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.

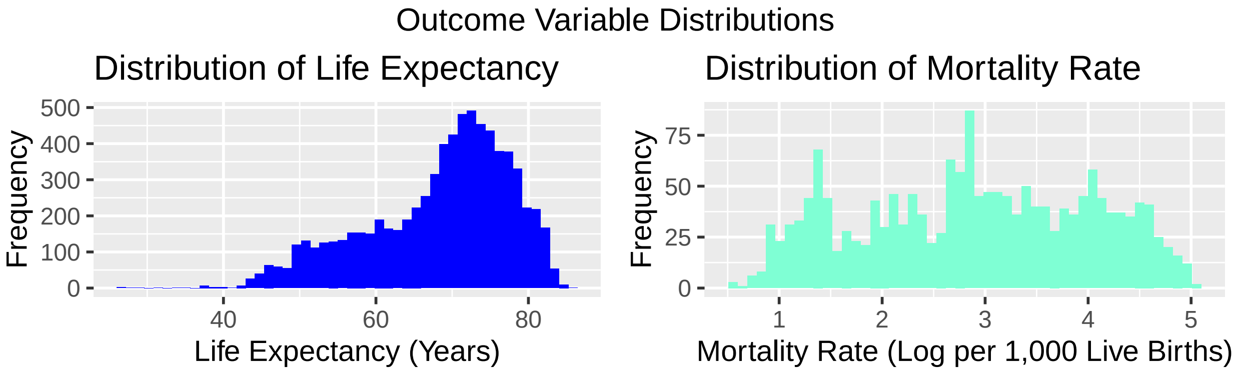
Data

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].

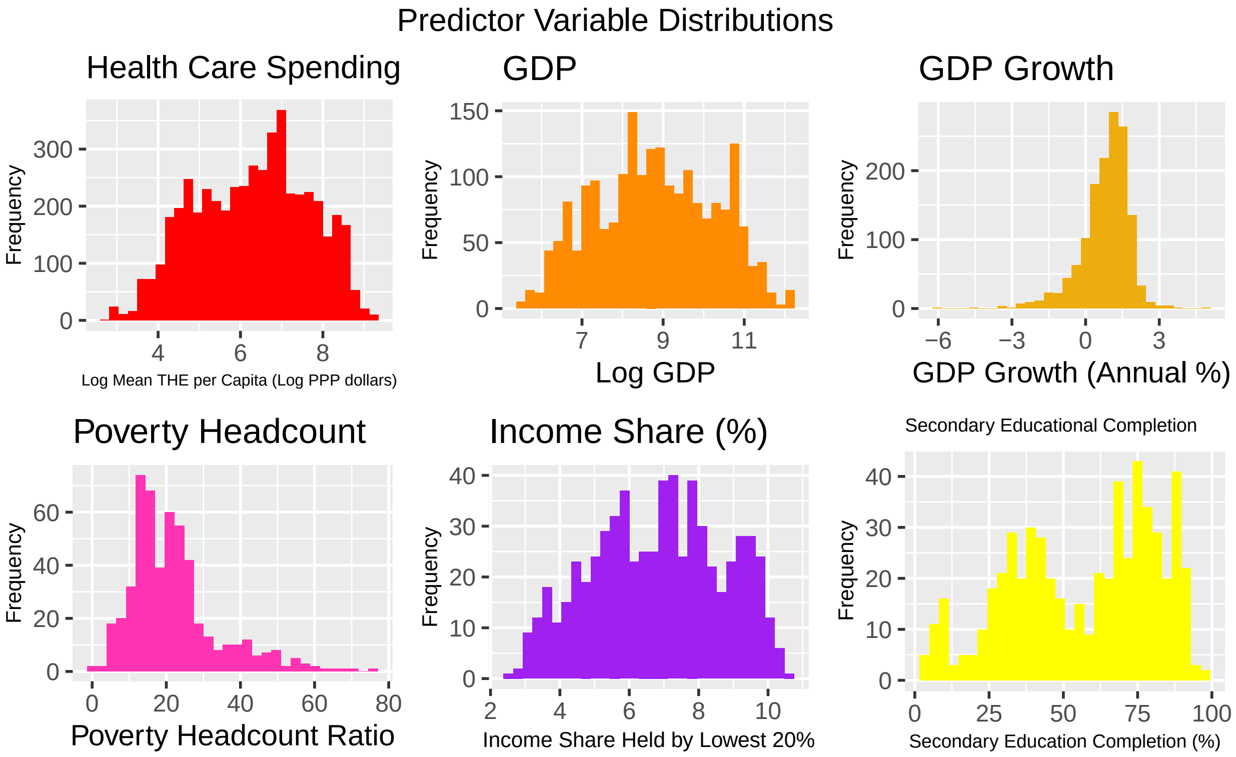
Data Wrangling and Transformations

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.

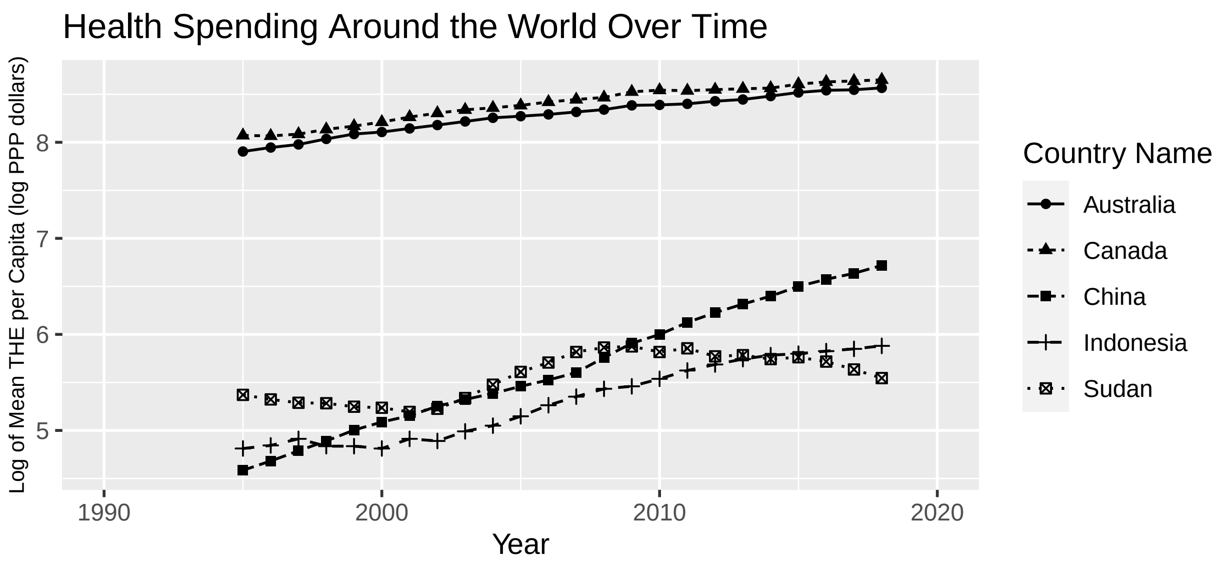
Since the analysis plan involved regression models, we looked at the distribution of our outcome variables, life expectancy and under-5 mortality rate. The life expectancy data looked approximately normal. However, the mortality rate data was log-transformed to produce a more normal distribution:



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.

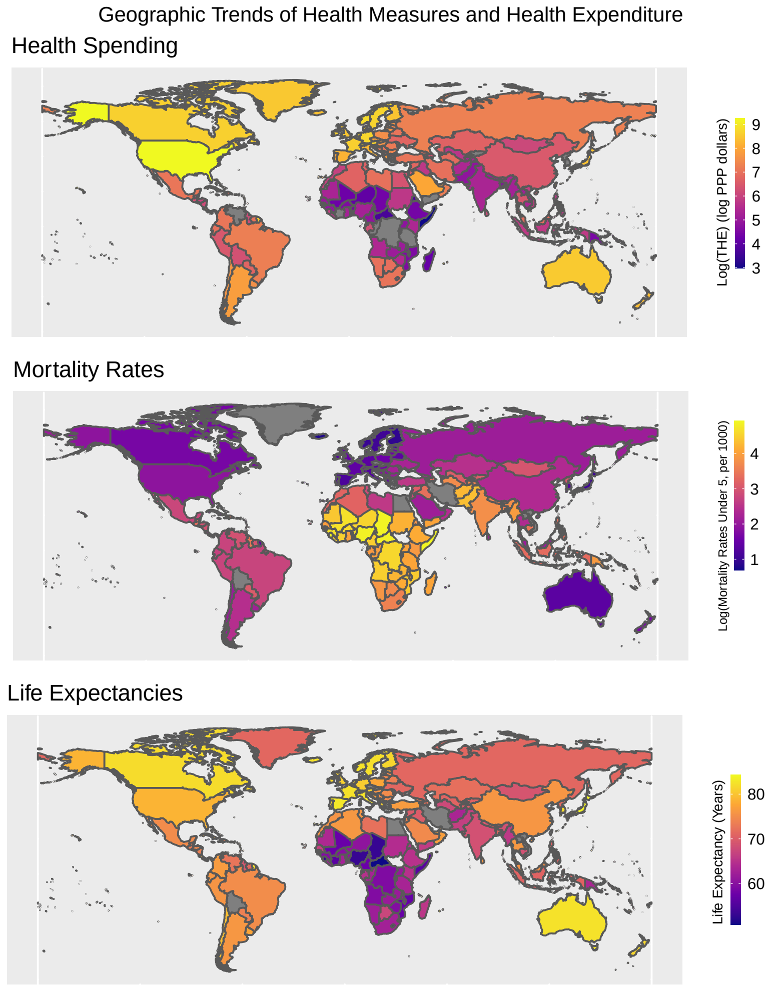


Then, we looked at healthcare spending over time in a few selected countries to see any trends in the data:



Health Spending Around the World Over Time

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

Health Spending

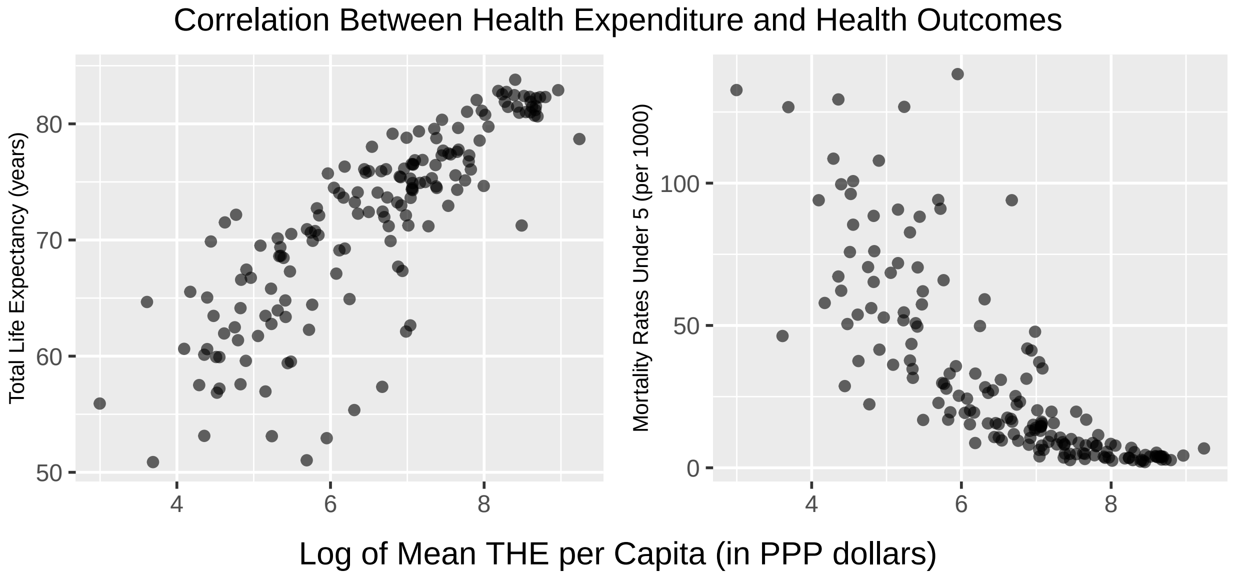
Mortality Rates

Life Expectancies

Log(THE) (log PPP dollars)
9
8
7
6
5
4
3
Log(Mortality Rates Under 5, per 1000)
4
3
2
1
Life Expectancy (Years)
80
70
60

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 $= \beta_0 + \beta_1 * log(\text{THE}) + \beta_2 * log(\text{GDP}) + \beta_3 * \text{GDP Growth} + \beta_4 * \text{Income Share} + \beta_5 *$ Poverty $+ \beta_6 *$ 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 and the adjusted R^2 for the life expectancy quadratic model is 0.896811. 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

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

	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

Given that this model has the highest adjusted R^2 value, we have decided to interpret the coefficient estimates and their significance. Based on the results above, log(THE) does not have a significant linear relationship to total life expectancy (p-value $= 0.4004 > 0.05$) and does not have a significant quadratic relationship to total life expectancy (p-value $= 0.0868 > 0.05$). log(GDP) does not have a significant linear relationship to total life expectancy (p-value $= 0.0997 > 0.05$). Poverty does not have a significant linear relationship to total life expectancy (p-value $= 0.4171 > 0.05$).

Education does have a significant linear relationship (p-value $\approx 0.0000 < 0.05$) and significant quadratic relationship (p-value $\approx 0.0000 < 0.05$) to total life expectancy. For every one increase in the percentage of the population 25+ that has at least completed upper secondary education, the total life expectancy of a country is expected to increase by 0.4655 years on average, holding all other covariates constant. Moreover, for every one increase in the percentage of the population 25+ that has at least completed upper secondary education squared, the total life expectancy of a country is expected to decrease by 0.0025 on average, holding all other covariates constant. This indicates that increasing educational attainment in a country is expected to have decreasing marginal returns on increasing total life expectancy on average.

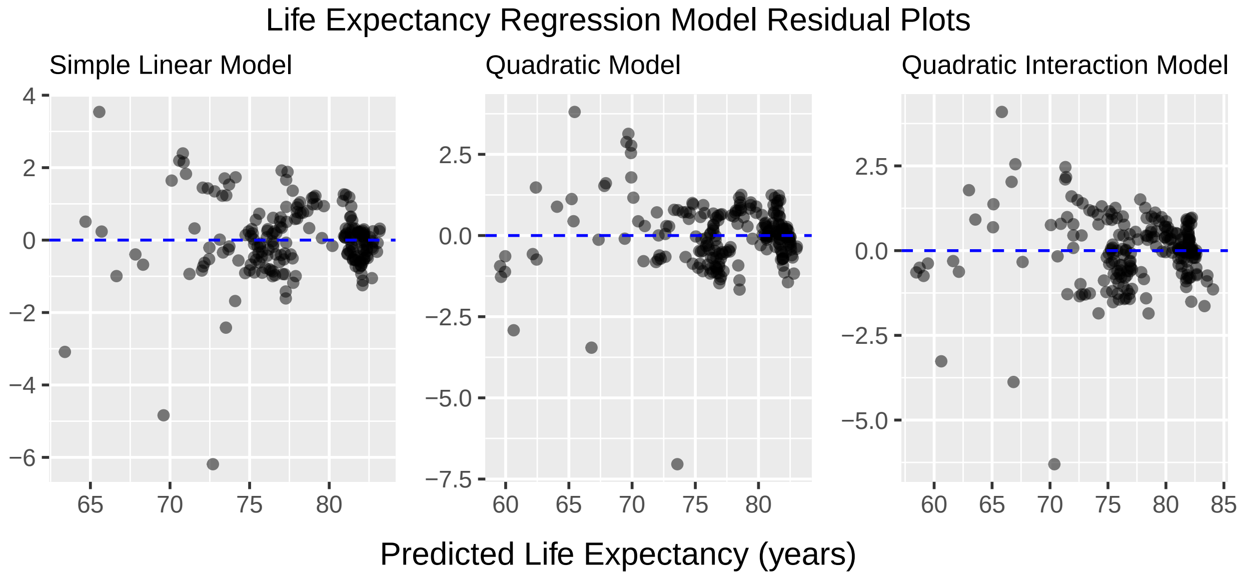
GDP growth does have a significant linear relationship to total life expectancy (p-value $= 0.0163 < 0.05$). For every one percentage increase in GDP growth, the total life expectancy of a country is expected to decrease by 0.3134 on average, holding all other covariates constant.

The interaction term between poverty and education does have a significant linear relationship to total life expectancy (p-value $\approx 0.0000 < 0.05$). Since $-0.0050 < 0$, the estimated effect of poverty and education on total life expectancy are negatively related, holding all other covariates constant. The interaction term between GDP growth and education does have a significant linear relationship to total life expectancy (p-value $= 0.0094 < 0.05$). Given that $0.0059 > 0$, the estimated effect of GDP growth and education on total life expectancy are positively related, holding all other covariates constant. The interaction term between poverty and GDP does have a significant linear relationship to total life expectancy (p-value = 0.0409 < 0.05). Because -0.0222 < 0, the estimated effect of poverty and GDP on total life expectancy are negatively related, holding all other covariates constant.

The interaction term between GDP and education does not have a significant linear relationship to total life expectancy (p-value = 0.0732 > 0.05). The interaction term between log(THE) and poverty does have a significant linear relationship to total life expectancy (p-value ≈ 0.0000 < 0.05). Since 0.1331 > 0, the estimated effect of log(THE) and poverty on total life expectancy are positively related, holding all other covariates constant.

The R^2 for the life expectancy quadratic interaction model is 0.9203884 and the adjusted R^2 for the life expectancy quadratic interaction model is 0.9155877. The higher adjusted Rˆ2 of this quadratic interaction model indicates that it is a better model than the simple linear model and quadratic model.

Residual Comparison



The residual plots for all three models are pretty similar. They all show fairly random distributions around 0, with some clustering at the higher predicted life expectancies. The simple linear model has the widest spread in residuals, followed by the quadratic regression with no interaction terms. Thus, the quadratic model with interaction terms is likely the best residual plot, which, along with its highest R^2, supports its case as the best model.

Mortality Rate Under the Age of 5 Simple Linear Model

log(Under-5 Mortality) $= \beta_0 + \beta_1 * log$(THE) $+ \beta_2 * log$(GDP) $+ \beta_3 *$ GDP Growth $+ \beta_4 *$ Income Share $+ \beta_5 *$ Poverty $+ \beta_6 *$ 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 and the adjusted R^2 for the mortality rate under the age of 5 simple linear model is 0.8467013.

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

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

	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 R^2 for the mortality rate under the age of 5 quadratic model is 0.8789936 and the adjusted R^2 for the mortality rate under the age of 5 quadratic model is 0.8732314. 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

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

	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

Similar to our approach for the life expectancy regression models, we are focusing our efforts on interpreting the model with the greatest R^2 value. Out of the 19 association terms, 13 are significant ($p < 0.05$, excluding β_0). Although there is no significant linear relationship between log(THE) and mortality rates (β_1, p = 0.793), there is a significant quadratic relationship (β_2, $p < 0.05$). Holding all other covariates constant, for every 1 increase in $\log(\text{THE})^2$, we expect, on average, an 0.120 increase in under-5 mortality rate. This is also the case with log(GDP), where there is no significant linear relationship (p = 0.183), but there is a significant quadratic relationship (β_2, $p < 0.05$). Holding all other covariates constant, for every 1 increase in $\log(\text{GDP})^2$, we expect, on average, an 0.041 increase in under-5 mortality rate. Income is another variable that shares this affect on mortality rate. The linear term is not significant (β_5, p = 0.758) while the quadratic term is (β_6, $p < 0.05$). Holding all other covariates constant, for every 1 increase in income, we expect, on average, an 0.041 increase in under-5 mortality rate. Income is another variable that shares this affect on mortality rate. For every 1 increase in $(\text{income share held by the bottom 20\%})^2$, we expect, on average, mortality rate under-5 to increase by 0.069.

Educational attainment has a significant effect on mortality rate as a linear and a quadratic predictor ($p < 0.05$). Holding all other covariates constant, for every 1% increase in secondary educational attainment, we expect, on average, an 0.113 decrease in under-5 mortality rate. Meanwhile, holding all other covariates constant, for every 1 increase in $(\text{secondary educational attainment})^2$, we expect, on average a near-zero change in under-5 mortality rate.

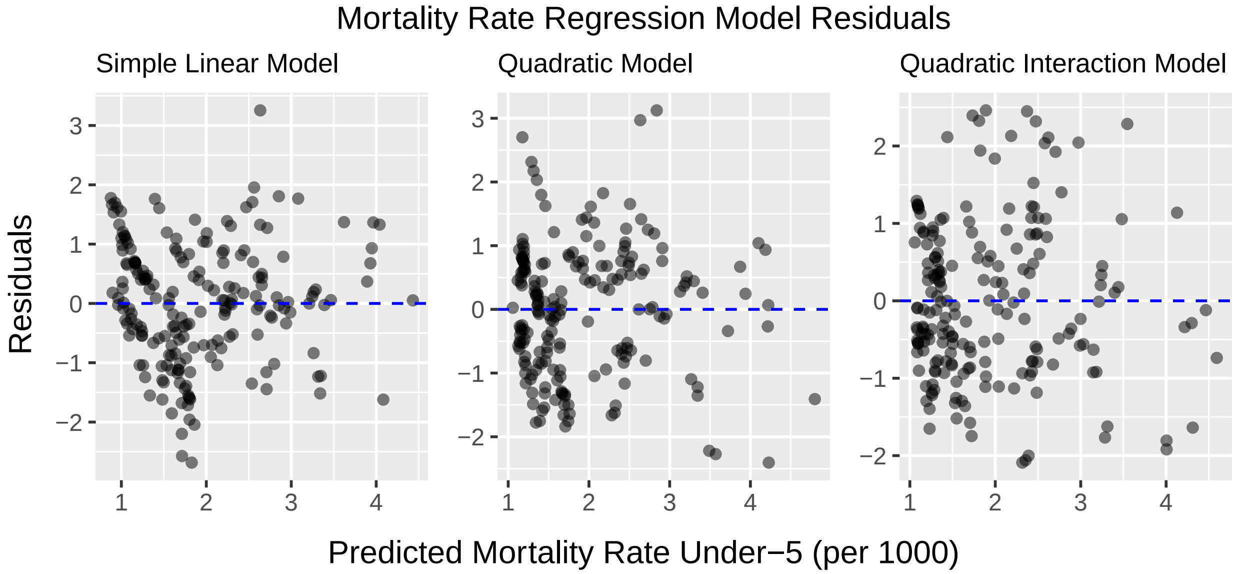
Poverty headcount has a significant linear correlation with under-5 mortality rates (β_{11}, $p < 0.05$) but not significant quadratic correlation (β_7, p = 0.052). Therefore, for every increase in poverty headcount by 1%, we expect, on average, a 0.080 decrease in under-5 mortality rate, with all other covariates constant.

There is a significant relationship between GDP growth (%) and under-5 mortality rates (β_{10}, $p < 0.05$). However, the value of the association is roughly 0, meaning that, although significant, there is minimal change in mortality rates when GDP growth changes.

The significant ($p < 0.05$) interaction terms include those between income share and secondary educational attainment (β_{12}), log(THE) and log(GDP) (β_{13}), income share and GDP growth (β_{14}), log(GDP) and secondary educational attainment (β_{16}), log(GDP) and income share (β_{17}), and income share and poverty headcount (β_{18}). β_{12}, β_{13}, β_{17}, and β_{18} are all positive, meaning that the estimated effect of the two variables in question on under-5 mortality rates are negatively related, holding all other covariates constant. In contrast, β_{18} is positive, indicating that the effects of income share held by the lower 20% and poverty headcount on under-5 mortality rate are positively related with other covariates constant. There is no significant interaction between secondary educational attainment and GDP growth (β_{15}, $p = 0.299$).

The R^2 for the mortality rate under the age of 5 quadratic interaction model is 0.9123402 and the adjusted R^2 for the mortality rate under the age of 5 quadratic interaction model is 0.9035743. 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



None of the three residual plots are particularly concerning. They are all randomly distributed around 0. For the linear and simple quadratic models, there appear to be more negative residuals at lower predicted rates and more positive residuals at the middle predicted rates. The spread of the residuals decreases as the model "complexity" increases (from linear to quadratic to interaction), which resembles the increase in R^2 with model "complexity," supporting the argument for the quadratic model with interaction terms being the best model.

Conclusions

Comparing the two final models, we see one major similarity arise, whether we use mortality rates under 5 or life expectancy. Secondary educational attainment has significant positive linear and quadratic effects on life expectancy and significant negative effects on under-5 mortality rates, both of which are indicators of improved health outcomes. Thus, as the percent of the population that has upper secondary education increases, health outcomes are expected to improve, on average. However, the quadratic term informs us that at a certain point, the benefits of increased secondary educational attainment on health outcomes levels off. Nonetheless, 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. This finding is consistent with other studies. Our results also seem to refute past studies analyzing GDP and total health expenditure. We did not find significant linear relationships between either log-transformed variable and the two health outcomes. There were significant relationships between $\log(\text{THE})^2$ and both life expectancy and under-5 mortality, but the different in their effects on improving health outcomes. There were no other variables with consistent effects across both models. We did determine multiple significant interaction effects between the predictors for both models. However, there were no consistent trends across the two models.

Limitations

One limitation of our analysis is that it only applies to countries that have data for the World Development Indicators and health outcomes (life expectancy and mortality rate under the age of 5) we are exploring. It is more difficult to find and gather accurate data in lower income countries, and thus that may limit the generalization of our results. Even though we used adjusted R^2 as a benchmark in model creation, we may still have a problem with overfitting because not all the included covariates in our best models are significant (p-value < 0.05). The coefficients in these models are also difficult to interpret given the complexity of the model (log transformations, quadratics, and interaction terms). The residual plot for the Life Expectancy Quadratic Interaction Model that we used in our final analysis (highest adjusted R^2) also had some clustering which was not ideal. Finally, we cannot make any statements about the causality between our selected World Development Indicators and health outcomes.

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] 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