Data Analysis Final Project

1.Introduction

This study aims to explore how long people live and how much education they get in different countries. I am particularly interested in how these two things are connected to each other and how other factors might affect this relationship. I will look at whether the differences in education between men and women, the wealth of a country measured by GNI per person, how many people use the internet, and how much a country spends on public health in 2014 play a role in this.

Research Question

In this analysis, I focus on the link between how long people live (life expectancy) and how long they go to school (education years). I want to see if this link changes when we consider differences like whether men and women go to school for the same amount of time, how rich the country is, how many people use the internet, and how much money the country spends on health. By understanding these connections, I hope to provide useful information for making better policies and strategies.

2.Data Description

2.1. Data Overview

In this study, I am using data that tells us about different things related to people's lives in various countries. The main thing I am looking at is life expectancy – that's how long people, on average, live in these countries. This is our dependent variable, which means it's what I am trying to understand and explain through our analysis.

2.2. Key Variables

- Life Expectancy (Dependent Variable)
- Mean Years of Schooling (Independent Variable)
- Other Important Factors (Control Variables):
    - Gender Differences in Schooling
    - Gross National Income (GNI) Per Capita
    - Internet Usage
    - Public Health Expenditure (% of GDP) for 2014

2.3. Data Source and Collection

This analysis uses data from the "Human Development Reports" on Kaggle, originally sourced from the United Nations Development Programme. This dataset includes global human development indicators, focusing on life expectancy, education, and socio-economic factors. It is reliable and comprehensive, covering a wide range of countries. Available under the CC0: Public Domain license, it allows unrestricted use for any purpose.

	Life_expectancy	Years_of_schooling	GNI_per_capita	Years_of_schooling_Female	Years_of_schooling_Male	Internet_users	Public_health_expenditure
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	71.454491	8.419162	17721.562874	8.072455	8.811976	47.935329	4.023952
std	8.400607	3.140887	19156.479671	3.469951	2.862403	28.623234	2.263200
min	48.900000	1.400000	587.000000	1.000000	2.000000	2.200000	0.800000
25%	65.500000	6.100000	3842.500000	4.950000	6.450000	21.000000	2.350000
50%	74.200000	8.800000	10789.000000	8.500000	8.900000	48.900000	3.500000
75%	77.250000	11.200000	24714.000000	10.950000	11.400000	72.250000	5.200000
max	83.700000	13.400000	129916.000000	13.400000	13.600000	98.200000	10.800000

ln_Years_of_schooling	ln_Years_of_schooling_Female	ln_Years_of_schooling_Male	ln_GNI_per_capita
167.000000	167.000000	167.000000	167.000000
2.038931	1.952372	2.111652	9.182883
0.467503	0.589485	0.383684	1.205373
0.336472	0.000000	0.693147	6.375025
1.808289	1.599337	1.864050	8.253878
2.174752	2.140066	2.186051	9.286282
2.415914	2.393329	2.433613	10.115118
2.595255	2.595255	2.610070	11.774643

3. Model

To understand the various factors influencing life expectancy, I employed five different linear regression models. Each model explores different aspects:

1. Life Expectancy vs log of Mean Years of Schooling and log of GNI per Capita.
2. Life Expectancy vs log of Mean Years of Schooling (Female).
3. Life Expectancy vs log of Mean Years of Schooling (Male).
4. Life Expectancy vs log of Mean Years of Schooling and Internet Usage.
5. Life Expectancy vs Mean Years of Schooling and Public Health Expenditure.
6. Life Expectancy vs log of Mean Years of Schooling

3.1. Why These Models?

Comprehensive Analysis: By using different models, we can compare the influence of various factors on life expectancy.
Gender-Specific Insights: Models 2 and 3 provide insights into how education impacts life expectancy differently for males and females.
Economic and Technological Factors: Models 1, 4, and 5 help in understanding the role of economic status, technology (internet usage), and health investment on life expectancy.

3.2. Result and Interpretation

The following table presents the results of the regression analysis, offering detailed explanations for each finding. The models are simple linear regressions with HC1 covariance type to ensure robustness. The detailed tables for each regression analysis can be found in Appendix 1.

3.2.1. Life Expectancy vs log of Mean Years of Schooling and log of GNI per Capita.

Based on the results of Model 1, where the coefficient for log(Years_of_schooling) is 4.56 and the coefficient for log(GNI_per_capita) is 4.22, and both are statistically significant at the 1% level, the interpretations are as follows:

1. Impact of Education Years (Coefficient of log(Years_of_schooling) = 4.56):

A one-unit increase in the natural log of the average years of schooling is associated with an approximate increase of 4.56 years in life expectancy.
The statistical significance of this coefficient suggests that the duration of education significantly influences life expectancy.

2. Impact of National Income (Coefficient of log(GNI_per_capita) = 4.22):

A one-unit increase in the natural log of Gross National Income per capita is associated with an approximate increase of 4.22 years in life expectancy.

The statistical significance of this coefficient indicates that the level of national income significantly affects life expectancy.

3.2.2. Life Expectancy vs log of Mean Years of Schooling (Female).
For Model 2, the coefficient for log(Years_of_schooling_Female) is 10.42 and is statistically significant at the 1% level, the interpretation is as follows:

Impact of Female Education Years (Coefficient of log(Years_of_schooling_Female) = 10.42):
A one-unit increase in the natural log of the average years of schooling for females is associated with an approximate increase of 10.42 years in life expectancy.
The coefficient being significantly large and statistically significant at the 1% level strongly suggests that the duration of female education has a substantial and positive impact on life expectancy.
This significant effect implies that improvements in female education could be particularly effective in increasing life expectancy.

3.2.3. Life Expectancy vs log of Mean Years of Schooling (Male).
Based on the results of Model 3, where the coefficient for log(Years_of_schooling_Male) is 15.48 and is statistically significant at the 1% level, the interpretation is as follows:

Impact of Male Education Years (Coefficient of log(Years_of_schooling_Male) = 15.48):
A one-unit increase in the natural log of the average years of schooling for males is associated with an approximate increase of 15.48 years in life expectancy.
The large magnitude of the coefficient, along with its statistical significance at the 1% level, indicates that the duration of male education has a profound and positive impact on life expectancy.
This substantial effect suggests that male education plays a critical role in determining life expectancy, possibly even more so than other factors.

3.2.4. Life Expectancy vs log of Mean Years of Schooling and Internet Usage.
Based on the results of Model 4, where the coefficient for internet usage is 0.239 and is statistically significant at the 1% level, the interpretation is as follows:

Interpretation of the Result
Impact of Internet Usage (Coefficient of Internet Usage = 0.239):
A one-unit increase in internet usage (presumably measured as a percentage of the population) is associated with an approximate increase of 0.239 years (about 2.39 months) in life expectancy.
The statistical significance of this coefficient at the 1% level indicates that internet usage has a positive and meaningful impact on life expectancy, although the magnitude of this impact is relatively small compared to factors like education.
This effect might reflect the broader implications of technological access and connectivity on health and well-being.

3.2.5. Life Expectancy vs Mean Years of Schooling and Public Health Expenditure.

Based on the results of Model 4, where the coefficient for public health expenditure is 1.97 and is statistically significant at the 1% level, the interpretation is as follows:

Impact of Public Health Expenditure (Coefficient of Public Health Expenditure = 1.97):
A one-unit increase in public health expenditure (presumably measured as a percentage of GDP) is associated with an approximate increase of 1.97 years in life expectancy.
The coefficient being large and statistically significant at the 1% level suggests that investment in public health has a substantial and positive impact on life expectancy.
This indicates that increased spending on health care and related public health services significantly contributes to improving the average lifespan of a population.

3.2.6. Life Expectancy vs log of Mean Years of Schooling

Based on the results of Model 6, where the coefficient for log(Years_of_schooling) is 13.11 and is statistically significant at the 1% level, the interpretation is as follows:

Impact of Education Years (Coefficient of log(Years_of_schooling) = 13.11):
A one-unit increase in the natural log of the average years of schooling is associated with an approximate increase of 13.11 years in life expectancy.
The large magnitude of the coefficient, combined with its statistical significance at the 1% level, indicates that the duration of education has a profound and positive impact on life expectancy.
This substantial effect suggests that education plays a critical role in determining life expectancy, highlighting the importance of educational policies and investments in enhancing public health and longevity.

4. Conclusion

In conclusion, Model 1, with the highest R-squared value, significantly impacts our understanding of life expectancy. It highlights the profound influence of education and economic factors, particularly the logarithmic values of Mean Years of Schooling and GNI per Capita. This model's findings suggest that enhancing educational opportunities and economic growth are crucial for improving public health outcomes.

The results emphasize the need for integrated policy approaches that focus on both education and economic development. Despite the strong correlations identified, it's important to remember that these do not imply causation. Future studies are needed to further explore these relationships. This analysis serves as a valuable guide for policymakers aiming to boost life expectancy through multifaceted strategies.

However, the analysis primarily faces the challenge of distinguishing correlation from causation, a common limitation in regression models. This means that while relationships between variables like education, economic status, and life expectancy are evident, the directionality and causative factors behind these associations remain uncertain. Additionally, the potential for omitted variable bias exists, where excluding key variables could skew the results. The linear nature of the models may also oversimplify the complex, potentially non-linear relationships inherent in socioeconomic data.

Appendix.

						Dependent variable: Life_expectancy
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	23.386***	51.118***	38.762***	59.979***	63.532***	44.735***
	(2.741)	(1.721)	(2.807)	(0.809)	(1.144)	(2.264)
ln_Years_of_schooling	4.564***					13.105***
	(1.281)					(1.029)
ln_GNI_per_capita	4.221***					
	(0.469)					
ln_Years_of_schooling_Female		10.416***				
		(0.800)				
ln_Years_of_schooling_Male			15.482***			
			(1.250)			
Internet_users				0.239***		
				(0.012)		
Public_health_expenditure					1.969***	
					(0.252)	
Observations	167	167	167	167	167	167
R^2	0.673	0.534	0.500	0.665	0.281	0.532
Adjusted R^2	0.669	0.531	0.497	0.663	0.277	0.529
Residual Std. Error	4.834 (df=164)	5.750 (df=165)	5.958 (df=165)	4.874 (df=165)	7.143 (df=165)	5.765 (df=165)
F Statistic	217.722*** (df=2; 164)	169.542*** (df=1; 165)	153.422*** (df=1; 165)	404.383*** (df=1; 165)	61.124*** (df=1; 165)	162.052*** (df=1; 165)
Note:						*p<0.1; **p<0.05; ***p<0.01