LMM Analysis of Life Expectancy Data from the GHO

Keaton Raymond
kraymond@uccs.edu

Abstract

This analysis aims to use Linear Mixed Effect Modeling (LMM) to analyze life expectancy data from the Global Health Observatory (GHO). Through this analysis, I will aim to answer two hypotheses: do more developed countries correlate to a higher life expectancy and does a higher average level of schooling correlate to a higher life expectancy.

Introduction

For this assignment, I was tasked with analyzing a dataset of my choice using Linear Mixed Effect Modeling (LMM). An LMM is a generalized linear model that utilizes both fixed and random effects. The benefits of LMM will be explained later with the explanation of the two hypotheses and the graphs. I chose a dataset [4] from the Global Health Observatory (GHO) which is under the World Health Organization (WHO) that gives metrics related to life expectancy for all countries from the years 2000 to 2015. Using LMM, I sought out to answer two hypotheses: do more developed countries correlate to a higher life expectancy, and does a higher average level of schooling correlate to a higher life expectancy.

Data

The data that I chose keeps track of the life expectancy and other related factors for all countries from the years 2000 to 2015. Some variables that it includes are adult mortality, infant deaths, alcohol consumption, vaccination statistics, and many others. The columns that I focused on include: the country, year, status (whether the country is developed or developing), life expectancy, and schooling (the average number of years of schooling). I chose to focus on ten different countries, because if I were to focus on all 193 countries, the scatterplots that I created would have been too cluttered and difficult to read (unless I was focusing on global averages). Of the ten chosen countries, I chose five that were developed and five that were developing to try and highlight the differences between developed and developing countries. The five developed countries were: Canada, Australia, Belgium, Germany, and the United States of America (swapped to the Netherlands for the second chart due to missing data). The five developing countries were: Afghanistan, Guatemala, Serbia, Pakistan, and Nigeria.

As for cleaning the data, the first thing that I noticed was that Canada was labeled as a developing country. This seemed wrong to me; however, this isn't current data, it is from 2000-2015, so I did some research and found that Canada has been a developed country since at least the 19th century [1]. Given this, I decided to change Canada from "developing" to "developed" in case I decided to add some sort of filter based on developing or developed to the chart. From there, the only other cleaning that I had to do was removing outliers for life expectancy. I spotted these outliers by looking at the initial graph and noticing anomalies. For example, in Nigeria the life expectancy in 2006 was 49, then it jumped up to 55 in 2007, then 59 in 2008, then back down to 51 in 2009. To remove outliers like this, I added some code to manually remove the years that I found were significant outliers where the life expectancy unexpectedly shot up for a few years then back down to the normal expected value. Just because this is all the cleaning that I did doesn't necessarily mean that the rest of the data was clean. There were a lot of columns that I chose not to use because the data was simply too dirty. For example, I tried to run an analysis on the GDP per capita against the percentage of GDP per capita that was used on healthcare. However, I noticed that for many countries, the percentage used for healthcare was well above 100% with some datapoints being upwards of 10,000%. Another analysis that I tried to run was life expectancy versus the average BMI; unfortunately, I found that the average BMI for some countries was well above what the actual average is. For example, I found that in America, the average BMI was around 63-70 for

the years 2000-2015. While those are technically valid values, that would mean that the average American weighs well over 500 pounds which didn't sound right. After some research, I found that the average BMI in 2015 was actually 29.1 for men and 29.6 for women [2] which is far from the values in the dataset. Due to errors like this, there were many columns that I could not use because they had either too many missing values or too many inaccurate values. That was something that influenced my final decision on what analyses to run because there were many columns that I couldn't use.

Charts and Hypotheses

The first hypothesis that I tested was that developed countries correlate to a higher life expectancy. To test this, I ran an LMM analysis on a scatterplot (Fig. 1) that plotted the life expectancy per year for each of the ten countries listed above with the outliers removed. The scatterplot has a color code for each of the countries where each country also has its own regression line. As for the LMM regression, I treated the year and country as fixed variables and the life expectancy as the random variable.

The second hypothesis that I tested was that more years of schooling correlates to a higher life expectancy. To test this, I ran an LMM analysis on a scatterplot (Fig. 2) that plotted the average years of schooling against the life expectancy for each of the ten countries. In this analysis, I replaced the Unites States of America with the Netherlands, because the United States of America was missing values for the average years of schooling. The scatterplot has a color code for each of the countries where each country also has its own regression line. As for the LMM regression, I treated the life expectancy and country as fixed variables and the schooling as the random variable.

Results and Analysis

Overall, I found that both of my hypotheses were confirmed: developed countries correlate to a higher life expectancy and more years of schooling also correlate to a higher life expectancy.

When looking at the life expectancy per year chart (Fig 1.), you can see that I have a regression line for each of the countries, then the blue line across the center of the chart is the regression line for the entire chart. The idea behind that blue regression line isn't to view some sort of trend or relationship, but instead to show that for data such as this, doing an overall regression is useless. For this kind of data, doing LMM is the best way to analyze this data. As for my hypothesis, you can see a grouping of five countries that all have similar life expectancies on the right; those were the five developed countries and, of the ten countries analyzed, they have the five highest life expectancies. When I first saw this, my initial thought was that a higher GDP correlated to a higher life expectancy; however, that doesn't seem to necessarily hold true for the developing countries as there are likely many more factors that influence life expectancy in developing countries other than GDP. For example, Afghanistan has a higher life expectancy than Nigeria by twelve to fifteen years on average; however, according to the World Bank [3], in 2015 Afghanistan had a GDP of 19.998 billion USD, while Nigeria had a GDP of 493.026 billion USD. This shows that while a country's state of development does have a correlation to the country's life expectancy, their GDP clearly does not, because Nigeria's GDP is substantially higher than Afghanistan's, but Afghanistan has a much higher life expectancy. On this chart, you are also able to see that the countries that have a lower average life expectancy also have a less steep slope than the countries with higher life expectancy, where a steeper slope means that the life expectancy has stayed more consistent. This means that while countries like Nigeria or Afghanistan have a lower average than countries like Australia or the United States of America, they are improving at a faster rate, which is overall a good thing.

When looking at the chart of years of schooling against life expectancy (Fig. 2), you can see that I have the regression lines for each of the countries as well as the blue line across the center of the chart, which is the regression line for all points on the chart. Like the last chart, the idea of the overall regression line isn't to show any sort of trend but is instead there to demonstrate the differences between the two methodologies. In this case, if we were to do an overall regression, it doesn't tell the entire story. When you see this chart, you can see that there is a light correlation between years of schooling and life expectancy. Viewing this chart, you can see that the five developing countries have the lowest level of schooling as well as the lowest life expectancy, while the five developed countries have the highest life expectancies and the highest level of schooling. All five of the developed countries have an average years of schooling of 16 years or above, which is the number of years required for a 4-year degree. Past this threshold of 16 years, there seems to be no correlation between schooling and life expectancy. However, below that threshold, there isn't an overall correlation across the five countries, while each individual country has a trend of more schooling correlating to a higher life expectancy. I believe that the reason for this is that most of the time the

reason countries have lower life expectancy is an abundance of poverty, while the more schooling somebody has, typically the more money they will make (with diminishing returns after a 4-year degree), which means that they are less likely to be below the poverty line.

Conclusion

Overall, I discovered a lot about how this analysis technique works and why it would be used. This is why I included the overall regression lines for each chart: to show the difference between a normal regression and LMM. Without LMM, there are conclusions that can be made; however, they are not fully accurate conclusions without considering the random effects.

References

[1] Johnston, M. (2022, July 13). *The economy of Canada: An Explainer*. Investopedia. Retrieved May 7, 2023, from https://www.investopedia.com/articles/investing/042315/fundamentals-how-canada-makes-its-money.asp#:~:text=Canada%20is%20a%20highly%20developed,mining%20companies%20in%20the%20world

[2] Thomas, N. (2018, December 20). *Are you heavier or shorter than the average American?* CNN. Retrieved May 7, 2023, from https://www.cnn.com/2018/12/20/health/us-average-height-weight-report/index.html#:~:text=In%202015%2D16%2C%20the%20average,men%20and%2028.2%20in%20women

[3] GDP *(constant 2015 US$)*. World Bank Open Data. (n.d.). Retrieved May 7, 2023, from https://data.worldbank.org/indicator/NY.GDP.MKTP.KD?end=2015&start=2014

[4] Akhil. (2019, May 2). Life expectancy (WHO). Kaggle. Retrieved May 7, 2023, from https://www.kaggle.com/datasets/augustus0498/life-expectancy-who

Appendix

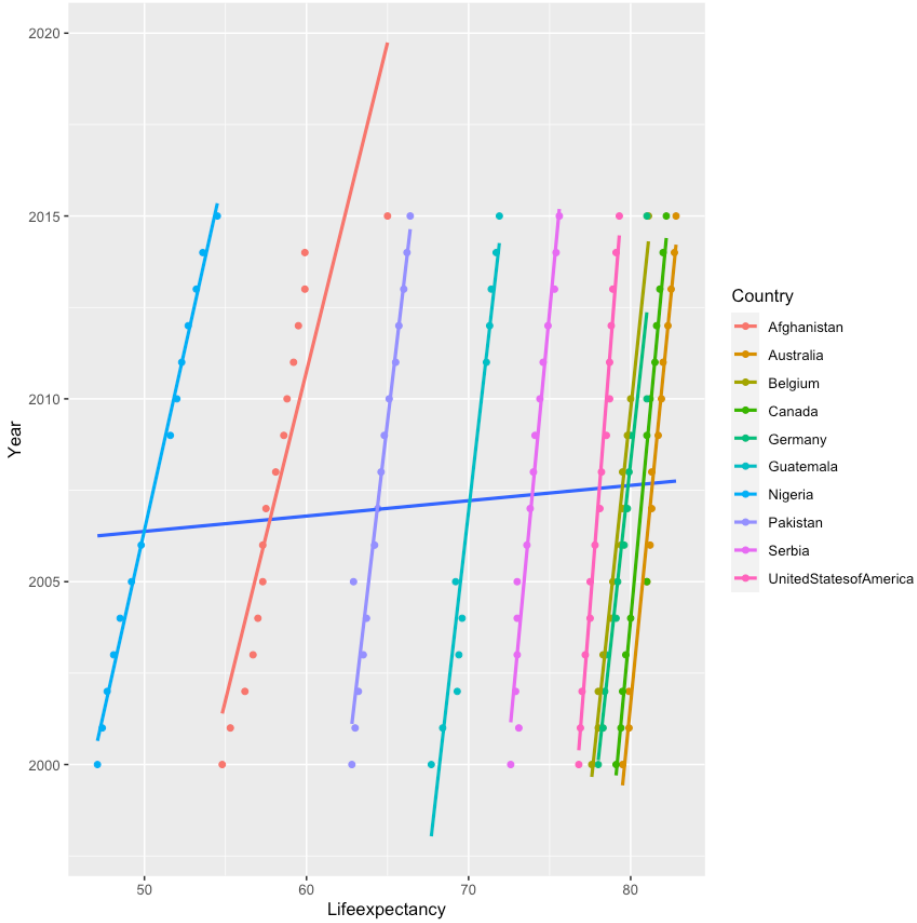

Figure 1: Life Expectancy vs. Year

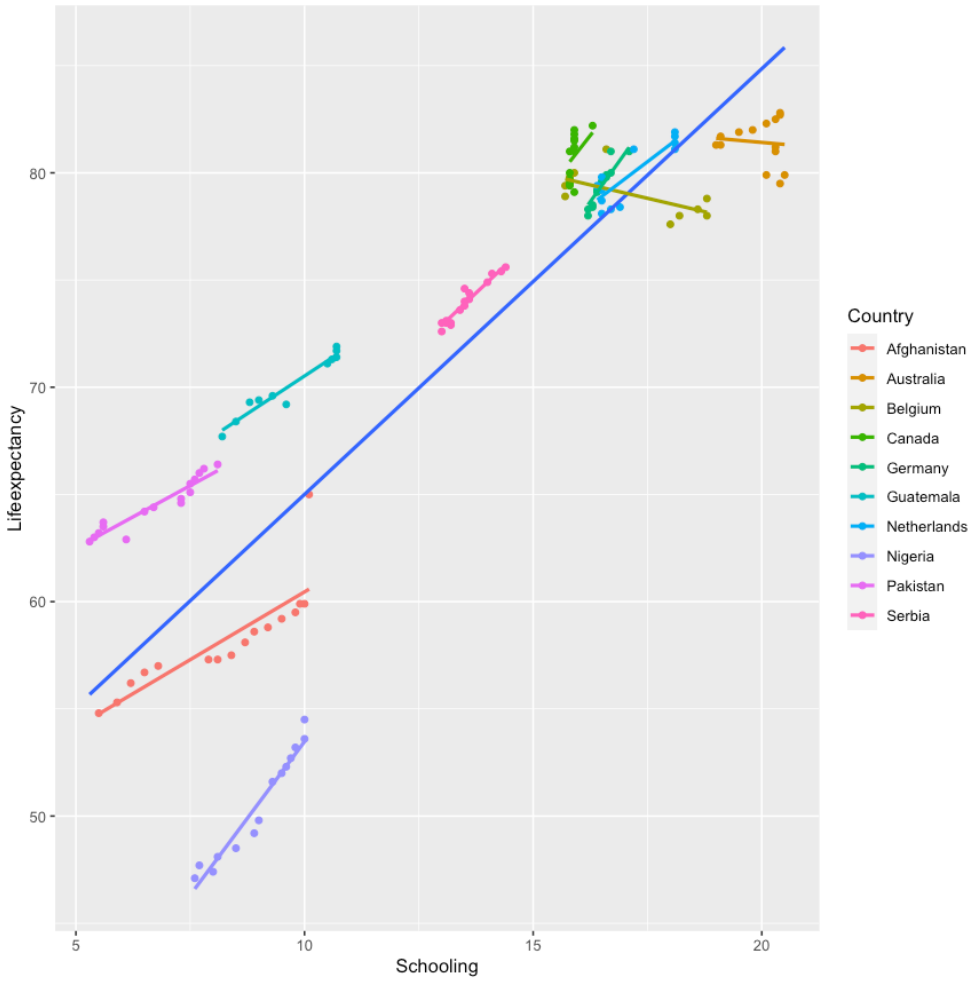

Figure 2: Schooling vs. Life Expectancy