Determinants of Health Outcomes, A Multivariate Analysis

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.

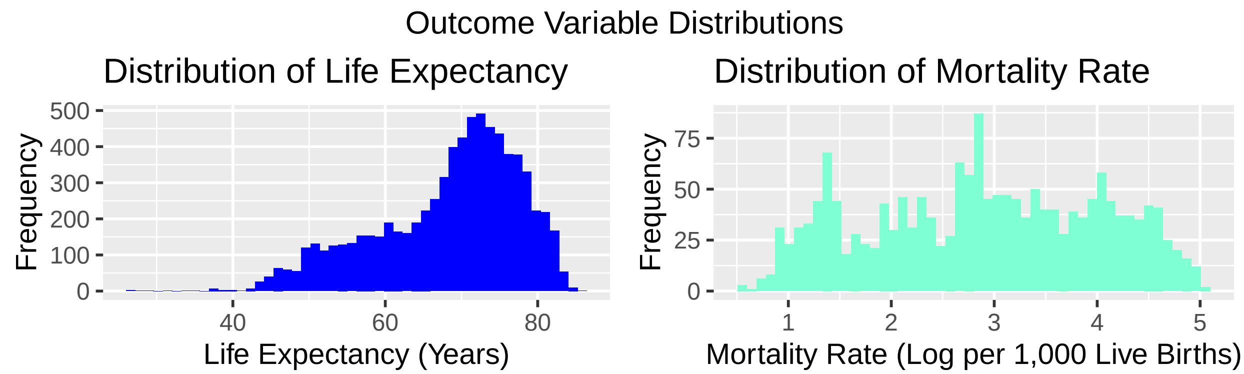
Data

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. 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].

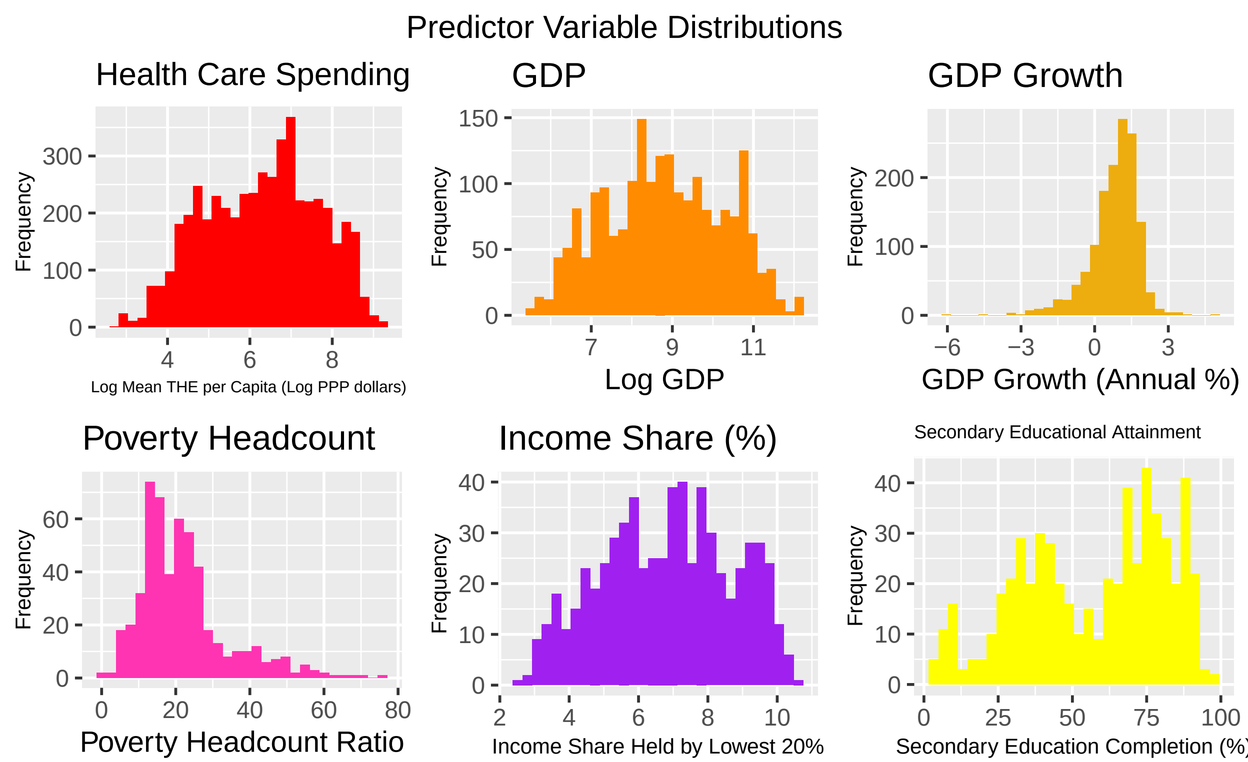
Data Wrangling and Transformations

The two data sets were modified such that the observations could be merged by country name and year. No observations were removed at this stage for missing data. Additionally, many of the measures included in the Global Health Data Exchange data set were not used but were not removed in case future analysis involved those measures.

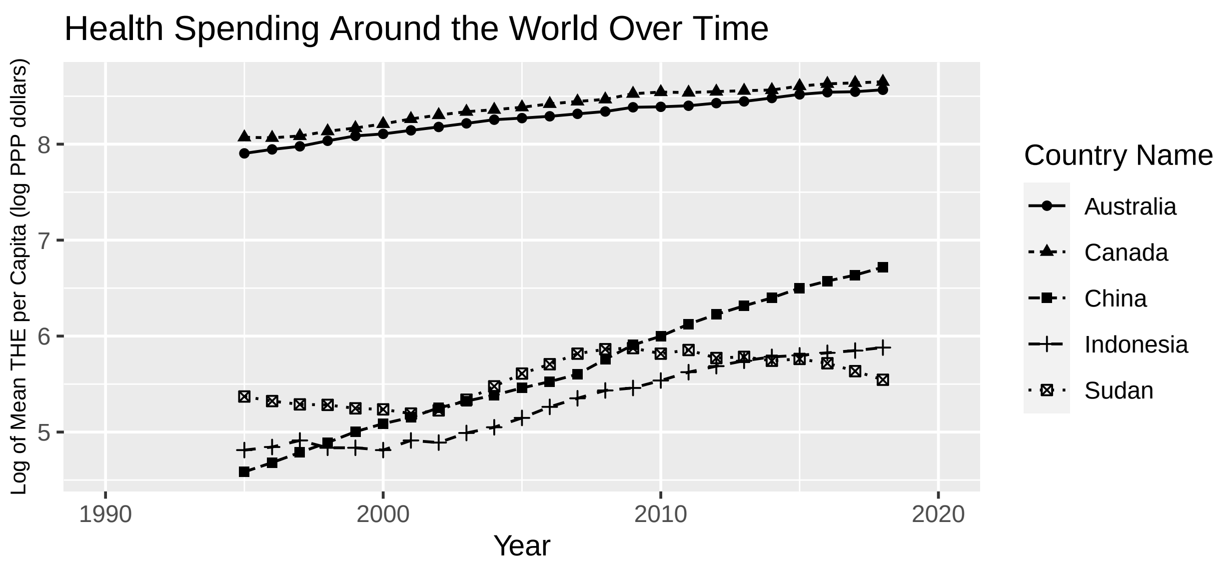
Since the analysis plan involved regression models, we looked at the distribution of our outcome variables, life expectancy and under-5 mortality rate. The life expectancy data looked approximately normal. However, the mortality rate data was log-transformed to produce a more normal distribution:



We then created visualizations of the predictor variables. Most were evenly distributed and generally resembled a normal distribution. However, THE and GDP were heavily skewed right. We found that applying a log transformation to both variables improved the results of our model.

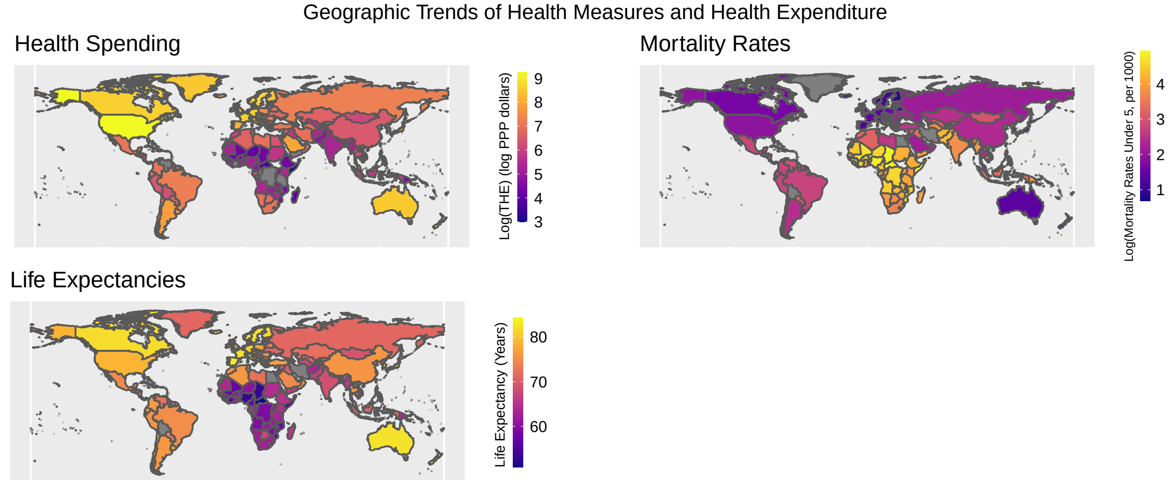


Then, we looked at healthcare spending over time in a few selected countries to see any trends in the data:



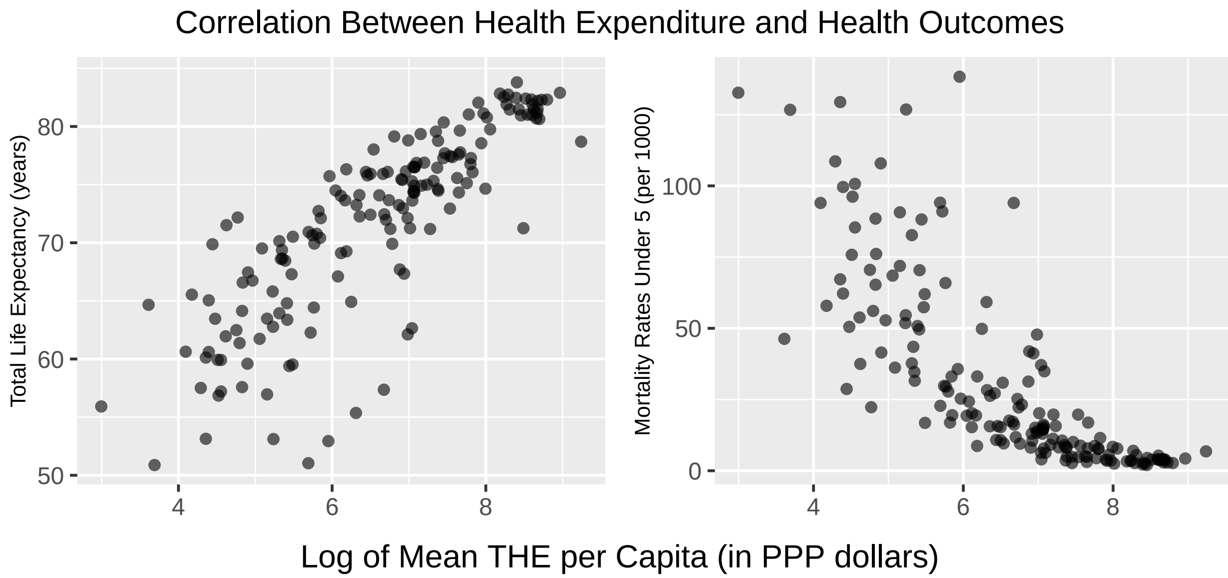
We found that healthcare spending generally increased over time in most countries, regardless of development status, and was also significantly larger in Western than Eastern countries. This would be confirmed when we mapped health care spending around the global as well as life expectancy to reveal geographic patterns./ /

Geographic Trends of Health Measures and Health Expenditure



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 were absent.

Finally, we wanted to visualize the relationship between healthcare spending and our health outcomes to gain a initial impression of whether past findings are consistent with our data:/ /



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 deciding what regression model to use, linear or polynomial.

Modeling

In our models, we decided to explore two commonly used health outcome measures: life expectancy and mortality rate under the age of 5. Since we had multiple continuous variables that were potentially correlated with the outcome variables, we decided to use regression to explore the associations. 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 R^2, and a quadratic model with interaction effects using only the covariates that increased adjusted R^2. We decided to utilize these 3 models to characterize linear, quadratic, and quadratic interaction relationships between the covariates and the health outcome variable (life expectancy and mortality rate under the age of 5). Adjusted R^2 and residual plots were used for assess each of these models and the quadratic interaction models for both health outcomes were eventually found to be the best fit.

Life Expectancy Simple Linear Model

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

	β_0	β_1	β_2	β_3	β_4	β_5	β_6
Value	54.0108	4.3215	-0.0868	-0.0714	-0.2327	-0.1031	-0.0457
p-value	0.0000	0.0000	0.3238	0.2155	0.0304	0.0000	0.0000
CI Lower	48.7562	3.9333	-0.2598	-0.1847	-0.4432	-0.1441	-0.0608
CI Higher	59.2654	4.7098	0.0862	0.0419	-0.0223	-0.0620	-0.0306

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.

Life Expectancy Quadratic Model

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

$\beta_6 * \text{Education}^2$

	β_0	β_1	β_2	β_3	β_4	β_5	β_6
Value	38.8376	7.0602	-0.2152	-0.1313	-0.0831	0.1907	-0.0020
p-value	0.0000	0.0000	0.0429	0.1559	0.0000	0.0000	0.0000
CI Lower	26.6853	4.0330	-0.4236	-0.3131	-0.1149	0.1085	-0.0027
CI Higher	50.9899	10.0874	-0.0069	0.0505	-0.0512	0.2728	-0.0013

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.

Life Expectancy Quadratic Interaction Model

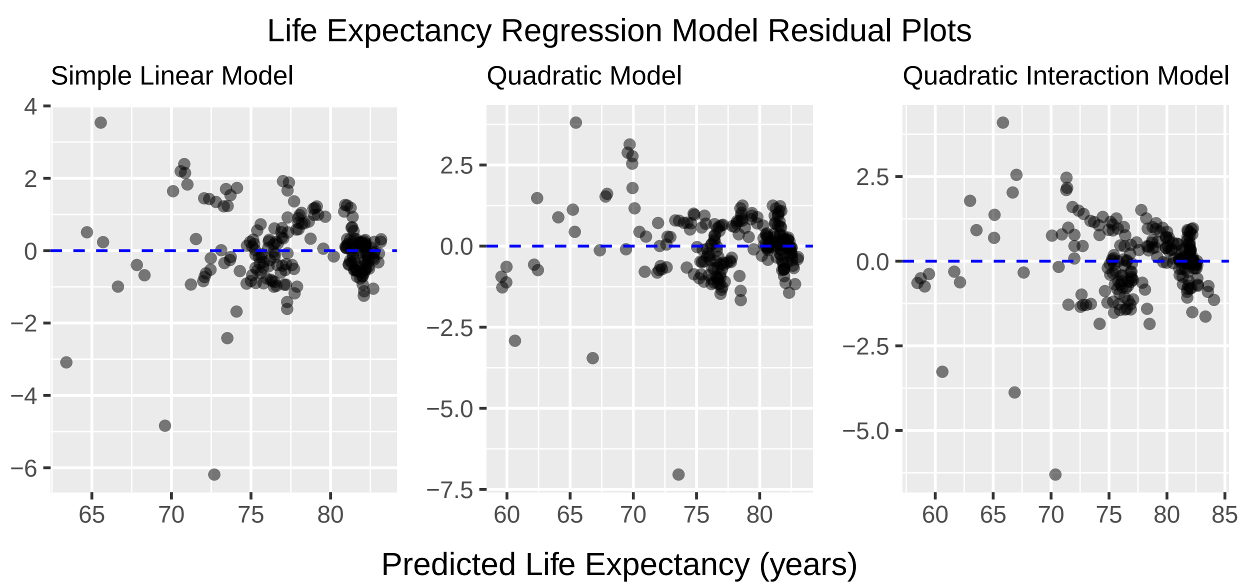
$\text{Total Life Expectancy} = \beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{THE})^2 + \beta_3 * log(\text{GDP}) + \beta_4 * \text{Poverty} + \beta_5 * \text{Education} + \beta_6 * \text{Education}^2 + \beta_7 * \text{GDP Growth} + \beta_8 * \text{Poverty} * \text{Education} + \beta_9 *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

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
Value	59.4595	-2.2205	0.2743	0.4514	-0.1820	0.4655	-0.0025	-0.3134	-0.0050	0.0059	-0.0222	-0.0060	0.1331
p-value	0.0000	0.4004	0.0868	0.0997	0.4171	0.0000	0.0000	0.0163	0.0000	0.0094	0.0409	0.0732	0.0000
CI Lower	37.9447	-7.4168	-0.0400	-0.0868	-0.6232	0.2836	-0.0032	-0.5683	-0.0066	0.0015	-0.0434	-0.0125	0.0805
CI Higher	80.9743	2.9757	0.5887	0.9896	0.2592	0.6475	-0.0018	-0.0584	-0.0033	0.0104	-0.0009	0.0006	0.1857

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.

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.

Residual Comparison



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.

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.

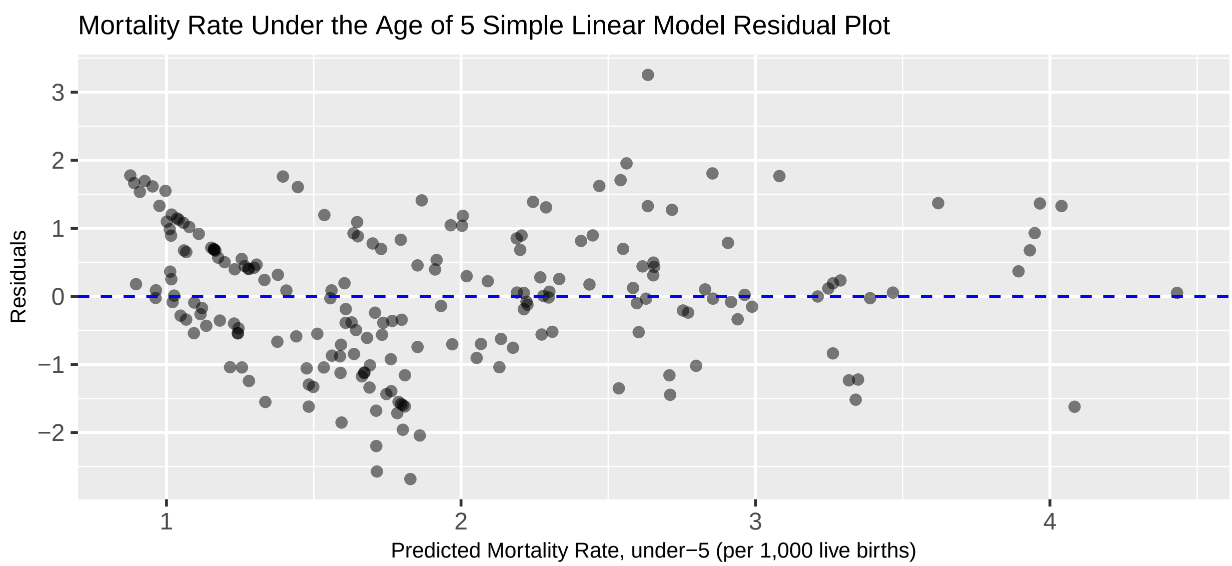
This residual scatter plot is an improvement from the quadratic model as the points are appear to less clustered and thus are more evenly distributed across the predicted years. However, the residual points seem to be less tightly centered/close to 0.

Mortality Rate Under the Age of 5 Simple Linear Model

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

	β_0	β_1	β_2	β_3	β_4	β_5	β_6
Value	4.6297	-0.5847	0.0726	0.0463	-0.0106	0.0109	-0.0076
p-value	0.0000	0.0000	0.0000	0.0000	0.5791	0.0036	0.0000
CI Lower	3.6932	-0.6539	0.0418	0.0261	-0.0481	0.0036	-0.0103
CI Higher	5.5662	-0.5155	0.1034	0.0665	0.0269	0.0183	-0.0049

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.

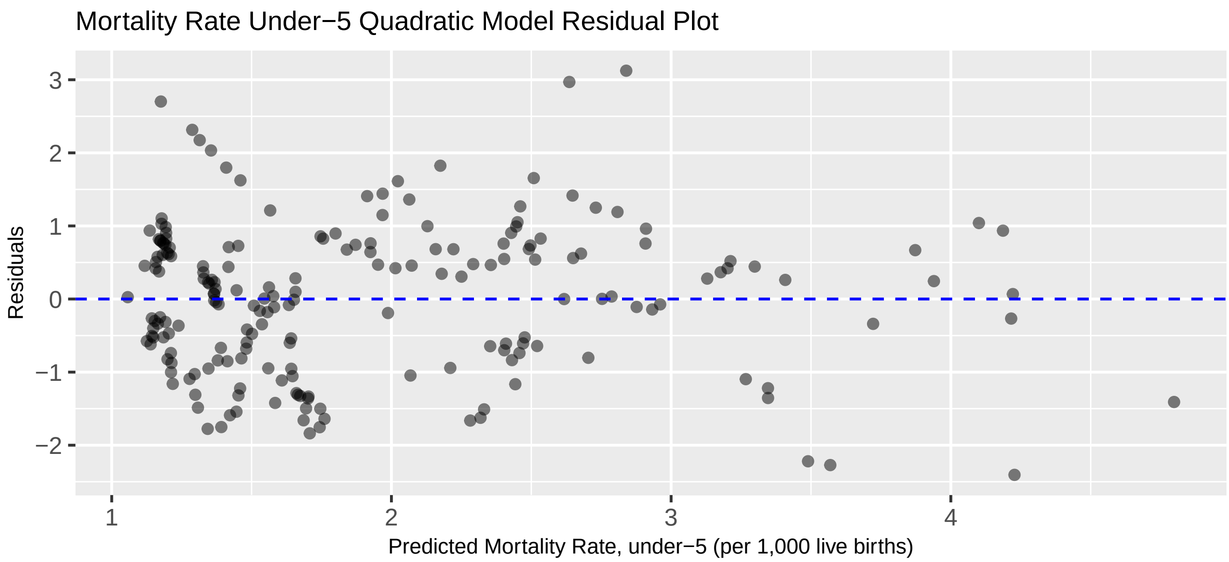
Mortality Rate Under the Age of 5 Quadratic Model

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

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

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Value	23.4436	-2.5355	0.1360	-0.5449	0.0115	-0.7722	0.0459	-0.0002	-0.0267	0.0002
p-value	0.0000	0.0000	0.0000	0.1520	0.1195	0.0000	0.0000	0.0019	0.0016	0.0078
CI Lower	13.3480	-3.2701	0.0847	-1.2921	-0.0030	-1.0298	0.0279	-0.0003	-0.0432	0.0001
CI Higher	33.5391	-1.8010	0.1874	0.2024	0.0261	-0.5145	0.0640	-0.0001	-0.0103	0.0003

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.

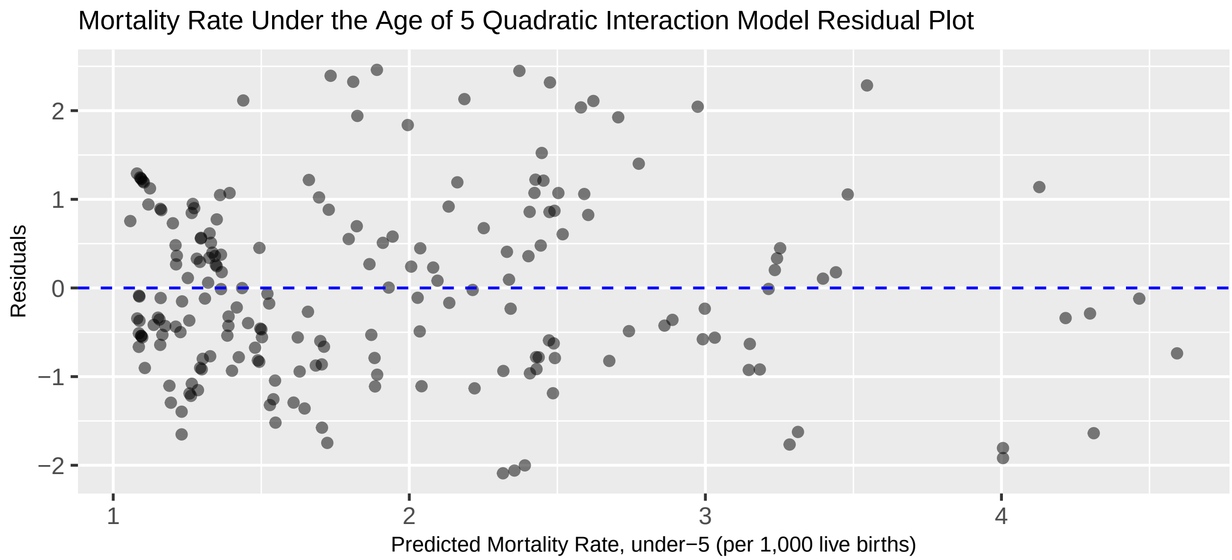
Mortality Rate Under the Age of 5 Quadratic Interaction Model

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

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.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}
Value	24.200	-0.178	0.120	-1.301	0.041	-0.161	0.069	0.001	-0.113	0	0.000	-0.080	-0.002	-0.079	0.000	0.000	0.003	-0.042	0.007
p-value	0.039	0.793	0.000	0.183	0.049	0.758	0.000	0.052	0.000	0	0.009	0.012	0.030	0.000	0.003	0.299	0.005	0.035	0.019
CI Lower	1.236	-1.511	0.058	-3.221	0.000	-1.193	0.044	0.000	-0.176	0	0.000	-0.142	-0.004	-0.123	0.000	0.000	0.001	-0.080	0.001
CI Higher	47.165	1.155	0.182	0.618	0.083	0.871	0.095	0.001	-0.051	0	0.000	-0.018	0.000	-0.035	0.000	0.000	0.006	-0.003	0.013

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.



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 R^2 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