[Proceedings of the Jepson Undergraduate Conference on International Economics](https://scholarworks.uni.edu/jucie)

Volume 3 | Article 3

7-2021

Average Life Expectancy Across Nations

Madison Eberhart
University of Northern Iowa

Average Life Expectancy Across Nations

Madison Eberhart

Department of Economics

University of Northern Iowa

Abstract

This paper analyzes life expectancy and the factors that influence it across several different countries worldwide. Data is collected from the World Bank DataBank. The cross-sectional data set contains variables for nation population, percentage of citizens with access to electricity, net income per capita, labor force participation rate, crude birth rate, population density, the percent of a nation's population that is female, and the dependent variable being average life expectancy. This regression gives insight as to how strongly average life expectancy is affected by the different independent variables selected. It is important to understand how different elements and factors correlate with the health and lifespan of a population.

Acknowledgements

I would like to thank all involved, especially Dr. Shahina Amin for organizing the conference, Dr. Imam Alam for being my faculty advisor, and Emily Howke for acting as my discussant and providing me with great feedback.

Average Life Expectancy Across Nations

I. Introduction

There are many factors that affect a population's quality of life, many of those associated with health and income. It is a general assumption that most people want to improve their quality of life. There are many different ways to define quality of life, but most are not quantifiable. In this study, average life expectancy in years is the variable closest to exemplifying quality of life.

I use a linear regression model based on previous papers and studies. I conduct the study with cross-sectional data coming from all countries with available data from the year 2017. My chosen independent variables are based on previous theory and my own knowledge. They are population, percentage of a country with access to electricity, net income per capita, labor force participation rate, crude birth rate, population density, and percentage of total population that is female. After running a regression analysis, the variables that are more correlated with average life expectancy are identified and may have a strong impact on the number.

II. Literature Review

It is important to research other studies that are conducted on similar topics. All of the papers I examine extensively choose independent variables that are similar in concept, but different in specific data points.

Regression Analysis on Life Expectancy (2019) has independent variables of birth rate, Environmental Performance Index (EPI)[1], gross domestic product (GDP), heart disease rate,

[1] Environment Performance Index provides a summary driven by data of the state of sustainability within a country.

population, area, population density, and stroke rate. At a five-percent significance level, all independent variables are found significant. The regression analysis overall has an $\bar{R}^2$ of 0.731.

The author finds that taking care of the environment has the largest impact on a country's life expectancy, shown by the coefficient of the EPI (Regression Analysis on Life Expectancy, 2019). He finds that he could have expanded the scope to cities instead of countries, which would give more accurate results to more specific groups of people or areas. The author also finds that he could have split the data to male and female categories for this life expectancy analysis, which leads me to adding the percentage of female population to my regression analysis.

Maity et al. (2017) use the following independent variables: Poverty Headcount Ratio at $1.00, GNI per capita (PPP[2]), expenditure on health per capita, lower secondary completion rates (%), physicians per 1000 (people), hospital beds per 1000 (people), adequacy of social protection, and Gini index[3]. At a five-percent significance level, all independent variables besides GNI per capita (PPP), lower secondary completion rates (%), physicians per 1000 (people), and Gini index are found significant. The regression analysis overall has an $\bar{R}^2$ of 0.7469.

Although the model Maity *et al.* (2017) use has statistically significant variables, they do not believe the variables chosen are the best measures of average life expectancy. The authors conclude by stating that the true influencers (gender, genetics, lifestyle, etc.) may be better-fitting independent variables. While most of the true influencers mentioned are not quantifiable, the percentage of a population that is one gender is observable. Jagger *et al.* (2005) calculates

---

[2] Purchasing Power Parity. An economic term for measuring prices at different locations based on the law of one price.

[3] The Gini index is a measure of the distribution of income across a population. A Gini index indicates greater inequality.

life expectancies of citizens of the European Union at fifty years of age and categorizes them by gender and country. The study finds that the average life expectancy of men is 65 years, and the average for women is 68 years (Jagger, et al., 2005).

A fourth study finds that drug consumption, as measured by per capita pharmaceutical expenditures, has a positive effect on population life expectancy at various ages (Shaw, Horrace, & Vogel, 2005). This discovery leads me to search for independent variables involving health and pharmaceutical statistics, however I'm unable to find sufficient data in my limited amount of time.

III. Model

After reviewing different studies on average life expectancy, the independent variables I have chosen to use for this study are population, access to electricity, net income per capita, labor force participation rate[4], birth rates, population density, and percentage of population that is female.

The first variable is population. Population is represented by 'POP' in the regression model. Some very populous countries have lower life expectancy, such as India. I still predict a positive coefficient due to significant theories in a previous study (Regression Analysis on Life Expectancy, 2019). I predict this statistic to be insignificant in my study.

A variable I chose based on my own curiosity was the percentage of a population with access to electricity. I predict that a high percentage of access to electricity has a significant positive impact on quality of life. This statistic is measured as the percentage of a country with

[4] The labor force participation rate is the percentage of a country's labor force that is employed.

access to electricity and is represented by 'ELECTRIC' in the regression model. I predict that this variable has a positive coefficient, as access to electricity is a crucial factor when weighing quality of life.

The third independent variable used is net income per capita. Net income per capita is measured in U.S. dollars and is represented by 'INC' in the regression model. A higher household income indicates easier access to necessities such as proper nutrition, housing, education, and medical professionals. I predict that the coefficient of net income per capita has a positive indicator. Maity et al. (2017) uses GNI per capita (PPP) as an independent variable and finds it to be insignificant. However, Jagger et al. (2005) finds that within countries, there is increasing evidence of lower life expectancy in less privileged and developed social groups and countries. Net income per capita has a positive relationship with a country's development and privilege, ceteris paribus.

Labor force participation rate is another independent variable that I use. The studies I read included nothing about employment nor labor rates. This variable is represented by 'LFRATE' in the regression model. I predict that the coefficient of the variable would have a positive indicator. A higher percentage of those willing and able to work means that the economy is running more efficiently than those with higher rates of unemployment, ceteris paribus. This is empirical data since there is no previous theory found that suggests otherwise.

Birth rate is a variable that was used in Regression Analysis on Life Expectancy (2019) and is significant with a negative indicator. The crude birth rate is the number of births per 1,000 people in a population in one year and is represented as 'BIRTH' in the regression model. The coefficient of the birth rate's predicted indication is negative. This follows the theory that high population growth is more likely to occur in less developed countries (Shaw, Horrace, & Vogel, 2005).

Population density is the people per square kilometer of land area in a country and is represented by 'POPDENS' in the regression model. This variable is significant in Regression Analysis on Life Expectancy (2019) and has a predicted negative coefficient. A lower population density indicates that people are able to live spatially distanced. A low population density could also imply a country is less urbanized and is more agricultural, which also indicates less development.

On average, women have a higher life expectancy than men, ceteris paribus (Jagger, et al., 2005). Because of this, I believe that a nation with a higher percentage of female citizens has a higher average life expectancy. The percentage of a country's population that is female is represented as 'FEMALE'. The predicted indicator of this coefficient is positive.

Other studies use many different variables, as well. I attempted to add some of these variables, however, I am unable to attain sufficient data due to time constraints of conducting the study. The independent variables would have included are Environmental Performance Index, hospital beds per 1,000 people, adequacy of social protection, heart disease rate, per capita pharmaceutical expenditures, and Gini index. The table below shows both the dependent and independent variables and their predicted indicator, if applicable:

Variable	Description	Predicted Indicator
LE	average life expectancy in years	dependent
POP	total population	positive
ELECTRIC	percentage of population with access to electricity	positive
INC	net income per capita in U.S. dollars	positive
LFRATE	percentage of labor force that is employed (labor force participation rate)	positive
BIRTH	number of births per 1,000 people in population	negative
POPDENS	people per square kilometer of land area	negative

FEMALE	percentage of a population that is female	positive

The true population model is shown as:

$$LE = \beta_0 + \beta_1 POP + \beta_2 ELECTRIC + \beta_3 INC + \beta_4 LFRATE + \beta_5 BIRTH + \beta_6 POPDENS + \beta_7 FEMALE + \Sigma$$

However, due to a true population being impossible to attain, I use an estimated linear regression. An Ordinary Least Squares (OLS) model is the best linear unbiased estimator (BLUE). I chose a simple linear regression model as the best fit and is shown as:

$$\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 POP + \hat{\beta}_2 ELECTRIC + \hat{\beta}_3 INC + \hat{\beta}_4 LFRATE + \hat{\beta}_5 BIRTH + \hat{\beta}_6 POPDENS + \hat{\beta}_7 FEMALE$$

IV. Data & Results

The table below are the summary statistics for the different variables used in the simple regression analysis:

Variable	Obs.	Mean	Std. Dev.	Min	Max
LE	106	75.38748	6.013504	53.95	84.29
POP	106	34.06056	54.80514	.094737	324.4595
ELECTRIC	106	93.11308	18.33926	12.7	100
INC	106	14978.5	16666.72	98.35886	64629.76
LFRATE	106	60.12818	10.20117	26.4867	88.05
BIRTH	106	16.57062	8.326765	7	46.54
POPDENS	106	208.3375	773.3972	2.004286	7915.73
FEMALE	106	50.19276	3.519459	24.33257	54.56482

An OLS linear regression analysis run using robust standard errors is depicted below:

LE	Coef.	Std. Err. Robust	t	P>|t|	[95% Conf. Interval]	
POP	-.0078524	.0044868	-1.75	0.083	-.0167563	.0010516
ELECTRIC	.0964852	.0269924	3.57	0.001	.0429197	.1500507
INC	.0001454	.0000215	6.76	0.000	.0001028	.0001881
LFRATE	-.0644495	.0321661	-2.00	0.048	-.1282821	-.0006169
BIRTH	-.3075895	.0655538	-4.69	0.000	-.4376789	-.1775002
POPDENS	.0001135	.000101	1.12	0.264	-.000087	.000314
FEMALE	.0315849	.0509021	0.62	0.536	-.0694287	.1325985

Number of obs. = 106　　R-squared = 0.8314
F(7, 98) = 76.05　　Root MSE = 2.5556
Prob > F = 0.0000

Of the variables in the regression model above, BIRTH has the largest coefficient by far. It is interpreted that for each additional birth per 1,000 people in a year, the average life expectancy in the country decreases by 0.3075895, ceteris paribus. The most significant variable in terms of accuracy is INC. The variable has the highest t-value and thus is the most likely to have a strong relationship with life expectancy. For every U.S. dollar increase in a country's net income per capita, their average life expectancy increases by 0.0001454 years. ELECTRIC and LFRATE also have t-values far enough from zero to consider significant by rule of thumb.

For this regression model, assume α[5] is equal to 0.05. At a five-percent significance level, I reject the null hypotheses of ELECTRIC, INC, and BIRTH, as these independent variables are significant at this level. POP, POPDENS, and FEMALE all failed to reject the null hypotheses at the five percent significance level. LFRATE is statistically significant in the opposite direction than predicted. Originally, I expected a higher labor force participation rate to lead to a higher life expectancy, ceteris paribus.

[5] The significance level.

In order for the OLS to be considered the best linear unbiased estimator, it must follow the Gauss-Markov theorem[6]. This means the regression analysis must meet the six classical assumptions, one being that no independent variable is a perfect linear function of any other independent variables in the model. In order to test this, I conducted a multicollinearity[7] test. There are no perfect multicollinearities between independent variables. Near-perfect multicollinearity does not violate the assumption, although the standard errors of the coefficients increase when this is present. In order to check if this is the case, I ran a variance inflation factor[8] test. There are no individual VIF tests above 3.66 (this belonging to crude birth rate), and the mean VIF is 1.83. Near-perfect multicollinearity did not take place in this regression.

Another assumption that must be met is that the error term has a constant variance throughout the regression. If there is heteroskedasticity[9], the OLS is not biased, but it may no longer be the "best" linear unbiased estimator because it may no longer have minimum variance among all other unbiased estimators. I ran a heteroscedasticity test to verify that there is likely homoskedasticity in the regression. I first test ran a Breuch-Pagan Test. If the p-value that corresponds to the Chi-square test statistic is below the significance level (in this case let's say 0.05), I can conclude that there is heteroskedasticity present in the data. The p-value corresponding to the Chi-square statistic is 0.1957. Although this value alone indicates there's no reason to think there is heteroskedasticity, I also ran White's test to correct for any possible heteroskedasticity. I received a p-value of 0.2370.

[6] Gauss-Markov theorem states that the OLS estimator has the lowest sampling variance of all linear unbiased estimators, assuming it meets the classical assumptions.
[7] Multicollinearity refers to a situation where two or more independent variables in a multiple regression model are highly related.
[8] Variance inflation factor (VIF) is a measure of the amount of multicollinearity within the independent variables.
[9] Heteroskedasticity is a circumstance in which the variance of the error is not constant along the regression line.

V. Conclusions

Overall, even though the model has statistically significant independent variables, some of the variables chosen are most likely not the best measure of average life expectancy. There is also a concern of omitted variables. With a dependent variable such as this, almost everything seems as if it would have a correlation with life expectancy. It is my belief that while some variables chosen are correct to be included (crude birth rate, access to electricity, net income per capita), there are many others that could have improved this model.

If I were to conduct this study again, I would do further research to find and include variables that I could not collect data for within the limited time of this study. These variables include the Environmental Performance Index, hospital beds per 1,000 people, adequacy of social protection, heart disease rate, per capita pharmaceutical expenditures, and Gini index. I would also look into including variables that quantify social and political stability, underlying health conditions, gender equality, urbanization rates, and percentage of those within a healthy weight range. After conducting this study, it is still uncertain what factors have direct causation with life expectancy, although the income, access to electricity, and birth rate variables appear to have a strong correlation.

References

Jagger, C., Gillies, C., Moscone, F., Cambois, E., Van Oyen, H., Nusselder, W., & Robine, J.-M. (2005). Inequalities in healthy life years in the 25 countries of the European Union in 2005: a cross-national meta-regression analysis. *The Lancet, 372*(9656), 2124-2131.

Maity, A., Rhenman, E., & Sanders, E. (2017). *Factors Explaining Average Life Expectancy: An Explanation Across Nations.*

Regression Analysis on Life Expectancy. (2019, December 13). Retrieved from Hackernoon: https://hackernoon.com/regression-analysis-on-life-expectancy-4wp34rf

Shaw, J. W., Horrace, W. C., & Vogel, R. J. (2005). The Determinants of Life Expectancy: An Analysis of the OECD Health Data. *Southern Economic Journal.*