# ESTIMATING LIFE EXPECTANCY USING GDP PER CAPITA

Zachary Howser

Department of Economics

University of California, Davis

June 5, 2022

## Abstract

Life expectancy has increased drastically in modern times yet remains largely varied among high and low GDP countries. Previous research on the topic has found positive correlations between health outcomes and income levels on both an individual and national level. This paper analyzes the effect of a country's GDP per capita on the country's average life expectancy at birth using a dataset compiled from sources including World Bank, United Nations, and Our World in Data. Single and multiple variable linear regressions with nonlinear function forms are used in this study, as well as a probit regression on a binary version of the dependent variable life expectancy at birth to estimate the effect of GDP per capita on life expectancy at birth. As expected based on previous studies, our findings indicate a positive relationship between the variables meaning an increase in GDP per capita leads to a corresponding increase in life expectancy.

## Introduction

Life expectancy has changed dramatically throughout the history of the world. In pre-modern times it was estimated to be around 30 years worldwide, which then increased as countries began to industrialize. Global inequality increased as rich countries had increasing health outcomes and poorer countries remained with lower health outcomes. The goal of this study is to analyze the factors that lead to increased life expectancy at birth that created this gap between countries with one central question, what is the effect of GDP per capita on life expectancy at birth?

This question is important in the global fight for equality because it allows us to examine the global inequality in health outcomes caused by higher levels of technology and better healthcare infrastructure that are associated with higher GDP per capita in a country. Studies on the subject of income's effect on health outcomes have been published and many of them share similar conclusions. Moore, Newman, and Fheili (1992) examined the relationship between income per capita and national health care expenditures across countries over a period of 15 years. The study found that income is the most important factor in determining health care spending for a country, explaining over ninety percent of variance across countries. This result is related to the question in this study because healthcare spending has a large impact on life expectancy at birth, and income per capita was found to be the main explanatory variable.

Another article that studies the relationship between healthcare expenditure and economic performance, V. Raghupathi and W. Raghupathi (2020), examined relationships between healthcare spending and many economic indicators such as; labor productivity, average personal income per capita, and spending on goods and services. It was found that in general there was a positive association between healthcare spending and the various economic indicators that were studied which outlined how governments could allocate healthcare spending to stimulate economic growth and personal well-being. This relates to the effect of income on life expectancy because similar to the previous study, health spending is largely related to expected age upon death and this study shows positive associations between income per capita and spending.

Hummer and Hernandez (2003) was a study that analyzed the mortality rates of adults over 25 in groups based on education level. The groups were: less than a high school degree, a high school degree, some college, and a bachelor's degree or higher. The study found that individuals with higher levels of education had lower mortality rates than individuals with less education, an example being that adults with a bachelor's degree have a remaining life expectancy of almost 10 years longer than adults that did not receive a high school diploma. Higher levels of education have the benefits of better jobs and higher income which are all factors that can contribute to a longer life expectancy.

A 2008 article authored by Jemal, Ward, Anderson, Murray, and Thun studied the socioeconomic inequalities in death rates for the years 1993 to 2001. It covered 25 to 64-year-olds of different races in 43 different states based on death certificate data available and found that those with lower levels of education had a death rate that increased in the 9 years studied and those with higher levels of education had a death rate that decreased over the 9 years. Further, the study showed that black men with less than a high school diploma had a death rate of almost 1.5 times white men with the same education level. This relates to our study because it shows again that education has the benefits of higher income and longer life expectancy, and it shows that the structural inequalities that affect blacks have a similar effect to a reduced income.

Chalhoub and Twomey (2018) studied income inequality and life expectancy through the lens of tobacco prevention public policies. The study found that tobacco prevention policies along with other policies that create an environment where tobacco use is "unattractive, expensive, and restricted" lower tobacco use rates and improve public health. With tobacco advertising targeting low-income areas and youths, this is an example of lower incomes leading to worse health outcomes and higher incomes being able to avoid tobacco use through policies and reduced use leading to better health outcomes.

A study conducted by Chetty et al. (2016) measured the relationship between income and mortality on an individual level. It covered over 1.4 billion person-year observations and concluded multiple things. The study found that life expectancy continues to increase as income goes up and that inequality in life expectancy has increased over time. The study concluded that the difference in life expectancy between the top percentage of earners and the bottom percentage of earners in the population varies greatly across locations and available healthcare infrastructure, and is increasing over time.

This study will be analyzing the effect of income on life expectancy on a national level instead of an individual level like many of the studies above. Given the relationships between health outcomes and economic indicators found in the studies outlined above, our hypothesis is that countries with a higher GDP per capita will have a higher average life expectancy at birth.

## Data

The focus of this study is to analyze the relationship between GDP per capita and life expectancy at birth. The dependent variable used in this study is life expectancy at birth and the independent variables used are GDP per capita, primary school completion rate, and income from natural resource exports as a percentage of GDP. We started with data from 173 countries and after cleaning the data by removing missing entries we were left with data from 161 countries ranging from 2000 to 2018 for a total sample size of 2,027 observations. Descriptions of each of the variables used for analysis and summary statistics for the dataset can be found in tables 1 and 2 of the appendix respectively.

The 'World Sustainability Dataset' being used for this study has been put together for the TrueCue Women+Data Hackathon and contains real-world data from World Bank, United Nations, and Our World in Data. This dataset, like any other dataset that works with real-world data, contains missing data values for many countries across different years. This issue has been resolved by cleaning the data to remove missing values for variables that will be analyzed which can lead to a bias toward or overrepresentation of certain countries in the analysis so this is to be kept in mind.

## Model

One step that must be completed before proceeding to linear regression models is making sure that the model meets the least-squares assumptions for causal inference in multiple regressions:

1. **Zero Conditional Mean Assumption**

   The zero conditional mean assumption tests that when given any value of the independent variables, the expected value for the error term is always 0. The way that this is tested in this study is through the residual plot of the multiple regression model in figure 1 of the appendix. Looking at the residual plot it is seen that the plotted points are scattered above and below zero with, leniently, no obvious pattern. Therefore, it can be concluded that the model passes the zero conditional mean assumption.

2. **Random Sampling**

This condition is met because the data has been collected by World Bank, United Nations, and Our World in Data, and all countries with data available are present in the data table. Observations have however been omitted due to a lack of data available for analysis variables.

3. **Large Outliers Unlikely**

   This condition means that no observations have values that are far outside the expected range of data, which would therefore make the results of the regression misleading. This is met because the data is collected and verified by World Bank, United Nations, and Our World in Data.

4. **No Perfect Colinearity**

   This condition means that there is no perfect linear relationship between any two regressors. The STATA correlate command is used to test this for each of the variables used which can be found in table 3 of the appendix. As seen in table 3, no correlation between regressors is equal to one which means that no correlation between regressors is perfect so this assumption holds true.

**Model 1:** Single Regression

$$LifeExpB = \beta_0 + \beta_1(GDPpc) + u$$

Model 1 demonstrates a single variable linear regression between life expectancy at birth and GDP per capita. The dependent variable in the model is *LifeExpB* and the independent variable is *GDPpc*.

**Model 2:** Multiple Regression 1

$$ln(LifeExpB) = \beta_0 + \beta_1(GDPpc) + \beta_2(NatResc) + \beta_3(ChildOutS) + u$$

Model 2 demonstrates a multiple-variable linear regression model between ln(life expectancy at birth) and: GDP per capita, % income from natural resources, and percentage of children not enrolled in primary school. The dependent variable in the model is ln*LifeExpB* and the independent variables are *GDPpc, NatResc,* and *ChildOutS*.

**Model 3:** Multiple Regression 2

$$LifeExpB \ = \beta_0 + \beta_1 ln(GDPpc) + \beta_2(NatResc) + \beta_3(ChildOutS) + u$$

Model 3 demonstrates a multiple variable linear regression model between life expectancy at birth and: ln(GDP per capita), % income from natural resources, and percentage of children not enrolled in primary school. The dependent variable in the model is *LifeExpB* and the independent variables are ln*GDPpc, NatResc,* and *ChildOutS*.

**Model 4:** Multiple Regression 3

$$LifeExpB \ = \beta_0 + \beta_1(GDPpc) + \beta_2(GDPpc\#GDPpc) + \beta_3(NatResc) + \beta_4(ChildOutS) + u$$

Model 4 demonstrates a multiple variable linear regression model between life expectancy at birth and: GDP per capita, GDP per capita squared, percent income from natural resources, and percentage of children not enrolled in primary school. The dependent variable in the model is *LifeExpB* and the independent variables are *GDPpc, GDPpc#GDPpc, NatResc, and ChildOutS.*

**Model 5:** Probit Regression

$$LifeExpBinary \ = \beta_0 + \beta_1(GDPpc) + u$$

In this regression we use a probit model to estimate the effects of GDP per capita on a binary variable created from life expectancy at birth that is equal to zero if life expectancy is less than or equal to the median, 73.13, and one otherwise. The independent variable in this model is

GDPpc and the dependent variable is LifeExpBinary. We use probit regression instead of ordinary least squares (OLS) regression for our binary outcome variable because the OLS regression may predict probabilities of less than zero or greater than one and the probit regression will not.

## Results

**Model 1:** Single Regression

STATA regression results can be found in the appendix as regression 1. The regression in STATA gives us estimates for the values of $\beta_0$ and $\beta_1$which can be used to create the following estimated equation:

$$LifeExpB \ = 67.383 \ + \ 0.000267(GDPpc)$$

This regression model can be interpreted as a one-dollar increase in GDP per capita leads to a rise in life expectancy at birth of 0.000267 years. The hypothesis testing included in this STATA regression tests the null hypothesis that the coefficient for the regressor is equal to zero against the alternative hypothesis that the coefficient for the regressor is not equal to zero. The t-statistic for the test is equal to 27.73 which is larger than the critical value of 1.96, therefore we reject the null hypothesis in favor of the alternative hypothesis that the coefficient value is different from zero at a 5% significance level. The regression has an R-squared value of 0.350 which means that the regressor only accounts for 35% of the variation in the dependent variable.

This regression has a potential endogeneity issue due to the omission of some explanatory variables in the regression, such as population, which can cause a correlation between the independent variable given and the error term. In this case, this could potentially be solved by a possible instrumental variable such as purchasing power parity of a nation's currency that has no direct correlation with population but has a correlation with GDP per capita in US currency.

**Model 2:** Multiple Regression 1

STATA regression results can be found in the appendix as regression 2. The regression in STATA gives us estimates for the values of $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ which can be used to create the following estimated equation:

$$lnLifeExpB\ = 4.275 + 2.57e^{-6}(GDPpc) - 0.00125(NatResc) - 0.00664(ChildOutS)$$

This regression model can be interpreted as a 1 dollar increase in GDP per capita leads to a life expectancy increase of 0.000257 percent, and a 1 unit increase in the other independent variables lead to decreases in ln(life expectancy at birth) with percents of 100 times the coefficients above. This regression has a larger R-squared value than the first at 0.589 which means more of the variance in the dependent variable is explained by the independent variables. This reduces the omitted variable bias of the regression; meaning there is less of an effect of the covariance between GDP per capita and other independent variables on the estimated $\beta_1$.

**Model 3:** Multiple Regression 2

STATA regression results can be found in the appendix as regression 3. The regression in STATA gives us estimates for the values of $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ which can be used to create the following estimated equation:

$$LifeExpB\ = 43.184 + 3.533(lnGDPpc) - 0.0791(NatResc) - 0.251(ChildOutS)$$

This regression model can be interpreted as a 1 percent increase in GDP per capita leads to a life expectancy increase of 0.03533 years, a relationship that can be seen in figure 2 of the appendix. The other regressors lead to decreases in life expectancy with coefficients that can be found above. This regression has a larger R-squared value than the first and second regressions at 0.730 which means more of the variance in the dependent variable is explained by the independent variables. Again the appearance of multiple independent variables reduces the effects of the omitted variable bias. For this regression, a hypothesis test testing that all coefficients for the regressors are equal to zero was run and the resulting F-statistic was 1777.92, which we compare to the critical value for 3 restrictions, 2.6, and we conclude that we can reject the null hypothesis that all coefficients equal zero in favor of the alternative hypothesis which is that the null hypothesis is false.

**Model 4:** Multiple Regression 3

STATA regression results can be found in the appendix as regression 3. The regression in STATA gives us estimates for the values of $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ which can be used to create the following estimated equation:

$$LifeExpB \ = 70.196 + 0.00436(GDPpc) - 3.44e^{-9}(GDPpc\#GDPpc)$$
$$- 0.0849(NatResc) - 0.371(ChildOutS)$$

We can interpret the effect of GDPpc and GDPpc$^2$ on LifeExpB in this regression model with the use of the margins command in STATA. With the margins command we learn that the marginal increase in life expectancy at the mean values of GDPpc, NatResc, and ChildOutS is 0.00340, which means that when GDP per capita is equal to $14,056.79, income from natural resources is equal to 6.37, and children out of primary school is equal to 7.21, a one-dollar increase in GDP per capita leads to a 0.00340-year increase in life expectancy at birth. This regression has an R-squared value of 0.671 which means that the variables explain a large amount of the variance in life expectancy and the number of independent variables reduces the omitted variable bias for the regression.

**Model 5:** Probit Regression

$$LifeExpBinary = -1.10 + 0.000129(GDPpc)$$

For this probit regression model we need to use the STATA margins command to allow us to approximate the effect of GDP per capita on the binary variable of life expectancy. Based on the estimated coefficient from the margins command at GDPpc equal to 10,000, 0.0000506, if GDPpc changes from 10,000 to 10,000 plus an amount ε then the probability that binary life expectancy will be equal to one increases by 10,000 multiplied by ε.

## Conclusion

The initial hypothesis was that an increase in income, measured by GDP per capita, would lead to a higher average life expectancy for a country. This makes sense because a country with a higher GDP per capita will likely have better access to health infrastructure and better standards of living. Based on the results of the multiple regressions models that have been run, we can conclude that the hypothesis was correct. There is a statistically significant positive relationship found between GDP per capita and life expectancy at birth found in multiple regressions that aimed to analyze this relationship as accurately as possible. In our regression with the largest variance in life expectancy explained by the independent variables, we found that all other factors constant, a 1 percent increase in GDP per capita leads to a life expectancy increase of 0.03533 years.

This relationship could be further studied to have a larger impact on the global fight for equality in health outcomes with the availability of further complete data, as this dataset only reaches 18 years with many incomplete data points, and the growth of the inequality between countries’ health outcomes has been prevalent for over 200 years.

# References

Chalhoub, T., & Twomey, M. (2018, August 6). *Income inequality and life expectancy*. Center for American Progress. Retrieved June 5, 2022, from https://www.americanprogress.org/article/income-inequality-life-expectancy/

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., & Cutler, D. (2016). The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA*, *315*(16), 1750–1766. https://doi.org/10.1001/jama.2016.4226

Hummer, R. A., & Hernandez, E. M. (2013). The Effect of Educational Attainment on Adult Mortality in the United States. *Population bulletin*, *68*(1), 1–16.

Jemal A, Ward E, Anderson RN, Murray T, Thun MJ. (2008). Widening of Socioeconomic Inequalities in U.S. Death Rates, 1993–2001.

Moore, W. J., Newman, R. J., & Fheili, M. (1992). Measuring the relationship between income and NHEs (national health expenditures). *Health care financing review*, *14*(1), 133–139.

Raghupathi, V., & Raghupathi, W. (2020). Healthcare Expenditure and Economic Performance: Insights From the United States Data. *Frontiers in public health*, *8*, 156. https://doi.org/10.3389/fpubh.2020.00156

# Appendix

**Table 1:** Description of Variables

| Variable Name | Description |
|---|---|
| LifeExpB | Value which captures the expected age a newborn baby will live to |
| GDPpc | GDP per capita (current US$) |
| NatResc | % Income accrued from natural resources (e.g. Exports) as a percentage of GDP |
| ChildOutS | Percentage of primary-school-age children who are not enrolled in primary or secondary school. Children in the official primary age group that are in preprimary education should be considered out of school. |

**Table 2:** Summary Statistics

| Variable | Obs | Mean | Std. dev. | Min | Max |
|---|---|---|---|---|---|
| LifeExpB | 2,027 | 71.13071 | 8.762917 | 42.595 | 84.93415 |
| GDPpc | 2,027 | 14058.79 | 19457.07 | 90.53243 | 118823.6 |
| NatResc | 2,027 | 6.368408 | 10.00112 | 0 | 81.94996 |
| ChildOutS | 2,027 | 7.214673 | 10.87436 | 0 | 69.4751 |

**Table 3:** Correlation of Regressors

| **Variables** | GDPpc | NatResc | ChildOutS |
|---|---|---|---|
| GDPpc | 1.0000 | | |
| NatResc | -0.0933 | 1.0000 | |
| ChildOutS | -0.3175 | 0.1989 | 1.0000 |

**Figure 1:** Residual Plot of Multiple Regression Model

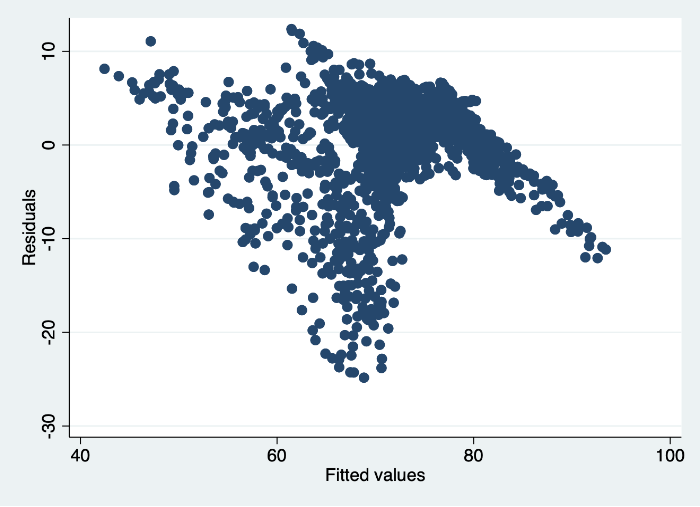
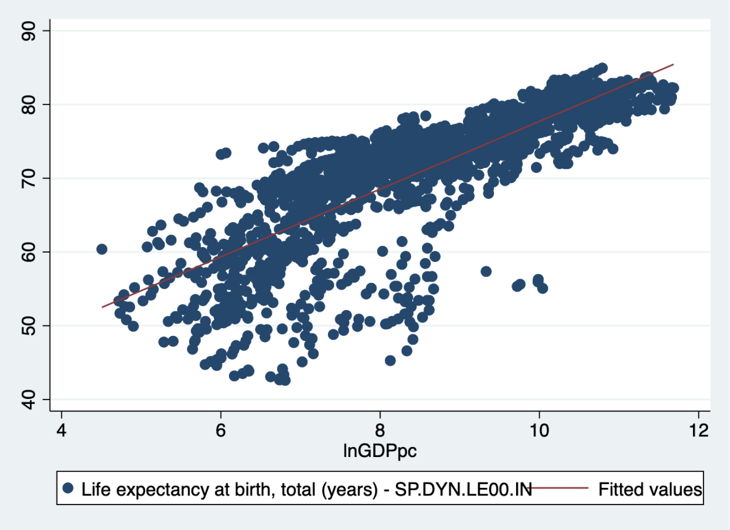


**Figure 2:** Scatter Plot of LifeExpB on lnGDPpc with Regression Line



**Regression 1:** LifeExpB on GDPpc

```
Linear regression                               Number of obs     =      2,027
                                                F(1, 2025)        =     769.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3503
                                                Root MSE          =     7.0648
----------------------------------------------------------------------------------------
                       |               Robust
              LifeExpB | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-----------------------+----------------------------------------------------------------
                 GDPpc |   .0002666   9.61e-06    27.73   0.000     .0002477    .0002854
                 _cons |   67.38301   .2236004   301.35   0.000      66.9445    67.82152
```

```
------------------------------------------------------------------------------
```

**Regression 2:** lnLifeExpB on GDPpc, NatResc, ChildOutS

```
Linear regression                               Number of obs     =      2,027
                                                F(3, 2023)        =     926.29
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5894
                                                Root MSE          =      .0861

-------------------------------------------------------------------------------
                        |               Robust
             lnLifeExpB | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
------------------------+------------------------------------------------------
                  GDPpc |   2.57e-06   1.03e-07    24.98   0.000     2.37e-06    2.77e-06
                NatResc |  -.0012478   .0001627    -7.67   0.000    -.0015668   -.0009288
              ChildOutS |  -.0066358   .0001957   -33.90   0.000    -.0070196   -.0062519
                  _cons |   4.275739   .0031268  1367.47   0.000     4.269607    4.281871
-------------------------------------------------------------------------------
```

**Regression 3:** LifeExpB on lnGDPpc, NatResc, ChildOutS

```
Linear regression                               Number of obs     =      2,027
                                                F(3, 2023)        =    1777.92
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7301
                                                Root MSE          =     4.5561

-------------------------------------------------------------------------------
                        |               Robust
               LifeExpB | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
------------------------+------------------------------------------------------
                lnGDPpc |   3.532766   .0816331    43.28   0.000     3.372673     3.69286
                NatResc |  -.0790607   .0097022    -8.15   0.000    -.0980881   -.0600333
              ChildOutS |  -.2512577   .0125097   -20.09   0.000    -.2757909   -.2267245
                  _cons |   43.18423   .8104423    53.28   0.000     41.59484    44.77362
-------------------------------------------------------------------------------
```

**Regression 4:** LifeExpB on GDPpc, GDPpc$^2$, NatResc, ChildOutS

```
Linear regression                               Number of obs     =      2,027
                                                F(3, 2022)        =          .
                                                Prob > F          =          .
                                                R-squared         =     0.6711
                                                Root MSE          =     5.0301


-------------------------------------------------------------------------------
                        |               Robust
               LifeExpB | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
------------------------+------------------------------------------------------
                  GDPpc |   .0004365   .0000144    30.22   0.000     .0004082    .0004648
      c.GDPpc#c.GDPpc |  -3.44e-09   2.00e-10   -17.21   0.000    -3.84e-09   -3.05e-09
                NatResc |  -.0849267   .0101292    -8.38   0.000    -.1047915    -.065062
              ChildOutS |  -.3707742   .0108643   -34.13   0.000    -.3920806   -.3494678
                  _cons |   70.19347   .2240071   313.35   0.000     69.75416    70.63278
-------------------------------------------------------------------------------
```