Final Report

due November 16, 2021 by 11:59 PM

The Gre8est Team: Roni Ochakovski, Matthew Wang, Judy Zhong

11/16/2021

Background

Numerous studies have analyzed the correlation between health spending and health outcomes, consistently finding a positive relationship between the two [1]. Other studies have investigated how GDP and educational attainment are associated with health outcomes and found that both are positive predictors of health [2]. Our study looks to build on past findings and conduct a multivariate analysis to better inform global leaders on focus areas for improving global health.

Like similar studies, we have chosen to analyze life expectancy and under-5 mortality rate as measures of health [1]. Period life expectancy at birth is often used as a measure of the overall health of a population. It is derived from the probabilities of people of certain age groups dying given the mortality rates of those age groups over a specific time frame. The probabilities are then used in a survival function to project the average age of death of a newborn of that time period [3]. Meanwhile, under-5 mortality rate reflects the probability of a child born in the year in question dying before the age of 5. It is represented as the number of predicted deaths per 1,000 live births [4].

Research Question

This analysis aims to determine the significance of relationships between a set of World Development Indicators and health outcomes, measured by life expectancy and under-5 mortality rate.

Data

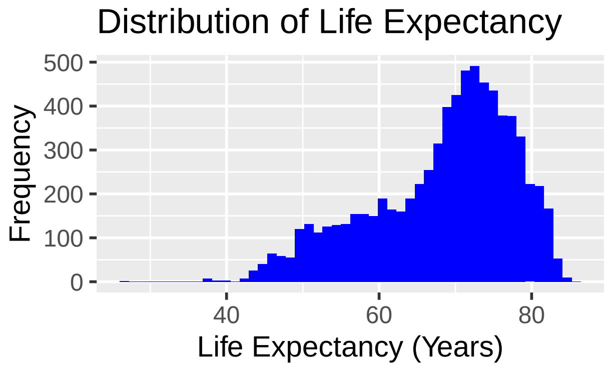
We are combining health spending data from the Global Health Data Exchange [5] and world economic/health-related data from the World Bank [6], ranging from 1990-2020. For these data sets, we will be analyzing data from the 204 countries and territories that are included in both databases. The health spending data was collected from a wide variety of sources that included program reports, budget data, national estimates, and National Health Accounts (NHAs). The variables that we are most concerned with are location ID, location name, year, and total health expenditure (THE) per capita in purchasing power parity (PPP) dollars (2020 USD). Purchasing power parity accounts for the differences in economic and standards of living between countries. The World Bank collects data through various sources, mostly through national statistical systems of member countries. The variables that we are most concerned with are country or area, total life expectancy at birth (years), male life expectancy at birth (years), female life expectancy at birth (years), GDP (current US dollars), GDP growth (annual %), income share held by lowest 20%, mortality rate under 5 (per 1,000 live births), poverty headcount ratio at national poverty lines (% of population), and educational attainment, at least completed upper secondary, population 25+, total (%) (cumulative) [6].

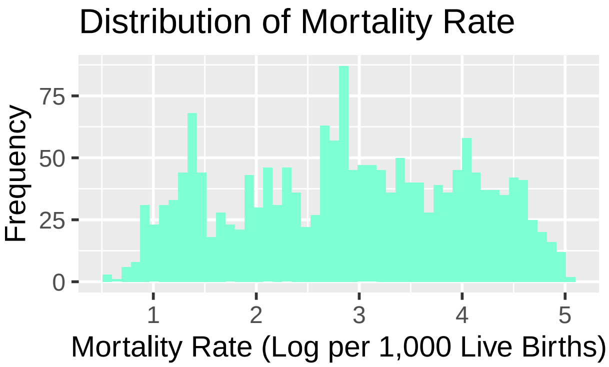
Descriptive Stats

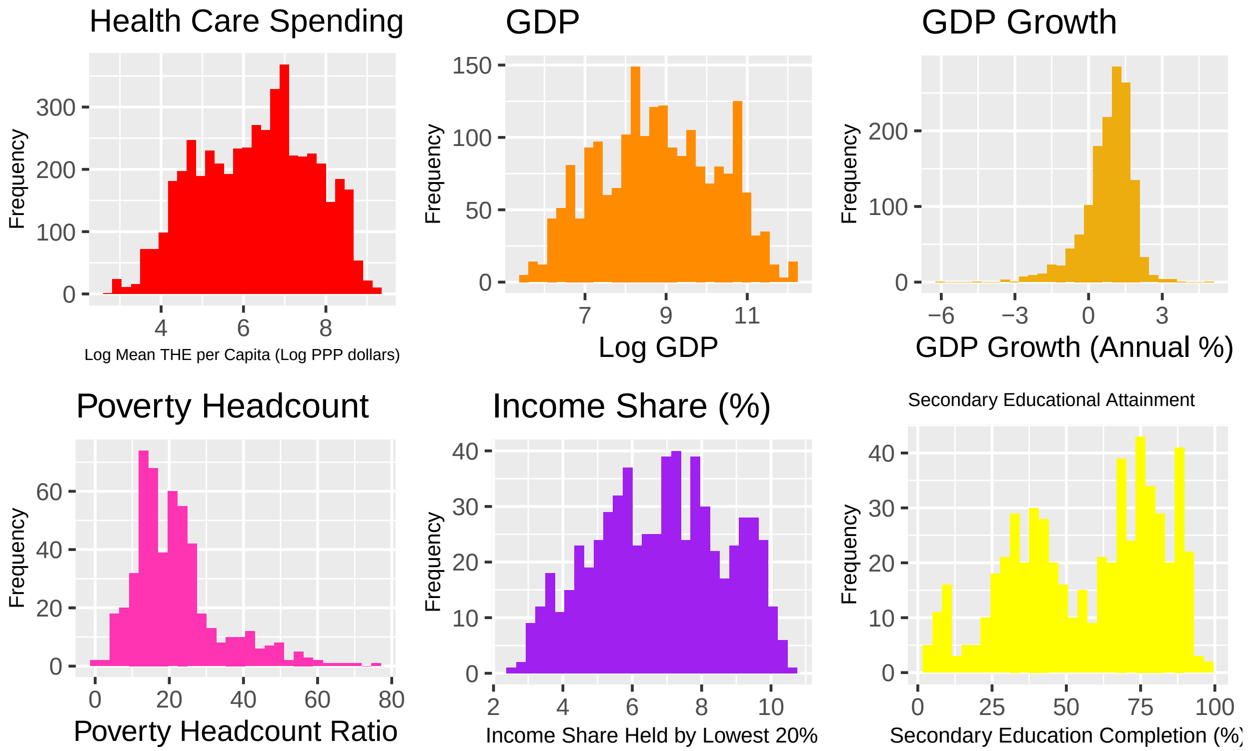
Following the data-cleaning, we ran some descriptive statistics to visualize the data. The first thing we noticed was that healthcare spending variation was extremely large across countries. To standardize the numbers across countries, we attempted to look at just the healthcare spending per person, but even that yielded huge variation across countries:

```
##  the_per_cap_ppp_mean
##  Min.   :   16
##  1st Qu.:  179
##  Median :  603
##  Mean   : 1207
##  3rd Qu.: 1534
##  Max.   :11027
```

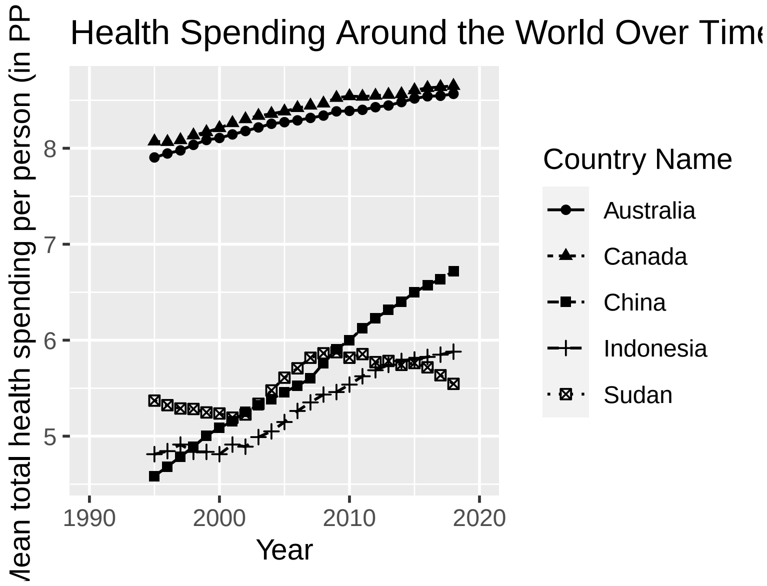
As you can see, the quartiles have enormous range between them. To standardize and reduce variation further, we took the log of the spending per person, and got an approximately normal distribution of healthcare spending across countries. We followed the same procedure to normalize mortality rate, healthcare spending, and GDP. We did so to satisfy any future assumptions given our relatively small sample size. The approximately normal distributions are shown below:



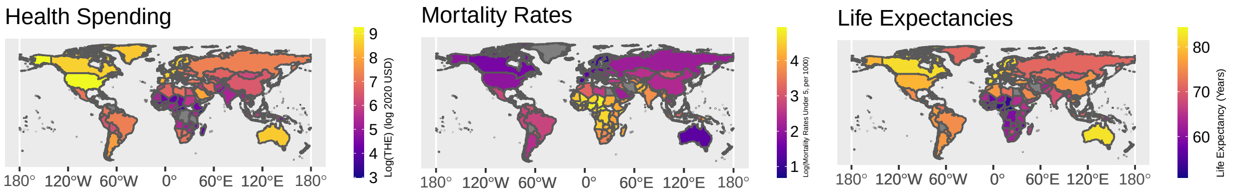




Then, we looked at healthcare spending over time in a few selected countries to see any trends in the data:

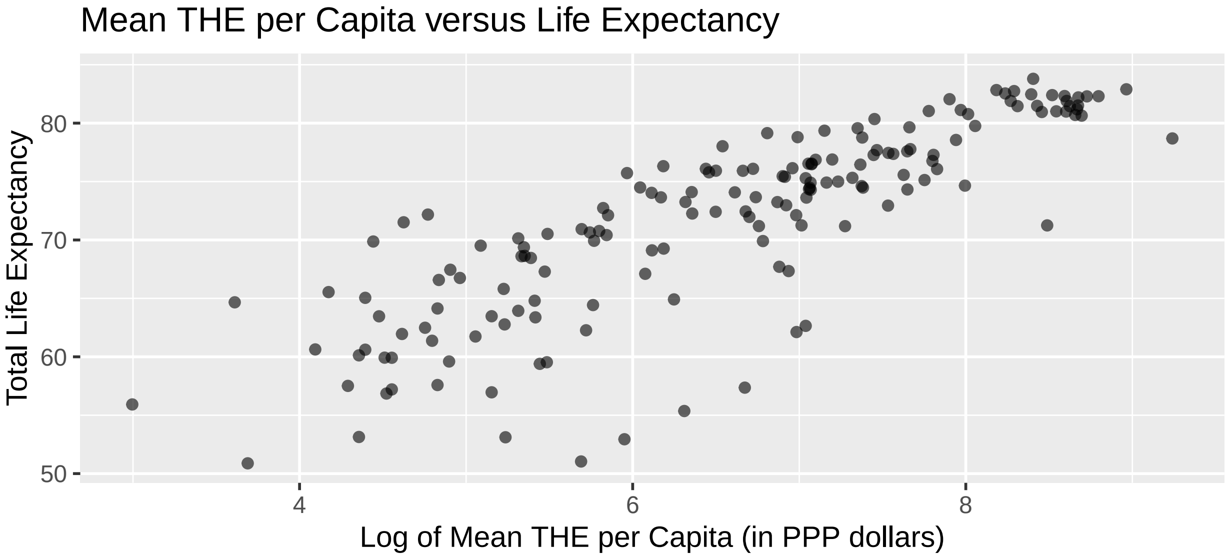


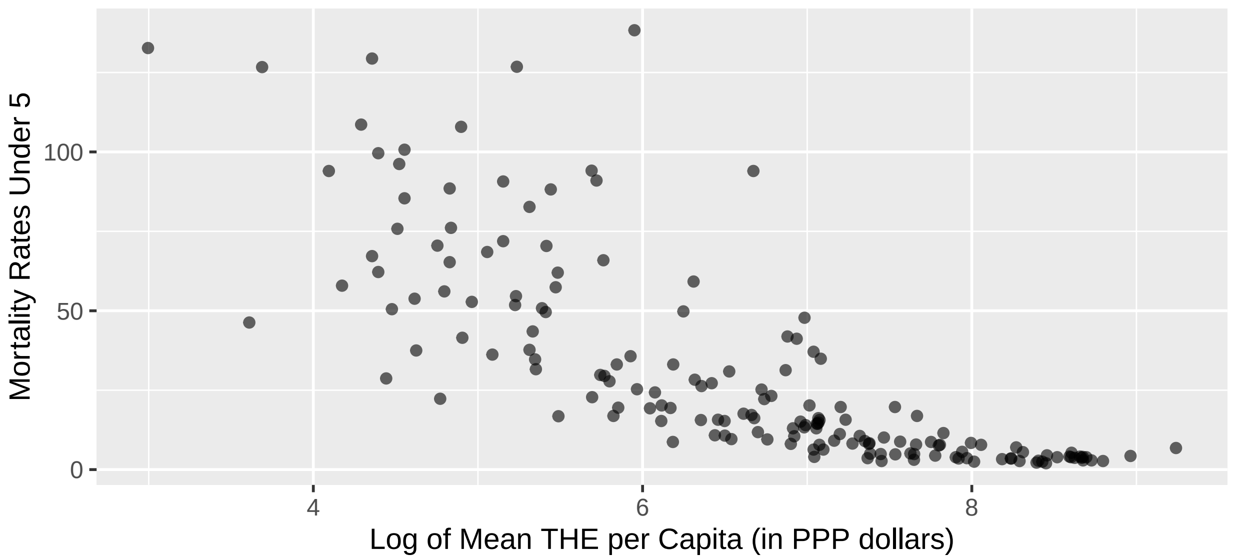
We found that healthcare spending generally increased over time in developed countries and was also significantly larger in Western than Eastern countries. This would be confirmed when we mapped health care spending. While mapping, we also mapped what would later be one of our response variables: life expectancy. We wanted to visualize life expectancy across countries to see any geographic patterns:


Health Spending
9
8
7
6
5
4
3
Log(THE) (log 2020 USD)
180° 120°W 60°W 0° 60°E 120°E 180°
Mortality Rates
4
3
2
1
Log(Mortality Rates Under 5, per 1000)
180° 120°W 60°W 0° 60°E 120°E 180°
Life Expectancies
80
70
60
Life Expectancy (Years)
180° 120°W 60°W 0° 60°E 120°E 180°

We saw the same pattern as in the line charts where Western countries had higher healthcare spending,higher life expectancy, and lower mortality rates. The same trend was generally true for developed compared to non-developed regions. We also began to notice the problem of missingness in our dataset where mortality rates and life expectancies for some countries (most notably the U.S.) were absent.

Finally, we wanted to visualize the relationship between healthcare spending and our health outcomes (being mortality rates under 5 and life expectancy). We used scatterplots to do so:





There is a clear positive linear relationship between healthcare spending and life expectancy but a more exponential-decay trend in healthcare spending and mortality rates under 5. Viewing these trends would come in helpful when modelling. Within the descriptive statistics portion of our project, we also explored healthcare spending by GDP, distributions of mortality rates, life expectancies, and fertility rates, male vs. female life expectancy distributions, and line graphs of health outcomes over time. However, the most significant graphs and visualizations are included above as these helped the most in understanding our data set and fitting the proper models. In addition to this, some outcomes such as fertility rates were excluded after data visualization due the missingness in that variable being to great to deal with.

Modelling

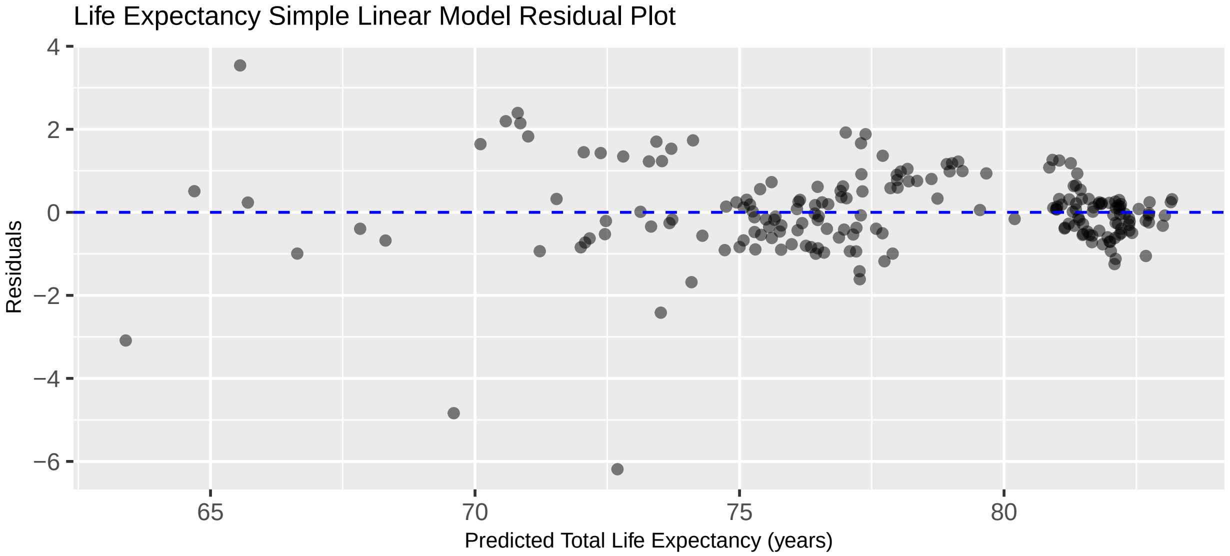
In our models, we decided to explore two commonly used health outcome measures: life expectancy and mortality rate under the age of 5. For each health outcome, we constructed 3 different models: a simple linear model with all the covariates we had identified, a quadratic model with only the covariates that increased adjusted Rˆ2, and a quadratic model with interaction effects using only the covariates that increased adjusted Rˆ2. Adjusted Rˆ2 and residual plots were used for assessing and selecting models.

Life Expectancy Simple Linear Model

Total Life Expectancy $= \beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{GDP}) + \beta_3 * \text{GDP Growth} + \beta_4 * \text{Income Share} + \beta_5 *$ Poverty $+ \beta_6 *$ Education

Term	Value	p-value	CI Lower	CI Higher
β_0	54.010778	0.000000	48.756156	59.265400
β_1	4.321537	0.000000	3.933269	4.709804
β_2	-0.086769	0.323785	-0.259771	0.086233
β_3	-0.071371	0.215534	-0.184661	0.041918
β_4	-0.232740	0.030374	-0.443188	-0.022292
β_5	-0.103055	0.000002	-0.144149	-0.061961
β_6	-0.045704	0.000000	-0.060810	-0.030599

The R^2 for the life expectancy simple linear model is 0.8387499 which means that 0.8387499 of the variation in life expectancy is explained by the model. The adjusted R^2 for the life expectancy simple linear model is 0.8337108 which means that 0.8337108 of the variation in life expectancy is explained by the model, adjusting for the number of variables included.



This residual scatter plot is concerning because the points in the plot appear to decrease for higher predicted life expectancies and thus are not randomly and evenly distributed around 0. There is also some clustering of the residual points occurring around 82.5 years in predicted total life expectancy.

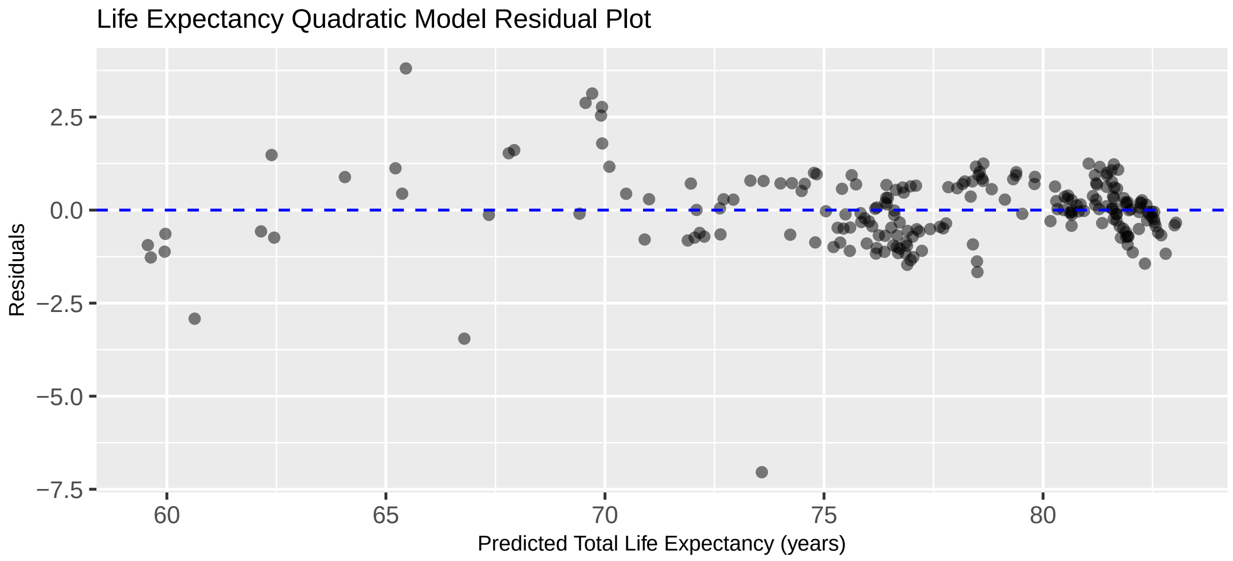
Life Expectancy Quadratic Model

Total Life Expectancy = 38.837594399 + 7.060190898*log(THE per capita in PPP dollars) - 0.215246470*(log(THE per capita in PPP dollars))^2 - 0.131323652*log(GDP) - 0.083085644*Poverty headcount ratio + 0.190683904*Educational attainment - 0.002025324*Educational attainment^2

Total Life Expectancy $= \beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{THE})^2 + \beta_3 * log(\text{GDP}) + \beta_4 * \text{Poverty} + \beta_5 * \text{Education} + \beta_6 * \text{Education}^2$

Term	Value	p-value	CI Lower	CI Higher
β_0	38.837594	0.000000	26.685308	50.989881
β_1	7.060191	0.000007	4.033017	10.087365
β_2	-0.215246	0.042947	-0.423594	-0.006899
β_3	-0.131324	0.155861	-0.313101	0.050454
β_4	-0.083086	0.000001	-0.114931	-0.051240
β_5	0.190684	0.000008	0.108539	0.272829
β_6	-0.002025	0.000000	-0.002725	-0.001325

The R^2 for the life expectancy quadratic model is 0.8997453 which means that 0.8997453 of the variation in life expectancy is explained by the model. The adjusted R^2 for the life expectancy quadratic model is 0.896811 which means that 0.896811 of the variation in life expectancy is explained by the model, adjusting for the number of variables included. The higher adjusted R^2 of this quadratic model indicates that it is a better model than the simple linear model.



This residual scatter plot is an improvement from the linear model as the points are more randomly and evenly distributed around 0, and there is less correlation between predicted value and residual. The residual points also seem to be more tightly centered/closer to 0. However, there still appears to be clustering around 82.5 predicted years as well as around 77.5 predicted years.

Life Expectancy Quadratic Interaction Model

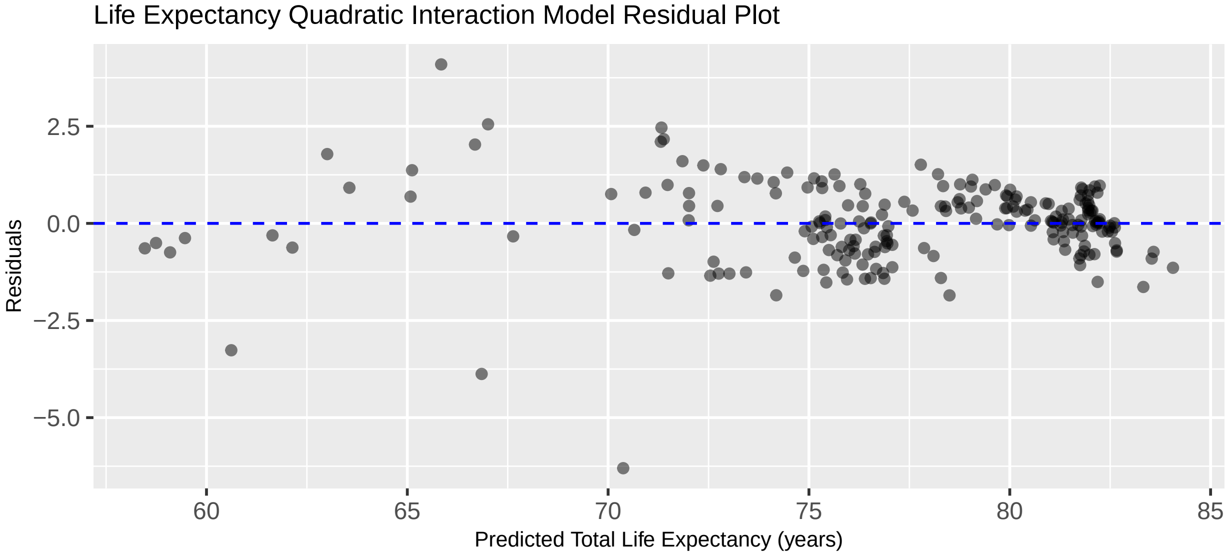
Total Life Expectancy = 59.46 - 2.22*log(THE per capita in PPP dollars) + 0.27*(log(THE per capita in PPP dollars))^2 + 0.46*log(GDP) -0.18*Poverty headcount ratio* + 0.46*Educational attainment - 0.002*Educational attainment^2 -0.313*GDP growth - 0.00495*Poverty headcount ratio * Educational attainment + 0.0059*Educational attainment * GDP growth - 0.022*log(GDP) * Poverty headcount ratio - 0.00597*log(GDP) * Educational attainment + 0.133*log(THE per capita in PPP dollars) * Poverty headcount ratio

Total Life Expectancy = $\beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{THE})^2 + \beta_3 * log(\text{GDP}) + \beta_4 * \text{Poverty} + \beta_5 * \text{Education} + \beta_6 * \text{Education}^2 + \beta_7 * \text{GDP Growth} + \beta_8 * \text{Poverty} * \text{Education} + \beta_9 * \text{GDP Growth} * \text{Education} + \beta_{10} * \text{Poverty} * \text{GDP} + \beta_{11} * \text{GDP} * \text{Education} + \beta_{12} * log(\text{THE}) * \text{Poverty}$

From this model, of the significant and interpretable coefficients, we see that for every percent of the population that has completed at least secondary education, we expect to see life expectancy to be higher by 0.46 on average.

Term	Value	p-value	CI Lower	CI Higher
β_0	59.459507	0.000000	37.944670	80.974345
β_1	-2.220538	0.400418	-7.416808	2.975731
β_2	0.274334	0.086833	-0.040031	0.588700
β_3	0.451402	0.099723	-0.086806	0.989610
β_4	-0.181952	0.417053	-0.623153	0.259249
β_5	0.465550	0.000001	0.283619	0.647481
β_6	-0.002480	0.000000	-0.003151	-0.001809
β_7	-0.313381	0.016256	-0.568348	-0.058414
β_8	-0.004955	0.000000	-0.006597	-0.003313
β_9	0.005941	0.009379	0.001475	0.010407
β_{10}	-0.022161	0.040949	-0.043402	-0.000921
β_{11}	-0.005979	0.073191	-0.012526	0.000567
β_{12}	0.133103	0.000001	0.080531	0.185674

The R^2 for the life expectancy quadratic interaction model is 0.9203884 which means that 0.9203884 of the variation in life expectancy is explained by the model. The adjusted R^2 for the life expectancy quadratic interaction model is 0.9155877 which means that 0.9155877 of the variation in life expectancy is explained by the model, adjusting for the number of variables included. The higher adjusted R^2 of this quadratic interaction model indicates that it is a better model than the simple linear model and quadratic model.



This residual scatter plot is an improvement from the quadratic model as the points are appear to less clustered and thus are more evenly distributed across the predicted years. However, the residual points seem to be less tightly centered/close to 0.

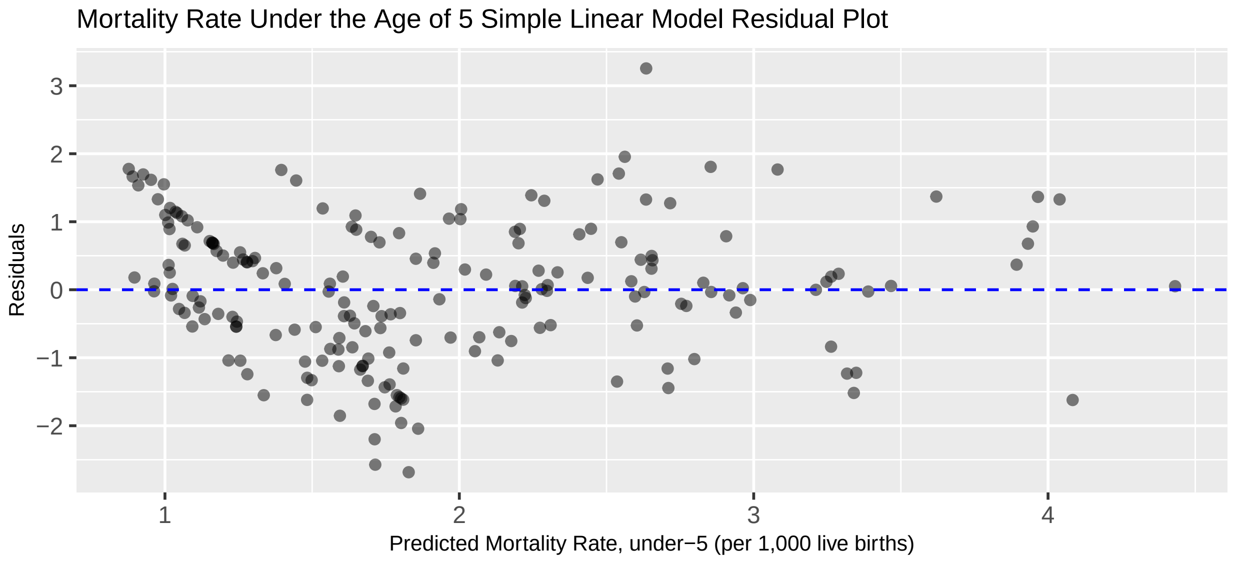
Mortality Rate Under the Age of 5 Simple Linear Model

log(Mortality Rate Under the Age of 5) = 4.629698850 - 0.584672973*log(THE per capita in PPP dollars) + 0.072586447*log(GDP) + 0.046301152*GDP growth - 0.010565241*Income share held by lowest 20% + 0.010933283*Poverty headcount ratio* - 0.007591069*Educational attainment

$log(\text{Under-5 Mortality} = \beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{GDP}) + \beta_3 * \text{GDP Growth} + \beta_4 * \text{Income Share} + \beta_5 * \text{Poverty} + \beta_6 * \text{Education}$

Term	Value	p-value	CI Lower	CI Higher
β_0	4.629699	0.000000	3.693192	5.566205
β_1	-0.584673	0.000000	-0.653872	-0.515474
β_2	0.072586	0.000006	0.041753	0.103420
β_3	0.046301	0.000011	0.026110	0.066492
β_4	-0.010565	0.579134	-0.048072	0.026942
β_5	0.010933	0.003635	0.003609	0.018257
β_6	-0.007591	0.000000	-0.010283	-0.004899

The R^2 for the mortality rate under the age of 5 simple linear model is 0.8513467 which means that 0.8513467 of the variation in mortality rate is explained by the model. The adjusted R^2 for the mortality rate under the age of 5 simple linear model is 0.8467013 which means that 0.8467013 of the variation in mortality rate is explained by the model, adjusting for the number of variables included.



This residual scatter plot is satisfactory because the points in the plot appear to be evenly and randomly distributed around 0.

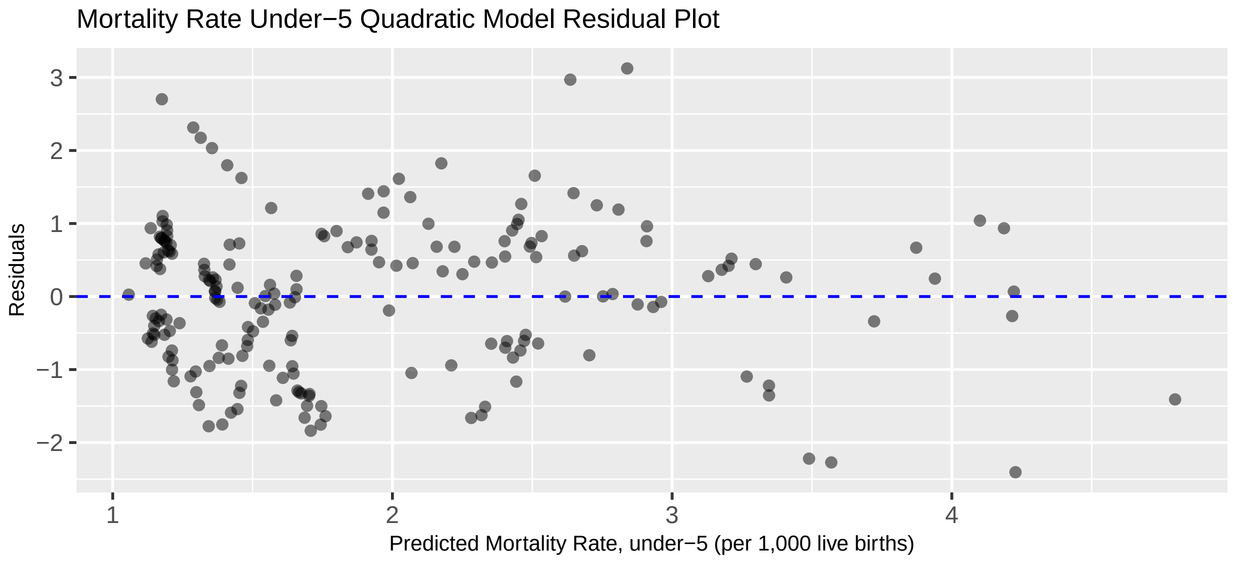
Mortality Rate Under the Age of 5 Quadratic Model

log(Mortality Rate Under the Age of 5) = 23.4435649786 - 2.5355361915*log(THE per capita in PPP dollars) + 0.1360493432*(log(THE per capita in PPP dollars))^2 - 0.5448553694*log(GDP)* + 0.0115264717*(log(GDP)*^{2} - 0.7721845449*Income share + 0.0459322976*Income share^2 - 0.0002119304*Poverty headcount ratio^2 - 0.0267254803*Educational attainment + 0.0001907620*Educational attainment^2

$log(\text{Under-5 Mortality} = \beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{THE})^2 + \beta_3 * log(\text{GDP}) + \beta_4 * log(\text{GDP})^2 + \beta_5 * \text{Income share} + \beta_6 * \text{Income share}^2 + \beta_7 * \text{Poverty}^2 + \beta_8 * \text{Education} + \beta_9 * \text{Education}^2$

Term	Value	p-value	CI Lower	CI Higher
β_0	23.443565	0.000008	13.348032	33.539098
β_1	-2.535536	0.000000	-3.270091	-1.800981
β_2	0.136049	0.000000	0.084684	0.187415
β_3	-0.544855	0.151977	-1.292064	0.202353
β_4	0.011526	0.119485	-0.003011	0.026064
β_5	-0.772185	0.000000	-1.029831	-0.514538
β_6	0.045932	0.000001	0.027880	0.063985
β_7	-0.000212	0.001939	-0.000345	-0.000079
β_8	-0.026725	0.001612	-0.043200	-0.010251
β_9	0.000191	0.007784	0.000051	0.000331

The R^2 for the mortality rate under the age of 5 quadratic model is 0.8789936 which means that 0.8789936 of the variation in mortality rate is explained by the model. The adjusted R^2 for the mortality rate under the age of 5 quadratic model is 0.8732314 which means that 0.8732314 of the variation in mortality rate is explained by the model, adjusting for the number of variables included. The higher adjusted R^2 of this quadratic model indicates that it is a better model than the simple linear model.



This residual scatter plot is slightly worse than that for the linear model. There appear to be more negative residuals at lower predicted rates and more positive residuals at the middle predicted rates. Overall, the residuals are generally randomly and evenly distributed around 0.

Mortality Rate Under the Age of 5 Quadratic Interaction Model

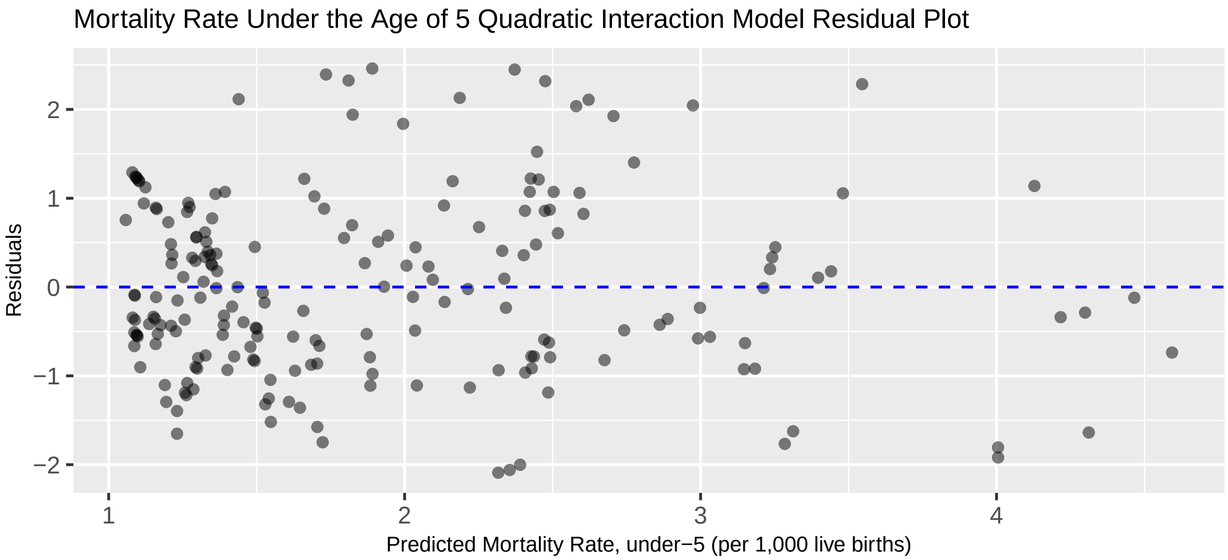
log(Mortality Rate Under the Age of 5) = 24.2 - 0.17*log(THE per capita in PPP dollars) + 0.12*(log(THE per capita in PPP dollars))^2 - 1.3*log(GDP) + 0.0041*(log(GDP))^2 - 0.16*Income share + 0.0069*Income share^2 + 0.00052*Poverty head count ratio^2 - 0.113*Educational attainment + 0.000328*Educational attainment^2 - 0.0796*Poverty head count ratio - 0.0019* Income share * Educational attainment - 0.079*log(THE per capita in PPP dollars) * log(GDP) + 0.0034*log(GDP) * Educational attainment -0.0415*log(GDP) * Income share + 0.0073*Income share * Poverty head count ratio

$log\text{Under-5 Mortality} = \beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{THE})^2 + \beta_3 * log(\text{GDP}) + \beta_4 * log(\text{GDP})^2 + \beta_5 * \text{Income} + \beta_6 * \text{Income}^2 + \beta_7 * \text{Poverty}^2 + \beta_8 * \text{Education} + \beta_9 * \text{Education}^2 + \beta_{10} * \text{GDP Growth} + \beta_{11} * \text{Poverty} + \beta_{12} * \text{Income} * \text{Education} + \beta_{13} * log(\text{THE}) * log(\text{GDP}) + \beta_{14} * \text{Income} * \text{GDP Growth} + \beta_{15} * \text{Education} * \text{GDP Growth} + \beta_{16} * log(\text{GDP}) * \text{Education} + \beta_{17} * log(\text{GDP}) * \text{Income} + \beta_{18} * \text{Income} * \text{Poverty}$

For this model, of the significant and interpretable coefficients, as the percent of the population with post-secondary education increased by 1 percent, we would expect the mortaly rate under the age of 5 to decrease by 0.11, on average. In addition to that, as poverty head count ratio increases by one, we expect the mortality rate under the age of 5 to decrease by 0.0796.

The R^2 for the mortality rate under the age of 5 quadratic interaction model is 0.9123402 which means that 0.9123402 of the variation in mortality rate is explained by the model. The adjusted R^2 for the mortality rate under the age of 5 quadratic interaction model is 0.9035743 which means that 0.9035743 of the variation in mortality rate is explained by the model, adjusting for the number of variables included. The higher adjusted R^2 of this quadratic interaction model indicates that it is a better model than the simple linear model and quadratic model.

Term	Value	p-value	CI Lower	CI Higher
β_0	24.200299	0.038997	1.235665	47.164933
β_1	-0.177918	0.792618	-1.511260	1.155424
β_2	0.119612	0.000191	0.057656	0.181569
β_3	-1.301485	0.182516	-3.220563	0.617593
β_4	0.041358	0.049115	0.000162	0.082554
β_5	-0.161374	0.757993	-1.193272	0.870523
β_6	0.069343	0.000000	0.043803	0.094883
β_7	0.000522	0.052249	-0.000005	0.001049
β_8	-0.113250	0.000443	-0.175683	-0.050817
β_9	0.000328	0.000006	0.000190	0.000467
β_{10}	0.000000	0.009027	0.000000	0.000000
β_{11}	-0.079616	0.012198	-0.141663	-0.017569
β_{12}	-0.001911	0.029713	-0.003631	-0.000190
β_{13}	-0.079274	0.000489	-0.123322	-0.035227
β_{14}	0.000000	0.003021	0.000000	0.000000
β_{15}	0.000000	0.299205	0.000000	0.000000
β_{16}	0.003432	0.004867	0.001057	0.005807
β_{17}	-0.041532	0.034617	-0.080027	-0.003038
β_{18}	0.007315	0.018890	0.001222	0.013407



This residual scatter plot is an improvement from the quadratic model as the points are appear to less clustered and thus are more evenly distributed across predicted mortality rates. The curvy shape of the residual points is also less prominent in this plot. However, the points seem to be less tightly centered/close to 0.

Conclusions

Comparing the two final models, we see one major similarity arise, whether we use mortality rates under 5 or life expectancy. As the percent of the population that has post-secondary education increases, health outcomes are expected to improve, on average. Therefore, the most significant and consistent conclusion we can take away is that there appears to be an association between educational levels in a country and their health outcomes.

Limitations

One limitation in our models is that although some of the interaction terms and normalized variables improved the models, their coefficients proved to be uninterpretable. In addition to that, the residual plot for the Life Expectancy Quadratic Interaction Model which we used for analysis had some clustering which was not ideal. The clustering was not extreme and the Rˆ2 was high enough to justify use of the model.

References

[1] Gallet CA, Doucouliagos H. The impact of healthcare spending on health outcomes: A meta-regression analysis. Soc Sci Med. 2017 Apr;179:9-17. doi: 10.1016/j.socscimed.2017.02.024. Epub 2017 Feb 20. PMID: 28237460.

[2] Lutz, Wolfgang and Endale Kebede. "Education and Health: Redrawing the Preston Curve." Population and Development Review, vol. 44, no. 2, 2018, pp. 343-61, doi:https://doi.org/10.1111/padr.12141.

[3] Luy, Marc et al. "Life Expectancy: Frequently Used, but Hardly Understood." Gerontology, vol. 66, no. 1, 2019, pp. 95-104, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7026938/.

[4] "Under-Five Mortality Rate (Probability of Dying by Age 5 Per 1000 Live Births)." Indicator Metadata Registry List. World Health Organization https://www.who.int/data/gho/indicator-metadata-registry/imr-details/7.

[5] [1] Global Burden of Disease Collaborative Network. Global Health Spending 1995-2018. Seattle, United States of America: Institute for Health Metrics and Evaluation (IHME), 2021.

[6] World Bank. World Development Indicators, The World Bank Group, 2021, https://databank.worldbank .org/source/world-development-indicators