An Analysis of Life Expectancy: Evidence from a Global Cross-Section

Statistics 525 – Group 5

December 3, 2019

Scott Cohn	Haotian Jiang
Ben Goslin	Ruja Kambli
Sophia Chan	Addis Gunst

Abstract: This paper explores the impact that various predictors have on life expectancy, measured in average years of life. We exploit a cross-section using 248 countries and use multiple linear regression techniques for analysis. We find that high birth rates and high rates of death from stroke are strong indicators of low life expectancy. Additionally, high Environmental Performance Index (EPI) correlates with higher life expectancy. Counter to our initial hypotheses, we find Gross Domestic Product (GDP) has minimal impact.

Introduction.

We base our modern attempt to live longer off of a metric called life expectancy. Life expectancy is a value determined by the average age of death for a certain time period or geographic location. Policy makers attempt to raise this value by promoting known mechanisms of health such as "clean" eating, frequent exercise, and medication. Medical science works to eradicate disease, limit malnutrition, and decrease stress as much as possible. It is known that these factors tend to decrease life expectancy. Our question then becomes, what factors contribute the most to life expectancy, *and*, further, what is the magnitude of those effects on a global scale?

In our analysis we seek to look for factors that affect life expectancy on a global level, comparing a cross-section of countries to their respective levels of contributing factors. When

choosing a topic, we discovered that the interests of our group are diverse and did not necessarily overlap. However, we all came to an agreement that we wanted our analysis to be about something meaningful. Thus, we decided to pursue the topic of life expectancy for its factors are universally important to understand for general health purposes.

Some of the literature that we consulted for this project shows that the most significant factor for life expectancy tends to be the GDP of a country, or a similar measure of overall country wealth. These included Chen and Ching (2000) who examined a multitude of different variables that may influence life expectancy, Shaw, Horrace, and Vogel (2005) who incorporated GDP into an analysis of pharmaceutical spending, and Ng (2019) who found that an increase in GDP would add half a year to an individual's life.

With that being said, most studies (including ours) found that there were other, more important factors that would lead to higher life expectancy. After all, The United States and China have the largest economies in the world and both national life expectancies fail to break the top ten. This may be partially due to how GDP is often multicollinear with other economic variables, or the existence of other factors. We expand on these ideas in the following sections.

Data.

We exploit a cross-sectional dataset from 2017 on life expectancy that was scraped from a Github user's overlapping project. There are 248 countries listed. These data include a number of metrics that seek to provide explanatory power to cross-country variation. *Birth Rate* is the number of births per 1,000 people in the population per minute. *Cancer Rate* is the death rate due to cancer per 100,000 people. We remark that this does not include frequency or rate of diagnosis. The *EPI*, or Environmental Performance Index, measures environmental health and ecosystem vitality. It illustrates how close countries are to meeting global environmental goals. The *GDP*, or Gross Domestic Product, measures the total value of goods produced and services provided in a country during one year. Here, we remark that GDP does not account for wealth

or production inequalities that are captured through other metrics. The *Health Expenditure* is the total health expenditure per capita in purchasing power parity (PPP) international U.S. dollars (not inflation-adjusted). The *Heart Disease Rate* and *Stroke Rate* is the death rate from heart disease and stroke rate, respectively, per 100,000 of the population. Finally, our dependent variable *Life Expectancy* is measured in years.

In cleaning the data, there were some missing values, but our degrees of freedom were high enough that we had no analytical problems associated with missing values. Further, these data required little transformation. A short function was written to capitalize the first letter of the country name for use in visuals (e.g. Figure 1). No further data transforms or mergers were made.

First, we looked at the life expectancy of the top and bottom 10 countries:

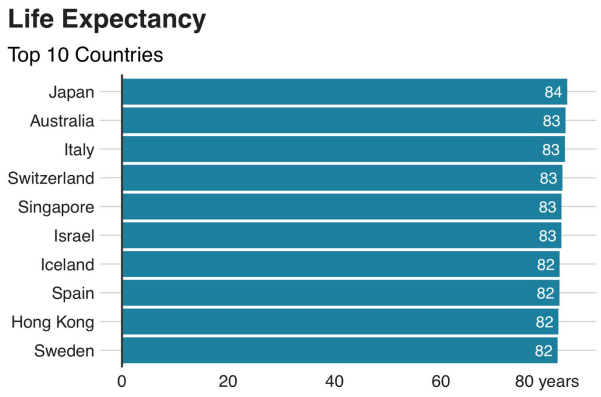

Figure 1a: Top 10 Life Expectancy by Country.

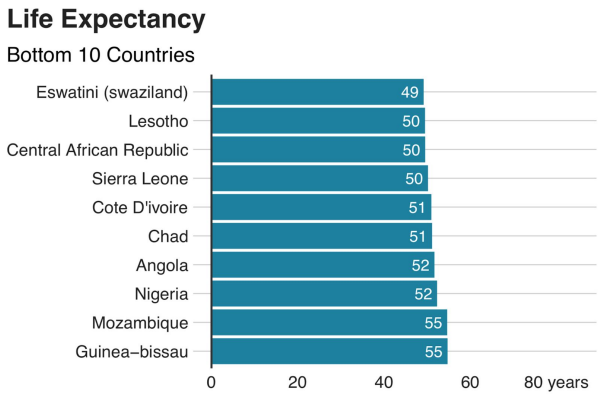

Figure 1b: Bottom 10 Life Expectancy by Country.

The countries that are in the bottom 10 are significantly poorer than the countries in the top 10. Moreover, the distribution of life expectancy global is left-skewed as seen in Figure 2. These visuals provide motivation for our analysis of the relevant factors.

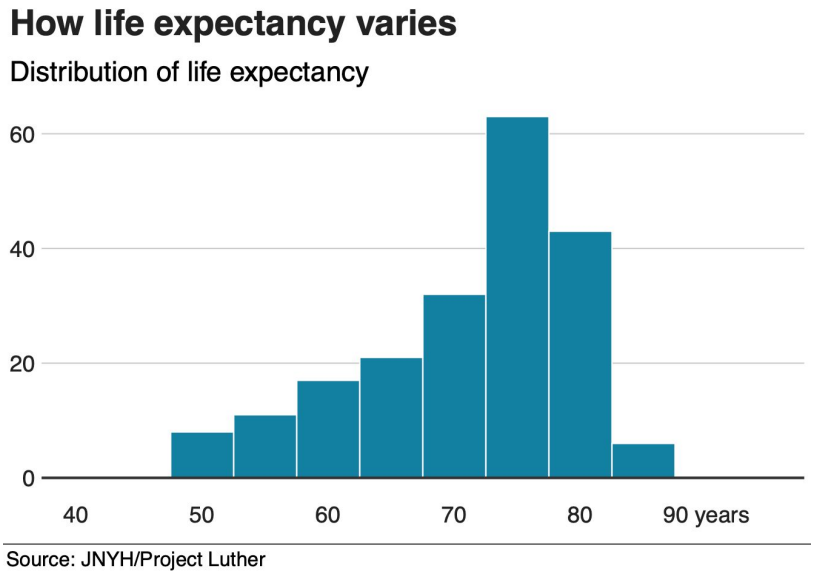

Figure 2: Histogram of Life Expectancy

While we do not have evidence to make any causal statements about the distribution of life expectancy yet, previous literature remarks the increasing trend in life expectancy in using panel data (Gapminder, 2019).

Next we seek to explore the relationships of our independent variables. The literature suggests that GDP is a strong indicator of life expectancy. We analyze this relationship in the following section. In Figure 3, observe that GDP exhibits a strong left skew. A careful zoom shows singular countries on the right-most tail of the distribution. Unsurprisingly, the United States and China rank among the countries with the highest GDP. It is likely that many of the countries shown in the top 10 life expectancies have smaller GDPs, and thus are not illuminated in a distribution of Gross Domestic Product. In completing a more rigorous analysis, we document our methods below.

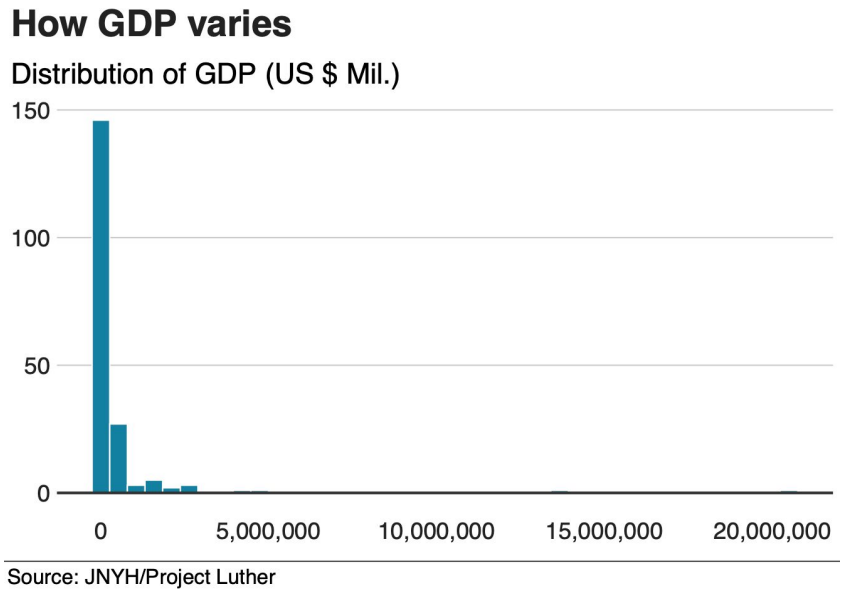

Figure 3: Histogram of Gross Domestic Product (GDP)

Methods.

We use a multiple linear regression model to explore these data because we have several explanatory variables for life expectancy. Modeling each variable against life expectancy would result in multiple models exhibiting omitted variable bias. Further, multiple regression allows a closed form expression where each independent variable can be fixed or manipulated to make predictions.

Analysis.

To begin our analysis, each variable was plotted against life expectancy individually to note any general patterns in the data.[1] The model we use is:

$$\text{Life Expectancy} = \beta_0 + \beta_1 \text{ Birth Rate} + \beta_2 \text{ Cancer Rate} + \beta_3 \text{ Heart Disease Rate} + \beta_4 \text{ Stroke Rate} + \beta_5 \text{ Health Expenditure} + \beta_6 \text{ EPI} + \beta_7 \text{ GDP} + \varepsilon$$

Next, we tested whether this model meets all of the linear regression assumptions listed below:

1. The regression function is linear (the relationship is linear).
2. The error terms have a constant variance

[1] See the technical appendix for the code that constructed these graphs.

3. The error terms are independent (there is no relationship among the error terms).
4. The error terms are normally distributed
5. There is no outlier in the data.
6. There are no important predictors that have been omitted from the model

First, we claim the function is linear in its parameters. A model is linear when each term is either a constant or the product of a parameter and a predictor. This is determined by graphing each independent variable against the dependent variable. The result is that the regression equation is linear in both its variables and parameters.

Second, we remark that the error terms have a non constant variance based on the residual plot as seen in Figure 4 below, based on our model. When plotting fitted values vs. residual errors, we can see a clear pattern in the plot that violates homoskedasticity, showing non constant variance. Namely, the higher the observed life expectancy of a country, the more accurate the model becomes. In other words, since error variance decreases as observed life expectancy increases, our model does not have constant variance. Thus, this diagnostic plot reveals that homoskedasticity is violated in our model.

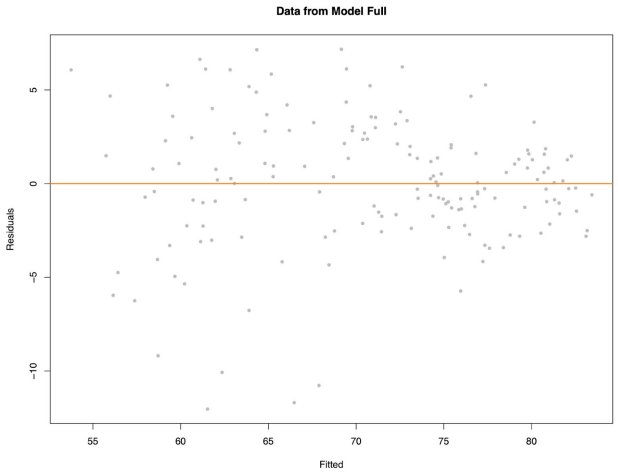

Figure 4: Fitted Values versus Residuals

Third, we claim there is no relationship between the error terms. First, we invoke a visual test to determine whether the values in Figure 4 appear to be randomly plotted. We

suspect that there may be an undesirable relationship here and invoke a more complete test. We employ a Breusch-Pagan[2] test for homoskedastcity:

$$\text{BP} = 20.003 \qquad \text{df} = 7 \qquad p\text{-value} = 0.006$$

We observe see a small p-value, so we reject the null hypothesis of homoskedasticity and heteroskedasticity assumed. Thus, the constant variance assumption is violated and so, the error terms are dependent. This matches our worries with the results of Figure 4. Hence, we caution interpretation of the p-values in Appendix A as heteroskedasticity tends to produce p-values that are smaller than they should be. This suggests that the variance of our coefficient estimates is larger than they ought to be if there was no sign of heteroskedasticity.

Fourth, we claim the error terms are normally distributed. We illustrate this claim in three ways. In Figure 5, below, we see the residuals are mostly distributed normally with a few potential outliers and a central spike. We can check this more completely with a Q-Q Plot.

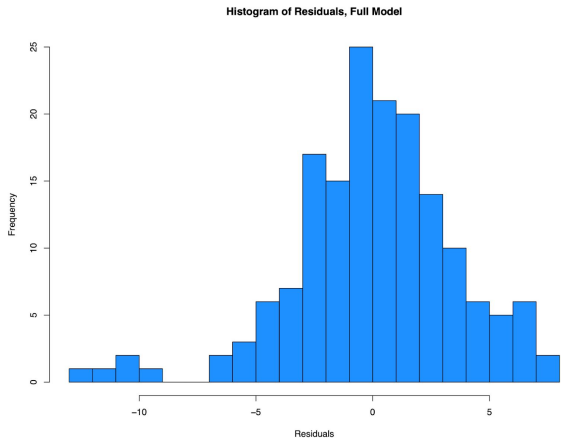

Figure 5: Histogram of the Residuals

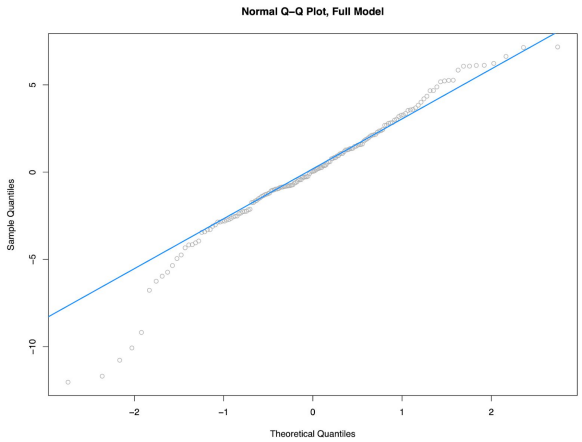

Figure 6: Q-Q Plot of the Model

Based on the Q-Q plot in Figure 6, we see that upon further exploration, this model may not be normally distributed. While the data falls on the line in the center of the graph, it curves

[2] Breush-Pagan Test: Test for heteroscedasticity of errors; $N*R^2$ where N is the sample size and R^2 is Coefficient of Determination of the regression of squared residuals

off with heavy tails at the extremities, implying that these data have more extreme values than would be expected if they came from a normal distribution. This visual cue leads to the third method.

The third way we can determine whether the error terms are normally distributed is with a Shapiro-Wilks[3] test. The Shapiro-Wilks test rejects the hypothesis of normality for small *p*-values. Our model returns a *p*-value of 0.0003. It follows that we have evidence to reject the null hypothesis of normality. That is, it is unlikely that the errors are normally distributed. This result is suggestive of other potential model specification problems. While this result is not necessarily problematic for our interpretations, it may bias later predictions.

Fifth, we claim there are no outliers in these data. We observe that outliers only occur in the case of Gross Domestic Product. The United States and China have significantly higher GDPs (more than 1.5 times the IQR) than all the other countries. This behavior, however, can be attributed to country size. When tested with simple removal of the outliers again, the results do not change. Hence, they are left in for completeness.

Finally, we claim there is no omitted variable bias that can be accounted for. Given the data we have, it suffices to show that no variable is correlated with another excluded variable. We use variation inflation factors (VIF) to determine this. All VIFs are less than 5. Since multicollinearity does not occur in these data, we claim that omitted variable bias does not occur.

Results.

The regression table is in Appendix A. Our results show that an increase in the birth rate negatively influences the life expectancy of a country. This follows from the fact that wealthier countries with higher life expectancies have lower rates of birth. The stroke rate variable is

[3] Shapiro-Wilks: Test if random sample comes from a normal distribution. Test gives a value W and if W is small, the sample is not normally distributed

negative and statistically significant. An increase in the stroke rate negatively impacts life expectancy. Further, we also find that the Environmental Performance Index (EPI) is statistically significant and positive. That is, as a country seeks to enhance its environmental performance, the life expectancy increases. Further analysis might show this is an instrument for public health or environmental quality. Counter to our hypothesis, GDP is not statistically significant. We hypothesize that this may occur due to the high number of small wealthy countries in the upper echelons of life expectancy that do not have large Gross Domestic Products.

Conclusion.

We seek to determine the effects of a variety of parameters of the life expectancy of various countries. We exploit a multinational cross-section and use multiple regression techniques. We find that in contrast to some literature, GDP is not a strong indicator of life expectancy. Rather, measures of health and environmental quality are more indicative of life expectancy. We conclude (despite heteroskedasticity) our model provides unbiased point estimates of life expectancy.

References.

Chen, M., & Ching, M. (2000, December 18). Penn Engineering. Retrieved November 11, 2019, from https://www.seas.upenn.edu/~ese302/Projects/Project_2.pdf.

Gapminder Data. Retrieved November 24, 2019, from https://www.gapminder.org/data/

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

Ng, James. "Regression Analysis on Life Expectancy." *Medium*, Towards Data Science, 9 Nov. 2019, from towardsdatascience.com/regression-analysis-on-life-expectancy-691477 5a77e2.

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*, *71*(4), 768–783. doi: 0.2307/20062079

Appendix A — Regression Table

Descriptive statistics/selected variables

	Dependent variable:
	Life Expectancy
Birth Rate	- 0.42***
	(0.04)
Cancer Rate	- 0.02
	(0.01)
Heart Disease Rate	0.002
	(0.01)
Stroke Rate	- 0.06***
	(0.01)
Health Expenditure	- 0.0002
	(0.0003)
EPI	0.18***
	(0.04)
GDP	0.0000
	(0.0000)
Constant	76.37***
	(2.96)
Observations	164
R2	0.83
Adjusted R2	0.83
Residual Std. Error	3.55 (df = 156)
F Statistic	110.82*** (df = 7; 156)
Note:	*p<0.1; **p<0.05; ***p<0.01
	p-values in parentheses.