What is the best model to predict life expectancy?

Analyzing the Life Expectancy dataset

```r
#Importing Libraries

library(dplyr)
library(tidyverse)
library(ggplot2)
library(psych)
library(leaps)


# Analyzing the 2 tables
# We merge the data into a single table

?merge
df <- merge(Life_Expectancy_per_Country, world_pop, by = "Country" )
view(df)

summary(df)

#There are N.A. values in the column Urban Population Percentage, so we drop it.
names(df)[names(df) == "Urban Pop%"] <- "urban"
life_exp <- subset(df, select = -c(urban))
View(life_exp)

#Exploring the final life_exp data set
dim(life_exp)
summary(life_exp)
class(life_exp)
lapply(life_exp, class)
```

At first look, we can see that the dataset are based on different countries.

Two data sets ‘Life Expectancy per country’ and ‘World population’ datasets are merged into one data frame by Country. We got all the common countries in the two tables.

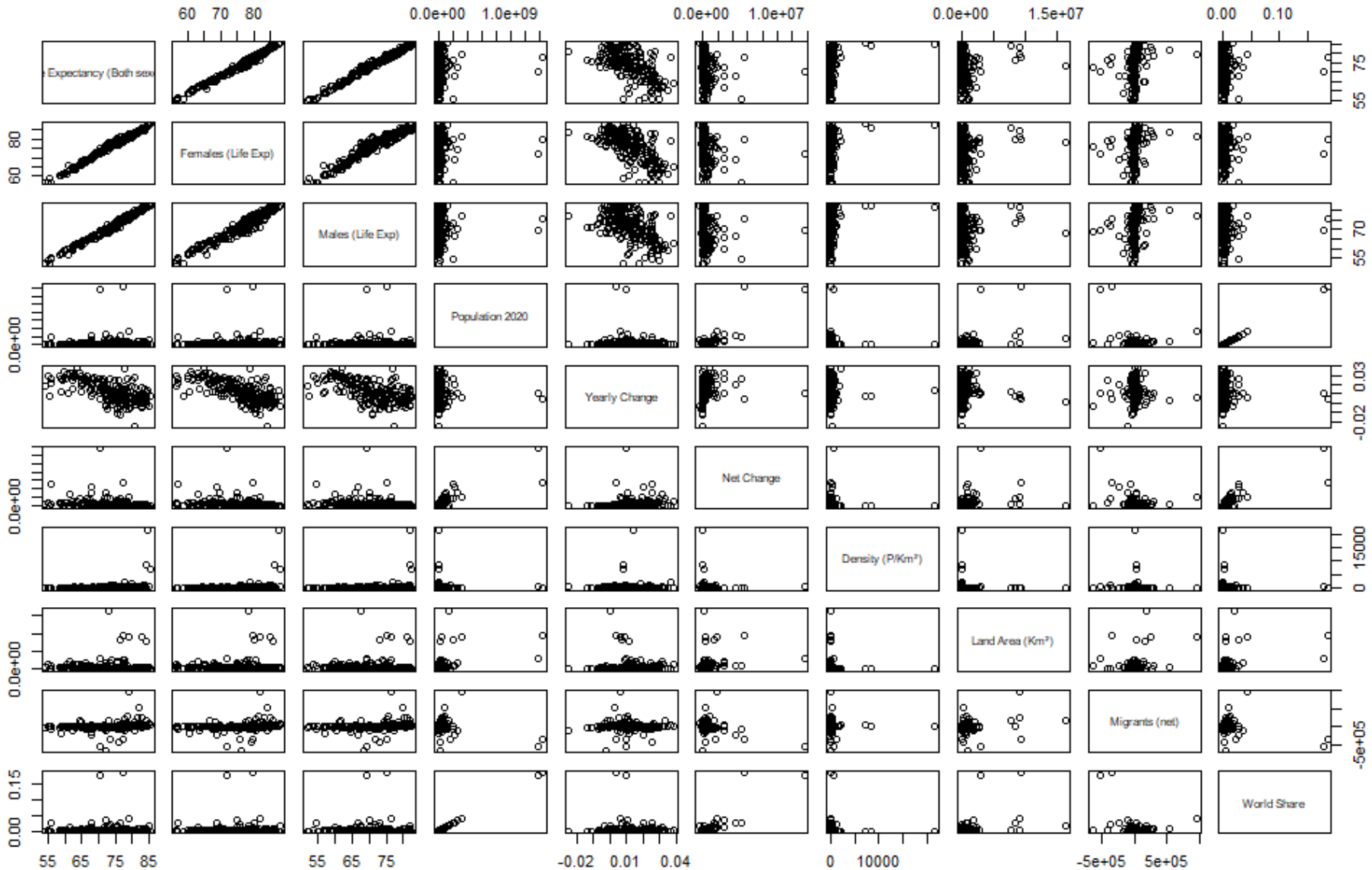
There are N.A. values in one column Urban Population percent, so we drop that particular column to make our data cleaner.

Now our final data frame consists of 202 rows and 13 variables.

	Country	Life Expectancy (Both sexes)	Females (Life Exp)	Males (Life Exp)	Population 2020	Yearly Change	Net Change	Density (P/Km²)	Land Area (Km²)	Migrants (net)	Fert. Rate	Med. Age	World Share
1	Afghanistan	65.98	67.59	64.47	38928346	0.0233	886592	60	652860	-62920	4.5999999999999996	18	0.0050
2	Albania	78.96	80.48	77.48	2877797	-0.0011	-3120	105	27400	-14000	1.6	36	0.0004

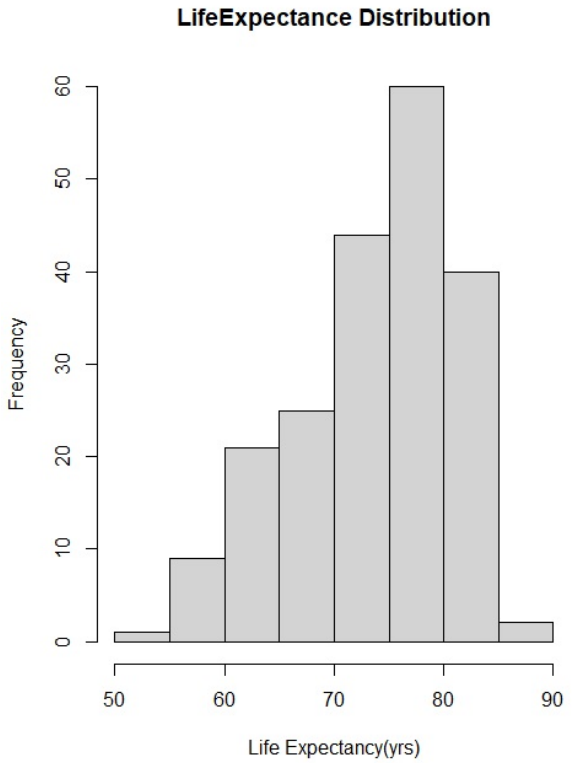
Comparing the pairs

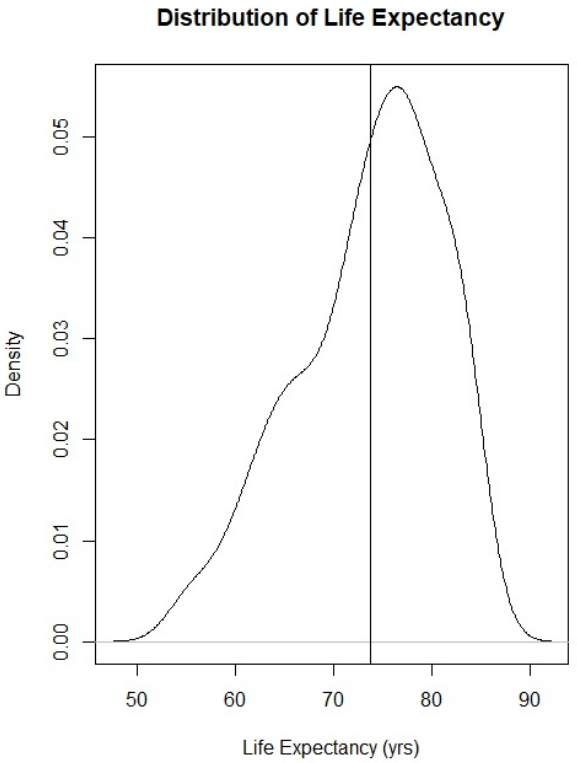
From this plot we can see the relationship among different variables.



Target Variable – Life Expectancy

```
# Target variable is Life Expectancy

# plotting a histogram
hist(life_exp$`Life Expectancy (Both sexes)`,
     main = "LifeExpectance Distribution",
     xlab = "Life Expectancy(yrs)")

# density plot
plot(density(life_exp$`Life Expectancy (Both sexes)`),
     main = "Distribution of Life Expectancy",
     xlab = "Life Expectancy (yrs)")
abline(v=mean(life_exp$`Life Expectancy (Both sexes)`))
```





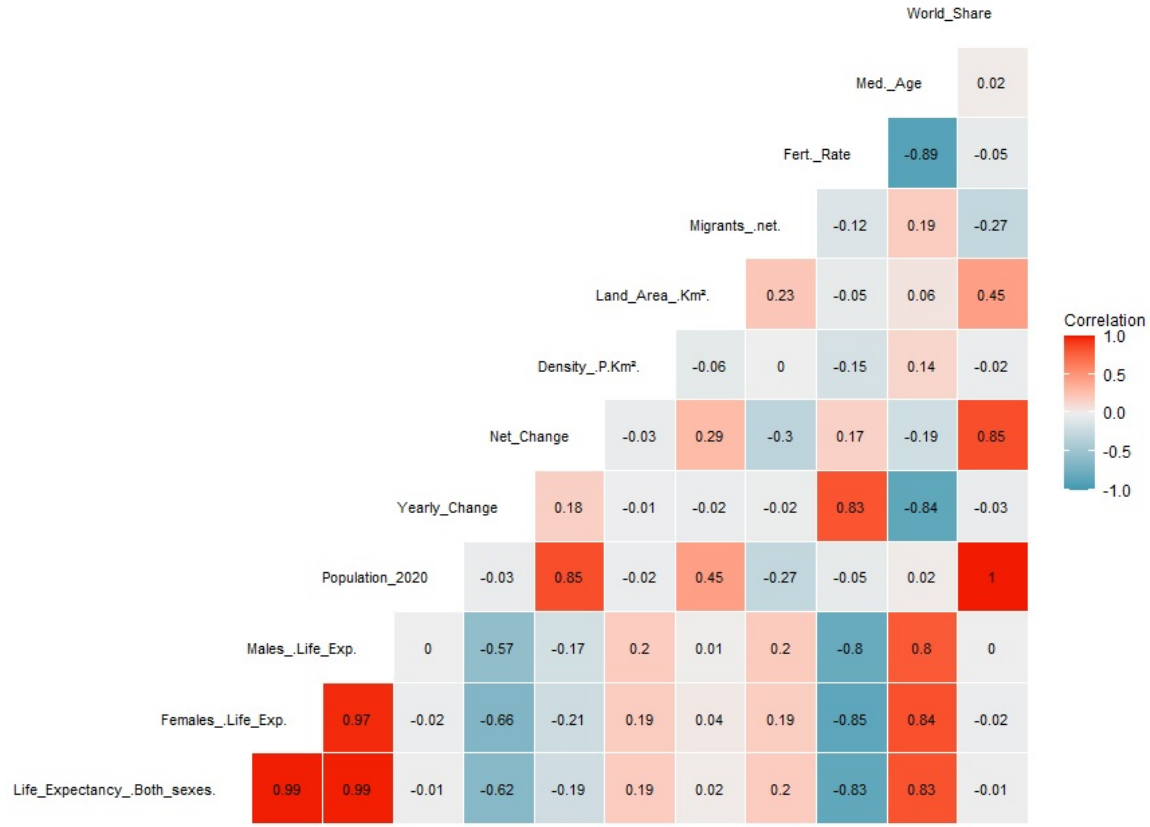
Since we need to tell the best model to predict Life Expectancy, the target variable for our evaluation would be Life Expectancy. We will compare it with all the other variables called Predictors.

Here we plotted a histogram for Life Expectancy and a Density curve.
We can see that the target variable is not perfectly normally distributed, instead it slightly skewed towards the left.

Correlation

```
#Correlation
cor(data_num, data_num$`Life Expectancy (Both sexes)`)

#selecting the variables that are numeric

data_num <- life_exp %>%
  select_if(is.numeric)

#Plotting the correlation
ggcorr(data_num,
       label = T,
       label_size = 3,
       label_round = 2,
       hjust = 1,
       size = 3,
       color = "black",
       layout.exp = 5,name = "Correlation")
```

```
                                [,1]
Life Expectancy (Both sexes)  1.00000000
Females (Life Exp)            0.99250671
Males (Life Exp)              0.99097355
Population 2020              -0.01124984
Yearly Change                -0.62201879
Net Change                   -0.19446438
Density (P/Km²)               0.19447763
Land Area (Km²)               0.02375794
Migrants (net)                0.19513195
Fert. Rate                   -0.83203989
Med. Age                      0.82903906
World Share                  -0.01110624
> |
```



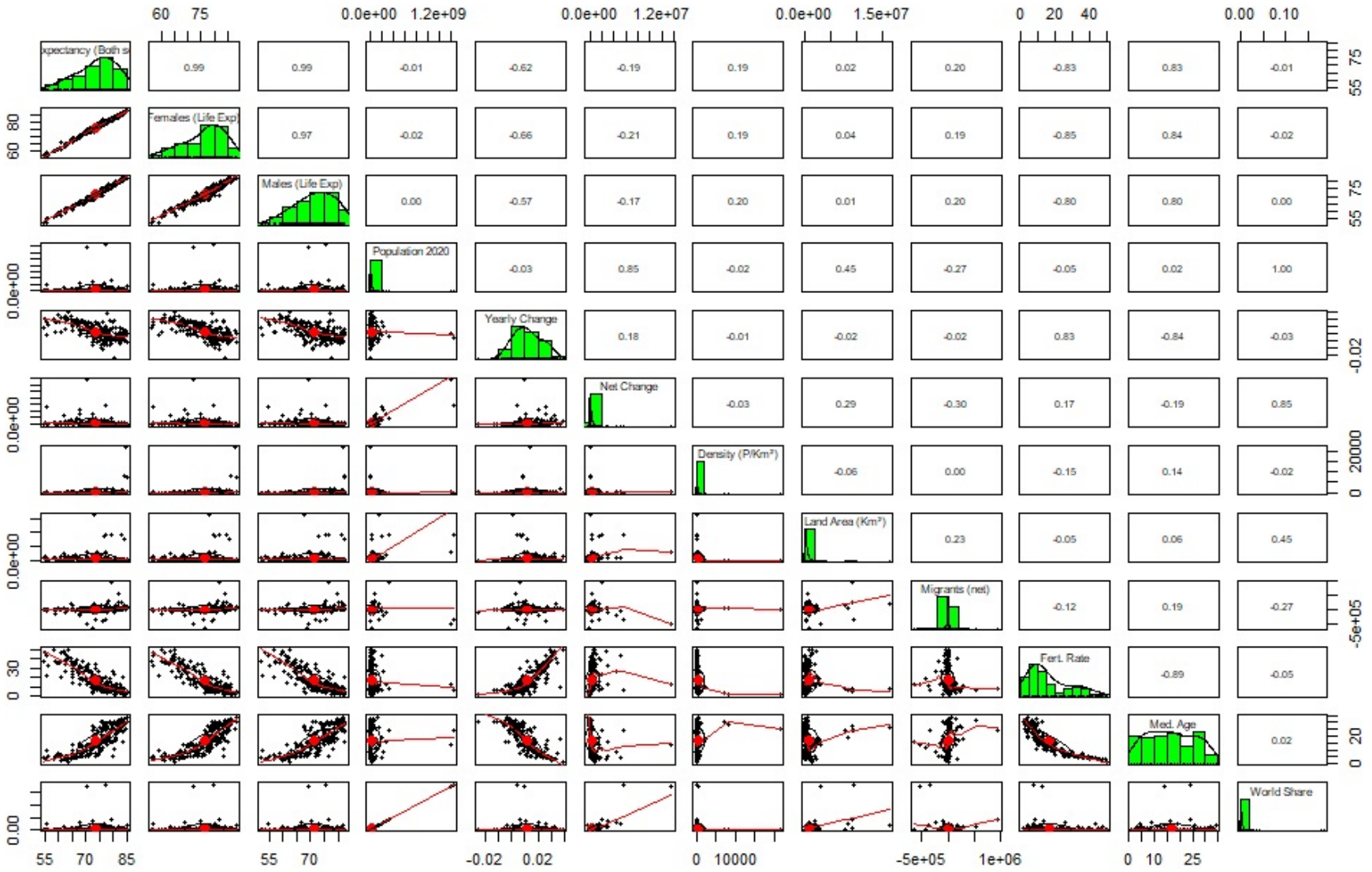
From plotting the correlation, we can see that the Life Expectancy variable is highly correlated with Life Expectancy of Males and Life Expectancy of Females with values of 0.990 and 0.992, respectively and Median Age with value of 0.829.

It is also negatively correlated to Yearly change with a value of 0.622 and Fertility Rate with value of 0.832.

Correlation

```
#Pearson Method
pairs.panels(data_num,
             method = "pearson",
             hist.col = "green",
             density = TRUE,
             ellipses = TRUE)
```

Here, the correlation is plotted using the Pearson method for better analysis. We can see that Life expectancy of Males; Life expectancy of Females and Fertility Rate, Median Age are highly correlated with Life Expectancy.



Variable Selection

Variable selection is the process of selecting the variables that should be present in the final model. This can be done in different ways – Forward selection, Backward elimination and Step-wise selection.

Forward Selection

```
#Variable Selection

#Forward Selection
nullmodel_1 <- lm(`Life Expectancy (Both sexes)`~ 1, data = data_num)
fullmodel_1 <- lm(`Life Expectancy (Both sexes)`~ ., data = data_num)

forward_1 <- step(nullmodel_1, scope=list(lower=nullmodel_1, upper=fullmodel_1)
                  direction='forward')

summary(forward_1)
```

```
Step:  AIC=-1008.3
`Life Expectancy (Both sexes)` ~ `Females (Life Exp)` + `Males (Life Exp)` +
    `Med. Age` + `Fert. Rate` + `Yearly Change`

                    Df  Sum of Sq    RSS     AIC
<none>                            1.2934 -1008.3
+ `World Share`      1 0.00196503 1.2915 -1006.6
+ `Population 2020`  1 0.00195444 1.2915 -1006.6
+ `Migrants (net)`   1 0.00167187 1.2918 -1006.6
+ `Density (P/Km²)`  1 0.00095390 1.2925 -1006.4
+ `Net Change`       1 0.00016740 1.2933 -1006.3
+ `Land Area (Km²)`  1 0.00011555 1.2933 -1006.3
```

The forward selection method suggests that we drop all the variables except Life Expectancy of Males and Females and Yearly Change, Median Age and Fertility Rate.

The adjusted R-square value here is 0.999

```
> summary(forward_1)

Call:
lm(formula = `Life Expectancy (Both sexes)` ~ `Females (Life Exp)` +
    `Males (Life Exp)` + `Med. Age` + `Fert. Rate` + `Yearly Change`,
    data = data_num)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32848 -0.03497  0.00355  0.04548  0.19768

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.225491   0.125716  -1.794  0.07441 .
`Females (Life Exp)`  0.510133   0.003613 141.180  < 2e-16 ***
`Males (Life Exp)`    0.490194   0.003475 141.066  < 2e-16 ***
`Med. Age`            0.005008   0.001746   2.868  0.00458 **
`Fert. Rate`          0.005389   0.001234   4.368 2.03e-05 ***
`Yearly Change`      -2.499022   1.190900  -2.098  0.03715 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08124 on 196 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 3.246e+05 on 5 and 196 DF,  p-value: < 2.2e-16
```

Backward Elimination

```
#Backward Elimination
backward_1 <- step(fullmodel_1,direction='backward')

summary(backward_1)
```

```
Step:  AIC=-1008.3
`Life Expectancy (Both sexes)` ~ `Females (Life Exp)` + `Males (Life Exp)` +
    `Yearly Change` + `Fert. Rate` + `Med. Age`

                     Df Sum of Sq     RSS      AIC
<none>                             1.293 -1008.30
- `Yearly Change`     1     0.029   1.322 -1005.81
- `Med. Age`          1     0.054   1.348 -1001.99
- `Fert. Rate`        1     0.126   1.419  -991.53
- `Males (Life Exp)`  1   131.322 132.615   -75.01
- `Females (Life Exp)` 1   131.533 132.826   -74.68
```

```
> summary(backward_1)

Call:
lm(formula = `Life Expectancy (Both sexes)` ~ `Females (Life Exp)` +
    `Males (Life Exp)` + `Yearly Change` + `Fert. Rate` + `Med. Age`,
    data = data_num)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32848 -0.03497  0.00355  0.04548  0.19768

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.225491   0.125716  -1.794  0.07441 .
`Females (Life Exp)`  0.510133   0.003613 141.180  < 2e-16 ***
`Males (Life Exp)`    0.490194   0.003475 141.066  < 2e-16 ***
`Yearly Change`      -2.499022   1.190900  -2.098  0.03715 *
`Fert. Rate`          0.005389   0.001234   4.368 2.03e-05 ***
`Med. Age`            0.005008   0.001746   2.868  0.00458 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08124 on 196 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 3.246e+05 on 5 and 196 DF,  p-value: < 2.2e-16
```

The backward elimination method also suggests that we drop all the variables except Life Expectancy of Males and Females and Yearly Change, Median Age and Fertility Rate.

The adjusted R-square value here is 0.999

Step-wise Selection

```
#Step - Wise Selection
stepwise_1 <- step(nullmodel_1, scope=list(lower=nullmodel_1, upper=fullmodel_1
                  direction='both')

summary(stepwise_1)
```

```
Step:  AIC=-1008.3
`Life Expectancy (Both sexes)` ~ `Females (Life Exp)` + `Males (Life Exp)` +
    `Med. Age` + `Fert. Rate` + `Yearly Change`

                     Df Sum of Sq     RSS      AIC
<none>                             1.293 -1008.30
+ `World Share`       1     0.002   1.291 -1006.60
+ `Population 2020`   1     0.002   1.291 -1006.60
+ `Migrants (net)`    1     0.002   1.292 -1006.56
+ `Density (P/Km²)`   1     0.001   1.292 -1006.44
+ `Net Change`        1     0.000   1.293 -1006.32
+ `Land Area (Km²)`   1     0.000   1.293 -1006.31
- `Yearly Change`     1     0.029   1.322 -1005.81
- `Med. Age`          1     0.054   1.348 -1001.99
- `Fert. Rate`        1     0.126   1.419  -991.53
- `Males (Life Exp)`  1   131.322 132.615   -75.01
- `Females (Life Exp)` 1  131.533 132.826   -74.68
```

The step-wise selection method for variable selection also suggests that we drop all the variables except Life Expectancy of Males and Females and Yearly Change, Median Age and Fertility Rate.

The adjusted R-square value here is 0.999

```
> summary(stepwise_1)

Call:
lm(formula = `Life Expectancy (Both sexes)` ~ `Females (Life Exp)` +
    `Males (Life Exp)` + `Med. Age` + `Fert. Rate` + `Yearly Change`
    data = data_num)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32848 -0.03497  0.00355  0.04548  0.19768

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.225491   0.125716  -1.794  0.07441 .
`Females (Life Exp)`  0.510133   0.003613 141.180  < 2e-16 ***
`Males (Life Exp)`    0.490194   0.003475 141.066  < 2e-16 ***
`Med. Age`            0.005008   0.001746   2.868  0.00458 **
`Fert. Rate`          0.005389   0.001234   4.368 2.03e-05 ***
`Yearly Change`      -2.499022   1.190900  -2.098  0.03715 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08124 on 196 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 3.246e+05 on 5 and 196 DF,  p-value: < 2.2e-16
```

Model – A

```r
## Model A - Null Model
model_a <- lm(data_num$`Life Expectancy (Both sexes)` ~ 1, data = data_num)

predict_a <- predict(model_a,  newdata = data_num)

summary(model_a)
plot(model_a)
step(model_a)
```

First, we do a linear regression model with just Life Expectancy as the dependent variable.

For testing the accuracy of the model, we also calculate the AIC value using step function and Adjusted R-squared values.

The values of AIC should be low and value of R-squared should be closer to 1.

Since this is a Null model, we can see that this model is not appropriate.

```
> step(model_a)
Start:  AIC=804.09
data_num$`Life Expectancy (Both sexes)` ~ 1


Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ 1, data = data_num)

Coefficients:
(Intercept)
       73.7
```

```
> summary(model_a)

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ 1, data = data_num)

Residuals:
    Min      1Q  Median      3Q     Max
-19.342  -5.012   1.328   5.460  11.588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.7022     0.5136   143.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.3 on 201 degrees of freedom
```

Model - B

```r
## Model B - Full Model
model_b <- lm(data_num$`Life Expectancy (Both sexes)` ~ .,data = data_num)
predict_b <- predict(model_b,  newdata = data_num)
summary(model_b)
plot(model_b)
step(model_b)
```

```
> step(model_b)
Start:  AIC=-998.34
data_num$`Life Expectancy (Both sexes)` ~ `Females (Life Exp)` +
    `Males (Life Exp)` + `Population 2020` + `Yearly Change` +
    `Net Change` + `Density (P/Km²)` + `Land Area (Km²)` + `Migrants (net)` +
    `Fert. Rate` + `Med. Age` + `World Share`

                      Df Sum of Sq     RSS      AIC
- `Migrants (net)`     1     0.001   1.281 -1000.20
- `Land Area (Km²)`    1     0.001   1.281 -1000.18
- `Density (P/Km²)`    1     0.002   1.282 -1000.09
- `Net Change`         1     0.002   1.283 -1000.00
- `Population 2020`    1     0.005   1.285  -999.60
- `World Share`        1     0.005   1.285  -999.59
<none>                             1.280  -998.34
- `Yearly Change`      1     0.031   1.311  -995.52
- `Med. Age`           1     0.045   1.325  -993.43
- `Fert. Rate`         1     0.115   1.395  -982.98
- `Females (Life Exp)` 1   122.805 124.085   -76.43
- `Males (Life Exp)`   1   123.554 124.835   -75.22
```

```
> summary(model_b)

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ ., data = data_num)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32363 -0.03535  0.00090  0.04175  0.19859

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.110e-01  1.293e-01  -1.632   0.1044
`Females (Life Exp)` 5.097e-01  3.775e-03 134.993  < 2e-16 ***
`Males (Life Exp)`   4.906e-01  3.623e-03 135.405  < 2e-16 ***
`Population 2020`    2.112e-08  2.521e-08   0.838   0.4031
`Yearly Change`     -2.743e+00  1.281e+00  -2.142   0.0334 *
`Net Change`         5.893e-09  1.041e-08   0.566   0.5721
`Density (P/Km²)`    1.725e-06  3.581e-06   0.482   0.6305
`Land Area (Km²)`    1.583e-09  4.072e-09   0.389   0.6980
`Migrants (net)`     2.065e-08  5.708e-08   0.362   0.7179
`Fert. Rate`         5.230e-03  1.267e-03   4.129 5.44e-05 ***
`Med. Age`           4.713e-03  1.833e-03   2.571   0.0109 *
`World Share`       -1.652e+02  1.963e+02  -0.841   0.4013
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08209 on 190 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 1.445e+05 on 11 and 190 DF,  p-value: < 2.2e-16
```

Here, we do a linear regression model with Life Expectancy as the dependent variable and all the remaining variables as independent.

From the summary of this model, we can see that the p-values of only 5 variables i.e., Life Expectancy of Males and Females and Yearly Change, Fertility Rate and Median Rate is significant.

For testing the accuracy of the model, we also calculate the R-squared values and AIC which comes out to 0.999 and -998.34 respectively.

Model - C

```
## Model C - Using Selected Variables from Forward Selection
model_c <- lm(data_num$`Life Expectancy (Both sexes)` ~ data_num$`Males (Life Exp)
              + data_num$`Females (Life Exp)` + data_num$`Yearly Change`
              + data_num$`Fert. Rate`+ data_num$`Med. Age`, data = data_num)
predict_c <- predict(model_c,  newdata = data_num)
summary(model_c)
plot(model_c)
step(model_c)
```

```
> step(model_c)
Start:  AIC=-1008.3
data_num$`Life Expectancy (Both sexes)` ~ data_num$`Males (Life Exp)` +
    data_num$`Females (Life Exp)` + data_num$`Yearly Change` +
    data_num$`Fert. Rate` + data_num$`Med. Age`

                              Df Sum of Sq     RSS      AIC
<none>                                      1.293 -1008.30
- data_num$`Yearly Change`     1     0.029   1.322 -1005.81
- data_num$`Med. Age`          1     0.054   1.348 -1001.99
- data_num$`Fert. Rate`        1     0.126   1.419  -991.53
- data_num$`Males (Life Exp)`  1   131.322 132.615   -75.01
- data_num$`Females (Life Exp)` 1   131.533 132.826   -74.68
```

```
> summary(model_c)

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Males (Life Exp)` +
    data_num$`Females (Life Exp)` + data_num$`Yearly Change` +
    data_num$`Fert. Rate` + data_num$`Med. Age`, data = data_num)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32848 -0.03497  0.00355  0.04548  0.19768

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.225491   0.125716  -1.794  0.07441 .
data_num$`Males (Life Exp)`   0.490194   0.003475 141.066  < 2e-16 ***
data_num$`Females (Life Exp)` 0.510133   0.003613 141.180  < 2e-16 ***
data_num$`Yearly Change`     -2.499022   1.190900  -2.098  0.03715 *
data_num$`Fert. Rate`         0.005389   0.001234   4.368 2.03e-05 ***
data_num$`Med. Age`           0.005008   0.001746   2.868  0.00458 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08124 on 196 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 3.246e+05 on 5 and 196 DF,  p-value: < 2.2e-16
```

Here, we do a linear regression model with Life Expectancy as the dependent variable, and we use predictors we got from the variable selection method. Since the 3 methods gave the same result, so we can form one model for this.

For testing the accuracy of the model, we also calculate the Adjusted R-squared values and AIC values from step method which comes out to be 0.999 and -1008.3

We can see that this model is better than the previous model because of the change in significance of Median Age.

Model D

```
## Model D
model_d <- lm(data_num$`Life Expectancy (Both sexes)` ~ data_num$`Yearly Change`)
predict_d <- predict(model_d,  newdata = data_num)
summary(model_d)
plot(model_d)
step(model_d)
```

```
> step(model_d)
Start:  AIC=707.26
data_num$`Life Expectancy (Both sexes)` ~ data_num$`Yearly Change`

                           Df Sum of Sq     RSS    AIC
<none>                                  6566.7 707.26
- data_num$`Yearly Change`  1    4144.1 10710.8 804.09

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Yearly Change`)

Coefficients:
            (Intercept)  data_num$`Yearly Change`
                   78.7                    -417.0
```

Here, we do a linear regression model with Life Expectancy as the dependent variable, and we use predictors as Yearly Change.

For testing the accuracy of the model, we calculate the Adjusted R-squared values and AIC values from step method which comes out to be 0.383 and 707.26, respectively.

```
> summary(model_d)

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Yearly Change`)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7172  -3.5113   0.0939   4.0820  14.3717

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 78.7030     0.6006  131.05   <2e-16 ***
data_num$`Yearly Change` -416.9741    37.1154  -11.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.73 on 200 degrees of freedom
Multiple R-squared:  0.3869,    Adjusted R-squared:  0.3838
F-statistic: 126.2 on 1 and 200 DF,  p-value: < 2.2e-16
```

Model E

```r
## Model E
model_e <- lm(data_num$`Life Expectancy (Both sexes)` ~ data_num$`Fert. Rate`)
predict_e <- predict(model_e,  newdata = data_num)
summary(model_e)
plot(model_e)
step(model_e)
```

Here, we do a linear regression model with Life Expectancy as the dependent variable, and we use predictors as Fertility Rate.

For testing the accuracy of the model, we also calculate the Adjusted R-squared values and AIC values from step model which comes out to be 0.690 and 568.01, respectively.

```
> step(model_e)
Start:  AIC=568.01
data_num$`Life Expectancy (Both sexes)` ~ data_num$`Fert. Rate`

                        Df Sum of Sq     RSS    AIC
<none>                               3295.8 568.01
- data_num$`Fert. Rate`  1      7415 10710.8 804.09

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Fert. Rate`)

Coefficients:
          (Intercept)  data_num$`Fert. Rate`
              82.0027                -0.4936
```

```
> summary(model_e)

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Fert. Rate`)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4939  -2.7607   0.1382   2.6769  11.3590

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            82.00266    0.48446  169.27   <2e-16 ***
data_num$`Fert. Rate` -0.49358    0.02327  -21.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.059 on 200 degrees of freedom
Multiple R-squared:  0.6923,    Adjusted R-squared:  0.6908
F-statistic:   450 on 1 and 200 DF,  p-value: < 2.2e-16
```

Model F

```
## Model F
model_f <- lm(data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age`)
predict_f <- predict(model_f,  newdata = data_num)
summary(model_f)
plot(model_f)
step(model_f)
```

Here, we do a linear regression model with Life Expectancy as the dependent variable, and we use predictors as Median Age.

For testing the accuracy of the model, we also calculate the Adjusted R-squared values and AIC values from step function which comes out to be 0.685 and 571.26, respectively.

```
> step(model_f)
Start:  AIC=571.26
data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age`

                      Df Sum of Sq     RSS    AIC
<none>                              3349.2 571.26
- data_num$`Med. Age`  1    7361.6 10710.8 804.09

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age`)

Coefficients:
        (Intercept)  data_num$`Med. Age`
             62.660                0.665


> summary(model_f)

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age`)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6605  -2.2918   0.1244   2.7106  13.1995

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          62.66043    0.60020  104.40   <2e-16 ***
data_num$`Med. Age`   0.66501    0.03172   20.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.092 on 200 degrees of freedom
Multiple R-squared:  0.6873,    Adjusted R-squared:  0.6857
F-statistic: 439.6 on 1 and 200 DF,  p-value: < 2.2e-16
```

Model G

```
## Model G
model_g <- lm(data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age` +
                data_num$`Fert. Rate`)
predict_g <- predict(model_g,  newdata = data_num)
summary(model_g)
plot(model_g)
step(model_g)
```

Here, we do a linear regression model with Life Expectancy as the dependent variable, and we use predictors as Median Age and Fertility Rate.

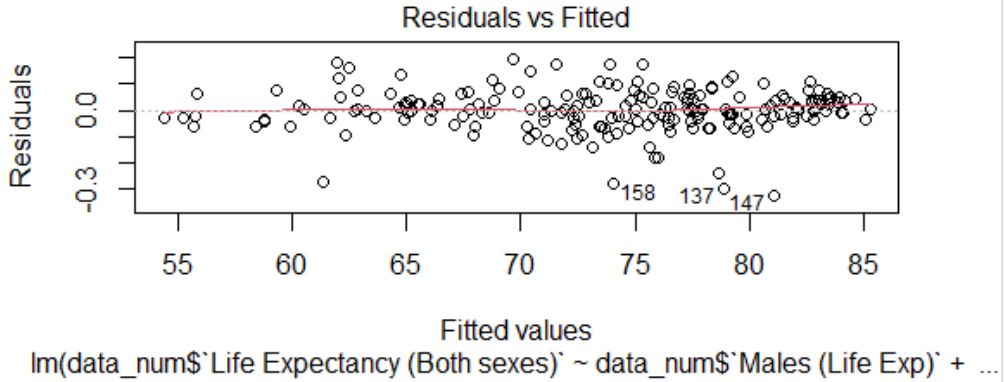
For testing the accuracy of the model, we also calculate AIC values and the Adjusted R-squared values from step method which comes out to be 543.48 and 0.727, respectively.

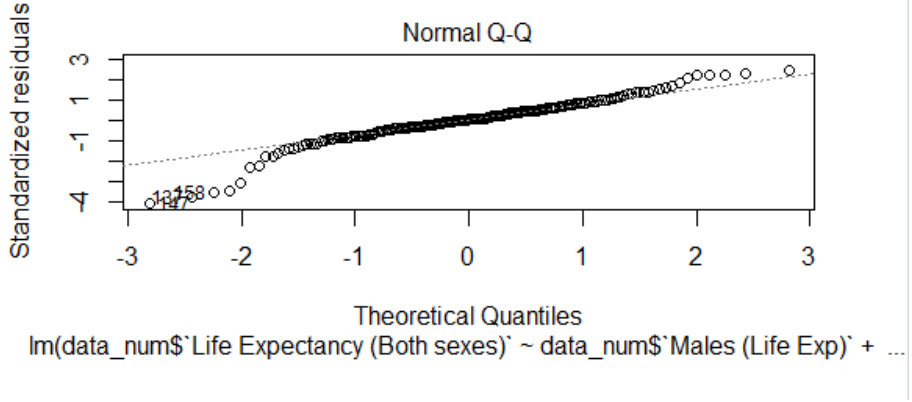
```
> step(model_g)
Start:  AIC=543.48
data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age` +
    data_num$`Fert. Rate`

                        Df Sum of Sq    RSS    AIC
<none>                               2890.2 543.48
- data_num$`Med. Age`    1    405.61 3295.8 568.01
- data_num$`Fert. Rate`  1    459.00 3349.2 571.26

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age` +
    data_num$`Fert. Rate`)

Coefficients:
          (Intercept)    data_num$`Med. Age`  data_num$`Fert. Rate`
              72.5494                 0.3417                -0.2688
```

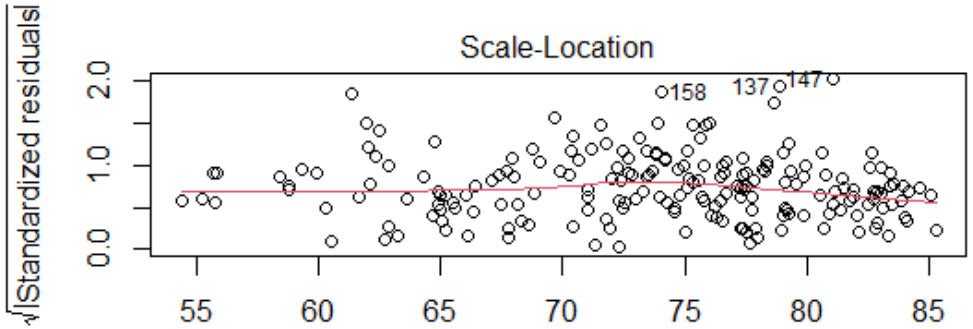
```
> summary(model_g)

Call:
lm(formula = data_num$`Life Expectancy (Both sexes)` ~ data_num$`Med. Age` +
    data_num$`Fert. Rate`)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4025  -2.2523   0.4327   2.4208  12.5083

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          72.54935    1.84573  39.307  < 2e-16 ***
data_num$`Med. Age`   0.34168    0.06466   5.285 3.29e-07 ***
data_num$`Fert. Rate` -0.26880    0.04782  -5.622 6.33e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.811 on 199 degrees of freedom
Multiple R-squared:  0.7302,    Adjusted R-squared:  0.7274
F-statistic: 269.2 on 2 and 199 DF,  p-value: < 2.2e-16
```

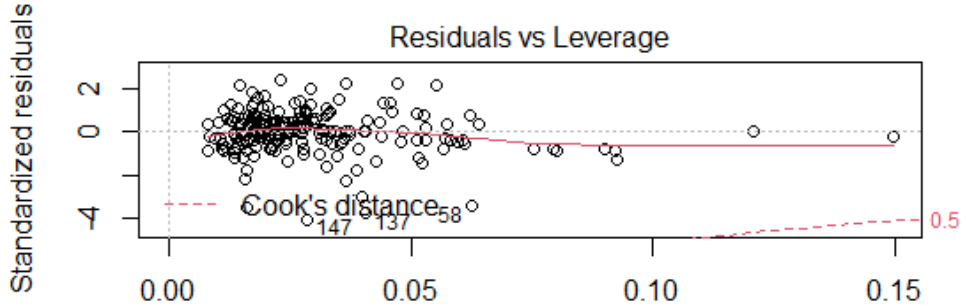
Residual Analysis for the best model - Model – C – Using Variable Selection

```
> plot(model_b)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```









Conclusion

```
> summary(model_a)$adj.r.squared
[1] 0
> summary(model_b)$adj.r.squared
[1] 0.9998735
> summary(model_c)$adj.r.squared
[1] 0.9998762
> summary(model_d)$adj.r.squared
[1] 0.3838419
> summary(model_e)$adj.r.squared
[1] 0.6907518
> summary(model_f)$adj.r.squared
[1] 0.6857423
> summary(model_g)$adj.r.squared
[1] 0.7274478
>
```

AIC Model a	804.09
AIC Model b	-998.34
AIC Model c	-1008.3
AIC Model d	707.26
AIC Model e	568.01
AIC Model f	571.26
AIC Model g	543.48

While we tried various types of regression models to predict the Life Expectancy from the dataset, we found out that the Model C i.e., the model where predictors are the ones selected through Variable Selection Method which are Life Expectancy for Males, Life Expectancy for Females, Yearly Change, Median Age and Fertility Rate with Life Expectancy as the target variable. This model yielded the best outcome with least AIC value of -1008.3 and greatest adjusted R-squared value of 0.999.

Thus, we can conclude that the best model to predict the Life Expectancy from the given datasets is the model done using variable selection method.

Which other variables could you include in the model to improve predictions of life expectancy?

As we saw in the earlier answer, when we performed variable selection, the most significant variables were chosen which were highly correlated to the target variable ‘Life Expectancy’. These variables were Life Expectancy for Males, Life Expectancy for Females and Yearly Change, Fertility rate and Median Value.

These are the variables which should be included in the model.

Apart from this dataset, some of the variables which should be included in the model to improve predictions of Life Expectancy could be Income, Adult Mortality and Schooling. These variables could have a positive effect on Life Expectancy. While some other variables which could be used are Alcohol consumption, Infant deaths, certain diseases. These variables will have a negative effect on Life expectancy.

Adding these variables could improve the models for predictions of life expectancy.