Understanding What is important in Allowing Us to Live Longer:
A Multiple Linear Regression Analysis on Life Expectancy

Alexander Martinez

The University of Arizona

Spring 2022

Abstract

Using a dataset that can be found on *Kaggle,* a Regression model was created to describe the average life expectancy in years based on data that was recorded between 2000 and 2012 of 183 countries around the planet. Life expectancy has increased tremendously over the last century due to breakthroughs in medicine, technology, diets, many other observable factors. Understanding the factors that dictate life expectancy can be beneficial to countries across the globe by giving light to the trends and patterns that are beneficial to their population. The analysis done in this paper attempts to recognize the political, social, and economic characteristics that enhance the health of a society, and it also calls attention to key insights that may support or be contrary to the current school of thought. The research performed in this paper defines components that have had a positive– and negative – effect on human life expectancy. This analysis aims to provide and analyze what the effects of monetary, social, and cultural practices have on life expectancy.

Point of Research – Background

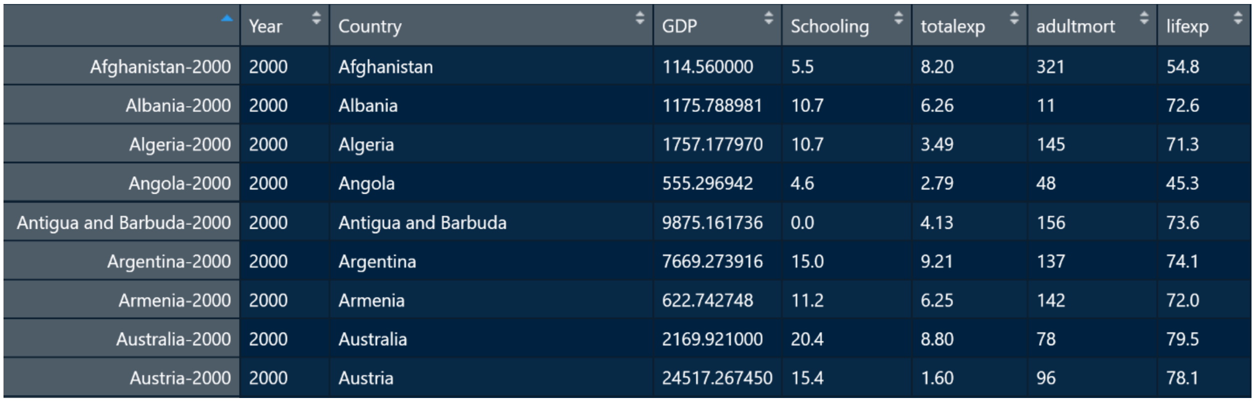
The World Health Organization (WHO), the organization responsible for the underlying dataset in this analysis, was founded in 1948, a few years after the creation of the United Nations. The World Health Organization is one of the most renowned health-related entities in the world, and in the organization's constitution they state their objective is to strive for "the attainment by all peoples of the highest possible level of health" (WHO, 1946). Because of their creation, there have been an exponential increase of efforts to distribute healthcare across the globe. Health problems that effect life expectancy come in numerous, distinct forms, namely, in the categories of disease and preexisting medical conditions. Moreover, they are caused by countless factors, such as obesity, environmental damage, poverty, malpractice, violence, and much more. These are all issues that are worth mentioning, and efforts to study these should not and will not come to a halt.

However, the analysis in this paper will study the external factors on life expectancy, and According to the OECD in their book, *Health at a Glance*, there are serious social, political, and economic contributors to life expectancy, and it has been a primary focus for research in addition to the aforementioned medical factors. The second chapter of the book begins by stating, “Countries with higher national income and health spending tend to have longer life expectancies.” (OECD, 2017). The goal is to gain insights on whether these external factors are worth allocating more of our resources towards, to achieve more knowledge on other indirect forces of life expectancy, and to ultimately shed a new light on the bigger picture on creating a more secured environment.

Data Retrieval and Description

The data was retrieved from Kaggle, a database community where analysts, data scientists, and interested persons interact with a wide array of datasets. The website is also responsible for hosting global competitions to allow beginners or data experts gain valuable insights on data by using machine learning techniques. The data was originally provided by the World Health Organization and initially had 22 features. The data consisted of 16 observations per country, each observation representing a different year from 2000 – 2015, and for the purposes of this analysis, the data was condensed based on social, economic, and political features. Additionally, to take advantage of the statistical and econometric attributes that come with it, which we will detail in the next few sections, the data was reformatted into panel data.

With panel data, it is best to have it balanced, and since smaller countries like Togo and Cabo Verde did not have recorded data in specific years, they were taken out of the dataset. Also, it is important to note that some of the years, such as 2015, did not have much data for several countries, so those years were not included in the data. As a solution, the data was split into intervals of two, starting in the year 2000. In other words, the data has an one observation for 183 countries, instead of the original 193 countries (*i = 183*) in 2000, 2002,…,2014 (*t = 8*) for a total of n = 1,464 observations (*8 x 183 = 1,464*). A snap shot of the what the dataset looks like can be viewed below.



	Year	Country	GDP	Schooling	totalexp	adultmort	lifexp
Afghanistan-2000	2000	Afghanistan	114.560000	5.5	8.20	321	54.8
Albania-2000	2000	Albania	1175.788981	10.7	6.26	11	72.6
Algeria-2000	2000	Algeria	1757.177970	10.7	3.49	145	71.3
Angola-2000	2000	Angola	555.296942	4.6	2.79	48	45.3
Antigua and Barbuda-2000	2000	Antigua and Barbuda	9875.161736	0.0	4.13	156	73.6
Argentina-2000	2000	Argentina	7669.273916	15.0	9.21	137	74.1
Armenia-2000	2000	Armenia	622.742748	11.2	6.25	142	72.0
Australia-2000	2000	Australia	2169.921000	20.4	8.80	78	79.5
Austria-2000	2000	Austria	24517.267450	15.4	1.60	96	78.1

Figure 1: First 9 rows of the dataset showing the year 2000

Response and Predictor Variables

The visual of the dataset shows not only the format of the data, but it also shows the features that were selected. To begin with the independent variables, GDP is the acronym for Gross Domestic Product, and it is measured per capita in US dollars. Schooling is the average number of years in education. The variable totalexp is defined as the general government expenditure on health as the percentage of total government expenditure (%). The adultmort variable describes the Adult Mortality Rates of both sexes (probability of dying between 15 and

60 years per 1000 population). The response variable, of course, is lifexp, which is the average life expectancy, measured in years. Finally, this is panel data, which means that the data is indexed by group (Country) and by time (Year).

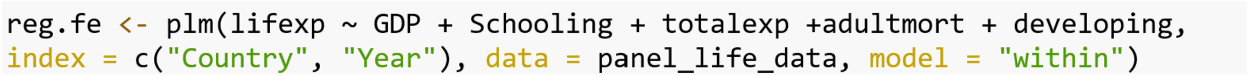
Selecting the Model

Dealing with panel data does have its advantages, and regression on panel data “may have the ability to mitigate omitted variable bias” (Zach, 2021). Importantly, the data used has an equal amount of time periods for all groups i.e., the data is balanced. There are unobservable factors that are distinct from country to country that may be present in the data, but an assumption can be made that these are constant across observations. It is incorrect to simply run a regression on the difference from the first time period (2000) to the last to the last (2014), while ignoring the likely important, unobserved factors. Since there are seemingly significant factors included in the years between the two time periods, there is a useful method known as fixed effects that allows for the addition of these control variables. In essence, the fixed effects model can be written as,

$$Y_{it} = \beta_1 X_{1,it} + . . . + \beta_k X_{k,it} + \alpha_i + u_{it}$$

with $i = 1, . . . , n$ and $t = 1, . . . , T$. The α_i are group-specific intercepts that account for heterogeneities throughout groups themselves. The idea behind fixed effects is when we use the “within” transformation, we get rid of the unobserved individual effect α_i to get closer to getting rid of omitted variable bias.

This a regression that is simple to perform with specific R software packages, using the “within” method found in the *plm()* function. Here, we must specify the indexes of country and year, and the formula is inputted like so:



```
reg.fe <- plm(lifexp ~ GDP + Schooling + totalexp +adultmort + developing,
index = c("Country", "Year"), data = panel_life_data, model = "within")
```

Figure 2: The Fixed Effect Model as Inputted in R

Reg.fe is the object that is being saved as our regression model. *Plm()* works just like *lm()*, but it recognizes that the data is panel data, and it treats the data accordingly. *Index* is the individual entity and time entity that is required to have panel data in the first place, and *panel_life_data* is the object that we assigned the dataset to. The last part, *model*, specifies the kind of model we want to use. Since fixed effects is performed with the "within" estimator, we use this option.

Assumptions and Standard Errors

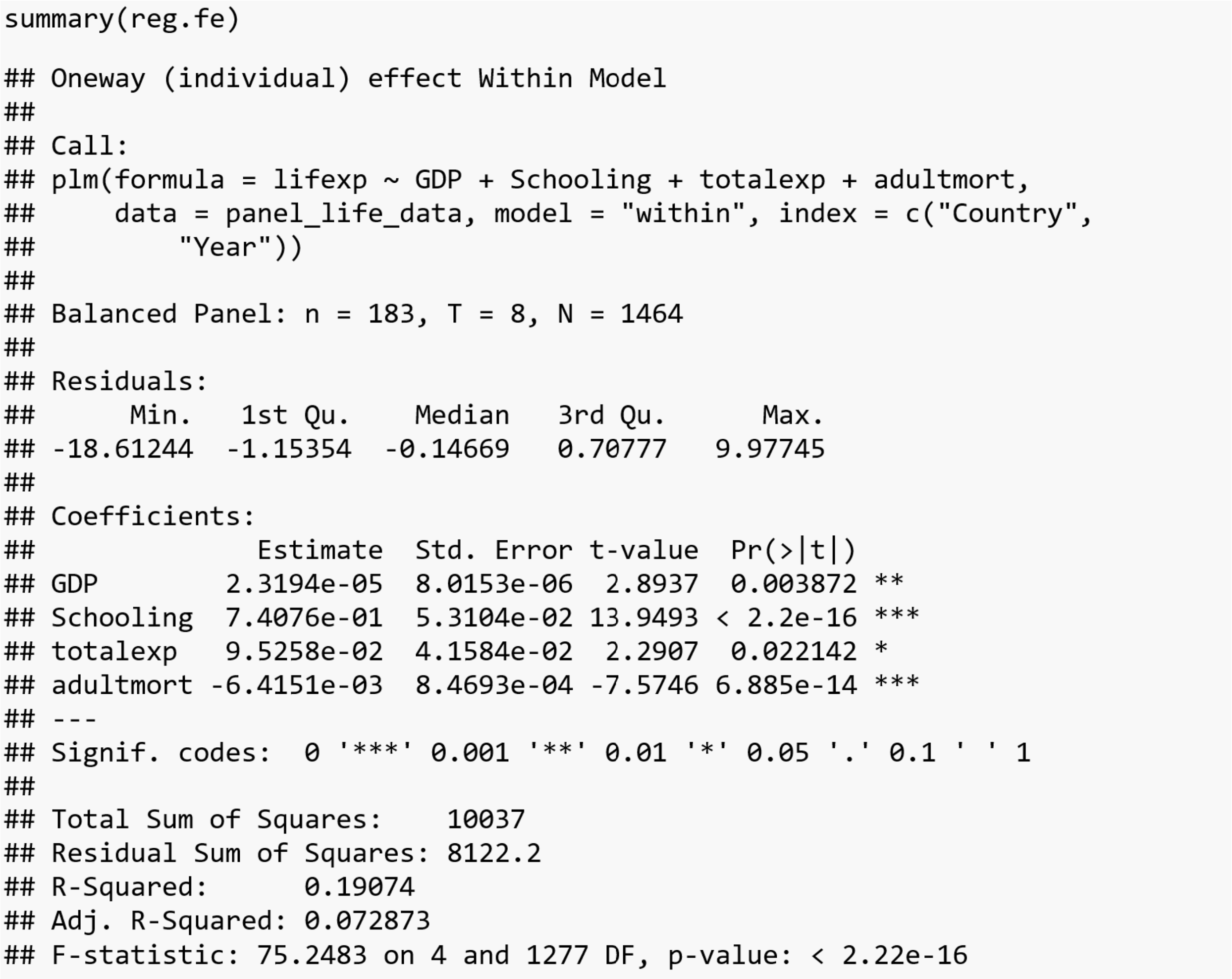
Like the Ordinary Least Squares model that many are familiar with, there are assumptions that are made when using this model. The list of these assumptions, in no order, are as follows, (retrieved from Econometrics with R)

1. The error term u_{it} has conditional mean 0, i.e., $E\ (u_{it} \mid X_{i1}, X_{i2}, . . . X_{iT})$.
2. $(X_{i1}, X_{i2}, . . ., X_{i3}, u_{i1}, . . ., u_{it}), i = 1, . . ., n$ are i.i.d. draws from their joint distribution.
3. Large outliers are unlikely, the fourth moment exists.
4. There is no perfect multicollinearity

The first assumption is a significant advantage that becomes more plausible with the fixed effects method on panel data; however, it is still undoubtedly difficult to be certain that omitted variable bias does not exist. On top of this, to make inferences, it is important to be accurate while doing so, and bias estimates could make or break a significance test. R will not initially take this into account; therefore, we need to acquire heteroskedasticity-robust standard errors to combat it. The next section will provide the model, test for heteroskedasticity, and find the heteroskedasticity-robust standard errors.

Evaluating the Model

Now that the model has been defined, the idea is to analyze the effects of the social, political, and economic climate on life expectancy. We do this by running the summary of the regression shown in Figure 2, and we want to see if the features that we had chosen had any significant effect on life expectancy. The model summary output is as follows:

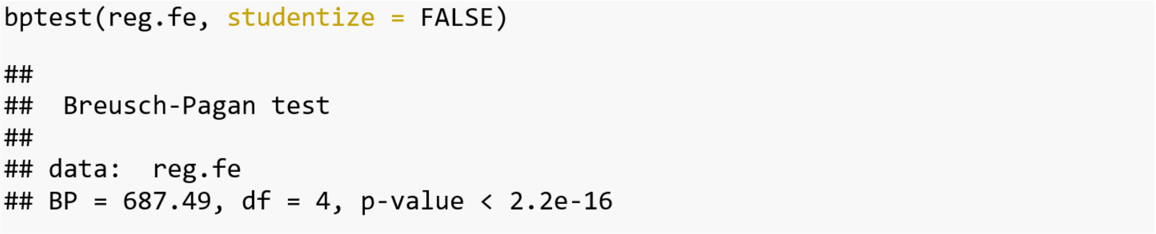


```
summary(reg.fe)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lifexp ~ GDP + Schooling + totalexp + adultmort,
##     data = panel_life_data, model = "within", index = c("Country",
##         "Year"))
##
## Balanced Panel: n = 183, T = 8, N = 1464
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -18.61244  -1.15354  -0.14669   0.70777   9.97745
##
## Coefficients:
##              Estimate  Std. Error t-value  Pr(>|t|)
## GDP        2.3194e-05  8.0153e-06  2.8937  0.003872 **
## Schooling  7.4076e-01  5.3104e-02 13.9493 < 2.2e-16 ***
## totalexp   9.5258e-02  4.1584e-02  2.2907  0.022142 *
## adultmort -6.4151e-03  8.4693e-04 -7.5746 6.885e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    10037
## Residual Sum of Squares: 8122.2
## R-Squared:      0.19074
## Adj. R-Squared: 0.072873
## F-statistic: 75.2483 on 4 and 1277 DF, p-value: < 2.22e-16
```

Figure 3: One-way Individual effect within model

Right away, we find that the model finds all the predictors to be significant, this is confirmed with the F-statistic showing overall model significance at the bottom. Now, while this may be convincing, we must be extra careful with what inferences we make about our data. First, we should confirm that heteroskedasticity does exist within the model. In most fixed effects cases this will be the case, but it is always important to cement any suspicions we may have in our analysis. A great way to do this is by performing a Breusch-Pagan test, of which the null hypothesis is that there is homoskedasticity, which is performed like so:



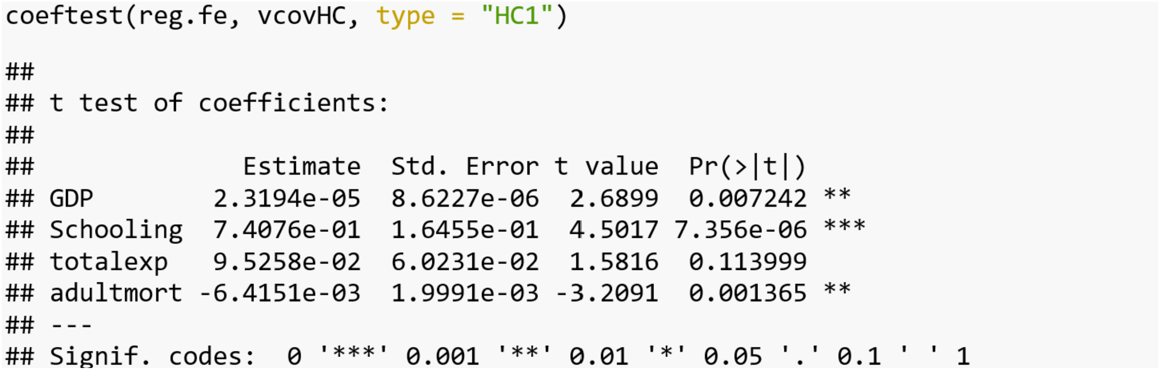
```
bptest(reg.fe, studentize = FALSE)

##
##  Breusch-Pagan test
##
## data:  reg.fe
## BP = 687.49, df = 4, p-value < 2.2e-16
```

Figure 4: Heteroskedasticity Test Performed in R

As we can see, we reject the null and do, in fact, conclude that there is heteroskedasticity.

To adjust accordingly, we acquire a summary that includes the appropriate standard errors. The process of doing this is simple in R, and can be done with the simple code using the *coeftest()* function found in the *lmtest* package:



```
coeftest(reg.fe, vcovHC, type = "HC1")

##
## t test of coefficients:
##
##              Estimate  Std. Error t value  Pr(>|t|)
## GDP        2.3194e-05  8.6227e-06  2.6899  0.007242 **
## Schooling  7.4076e-01  1.6455e-01  4.5017 7.356e-06 ***
## totalexp   9.5258e-02  6.0231e-02  1.5816  0.113999
## adultmort -6.4151e-03  1.9991e-03 -3.2091  0.001365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5:Summary with Heteroskedasticity-Robust Standard Errors

vcovHC is a formula that calculate different variance covariance matrices for different kinds of standard errors. In this case, we will use *HC1*. After that, we are now ready to make sound conclusions, and if we had not considered to adjust our standard errors, we would have concluded that total expenditure on health as a percent of the total government expenditure has a significant negative effect on average life expectancy, which is not true.

Conclusion

As any model offers, there are key conclusions to draw from the model, and it begins with the evaluation of the coefficients. Holding all else constant, the model indicates that GDP per capita has a significant positive effect on life expectancy, and, for every 1 unit increase in GDP, there is an ever-so-slightly increase in life expectancy of about .00002 years. This makes sense because 1 USD will not make a huge difference, but an increase of 1,000 USD will increase life expectancy buy about .02 years. The next is schooling, which indicates, holding all else constant, a 1 unit increase in school years, will increase life expectancy of about .74 years. With the adjustment to the standard errors, total expenditure spent on health as a percent of total government expenditure does not hold statistical significance to effect life expectancy.

Lastly, holding all else constant, a 1 unit increase in adult mortalities decreases the life expectancy by about .006415 years.

If we apply common sense, we can see how these estimates came to be. If we think about GDP per capita increasing, more money will allow one to allocate more access and resources to health-related benefits, which, in turn, would increase the average life expectancy. If one receives more schooling, two things could be the reason for it increasing life expectancy. The first is the fact that if one goes to school for more years, that means, by default, they have been alive for more years, increasing life expectancy. The second is that if the society is a developed society, per se, they will most likely have access to more education, healthcare, food, etc., and that would allow them to live a longer, healthier life. Adult mortalities having a negative effect on life expectancy makes sense too because if more people are dying as an adult, the conditions may be unfriendly, leading to premature deaths, and that directly decreases the life expectancy.

Total expenditure on health not having a significant effect on life expectancy is more difficult to understand, which makes us wonder if there is a better model that may be able to account for this. It could be several issues arising, such as omitted variable bias, but it could also be that total expenditure on health does not increase the life expectancy in many cases. Countries could be misreporting, or the model could have been made to do a better job at accounting for the time and entity fixed effect in a two-way model. The dataset did include immunization data, which could be a necessary variable in a model, such as this one.

In sum, political, social, and economic do seem to be statistically significant in effecting total life expectancy, so efforts to go beyond the medical field to optimize a safer world for humanity should be encouraged. Strict policies and harsh conditions can indirectly decrease the well-being of everyday people, according to this fixed effects model, and we must highlight this important observation for the future.

Citations

Data found here: KumarRajarshi. “Life Expectancy (WHO).” *Kaggle*, 10 Feb. 2018, https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?select=Life%2BExpectancy%2BData.csv.

Christoph Hanck, Martin Arnold. “Introduction to Econometrics with R.” *10.3 Fixed Effects Regression*, 6 Oct. 2021, https://www.econometrics-with-r.org/10.3-fixed-effects-regression.html.

Constitution of the World Health organization1. https://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1.

Zach. “Omitted Variable Bias: Definition & Examples.” *Statology*, 19 Feb. 2021, https://www.statology.org/omitted-variable-bias/.

OECD Ilibrary | Health at a Glance 2017: OECD Indicators. https://www.oecd-ilibrary.org/docserver/health_glance-2017-en.pdf.

Heiss, Florian. *Using R for Introductory Econometrics*. Florian Heiss, 2020.