Final Paper

STOR 320.(01) Group 11

April 28, 2023

INTRODUCTION

The world is a vastly different place with different terrains, populations, and cultures. From the Pyramids of Egypt to the Great Wall of China, wonders of the world have attracted people from all backgrounds. However, other "wonders" of the world are present in our everyday lives, and although they attract less attention than the physical wonders of the world, they are arguably more relevant to understanding human life. Human health and knowledge are two of the less obvious wonders that surround our world as they continuously impact our daily lives. We thought it would be interesting to delve deeper into these "wonders", analyzing differences around the world, specifically how location and various related factors might affect our cumulative health and relative improvement.

The most natural curiosity that surrounds human health is the matter of life expectancy. There has been an obvious positive trend of life expectancy over time due to the progression of medical resources and treatments available to society as a whole. However, if society as a whole has advanced so much, why is there still so much variability in life expectancy over the globe? This is the first question we decided to tackle, specifically: what factors impact life expectancy the most? To analyze this question, we decided to create a predictive model to estimate life expectancy utilizing global data. There are many studies already analyzing correlations between various human health factors and life expectancy, but we wanted to take a step further, delving into the details of what exactly predicts life expectancy the best (including interactions between multiple human health factors). This analysis is one that could easily impact and possibly improve global health; if we are able to pinpoint exactly what predicts a greater life expectancy, we can focus on discrepancies concerning those factors in particular to strengthen overall human health and longevity. So, although this is a natural and common curiosity, it is absolutely one worth analyzing further.

We've all heard the statement "knowledge is power" at some point or another. To tackle the other "wonder" of the world that is knowledge, we decided to contextualize this common phrase in the realm of resource utilization. In specific, does the Income Composition of Resources (ICOR) of a country seem to correspond to the average amount of schooling in that country? ICOR is measured on a Human Development scale from 0 to 1, and is typically used to describe the productivity of resource utilization in a country. These are factors that are not as commonly explored when observing global data, so we thought it would be interesting to find a link between the two, justifying the idea that knowledge is, in fact, power. Productive resource utilization would be the power we are referring to in this analysis, as countries that find more productive ways to utilize resources are often more successful in their advancements and ambitions.

Overall, our goal of this analysis is to discover less obvious "wonders" that can improve the world, whether that be through human health, knowledge, or some other factor.

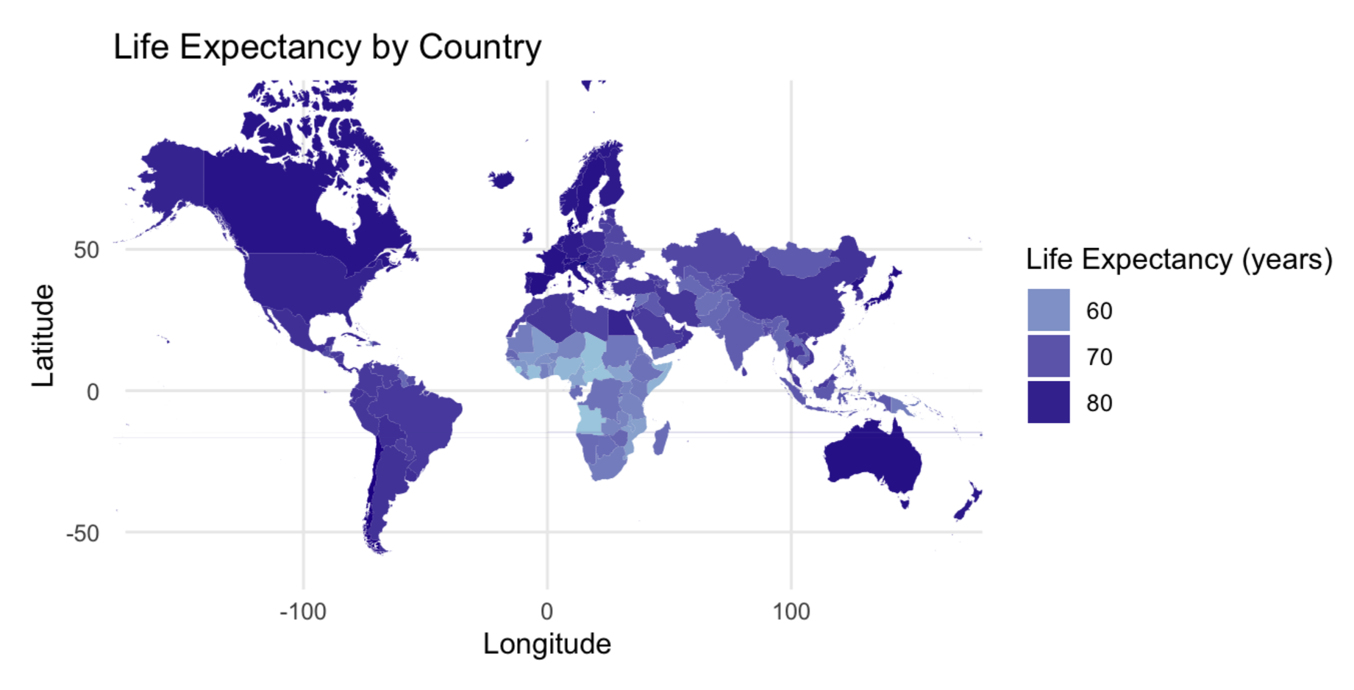
DATA

The original dataset we used as a base for this analysis came from Kaggle, specifically an author named Kumarrajarshi. This author compiled human health data from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO) for 193 countries over 15 years. The author merged this human health data with its corresponding economic data from the United Nations website to create the original life expectancy dataset we found on Kaggle. There are a total of 22 variables found in this dataset with 2938 observations. Each observation represents an individual country's health and economic data for a single year between 2000 and 2015. Some of the variables we utilized from this original dataset are *Country*, *Year*, *Status*, *Life expectancy*, *Adult Mortality*, *BMI*, *ICOR*, and *Schooling*. We left out *Infant deaths*, *Alcohol*, *% expenditure*, *Hepatitis B*, *Measles*, *Polio*, *Total Expenditure*, *Diphtheria*, *HIV/AIDS*, *thinness 1-19 years*, and *thinness 5-9 years* due to discrepancies we found in the original dataset concerning these variables.

To further explain the variables we used in our analysis, *Country* and *Year* are simply the name of the country and the year in which the observation was recorded. *Status* is intuitively the status of the given country, defined as either "Developed" or "Developing". *Adult Mortality* is the rate of death in both sexes, given as a probability of dying between 15 and 60 years per 1000 people in the population. *BMI*, *Life expectancy*, and *Schooling* are all averages over a given year, respectively measuring the average body mass index of the population, average age of life expectancy, and average number of years of schooling within a given country. *ICOR* was explained in the above section as a Human Development index between 0 and 1 depending on the productivity of resource utilization in the designated country. These are a good number of variables, but in order to extend our analysis and make it as accurate as possible, we decided to merge in multiple other datasets from World Bank, each representing an additional factor we wished to use in our analysis.

Health Expenditure Per Capita, which was taken from the WHO database, is the amount each country spends on health (divided by the total population). *Health Expenditure %GDP*, which was also taken from the WHO database, is the amount each country spends on health as a percent of its overall GDP. *Under-5 deaths* was taken from the UN Inter-agency Group for Child Mortality Estimation and estimates the deaths per 1,000 live births. *Population* displays the total population in each country and was taken from multiple sources such as the U.S. Census, UN population division, and Eurostat: Demographic Statistics. *Area (km^2)* is the area of each company (in Kilomters squares), and was taken from the Food and Agriculture Organization. We used *Area (km^2)* to calculate the *Population Density* of each country. Lastly, *GDP (Billions USD)* was taken from World Bank data and recorded in billions of US dollars.

We used all of the variables above in our model analysis to predict life expectancy besides *Country*, *Status*, and *Year*. We did not consider these variables to have a measurable effect on life expectancy when creating a predictive model, as they are categorical variables. However, these variables played an important role in the inspiration of our model prediction analysis. In the graph below, life expectancy is denoted by shading on a world map in the year 2015. In this graph, *Country* plays a vital role since life expectancy is plotted on a world map with specific use of that variable. This map displays the curiosity that inspired our model prediction analysis, as we observed the stark contrast in life expectancy around the globe. The graph can be read by observing the legend assigning a different life expectancy range for each shade on the world map. For example, it seems Canada has one of the highest life expectancy values since it is shaded as one of the darkest countries and, in contrast, Chad (a country in Africa) seems to have one of the lowest life expectany values since it is shaded as one of the lightest countries.



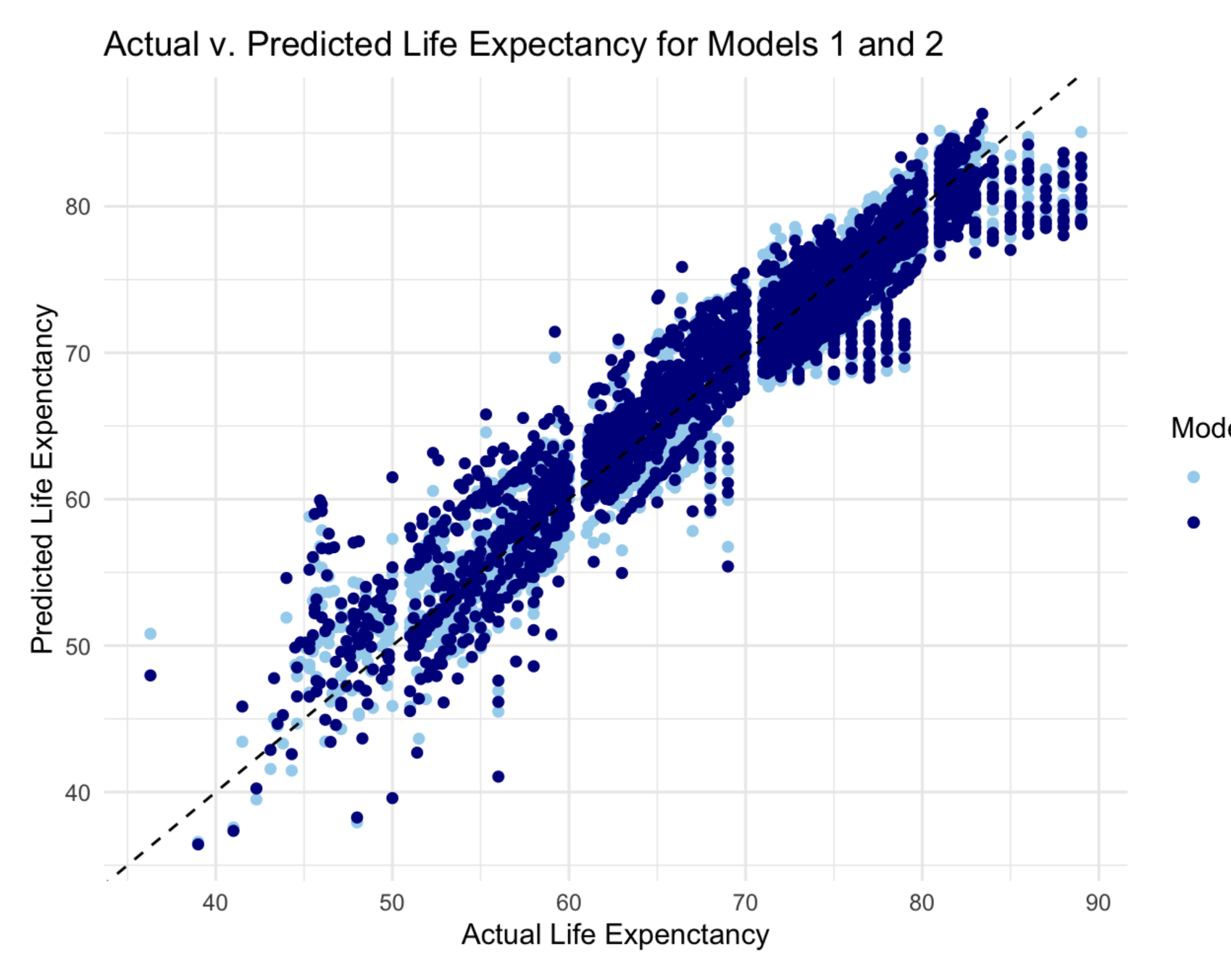
For our second question, we honed in on the variables *ICOR*, *Schooling*, *Country* and *Status* in order to focus on a possible correspondence between ICOR and years of Schooling based on individual countries and their corresponding status. However, to get the most accurate values for ICOR and Schooling to use in our analysis, we decided to take the average of both ICOR and Schooling from 2000-2015 in each country to create two new variables: *averageICOR* and *averageSchooling*. We also wanted to visualize this data on a world map, so we had to join the dataset with the 'continent3' dataset in r that contains longitude and latitude information (*long* and *lat*) for each country. There was a lot of manual manipulation and mutation in this step since countries were named differently in the 'continent3' dataset than they were in the original dataset. Also, we ran into some issues with missing ICOR and Schooling data in certain countries (specifically from the US and Russia), but overall this method seemed to work well. Therefore, the dataset we used for this analysis in particular looked similar to the one below, allowing us to easily analyze the possible similarities between ICOR and Schooling in different countries.

Country	Status	averageSchooling	averageICOR	long	lat
Afghanistan	Developing	8.2125	0.415375	74.89131	37.23164
Afghanistan	Developing	8.2125	0.415375	74.84023	37.22505
Afghanistan	Developing	8.2125	0.415375	74.76738	37.24917
Afghanistan	Developing	8.2125	0.415375	74.73896	37.28564
Afghanistan	Developing	8.2125	0.415375	74.72666	37.29072
Afghanistan	Developing	8.2125	0.415375	74.66895	37.26670

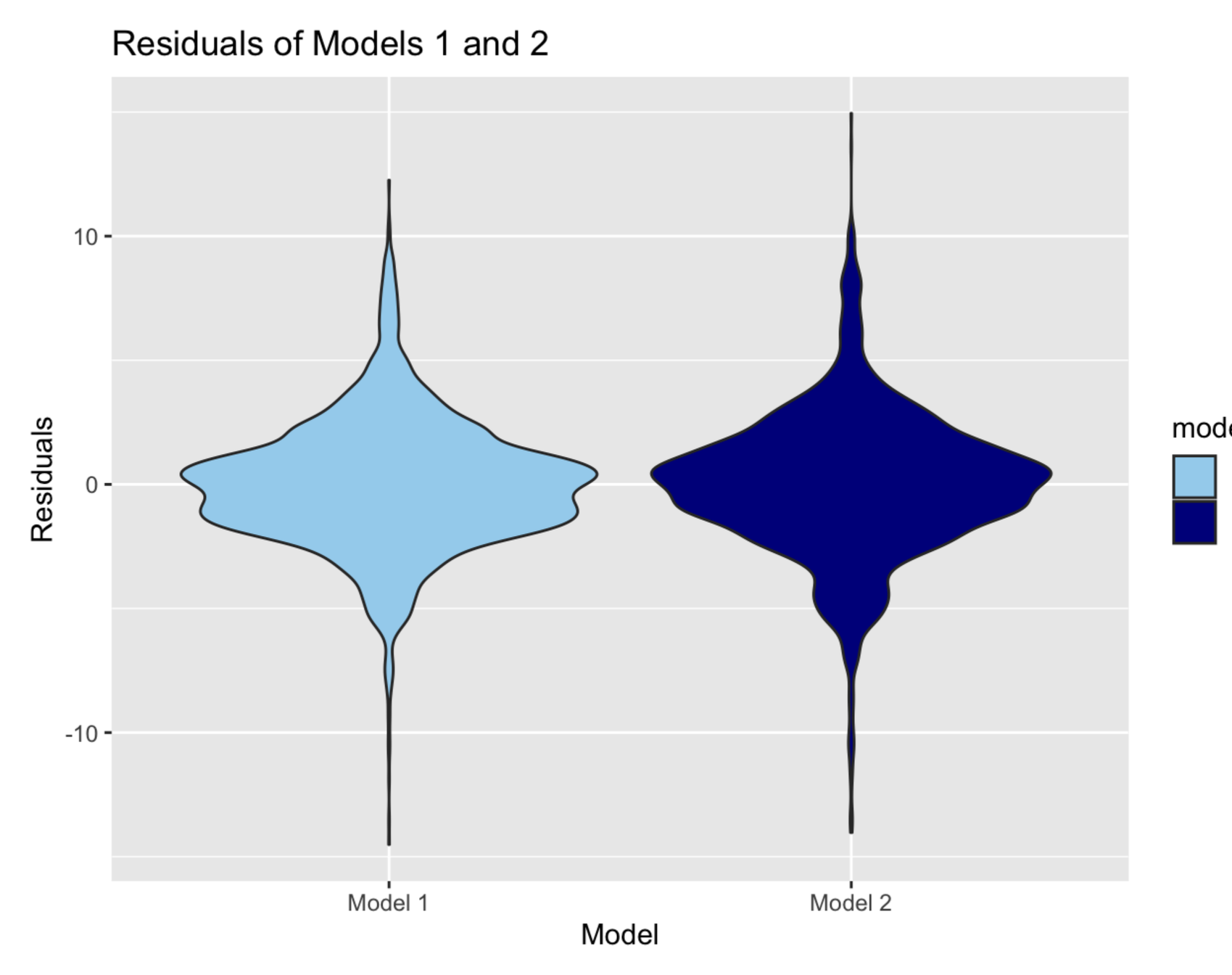
RESULTS

Our first goal was to identify which variables are most important when considering life expectancy. To achieve our goal, we decided to create a predictive model for life expectancy, which would contain the variables of importance to predicting life expectancy. Our first step in creating this predictive model was deciding whether or not we wanted to incorporate the interactions between the 11 variables in our dataset, listed above as *Adult mortality*, *BMI*, *ICOR*, *Schooling*, *Health Expenditure Per Capita*, *Health Expenditure %GDP*, *Under-5 deaths*, *Population*, *Population Density*, *Area (km^2)*,and *GDP (Billions USD)*. We used a method called the Elastic Net to help us make this decision. In our case, we performed the Elastic Net twice, once using a matrix consisting of all possible interactions between our variables (we found there to be a total of 66 combinations) and the other matrix containing just the 11 variables in our dataset.

We used alpha values of 0, 0.25, 0.50, 0.75, and 1 to perform this predictive analysis. These alpha values determine which variables are included in the respective models, depending on if those variables are significant when minimizing predictive error for that alpha value. Then, for each matrix, we selected the "best model" out of the five created (from the five different alpha values), which we identified by the model with the lowest error value. Now we have two models: the "best model" with interactions and the "best model" without. In order to lessen confusion, we will refer to Model 1 as the model including interactions between variables and Model 2 as the model without interactions. This brings us back to our question; which of the models is "better"; the one with interactions or the one with our original variables? To visualize this and help us make our decision, we plotted Model 1 and Model 2 on the same plot displaying the predictions for life expectancy the models created versus actual values of life expectancy from the data.



As seen above, Model 1 seems to have more accurate predictive values for life expectancy than Model 2 does (since Model 1's points are more centralized to the dashed line on the graph, which signifies accurate predictions). To investigate this inference further, we also visualized the residuals of both models, shown in the below violin plots. Model 2 seems to have a larger spread of residuals and more nonzero residual values than Model 1 does, signifying it is less accurate when predicting life expectancy. Therefore, we found our answer to our first question: the model with interactions predicts life expectancy more accurately than the one without.



Now that we know the model containing interactions is a more accurate predictor for life expectancy than the model without interactions, we can take a step further by comparing the interaction model created by the Elastic Net with two more interaction models created using stepwise techniques. This will allow us to narrow down the most important variables for predicting life expectancy even more than we already have. We used both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to create the two additional interaction models. The AIC stepwise selection method is a technique with the goal to estimate the likelihood of a model to predict future values, where the BIC stepwise selection method is a technique that measures the trade-off between model fit and complexity. These methods only kept variables and interactions that maximized the above respective goals for the AIC and BIC techniques. After completing these stepwise selections, we now have three models to assess:

Model 1 (the original interaction model created from the Elastic Net): *Adult Mortality* + *BMI* + *ICOR* + *Schooling* + *Health Expenditure Per Capita* + *Under-5 deaths* + *Area (km^2)* + *GDP (Billions USD)* + *Adult Mortality* : *Schooling* + *Adult Mortality* : *Health Expenditure Per Capita* + *Adult Mortality* : *Under-5 deaths* + *Adult Mortality* : *Health Expenditure %GDP* + *Adult Mortality* : *Population* + *Adult Mortality* : *Area (km^2)* + *Adult Mortality* : *Population Density* + *Adult Mortality* : *GDP (Billions USD)* + *BMI* : *Health Expenditure %GDP* + *BMI* : *Population* + *BMI* : *Area (km^2)* + *BMI* : *Population Density* + *Schooling* : *Health Expenditure Per Capita* + *Schooling* : *Under-5 deaths* + *Schooling* : *Health Expenditure %GDP* + *Health Expenditure Per Capita* : *Under-5 deaths* + *Health Expenditure Per Capita* : *Health Expenditure %GDP* + *Health Expenditure Per Capita* : *Area (km^2)* + *Under-5 deaths* : *Health Expenditure %GDP* + *Under-5 deaths* : *Population Density* + *Under-5 deaths* : *GDP (Billions USD)* + *Health Expenditure %GDP* : *Area (km^2)* + *Population* : *GDP (Billions USD)*

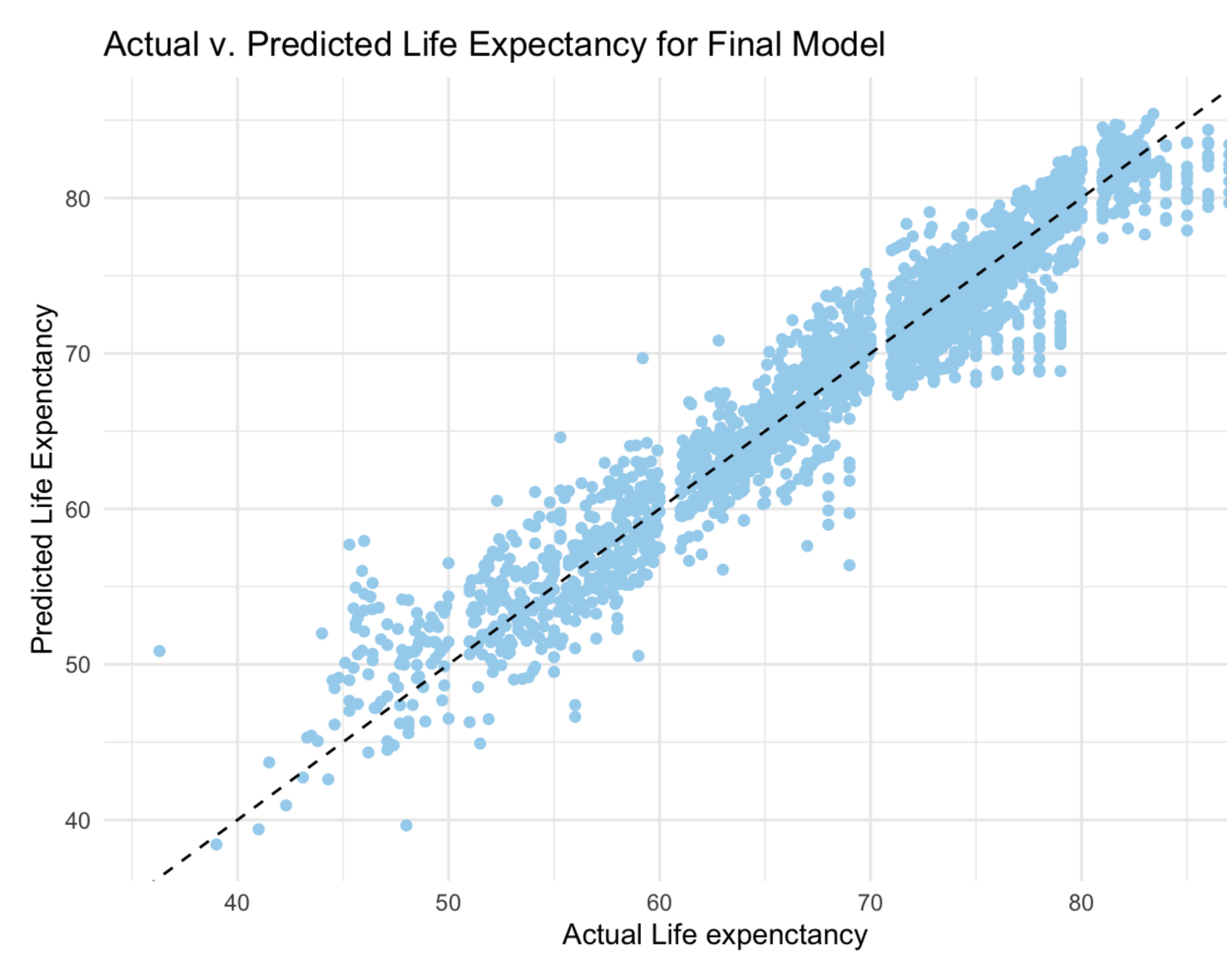
Model 2 (the interaction model created using AIC stepwise selection): *Adult Mortality* + *BMI* + *ICOR* + *Schooling* + *Health Expenditure Per Capita* + *Under-5 deaths* + *Health Expenditure %GDP* + *Population* + *Area (km^2)* + *Population Density* + *GDP (Billions USD)* + *Adult Mortality* : *Schooling* + *Adult Mortality* : *Health Expenditure Per Capita* + *Adult Mortality* : *Under-5 deaths* + *Adult Mortality* : *Health Expenditure %GDP* + *Adult Mortality* : *Population Density* + *Adult Mortality* : *GDP (Billions USD)* + *BMI* : *Schooling* + *BMI* : *Under-5 deaths* + *BMI* : *Health Expenditure %GDP* + *BMI* : *Population* + *BMI* : *Area (km^2)* + *BMI* : *Population Density* + *ICOR* : *Area (km^2)* + *ICOR* : *GDP (Billions USD)* + *Schooling* : *Population* + *Schooling* : *Under-5 deaths* + *Schooling* : *Health Expenditure %GDP* + *Health Expenditure Per Capita* : *Under-5 deaths* + *Health Expenditure Per Capita* : *Health Expenditure %GDP* + *Health Expenditure Per Capita* : *Population Density* + *Health Expenditure Per Capita* : *Area (km^2)* + *Under-5 deaths* : *Health Expenditure %GDP* + *Under-5 deaths* : *Population* + *Under-5 deaths* : *GDP (Billions USD)* + *Under-5 deaths* : *Area (km^2)* + *Health Expenditure %GDP* : *Area (km^2)* + *Health Expenditure %GDP* : *Population Density* + *Health Expenditure %GDP* : *GDP (Billions USD)* + *Population* : *Area (km^2)* + *Population Density* : *GDP (Billions USD)*

Model 3 (the interaction model created using BIC stepwise selection): *Adult Mortality* + *BMI* + *ICOR* + *Schooling* + *Health Expenditure Per Capita* + *Under-5 deaths* + *Health Expenditure %GDP* + *Population* + *Area (km^2)* + *Population Density* + *GDP (Billions USD)* + *Adult Mortality* : *Under-5 deaths* + *Adult Mortality* : *Population Density* + *Adult Mortality* : *GDP (Billions USD)* + *BMI* : *Schooling* + *ICOR* : *Area (km^2)* + *Schooling* : *Population* + *Schooling* : *Under-5 deaths* + *Schooling* : *Health Expenditure %GDP* + *Health Expenditure Per Capita* : *Under-5 deaths* + *Health Expenditure Per Capita* : *Area (km^2)* + *Under-5 deaths* : *Population* + *Under-5 deaths* : *Area (km^2)* + *Health Expenditure %GDP* : *Area (km^2)* + *Health Expenditure %GDP* : *GDP (Billions USD)* + *Population* : *Area (km^2)*

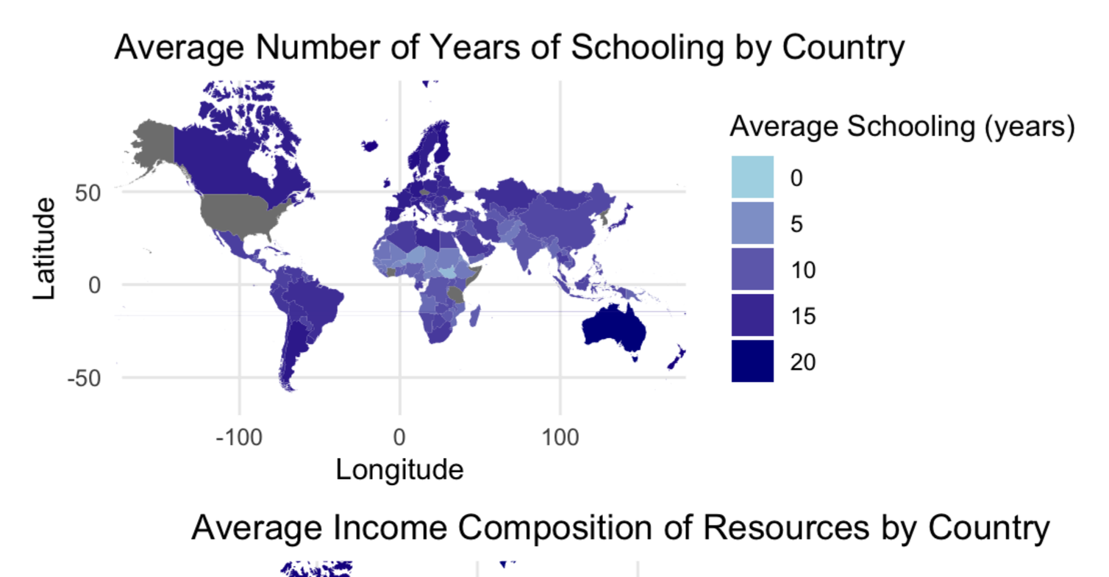
To determine which of these models predicts life expectancy the best, we randomly assigned 85% of the data as training data and the rest as testing data to use as cross-validation. Utilizing this training and testing data, we calculated the root mean squared error (RMSE) of each of the models displayed below (with the RMSE applied to the testing data specifically). We notice all models have similar RMSE values, however Model 2 (the model created using the AIC stepwise technique) has the lowest.

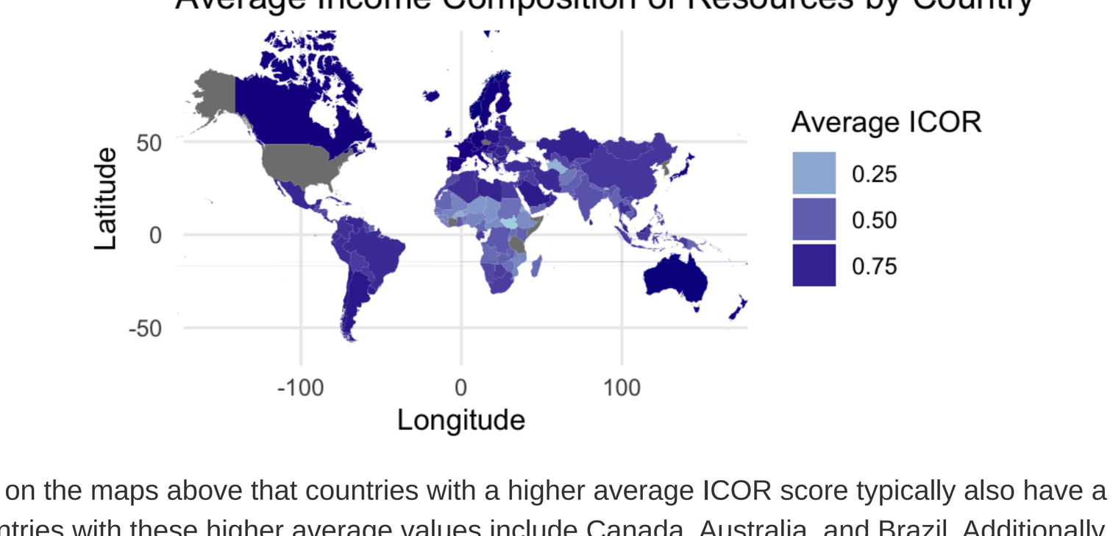
RMSE	Models
2.67	Model 1
2.64	Model 2
2.68	Model 3

Below is the plot of our final model (Model 2) for predicting life expectancy. The points represent the predictive values of life expectancy computed by the model compared to the actual values of life expectancy taken from the data. We notice the points are relatively close to the dashed line plotted, which again represents accurate predictions of life expectancy, so this confirms Model 2 does what it is supposed to do concerning predictions of life expectancy.



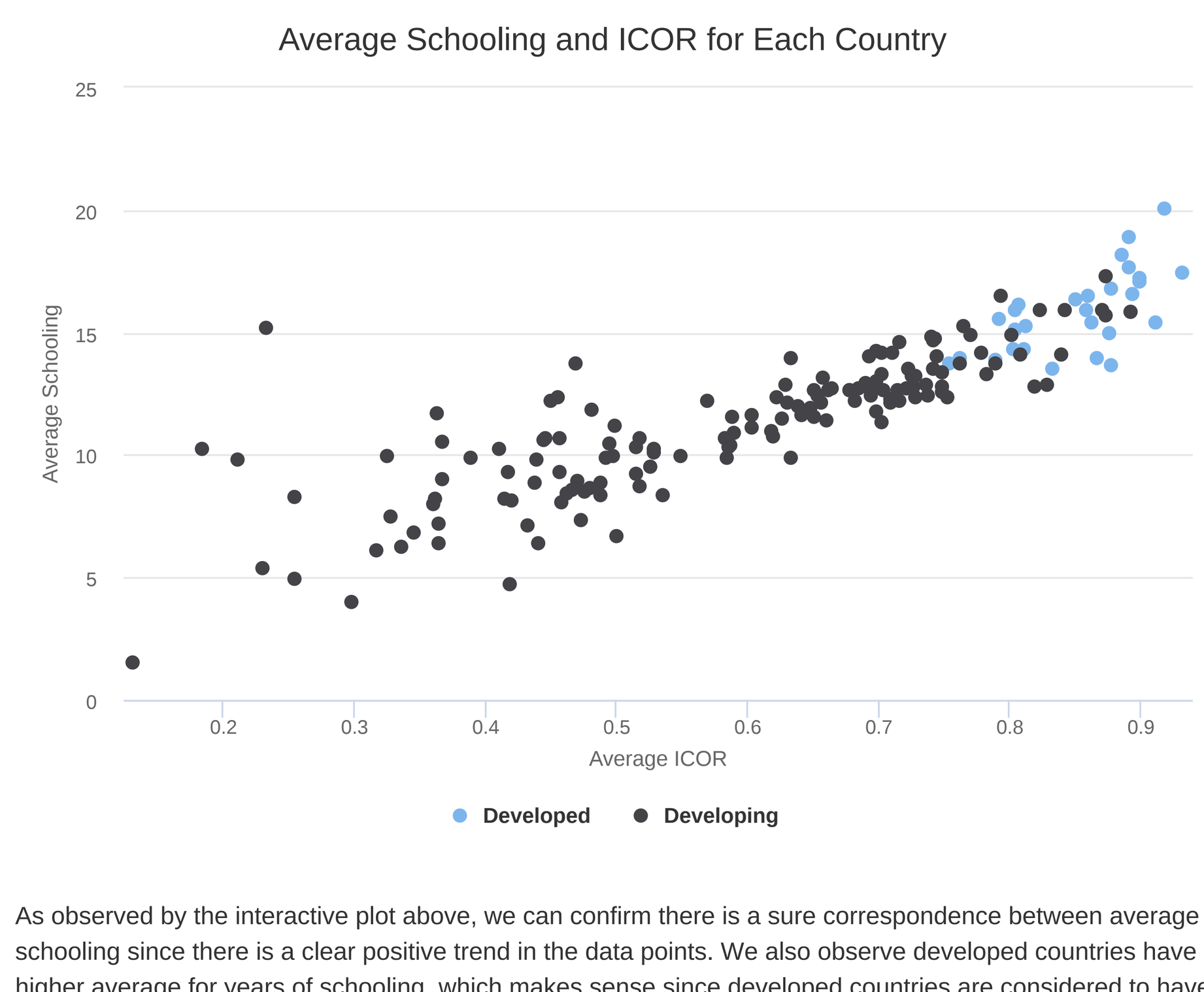
Our second goal was to observe whether the Income Composition of Resources (ICOR) of a country seems to correspond to the average amount of schooling in that country. In order to tackle this, we started with the dataset described in the section above, with the variables *Country*, *Status*, *lat*, *long*, *averageICOR* and *averageSchooling*. We used this dataset to plot *averageICOR* and *averageSchooling* values on separate world heat maps so we could observe any similar patterns between the two. In the heat map labeled "Average Number of Years of Schooling by Country", there are values from 0 average years to 20 average years of Schooling plotted, where the darker the shade of blue on the map denotes more years of schooling. In the other heat map, labeled "Average Income Composition of Resources by Country", there are ICOR scores from 0.25 to upwards of 0.75, where the darker the shade of blue on the map denotes the larger the ICOR score. Each country displays a different shade of blue corresponding to their average ICOR and Schooling values on their respective maps. As mentioned above, we are lacking data on a few countries, the most noticeable being the US and Russia, which is why they are not pictured on the maps below.





We notice on the maps above that countries with a higher average ICOR score typically also have a higher average for years of schooling. Some of the countries with these higher average values include Canada, Australia, and Brazil. Additionally, we see the countries that have a lower average ICOR score also typically have a lower average for years of schooling, some of which include South Sudan, Turkmenistan, and Burkina Faso. After visually observing these patterns, we can make an inference that average ICOR and average years of schooling have corresponding values when considering individual countries.

After making the inference that ICOR and schooling have corresponding values concerning individual countries, we wanted to dig deeper and observe if there was a difference in this correspondence depending on the status of the country. To do so, we created an interactive scatter plot with average ICOR scores on the x-axis and average years of schooling on the y-axis. There are two different colors for points, one denoting developed countries and the other denoting developing countries. Each point on the plot below denotes a different country.



As observed by the interactive plot above, we can confirm there is a sure correspondence between average ICOR scores and average years of schooling since there is a clear positive trend in the data points. We also observe developed countries have a higher average ICOR score and higher average for years of schooling, which makes sense since developed countries are considered to have higher productivity when it comes to resource utilization.

CONCLUSION

Through this analysis, we sought to investigate two questions; what are the most important variables to consider when predicting life expectancy, and is there a correspondence between ICOR and years of schooling? When coming up with a predictive model for life expectancy, we found the most important variables include *Adult Mortality* + *BMI* + *ICOR* + *Schooling* + *Health Expenditure Per Capita* + *Under-5 deaths* + *Health Expenditure %GDP* + *Population* + *Area (km^2)* + *Population Density* + *GDP (Billions USD)*, and various interactions between these variables. When thinking about the real-world application of this analysis, we could use these factors to improve life expectancy globally. One solution that could stem from this analysis is increasing health expenditure in countries, (as this variable proved to be important in our most accurate model) specifically in those that have a particularly low life expectancy. Another solution that could stem from this analysis is increasing income composition of resources, or the productivity of resource utilization in countries (since this is another variable vital to our most accurate model). This could be done by increasing the average years of schooling in each country, as we found the two have a positive trend in our second analysis.

Increasing the ICOR or number of years of schooling in a country could also help countries reach developed status, since in our second analysis we observed developed countries typically have higher average ICOR scores and a higher average for years of schooling. We have then solidified our original argument that "knowledge is power" to an extent. Our results seem to support that the more educated a country is, the more productive they are with their resource utilization, they typically have a developed status, and based on the model we created in our first analysis, they would also have a greater life expectancy. Then, resources should be directed specifically towards education and health expenditure in developing countries. These factors are the "wonders" we were looking for originally to promote global improvement.

The next step to this analysis would be analyzing where countries' resources are already directed and how to re-arrange them to be more productive, promoting education and health expenditure in specific. By helping improve developing countries, our global economy and quality of life could improve greatly, so this is an obvious topic everyone should care about. Another angle we could attack this idea is by analyzing what variable has the most impact on the status of a country and focus on re-directing resources to improving that factor. Overall, there are so many ways to looking at issues such as improving the status of countries and the world in general, but this analysis provides a minuscule look at how to do that.