# A Panel Data Study of the Determinants of Life Expectancy Across Sub-Saharan African Countries

**Shritej Shrikant Chavan**[1*], **William Chase Guyton**[1], **Varun Golas**[1], **Anusha Chennapragada**[1], **Vedita K**[1]

[1] Group 4, Naveen Jindal School of Management, Masters student Data Science cohort, The University of Texas at Dallas, Richardson, Texas, United States

**Abstract**

Sub-Saharan African countries rank the lowest in life expectancy. For these developing countries, with their recent economic growth and rising GDP per capita, it's crucial to understand their key determinants of life expectancy. In our research, we attempt to understand the major socio-economic, demographics, and environmental factors on life expectancy in 19 Sub-Saharan African countries (see Appendix B) for the period 2005-2019. We established through panel data techniques the effect of government expenditure in the healthcare and education sector on life expectancy. We also quantified how much impact some social factors like corruption and the unemployment rate have on the average life expectancy. Our empirical findings can be leveraged by these counties for the effective allocation of funds and resources to elevate their low life expectancy, therein improving longevity and enhancing their economic productivity.

***Keywords***: *Life Expectancy; Panel Data; SSA; Fixed Effects; GDP*

## 1. Introduction

Life expectancy at birth indicates the number of years a newborn infant would live *if* prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.[1] It is a key metric for assessing population health. It is broader than infant mortality (*the death of children under the age of one*) and child mortality (*the death of children under the age of five*) since it encapsulates mortality along the entire life course.

Life expectancy has burgeoned since the advent of industrialization in the early 1900s and the world average has now more than doubled to 70 years (refer Figure 1) [2]. Yet, we still see inequality in life expectancy across and within countries. In 2020, Monaco sits at the top with the highest life expectancy at 86 years whereas some countries in the Sub-Saharan African (SSA) region, for example, the Central African Republic, have a life expectancy of 54.6 years [1]. That means that a person born in Monaco in 2020 is expected to live 32 years longer than if he/she was born in the Central African Republic, assuming that other factors remain constant.

The study by Acemoglu and Johnson demonstrated the relationship between increased life expectancy and improvement in economic growth (GDP per capita), controlling for country fixed effects [3]. In the table below, we have shown how life expectancy varies between high-income and low-income countries. However, further analysis is necessary for determining how the allocation of a country's wealth through certain investments in healthcare, education, environmental management, and also some socio-economic factors have an overall effect in determining average life expectancy.

[*]**Corresponding author.** *E-mail address: shritej24c@gmail.com*

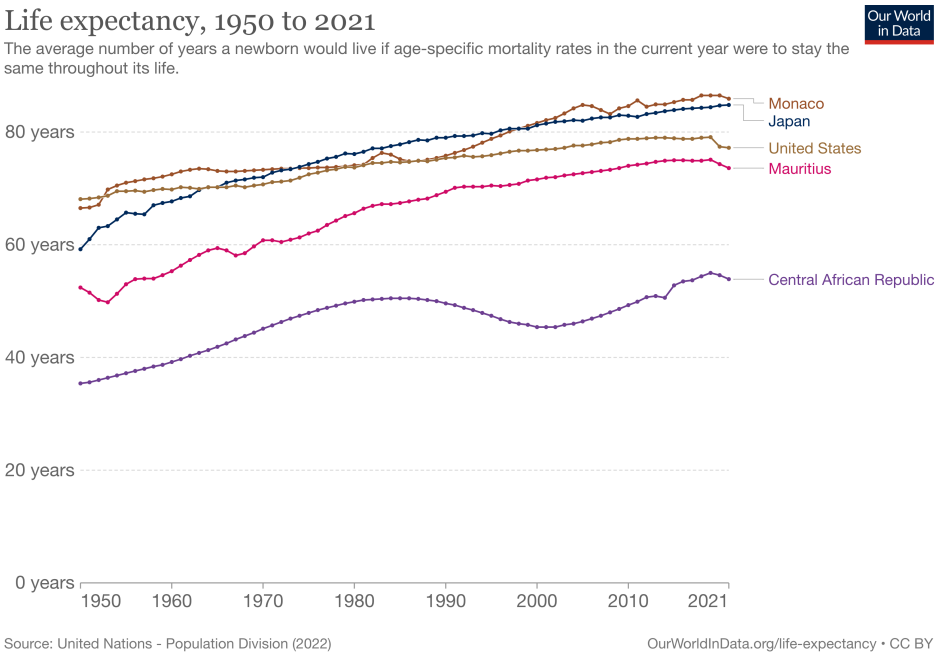


**Figure 1.** Panel Data on Life Expectancy for 5 countries from 1950 - 2021

The Sub-Saharan African region experiences the lowest life expectancy at birth compared to other regions over the past 3 decades. In the year 2019, the Middle East and North Africa had an average life expectancy of 74.27 years compared to the SSA region's average life expectancy of 61.6 years. From 2, we can see the variation of average life expectancy across various regions around the world for the last 3 decades. Moreover, SSA countries have consistently ranked as the lowest-earning countries in terms of GDP per capita. Therefore, there is a huge scope for improvement in life expectancy in SSA countries and hence our research focuses on the 40 Sub-Saharan African (SSA) countries with the lowest GDP per capita.

## 2. Literature Review

The existing literature on life expectancy is predominately focused on 2 perspectives. First, understanding the effect of life expectancy on economic productivity and second, the determinants of Life expectancy in general. Additionally, even though the evidence differs, the results are independent of the methodology employed. This research has leveraged methodologies such as quantile regression ([4]), vector auto-regression ([5]), and panel data techniques. [4] analyzed the relationship between economic productivity and life expectancy using vast data from 148 countries from the year 1970-2010. Further, they improved their analysis using a quantile regression approach, which showed that the effect of income on life expectancy was more profound in low-income countries compared to high-income countries.

Literature like [6] focused on the socio-economic determinants of life expectancy for 91 developing countries using multiple regression and probit regression. Socio-economic factors like per capita income, education, health expenditure, and urbanization turned out to be statistically insignificant, which proved that in developing countries these factors cannot always be prominent in determining life expectancy. Moreover, [7] approaches the problem and attempts to quantify the effect of Foreign Aid on life expectancy for 34 low-income countries over ten years (2005- 2015).

Regardless, none of the mentioned literature approaches the problem from a granular perspective. [8] attempts to quantify the effects of improved water supply, ability to meet basic needs, sanitation facility, and prevalence of pollution on life expectancy. This research implements a random effects model on panel data of 8 SSA countries from 2000-2015. The results in this literature were significant and the explanatory variables demonstrated as expected effect. Furthermore, [9] analyses what effect economic and educational development have on life expectancy in 38 African countries using the Least Squares Dummy Variables methodology.

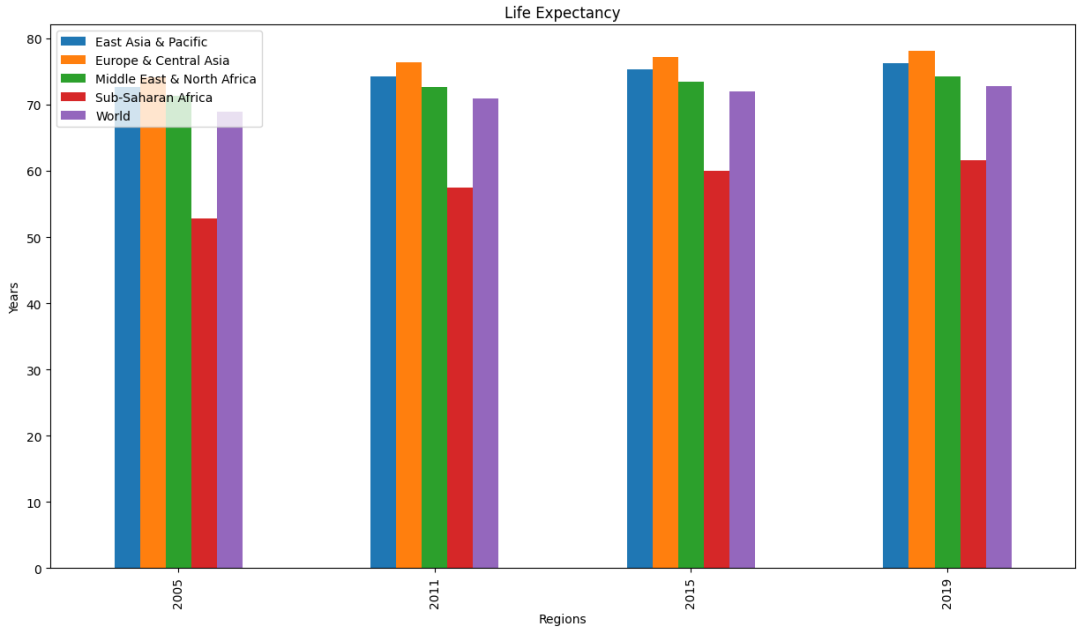


**Figure 2.** Life Expectancy in regions around the world over 3 decades

## 3. Research Questions

After reviewing the rich existing literature on Life Expectancy, we realized the lack of concrete research on understanding the impact of all-encompassing determinants that covers socio-economic and environmental factors for SSA countries using Panel Data techniques. Hence, we tried to address this inadequacy through our research. In this paper, we aim to have a better understanding of factors affecting life expectancy in the SSA region for an efficient policy-making process, and better allocation of funds and resources in addressing the prevalence of low life expectancy in Sub-Saharan Africa. To achieve that we attempt to answer the following questions in this research:

1. What's the Impact of Expenditure on Health and Education (% of GDP) on Life Expectancy?
2. How does the prevalence of undernourishment and communicable disease Affect Life Expectancy?
3. Do factors like corruption and unemployment rate impact life expectancy? If yes, quantify
4. Increase in CO2 emissions decrease life expectancy? Is it significant?

**Determinants of Life Expectancy considered:** Refer to Appendix A for definitions

1. Health Expenditure (% of GDP)
2. Education Expenditure (% of GDP)
3. Unemployment (% total labor force)
4. Corruption (CPIA rating)
5. Disability-Adjusted Life Years (DALYs) due to Communicable diseases
6. Prevalence of Undernourishment (% of population)
7. Carbon dioxide emissions (kiloton)

## 4. Data

We chose to focus only on Sub-Saharan African countries since they have low GDP per capita and low life expectancy countries. Hence, understanding what factors contribute majorly to their life expectancy is crucial for elevating their average in general. Our Data was gathered and collated predominately from these two sources: [1] and [2]. Out of 40+ SSA countries, due to a lot of missing values for the indicators

chosen, we focused on only 19 countries in that region. Also, the CPIA rating for corruption wasn't recorded until the year 2005. Moreover, we intend to exclude to account for any effect of COVID-19 on life expectancy. Hence, we restricted our analysis to data from 2005-2019. Furthermore, indicators like Sanitation, Water Supply, and Living standards were unavailable for a lot of countries. So our study considered 19 SSA countries (refer Appendix B) from 2005-2019.

The descriptive statistics as presented in Table 4 (see appendix) provide information about the means and standard deviations of variables used. From the table, we can see that the minimum life expectancy is 42.6 years and the maximum is 69.02 years. C (see appendix) shows the correlation matrix for the variables; some with high correlation and some with low correlation. For example, there is a high negative correlation between unemployment and disability years due to Communicable diseases, between undernourishment and CO2 emissions, and between health expenditure and communicable diseases.

## 5. Empirical Method

First and foremost, we tried to find the correlation of independent variables with Life Expectancy. Correlation between life expectancy and the log functional form of GDP for Mauritius from the year 1976-2019. Refer to the figure below 3

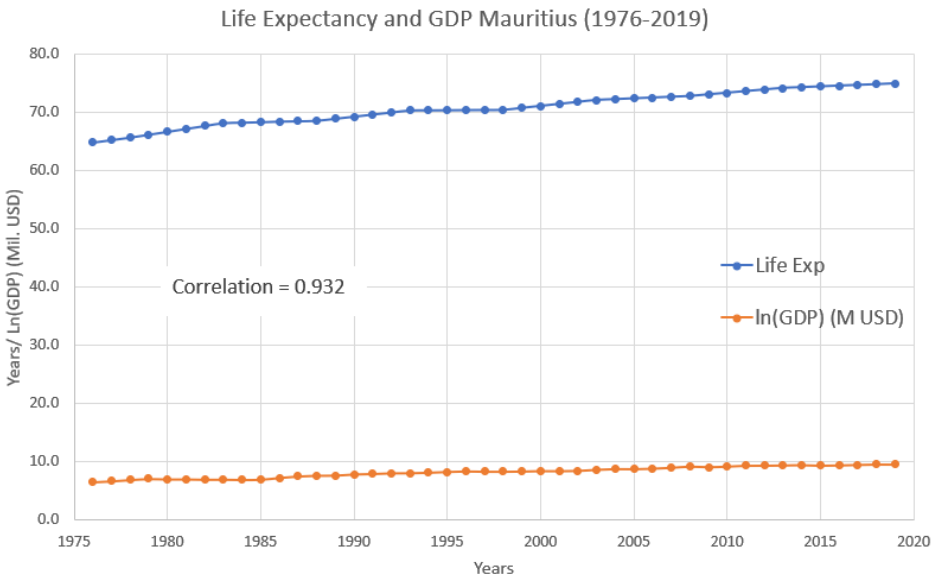


**Figure 3.** Spurious correlation between GDP and Life Expectancy of Mauritius

Though we found an almost perfect correlation between Log(GDP) and life expectancy, this analysis fails to explain the causation, and moreover, this can be a case of spurious correlation as we haven't differenced the time series. Hence, relying on a correlation between the independent and the dependent variable is not in our best interests.

### 5.1. Pooled OLS

For accurate, appropriate interpretations of the effects of some exogenous determinants, we considered fitting a Pooled OLS model on our Panel Data where N (countries) = 19 and T = 15. Our Pooled OLS is as follows:

$$\begin{aligned} \mathrm{LE}_{i,t} = \beta_0 + \beta_1\, unr_{i,t} + \beta_2\, log(co_2)_{i,t} + \beta_3\, health_{i,t} + \beta_4\, educ_{i,t} \\ +\beta_5\, unemp_{i,t}\beta_6\, crrptn_{i,t} + \beta_7\, log(comm)_{i,t} + U_{i,t} \end{aligned} \tag{1}$$

(2)

where,

1. unr - undernourished (% of population)
2. co2 - CO2 emissions in kiloton
3. health - Health expenditure (% of GDP)
4. unemp - unemployment (% of total labor force)
5. crrptn - (CPIA) corruption in the public sector rating (1=low to 6=high)
6. Comm - Disability-adjusted life-years due to communicable diseases per 100,000 individuals
7. educ - Education expenditure (% of GDP)

#### 5.1.1. Assumption for Pooled OLS

- Linear in Parameters - Linearity is ensured by taking functional forms of some variables
- No perfect collinearity - Pearson's Correlation Matrix reveals no perfect collinearity between the chosen indicators (see Appendix C)
- Homoscedasticity - Can be evaluated by various tests like the White test, Breusch-Pagan test, or just by plotting Residuals vs Fitted Values and observing their variance (see Appendix 4)
- No Serial Correlation among error term (residuals) - Durbin-Watson test, Ljung-Box test (see Appendix 5)
- Contemporaneous Exogeneity - Ensures that estimators are consistent

### 5.2. Individual Fixed Effects

To overcome the limitations of the Pooled OLS model we implemented a fixed effects model, where we introduced an alpha term that is fixed over time and is constant for each individual in the given Panel Data. Our Individual Fixed Effects model is as follows:

$$\begin{aligned} \text{LE}_{i,t} = \beta_0 + \beta_1\, unr_{i,t} + \beta_2\, log(co_2)_{i,t} + \beta_3\, health_{i,t} + \beta_4\, educ_{i,t} \\ +\beta_5\, unemp_{i,t}\beta_6\, crrptn_{i,t} + \beta_7\, log(comm)_{i,t} + \alpha_i + u_{i,t} \end{aligned} \tag{3}$$

(4)

where the variables remain the same as in the Pooled OLS model.

We know that fixed effects is always a consistent estimator but an inefficient estimator. Also, Fixed effects assume that the individual effects remain fixed over time.

### 5.3. Random Effects

In Random effects, the equation of the panel data model is the same as the fixed effect, however, in contrast to the fixed effects, it assumes that individual effects vary over time. Though random effects is not always consistent, if the covariance between alpha and independent variables is zero then Random Effects is efficient and consistent and hence it's preferred over Fixed Effects.

$$\begin{aligned} \text{LE}_{i,t} = \beta_0 + \beta_1\, unr_{i,t} + \beta_2\, log(co_2)_{i,t} + \beta_3\, health_{i,t} + \beta_4\, educ_{i,t} \\ +\beta_5\, unemp_{i,t}\beta_6\, crrptn_{i,t} + \beta_7\, log(comm)_{i,t} + \alpha_i + u_{i,t} \end{aligned} \tag{5}$$

(6)

## 6. Results and Discussion

### 6.1. Pooled OLS

Correlation Matrix (see Appendix C) suggests that there's no perfect correlation between explanatory variables, therefore our 2nd assumption for Pooled OLS is satisfied. Testing for Heteroskedasticity using White Test and Breusch-Pagan test, we rejected the null in both which meant that there's heteroskedasticity in our data 4. Hence, we fit the Pooled OLS model using Driscoll Kraay standard errors. Moreover, to test the serial correlation between residual errors, we got Durbin-Watson statistics = 1.64. Since it's slightly less than 2 it suggests slight positive autocorrelation between the errors. In addition to that, the Ljung-Box reveals no autocorrelation for all lags up to 10, which can be referred to herein Table 5. To conclude our Pooled OLS analysis we got results that are as follows in Table 1

**Table 1.** Pooled OLS Results

| | **Parameter** | **HAC Std. Error** | **P-value** |
|---|---|---|---|
| const | 79.539 | 2.921351 | 0.0000 |
| Undernourishment | 0.1048 | 0.025857 | 0.001 |
| log_CO2 | 9.9371 | 0.7220104 | 0.0000 |
| Health | -0.0245 | 0.091807 | 0.794 |
| Educ | 0.1296 | 0.1293494 | 0.333 |
| Unemployment | -0.4614 | 0.0369886 | 0.0000 |
| Corruption | 2.0094 | 0.2164577 | 0.0000 |
| log_Communicable | -9.1163 | 0.5077531 | 0.0000 |
| R-squared = 0.6638 | | | |
| Log-likelihood = -759.38 | | | |
| F-statistic (robust): 34.400 | | | |

We observe that our main 2 determinants; health and education came out to be statistically insignificant and the estimated effect of other variables is not as expected. Moreover, Pooled OLS has obvious limitations while working with Panel Data. Let's see the results of the Fixed effects model.

### 6.2. Individual Fixed Effects

Given there's heteroskedasticity in our data, we build our individual fixed effects model using Driscoll-Kraay HAC estimator and evaluated our results which are as follows Table 2

**Table 2.** Individual Fixed Effect Results

| | **Parameter** | **HAC Std Error** | **P-Values** |
|---|---|---|---|
| const | 289.23 | 16.685 | 0.0000 |
| Undernourishment | 0.0015 | 0.0247 | 0.952 |
| log_CO2 | 7.0380 | 0.5539 | 0.0000 |
| Health | 0.2408 | 0.0392 | 0.0000 |
| Educ | 0.2301 | 0.0705 | 0.0013 |
| Unemployment | -0.1714 | 0.0491 | 0.0006 |
| Corruption | -1.1019 | 0.2518 | 0.0000 |
| log_Communicable | -37.482 | 2.2560 | 0.0000 |
| R-squared: 0.9040 | | | |
| Log-likelihood: -383.37 | | | |
| F-statistic (robust): 2485.47 | | | |

As shown in the above table, we can see that except for undernourishment all the variables are statistically significant. Moreover, variables undernourishment, health, Educ have positive effects on life expectancy, and variables unemployment, corruption, and log_communicable have negative effects, which is as expected. Only log_CO2 has a positive effect which doesn't make sense. Most importantly, the R-squared of the F.E. is 0.904 which is remarkable and the determinants we considered do a good job of explaining the variation in Life Expectancy.

### 6.3. Random Effects

Similarly, given there's heteroskedasticity in our data, we build our random effects model using the HAC estimator and evaluated our results which are as follows Table 3

**Table 3.** Random Effects Model Results

| | **Parameter** | **HAC Std Error** | **P-Values** |
|---|---|---|---|
| const | 217.44 | 18.086 | 0.0000 |
| Undernourishment | 0.0087 | 0.0218 | 0.697 |
| log_CO2 | 9.3950 | 0.5744 | 0.0000 |
| Health | 0.2340 | 0.07882 | 0.010 |
| Educ | 0.2213 | 0.0619 | 0.003 |
| Unemployment | -0.3192 | 0.08931 | 0.003 |
| Corruption | -0.8805 | 0.4871 | 0.0092 |
| log_Communicable | -28.051 | 2.7267 | 0.0000 |
| R-squared: 0.8486 | | | |
| Log-likelihood: -450.18 | | | |
| F-statistic (robust): 221.73 | | | |

Observing the results of the Random Effects, we can see that they almost mirror the F.E results, varying slightly in the estimates. Here, the R-squared is pretty decent as well. After eliminating Pooled OLS as our choice of model, we evaluated for endogeneity to determine the best model between F.E and R.E. Using Hausman test to test for covariance between alpha and independent variables. We rejected the null hypothesis, hence we chose Fixed effects as our final econometrics model to determine the significant factors and their effect on Life expectancy. Results of the Hausman test can be seen in the Appendix section here 6

## 7. Conclusion

### 7.1. Final Econometrics Model

After careful evaluation, we found that the fixed effects model is a perfect fit for our panel data. So here's what our model looks like;

$$LE_{i,t} = 289.23 + 0.0015\,unr_{i,t} + 7.04\,log(co_2)_{i,t} + 0.24\,health_{i,t} + 0.23\,educ_{i,t} - 0.17\,unemp_{i,t} \\ -1.10\,crrptn_{i,t} - 37.48\,log(comm)_{i,t} + \alpha_i + u_{i,t} \tag{7}$$

**R-squared = 0.9040**

From Fixed Effects results - Except for Prevalence of Undernourishment all variables are significant (below results are concluded by keeping other factors constant)

1. 1% Increase in CO2 emissions increases Life Expectancy by 0.07 years (not expected)
2. 1% points increase in health and education expenditure (% of GDP) cause life expectancy to increase by 0.24 and 0.23 years respectively (expected)
3. 1% point increase in unemployment (% of total labor force) decreases life expectancy by .17 years (expected)
4. 1 point increase in the CPIA rating decreases the life expectancy by 1.1 years (expected)
5. 1% increase in disability-adjusted life year per 100,000 individuals, decreases L.E. by 0.37 years (expected)

Most of our independent variables showed expected significant behavior. An increase in health and education expenditure and a decrease in the unemployment rate and corruption should have a positive ceteris paribus effect on life expectancy. An increase in disability-adjusted life years specifically

due to communicable diseases puts the population in danger and lowers the life expectancy of that respective country. The opposite effect was observed in disability-adjusted life years specifically due to non-communicable diseases & injuries. CO2 output has a positive effect on life expectancy may be due to industrialization which we didn't control for. Increased industrialization might suggest overall progress and development of a nation hence better life expectancy. Our analysis shows what government can do to uplift the life expectancy in these low-income countries and how can it direct the efforts in order to achieve this goal.

## 8. Limitations

Further, we can control for endogeneity by introducing Instrument Variables and performing 2SLS. First, due to the lack of availability of time-series data on certain factors such as doctor-patient ratio, public vs private healthcare services, and efficiency we cannot control for these factors. We are not analyzing the causal impact of any policies that might have been in effect during the time period. Certain factors might have data quality issues since some countries have good reporting systems while others rely on estimation methods and data quality varies over countries. Completing this study with a details analysis of these determinants (probably at micro level analysis) will be important for the purposes of effective policy making.

## 9. Acknowledgement

We thank Professor Quanquan Liu for guiding us in understanding the various concepts of Econometrics and for being an instructor and mentor throughout this research.

## References

[1] W. B. (2017). World development indicators. [Online]. Available: https://databank.worldbank.org

[2] M. Roser, E. Ortiz-Ospina, and H. Ritchie, "Life expectancy," *Our World in Data*, 2013, https://ourworldindata.org/life-expectancy.

[3] A. Daron and S. Johnson, "Disease and development: The effect of life expectancy on economic growth," *Journal of Political Economy*, vol. 115, no. 6, 2007.

[4] M. Linden and D. Ray, "Aggregative bias-correcting approach to the health-income relationship: Life-expectancy and gross domestic product per capita in 148 countries," *Economic Modelling*, vol. 61, no. 1, pp. 126–136, 1970-2010.

[5] A. Bergh and T. Nilsson, "Good for living? on the relationship between globalization and life expectancy," *World Development*, vol. 38, no. 9, pp. 1191–1203, 2010.

[6] M. Kabir, "Determinants of life expectancy in developing countries," *Journal of Developing Areas*, vol. 41, no. 2, pp. 185–204, 2008.

[7] T. Rizzo, "A panel data study of the determinants of life expectancy in low income countries," *Honors Thesis*, 2019.

[8] P. O. Timothy, "Macroeconomic implications of low life expectancy in sub-saharan africa nations: A panel technique approach," *Journal of Developing Areas*, vol. 41, no. 2, pp. 185–204, 2008.

[9] M. GUISAN and P. EXPOSITO, "Life expectancy, education and development in african countries 1980-2014: Improvements and international comparisons," *Applied Econometrics and International Development*, vol. 16, no. 2, pp. 88–195, 2016.

# Appendices

## A. Definitions

1. Health Expenditure (% of GDP) - Level of current health expenditure expressed as a percentage of GDP. Estimates of current health expenditures include healthcare goods and services consumed

during each year. This indicator does not include capital health expenditures such as buildings, machinery, IT, and stocks of vaccines for emergencies or outbreaks

2. Education Expenditure (% of GDP) - General government expenditure on education (current, capital, and transfers) is expressed as a percentage of GDP. It includes expenditures funded by transfers from international sources to the government. General government usually refers to local, regional and central governments

3. Unemployment (% total labor force) - Unemployment refers to the % share of the labor force that is without work but available for and seeking employment

4. Corruption (CPIA rating) - Transparency, accountability, and corruption in the public sector assess the extent to which the executive can be held accountable for its use of funds and for the results of its actions by the electorate and by the legislature and judiciary, and the extent to which public employees within the executive are required to account for administrative decisions, use of resources, and results obtained.

5. Disability-Adjusted Life Years (DALYs) due to Communicable diseases - One DALY represents the loss of the equivalent of one year of full health. DALYs for a communicable disease or health condition is the sum of the years of life lost to due to premature mortality (YLLs) and the years lived with a disability (YLDs) due to prevalent cases of the disease in a population

6. Prevalence of Undernourishment (% of the population) - Prevalence of undernourishment is the percentage of the population whose habitual food consumption is insufficient to provide the dietary energy levels that are required to maintain a normally active and healthy life

7. Carbon dioxide emissions (kiloton) - Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during the consumption of solid, liquid, and gas fuels and gas flaring

## B. Countries (N)

[Benin, Burkina Faso, Central African Republic, Cameroon, Ethiopia, Ghana, Kenya, Lesotho, Madagascar, Mali, Mozambique, Mauritania, Malawi, Rwanda, Senegal, Sierra Leone, Chad, Togo, Tanzania]

## C. Correlation Matrix

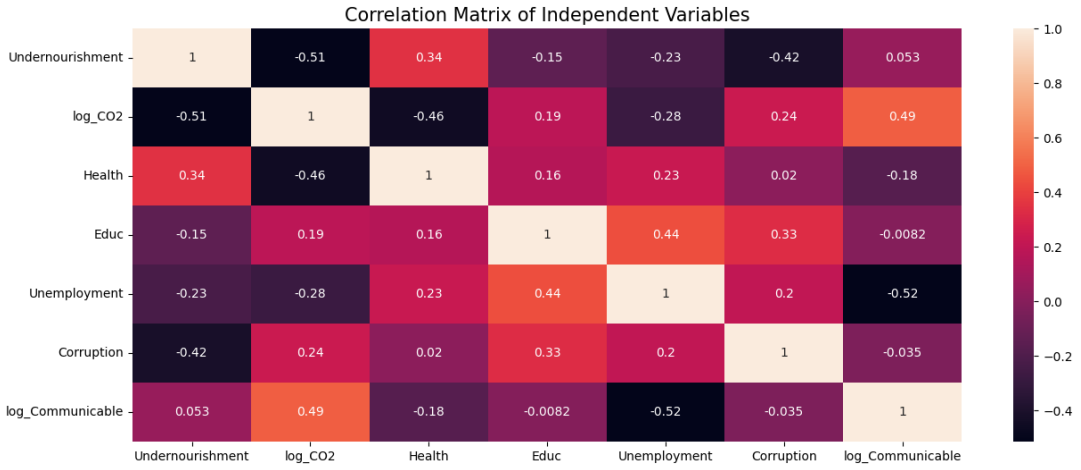


**Figure 4.** Life Expectancy in regions around the world over 3 decades

## D. Descriptive Statistics

| Variable | Unit | Country | Yrs | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|---|
| Undernourishment | percentage | 19 | 15 | 20.75 | 11.38 | 3.1 | 49.9 |
| CO2 | kiloton | 19 | 15 | 4724.74 | 4687.9 | 120 | 22280 |
| Health | percentage | 19 | 15 | 5.45 | 2.55 | 2.39 | 20.41 |
| Life-Expectancy | Years | 19 | 15 | 57.99 | 6 | 42.6 | 69.02 |
| Educ | percentage | 19 | 15 | 4.04 | 1.76 | 1.11 | 12.33 |
| Unemployment | percentage | 19 | 15 | 5.34 | 5.56 | 0.6 | 31.31 |
| Corruption | CPIA rating | 19 | 15 | 2.96 | 0.47 | 2 | 4 |
| Communicable | Years | 19 | 15 | 8163658.8 | 6727729.82 | 626351.4 | 40790843 |

## E. Heteroskedasticity tests

**Table 4.** Results of Tests for Heteroscedasticity

| Test | LM (p-value) | F p-value |
|---|---|---|
| White-Test | 1.46e-21 | 3.02e-37 |
| Breusch-Pagan-Test | 1.62e-09 | 1.73e-10 |

## F. Ljung Box test

**Table 5.** Ljung Box Tests Results

| lb_stat | lb_pvalue |
|---|---|
| 8.955250 | 0.002767 |
| 10.913670 | 0.004267 |
| 11.112064 | 0.011135 |
| 15.901298 | 0.003155 |
| 23.570328 | 0.000263 |
| 23.702189 | 0.000592 |
| 24.601226 | 0.000893 |
| 24.992508 | 0.001559 |
| 32.143108 | 0.000188 |
| 38.951385 | 0.000026 |

## G. Hausman test

The test evaluates the consistency of an estimator when compared to an alternative, less efficient estimator which is already known to be consistent

**Table 6.** Hausman test

| Chi-Squared | Degrees of freedom | p-value |
|---|---|---|
| 91.1162 | 8 | 2.759e-16 |

.