Report about The Human Development Index for year 2021

Course:

Regression Analysis SE304

Presented By:

Student Name	ID
Mareez Maher Maawad	5200524
Shahd Hesham Elsayed Hassan	5200384

Under the Supervision of:

Dr. Amira El Ayouti

Content:

Introduction:

In this report, we'll be using the R program to perform a multiple linear regression model using the data provided in the annual HDI report. Having the life expectancy at birth as the response variable (Y) and 7 explanatory variables (X's) which are:

1. Human Development Index Value
2. Human Development Index Rank 2021
3. Expected Years of Schooling
4. Mean Years of Schooling
5. Gross National Income per Capita (GNI)
6. Gross National Income Per capita Rank minus HDI rank
7. Human Development Index Rank 2020

Our aim is to fit the model to determine the true values and the effect of each explanatory variable on y, as well as which explanatory variables are worthwhile to be included in the model. Before beginning any analysis, we must ensure that the model assumptions are valid, so we can proceed with our analysis.

A sample of 50 countries was selected at random, and a multiple linear regression analysis was conducted using the above explanatory variables.

Descriptive Statistics:

Descriptive statistics help us simplify large amounts of data in a sensible way. The following table shows some statistical measures for the response and explanatory variables in the data, which may help us in visualizing the data available.

Variables	Mean	Variance	1st quartile	3rd quartile	Median
Life Expectancy at Birth	69.77	65.86025	63.42	74.45	70.25
HDI value	0.7012	0.02390332	0.5765	0.8057	0.725
HDI rank 2021	102.78	3033.073	60.25	151.25	100.5
Expected Years of Schooling	13.31	9.173404	11.9	15.35	13.10
Mean Years of Schooling	8.536	10.58072	5.6	11.300	9
GNI	18104	406791162	4316	20945	12510
GNI rank – HDI rank	-2.44	209.68	-8	7.5	-0.5
HDI rank 2020	102.2	3032.069	60.5	151.8	100

Statistical Analysis:

1) Life Expectancy at Birth vs Expected Years of Schooling

In Figure 1, the correlation coefficient calculated was **0.7895455.**

We can conclude a strong linear relationship between Life Expectancy at Birth (Y) and Expected Years of Schooling (x_{i1}).

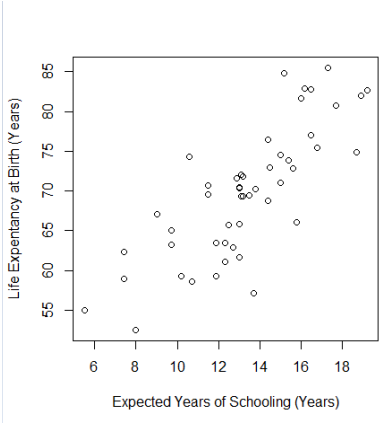

Figure 1: Relation between Y and Expected Years of Schooling

2) Life Expectancy at Birth vs GNI per Capita minus HDI rank

Figure 2 shows a slight indication of the existence of a pattern when plotting GNI per Capita minus HDI rank (x_{i2}) against the Life Expectancy at Birth (Y). In addition, the calculated correlation coefficient is equal to **0.3861718**, which indicates a weak but positive relationship between the response and explanatory variable.

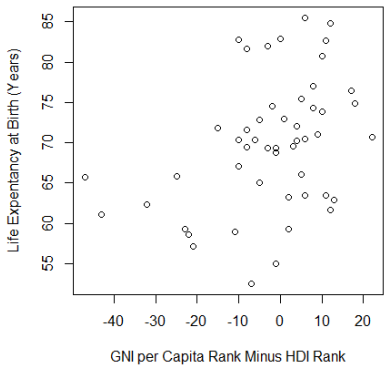

Figure2: Relation between Y and GNI rank-HDI rank

3) Life Expectancy at Birth vs GNI per Capita

In Figure 3, **r =0.80777** which indicates a strong positive linear relationship between Life Expectancy at Birth (Y) and the Gross National Income per Capita (x_{i3}).

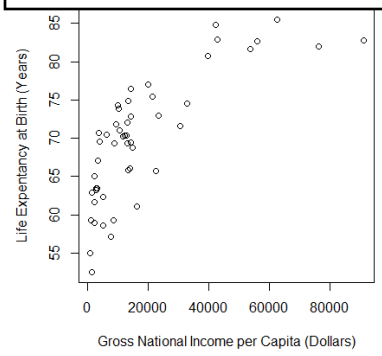

Figure3: Relation between Y and GNI

4)Life Expectancy at Birth vs Mean Years of Schooling

The calculated Pearson Correlation Coefficient is equal to **0.7479580,** which indicates a strong to a moderate positive linear relationship between the Mean Years of Schooling (x_{i5}) and the Life Expectancy at Birth.

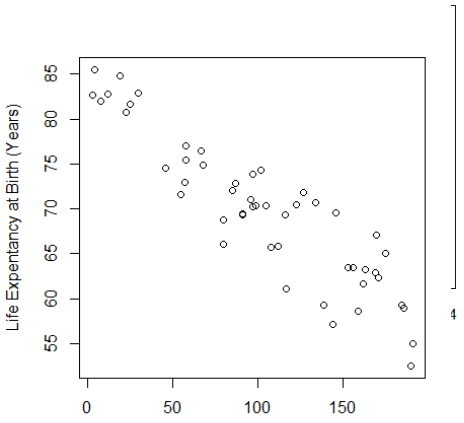

Figure4: Relation between Y and Mean Years of Schooling

5)Life Expectancy at Birth vs Human Development Index Value

As it's obvious from the scatter plot in figure 5, there happens to be a strong linear positive relation between (Y) and HDI value (x_{i4}). The correlation coefficient is equal to **0.9093036.**

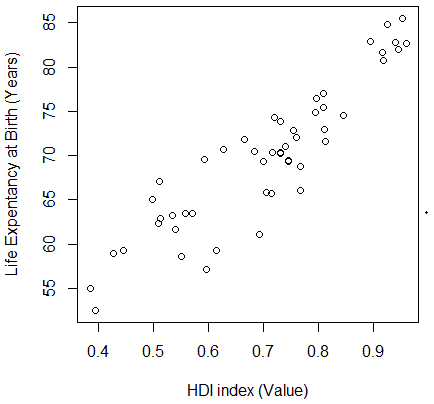

Figure5: Relation between Y and HDI value

6) Life Expectancy at Birth vs Human Development Index Rank 2020

As it's clear from the scatter plot, there is a negative linear relationship between HDI rank for the year 2020 (X6) and the Life Expectancy at Birth (Y). The calculated value of r = **-0.906189,** makes it more certain about the negative linear relation between (Y) and (x_{i16}).

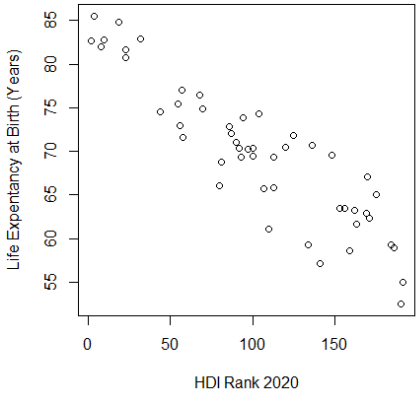

Figure6: Relation between Y and HDI rank for year 2020

7) Life Expectancy at Birth vs Human Development Index Rank 2021

When plotting Life Expectancy at Birth (Y) and HDI Rank for the year 2021 (x_{i7}), we've got the correlation coefficient to be **-0.909933**, which points to the negative strong linear relationship between (Y) and (x_{i7}).

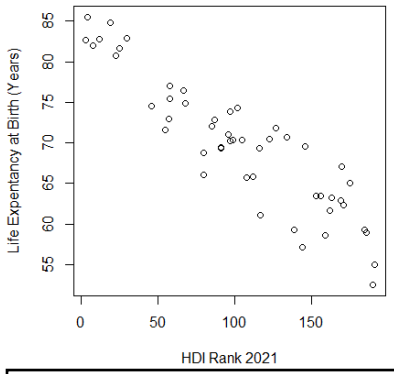

Figure 7: Relation between Y and HDI rank for year 2021

Checking Model Assumptions for the Deterministic Part

Assumptions regarding the deterministic part of the model we checked before fitting the model

1)The parameters are linear in the deterministic part of the model.

- This was shown when we plotted each explanatory variable with the response variable and the plots revealed a linear relationship between the X's and Y

2)The explanatory variables are recorded without error.

- Since the data for the explanatory variables were recorded by a trusted source, so this assumption is valid.

3)The explanatory variables are fixed in repeated samples.

4) Reasonable variation in the values of explanatory variables.

- Since every independent variable takes on different values in the data, so this assumption is satisfied.

5) The sample size must be greater than the number of parameters to be estimated.

- Sample Size = 50, Parameters = 8 (including the intercept).

6) No multicollinearity between the explanatory variables in multiple regression models.

- We'll be checking this assumption in the next page.

Multicollinearity:

As it's clear from figure 8, there is a negative strong linear relationship between the HDI rank for the year 2020 and all the other explanatory variables for example, between the HDI rank 2020 and the HDI value the correlation is **-0.990.**

We saw also there is a negative strong linear relationship between HDI rank 2021 and all the other explanatory as between HDI rank 2021 and HDI the correlation is **-0.990**

Moreover, it's shown that there is a positive strong linear relationship between the Mean Years of Schooling and the HDI value.

Since that we can't remove the 3 variables at once we remove them one by one.

First, we decided to remove the HDI rank for the year 2020 and the HDI rank for the year 2021.

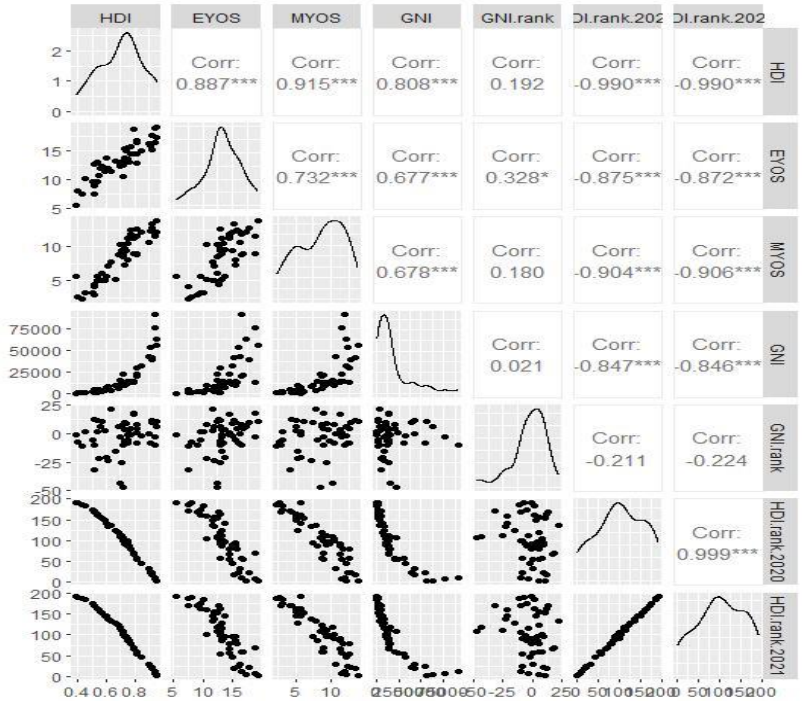

Figure8: Checking Multicollinearity

Second, regarding multicollinearity between Mean Years of Schooling and HDI value. In order to solve this issue, we decided to test the significance of both and see which one of them has a stronger relationship with Y in order to eliminate the other, and we concluded the following:

Both explanatory variables were significant.

HDI value has the strongest relationship with Life Expectancy at Birth with **r = 0.909**, as we computed previously.

And for the Mean Years of Schooling, a value of r = **0.748** was computed.

So, we decided to remove the Mean Years of Schooling from the model.

```
> cor.test(sample$Life.exp,sample$HDI)

        Pearson's product-moment correlation

data:  sample$Life.exp and sample$HDI
t = 15.139, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8447692 0.9477687
sample estimates:
      cor
0.9093036

> cor.test(sample$Life.exp, sample$MYOS)

        Pearson's product-moment correlation

data:  sample$Life.exp and sample$MYOS
t = 7.8072, df = 48, p-value = 4.312e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5930863 0.8494555
sample estimates:
      cor
0.747958
```

Figure 9: Correlation Test for the HDI value and Mean Years of Schooling

Fit the Model

After removing the multicollinearity, we'll be fitting our model with 4 explanatory variables only (Human Development Index Value, Expected Years of Schooling, Gross National Income Per capita, Gross National Income Rank minus Human Development Index Rank). In our data, we referred to these variables by (HDI, EYOS, GNI, GNI rank)

We can write the model in form of:

$$E(Life\ expectancy\ /xi) = \beta 0 + \beta 1\ (\ HDI\ ) + \beta 2\ (EYOS\ ) + \beta 3(GNI) + \beta 4(GNI\ rank)$$

$$E(Y\ /xi) = \beta 0 + \beta 1\ (\ x_{i1}) + \beta 2\ (x_{i2}\ ) + \beta 3(x_{i3}) + \beta 4(x_{i14})$$

Methods to Reach the Best Fitted Model

I. Backward Elimination Method:

Our aim is to know which variables will stay in the model and which variables will be excluded from the model based on backward elimination, which removes the most insignificant variables that have the lowest marginal contribution. Taking into consideration that this method starts with the full model. After applying the backward elimination method using R, we reached the conclusion that all four explanatory variables we've chosen to keep from the beginning (Human Development Index Value (HDI), Gross National Income Per capita Rank minus HDI rank (GNI rank), Expected Years of Schooling (EYOS), Gross National Income per Capita (GNI)), are all significant and should be added to the model.

```
> summary(backward.model1)

Call:
lm(formula = Life.exp ~ HDI + EYOS + GNI + GNI.rank, data = sample)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1244 -1.5636 -0.1302  1.3389  6.3739

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.383e+01  2.690e+00  16.294  < 2e-16 ***
HDI          4.854e+01  6.779e+00   7.161 5.88e-09 ***
EYOS        -6.846e-01  2.926e-01  -2.340   0.0238 *
GNI          7.755e-05  3.322e-05   2.334   0.0241 *
GNI.rank     1.618e-01  2.934e-02   5.514 1.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.683 on 45 degrees of freedom
Multiple R-squared:  0.8996,    Adjusted R-squared:  0.8907
F-statistic: 100.8 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
> backward.model1<-step(Model,direction = "backward")
Start:  AIC=103.43
Life.exp ~ HDI + EYOS + GNI + GNI.rank

           Df Sum of Sq    RSS    AIC
<none>                  323.97 103.43
- GNI       1     39.23 363.20 107.15
- EYOS      1     39.43 363.40 107.17
- GNI.rank  1    218.91 542.88 127.24
- HDI       1    369.18 693.16 139.46
```

Figure 10 and 11: Fitted Model of Backward Elimination Method

II. Forward selection method:

Unlike the backward elimination method, the forward selection technique starts with the null model with the intercept only. It depends on choosing the highest R-squared model and conducting the marginal contribution of adding X's to the model.

We fitted the model by the same four variables chosen before and it resulted in the same results conducted by the backward elimination method.

```
> summary(forward.model1)

Call:
lm(formula = Life.exp ~ HDI + GNI.rank + EYOS + GNI, data = noc)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1244 -1.5636 -0.1302  1.3389  6.3739

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.383e+01  2.690e+00  16.294  < 2e-16 ***
HDI          4.854e+01  6.779e+00   7.161 5.88e-09 ***
GNI.rank     1.618e-01  2.934e-02   5.514 1.63e-06 ***
EYOS        -6.846e-01  2.926e-01  -2.340   0.0238 *
GNI          7.755e-05  3.322e-05   2.334   0.0241 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.683 on 45 degrees of freedom
Multiple R-squared:  0.8996,    Adjusted R-squared:  0.8907
F-statistic: 100.8 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
Step:  AIC=87.93
Life.exp ~ HDI + GNI.rank

       Df Sum of Sq    RSS    AIC
+ GNI   1    35.063 233.57 83.358
+ EYOS  1    32.762 235.88 83.819
<none>              268.64 87.931

Step:  AIC=83.36
Life.exp ~ HDI + GNI.rank + GNI

       Df Sum of Sq    RSS    AIC
+ EYOS  1    30.097 203.48 78.874
<none>              233.57 83.358

Step:  AIC=78.87
Life.exp ~ HDI + GNI.rank + GNI + EYOS
```

Figures 12, 13: Results Conducted by Forward Selection Method

III. Using Stepwise Method:

The stepwise technique is considered a mixture between the forward selection method and backward elimination method, where it keeps on adding the independent variable which has the highest contribution to the data, then removes the independent variable that has the lowest marginal contribution to the data, until it reaches the best-fitted model for this data.

```
> summary(stepwisemodel)

Call:
lm(formula = Life.exp ~ HDI + GNI.rank + EYOS + GNI, data = noc)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1244 -1.5636 -0.1302  1.3389  6.3739

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.383e+01  2.690e+00  16.294  < 2e-16 ***
HDI          4.854e+01  6.779e+00   7.161 5.88e-09 ***
GNI.rank     1.618e-01  2.934e-02   5.514 1.63e-06 ***
EYOS        -6.846e-01  2.926e-01  -2.340   0.0238 *
GNI          7.755e-05  3.322e-05   2.334   0.0241 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.683 on 45 degrees of freedom
Multiple R-squared:  0.8996,    Adjusted R-squared:  0.8907
F-statistic: 100.8 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
Step:  AIC=78.87
Life.exp ~ HDI + GNI.rank + GNI + EYOS

           Df Sum of Sq    RSS     AIC
<none>                  203.48  78.874
- EYOS      1     30.10 233.58  83.358
- GNI       1     32.40 235.88  83.819
- GNI.rank  1    242.80 446.28 113.788
- HDI       1    380.18 583.66 126.401
```

Figures 14, 15: Results Conducted by Stepwise Method

General conclusion on the 3 methods (backward, forward, stepwise)

We have seen that the three methods give us the same results R-squared adjusted which is equal to **0.8907** and this is a good indicator that proves that the explanatory variables we've decided to include from the beginning, are significant and we should be added to the model, and it clarifies that there is no correlation between the explanatory variables (the absence of multicollinearity).

$$E\ (\hat{y}) = \hat{\beta}0 + \widehat{\beta}1\ x_{i1} + \hat{\beta}2\ x_{i2} + \hat{\beta}3x_{i3} + \hat{\beta}4\ x_{i4}$$

$$E(Y) = 43.83 - 0.6846(\ x_{i1}) + 0.1168\ (x_{i2}) + 0.00007755(x_{i3}) + 48.54(x_{i4})$$

Interpretation:

1. $\hat{\beta}0$: For zero units of all explanatory variables, the average of the life expectancy at birth (Y) is 43.83
2. $\widehat{\beta}1$: As the Expected Years of Schooling increase by one year(unit), the average of the life expectancy at birth decreases by 0.6846 units, holding $x_{i2}\ , x_{i3}, and\ x_{i4}$ constant.
3. $\hat{\beta}2$: As the GNI rank increase by 1 unit, the average increase in the life expectancy at birth is equal to 0.1168 units, holding x_{i3}, x_{i4}, x_{i1} constant.
4. $\hat{\beta}3$:As the GNI increase by 1 dollar, the average increase in Y is equal to0.00007755 units, holding all other explanatory variables constant.
5. $\hat{\beta}4$: As the HDI value increase by 1 unit, the average increase in Y is equal to 48.54 units, holding all other explanatory variables constant.

Checking Model Assumptions for the Random Part:

After fitting the model, we'll start to check the assumptions related to the random part:

1)The expected value of the random error is zero

- $E(\epsilon i\ ) = 0$
- Observations are randomly scattered above and below the zero line and there is no pattern appearing in the plot, so this condition is satisfied.

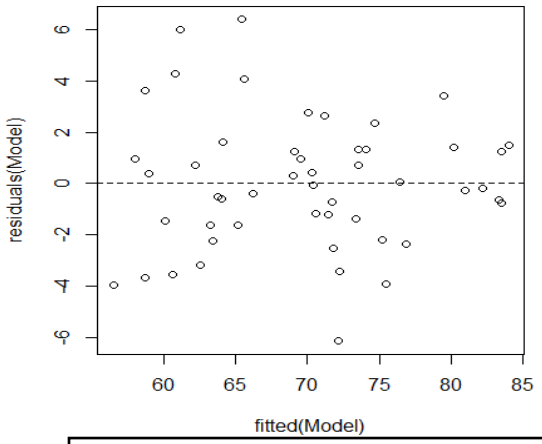

Figure 16: Fitted vs Residuals

2)The variability in the random error is constant

- Variation doesn't appear to be very constant as there is a little fanning in the middle of the plot of residuals (Figure 10).
- Therefore, we'll try to solve this issue by taking the log and square root transformations. But, since there are some explanatory variables that take on negative values, we'll only take the log and square root for the response variable.
- **Log Transformation:** $E(\log(Y)) = \beta0\ + \beta1\ (\ x_{i1}) + \beta2\ (x_{i2}\ ) + \beta3(x_{i3}) + \beta4(x_{i14})$
- In figure 21 and 22 in below appendix, it appears that the log transformation didn't solve the variance problem, as it still appears to be some fanning in in the data. Also, it created mor outliers in the data when plotting the Q-Q plot. It resulted in deviations from the equality line, so the normality assumption is dubious.
- **Square Root Transformation:** $E(\text{sqrt}(Y)) = \beta0\ + \beta1\ (\ x_{i1}) + \beta2\ (x_{i2}\ ) + \beta3(x_{i3}) + \beta4(x_{i14})$

- In figure 23 and 24 in the appendix, it's shown that the same conclusion conducted by the log transformation appeared once again, as there still occur fanning in, and there is deviations from the equality line.
- **Decision:** After trying both transformations, we've decided to stay along with no transformations at all, because the normal fitted model is the one with the least problems when comparing it with both transformations.

3) The errors are independent (no autocorrelation), and this is clear from the context as all countries are independent from each other in the sample we've selected.

4) Normality Assumption:

When drawing the Q-Q plot and the histogram, it's clear that there are three outliers, while all other points seem to lie on the equality line or near it. Additionally, by looking at the histogram, data seems to be somehow symmetric in figure 17.

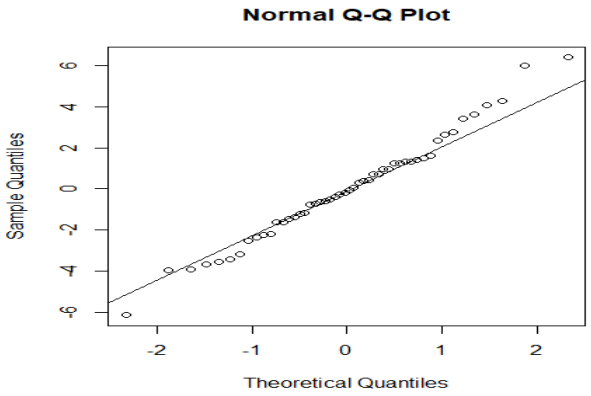

Figure 17: Normal Q-Q plot

Now we must check if these three outliers are influential or not in order to take a decision about keeping or leaving them in the data.

Using Cook's Distance, we have seen the result in figure 12 as these are not influential observations, therefore, we decided to keep them in our sample as n will be larger than the number of parameters which is assumption 5 in the deterministic part of the model (the larger the sample size, the better), and by keeping them we'll get degrees of freedom to be larger.

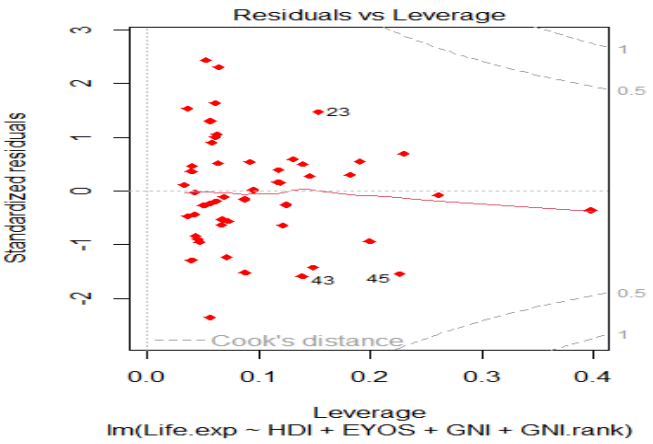

Figure 18: Cook's Distance

Assessing the Goodness of Fit:

R-squared adjusted is 0.8907 (for all methods) which is considered the goodness of fit measure for our case. This value shows that the model is a good fit for the data since R-squared adjusted is close to 1, Only 89.07 % of the variability in the Life Expectancy at Birth (Y) is explained by the fitted multiple linear regression model with the Expected Years of Schooling (x_{i1}), Gross National Income Per capita Rank minus HDI rank (x_{i2}), Gross National Income per Capita (x_{i13}), Human Development Index Value (x_{i4}), as predictors.

Testing the Significance of the Model:

$$H0 = \beta1 = \beta2 = \beta3 = \beta4 = 0$$

$$H1: At\ least\ one\ \beta\ is\ not\ equal\ zero$$

F-statistic = 100.8

Since the value of F is very large, this indicates that the model is a good for the data and that the model explains a significant amount of variability in the response variable. Also, since the p-values are all less than 0.05 therefore, we decide to Reject H0.

```
Call:
lm(formula = Life.exp ~ HDI + EYOS + GNI + GNI.rank, data = sample)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1244 -1.5636 -0.1302  1.3389  6.3739

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.383e+01  2.690e+00  16.294  < 2e-16 ***
HDI          4.854e+01  6.779e+00   7.161 5.88e-09 ***
EYOS        -6.846e-01  2.926e-01  -2.340   0.0238 *
GNI          7.755e-05  3.322e-05   2.334   0.0241 *
GNI.rank     1.618e-01  2.934e-02   5.514 1.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.683 on 45 degrees of freedom
Multiple R-squared:  0.8996,    Adjusted R-squared:  0.8907
F-statistic: 100.8 on 4 and 45 DF,  p-value: < 2.2e-16
```

Figure 19: Model Summary

We conclude that, $\beta's$ are statistically significantly different from zero.

Conclusion:

The main aim of this report was to study the effect of the 7 explanatory variables on the response variable (Life Expectancy at Birth), using RStudio.

We made some analysis and conducted the HDI rank for the year 2020, HDI rank for the year 2021, Mean years of schooling caused multicollinearity, and by this way, we satisfied the assumptions of the deterministic part of the model. Then we started fitting the model with the remaining 4 explanatory variables. Then we checked the assumptions of the random part of the model we have noticed that not all assumptions were valid, so we performed some transformations and conducted the graph of Cook's Distance in order to take a decision whether to remove the outliers or not. The outliers weren't influential, so we decided to keep them. Moreover, one of the main reasons that we decided to keep the outliers for was to increase the degrees of freedom and the sample size. We have used different methods to decide the best- fitted model for our data, and all methods agreed on the same model. R-squared adjusted indicated that the model is a good fit for the data as its value was close to 1. Afterwards, we tested the significance of the model were the F-value appeared to be a very large value, and that assured that the right decision at the beginning when we removed multicollinearity.

Appendix:

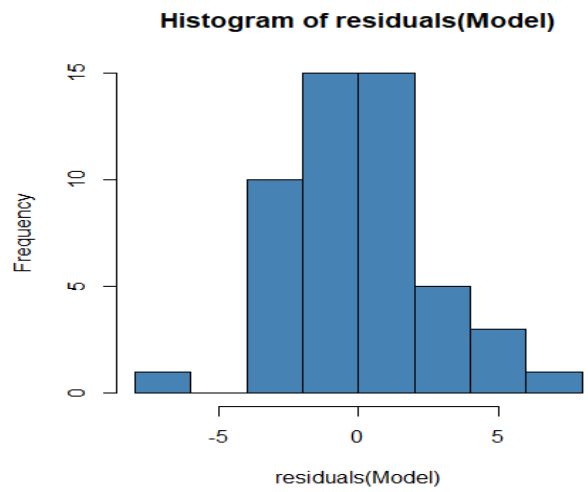

Figure 20: Histogram for the Fitted Model

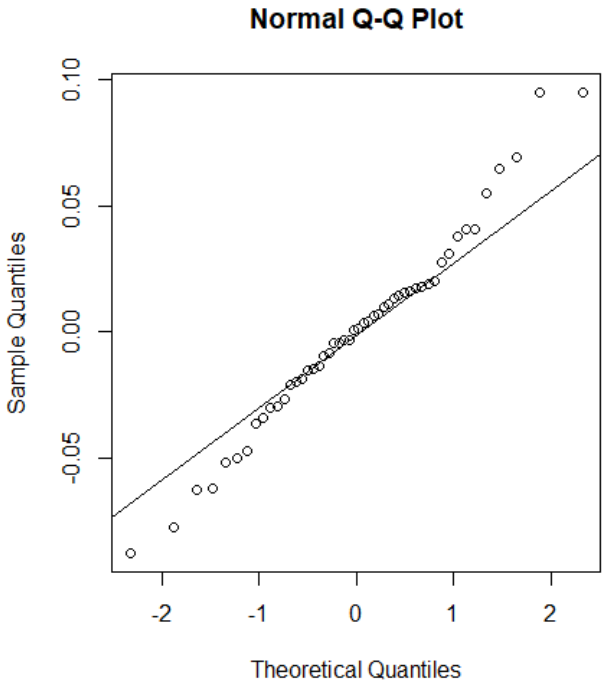

Figure 21: Q-Q plot for the Log Transformation Model

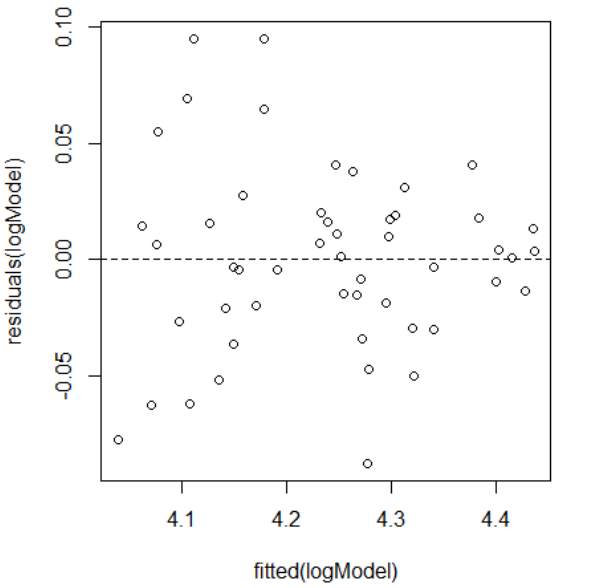

Figure 22: Fitted vs Residuals for the Log transformation Model

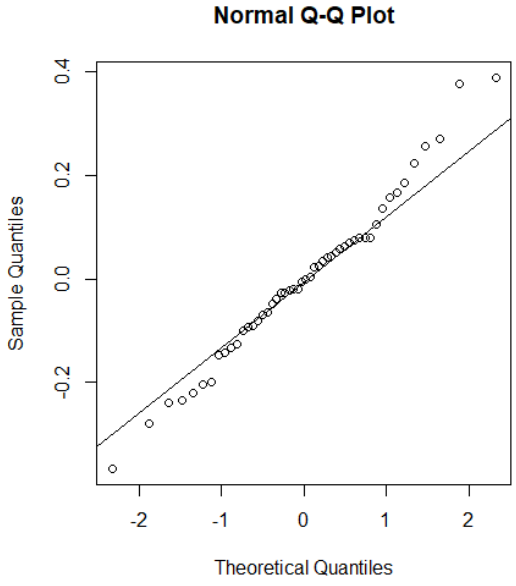

Figure 23: Q-Q plot for sqrt transformation model

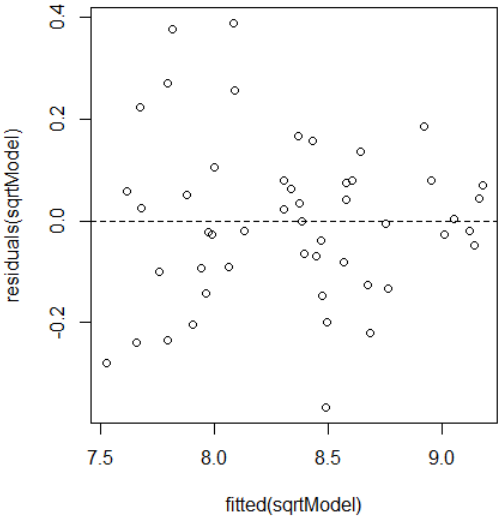

Figure 23: Fitted vs Residuals for Sqrt Transformation Model