Regression Analysis Project

Names and codes:

Group 12 – English section

Haneen Ahmed Elgabry (5200199)

Nour Mohamed Eletreby (5200917)

Nada Hussein Mohamed (5200699)

Course: Regression Analysis – Statistics Major

Submitted to: Dr Amira Alayouty

Introduction

In this model we have one response variable which is (Healthy life expectancy at birth years) and seven explanatory variables which are (HDI rank, Human Development Index (HDI), Expected years of schooling, Gross Domestic Product per capita, mean years of schooling, GNI per capita rank minus HDI rank and HDIrankfor2020) ,we ignored the country variable as it is a qualitative variable .and we aim to find the best linear relationship between this response variable and those seven explanatory variables using multiple regression analysis on a random sample of 50 countries. In this project we will drop some insignificant explanatory variables if it is found that there is multicollinearity between some of those explanatory variables. We will use one of the automatic procedures for example (Forward regression, Backward regression or Stepwise regression) methods. We will find the best multiple regression model using one of those methods, then we will assess the goodness of this fitted model and finally check the validity of the error part assumptions.

Descriptive statistics

First, we are going to do some descriptive statistics for the response variable along with other independent variables. The mean of life expectancy at birth years is 71.22 which means that the average of years is 71.22, the variance is 67.568 , the median which is the value in the middle of the order ascending or descending is 73, the range of our observations is (82.9-52.7=30.2), the first quartile is 65.40 which means that 25% of the data have life expectancy at birth of 65.40 years or lower. The third quartile is 77.33 which means that 75% of the data have healthy life expectancy at birth of 77.33 years or higher.

Descriptive statistics for the explanatory variables:

1- The mean of HDI rank 2021 is 97.08 and the median is 84.5, the first quartile is 43, variance 3588.32 and third quartile is 154.25
2- The mean of Gross national income (GNI) per capita is 21973.06 and the median is 12654, variance is 530286255, the first quartile is 4131 and third quartile is 36737.
3- The mean of Human Development Index (HDI) is 0.7117 and the median is 0.7575, variance is0.02784, the first quartile is 0.5647 and third quartile is 0.8528.
4- The mean of Expected years of schooling 13.31 and the median is 13.45., variance is 9.712, the first quartile is 10.75 and third quartile is 15.78.
5- The mean of mean years of schooling is 8.578 and the median is 9.4, variance is 11.51, the first quartile is 5.75 and third quartile is 11.225.
6- The mean of GNI per capita rank minus HDI rank is -0.9 and the median is 0.5, variance is 225.398, the first quartile is -8 and third quartile is 7.75.
7- The mean of HDI rank for 2020 is 97.5 and the median is 84.5, variance is 3557.888, the first quartile is 44.25 and third quartile is 153.25.

Statistical Analysis

This is the variable under study, which we want to study the relationship between it and the rest of the other independent variables affecting it.

- Independent variables: which affects the dependent variable

HDI rank, Human Development Index (HDI), Expected years of schooling, Gross Domestic Product per capita, mean years of schooling, GNI per capita rank minus HDI rank and HDIrankfor2020.

- Common Applications: Regression is used to look for significant relationships between two variables or predict a value of one variable for given values of the others

Data: The data set contains details of 50 countries and the aim of this study is Modeling the relationship between life expectancy and other variables in the data set using linear regression for this random sample of 50 countries

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. It is parametric in nature because it makes certain assumptions based on the data set. If the data set follows those assumptions, regression gives incredible results. And these assumptions are:

- existence of a linear and additive relationship between dependent variable and other independent variables. We mean by additive that the effect of X on Y is independent of other variables.

- No correlation among independent variables. Presence of correlation in independent variables lead to Multicollinearity. If variables are correlated, it becomes extremely difficult for the model to determine the true effect of independent variables on dependent variable and this This causes problems with the analysis and interpretation

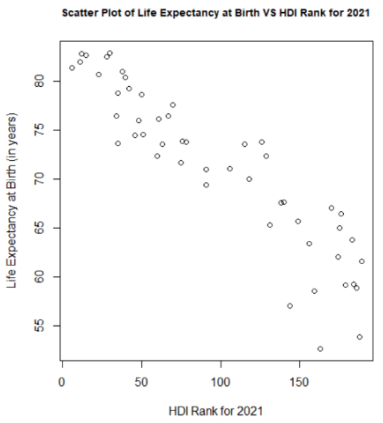
- The error terms must possess constant variance (Homoscedasticity). Absence of constant variance leads to Heteroscedasticity. And the expected value of the random error is zero.

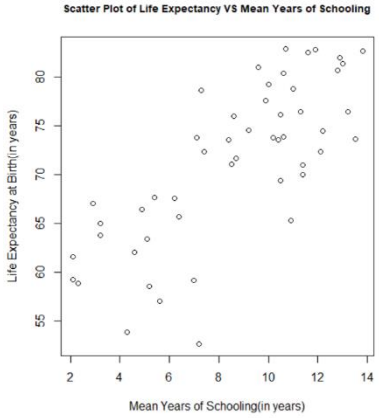
- The error terms must be uncorrelated, and the dependent variable and the error terms must possess a normal distribution. Presence of correlation in error terms is known as Autocorrelation. It affects the regression coefficients and standard error values since they assume of uncorrelated error terms

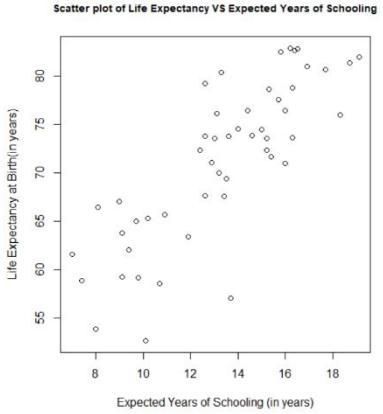
- The assumptions regarding the deterministic part are all valid except the assumptions of multicollinearity

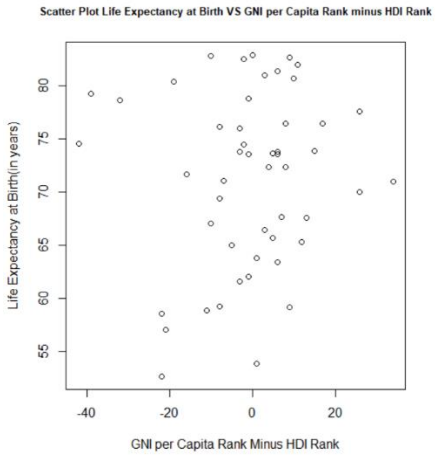
Firstly, we will start by checking the deterministic part of the model before we fit the model, we will check the relationship between the dependent variable and all the explanatory variables, then the relationship between the explanatory variables with each other's to investigate the multicollinearity we will display the coefficient of correlation of the variables So that if the correlation coefficient (r) approaches 1 , then there is a multicollinearity and we must drop one

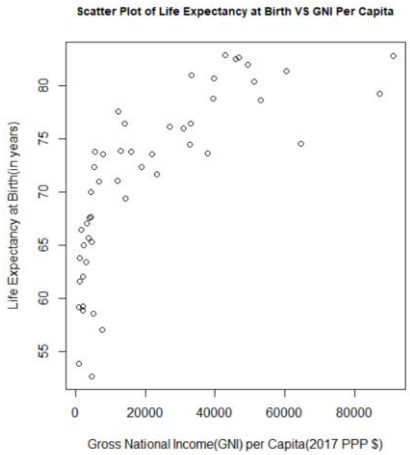
of them (we will drop the variable which has a less correlation coefficient with response variable).

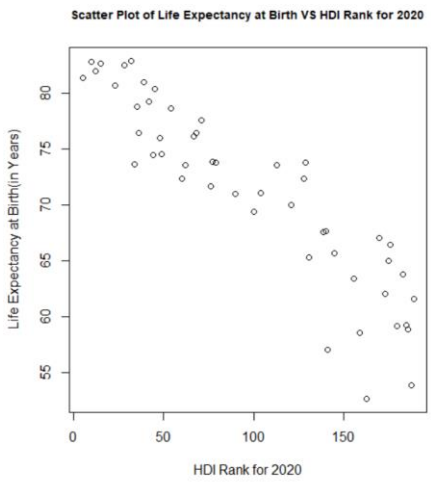


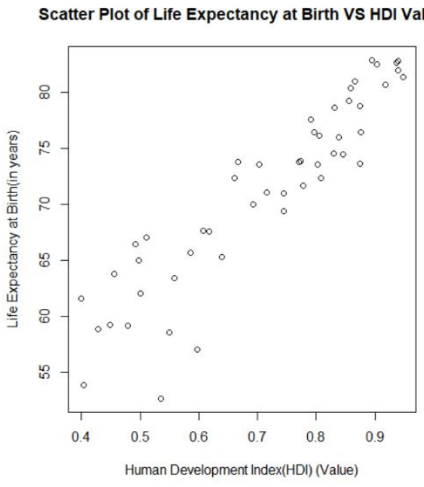












By plotting the response variable with each explanatory variable, we found out that there are no non-linear patterns between any pair of variables, all of them are linear but varies in strength.

And because there is no intuitive natural order that we know about the explanatory variable. (Explanatory variables don't involve a time component.) Thus, no need for a plot of the residuals versus observations order.

To check multicollineraity between explanatory variables we calculated coefficient of correlation:

r(life expectancy , hdi rank)= -0.9142179
r(life expectancy , HDI)= 0.9191413
r(life expectancy ,expected years of schooling)= 0.8143237
r(life expectancy ,mean years of schooling)= 0.7622374
r(life expectancy , GNI PER CAPITA)= 0.7544195
r(life expectancy , GNI minus HDI)= 0.1059691
r(life expectancy , hdi rank 2020)= -0.9116978
r (hdi randk , hdi index)= -0.9929606
r (hdi rank , expected years school) = -0.9143231
r (hdi rank , mean years school)=-0.8892131
r(hdi rank,GNI per capita) =-0.8265752
r(hdi rank , GNI minus HDI rank)= -0.006635302
r (hdi rank , hdi rank 2020)= 0.9993346
r (HumanDevelopmentIndex.HDI,expected years of scholl)= 0.9277247
r (HumanDevelopmentIndex.HDI,mean years of school)= 0.9058095
r (HumanDevelopmentIndex.HDI,gross national income GNI per capita)= 0.7944003
r (HumanDevelopmentIndex.HDI,GNI minus HDI)= 0.03706069
r (HumanDevelopmentIndex.HDI,HDI rank 2020)= -0.9923674
r(Expectedyearsofschooling,mean years of school)= 0.8294713
r(Expectedyearsofschooling,GNI PER CAPITA)= 0.6399038
r(Expectedyearsofschooling, GNI minus HDI)= 0.1905102
r(Expectedyearsofschooling,HDI rank 2020)= -0.9163945
r (Meanyearsofschooling,GNI PER CAPITA)= 0.6136433

r (Meanyearsofschooling,GNI minus HDI)= 0.2636392
r (Meanyearsofschooling,HDI rank 2020)= -0.8881409
r (GNI PER CAPIA,GNI minus HDI)= -0.3825981
r (GNI PER CAPIA,hdi rank 2020)= -0.8292478
r(GNI minus HDI , HDI rank 2020) = -0.007554685

-we are going to remove HDI rank 2021, HDI rank 2020 and mean years of schooling because of the multicollinearity.

Now and after removing multicollinearity between variables we will start fitting the model and check the assumptions after running it we get:

$E(\text{lifeexpectancyatbirth}) = \beta 0 + \beta 1\ \text{HumanDevelopmentIndex.HDI.} + \beta 2(\text{Grossnationalincome.GNI.percapita}) + \beta 3(\text{Expectedyearsofschooling}) + \beta 4(\text{GNIpercapitarankminusHDIrank})$

-after we did backward elimination, forward selection and stepwise we found out that these 3 procedures they lead us to 3 different models which means that there may still be explanatory variables that have multicollinearity.

After re-checking the multicollinearity, we found out that we must remove human development index.

So, we are going to do backward elimination, forward selection and stepwise on the model after removing HDI rank 2021, HDI rank 2020 and mean years of schooling and HDI index the 3 procedures lead us to the 3 same models and the model(check the appendix for the details of the procedure) is :

E(lifeexpectancyatbirth)=β0+β1(Grossnationalincome.GNI.percapita) + β2(Expectedyearsofschooling)+ β3(GNIpercapitarankminusHDIrank)

E(lifeexpectancyatbirth)= 53.959017+0.000227 (Grossnationalincome.GNI.percapita) + 0.932783 (Expectedyearsofschooling)+ 0.154356 (GNIpercapitarankminusHDIrank)

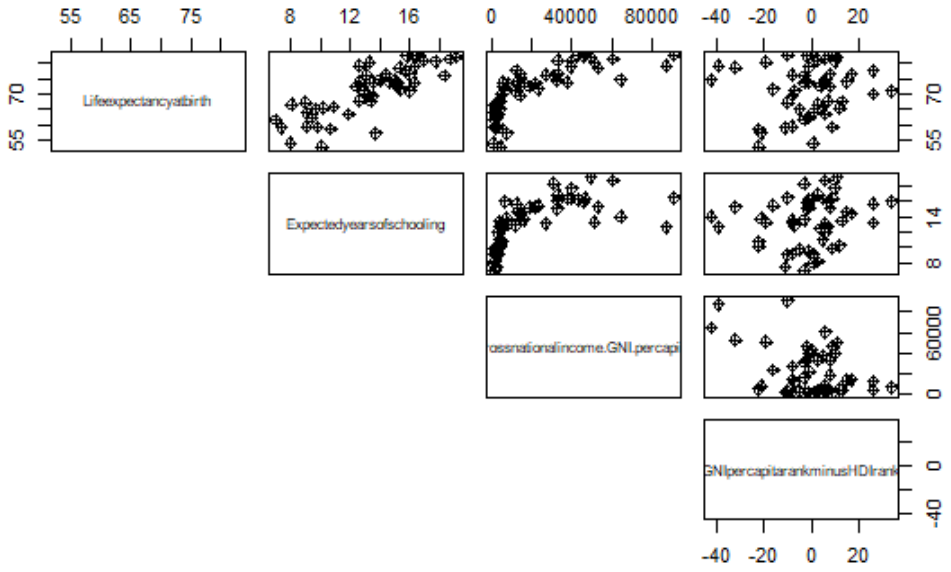


Figure 1

1

By observing this figure 1, we find out:

- *Life expectancy at birth* and *expected years of schooling* appear to have a strong positive linear correlation
- *Life expectancy at birth* and *GNI per capita* appear to have a strong positive linear correlation
- *Life expectancy at birth and GNI minus HDI rank* appear to have a weak positive linear correlation
- There is no multicollinearity between explanatory variables.

<u>Rechecking the assumptions of the model:</u>

First: Assumptions concerning the deterministic part of the model (checked before we fit the model):

Assumption 1. The parameters are linear in the deterministic part of the model. That is, it makes sense to model the relationship between Y and X with a linear function. And that's what we can observe in figure 1.

Assumption 2. The values of the explanatory variables are recorded without error. By checking the nature of the experiment, the explanatory variables are measured precisely.

Assumption 3. The explanatory variables are fixed in repeated samples.

Assumption 4. Reasonable variation in the values of the explanatory variables, as each explanatory variable takes different values in the data.

Assumption 5. The sample size is greater than the number of parameters to be estimated ($n=50>p=3$), This assumption is important for the estimation of β, so that there are enough degrees of freedom to estimate the model parameters.

Assumption 6. No multicollinearity between the explanatory variables (what we can observe in figure 1)

-So, all the assumptions regarding the deterministic part of the model are valid.

Second: Assumptions concerning the random part of the model (will be checked after we fit the model):

Assumption 7. The expected value of the random error is zero, this implies that the deterministic part of the model captures all the non-random structure in the data.

Assumption 8: The variance of the residuals should be consistent for all observations: This preferred condition is known as homoskedasticity. Violation of this assumption is known as heteroskedasticity.

Assumption 9. The errors are independent (We don't have here a time series and hence no need for the plot of the residuals versus the order of the observations).

To check if assumptions 7,8 is met, we can create a *fitted value vs. residual plot:* Figure 2 displays the plot of standardized residuals versus fitted values, where we can see that the points are evenly scattered above and below the zero line. The points are slightly grouped together towards the upper fitted values. However, this effect is not very marked and thus it is reasonable to assume that the linear relationship between life expectancy at birth and its explanatory variables captures the non-random structure in the data. This suggests that it is reasonable to assume that the random errors have mean zero and hence that a linear regression model is appropriate in this case. The vertical variation of the points in

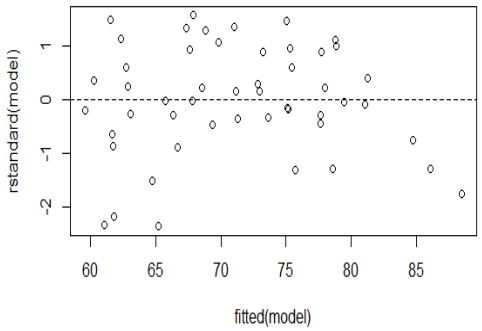
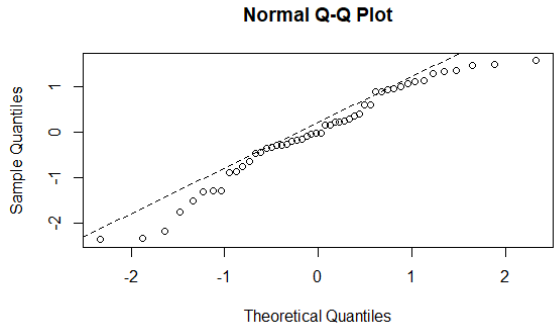


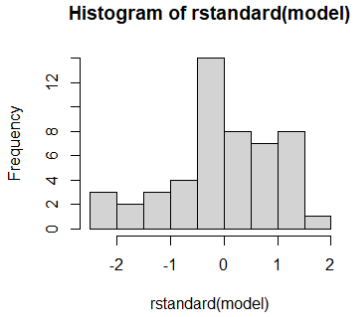
Figure 2

Figure 2 is constant across the range of the fitted values and does not seem to depend on the fitted values. Hence, it seems reasonable to assume that the random errors have constant variance. (Assumptions 7,8 is valid)

Assumption 10. The errors must be Normally distributed, i.e. I ~ N (0, σ2) for all I = 1, . . ., n. it is very important for making inference, as we observed figure 3, we found out that this

assumption is not perfectly valid, so we tried to do a log transformation but it didn't do a major difference and by observing figure 4 Although the distribution is slightly left skewed, it isn't abnormal enough to cause any major concerns so we will





continue with our model .

figure 3 figure 4

-so, the assumptions regarding the error term are perfectly valid except that for the normal distribution but we will continue with our model.

Test the significance of the model:

H0: b1=b2=b3=0

H1=at least one B is different from 0.

Table (1) ANOVE TABLE for the model;

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Expectedyearsofschooling	1	2195.51	2195.51	150.7511	4.042e-16 ***
Grossnationalincome.GNI.percapita	1	305.24	305.24	20.9591	3.566e-05 ***
GNIpercapitarankminusHDIrank	1	140.17	140.17	9.6248	0.003277 **
Residuals	46	669.93	14.56		

All p-values are less than alpha (0.05). Therefore, reject H0 and move on to the alternative hypothesis.
All Beta values are statistically significantly different from zero, indicating that the explanatory variables in the Model have a significant role in explaining the non-random variability occurring in the values of y, and that the should remain in the model.

The final model between a dependent variable and three independent variables:

1)E(lifeexpectancyatbirth)=β0+ β1(Grossnationalincome.GNI.percapita) + β2(Expectedyearsofschooling)+ β3(GNIpercapitarankminusHDIrank)

2)E(lifeexpectancyatbirth)= 53.959017+0.000227 (Grossnationalincome.GNI.percapita) + 0.932783 (Expectedyearsofschooling)+ 0.154356 (GNIpercapitarankminusHDIrank)

Table (2) summary of the model:

Residuals:

Min	1Q	Median	3Q	Max
-8.3717	-1.7361	-0.0888	3.2845	5.8848

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.396e+01	3.239e+00	16.657	< 2e-16	***
Expectedyearsofschooling	9.328e-01	2.882e-01	3.237	0.00224	**
Grossnationalincome.GNI.percapita	2.270e-04	4.144e-05	5.478	1.74e-06	***
GNIpercapitarankminusHDIrank	1.544e-01	4.975e-02	3.102	0.00328	**

Residual standard error: 3.816 on 46 degrees of freedom
Multiple R-squared: 0.7977, Adjusted R-squared: 0.7845
F-statistic: 60.44 on 3 and 46 DF, p-value: 5.415e-16

From the above R output, we can see that the regression equation is:
Life expectancy at birth= 53.96+0.9328expectedyearsofschooling+0.227GNIpercapita+0.1544GNIminusHDI

- **Intercept:** It's the prediction made by model when all the independent variables are set to zero, and in our model when there is zero expected years of schooling / GNI per capita / GNI minus HDI rank the life expectancy at birth = 53.96.
- **Estimate:** This represents regression coefficients for respective variables. It's the value of slope. For example Expected years of schooling is statistically significant at the 0.10 significance level. In particular, the coefficient from the model output tells is that a one unit increase in *expected years of schooling* is associated with a 0,9328 unit increase, on average, in *life expectancy at birth*, assuming *GNI* and *GNI minus HDI rank* are held constant
- **Std. Error:** This determines the level of variability associated with the estimates. Smaller the standard error of an estimate is, more accurate will be the predictions.
- **T-value:** t statistic is generally used to determine variable significance. if a variable is significantly adding information to the model. T-value > 2 suggests the variable is significant. we used it as an optional value as the same information can be extracted from the p value.
- **P-value:** It's the probability value of respective variables determining their significance in the model. P-value < 0.05 is always desirable. And we noticed that there are four variables that their p value of them greater than 0.05

- **F Statistics:** It evaluates the overall significance of the model. It is the ratio of explained variance by the model by unexplained variance. It compares the full model with an intercept only (no predictors) model. Its value can range between zero and any arbitrary large number. Naturally, higher the F statistics, better the model. And here it is equal 60.44 and its corresponding p-value is 5.415e-16 This indicates that the overall model is statistically significant. In other words, the regression model is useful (good fit).

Assessing the Goodness of Fit of the Model:

To assess how "good" the regression model fits the data, we can look at a couple different metrics:

Multiple R-Squared: This measures the strength of the linear relationship between the predictor variables and the response variable. A multiple R-squared of 1 indicates a perfect linear relationship while a multiple R-squared of 0 indicates no linear relationship whatsoever. Multiple R is also the square root of R-squared, which is the proportion of the variance in the response variable that can be explained by the predictor variables. In this example, the multiple R-squared is **0.7977**, thus 79.77% of the variance in *life expectancy at birth* can be explained by the predictors in the model, thus the model is a good fit.

- **Residual Standard Error**: This measures the average distance that the observed values fall from the regression line. In this example, the observed values fall an average of **3.816 units** from the regression line**.**

Analysis of variance:

The same as table (1):

Analysis of Variance Table
Response: Lifeexpectancyatbirth

	Df	Sum Sq	Mean Sq
Expectedyearsofschooling	1	2195.51	2195.51
Grossnationalincome.GNI.percapita	1	305.24	305.24
GNIpercapitarankminusHDIrank	1	140.17	140.17
Residuals	46	669.93	14.56

	F value	Pr(>F)
Expectedyearsofschooling	150.7511	4.042e-16
Grossnationalincome.GNI.percapita	20.9591	3.566e-05
GNIpercapitarankminusHDIrank	9.6248	0.003277
Residuals		

-The above table shows 4 rows relating to different sources of variation and several columns containing calculated values related to each source of variance.

Row 1,2,3 relates to variation between the means of the groups; the values are almost always either referred to as "between group" terms or are identified by the 'grouping factor'.

Row 4 refers to variation within each group

Sum of squares (SS) The SS terms are calculated by adding a series of squared error terms. In the case of within-group SS term SS w, we are interested in the differences between the individual data points and the mean of the group to which they belong.

Degrees of freedom for the one-way ANOVA in the table, the total number of data points is N and the number of groups of data is k. The total number of degrees of freedom is N – 1, just as for a simple data set of size N. There are k different groups and therefore, k – 1 degrees of freedom for the between-group effect. The degrees of freedom associated with the within-group SS term is the difference between the two values, N – k.

Mean squares MS The mean squares are the key term in classical ANOVA. They are variances, calculated by dividing the between- and within-group sum of squares by the appropriate number of degrees of freedom.

F-value: The F value is a test statistic that allows to test whether there is sufficient evidence that at least one of the model parameters (except the intercept) is not zero. In SLRM, this test is equivalent to testing the significance of β1 (i.e., H0: β1 = 0 versus H0: β1 6= 0). The f-stat for b1 is 150.75 which indicates that expected years of schooling is sufficient.

Conclusion

We used the R program to obtain the results shown above, and we conclude from this that the dependent variable, the life expectancy at birth, is significantly affected by the Gross national income (GNI)per capita, Expected years of schooling, and GNI per capita rank minus HDI rank. and this means that changes in the values of life expectancy at birth are well explained by the multiple linear regression model above for Through the independent variable. An important thing that indicate the improvement of the model - comparing to the 1st model that contains 7 explanatory variables- is that the (Residual standard error) has a higher value in our model which is equal to 3.8 compared to the 1st model which is 1.18 , which indicates that the final model is more significant than the 1st one, therefore in the end we concluded this good model to describe the changes in the dependent variable by the independent variables.

APPENDIX

Here we want to state the forward, backward and stepwise regression for our model with their summaries

Forward method:

```
Start:  AIC=211.65
Lifeexpectancyatbirth ~ 1

                                Df Sum of Sq    RSS    AIC
+ Expectedyearsofschooling           1   2195.51 1115.4 159.25
+ Grossnationalincome.GNI.percapita  1   1884.37 1426.5 171.55
<none>                                       3310.9 211.65
+ GNIpercapitarankminusHDIrank       1     37.18 3273.7 213.08
Step:  AIC=159.25
Lifeexpectancyatbirth ~ Expectedyearsofschooling

                                Df Sum of Sq     RSS    AIC
+ Grossnationalincome.GNI.percapita  1   305.244  810.11 145.26
<none>                                       1115.35 159.25
+ GNIpercapitarankminusHDIrank       1     8.305 1107.05 160.87
Step:  AIC=145.26
Lifeexpectancyatbirth ~ Expectedyearsofschooling + Grossnationalincome.GNI.percapita

                                Df Sum of Sq    RSS    AIC
+ GNIpercapitarankminusHDIrank  1    140.17 669.93 137.76
<none>                                       810.11 145.26
Step:  AIC=137.76
Lifeexpectancyatbirth ~ Expectedyearsofschooling + Grossnationalincome.GNI.percapita +
    GNIpercapitarankminusHDIrank
Call:
lm(formula = Lifeexpectancyatbirth ~ Expectedyearsofschooling +
    Grossnationalincome.GNI.percapita + GNIpercapitarankminusHDIrank,
    data = sample)
Coefficients:
                  (Intercept)          Expectedyearsofschooling
                    53.959017                          0.932783
Grossnationalincome.GNI.percapita       GNIpercapitarankminusHDIrank
                     0.000227                          0.154356
Start:  AIC=211.65
Lifeexpectancyatbirth ~ 1

                                Df Sum of Sq    RSS    AIC
+ Expectedyearsofschooling           1   2195.51 1115.4 159.25
+ Grossnationalincome.GNI.percapita  1   1884.37 1426.5 171.55
<none>                                       3310.9 211.65
```

```
+ GNIpercapitarankminusHDIrank       1     37.18 3273.7 213.08
Step:  AIC=159.25
Lifeexpectancyatbirth ~ Expectedyearsofschooling

                                Df Sum of Sq     RSS     AIC
+ Grossnationalincome.GNI.percapita  1   305.244  810.11 145.26
<none>                                       1115.35 159.25
+ GNIpercapitarankminusHDIrank       1     8.305 1107.05 160.87
Step:  AIC=145.26
Lifeexpectancyatbirth ~ Expectedyearsofschooling + Grossnationalincome.GNI.percapita

                            Df Sum of Sq    RSS     AIC
+ GNIpercapitarankminusHDIrank  1    140.17 669.93 137.76
<none>                                   810.11 145.26
Step:  AIC=137.76
Lifeexpectancyatbirth ~ Expectedyearsofschooling + Grossnationalincome.GNI.percapita +
    GNIpercapitarankminusHDIrank
summary(forward.model)
Call:
lm(formula = Lifeexpectancyatbirth ~ Expectedyearsofschooling +
    Grossnationalincome.GNI.percapita + GNIpercapitarankminusHDIrank,
    data = sample)
Residuals:
    Min      1Q  Median      3Q     Max
-8.3717 -1.7361 -0.0888  3.2845  5.8848
Coefficients:
                               Estimate Std. Error t value
(Intercept)                   5.396e+01  3.239e+00  16.657
Expectedyearsofschooling      9.328e-01  2.882e-01   3.237
Grossnationalincome.GNI.percapita 2.270e-04  4.144e-05   5.478
GNIpercapitarankminusHDIrank      1.544e-01  4.975e-02   3.102
                              Pr(>|t|)
(Intercept)                    < 2e-16 ***
Expectedyearsofschooling       0.00224 **
Grossnationalincome.GNI.percapita 1.74e-06 ***
GNIpercapitarankminusHDIrank       0.00328 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.816 on 46 degrees of freedom
Multiple R-squared:  0.7977,             Adjusted R-squared:  0.7845
F-statistic: 60.44 on 3 and 46 DF,  p-value: 5.415e-16
BACKWARD:

 Start:  AIC=137.76
 Lifeexpectancyatbirth ~ Expectedyearsofschooling + Grossnationalincome.GNI.percapita +
```

GNIpercapitarankminusHDIrank

```
                                Df Sum of Sq     RSS     AIC
<none>                                       669.93 137.76
- GNIpercapitarankminusHDIrank       1    140.17  810.11 145.26
- Expectedyearsofschooling           1    152.59  822.53 146.02
- Grossnationalincome.GNI.percapita  1    437.11 1107.05 160.87
```

summary(backward.model)

```
Call:
lm(formula = Lifeexpectancyatbirth ~ Expectedyearsofschooling +
    Grossnationalincome.GNI.percapita + GNIpercapitarankminusHDIrank,
    data = sample)
Residuals:
    Min      1Q  Median      3Q     Max
-8.3717 -1.7361 -0.0888  3.2845  5.8848

Coefficients:
                                Estimate Std. Error t value
(Intercept)                     5.396e+01  3.239e+00  16.657
Expectedyearsofschooling        9.328e-01  2.882e-01   3.237
Grossnationalincome.GNI.percapita 2.270e-04  4.144e-05   5.478
GNIpercapitarankminusHDIrank    1.544e-01  4.975e-02   3.102
                                Pr(>|t|)
(Intercept)                      < 2e-16 ***
Expectedyearsofschooling         0.00224 **
Grossnationalincome.GNI.percapita 1.74e-06 ***
GNIpercapitarankminusHDIrank     0.00328 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.816 on 46 degrees of freedom
Multiple R-squared:  0.7977,             Adjusted R-squared:  0.7845
F-statistic: 60.44 on 3 and 46 DF,  p-value: 5.415e-16
STEP:

 Start:  AIC=211.65
 Lifeexpectancyatbirth ~ 1

                                Df Sum of Sq    RSS     AIC
 + Expectedyearsofschooling         1   2195.51 1115.4 159.25
 + Grossnationalincome.GNI.percapita  1   1884.37 1426.5 171.55
 <none>                                      3310.9 211.65
```

```
+ GNIpercapitarankminusHDIrank       1     37.18 3273.7 213.08
Step:  AIC=159.25
Lifeexpectancyatbirth ~ Expectedyearsofschooling

                                Df Sum of Sq    RSS    AIC
+ Grossnationalincome.GNI.percapita  1    305.24  810.1 145.26
<none>                                      1115.4 159.25
+ GNIpercapitarankminusHDIrank       1      8.31 1107.0 160.87
- Expectedyearsofschooling           1   2195.51 3310.9 211.65
Step:  AIC=145.26
Lifeexpectancyatbirth ~ Expectedyearsofschooling + Grossnationalincome.GNI.percapita

                                Df Sum of Sq     RSS    AIC
+ GNIpercapitarankminusHDIrank       1    140.17  669.93 137.76
<none>                                       810.11 145.26
- Grossnationalincome.GNI.percapita  1    305.24 1115.35 159.25
- Expectedyearsofschooling           1    616.38 1426.49 171.55
Step:  AIC=137.76
Lifeexpectancyatbirth ~ Expectedyearsofschooling + Grossnationalincome.GNI.percapita +
    GNIpercapitarankminusHDIrank

                                Df Sum of Sq     RSS    AIC
<none>                                       669.93 137.76
- GNIpercapitarankminusHDIrank       1    140.17  810.11 145.26
- Expectedyearsofschooling           1    152.59  822.53 146.02
- Grossnationalincome.GNI.percapita  1    437.11 1107.05 160.87
```

summary(stepwise.model)

```
Call:
lm(formula = Lifeexpectancyatbirth ~ Expectedyearsofschooling +
    Grossnationalincome.GNI.percapita + GNIpercapitarankminusHDIrank,
    data = sample)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3717 -1.7361 -0.0888  3.2845  5.8848

Coefficients:
                                  Estimate Std. Error t value
(Intercept)                      5.396e+01  3.239e+00  16.657
Expectedyearsofschooling         9.328e-01  2.882e-01   3.237
Grossnationalincome.GNI.percapita 2.270e-04  4.144e-05   5.478
```

```
GNIpercapitarankminusHDIrank      1.544e-01  4.975e-02   3.102
                              Pr(>|t|)
(Intercept)                     < 2e-16 ***
Expectedyearsofschooling          0.00224 **
Grossnationalincome.GNI.percapita 1.74e-06 ***
GNIpercapitarankminusHDIrank       0.00328 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.816 on 46 degrees of freedom
Multiple R-squared:  0.7977,             Adjusted R-squared:  0.7845
F-statistic: 60.44 on 3 and 46 DF,  p-value: 5.415e-16
```