Mahrukh Jaura
Introduction to Econometrics
Professor Foster
December 10th, 2021

Observing the Impact of Health Expenditure on Life Expectancy, Internationally

Introduction

Primarily depicting the average duration of time that an individual is expected to live until, life expectancy also serves as a general representation of population health. Following the early 19th century, life expectancy, while gradually developing internationally, significantly escalated in the early industrialized countries, thus, contributing to a health-distribution inequality. A result of the substantial health improvements achieved by the affluent regions, countries in abundance of wealth fostered and maintained good health whereas impoverished countries experienced health of a persistently poor state. To illustrate, as of 2021, and for the preceding several years, the Central African Republic holds the lowest life expectancy of 54.36 whereas, Japan upholds a life expectancy of 85.03, illuminating the inequality of life expectancy.

While the global inequality in health has been continuously stated to influence the life expectancy accordingly, this paper seeks to observe the significance of health expenditure, as well as recognize additional factors that impact life expectancy, at an international level. Specifically, through ordinary least squares (OLS) regressions, single and multiple, this paper strives to examine the relationship that variables of considerable relevance, may have on life expectancy. Variables such as, GDP per capita, population, poverty headcount, government expenditure on education, primary school enrollment, income inequality and the number of physicians per 1,000 individuals. Rationalizing the input of education related variables, studies with varying datasets have observed an association between educational attainment and adult mortality, where individuals with higher educational attainments can be observed to have lower mortality rates. Therefore, each of the independent variables selected are expected to depict varying levels of significance in relation to life expectancy, the dependent variable, through the OLS regressions. This paper will also demonstrate adherence to the Gauss Markov Assumptions since violation of the assumptions may reduce the validity of results outputted by the OLS models.

Preliminary

Before advancing to the body of this paper, it is of interest to detail over the initial attempt of determining significant variables in regard to life expectancy. Originally, the focus of this paper pertained to New York City and data consisted of a five-year period of the 60 community districts, which compose the city. I attempted to establish a relationship between income and life expectancy since one of the first studies observing life expectancy, linked a relationship to income inequality while identifying factors related to small area variation within the association. After evaluating factors associated with differences in life expectancy and estimating said expectancy by household income percentile and geographic area, the study found that higher income was associated with higher life expectancy throughout the income distribution.

Using data from Citizens' Committee for Children (CCC) of New York, I obtained data on the income diversity ratio, poverty level, monthly rent, unemployment rate, educational attainment, graduation rate, household income, population living in concentrated poverty, public assistance, SNAP, uninsured individuals and unemployment by teen and youth. Data for each variable was individually downloaded and subsetted to years 2014-2018 (to match the timeframe of data available on life expectancy for the community districts).

The following models depict the relationship observed between life expectancy and the explanatory variables chosen.

Single Regression Model:

$$\text{LifeExpect1} = \beta 0 + \beta 1(\text{IDR2\$Data}) + u$$

$$\text{LifeExpect1} = 81.46033 + 0.04589(\text{IDR2\$Data})$$

Results:

Residual standard error: 2.815 on 298 degrees of freedom

Multiple R-squared: 0.0006229, Adjusted R-squared: -0.002731

F-statistic: 0.1858 on 1 and 298 DF, p-value: 0.6668

In the single regression model, the dependent variable is life expectancy, and the independent variable is the income diversity ratio. The adjusted R-squared reflects the fit of the model where a higher value generally indicates a better fit. This model has an adjusted R-squared value of (-)0.002731 which demonstrates that the model selected does not follow the trend of the data,

ergo, leading to a worse fit than a horizontal line. The negative adjusted R-squared value can also be the result of constraints on either the intercept or the slope of the linear regression line. Hence, the p-level for the income diversity ratio is not statistically significant. Observing the absence of significant results, I introduced 28 additional independent variables in the expectation of attaining significant relationships.

Multiple Regression Model:

$$\text{LifeExpect1} = \beta 0 + \beta 1(\text{IDR2\$Data}) + \beta 2(\text{Poverty2\$Data}) + \beta 3(\text{MRent\_500\$Data}) + \beta 4(\text{MRent\_999\$Data}) + \beta 5(\text{MRent\_1k\$Data}) + \beta 6(\text{MRent\_1500\$Data}) + \beta 7(\text{MRent\_2k\$Data}) + \beta 8(\text{UnER1\$Data}) + \beta 9(\text{Educ\_nohs\$Data}) + \beta 10(\text{Educ\_SC\$Data}) + \beta 11(\text{Educ\_AD\$Data}) + \beta 12(\text{Educ\_BD\$Data}) + \beta 13(\text{ConPov1\$Data}) + \beta 14(\text{Graduation\_Rate\$Data}) + \beta 15(\text{HH\_Inc\_100k\$Data}) + \beta 16(\text{HH\_Inc\_15k\$Data}) + \beta 17(\text{HH\_Inc\_200k\$Data}) + \beta 18(\text{HH\_Inc\_25k\$Data}) + \beta 19(\text{HH\_Inc\_50k\$Data}) + \beta 20(\text{HH\_Inc\_75k\$Data}) + \beta 21(\text{HH\_Inc\_Under15k\$Data}) + \beta 22(\text{PPLCV1\$Data}) + \beta 23(\text{Pub\_Assist1\$Data}) + \beta 24(\text{Snap\_HH\$Data}) + \beta 25(\text{Uninsured\_Adults\$Data}) + \beta 26(\text{Uninsured\_All\$Data}) + \beta 27(\text{Uninsured\_Child\$Data}) + \beta 28(\text{Unemploy\_Teen1\$Data}) + \beta 29(\text{Unemploy\_Youth1\$Data}) + u$$

$$\text{LifeExpect1} = 8.146e{+}01 + -1.122e{-}01(\text{IDR2\$Data}) + -1.723e{-}07(\text{Poverty2\$Data}) + -1.105e{-}05(\text{MRent\_500\$Data}) + 7.587e{-}05(\text{MRent\_999\$Data}) + -5.544e{-}05(\text{MRent\_1k\$Data}) + -2.382e{-}05(\text{MRent\_1500\$Data}) + 9.099e{-}05\ (\text{MRent\_2k\$Data}) + -1.110e{-}01(\text{UnER1\$Data}) + -2.388e{-}05(\text{Educ\_nohs\$Data}) + 3.128e{-}05(\text{Educ\_SC\$Data}) + -2.024e{-}06(\text{Educ\_AD\$Data}) + -1.896e{-}06(\text{Educ\_BD\$Data}) + --1.845e{-}04(\text{ConPov1\$Data}) + -7.831e{-}01(\text{Graduation\_Rate\$Data}) + 2.395e{-}06(\text{HH\_Inc\_100k\$Data}) + -4.716e{-}05(\text{HH\_Inc\_15k\$Data}) + 3.741e{-}05(\text{HH\_Inc\_200k\$Data}) + 3.293e{-}04(\text{HH\_Inc\_25k\$Data}) + -3.078e{-}04(\text{HH\_Inc\_50k\$Data}) + 1.144e{-}04(\text{HH\_Inc\_75k\$Data}) + 1.024e{-}04(\text{HH\_Inc\_Under15k\$Data}) + 3.182e{-}04(\text{PPLCV1\$Data}) + -2.053e{+}00(\text{Pub\_Assist1\$Data}) + -1.715e{-}05(\text{Snap\_HH\$Data}) + 7.866e{+}01(\text{Uninsured\_Adults\$Data}) + -1.065e{-}02(\text{Uninsured\_All\$Data}) + 2.465e{+}01(\text{Uninsured\_Child\$Data}) + -2.052e{+}00(\text{Unemploy\_Teen1\$Data}) + -5.188e{+}00(\text{Unemploy\_Youth1\$Data})$$

Results:

Residual standard error: 2.642 on 270 degrees of freedom

Multiple R-squared: 0.2024, Adjusted R-squared: 0.1167

F-statistic: 2.363 on 29 and 270 DF, p-value: 0.00019

In this model, the dependent variable is life expectancy, and the independent variables are the income diversity ratio, poverty level, monthly rent, unemployment rate, educational attainment, graduation rate, household income, population living in concentrated poverty, public assistance, SNAP, uninsured individuals and unemployment by teen and youth. The adjusted R-squared value is 0.1167 which indicates that majority of the variables are not a good fit for the model. The variables with a statistically significant p-value are MRent_2k$Data (monthly rent of $2,000 or more), UnER1$Data, (the unemployment rate), HH_Inc_50k$Data, (household income of $50k - $74,999), and Uninsured_All$Data (individuals uninsured). The regression output can be observed in Table 1.

To interpret the significance of the variables, for every increase in the number of people paying above 2k rent monthly, everything else held constant, the life expectancy can be expected to change by .00009099. For every increase in the unemployment rate, everything else held constant, the life expectancy can be expected to decrease by .11101. For every increase in the number of people with a household income of $50k - $74,999, with everything else held constant, the life expectancy can be expected to decrease by .0003078. Finally, for every increase in the number of people uninsured, the life expectancy can be expected to decrease by .001065. The surprising output of these results is the negative coefficient estimate for HH_Inc_50k$Data, a coefficient estimate, if significant, I had expected to be positive.

Table 1: OLS Regression Models Output

Independent Variables	Model 1 - Single Regression	Model 2 -Multiple Regression
(Intercept)	81.46033 (***)	84.0601 (***)
IDR2$Data	0.04589	-1.122e-01
Poverty2$Data	--	-1.723e-07
MRent_500$Data	--	-1.105e-05
MRent_999$Data	--	7.587e-05
MRent_1k$Data	--	-5.544e-05
MRent_1500$Data	--	-2.382e-05
MRent_2k$Data	--	9.099e-05 (**)
UnER1$Data	--	-1.110e-01 (.)
Educ_nohs$Data	--	-2.388e-05
Educ_SC$Data	--	3.128e-05
Educ_AD$Data	--	-2.024e-06
Educ_BD$Data	--	-1.896e-06

ConPov1$Data	--	-1.845e-04
Graduation_Rate$Data	--	-7.831e-01
HH_Inc_100k$Data	--	2.395e-06
HH_Inc_15k$Data	--	-4.716e-05
HH_Inc_200k$Data	--	3.741e-05
HH_Inc_25k$Data	--	3.293e-04
HH_Inc_50k$Data	--	-3.078e-04 (*)
HH_Inc_75k$Data	--	1.144e-04
HH_Inc_Under15k$Data	--	1.024e-04
PPLCV1$Data	--	3.182e-04
Pub_Assist1$Data	--	-2.053e+00
Snap_HH$Data	--	-1.715e-05
Uninsured_Adults$Data	--	7.866e+01
Uninsured_All$Data	--	-1.065e-02 (.)
Uninsured_Child$Data	--	2.465e+01
Unemploy_Teen1$Data	--	-2.052e+00
Unemploy_Youth1$Data	--	-5.188e+00

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

After speculating over the results, I realized that my lack of significant results may arise from my lack of relevant variables. Following this, I changed my data source to the World Bank Data, where I would have access to datasets with variables that I hypothesized were more germane to life expectancy. For this reason, I ceased operations on these models and ultimately, started over. The process and results are outlined as follow.

Literature Review

An international study (Kim, et al 2017) analyzed the relationship between public health expenditure and national outcomes amongst developed countries from 1973 through 2000, since the exact nature of the relationship between health expenditure and health outcome remained unclear and undefined. To explore the impact of public health expenditure on the public health outcome, data for 17 developed countries was integrated using OECD statistics, the World Health Organization database, & a Quality of Government Study dataset. With public health outcomes serving as the dependent variable, public health expenditure as the independent variable, infant mortality rate & life expectancy at birth were utilized as indicators. Control variables included GDP per capita, the Gini coefficient, unemployment rates, and the rate of the aging population. A regression found a statistically significant association between government

health expenditure and public health outcomes. Specifically, there was a negative relationship between government health expenditure and infant mortality rate, and a positive relationship between government health expenditure and life expectancy at birth. Concluding that an increased government spending on goods and services pertaining to medicine & health will aid in establishing better overall health.

A second study investigating the relationship between health care expenditure and life expectancy, used Nigeria (Isaac, et al, 2017) as the focal point of their study. Examining empirical evidence of the impact of public health expenditure on life expectancy for years, 1981 through 2014, and employing the recent bounds testing cointegration approach, the study found that there is a significant long-run relationship between life expectancy, public health expenditure and primary school enrollment, in Nigeria. Specifically, primary school enrollment introduced awareness regarding health and health-related issues and while found to be insignificant in the short run, was significant in the long run, in terms of increasing the life expectancy at birth.

Data

For this project, I decided to analyze life expectancy and health expenditure, for which I used data from 186 countries. Life expectancy, the dependent variable, is the average period that a person may expect to live and is often used as a measure of a population's overall health. The primary explanatory variable is health expenditure, which includes all expenditures for the provision of health services and aid designated for health-related activities. I chose this variable as my primary explanatory variable after reading about the importance it plays on life expectancy and the prior association found in the studies mentioned above. For this single regression, I would expect a high level of significance to result between life expectancy and health expenditure. The following table depicts summary statistics for all variables used.

Table 2: Summary Statistics

Variables	Mean	SD	Min	Max
Life Expectancy	72.19859	7.623334	52.24	84.10
Health Expend.	6.594637	2.6887469	2.27	17.004
GDP per Cap.	1.977667	2.790099	-9.1677	7.9416
Log(total pop.)	16.14828	2.12692	11.47	21.56
Pov. at $1.9	4.13243243	9.6848576	0.00	49.4
Educ. Expend.	14.6616671	4.529254	5.117	30.151
Prim. School En.	103.453152	10.6004213	69.6	145.49
Gini Coeff.	36.03188	7.512489	24.20	56.30
Physicians_1k	2.065206	1.627251	0.0008	8.2950

Method

The World Bank Data offers access to resources such as an open data catalog, databank, microdata library, world development indicators, open finances, open data toolkit, amongst others, for the purpose of providing quality statistical data for use. The data comes from the statistical systems of member countries and the quality of the data depends on the performance of each country's national systems. By working alongside the international statistical community, which include the United Nations agencies, Organization for Economic Co-Operation and Development, the International Monetary Fund, regional development banks, and donors, the World Bank can establish data exchange processes/methods thus, compiling international data sets by assembling and analyzing online data.

From the World Bank, I downloaded data on life expectancy, government expenditure on health, GDP per capita, population, poverty headcount ratio, government expenditure on education, primary school enrollment, the Gini coefficient, and the number of physicians per 1,000 people. The variables were chosen based on their prior linkage to life expectancy, demonstrated in several studies, and for their expected significance. After downloading each dataset, a painstaking task was endured of eliminating countries that omitted data for each year and additional miscellaneous items included. This resulted in reducing the number of observations

from 266 to 186. A complete list of the final countries included in the study can be found in Appendix I.

After downloading the datasets and omitting unrelated rows interfering with the data, I opened each data file in RStudio, where I was able to perform regressions on variables from differing files by stating the data set followed by a dollar sign, which allowed access to one variable (column) within the data, for each variable in the regressions. I was flabbergasted at the discovery of being able to conduct such a regression. For majority of the datasets used, information was available until 2019. Not all of the datasets included information for 2019 or 2018 which is why I chose to utilize data from 2017. For the purpose of this project, I used Ordinary Least Squares regression (OLS); a technique used for estimating coefficients of linear regression equations, which depict the relationship between one or multiple independent variables and a dependent variable. The equation for the OLS regression is:

$$Y = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n + u$$

Where Y, is the response variable or the predicted outcome value in a linear function of the regressors, $\beta 0$ represents the intercept, $\beta_n x_n$ depict regression coefficients for the corresponding independent variables.

Before continuing to the regression models, I will go over the assumptions made by OLS, otherwise known as the Gauss-Markov conditions, since violation of the assumptions may reduce the validity of the results produced by the model.

Gauss-Markov Assumptions:

1. **Linearity**: This assumption states that the relationship between the dependent and the independent variable should be linear. An assumption that can be confirmed through scatter plots, as depicted in Figures 1, 2, 3 and 4.
2. **No Multicollinearity**: Multicollinearity occurs when the independent variables are highly correlated, that is, the independent variables depend on one another. Therefore, a multiple linear regression assumes that there is no multicollinearity within the data. This assumption is confirmed through observation of correlation of regressors and can be seen in Table 4.
3. **Random Sampling**: The data should meet the condition of being randomly sampled from the population and given that the data used in this project is from the World Bank, this assumption is fulfilled.

4. **Zero Conditional Mean**: The error has an expected value of zero, given any values of the independent variables. This can be observed through a plot of residuals as seen in Figure 5.
5. **Homoskedasticity**: The error has the same variance given any value of the explanatory variable. This can also be observed through observation of Figure 5.

Since the data meets the assumptions detailed above, the following sections detail the application of the regression models serviced in this project.

Model 1 - Single Regression Model

$$LIFE\_EXPECTANCY\_5\_years\$`2017` = \beta 0 + \beta 1(HEALTH\_EXPENDITURE\$`2017`) + u$$

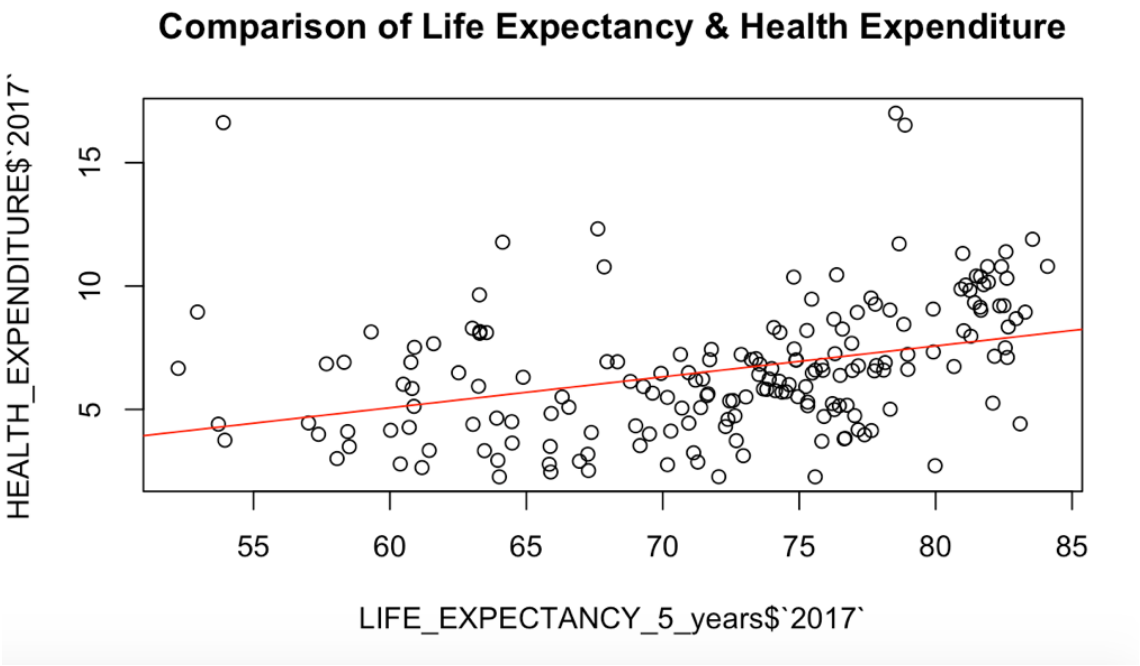
LIFE_EXPECTANCY_5_years$`2017` = 65.5062 + 1.0112 (HEALTH_EXPENDITURE$`2017`)

Results

Residual standard error: 7.156 on 183 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.1267, Adjusted R-squared: 0.122
F-statistic: 26.56 on 1 and 183 DF, p-value: 6.584e-07

This model observes a p-value of 6.584e-07 which is statistically significant at 0.001, indicating strong evidence against the null hypothesis, that there is no relationship between the dependent and the independent variable, i.e., there is no relationship between life expectancy and health expenditures. Since this regression model contains an independent variable that is statistically significant at 0.001, I assumed that the adjusted R-squared value would be closer to 1 than 0, however, that adjusted R-squared value reads 0.122. This reveals that while the independent variable is correlated with the dependent variable, it does not explain majority of the variability within the dependent variable. Figure 1 depicts the relationship between life expectancy and health expenditure, based on the data used for the models. To interpret, the coefficient in this single regression model is the change in the dependent variable due to the change in one unit of the independent variable. To elaborate, each increase in the level of current health expenditure, expressed as a percentage of GDP, will increase the life expectancy by 1.0112.

Figure 1: A Comparison of Life Expectancy & Health Expenditure Globally for 2017



The R-squared value is significant because it measures the scatter of the data around the regression line and higher variability around the regression line will produce a lower R-squared value. Figure 1, more specifically, illustrates that as the value of health expenditure increases, the value for life expectancy also increases, however, the spread of the data points suggests that health expenditure alone, does not explain much of the variability in life expectancy.

The second model to be presented is a multiple regression model where the dependent variable is life expectancy and the independent variables are health expenditure, GDP per capita, total population, poverty headcount, education expenditure, primary school enrollment, the Gini coefficient, and the number of physicians per 1,000 people. I appointed these datasets to perform as the independent variables due to the reference made to their link on life expectancy, within several studies. Data was collected from the World Bank and a more detailed description of each of the variables can be found in Table 3.

Table 3: Description of Variables

Variable	Description	Units
Life expectancy	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.	Years
Health expenditure	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.	Percentage
GDP per capita	Annual percentage growth rate of GDP per capita based on constant local currency. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.	Percentage
Log(total population)	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.	Number
Poverty headcount	Poverty headcount ratio at $1.90 a day is the percentage of the population living on less than $1.90 a day at 2011 international prices.	Percent
Education expenditure	General government expenditure on education (current, capital, and transfers) is expressed as a percentage of total general government expenditure on all sectors (including health, education, social services, etc.).	Percent
Primary school enrollment	Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the	Percent

	population of the age group that officially corresponds to the level of education shown.	
Gini coefficient	Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. Thus, a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.	Number
Physicians	Physicians include generalist and specialist medical practitioners per 1,000 people.	Number

Model 2 - Multiple Regression Model

$$LIFE\_EXPECTANCY\_5\_years\$`2017` = \beta 0 + \beta 1(HEALTH\_EXPENDITURE\$`2017`) + \beta 2(GDP\_Per\_Capita\$`2017`) + \beta 3(log(TOTAL\_POPULATION\$`2017`)) + \beta 4(Poverty\_Headcount\_at\_1\_9\$`2017`) + \beta 5(EDUC\_EXPENDITURE\$`2017`) + \beta 6(School\_Enrollment\_Primary\_\$`2017) + \beta 7(Gini\_Coeff\$`2017`) + \beta 8(Physicians\_1k\$`2017`) + u$$

LIFE_EXPECTANCY_5_years$`2017` = 103.85641 + 0.78256 (HEALTH_EXPENDITURE$`2017`) + (-)0.88943(GDP_Per_Capita$`2017`) + (-)0.38188 (log(TOTAL_POPULATION$`2017`)) + (-)0.35896(Poverty_Headcount_at_1_9$`2017`) + (-)0.11978(EDUC_EXPENDITURE$`2017`) + (-)0.21405(School_Enrollment_Primary_$`2017) + 0.02404(Gini_Coeff$`2017`) + 0.13421(Physicians_1k$`2017`)

Results

Residual standard error: 2.883 on 32 degrees of freedom

(145 observations deleted due to missingness)

Multiple R-squared: 0.7317, Adjusted R-squared: 0.6647

F-statistic: 10.91 on 8 and 32 DF, p-value: 2.916e-07

Notably, this model omits 145 observations due to limited information recorded or available for the poverty headcount ration, the Gini Index, and the number of physicians per a group of 1,000. Despite the smaller sample size, the multiple R-squared value is 0.7317 and the adjusted R-squared value is 0.6647. The increase in the adjusted R-squared value, compared to the single regression model, indicates that the variables selected in this model improve the model fit by explaining the majority of the variability for the dependent variable. To confirm this, for this model, the intercept and the poverty headcount ratio are significant at 0.001, GDP per capita is significant at 0.01, and health expenditure & primary school enrollment are significant at 0.05. Four of the variables, log(TOTAL_POPULATION$`2017`), EDUC_EXPENDITURE$`2017`, Gini_Coeff$`2017`, Physicians_1k$`2017`, did not depict any significance which is why the third model omits the input of those variables.

To interpret the significance of the results from this model individually, with all else held constant, each increase in the current health expenditure, expressed as a percentage of GDP, will increase the life expectancy by 0.78256. Each increase in GDP per capita growth, expressed as an annual percentage, ceteris paribus, will decrease the life expectancy by -0.88943; a surprising result since an increase in GDP per capita growth is associated with an increase in socioeconomic development and thus, the prolongation of longevity. An increase in the poverty headcount ratio at $1.90 a day, as a percent of the population, ceteris paribus, will decrease life expectancy by 0.35896. To clarify, poverty headcount ratio at $1.90 a day is the percentage of the population living o less than $1.90 a day at 2011 international prices, as stated in Table 2. Therefore, it is reasonable that an increase in this ratio will decrease life expectancy. Further, an increase in the primary completion rate or the gross intake ratio to the last grade of primary education, ceteris paribus, will decrease the life expectancy by 0.21405. Another surprising result since primary school is the introduction to education regarding physical health importance.

Model 3 – Multiple Regression Model

$$LIFE\_EXPECTANCY\_5\_years\$`2017` = \beta 0 + \beta 1(HEALTH\_EXPENDITURE\$`2017`) + \beta 2(GDP\_Per\_Capita\$`2017`) + \beta 3(Poverty\_Headcount\_at\_1\_9\$`2017`) + \beta 4(School\_Enrollment\_Primary\_\$`2017) + u$$

LIFE_EXPECTANCY_5_years$`2017` = 83.491441 + 0.870650 (HEALTH_EXPENDITURE$`2017`) + 0.002858 (GDP_Per_Capita$`2017`) + (-)0.374792(Poverty_Headcount_at_1_9$`2017`) + (-)0.119999(School_Enrollment_Primary_$`2017)

Results

Residual standard error: 3.685 on 66 degrees of freedom
Multiple R-squared: 0.6242, Adjusted R-squared: 0.6014
F-statistic: 27.41 on 4 and 66 DF, p-value: 2.034e-13

For note, this model omits 115 observations due to missing data. The multiple R-squared value is 0.6242 and the adjusted R-squared value is 0.6014, indicating that the variables in this model are a good fit and explain much of the variability of the dependent variable. The intercept, HEALTH_EXPENDITURE$`2017` and Poverty_Headcount_at_1_9$`2017` are significant at 0.001, which allows us to reject the null hypothesis for each of these variables that there is no significant relationship between them and the dependent variable. To interpret the significant results of this model, an increase in the health expenditure, expressed as a percentage of GDP, ceteris paribus, will increase the life expectancy by 0.870650. An increase in the primary school completion rate, ceteris paribus, will decrease the life expectancy by 0.374792.

The results of all 3 regressions can be seen in Table 4.

Table 4: Regression Model Output Results

Independent Variables	Model 1	Model 2	Model 3
HEALTH_EXPENDITURE$`2017'	1.0112(***)	0.78256(*)	0.870650(***)
GDP_Per_Capita$`2017'	--	-0.88943(**)	0.002858
log(TOTAL_POPULATION$`2017`)	--	-0.38188	--
Poverty_Headcount_at_1_9$`2017`	--	-0.35896(***)	- 0.374792(***)
EDUC_EXPENDITURE$`2017`	--	-0.11978	--
School_Enrollment_Primary_$`2017	--	-0.21405(*)	- 0.119999(.)
Gini_Coeff$`2017`	--	0.02404	--
Physicians_1k$`2017`	--	0.13421	--

The following figures portray the correlation of the independent variables that depicted significance in relation to life expectancy and corresponding tables with correlation of regressors.

Figure 2: A Comparison of Life Expectancy and GDP Per Capita

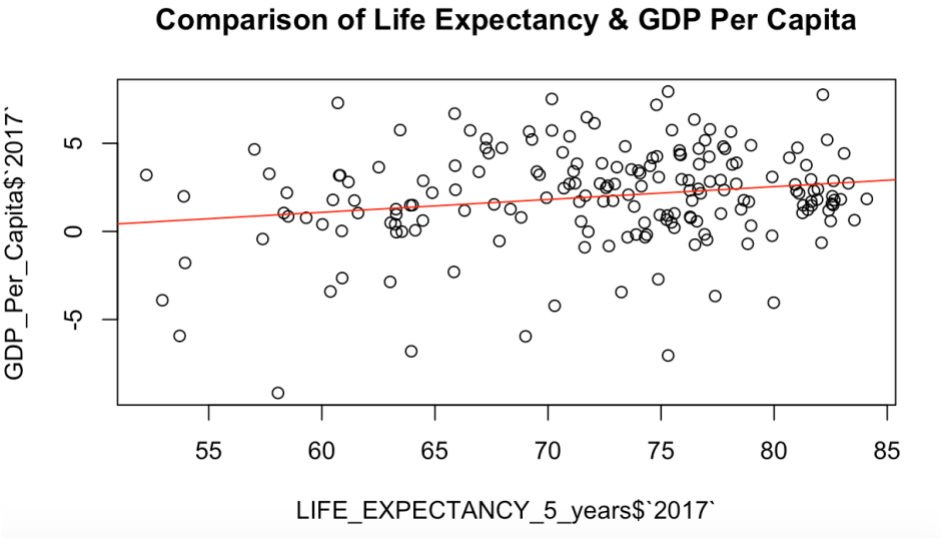


Figure 3: A Comparison of Life Expectancy and Primary School Completion Rate

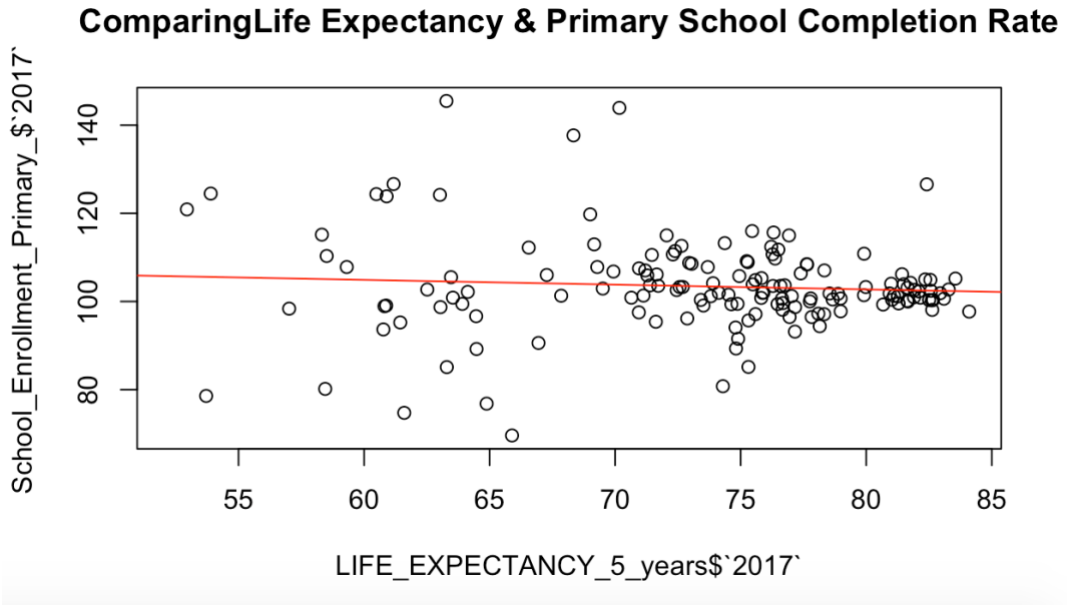


Figure 4: A Comparison of Life Expectancy and Poverty Headcount at $1.90 A Day

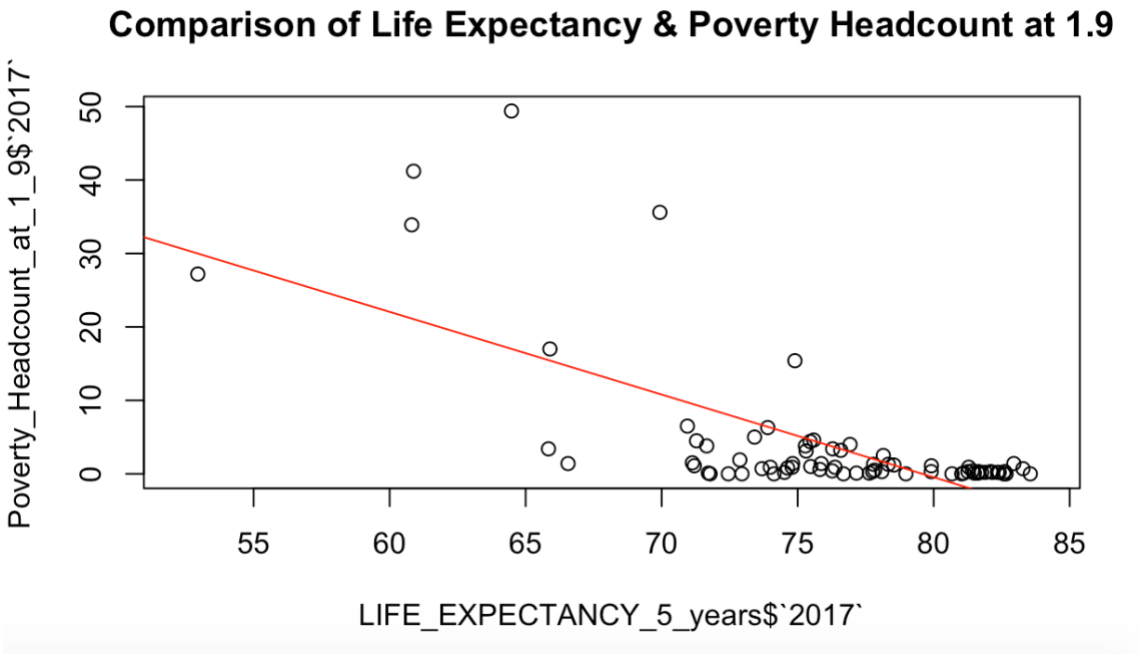


Figure 5: Performance Analytics: Correlation Matrix Chart

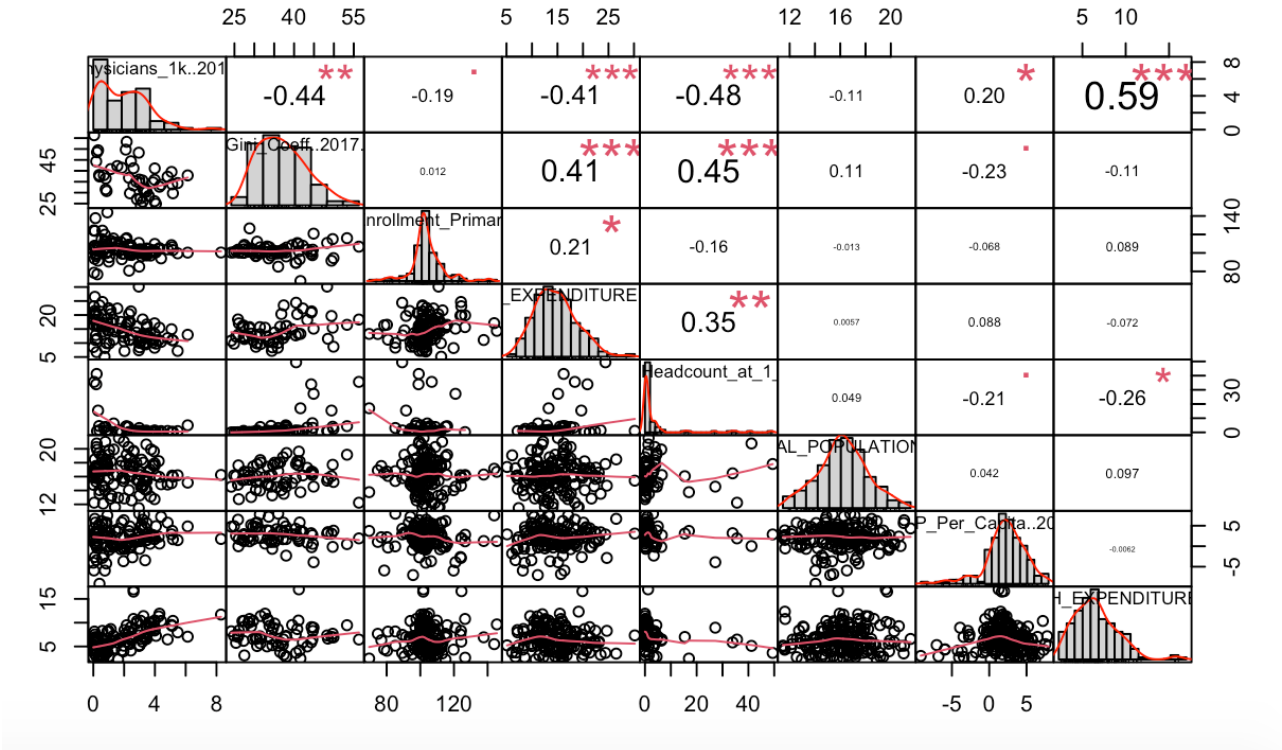


Table 5: Correlation of Regressors

	Phys_1k	*Gini_coef*	*Prim_school*	*Educ_exp*	*Pov_1.9*	*log(pop)*	*Gdp_cap*	*Health_exp*
Phys_1k	**1.00**							
Gini_coef	-0.43572	**1.00**						
Prim_school	-0.09277	0.339309	**1.00**					
Educ_exp	-0.52316	0.531674	0.25458166	**1.00**				
Pov_1.9	-0.49621	0.5663392	-0.05647337	0.2822756	**1.00**			
log(pop)	-0.15584	0.1276372	0.27448835	-0.125151	-0.1723	**1.00**		
Gdp_cap	0.06080	-0.320835	-0.43577564	-0.148942	-0.0711	-0.15293	**1.00**	
Health_exp	0.53868	-0.160040	0.16978748	-0.218942	-0.2170	0.15024	-0.39575	**1.00**

Conclusion

In concluding remarks, my initial hypothesis that there would be a significant relationship between health expenditure and life expectancy albeit true, presented itself with limited explanation regarding variability on the dependent variable. With this in mind, I inputted additional variables in the following model which presented a higher adjusted R-squared value thus, indicating a more appropriate fit of a model. I refined this model to include only the significant variables however, this lowered the adjusted R-squared demonstrating that removing certain variables from the model would lower the ability to explain the variation of life expectancy. To note, the negative significant correlation estimates of certain variables in Model 2 were surprising, in the least. To address limitations, I believe conducting a thorough investigation of all variables that might influence life expectancy would aid in achieving a higher adjusted R-squared value, for the purpose of being able to explain most of the variability of life expectancy.

References

Kim, Tae K., Lane, Shannon R. “Government Health Expenditure and Public Health Outcomes: A Comparative Study among 17 Countries and Implications for US Health Care Reform.” *American International Journal of Contemporary Research*, Sept. 2013, http://www.aijcrnet.com/journals/Vol_3_No_9_September_2013/2.pdf.

“Correlation Matrix : A Quick Start Guide to Analyze, Format and Visualize a Correlation Matrix Using R Software.” *STHDA*, http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software.

“Life Expectancy of the World Population.” *Worldometer*, https://www.worldometers.info/demographics/life-expectancy/.

Roser, Max, et al. “Life Expectancy.” *Our World in Data*, 23 May 2013, https://ourworldindata.org/life-expectancy.

Hummer, Robert A., Hernandez, Elaine M. (2013). “TheEffects of Educational Attainment on Adult Mortality in the United States,” *Popul Bull.*

Suzuki, Emi, et al. “World Bank Open Data.” *Data*, 20 Dec. 2021, https://data.worldbank.org/.

“Keeping Track Online.” *Download Data*, https://data.cccnewyork.org/data/download#0,1,3,4,6,8,10,13,21/0,3,81,1180,1294,1324,1386,1448,1493.