Team 38 Project Final Report

Bill Coningsby Kevin Holmes Miller Love Badal Rastogi Peggy Wu

Introduction

Healthcare spending globally varies significantly between countries, both in magnitude and form. It is often the single largest expense a national government takes on, and the decision of how much to spend on healthcare and what form that spending takes on can have significant impacts on health outcomes, especially life expectancy. As a result, decisions related to healthcare spending should be carefully considered.

In this project, we will examine healthcare spending across 192 countries from the years 2000-2015. We will attempt to identify the most effective methods of spending for improving life expectancy for countries in each of four income groups. We will also examine the differences in spending between each of the income groups.

Literature Review

Given the importance of this topic globally, there is a significant body of research that has been focused on this area. Rahman et. al. found that when resources in the public sector are used inefficiently, the likely outcomes on health status will not be achieved as expected based on the relationship between total expenditure and life expectancy. This paper helped our project group examine the effects of public versus private health expenditures on health outcomes as indicated by increased life expectancies. This research indicates that while in some countries additional public expenditure has a positive effect on life expectancy, improperly used public funds, or countries with copious corruption among public officials, can actually lead to public funding having an inverse relationship with some life expectancy metrics (as studied in this cohort of South Asian and South East Asian countries) (Rahman et al., 2018).

Similarly, Squires and Anderson show the importance of efficient healthcare spending. They show that many wealthy countries achieve better health outcomes than the United States despite lower overall spending and conclude that among wealthy countries, the US ranks last in healthcare outcomes. This is relevant to our project because despite spending being highly correlated and capable of predicting healthcare outcomes on average, other data points that may not have been available to us from our sources also play major roles in outcomes when segmenting within a country's population. Such data are cited in the reviewed literature as equity in care, administrative efficiency, access to care, and the types of diseases (such as chronic conditions) that are more prevalent in countries such as the US (Squires & Anderson, 2015).

Other research has focused on the impacts of improving demographic conditions on life expectancies. Zarulli et al. found that among countries with lower levels of education, decreasing unemployment and income inequality increases average life expectancy, without increasing health expenditure levels. This study aims to understand what aspects of different cohorts within a population determine the efficiency of a healthcare system, which in turn differentiates countries with similar healthcare expenditures and their ultimate life expectancy outcomes. This study and paper are directly influential to our project because it uses the same dataset from the WHO we are using. It also provides methods of variable selection that we have mirrored such as removing highly collinear and correlated variables from the data. Although we do not plan to pursue these details within the scope of this class project, the paper also presents other data which could be included in future models to describe and explain the socioeconomic and education factors that provide further segmentation and explanations for differing life expectancy predictions amongst cohorts of countries with similar overall expenditures (Zarulli et al., 2021).

Data Set

Our data set is the combination of two individual data sets, both of which originate from the World Health Organization (WHO). Dataset 1 is National Life Expectancy data for 193 countries from years 2000-2015. In addition to life expectancy, this data set reports infant mortality rates, adult mortality rates, and the under-five death rate. It also contains demographic information relating to population size, average years of schooling, alcohol consumption, average body mass index (BMI), and the percentage of children considered thin, both for ages 5-9 and 1-19. Finally, the data set reports disease incidence for several important diseases, including HIV/AIDS, Hepatitis B, Measles, Diphtheria, and Polio.

Dataset 2 is Global Healthcare Expenditure data for 192 countries from 2000-2019. This data set contains 2930 features sorted into broad categories of aggregated healthcare spending, healthcare financing schemes & sources, specific disease spending, and macroeconomic data. There is also an income group category, which groups countries into one of four categories defined by the World Bank using Gross National Income per Capita in USD. Most features are reported in several different formats. For example, in addition to a raw dollar value in local currency, Current Health Expenditure is reported as a percentage of gdp, per capita in US Dollars, and per capita in terms of Purchasing Power Parity. Other features are reported as a percentage of Current Health Expenditure or a percentage of another aggregate spending category.

Combined, our data contains 2762 records representing 192 countries from 2000-2015. Initial exploration of the data indicated many variables were sparsely populated among the records, so our first step to clean the data was to remove variables missing data or values in more than 10% of the records. For the

remaining missing data we have imputed the values for the columns utilizing the MICE package, which utilizes advanced regression techniques such as Random Forest to predict and impute any missing values in the data.

Initial Findings

Initial exploratory analysis revealed several interesting insights into the data. First, nearly all spending categories have a logarithmic relationship with Life Expectancy, indicating additional healthcare spending has a diminishing marginal return. Figure 1 below indicates this relationship. Figure 1 Also highlights the disparities of spending and outcomes between income groups, with most high income countries spending significantly more on healthcare than any other income groups.

Figure 1 - Relationship between Life Expectancy and per Capita Healthcare Expenditure

Additionally, our exploratory analysis showed clear divisions in life expectancy based on the country's income group. Figure 2 shows the median life expectancy is highest in high-income countries and lowest in low-income countries, which is an expected result. More interesting is the differences in variability of life expectancy in income groups. The Low, Low-Mid, and Up-Mid income groups show a large range of life expectancies, while the High income group has a tight range of values. The Up-Mid group is particularly interesting, as it has a very narrow interquartile range, but a long tail of life-expectancies below the median. It is possible these low-end values are outliers caused by specific events (such as war or local disease outbreaks). Or it is possible some of these results do line up with the findings of prior researchers in our literature review which identified that other factors such as access to care and inequity can lead to lower life expectancies even amongst the upper income and high expenditure nations such as the United States.

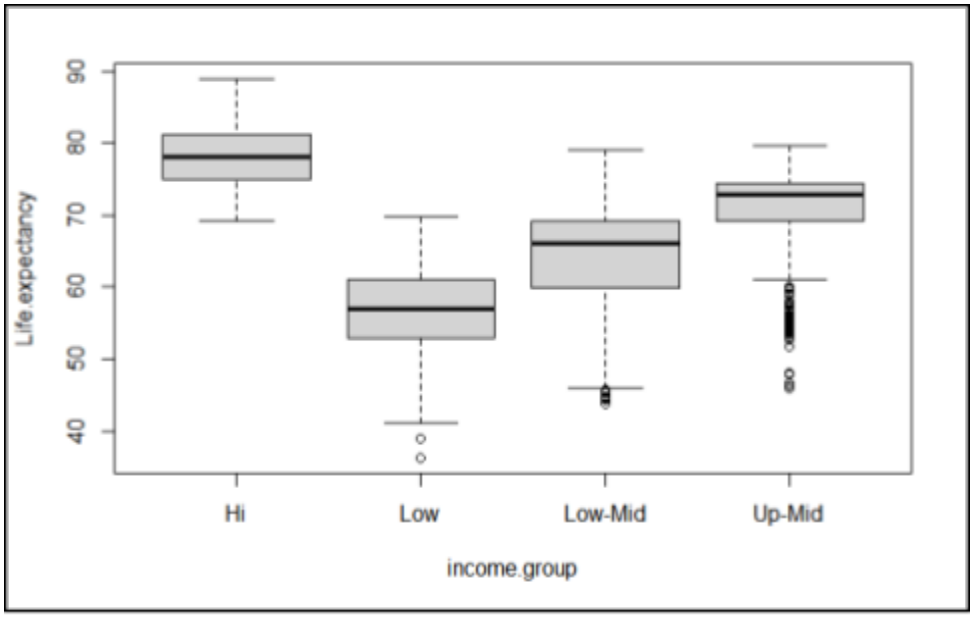

Figure 2 - Life Expectancy Boxplots by Income Group

Hypotheses

Our analysis is based on several hypotheses. First, most broadly, we expect that countries that spend more on healthcare will have longer life expectancies. Figure 1 appears to support this hypothesis, but we would also like to show that this apparent conclusion is statistically significant. We also anticipate the type of spending to be correlated with life expectancy, with countries having a larger fraction of public (vs. private) spending having longer life expectancies.

The bulk of our analysis is focused on spending within and between income cohorts. We expect the most effective type of spending to vary between the income cohorts. For example, we expect external spending is likely to be highly influential in low income countries, while public spending (such as social insurance schemes) to be most influential in High income countries. For the Low-mid and Up-Mid groups, some private expenditure is likely to be an important factor.

Approach

Our approach relies on multiple linear regression to investigate our hypotheses. We started with an interaction model using total healthcare spending per capita, income group, and an interaction term combining the two. Total healthcare spending used a logarithmic transformation due to the findings in our initial analysis discussed above which showed the logarithmic relationship between dependent and independent variables. Despite some skewness in the histogram of the residuals (in figure 3 below), the fit of this model is acceptable for inferential analysis. This model (also in figure 3), indicates a clear difference in the

relationship between healthcare spending and life expectancy in each income group. Most notably, the interaction terms for the base case (High income countries), Low income countries, and Upper-Middle income countries are all significant at the 5% level, meaning spending increases (or decreases) have different effects depending on the income group. Additionally, the high adjusted R-squared indicates healthcare spending and income group (which represents a large amount of demographic and economic detail), do a good job of explaining differences in life expectancy.

Interaction Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44.7267	2.1717	20.595	< 2e-16 ***
log(che_pc_ppp)	4.4046	0.2851	15.450	< 2e-16 ***
income.groupLow	1.4380	2.6575	0.541	0.5885
income.groupLow-Mid	-5.2861	2.6211	-2.017	0.0438 *
income.groupUp-Mid	13.6713	2.9834	4.582 4	.80e-06 ***
log(che_pc_ppp):income.groupLow	-1.9855	0.4490	-4.422	1.02e-05 ***
log(che_pc_ppp):income.groupLow-Mid	0.5152	0.4036	1.276	0.2019
log(che_pc_ppp):income.groupUp-Mid	-2.3782	0.4313	-5.514	3.83e-08 ***

Residual standard error: 5.281 on 2754 degrees of freedom
Multiple R-squared: 0.6873,
Adjusted R-squared: 0.6865
F-statistic: 864.6 on 7 and 2754 DF p-value: < 2.2e-16

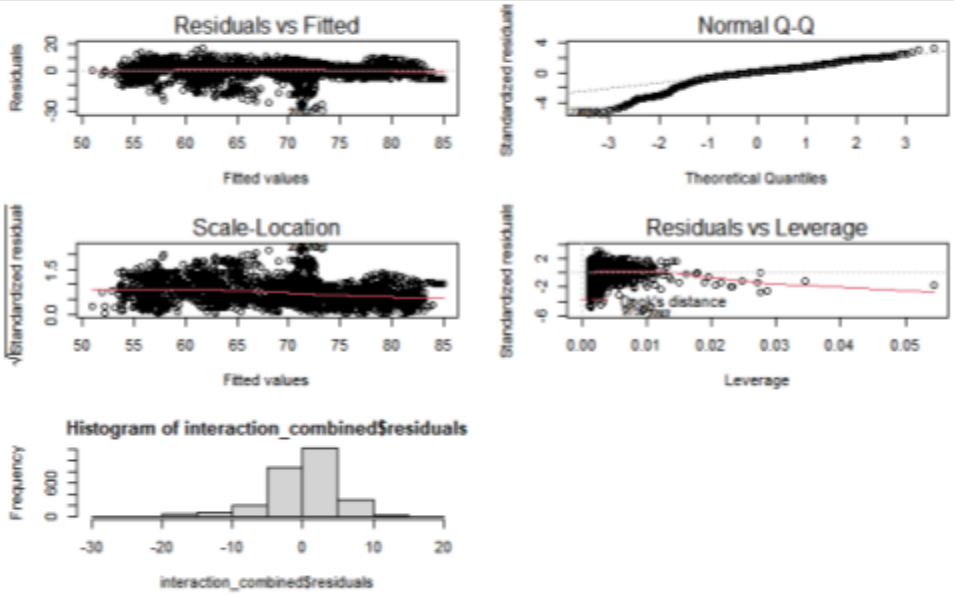

Figure 3 - Simple Interaction Model and Residual Plots

As the interaction model confirmed our hypothesis that healthcare spending has different effects in the income groups, we then developed individual regression models for each income cohort. Variable selection for these models followed a primarily qualitative process. Due to the highly correlated nature of our data (many variables are a component of another variable or are derivatives of other variables), careful variable selection is critical to avoid excess multicollinearity in our final model. We decided to limit most spending categories to the versions that are provided in terms of per-capita purchasing power parity, in order to control for population and price variation effects. We also carefully selected subset variables to include the most detail available in the data. For example, if the variables hf1 (Government and compulsory financing schemes) and hf11 (Government schemes, a subset of hf1) were available, hf1 was used in the model. If hf1, hf11, and hf12 (Compulsory insurance schemes) were available, hf1 was omitted from the model in favor of hf11 and hf12. Data cleaning handled the remainder of our variable selection, as the surviving non-sparse features left 9-14 variables available for the models.

We also briefly explored utilizing LASSO regression to aid in variable selection, but ultimately decided against it, as LASSO does not consider variable importance when eliminating correlated features. In the end, our qualitative method ensured effective models were constructed for each of the income cohorts.

Regression models for each income cohort were constructed using the variables selected in the process described above. For the high-income cohort, 9 variables were selected for the model. The initial fit of a strictly linear model was poor, as all variables have logarithmic relationship with life expectancy. Log transformations of the independent variables significantly improved the fit of the model. There remains some left skew in the distribution of the residuals (see figure 4 below), but overall this is a well fit regression model.

Only direct government transfers (fs1) and insurance contributions (fs3) are significant in this model, which may indicate that the overall level of healthcare spending is more important than specific spending policies for high income countries.

High Income Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	52.00162	1.25715	41.365	< 2e-16 ***
log1p(pvtd_ppp_pc)	0.39189	0.68740	0.570	0.5688
log1p(fs1_ppp_pc)	2.14722	0.27582	7.785	2.09e-14 ***
log1p(fs3_ppp_pc)	0.39940	0.16744	2.385	0.0173 *
log1p(fs5_ppp_pc)	-0.21827	0.13611	-1.604	0.1092
log1p(fs6_ppp_pc)	1.82529	1.15135	1.585	0.1133
log1p(hf11_ppp_pc)	-0.36856	0.22962	-1.605	0.1089
log1p(hf12_ppp_pc)	0.02081	0.16291	0.128	0.8984
log1p(hf2_ppp_pc)	0.07120	0.16336	0.436	0.6631
log1p(hf3_ppp_pc)	-0.10326	0.90373	-0.114	0.9091

Residual standard error: 3.059 on 826 degrees of freedom
Multiple R-squared: 0.4342, Adjusted R-squared: 0.428
F-statistic: 70.42 on 9 and 826 DF, p-value: < 2.2e-16

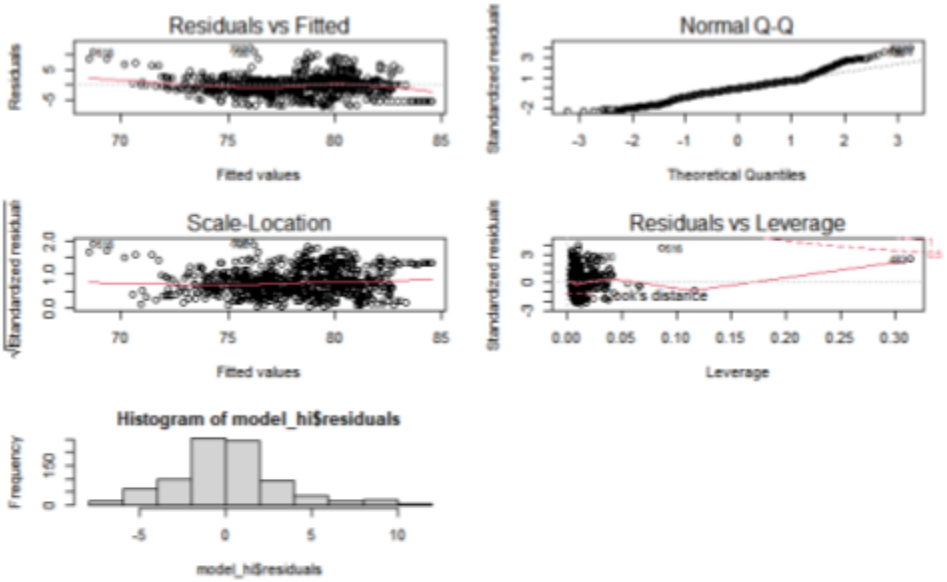

Figure 4 - High Income Model and Residual Plots

The Upper-Middle Income model was also developed with 9 variables, which were all log-transformed based on our experience with the High-income model. This model has the worst fit of all of the models, which is caused by some bimodality in the underlying data. Despite this, the fit is still reasonably good, as the residual plots (included in figure 5) show the bulk of the residuals are homoscedastic, uncorrelated, and approximately normal.

This model has several significant variables at the 5% level: Domestic Private Expenditure (pvtd), Direct Government transfers (fs1), Compulsory Contributory Health Insurance schemes (hf12), Compulsory Medical Savings Accounts (hf13), and Household Out-of-Pocket payments (hf3). Notably, Domestic Private

Expenditure has a negative coefficient, which indicates increases in private healthcare spending are associated with reductions in life expectancy. This is likely due to exogenous impacts that lead to countries increasing private health spending (such as economic downturns that affect government budgets), but it may also serve as a warning sign to countries to avoid shifting health spending to private sources.

Upper-Middle Income Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	65.3862	1.5056	43.428	< 2e-16 ***
log1p(pvtd_ppp_pc)	-6.3683	0.7795	-8.170	1.29e-15 ***
log1p(fs1_ppp_pc)	1.6400	0.5142	3.189	0.00148 **
log1p(fs3_ppp_pc)	0.3698	0.2495	1.482	0.13873
log1p(fs6_ppp_pc)	0.7891	1.3987	0.564	0.57280
log1p(hf11_ppp_pc)	-0.3041	0.4809	-0.632	0.52737
log1p(hf12_ppp_pc)	0.5193	0.2505	2.073	0.03848 *
log1p(hf13_ppp_pc)	4.3324	1.4810	2.925	0.00354 **
log1p(hf2_ppp_pc)	-0.2486	0.2006	-1.240	0.21552
log1p(hf3_ppp_pc)	5.3676	1.3158	4.079	4.99e-05 ***

Residual standard error: 4.537 on 759 degrees of freedom
Multiple R-squared: 0.3646, Adjusted R-squared: 0.3571
F-statistic: 48.4 on 9 and 759 DF, p-value: < 2.2e-16

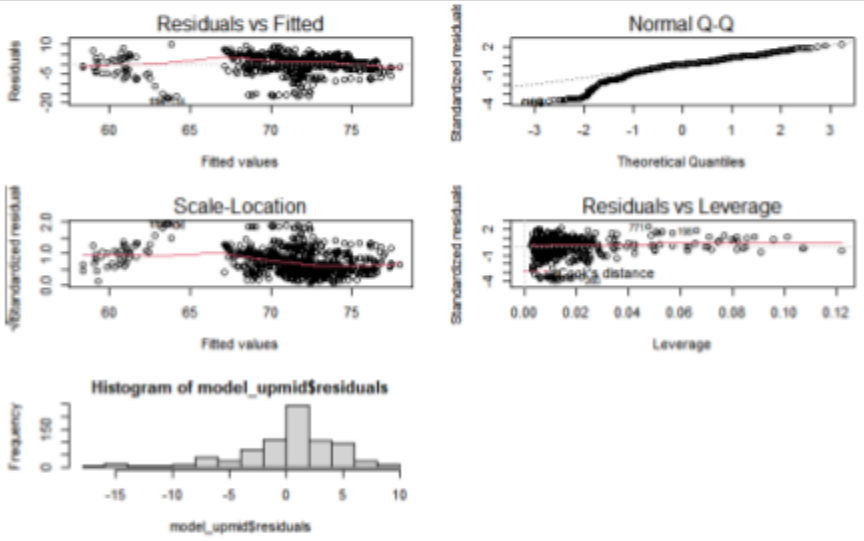

Figure 5 - Upper-Middle Income Model and Residual Plots

The Lower-Middle Income model was developed with 14 log-transformed variables. The fit of this model is relatively good, though there is some slight skewness and heteroscedasticity in the residuals. Because a Box-Cox transformation did not improve this issue and because the fit of the model is otherwise okay, we determined the fit of the model is acceptable for analysis.

This model has the most significant variables among four models, with 11 of the 14 variables significant at the 5% level. This indicates spending decisions are very important in this cohort, and the mix of negative and positive coefficients among the significant variables indicates there are "good" and "bad" spending decisions for countries in this cohort.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	52.2023	1.8693	27.926	< 2e-16 ***
log1p(ext_ppp_pc)	-2.6373	0.7258	-3.634	0.000302 ***
log1p(pvtd_ppp_pc)	6.6945	3.9233	1.706	0.088428 .
log1p(fs1_ppp_pc)	0.8112	0.5736	1.414	0.157742
log1p(fs2_ppp_pc)	1.2224	0.4994	2.448	0.014654 *
log1p(fs3_ppp_pc)	2.5871	1.0576	2.446	0.014706 *
log1p(fs4_ppp_pc)	1.8931	0.8539	2.217	0.026980 *
log1p(fs5_ppp_pc)	-19.8650	6.4563	-3.077	0.002182 **
log1p(fs6_ppp_pc)	-9.7141	3.7541	-2.588	0.009886 **
log1p(fs7_ppp_pc)	1.8515	0.4881	3.794	0.000163 ***
log1p(hf11_ppp_pc)	2.4334	0.6462	3.766	0.000181 ***
log1p(hf12_ppp_pc)	-1.2065	1.0185	-1.185	0.236616
log1p(hf21_ppp_pc)	18.8661	6.4640	2.919	0.003640 **
log1p(hf22_ppp_pc)	-0.7680	0.3639	-2.110	0.035206 *
log1p(hf3_ppp_pc)	3.2914	1.0650	3.091	0.002085 **

Residual standard error: 5.56 on 634 degrees of freedom
Multiple R-squared: 0.3987, Adjusted R-squared: 0.3855
F-statistic: 30.03 on 14 and 634 DF, p-value: < 2.2e-16

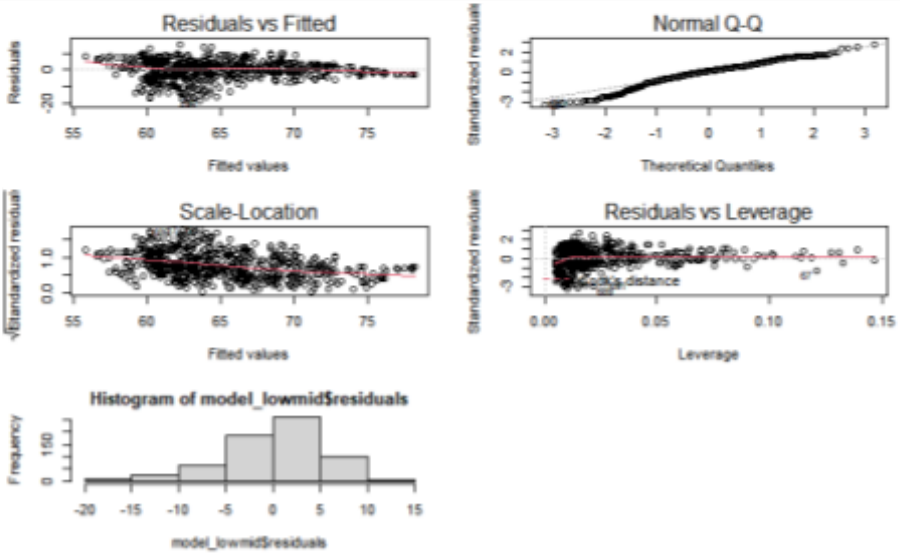

Figure 6 - Lower-Middle Income Model and Residual Plots

The Low income model was developed with 10 log-transformed variables. The fit of this model is the best of the four income cohort models, with homoscedastic residuals that are uncorrelated and approximately normal.

Despite the good fit of the model, its explanatory power is much lower than the other income cohort models. Additionally, most variables in the model are not significant - only Government Financing schemes (hf1) and Voluntary Healthcare Payment Schemes are significant. This could indicate that factors not considered in our modeling, such as disease, famine, war, or natural disasters, could have a larger impact on life expectancies than spending in low income countries. Countries in this cohort should likely focus their spending efforts on areas that will reduce or eliminate those exogenous factors.

Low Income Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	41.2254	1.7662	23.341	< 2e-16 ***
log1p(ext_ppp_pc)	-0.6287	1.1526	-0.545	0.58579
log1p(oop_pc_ppp)	0.6496	3.0335	0.214	0.83055
log1p(pvtd_ppp_pc)	-0.6955	5.7204	-0.122	0.90330
log1p(fs1_ppp_pc)	-1.6449	1.3357	-1.232	0.21891
log1p(fs2_ppp_pc)	0.1147	0.7065	0.162	0.87106
log1p(fs4_ppp_pc)	2.9841	2.7990	1.066	0.28706
log1p(fs6_ppp_pc)	1.6511	5.7384	0.288	0.77372
log1p(fs7_ppp_pc)	-1.5576	0.8747	-1.781	0.07578 .
log1p(hf1_ppp_pc)	3.4431	1.6722	2.059	0.04019 *
log1p(hf2_ppp_pc)	3.3981	1.1289	3.010	0.00279 **

Residual standard error: 5.305 on 368 degrees of freedom
Multiple R-squared: 0.2251, Adjusted R-squared: 0.204
F-statistic: 10.69 on 10 and 368 DF, p-value: 5.829e-16

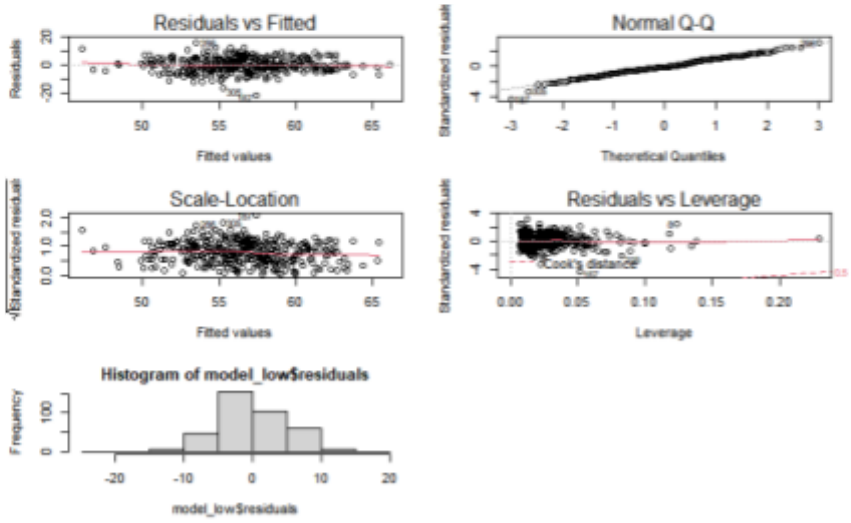

Figure 7 - Low Income Model and Residual Plots

Results and Conclusions

Our models give us an idea of which variables are most important and how they affect a population's life expectancy within an income cohort. They also confirm our hypothesis that effective spending strategies vary between different income groups. We see that for the high income cohort of countries, transfers from government domestic revenue and social insurance programs are significant in the model but that overall spending is the main contributor to health outcomes in this model as no specific financing scheme was significant.

For Upper-Middle income countries, the model shows increased private financing is negatively impacting life expectancy, meaning public expenditure is more effective at increased life expectancy for this cohort. Surprisingly for the upper-middle tier, increased out of pocket payments were also correlated with increases in life expectancy, however the impact from private financing still creates an overall net negative effect.

The Lower-Middle income model was very interesting. For this cohort, government financing schemes (policies written to improve the standard of living of all citizens), voluntary health insurance schemes, and household out of pocket payments are positively correlated with increased life expectancy. However, countries in this cohort may want to avoid relying on NPISH financing schemes, as increases in non-profit financing appear to have a negative effect on life expectancy in this cohort.

The Low income cohort model is not as capable of evaluating the main drivers of life expectancy as its related to healthcare. This could indicate that we need to do additional work in the future to add new data sources and variables to determine the best course of action for Low income countries.

Our research shows healthcare spending is a complex topic that requires close examinations by countries to ensure the most efficient strategies are selected. While the overall level of spending is the most important decision a country can make, marginal improvements can be gained by structuring health spending based on the economic condition of the country.

Unexpected Challenges

We faced a few challenges in our analysis during the course of this project. First, our initial proposal had indicated an interest in understanding the impact of spending on specific diseases on life expectancy. However, further investigation into our identified datasets showed us that there was only 3 years of usable data available for disease specific spending (i.e. 3 records per country), which was not enough data for meaningful analysis.

As previously noted, our group also had to deal with large amounts of multicollinearity in our data. Variable selection is critical to combatting this issue, so

significant time and discussion was spent on this topic. We used a variety of statistical techniques to identify and remove highly collinear variables within our dataset, but it is possible we could have removed more or created blended variables in cases of thin data if time had permitted.

Unfinished Business

We explored other modeling techniques during our analysis, but abandoned those attempts when we narrowed our project to an inferential analysis. Principal Component Analysis was a strong contender due to its strengths in feature reduction, but the loss of explanatory power in PCA models caused us to rule it out. Future research on this topic could consider a PCA approach for building a predictive model, which could be useful to both governments and private industries, such as insurance, for risk avoidance.

Future research could also pursue additional data sources to include in similar models. The model for Low income countries seemed to indicate the largest effects on life expectancy were not captured in the data available to us, so identifying and collecting the data needed for more extensive evaluation would be a good extension of this project.

Finally, given additional demographic data from each cohort, it would also be interesting to use age groups as interaction variables to determine how various age groups within the different income tiers are impacted by different diseases or sources of expenditure.

Works Cited

Squires, D. & Anderson, C. (2015). U.S. Health Care from a Global Perspective: Spending, Use of Services, Prices, and Health in 13 Countries. Commonwealth Fund. https://doi.org/10.26099/77tf-5060

Rahman, M. M., Khanam, R., & Rahman, M. (2018). Health Care Expenditure and Health Outcome Nexus: New evidence from the SAARC-asean region. Globalization and Health, 14(1). https://doi.org/10.1186/s12992-018-0430-1

Zarulli, V., Sopina, E., Toffolutti, V., & Lenart, A. (2021). Health Care System Efficiency and life expectancy: A 140-country study. PLOS ONE, 16(7). https://doi.org/10.1371/journal.pone.0253450