Faculty of Economics and Political Science

Cairo University

English Section

Analyzing life expectancy thorough a regression model

Student's Names:

Khaled Mohamed Aboul-Azm - 5190671

Hania Ahmed Mohamed Abd El Rahman El Demery - 5190593

Abdelrahman Ashraf Abdelmohsen El-sayed -5190312

Maryam Adnan Ahmed -5180934

Under the Supervision of:

Prof. Amira EL-Ayouti

Teacher assistant:

Dr.Reham El-Shaer

LIFE EXPECTANCY

Statistical Report

ABSTRACT

In this report we use data on life expectancy to try and predict it using Regression models. Using variables from HDI to Unemployment to GDP per capita.

Regression Analysis

Introduction

The objective of this paper is to test the relationship between the life expectancy at birth and some variables of interest (Ranking on the Human Development Index, Total population in millions, Population with at least some secondary education "percentage age 25 and over", GDP per capita, Unemployment rate "% of labor force", Urban population relative to total population (%), Current health spending "% of GDP", Government spending on education "% of GDP"). This paper establishes a model that is able to predict Life expectancy through the unemployment rate, population with at least some secondary education ages 25 and older, GDP per capita, and government expenditure on education. While fixing all the assumptions necessary, removing influential outliers, and selecting the optimal model.

Data collection

"All reasonable precautions have been taken by the HDRO to verify the information contained in this report". This quote is directly extracted from the 2019 human development report ,It signifies the validity of the first 5 assumptions of the deterministic part of the model .A closer look into the data itself we can determine that it is a cross sectional data ,which is data collected at the same time point, therefore we assumed the validity of error independence assumption. Later on we will test the validity of the No-Autocorrelation assumption.

Variable description

Ranking on the Human Development Index:

HDI is a composite index of a country's achievements in basic three aspects; health(measured by life expectancy at birth), knowledge(measured by mean years of education) and standard of living(measured by GDP per capita) indicators. So the ranking of countries on this indicators is based on these basic indicators.

Total population in millions:

This represents the number of people in each country

Population with at least some secondary education (percentage age 25 and over):

This represents the number of people living in each country with secondary education

Unemployment rate (% of labor force):

It measures the unemployed people divided by the people in the labor force(Labor Force= unemployed+employed).

GDP per capita:

It measures each country's economic output per person(GDP per capita=GDP/Population).

Urban population relative to total population (%):

It represents the people living in Urban areas divided by total population in the country

Current health spending (% of GDP):

It represents the final consumption of health care goods and services and the percentage it represents of GDP.

Government spending on education (% of GDP):

Percentage of government expenditure of the GPD

Variable Designations

VARIABLE NAME	DESIGNATION
Healthy life expectancy at birth years	$Lifex$
Rank on the human development index	$RANKHDI$
Total population in million	$Tpopulation$
Population with at least some secondary education ages 25 and older	$edu25+$
Gross Domestic Product per capita	$GDPcap$
Unemployment rate of labour force	$Unemploy$
Urban population to the total population	$urbanratio$
Current health expenditure of GDP	$HEAlTH$
Government expenditure on education of GDP	$Education$

Descriptive Statistics

The following table shows the Structure of our Dataset. As we can see all of our variables are numeric so most of the following testing will proceed with ease. n= 35

```
tibble [35 x 9] (S3: tbl_df/tbl/data.frame)
 $ LIFEX      : num [1:35] 70.2 70.9 69 69.3 69.5 73 71.7 66.3 67.8 65.8 ...
 $ RANKHDI    : num [1:35] 1 8 14 15 17 19 26 36 45 47 ...
 $ Tpopulation: num [1:35] 5.34 9.97 4.74 67.14 11.48 ...
 $ edu25+     : num [1:35] 95.4 88.9 96.9 84.5 84.8 ...
 $ GDPcap     : num [1:35] 65441 47194 36355 40158 43218 ...
 $ Unemploy   : num [1:35] 3.9 6.4 4.5 4 6.3 2.5 9.19 6.8 1 3.1 ...
 $ urbanratio : num [1:35] 82.2 87.4 86.5 83.4 98 91.6 80.4 53.7 89.3 84.5 ...
 $ HEAlTH     : num [1:35] 10.5 10.93 9.22 9.76 10.04 ...
 $ Education  : num [1:35] 7.55 7.55 6.3 5.54 6.55 ...
```

Figure A.1

The following tables are summary statistics for all the variables given

LIFEX	RANKHDI	Tpopulation	edu25+
Min. :51.50	Min. : 1.0	Min. : 0.1961	Min. : 6.026
1st Qu.:57.00	1st Qu.: 46.0	1st Qu.: 5.3955	1st Qu.:40.661
Median :64.90	Median :108.0	Median : 11.4822	Median :62.240
Mean :62.93	Mean :100.2	Mean : 25.5154	Mean :60.231
3rd Qu.:68.20	3rd Qu.:150.5	3rd Qu.: 32.2328	3rd Qu.:84.630
Max. :73.00	Max. :187.0	Max. :127.2022	Max. :99.904

Figure B.1

GDPcap	Unemploy	urbanratio	HEAlTH	Education
Min. : 660.3	Min. : 0.700	Min. :13.00	Min. : 3.113	Min. :1.470
1st Qu.: 2685.7	1st Qu.: 2.950	1st Qu.:39.05	1st Qu.: 5.311	1st Qu.:3.978
Median :10412.1	Median : 4.500	Median :56.10	Median : 6.616	Median :4.657
Mean :16717.4	Mean : 6.334	Mean :57.21	Mean : 7.194	Mean :4.852
3rd Qu.:33840.4	3rd Qu.: 7.750	3rd Qu.:81.30	3rd Qu.: 8.812	3rd Qu.:6.215
Max. :65441.3	Max. :27.000	Max. :98.00	Max. :16.532	Max. :7.552

Figure B.2

Variable standard deviation

Lifex	*RANKHDI*	*Tpopulation*	*edu25*	*GDPcap*	*Unemploy*	*urbanratio*	*HEAlTH*	*Education*
6.713	59.552	31.84	29.02	15793.5	5.63	24.08	2.84	1.60

Methodology

Initial model

- We decided to fit a model using all variables to try to appropriately predict life expectancy based on our data:

$$Lifex = RANKHDI + Tpopulation + edu25 + GDPcap + Unemploy + urbanratio + HEAlTH + Education$$

Since this model haven't met any of the assumptions yet, it would not be appropriate to add the coefficient values.

- The first 5 assumptions concerning the deterministic part are valid.
  There is no evidence to suggest a violation in the deterministic assumptions. (Covered in the data collection part of the report). The Model is linear in the parameters, no evidence that explanatory variables are recorded with error, variation in values of explanatory variables and (n =35) > (p = 8).
- Checking the multicollinearity assumption between the explanatory variables, we have done an informal test by getting the correlation coefficient of each one of the variables with the rest. Visually represented in the heatmap below.(Figure 1)

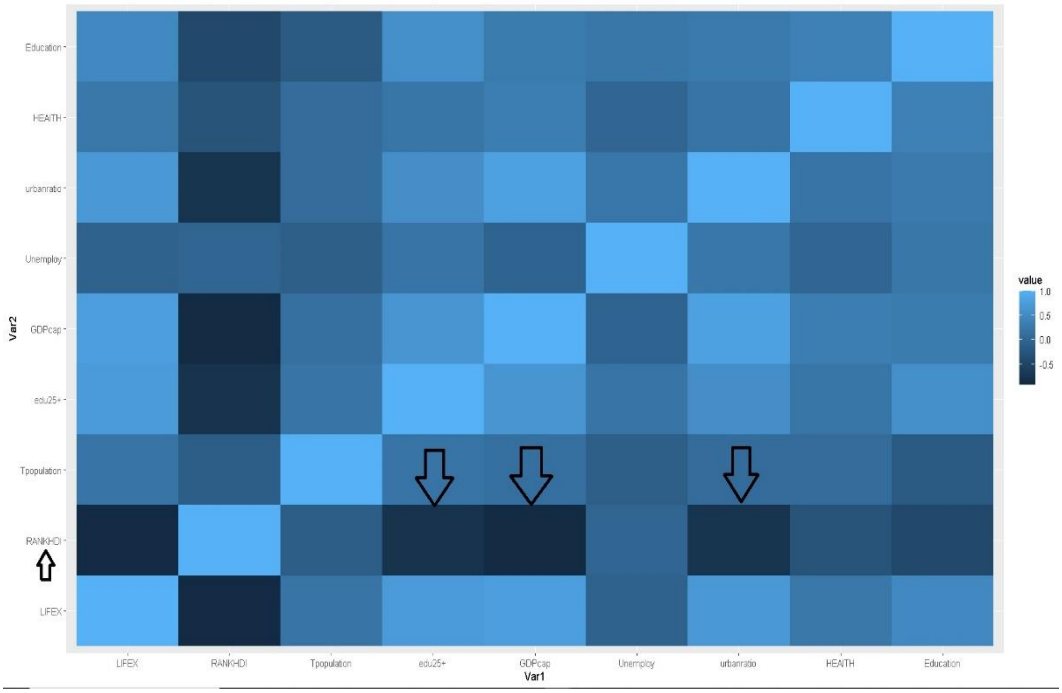

Figure 1 (Heatmap)

- We noticed that the RANKHDI variable is linearly dependent on two other variables, educated 25+ and GDP per capita, with a correlation coefficient of $r > 0.8$, implying that RANKHDI and the other explanatory variables have a strong positive linear relationship, and that the RANKHDI variable is causing multicollinearity in our model. By estimating the VIF values for our model, we can see that the RANKHDI has a VIF value greater than 10, indicating that it is causing multicollinearity. As a result, we eliminated the variable.

- All assumptions regarding the deterministic part of the model are satisfied

.

Let's check the assumptions regarding the random part of the model.

- Check that $E(i) = 0$ and $V(i) = \sigma 2$ for all $i = 1, \cdot \cdot \cdot 35$

For which the residuals diagnostics plots are as follows: (figure 2 and 3, respectively)

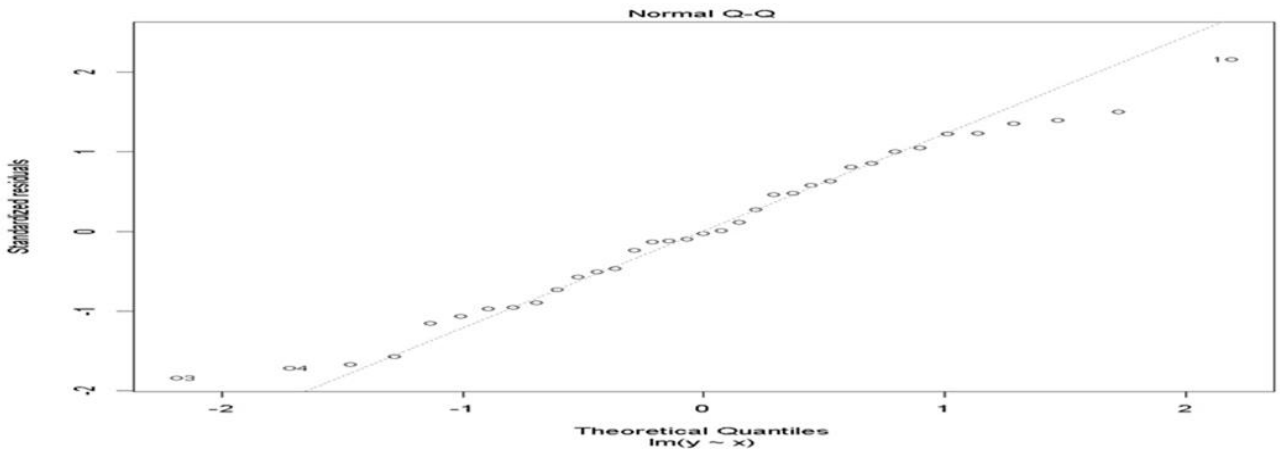

Figure 2

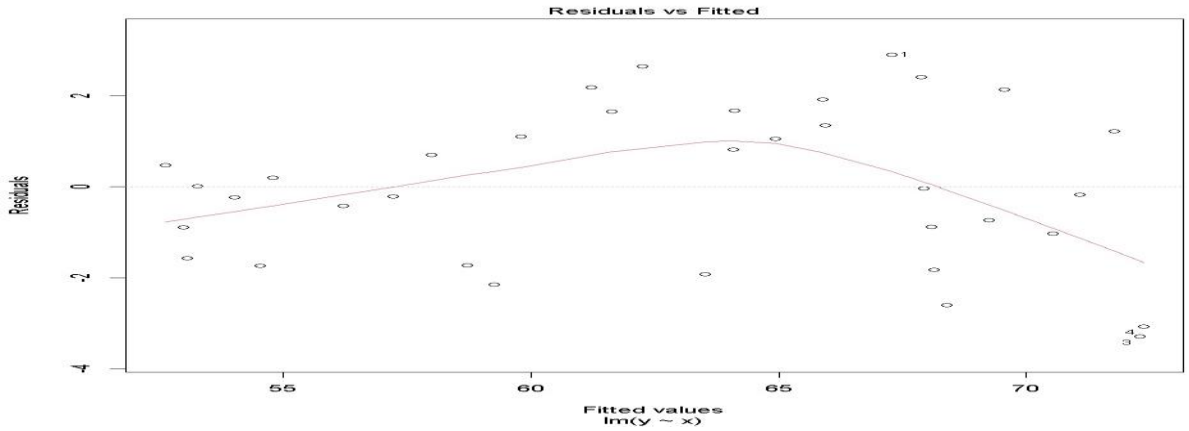

Figure 3

In the plot of residuals versus fitted values (Figure 3), the points seem to be fairly scattered below and above the zero line and there is an obvious change in the variance of the residuals. Indicating that the assumption of E (i) = 0 ∀i = 1, · · · 35 is valid, however that of homoscedasticity is not, i.e. V (i) = σ^2 i . The deterministic part of the model captures the non-random structure in the data but the error scale of variability is not constant at all values of the covariate. Later on, we transformed our model to try to approach a homoscedastic model, which seemed to decrease the issue but not totally fix it. Therefore, we conducted a formal heteroscedasticity test (Breusch pagan statistic), which yielded a insignificant result, confirming that our model doesn't suffer from heteroscedasticity.

- Check the assumption of independent error (no-auto correlation).

Because our data is cross-sectional, there appears to be no intuitive natural order in the explanatory variables (no lagged variables). Furthermore, we utilized the Durbin-Watson statistic to prove that our model does not suffer from autocorrelation. Thus, there is no need for a plot of the residuals versus observations order. So, we can assume independence between the errors.

- Check the assumption of normally distributed errors using the Normal Q-Q plot of the residuals.

In the normal Q-Q plot, there are some deviations from the equality line and the results indicate the presence of two potential outliers (figure2), hence the normality assumption is dubious (a large sample size would be preferred for taking a better conclusion). Since influential outliers caused the normality assumption to be violated, we had to remove them from our data. Utilizing Cook's distance, we were able to locate three influential outliers, which are South Africa, Sao Tome, and Sierra Leone. Which resonates logically as these countries suffer from abnormal conditions regard some of the variables. Afterwards, excluded them from our data. Removing those outliers seemed to decrease these deviations, however it was not sufficient, due to the fact that most variables are skewed .

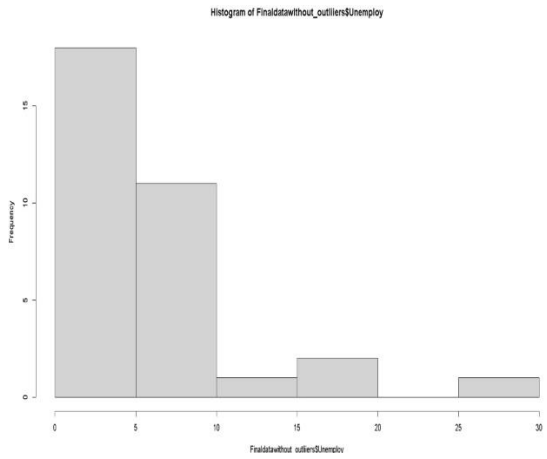

Figure 4 (Histogram unemployment)

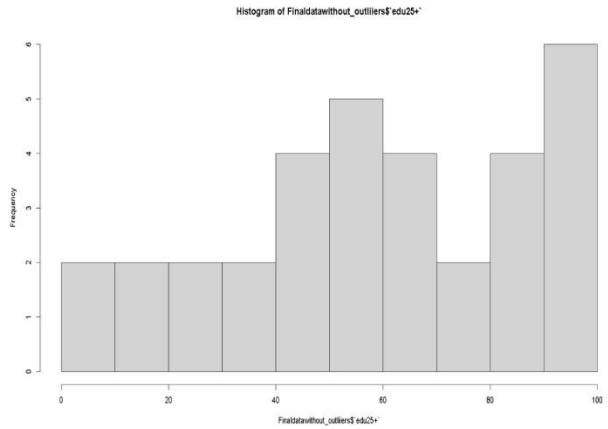

Figure 5 (Histogram edu25+)

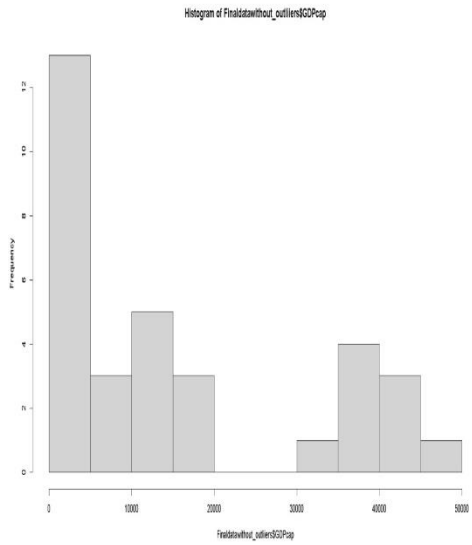

Figure 6 (Histogram GDPcap)

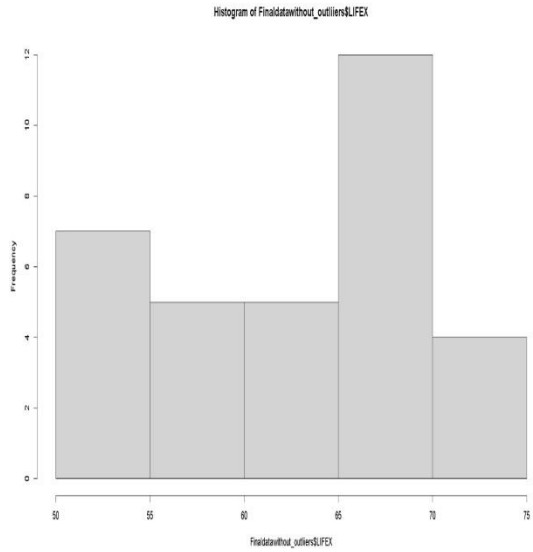

Figure 7 (Histogram Lifeex)

All distributions are skewed (before log transformation)

In order to solve these problems, we will use "log transformation" for both the response and the explanatory variables. As it is very useful for reducing both the non-constant variance and the skewness in the data when most of the data appears to be grouped at low values of x or y or both.

Fitting this model

$$\log(lifex) = log(Unemploy) + log(edu25) + log(GDPcap) + log(Education) + \log(Tpopulation) + \log(HEAlTH) + \log(urbanratio)$$

All assumptions are now valid with this transformation (they look reasonably satisfied) after removing the influential observations (outliers). It is important to reiterate that our sample size (n) is small, hence most of the deviations seen below would not be present with a larger sample size.

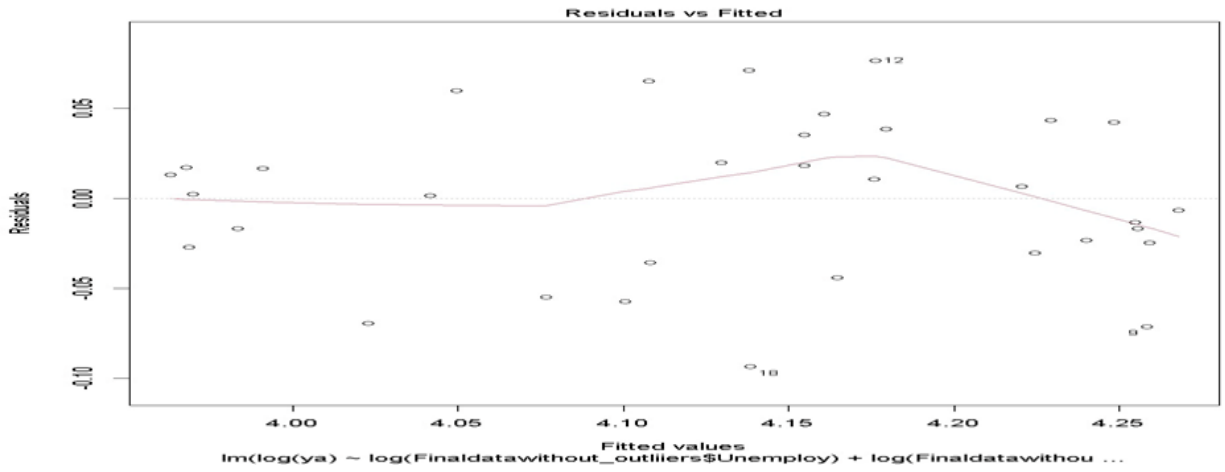

Figure 8 (Residual vs. Fitted)

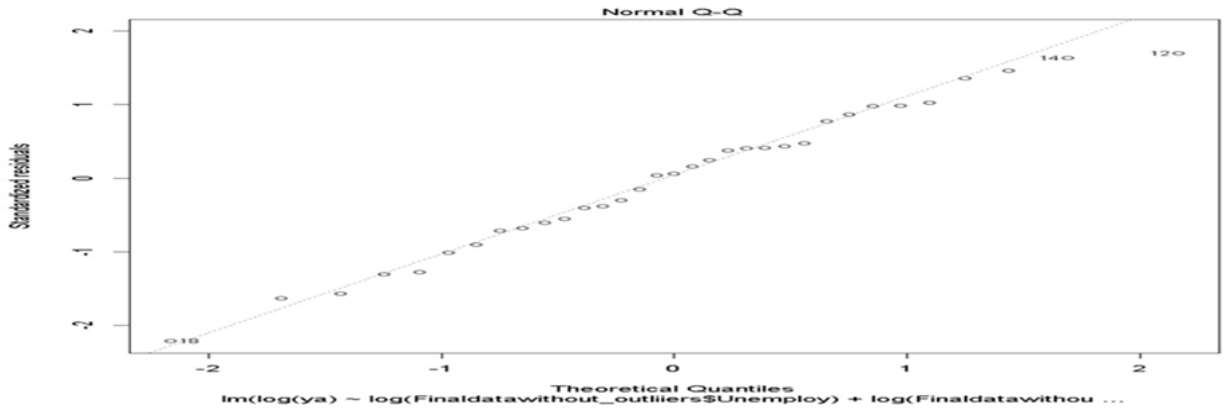

Figure 9 (QQ-plot)

- Model Selection

Now that we have satisfied all of our assumptions, we have to start filtering out of our model the insignificant variables and selecting the best model. Utilizing the stepwise regression and best subset method, we compared our models` AIC, BIC, and $\boldsymbol{R^2}$adj values, looking for the highest $\boldsymbol{R^2}$adj and the lowest AIC and BIC values. Keep in mind, that these aren't the only criteria we also have to consider the significance of the variables. As a result, we dropped the following variables: $Tpopulation$, $HEAlTH$, and $urbanratio$.
With an F-statistic = 45.48 on 4 and 27 DF and p-value = 1.287e-11, confirming that our model is significant
Hence , our optimal model is as follows :

$$\text{E}(\log lifex)) = 3.495853 - 0.021409\ log(Unemploy) + 0.042395 log(edu25) + 0.048215 log(GDPcap) + 0.05038\ log(Education)$$

With the following AIC , BIC , and R^2adj values:

$\boldsymbol{R^2}$**adj**	**AIC**	**sBIC**
0.8516	-106.9407	-196.2768

AIC: Akaike Information Criteria

sBIC: Sawa's Bayesian Information Criteria

Interpreting the Model

Our final model explains approximately 85% of all variations in life expectancy while satisfying all assumptions. A country's life expectancy, if unaffected by the rest of the variables, is approximately 33 years. Since we used the LOG transformation on both the Y and X variables, we can interpret the coefficients as the following. Ceteris paribus, a 1% increase in the unemployment rate is associated with a 0.021% decrease in life expectancy. Furthermore, a 1% increase in the population with at least some secondary education, ages 25 and older, would lead to an increase of 0.4% in life expectancy, holding other factors constant. Additionaly, a 1% increase in the GDP per capita would lead to an increase of 0.048% in life expectancy, holding other factors constant .Finally, a 1% increase in government expenditure on education is associated with a 0.05% increase in life expectancy , holding others factors constant.

Conclusion

In conclusion , this paper establishes a model that is able to predict Life expectancy through the unemployment rate , population with at least some secondary education ages 25 and older, GDP per capita , and government expenditure on education. We detected the effect of each of these variables on life expectancy . finally reaching our optimal model which is as follows;

$$log\ (lifeex) = 3.495853 - 0.021409\ log(Unemploy) + 0.042395log(edu25) + 0.048215log(GDPcap) + 0.05038\ log(Education + Ui$$