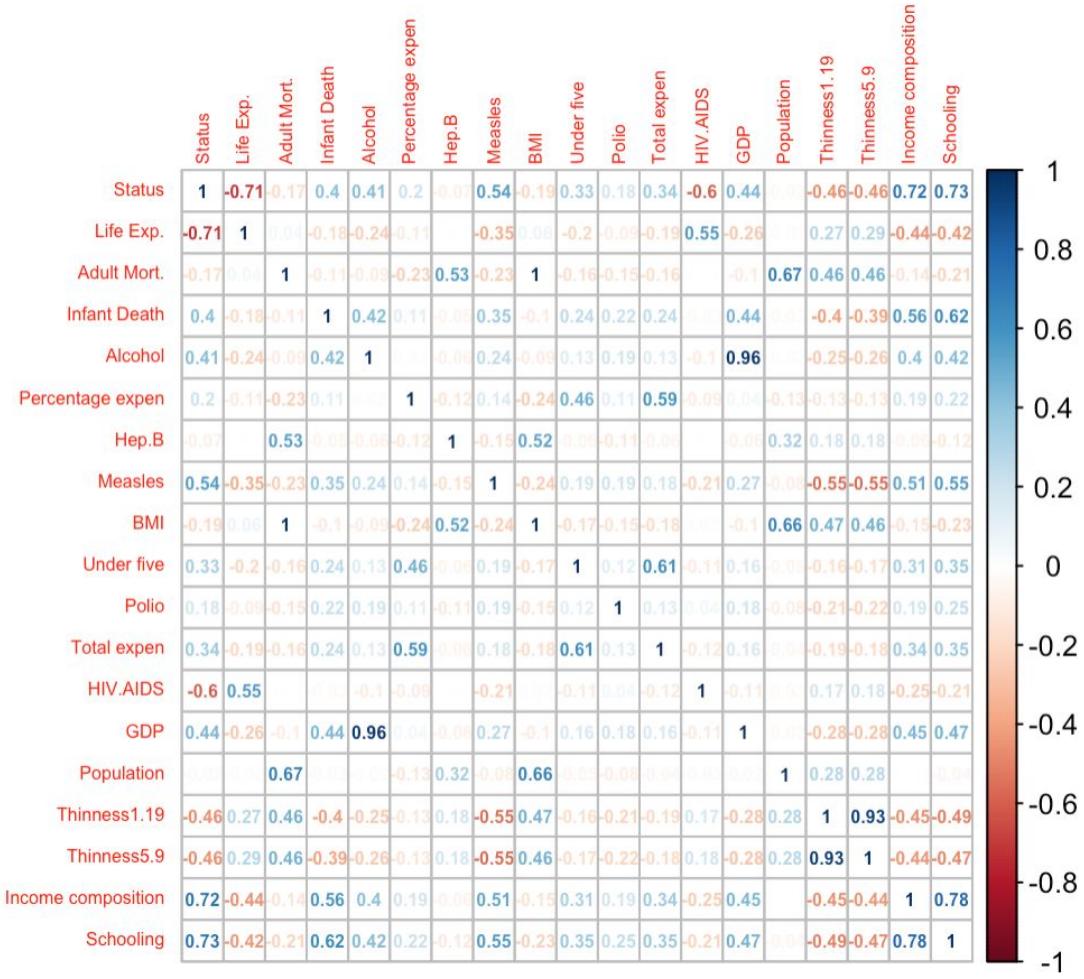
thinness 5-9 years

Income.composition.of .resources

Schooling

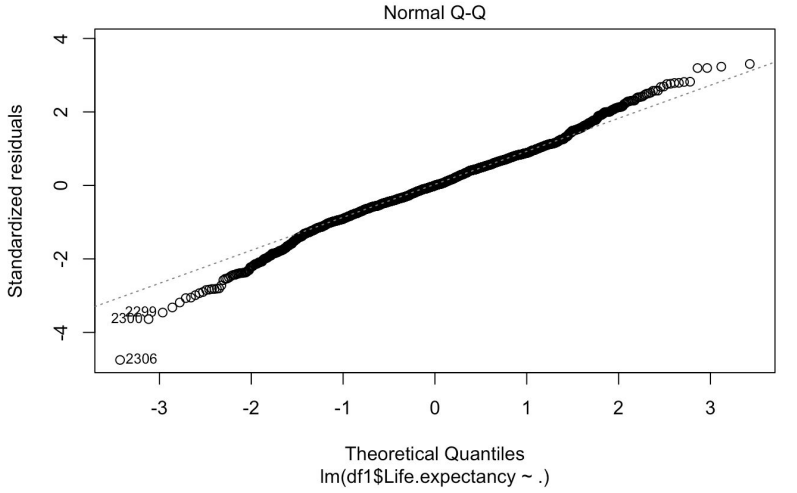
Track of the health status as well as many other related factors for 193 countries from 2000-2015.

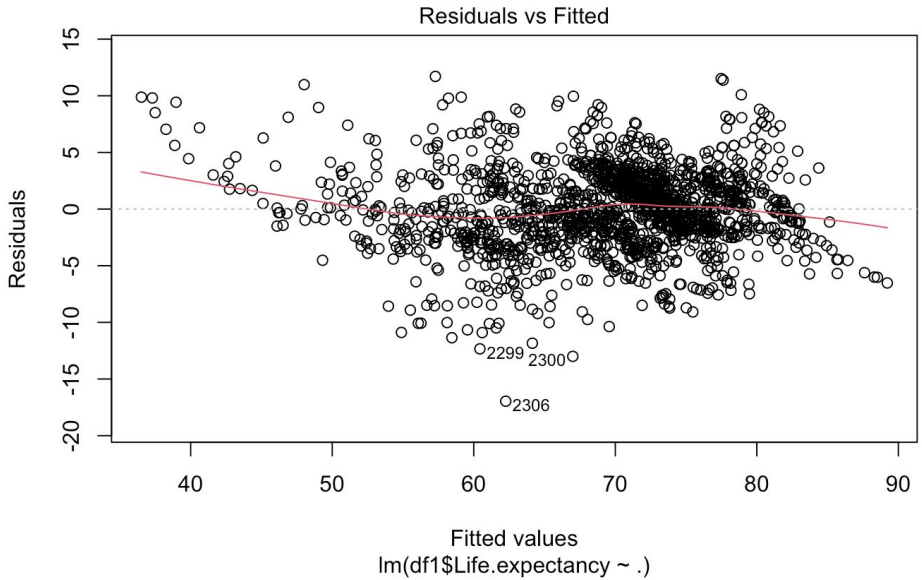
```
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                         : Factor w/ 2 levels "Developed","Developing": 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 ...
```


Status
Life Exp.
Adult Mort.
Infant Death
Alcohol
Percentage expen
Hep.B
Measles
BMI
Under five
Polio
Total expen
HIV.AIDS
GDP
Population
Thinness1.19
Thinness5.9
Income composition
Schooling
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

# Read life data

```r
# Read Life dataset
```{r}
life<-read.csv("Life Expectancy Data.csv")
df<-life[,-c(1,2)]
df1<-na.omit(df)# remove NA
full<-lm(df1$Life.expectancy~.,data=df1)
plot(full)
```



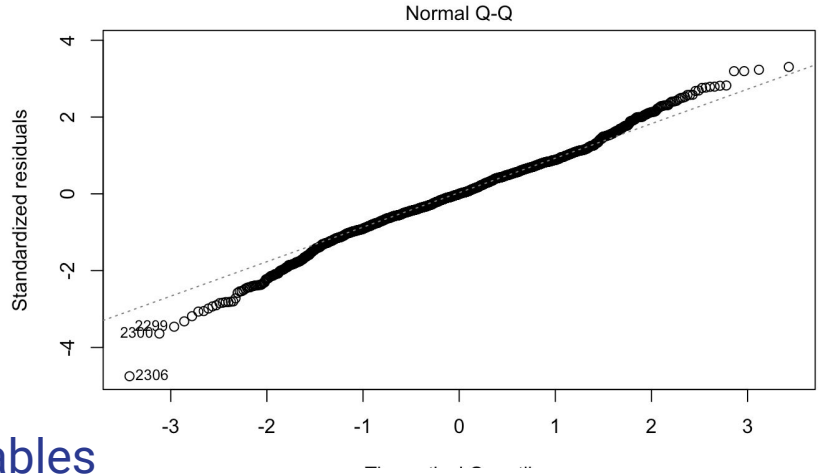


```

## Split datasets into train/test

```
sampleindex<-sample(nrow(df2),0.75*nrow(df2))
dftrain<-df2[sampleindex,] # 75% of datasets
dftest<-df2[-sampleindex,] # 25% of datasets
```

## Remove NA

```
df<-na.omit(df)# remove NA
```

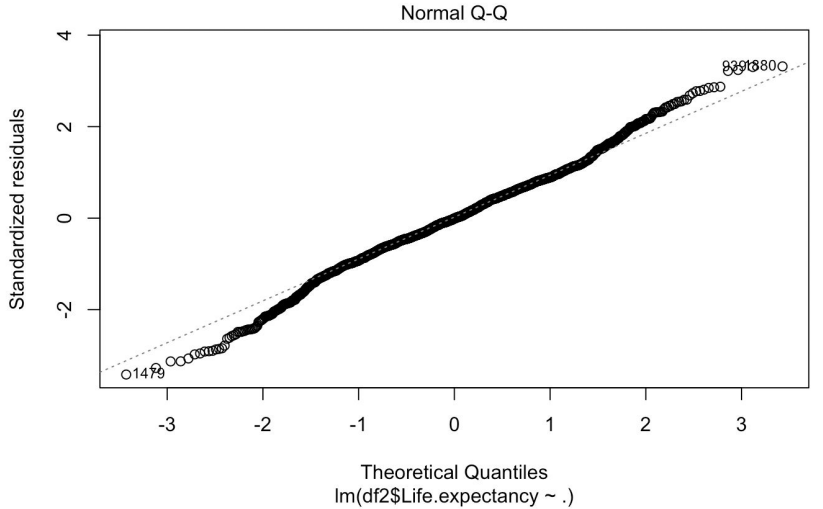



## Convert Categorical variables into Dummy Variables

```
levels(df1$Status) <- c(1,0)
# Change catagorical variable into dummy variables.
```

## Remove outliers

```
df2<-na.omit(df1[-c(2299,2300,2306),])
# Remove outliers
```



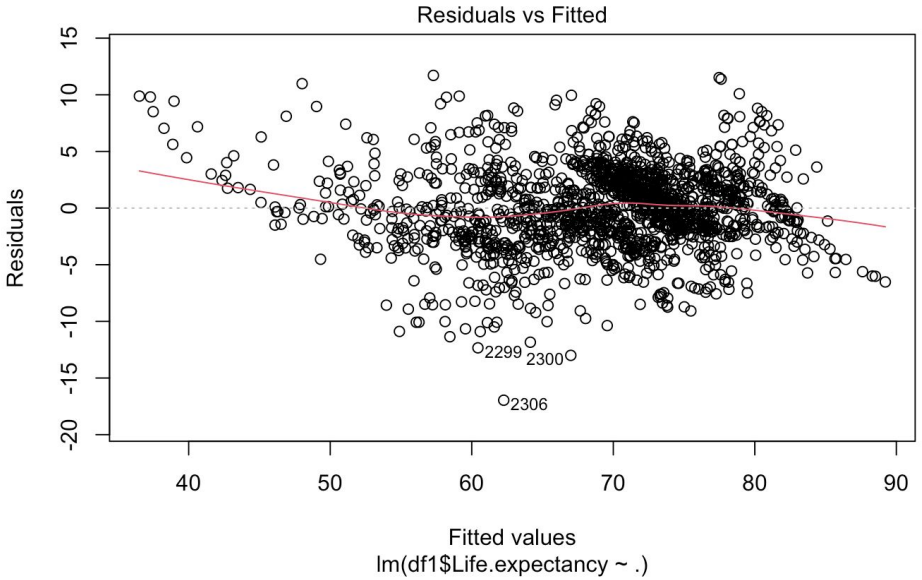
# BP test

```
library(lmtest)
bptest(full1,studentize=FALSE)
qchisq(0.99, 19)
```

```
        Breusch-Pagan test

data:  full1
BP = 259.06, df = 19, p-value < 2.2e-16

[1] 36.19087
```



# Stepwise Regression (Forward)

```
library(olsrr)
ols_step_forward_aic(full, details = TRUE)
```

```
Forward Selection Method
------------------------

Candidate Terms:

1 . Status
2 . Adult.Mortality
3 . infant.deaths
4 . Alcohol
5 . percentage.expenditure
6 . Hepatitis.B
7 . Measles
8 . BMI
9 . under.five.deaths
10 . Polio
11 . Total.expenditure
12 . Diphtheria
13 . HIV.AIDS
14 . GDP
15 . Population
16 . thinness..1.19.years
17 . thinness.5.9.years
18 . Income.composition.of.resources
19 . Schooling

 Step 0: AIC = 11853.8
 df1$Life.expectancy ~ 1
```

```
Step 0: AIC = 11853.8
df1$Life.expectancy ~ 1
```

| Variable | DF | AIC | Sum Sq | RSS | R-Sq | Adj. R-Sq |
|---|---|---|---|---|---|---|
| Schooling | 1 | 10612.716 | 67519.815 | 60009.496 | 0.529 | 0.529 |
| Income.composition.of.resources | 1 | 10645.625 | 66310.154 | 61219.157 | 0.520 | 0.520 |
| Adult.Mortality | 1 | 10733.977 | 62940.644 | 64588.667 | 0.494 | 0.493 |
| HIV.AIDS | 1 | 11143.554 | 44730.119 | 82799.193 | 0.351 | 0.350 |
| BMI | 1 | 11282.166 | 37469.270 | 90060.041 | 0.294 | 0.293 |
| thinness..1.19.years | 1 | 11467.899 | 26732.160 | 100797.151 | 0.210 | 0.209 |
| thinness.5.9.years | 1 | 11468.529 | 26693.649 | 100835.662 | 0.209 | 0.209 |
| Status | 1 | 11495.921 | 25004.633 | 102524.678 | 0.196 | 0.196 |
| GDP | 1 | 11498.595 | 24838.239 | 102691.073 | 0.195 | 0.194 |
| percentage.expenditure | 1 | 11552.916 | 21399.088 | 106130.223 | 0.168 | 0.167 |
| Alcohol | 1 | 11564.006 | 20682.965 | 106846.347 | 0.162 | 0.162 |
| Diphtheria | 1 | 11651.539 | 14858.059 | 112671.252 | 0.117 | 0.116 |
| Polio | 1 | 11668.964 | 13661.147 | 113868.164 | 0.107 | 0.107 |
| Hepatitis.B | 1 | 11788.532 | 5097.871 | 122431.440 | 0.040 | 0.039 |
| under.five.deaths | 1 | 11793.691 | 4714.242 | 122815.070 | 0.037 | 0.036 |
| Total.expenditure | 1 | 11804.681 | 3892.992 | 123636.319 | 0.031 | 0.030 |
| infant.deaths | 1 | 11807.978 | 3645.547 | 123883.765 | 0.029 | 0.028 |
| Measles | 1 | 11847.961 | 605.078 | 126924.233 | 0.005 | 0.004 |
| Population | 1 | 11854.983 | 63.447 | 127465.864 | 0.000 | 0.000 |

```
Step 1 : AIC = 10612.72
df1$Life.expectancy ~ Schooling
```

| Variable | DF | AIC | Sum Sq | RSS | R-Sq | Adj. R-Sq |
|---|---|---|---|---|---|---|
| HIV.AIDS | 1 | 9696.327 | 25626.363 | 34383.133 | 0.730 | 0.730 |
| Adult.Mortality | 1 | 9757.854 | 24319.236 | 35690.260 | 0.720 | 0.720 |
| Income.composition.of.resources | 1 | 10395.281 | 7477.022 | 52532.473 | 0.588 | 0.588 |
| BMI | 1 | 10514.885 | 3525.202 | 56484.294 | 0.557 | 0.557 |
| thinness.5.9.years | 1 | 10555.318 | 2123.084 | 57886.412 | 0.546 | 0.546 |
| thinness..1.19.years | 1 | 10567.462 | 1695.222 | 58314.274 | 0.543 | 0.542 |
| GDP | 1 | 10568.456 | 1660.047 | 58349.449 | 0.542 | 0.542 |
| percentage.expenditure | 1 | 10569.291 | 1630.518 | 58378.978 | 0.542 | 0.542 |
| Diphtheria | 1 | 10584.642 | 1084.522 | 58924.974 | 0.538 | 0.537 |
| Status | 1 | 10591.357 | 844.058 | 59165.438 | 0.536 | 0.536 |
| Polio | 1 | 10593.577 | 764.357 | 59245.139 | 0.535 | 0.535 |
| Alcohol | 1 | 10602.588 | 439.715 | 59569.781 | 0.533 | 0.532 |
| Hepatitis.B | 1 | 10607.792 | 251.439 | 59758.057 | 0.531 | 0.531 |
| under.five.deaths | 1 | 10611.857 | 103.953 | 59905.543 | 0.530 | 0.530 |
| Measles | 1 | 10613.887 | 30.166 | 59979.329 | 0.530 | 0.529 |
| infant.deaths | 1 | 10614.086 | 22.906 | 59986.590 | 0.530 | 0.529 |
| Population | 1 | 10614.542 | 6.308 | 60003.187 | 0.529 | 0.529 |
| Total.expenditure | 1 | 10614.689 | 0.964 | 60008.532 | 0.529 | 0.529 |

```
+ HIV.AIDS
```

```
 Step 12 : AIC = 8909.036
 df1$Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
percentage.expenditure + BMI + Diphtheria + under.five.deaths + infant.deaths + Status + Alcohol +
thinness.5.9.years

-------------------------------------------------------------------------------
Variable                DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
-------------------------------------------------------------------------------
Total.expenditure        1    8907.600    43.856    21029.340    0.835        0.834
Polio                    1    8909.227    23.101    21050.095    0.835        0.834
Hepatitis.B              1    8909.724    16.751    21056.445    0.835        0.834
Measles                  1    8910.471     7.218    21065.978    0.835        0.834
Population               1    8910.960     0.966    21072.230    0.835        0.833
GDP                      1    8910.972     0.814    21072.383    0.835        0.833
thinness..1.19.years     1    8911.010     0.329    21072.867    0.835        0.833
-------------------------------------------------------------------------------

+ Total.expenditure


 Step 13 : AIC = 8907.6
 df1$Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
percentage.expenditure + BMI + Diphtheria + under.five.deaths + infant.deaths + Status + Alcohol +
thinness.5.9.years + Total.expenditure

-------------------------------------------------------------------------------
Variable                DF      AIC       Sum Sq       RSS       R-Sq     Adj. R-Sq
-------------------------------------------------------------------------------
Polio                    1    8907.851    22.292    21007.048    0.835        0.834
Hepatitis.B              1    8908.097    19.158    21010.183    0.835        0.834
Measles                  1    8909.150     5.740    21023.600    0.835        0.834
GDP                      1    8909.498     1.303    21028.037    0.835        0.834
Population               1    8909.525     0.960    21028.380    0.835        0.834
thinness..1.19.years     1    8909.567     0.428    21028.912    0.835        0.834
-------------------------------------------------------------------------------


No more variables to be added.
```

```
No more variables to be added.

Variables Entered:

+ Schooling
+ HIV.AIDS
+ Adult.Mortality
+ Income.composition.of.resources
+ percentage.expenditure
+ BMI
+ Diphtheria
+ under.five.deaths
+ infant.deaths
+ Status
+ Alcohol
+ thinness.5.9.years
+ Total.expenditure
```

Final Model Output

Model Summary

| | | | |
|---|---|---|---|
| R | 0.914 | RMSE | 3.586 |
| R-Squared | 0.835 | Coef. Var | 5.175 |
| Adj. R-Squared | 0.834 | MSE | 12.862 |
| Pred R-Squared | 0.831 | MAE | 2.737 |

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

| | Sum of Squares | DF | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 106499.971 | 13 | 8192.305 | 636.94 | 0.0000 |
| Residual | 21029.340 | 1635 | 12.862 | | |
| Total | 127529.311 | 1648 | | | |

Parameter Estimates

| model | Beta | Std. Error | Std. Beta | t | Sig | lower | upper |
|---|---|---|---|---|---|---|---|
| (Intercept) | 54.437 | 0.813 | | 66.952 | 0.000 | 52.842 | 56.032 |
| Schooling | 0.874 | 0.059 | 0.278 | 14.897 | 0.000 | 0.759 | 0.989 |
| HIV.AIDS | -0.436 | 0.018 | -0.299 | -24.479 | 0.000 | -0.471 | -0.401 |
| Adult.Mortality | -0.017 | 0.001 | -0.238 | -17.669 | 0.000 | -0.019 | -0.015 |
| Income.composition.of.resources | 9.871 | 0.829 | 0.205 | 11.910 | 0.000 | 8.245 | 11.496 |
| percentage.expenditure | 0.000 | 0.000 | 0.085 | 7.196 | 0.000 | 0.000 | 0.001 |
| BMI | 0.034 | 0.006 | 0.076 | 5.662 | 0.000 | 0.022 | 0.045 |
| Diphtheria | 0.015 | 0.005 | 0.037 | 3.291 | 0.001 | 0.006 | 0.024 |
| under.five.deaths | -0.070 | 0.007 | -1.287 | -9.402 | 0.000 | -0.084 | -0.055 |
| infant.deaths | 0.092 | 0.010 | 1.262 | 9.229 | 0.000 | 0.072 | 0.111 |
| StatusDeveloping | -0.946 | 0.336 | -0.038 | -2.817 | 0.005 | -1.605 | -0.287 |
| Alcohol | -0.087 | 0.033 | -0.040 | -2.654 | 0.008 | -0.152 | -0.023 |
| thinness.5.9.years | -0.057 | 0.026 | -0.030 | -2.158 | 0.031 | -0.109 | -0.005 |
| Total.expenditure | 0.075 | 0.041 | 0.020 | 1.847 | 0.065 | -0.005 | 0.154 |

# Stepwise Regression (Backward)

```r
ols_step_backward_aic(full, details = TRUE)
```

```
 Step 0: AIC = 8914.947
 df1$Life.expectancy ~ Status + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure +
Hepatitis.B + Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP +
Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources + Schooling

-------------------------------------------------------------------------------------------
Variable                          DF        AIC         Sum Sq         RSS       R-Sq      Adj. R-Sq
-------------------------------------------------------------------------------------------
thinness..1.19.years              1      8913.003          0.703    20970.790     0.836          0.834
GDP                               1      8913.043          1.221    20971.308     0.836          0.834
Population                        1      8913.084          1.738    20971.824     0.836          0.834
Measles                           1      8913.486          6.846    20976.932     0.836          0.834
thinness.5.9.years                1      8913.799         10.830    20980.917     0.835          0.834
Hepatitis.B                       1      8915.124         27.692    20997.779     0.835          0.834
Polio                             1      8915.347         30.539    21000.626     0.835          0.834
Total.expenditure                 1      8916.465         44.776    21014.862     0.835          0.833
percentage.expenditure            1      8917.153         53.548    21023.635     0.835          0.833
Diphtheria                        1      8919.330         81.323    21051.410     0.835          0.833
Alcohol                           1      8920.618         97.770    21067.857     0.835          0.833
Status                            1      8921.238        105.701    21075.788     0.835          0.833
BMI                               1      8944.700        407.704    21377.791     0.832          0.831
infant.deaths                     1      8989.135        991.593    21961.680     0.828          0.826
under.five.deaths                 1      8995.113       1071.356    22041.442     0.827          0.825
Income.composition.of.resources   1      9048.129       1791.519    22761.606     0.822          0.820
Schooling                         1      9115.385       2739.056    23709.143     0.814          0.812
Adult.Mortality                   1      9197.540       3950.190    24920.277     0.805          0.802
HIV.AIDS                          1      9429.874       7720.690    28690.777     0.775          0.773
-------------------------------------------------------------------------------------------


Variables Removed:

x thinness..1.19.years
```

```
  Step 3 : AIC = 8909.243
 df1$Life.expectancy ~ Status + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure +
Hepatitis.B + Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS +
thinness.5.9.years + Income.composition.of.resources + Schooling

------------------------------------------------------------------------------------------------
Variable                         DF      AIC         Sum Sq        RSS      R-Sq     Adj. R-Sq
------------------------------------------------------------------------------------------------
Measles                          1     8907.733       6.226    20980.073    0.835        0.834
Hepatitis.B                      1     8909.390      27.326    21001.173    0.835        0.834
Polio                            1     8909.626      30.328    21004.175    0.835        0.834
Total.expenditure                1     8910.718      44.246    21018.093    0.835        0.834
thinness.5.9.years               1     8912.118      62.092    21035.939    0.835        0.834
Diphtheria                       1     8913.600      81.004    21054.851    0.835        0.833
Alcohol                          1     8914.849      96.961    21070.808    0.835        0.833
Status                           1     8915.672     107.485    21081.331    0.835        0.833
BMI                              1     8939.050     408.485    21382.332    0.832        0.831
percentage.expenditure           1     8956.722     638.861    21612.708    0.831        0.829
infant.deaths                    1     8986.578    1033.738    22007.585    0.827        0.826
under.five.deaths                1     8990.998    1092.811    22066.658    0.827        0.825
Income.composition.of.resources  1     9044.087    1814.789    22788.636    0.821        0.820
Schooling                        1     9112.931    2786.334    23760.181    0.814        0.812
Adult.Mortality                  1     9192.551    3961.723    24935.570    0.804        0.803
HIV.AIDS                         1     9424.208    7722.747    28696.594    0.775        0.773
------------------------------------------------------------------------------------------------

x Measles
```

```
  Step 4 : AIC = 8907.733
 df1$Life.expectancy ~ Status + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure +
Hepatitis.B + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + thinness.5.9.years
+ Income.composition.of.resources + Schooling

------------------------------------------------------------------------------------------------
Variable                         DF      AIC         Sum Sq        RSS      R-Sq     Adj. R-Sq
------------------------------------------------------------------------------------------------
Hepatitis.B                      1     8907.851      26.976    21007.048    0.835        0.834
Polio                            1     8908.097      30.110    21010.183    0.835        0.834
Total.expenditure                1     8909.328      45.798    21025.870    0.835        0.834
thinness.5.9.years               1     8910.267      57.773    21037.846    0.835        0.834
Diphtheria                       1     8912.086      80.996    21061.069    0.835        0.833
Alcohol                          1     8913.519      99.307    21079.380    0.835        0.833
Status                           1     8914.261     108.786    21088.859    0.835        0.833
BMI                              1     8938.570     421.976    21402.049    0.832        0.831
percentage.expenditure           1     8955.382     641.286    21621.358    0.830        0.829
infant.deaths                    1     8987.346    1064.482    22044.555    0.827        0.826
under.five.deaths                1     8990.634    1108.484    22088.557    0.827        0.825
Income.composition.of.resources  1     9042.564    1815.168    22795.241    0.821        0.820
Schooling                        1     9112.071    2796.544    23776.617    0.814        0.812
Adult.Mortality                  1     9190.784    3959.015    24939.087    0.804        0.803
HIV.AIDS                         1     9423.364    7736.637    28716.710    0.775        0.773
------------------------------------------------------------------------------------------------


No more variables to be removed.
```

Variables Removed:

x thinness..1.19.years
x GDP
x Population
x Measles

| model | Beta |
|---|---|
| (Intercept) | 54.388 |
| StatusDeveloping | -0.980 |
| Adult.Mortality | -0.017 |
| infant.deaths | 0.091 |
| Alcohol | -0.092 |
| percentage.expenditure | 0.000 |
| Hepatitis.B | -0.006 |
| BMI | 0.034 |
| under.five.deaths | -0.069 |
| Polio | 0.008 |
| Total.expenditure | 0.077 |
| Diphtheria | 0.015 |
| HIV.AIDS | -0.437 |
| thinness.5.9.years | -0.056 |
| Income.composition.of.resources | 9.848 |
| Schooling | 0.869 |

| model | Beta |
|---|---|
| (Intercept) | 54.437 |
| Schooling | 0.874 |
| HIV.AIDS | -0.436 |
| Adult.Mortality | -0.017 |
| Income.composition.of.resources | 9.871 |
| percentage.expenditure | 0.000 |
| BMI | 0.034 |
| Diphtheria | 0.015 |
| under.five.deaths | -0.070 |
| infant.deaths | 0.092 |
| StatusDeveloping | -0.946 |
| Alcohol | -0.087 |
| thinness.5.9.years | -0.057 |
| Total.expenditure | 0.075 |

# Stepwise Regression (Both direction)

```r
ols_step_both_aic(full, details=T)
```

```
 Step 10 : AIC = 8915.624
 df1$Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
percentage.expenditure + BMI + Diphtheria + under.five.deaths + infant.deaths + Status

                                Remove Existing Variables
-----------------------------------------------------------------------------------------------------
Variable                          DF      AIC         Sum Sq         RSS      R-Sq     Adj. R-Sq
-----------------------------------------------------------------------------------------------------
Status                             1    8918.221    106261.163    21268.148    0.833         0.832
Diphtheria                         1    8923.872    106188.155    21341.156    0.833         0.832
BMI                                1    8960.588    105707.642    21821.669    0.829         0.828
percentage.expenditure             1    8964.911    105650.365    21878.947    0.828         0.827
infant.deaths                      1    9002.614    105144.365    22384.947    0.824         0.824
under.five.deaths                  1    9007.382    105079.535    22449.776    0.824         0.823
Income.composition.of.resources    1    9047.856    104521.707    23007.604    0.820         0.819
Schooling                          1    9125.036    103419.254    24110.058    0.811         0.810
Adult.Mortality                    1    9217.875    102022.921    25506.391    0.800         0.799
HIV.AIDS                           1    9440.773     98331.282    29198.029    0.771         0.770
-----------------------------------------------------------------------------------------------------

                                Enter New Variables
-----------------------------------------------------------------------------------------------
Variable                DF      AIC         Sum Sq         RSS      R-Sq     Adj. R-Sq
-----------------------------------------------------------------------------------------------
Alcohol                  1    8912.288    106388.885    21140.426    0.834         0.833
thinness.5.9.years       1    8913.882    106368.439    21160.872    0.834         0.833
Total.expenditure        1    8913.949    106367.585    21161.727    0.834         0.833
thinness..1.19.years     1    8914.867    106355.802    21173.509    0.834         0.833
Polio                    1    8916.162    106339.166    21190.145    0.834         0.833
Hepatitis.B              1    8916.385    106336.300    21193.011    0.834         0.833
Measles                  1    8917.304    106324.484    21204.827    0.834         0.833
Population               1    8917.538    106321.469    21207.842    0.834         0.833
GDP                      1    8917.548    106321.349    21207.963    0.834         0.833
-----------------------------------------------------------------------------------------------

+ Alcohol
```

Parameter Estimates

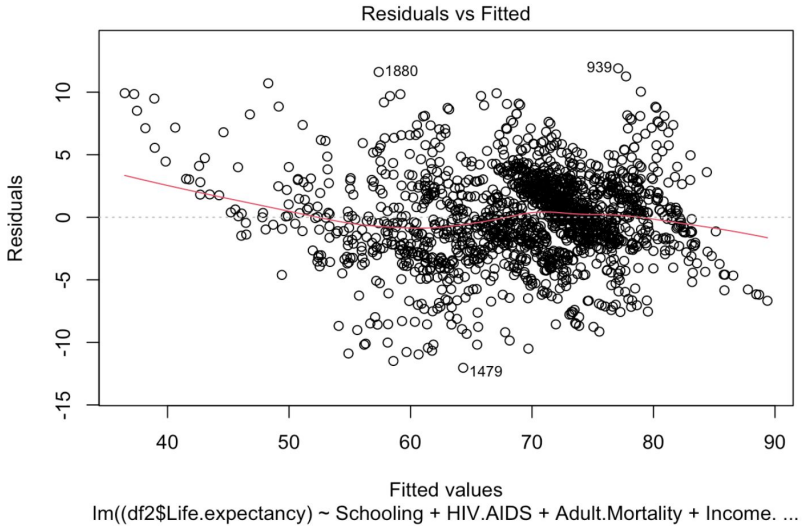
| model | Beta | Std. Error | Std. Beta | t | Sig | lower | upper |
|---|---|---|---|---|---|---|---|
| (Intercept) | 54.437 | 0.813 | | 66.952 | 0.000 | 52.842 | 56.032 |
| Schooling | 0.874 | 0.059 | 0.278 | 14.897 | 0.000 | 0.759 | 0.989 |
| HIV.AIDS | -0.436 | 0.018 | -0.299 | -24.479 | 0.000 | -0.471 | -0.401 |
| Adult.Mortality | -0.017 | 0.001 | -0.238 | -17.669 | 0.000 | -0.019 | -0.015 |
| Income.composition.of.resources | 9.871 | 0.829 | 0.205 | 11.910 | 0.000 | 8.245 | 11.496 |
| percentage.expenditure | 0.000 | 0.000 | 0.085 | 7.196 | 0.000 | 0.000 | 0.001 |
| BMI | 0.034 | 0.006 | 0.076 | 5.662 | 0.000 | 0.022 | 0.045 |
| Diphtheria | 0.015 | 0. | | | | | 024 |
| under.five.deaths | -0.070 | 0. | | | | | 055 |
| infant.deaths | 0.092 | 0. | | | | | .11 |
| StatusDeveloping | -0.946 | 0. | | | | | 87 |
| Alcohol | -0.087 | 0. | | | | | 023 |
| thinness.5.9.years | -0.057 | 0. | | | | | 005 |
| Total.expenditure | 0.075 | 0. | | | | | .54 |

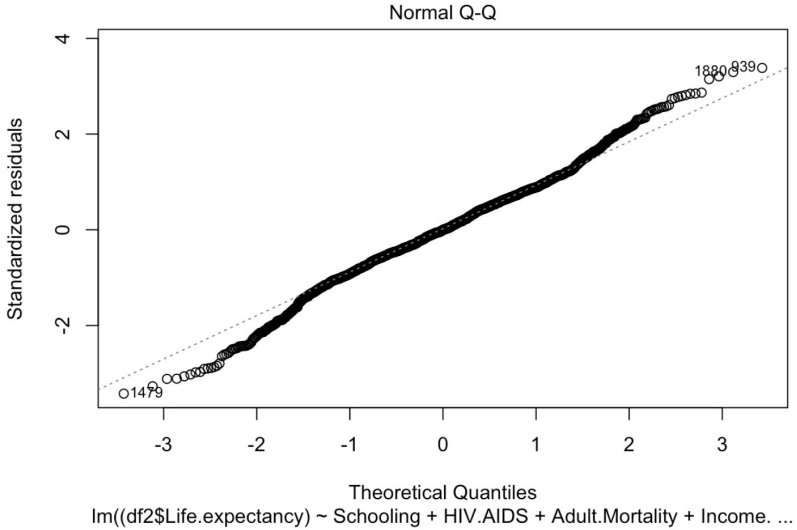
| model | Beta |
|---|---|
| (Intercept) | 54.437 |
| Schooling | 0.874 |
| HIV.AIDS | -0.436 |
| Adult.Mortality | -0.017 |
| Income.composition.of.resources | 9.871 |
| percentage.expenditure | 0.000 |
| BMI | 0.034 |
| Diphtheria | 0.015 |
| under.five.deaths | -0.070 |
| infant.deaths | 0.092 |
| StatusDeveloping | -0.946 |
| Alcohol | -0.087 |
| thinness.5.9.years | -0.057 |
| Total.expenditure | 0.075 |

```{r}
modelfor<-lm((df2$Life.expectancy) ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + percentage.expenditure +
    BMI + Diphtheria + under.five.deaths + infant.deaths + Status +
    Alcohol + thinness.5.9.years + Total.expenditure, data = df2)
```

# Plots after selection

```
plot(modelfor)
```



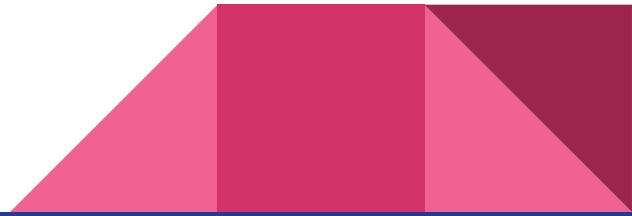
# BP Test

```
bptest(modelfor,studentize=FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  modelfor
## BP = 251.16, df = 13, p-value < 2.2e-16
```

```
qchisq(0.99, 13)
```

```
## [1] 27.68825
```



# Bonferroni outlier test

```r
stud.del.res <- rstudent(modelfor)
head(stud.del.res)
a<-0.10
n<-nrow(df1)
p<-ncol(df1)
stud.del.res[which(stud.del.res>qt(1-a/2/n,n-p-1))==F]
```

```
         1          2          3          4          5          6
 0.3117773 -1.0384853 -1.0039467 -1.0675144 -0.9851302 -0.9945638
named numeric(0)
```

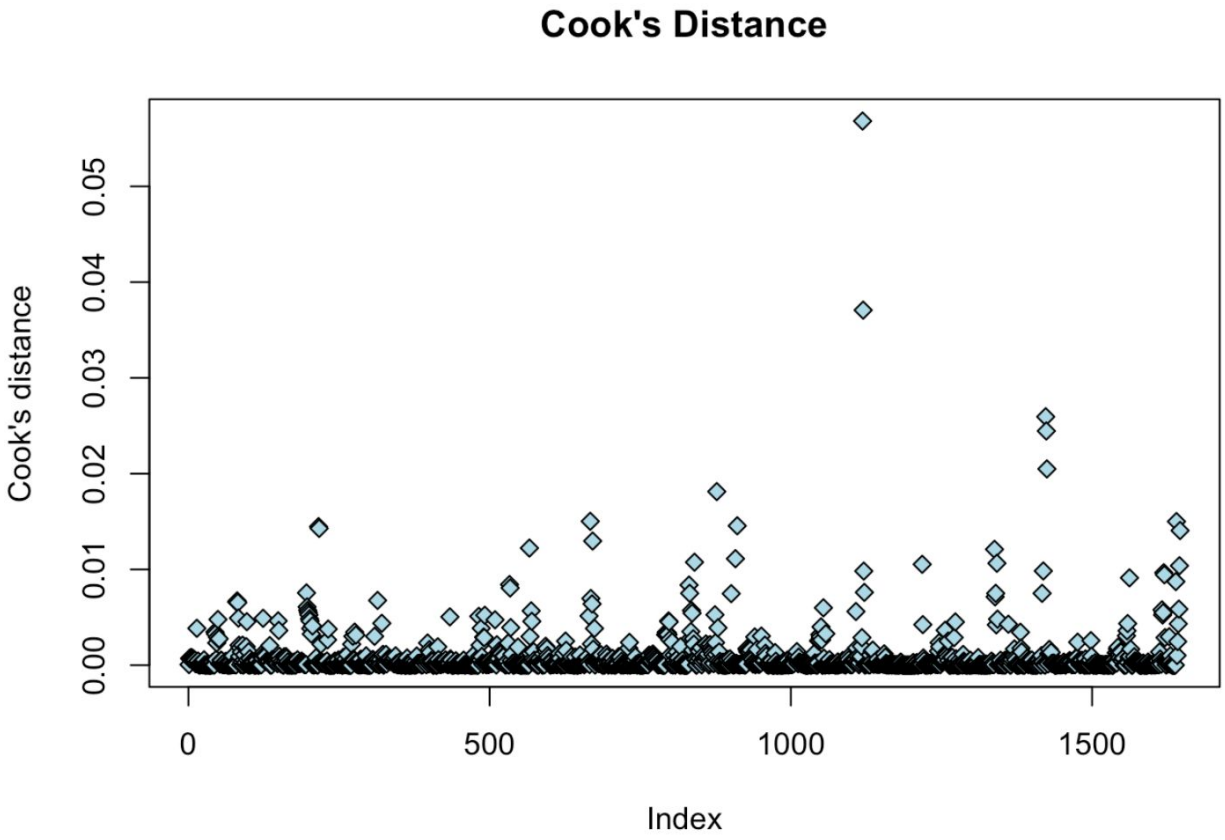
# Cook’s distance

```{r}
which(cooks.distance(modelfor)>0.05)
```

```
 1901
```

```
dffits(modelfor1)[1119]
p=13
n=1646
2*sqrt(p/n)# DFFITS
```

```
      1901
 0.09394921
 [1] 0.1777406
```



# Check VIF

```r
library(car)
vif(modelfor)
#pairs(cor(df1[,-1]))
```

```
                Schooling                   HIV.AIDS
                 3.441251                   1.482517
          Adult.Mortality Income.composition.of.resources
                 1.812906                   2.943903
  percentage.expenditure                        BMI
                 1.388648                   1.766704
              Diphtheria           under.five.deaths
                 1.219989                 185.911780
           infant.deaths                     Status
                185.415311                  1.811154
                 Alcohol           thinness.5.9.years
                 2.259792                   1.931070
       Total.expenditure
                 1.116564
```

# New model

```r
modelfor1<-lm((df2$Life.expectancy) ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + percentage.expenditure +
    BMI + Diphtheria + infant.deaths + Status +
    Alcohol + thinness.5.9.years + Total.expenditure, data = df2)
```

# VIF for new model

| | | |
|---|---|---|
| Schooling | Adult.Mortality | HIV.AIDS |
| 3.199916 | 1.766745 | 1.466934 |
| Income.composition.of.resources | percentage.expenditure | Diphtheria |
| 2.966306 | 1.403594 | 1.734698 |
| BMI | Total.expenditure | infant.deaths |
| 1.704111 | 1.109511 | 1.331757 |
| thinness.5.9.years | Status | Hepatitis.B |
| 1.838359 | 1.537646 | 1.637179 |

```
Call:
lm(formula = (df2$Life.expectancy) ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + percentage.expenditure +
    BMI + Diphtheria + infant.deaths + Status + Alcohol + thinness.5.9.years +
    Total.expenditure, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8022  -2.0964   0.0381   2.3533  11.9078

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.351e+01  8.165e-01  65.537  < 2e-16 ***
Schooling                       9.073e-01  5.931e-02  15.298  < 2e-16 ***
HIV.AIDS                       -4.381e-01  1.806e-02 -24.256  < 2e-16 ***
Adult.Mortality                -1.794e-02  9.568e-04 -18.753  < 2e-16 ***
Income.composition.of.resources  1.026e+01  8.377e-01  12.249  < 2e-16 ***
percentage.expenditure          4.165e-04  5.994e-05   6.949 5.31e-12 ***
BMI                             3.401e-02  6.027e-03   5.642 1.97e-08 ***
Diphtheria                      2.243e-02  4.510e-03   4.973 7.28e-07 ***
infant.deaths                  -1.471e-03  8.570e-04  -1.717   0.0862 .
StatusDeveloping               -9.460e-01  3.403e-01  -2.780   0.0055 **
Alcohol                        -1.390e-01  3.293e-02  -4.222 2.55e-05 ***
thinness.5.9.years             -4.354e-02  2.670e-02  -1.631   0.1031
Total.expenditure               7.550e-02  4.132e-02   1.827   0.0679 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.633 on 1633 degrees of freedom
Multiple R-squared:  0.8293,    Adjusted R-squared:  0.8281
F-statistic: 661.3 on 12 and 1633 DF,  p-value: < 2.2e-16
```

# Summary

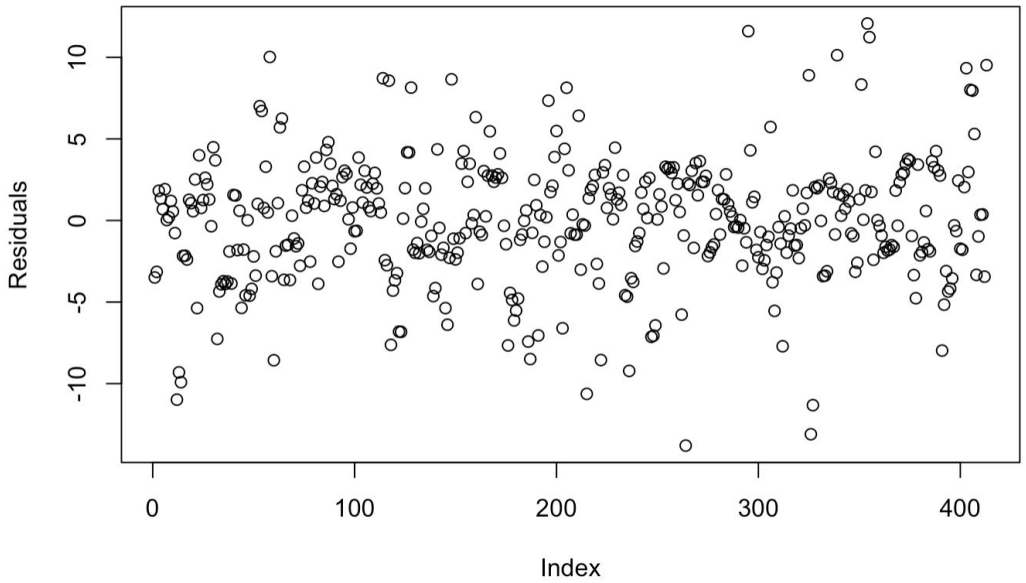
```
Call:
lm(formula = (df2$Life.expectancy) ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + percentage.expenditure +
    BMI + Diphtheria + infant.deaths + Status + Alcohol + thinness.5.9.years +
    Total.expenditure, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8022  -2.0964   0.0381   2.3533  11.9078

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.351e+01  8.165e-01  65.537  < 2e-16 ***
Schooling                        9.073e-01  5.931e-02  15.298  < 2e-16 ***
HIV.AIDS                        -4.381e-01  1.806e-02 -24.256  < 2e-16 ***
Adult.Mortality                 -1.794e-02  9.568e-04 -18.753  < 2e-16 ***
Income.composition.of.resources  1.026e+01  8.377e-01  12.249  < 2e-16 ***
percentage.expenditure           4.165e-04  5.994e-05   6.949 5.31e-12 ***
BMI                              3.401e-02  6.027e-03   5.642 1.97e-08 ***
Diphtheria                       2.243e-02  4.510e-03   4.973 7.28e-07 ***
infant.deaths                   -1.471e-03  8.570e-04  -1.717   0.0862 .
StatusDeveloping                -9.460e-01  3.403e-01  -2.780   0.0055 **
Alcohol                         -1.390e-01  3.293e-02  -4.222 2.55e-05 ***
thinness.5.9.years              -4.354e-02  2.670e-02  -1.631   0.1031
Total.expenditure                7.550e-02  4.132e-02   1.827   0.0679 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.633 on 1633 degrees of freedom
Multiple R-squared:  0.8293,    Adjusted R-squared:  0.8281
F-statistic: 661.3 on 12 and 1633 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = df1$Life.expectancy ~ ., data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-16.9597  -2.0621  -0.0147   2.2751  11.7115

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.445e+01  8.400e-01  64.822  < 2e-16 ***
StatusDeveloping                -9.684e-01  3.379e-01  -2.865  0.00422 **
Adult.Mortality                 -1.663e-02  9.494e-04 -17.517  < 2e-16 ***
infant.deaths                    9.350e-02  1.065e-02   8.777  < 2e-16 ***
Alcohol                         -9.140e-02  3.316e-02  -2.756  0.00592 **
percentage.expenditure           3.673e-04  1.801e-04   2.040  0.04156 *
Hepatitis.B                     -6.525e-03  4.449e-03  -1.467  0.14265
Measles                         -7.865e-06  1.079e-05  -0.729  0.46597
BMI                              3.376e-02  5.998e-03   5.628 2.15e-08 ***
under.five.deaths               -7.035e-02  7.711e-03  -9.123  < 2e-16 ***
Polio                            7.935e-03  5.152e-03   1.540  0.12370
Total.expenditure                7.586e-02  4.067e-02   1.865  0.06236 .
Diphtheria                       1.490e-02  5.928e-03   2.513  0.01205 *
HIV.AIDS                        -4.370e-01  1.784e-02 -24.490  < 2e-16 ***
GDP                              8.738e-06  2.837e-05   0.308  0.75813
Population                      -6.425e-10  1.749e-09  -0.367  0.71337
thinness..1.19.years            -1.238e-02  5.300e-02  -0.234  0.81527
thinness.5.9.years              -4.798e-02  5.231e-02  -0.917  0.35917
Income.composition.of.resources  9.817e+00  8.321e-01  11.797  < 2e-16 ***
Schooling                        8.665e-01  5.940e-02  14.587  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.588 on 1629 degrees of freedom
Multiple R-squared:  0.8356,    Adjusted R-squared:  0.8336
F-statistic: 435.7 on 19 and 1629 DF,  p-value: < 2.2e-16
```

# Prediction

```r
x<-cbind(rep(1,nrow(dftest)),dftest[,c("Schooling", "Adult.Mortality" ,
    "HIV.AIDS" , "Income.composition.of.resources" , "percentage.expenditure" ,
    "Diphtheria", "BMI" , "Total.expenditure" , "infant.deaths" , "thinness.5.9.years" ,
    "Status" , "Hepatitis.B")])
x1<-data.matrix(x)
residualtest<-as.matrix(modelforward1$coefficients)
yhat<-x1%*%residualtest
sum((dftest$Life.expectancy-yhat)^2)
```

```
[1] 6019.014
```



# Improvement

- Rescale some of the predictors
- Further investigation on BP test

```
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ 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 ...
```

# Thank you !