FINAL PAPER

GROUP 24

Creator: Kevin Lu

Interpreter: Aditi Chiney

Orator: Jessica Barta

Deliverer: Julia Mengxuan Yu

INTRODUCTION

Group 24 chose the World Health Statistics 2020 dataset out of a broad interest in global health and a sense of deep curiosity. We wondered: What is the state of global health? Which countries are becoming healthier, and which countries are becoming less healthy? Do health outcomes differ by gender or race? Due to its wide scope and reputable source (the WHO), the dataset had the potential to answer real, meaningful questions about the health of our world. This made narrowing down potential questions a challenge! After the exploratory data analysis, though, our group decided on two main questions:

- Which health variables are most significant in predicting female HALE?
- How has universal healthcare coverage (UHC) changed over time on a global scale and can we predict how UHC will change into the future?

The group decided on the first question based on an awareness of the worldwide disparities that often exist between men's and women's health. A report by the WHO on Women and Gender Equality notes that in nearly all societies, women possess less property, wealth, and land than men, yet take on the majority of care-related tasks that involve maintaining the physical health and well-being of family and friends. In some societies, restriction of women's activity, education, and reproductive capacities are perceived as natural, and legal systems may fail to punish or even encourage violence against women. For example, from summer 2016-2017, the Supreme Court of the Russian Federation signed several bills into law that made the penalties for domestic violence administrative offenses rather than criminal offenses. Overall, there are many different factors (not just domestic violence) that can influence women's health, and therefore their lifespan, within a country. Which predictors are significant? Which will best predict female HALE? Knowing the answer to these questions could potentially give researchers and policymakers a place to start when researching how best to improve women's status and well-being around the world.

While investigating the first question, the group found that UHC, or universal healthcare coverage, was a significant predictor of women's HALE. (For background, the WHO assigns a UHC index to indicate, among other things, how accessible and affordable healthcare is in a country.) So our group developed a global map showing the distribution of coverage over time. We also constructed a simple model to project UHC indices into the future. The group was aware that healthcare access tends to improve as a country develops, but this is not always the case. Some countries have free healthcare or ensure healthcare access for all their citizens, while others do not. Some countries are a great deal wealthier than others. So which countries would have better healthcare coverage scores? Was healthcare coverage greater in wealthier countries? Was it worse in still-developing countries? Was excellent healthcare coverage a mostly western phenomenon? Since the WHO

dataset did not contain information on countries' wealth, government type or strategy for providing healthcare, this map would just begin to illustrate patterns in healthcare. Such patterns could be used to construct predictive models later using a more complete dataset.

DATA

We discovered our dataset on Kaggle. A user named "Zeus" uploaded it to Kaggle 3 months ago. However, the true origin of the dataset is from the World Health Organization (WHO). The WHO runs a global health monitoring program called the Global Health Observatory (GHO). Health data is compiled by the WHO GHO from the various publications and databases managed by the WHO, some partner United Nations agencies, and the individual WHO Member States' health monitoring agencies. At the end of every year, the GHO releases a formal report on annual trends since 2000 and makes health data publicly available. The user Zeus appeared to have downloaded a portion of this public data from the 2020 release (approximately 25 variables), partially cleaned the data, and uploaded this data as a set of CSV files to Kaggle, where we encountered it. We chose to download 10 variables that we found most interesting and useful from Zeus' post and utilize it for our project.

These variables include the following:

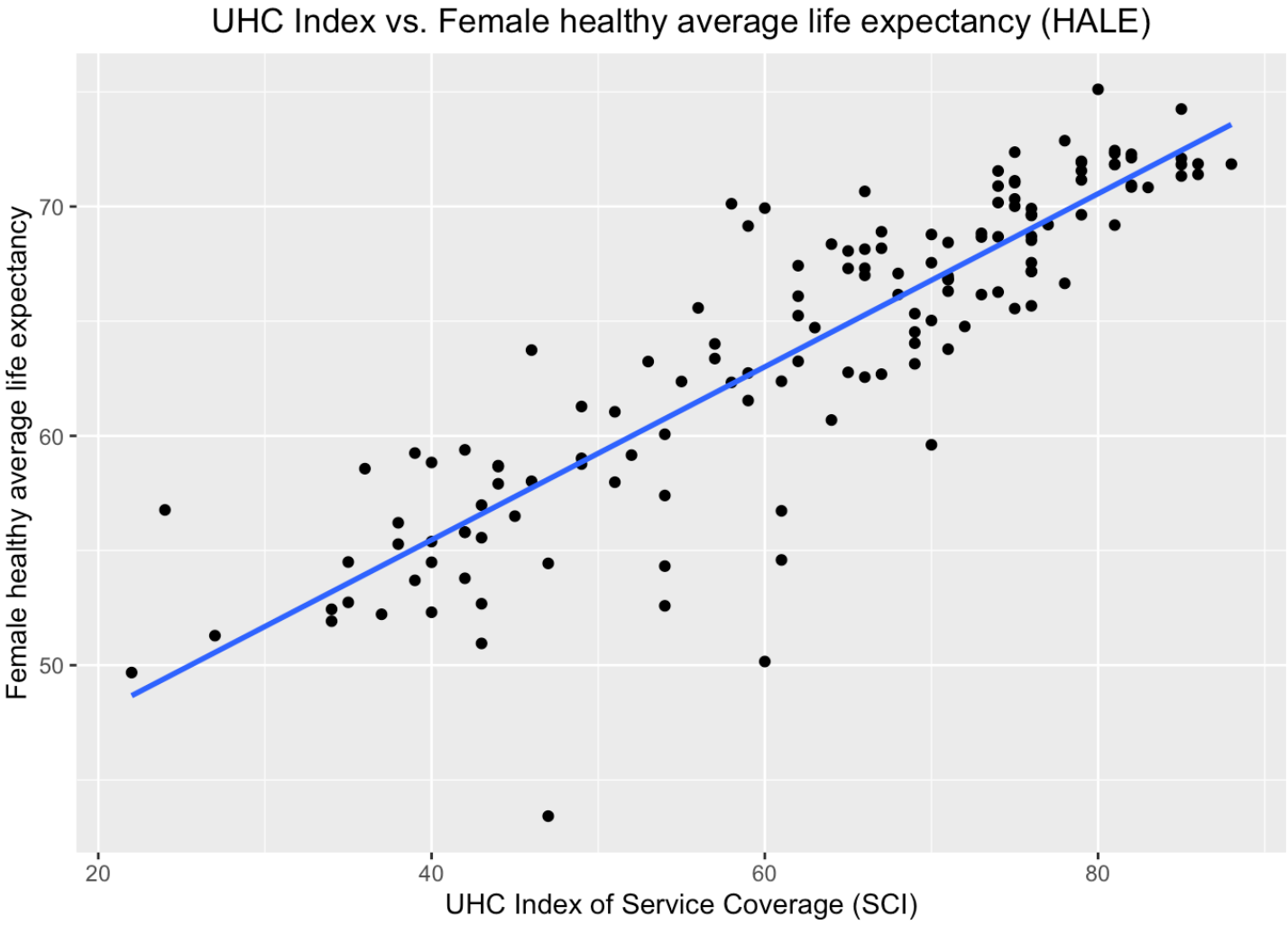
- ambient/household air pollution-attributable death rate (per 100,000 population)
- probability (as %) of dying between ages 30-70 from cardiovascular disease, cancer, diabetes, or chronic respiratory disease
- total (recorded & unrecorded) alcohol per capita (ages 15+) consumption
- crude suicide rates (per 100,000 population)
- population (%) using at least basic sanitation services
- healthy average life expectancy (HALE) at birth in years
- number of medical doctors (per 10,000 population)
- number of nursing and midwifery personnel (per 10,000 population)
- universal healthcare (UHC) index of essential services coverage
- adolescent birth rate (ABR) (per 1000 women aged 15-19)

An important note is that for each variable, data is not consistently available for every single year since 2000 nor for every country in the world. Each variable also had its own CSV file when Zeus uploaded it; for these reasons, the number of observations for each variable was different. However, for the purposes of our report, we focused on the year 2015 as data was available for most variables for this year. The number of observations for our 5 main variables of interest was 59 (i.e. data from 2015 was available for 59 countries for 5 variables). The table below shows the head of the dataset in which we combined our 5 main variables of interest.

Location	Year	HALE_Female	UHC	Adolescent_Birth_Rate	Medical_Doctors	Nursing_And_Midwife
Armenia	2015	68.14	66	24.3	29.14	49.54
Australia	2015	71.40	86	11.9	34.89	122.00
Austria	2015	71.56	79	7.6	51.04	70.14
Bahrain	2015	65.55	75	14.6	9.26	24.94
Bangladesh	2015	63.74	46	75.0	4.86	2.75

Belarus	2015	68.68	74	18.1	51.91	110.00

We selected our 5 main variables of interest based on our EDA findings and our personal interests in women's health. In our EDA, we plotted *female HALE* against all 9 other variables in our dataset and found positive linear trends for 4 of those 9, which were *UHC index*, *ABR*, *medical doctors* and *nurses and midwives*. One notably strong positive linear trend is shown below, between the UHC index and female HALE. Based on these results, we decided to build a model that would allow us to predict female HALE globally in 2015. We also created a map to visualize trends in female HALE over time on a global scale and created a model to predict female HALE into future years.



RESULTS

Part 1

The goal of our first question was to predict female HALE using the 4 variables in our dataset which we determined to be relevant in our EDA. To recap, these variables included the universal healthcare coverage (UHC) index, the adolescent birth rate (ABR), the number of medical doctors per 10,000 population, and the number of nurses and midwives per 10,000 population.

We began by splitting our data through random sampling into train and test sets (70% train, 30% test). We then created a linear model containing all four regressors: Female HALE = UHC Index + Adolescent Birth Rate + Medical Doctors + Nurses_And_Midwives. When we looked at regression statistics for this model, the adjusted

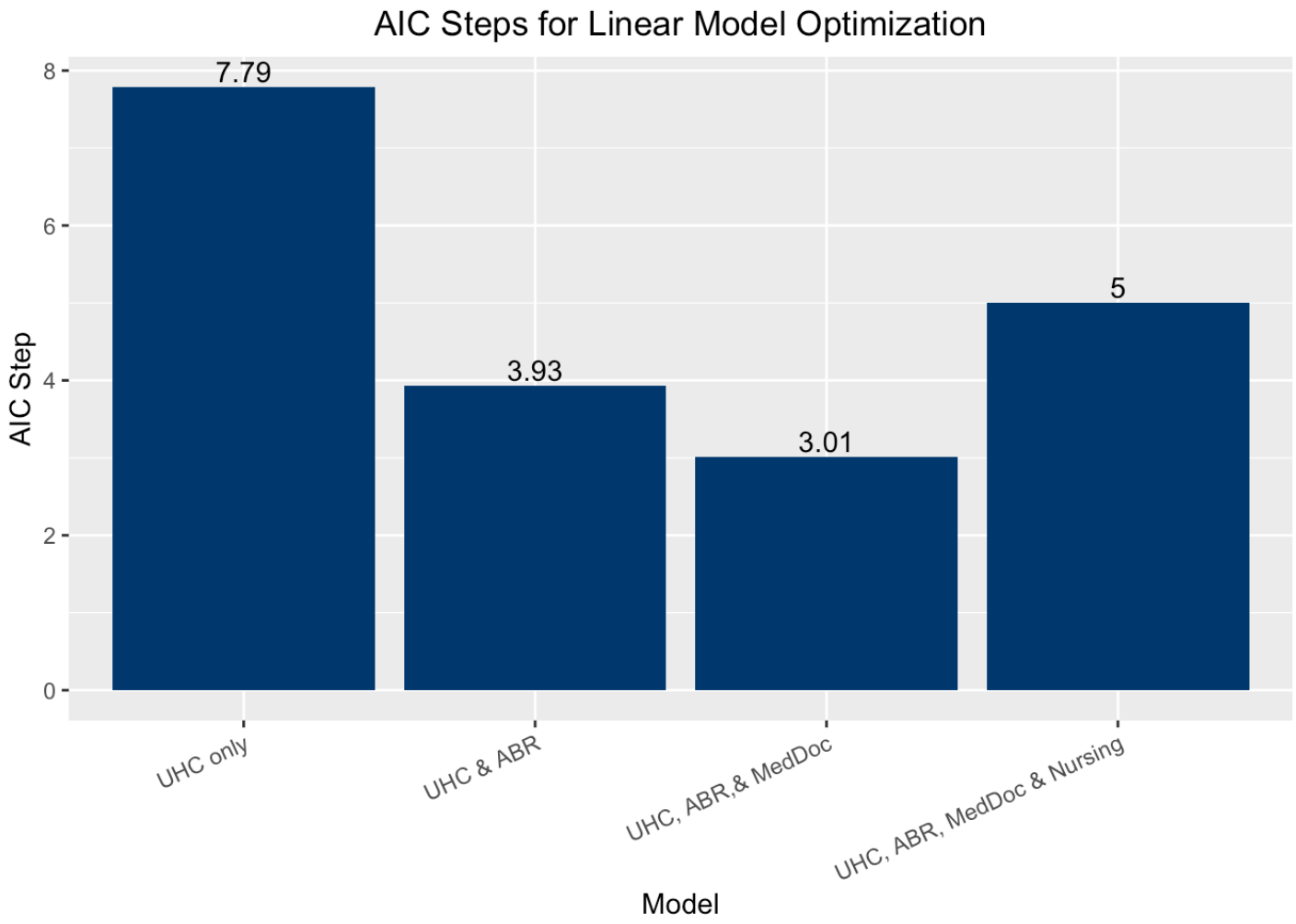
R-squared value and mean absolute error of the residuals for this model were 0.8535 and 1.069. However, when we looked at the collinearity matrix for the model, we found that the model displayed collinearity between certain variables. We also had some concerns about overfitting in our model.

To combat this, we decided to look into a measure called the Akaike information criterion (AIC). AIC estimates prediction error and the relative amount of information lost by a model. The less information lost, the better the model. One reason AIC was particularly well-suited for our needs was that AIC deals with the trade-off between how well the model is fitted and how simple the model is. This trade-off is crucial in avoiding over- and under-fitting in our model. Our goal was to determine a model that minimized AIC.

With AIC in hand, we went back to the drawing board. We decided to start with a linear model with NO regressors: Female Hale = 1. This produced an AIC of 269.06, an extremely high AIC as expected. We then tested 4 additional models with a step approach, adding one regressor at a time and assessing the resulting AIC. The models are listed below:

- Female HALE = UHC
- Female HALE = UHC + ABR
- Female HALE = UHC + ABR + Medical_Doctors
- Female HALE = UHC + ABR + Medical_Doctors + Nursing_And_Midwife

The plot below displays the AIC for each of the four models. We saw a decrease in AIC from 269.06 down to the minimum of 3.01 after adding UHC, ABR, and Medical_Doctors to the model. We confirmed that the model with these three variables was the optimal model by checking that the AIC of the model with all four regressors was greater than 3.01, which it was (AIC = 5.00).

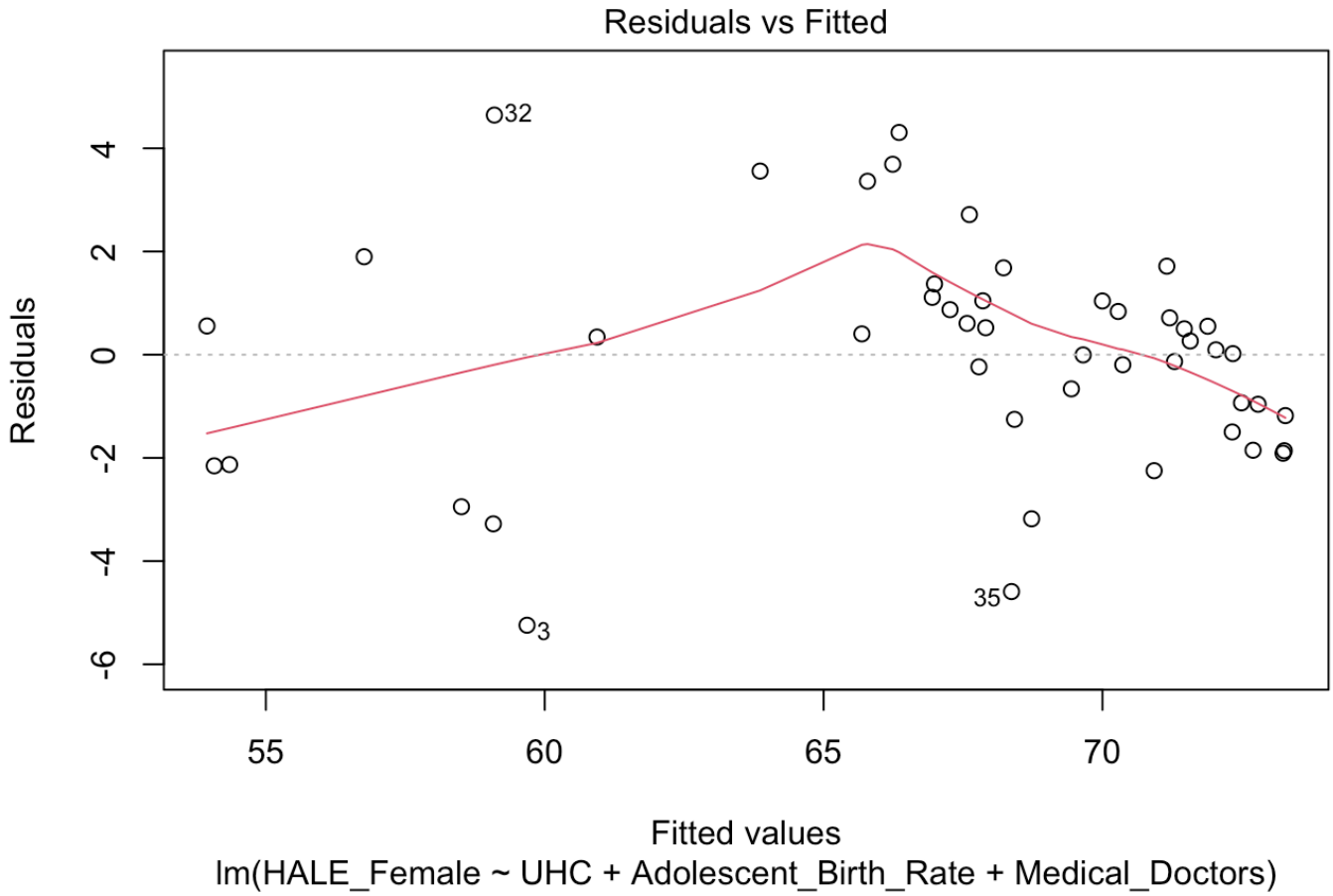


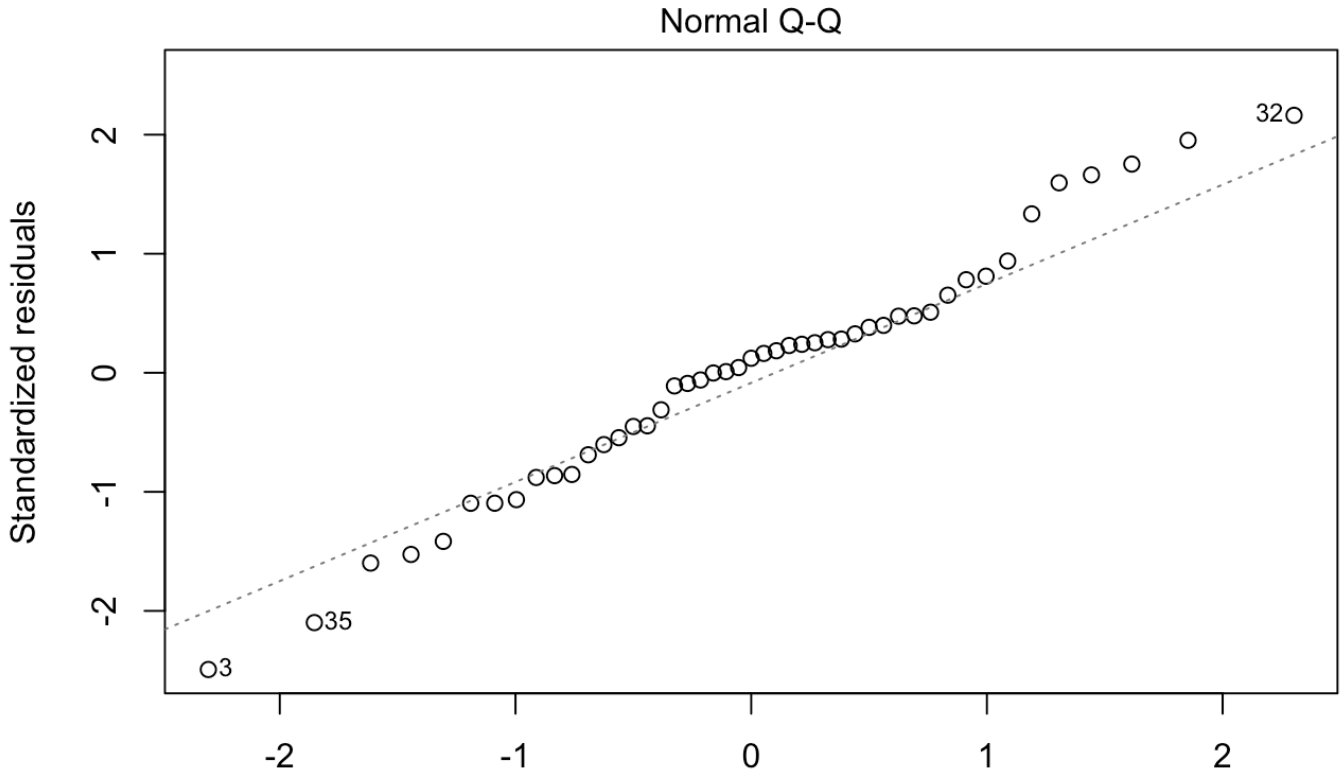
Below is a table showing coefficients, error, and p-values for our optimal model. One interesting note was that only UHC was significant at alpha = 0.05. ABR and Medical_Doctors was only significant at alpha = 0.10, which is a poor significance level. We found this interesting because when we looked at models which only contained ABR or Medical_Doctors individually, we did see significance for those regressors. However, we decided that since our key metric of interest was minimized, our model was reasonably well fitted. Our adjusted R-squared value and mean absolute error for the optimal model was 0.8569 and 1.086 respectively. Note the trade-off between AIC and mean absolute error between our optimal model and the model with all four regressors: we decided that a model with a lower AIC was a better model.

term	estimate	std.error	statistic	p.value
(Intercept)	48.6886341	2.9644818	16.423995	0.0000000
UHC	0.2652178	0.0412255	6.433343	0.0000001
Adolescent_Birth_Rate	-0.0277846	0.0149460	-1.859003	0.0698790
Medical_Doctors	0.0600176	0.0347141	1.728912	0.0909995

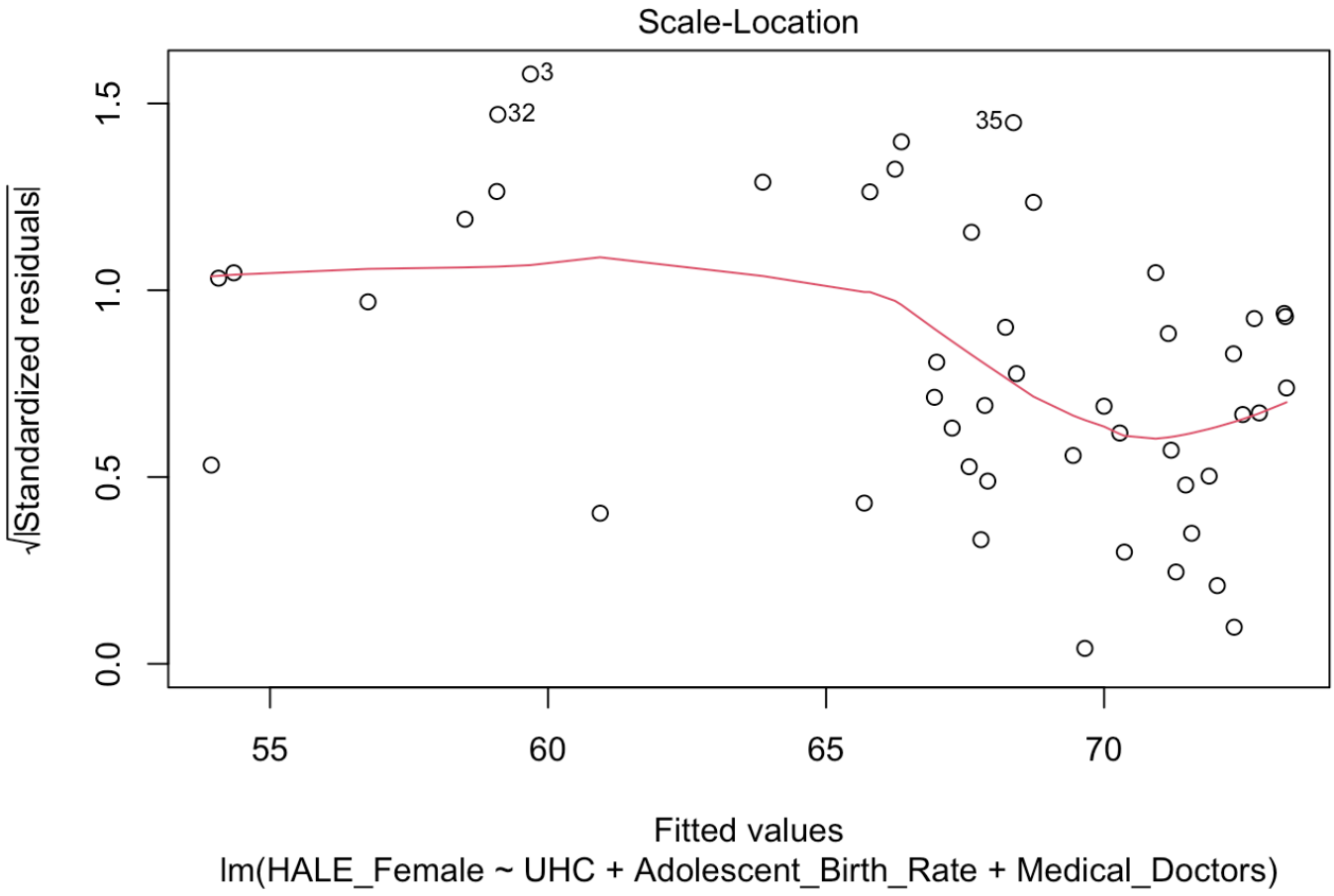
We looked at four diagnostic plots for our optimal model to determine whether there were issues with over- or under-estimation or significant clustering; these plots are shown below. Our residuals vs. fitted plot showed mild heteroscedasticity but no notable nonlinear trend among residuals. The scale-location plot showed similar results for the standardized residuals, with some evidence of increased error at upper extremes. The normal Q-

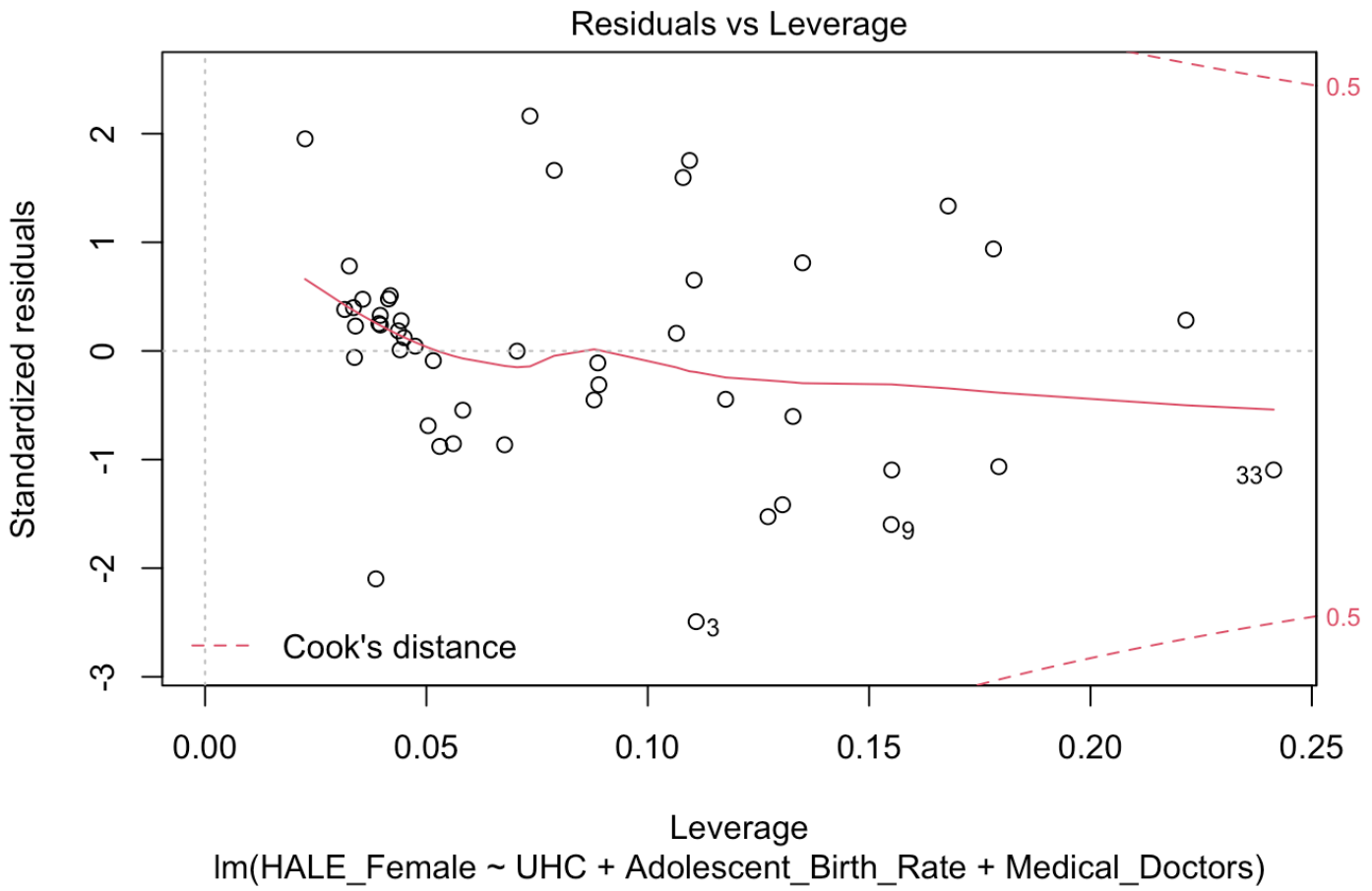
Q plot showed that residuals follow a somewhat normal distribution, with slight clustering in the middle of the distribution and some skewing at the lower and upper extremes. Our residuals vs. leverage plot displayed no alarming lines for Cook's distance scores, indicating that there were no major influential cases/observations that would need to be excluded to improve the model.


Residuals vs Fitted
Residuals
Fitted values
lm(HALE_Female ~ UHC + Adolescent_Birth_Rate + Medical_Doctors)
32
35
3
-6
-4
-2
0
2
4
55
60
65
70

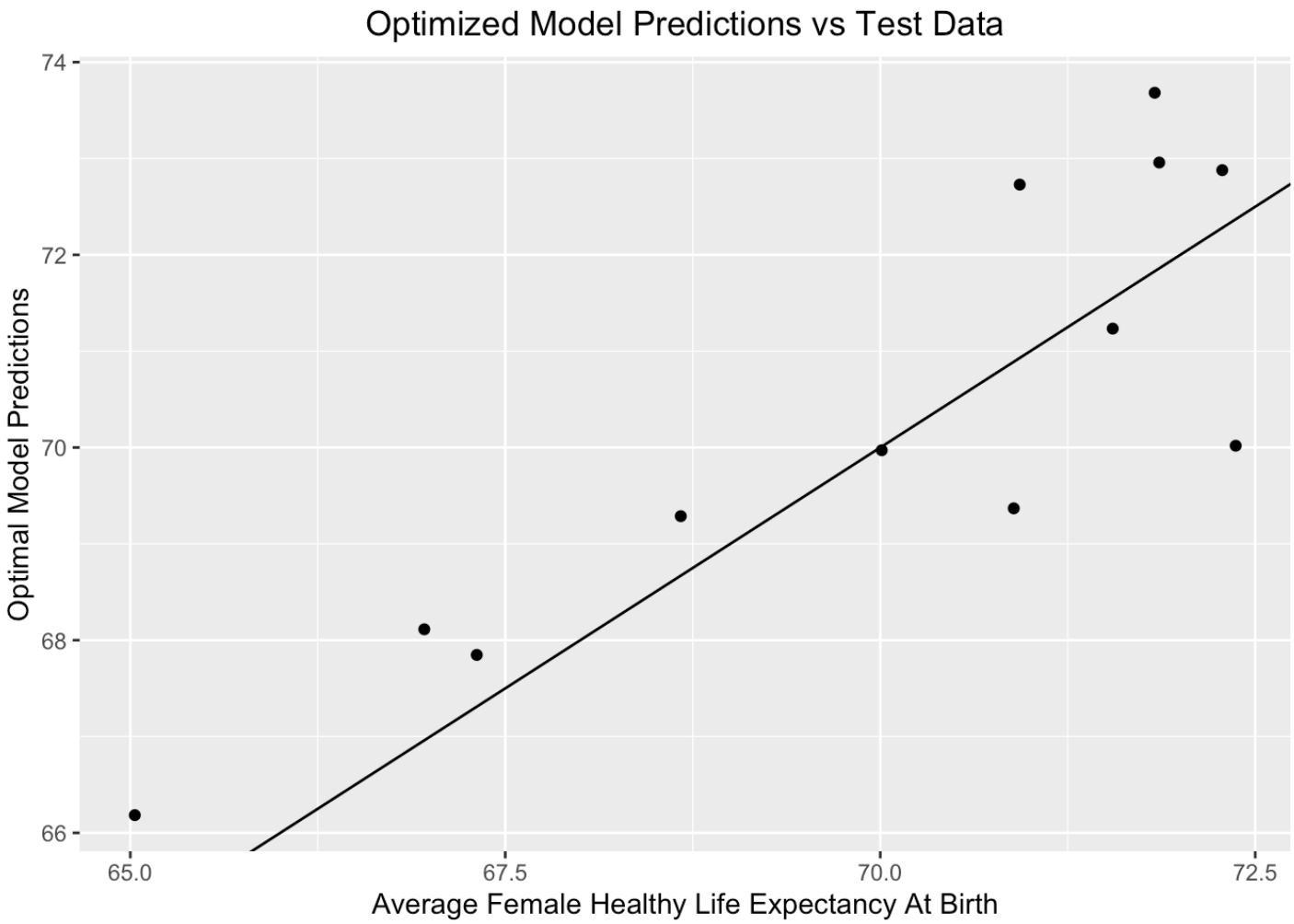

Normal Q-Q
Standardized residuals
32
35
3
-2
-1
0
1
2

Theoretical Quantiles
lm(HALE_Female ~ UHC + Adolescent_Birth_Rate + Medical_Doctors)





Finally, to better visualize the accuracy of our model when tested against the test data, we plotted the data of predicted values from the model versus the actual values in our test data set in a scatter plot (see below). We added a 45-degree line to represent a perfect model. As you can see, our model generally does a good job of predicting female HALE, as the predictions hover closely to the line. However, we see that at the upper extremes, the optimal model tends to slightly overestimate more often than it underestimates female HALE.



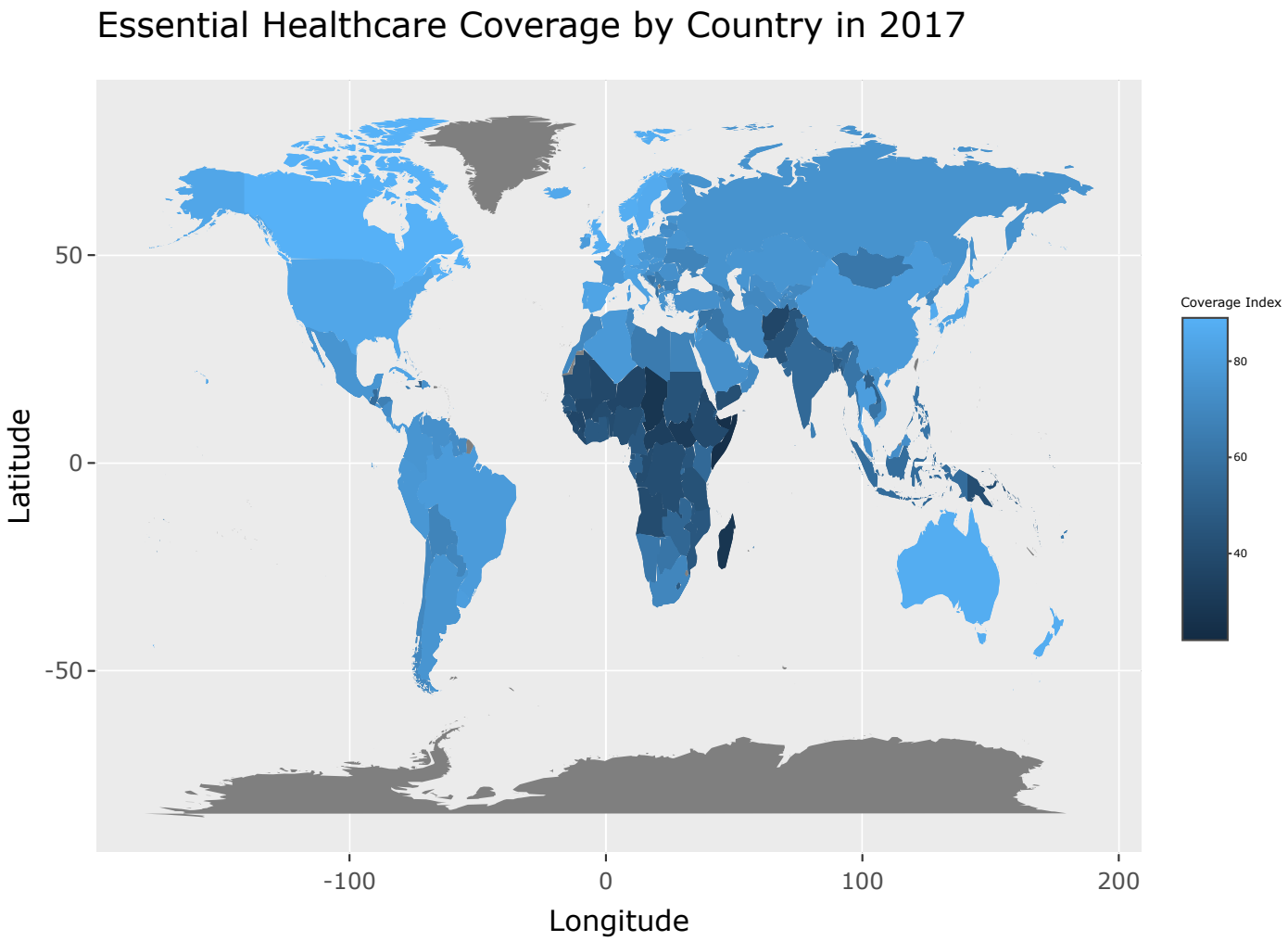
Part 2

To answer the first part of Question 2, we decided to develop a global map of UHC, or universal healthcare coverage, over time. We chose UHC because it was the most significant predictor of female HALE. To begin, we imported a publicly available global basemap shapefile from `rworldmap`. This dataset contains columns for latitude, longitude, and country. The dataset had enough geographical data points for R to map out the outlines of each country. After cleaning the dataset with the UHC coverage data, we joined the two datasets together using the `full_join` function.

Next, we used the `geom_polygon` function in `ggplot2` to visualize our distribution. When it became apparent that some countries' data was missing, we went back to each dataset and renamed the countries that had differing names and thus were overlooked during the join. These countries showed up as grey outlines in our map due to the missing data. Countries had different names in each dataset due to a variety of factors, such as reporting longer titles for each country (e.g. "Republic of Korea" vs. "South Korea") or recent changes in country names (e.g. "Czechia vs. "Czech Republic"). Some areas were not countries at all, such as part of the Sahara, and so they remained grey. (Taiwan was simply missing from the UHC dataset, but the country has excellent universal healthcare coverage.)

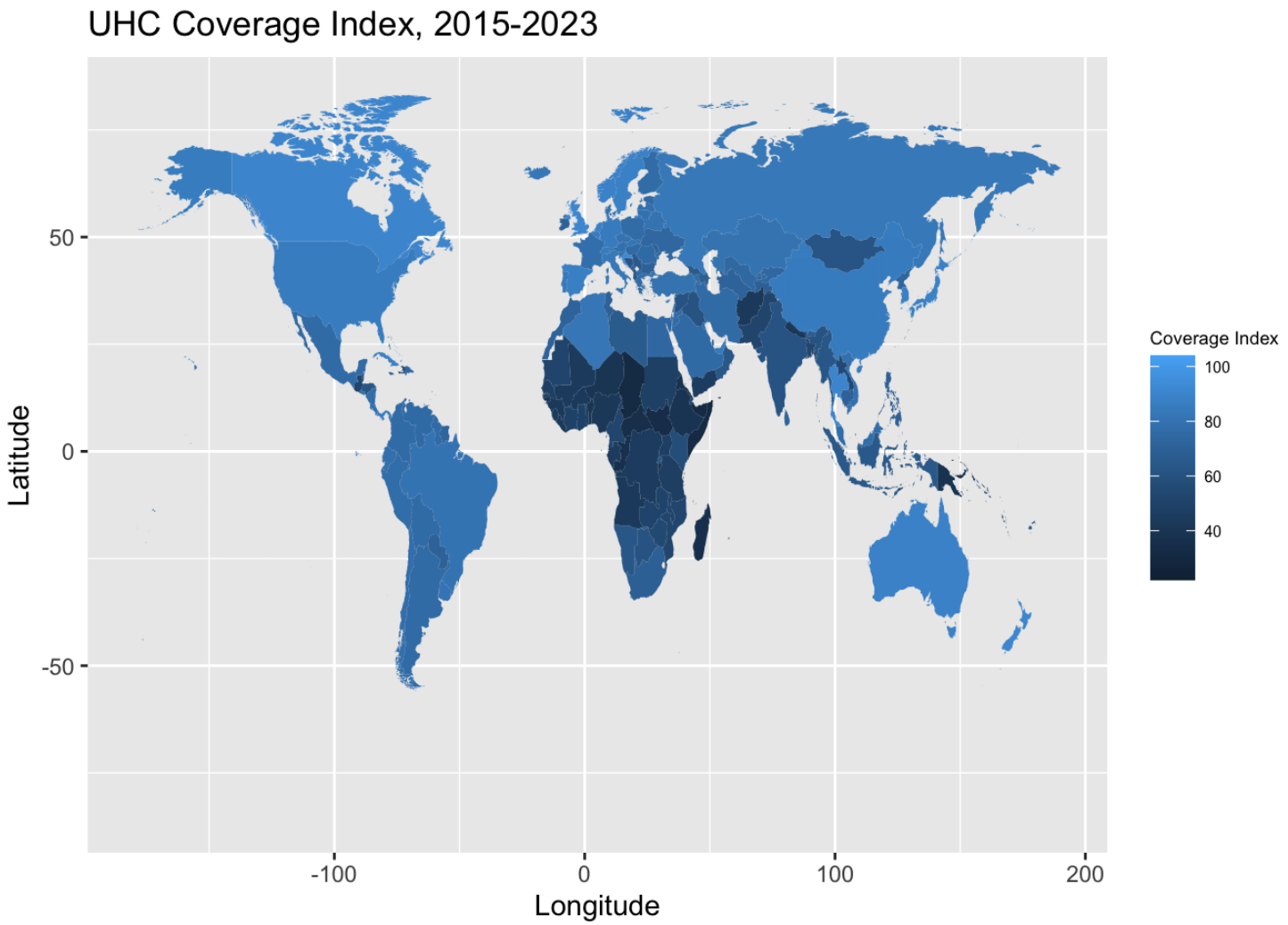
After this, we re-plotted the dataset with the formerly missing areas included and colored in blue to indicate their healthcare coverage data. We experimented with several visualizations after this point. First, we downloaded the `plotly` package and used it to make an interactive map of the distribution of healthcare coverage for the most recent year in the dataset, 2017. The `plotly` package allows users to zoom in on portions of the map, and click and drag to navigate. Next, we plotted a map that showed data from both 2015

and 2017 to compare. After this, we plotted a map that showed all this data, but was animated so that it would gradually transition from 2015 coverage to 2017 coverage. We used the `gganimation` package to complete the animation. We used the `transition_manual` argument and specified that the transitions should depend on the time data in the dataset. The `plotly` map is shown below for 2017, the most recent year in the dataset. (Note: Since this report was submitted as a PDF, we are attaching a Github link at the bottom so that you can view the `plotly` map and the GIF animation.Please click the link to download our RMarkdown file and knit to HTML to view the maps.)



Next, we decided to plot changes in healthcare coverage between 2015 and 2017. First, we took the difference in coverage and added a new column in the original dataset using the “mutate” function. Then, we plotted the changes the same way we plotted the original world maps (so no animation this time). If you look closely, you can spot the country that experienced the greatest change in UHC coverage–it’s Moldova!

Finally, to answer the second part of our Question 2, we decided to project changes in UHC over time, into the future, using a simple predictive model. Instead of using the years as integers for the modeling, we mutated the dataset to include a column showing the number of years after 2015, which was the regressor used. For every country, we created a linear model for its UHC coverage from 2015 and 2017, and then projected this model on the years 2019, 2021, and 2023. After adding the additional years to the dataset, we went through the same steps of renaming countries, joining the datasets, plotting using `geom_polygon` , and animating using `transition_manual` . A GIF animation of the map from 2015 to 2023 is below. Watch closely to see the changes!



Since we only had two years' worth of data in the dataset (data for 2015 and data for 2017), our projections are less reliable than they otherwise would be, especially because we are predicting until 2023. The predictions for the earlier years are likely the most accurate, and would begin to show increased error the farther out we project. However, they are still useful for showing which areas in the world that are most in need of healthcare coverage might get better over time, into the future. We see a general trend that as the years progress, UHC coverage increases across the globe.

CONCLUSION

In this project, the group investigated two questions. The first was: Which health variables are most significant in predicting female HALE? The second was: How has universal healthcare coverage changed over time on a global scale and can we predict how UHC will change into the future? In answering. For our first question, after finding the best linear model with three variables (UHC, adolescent birth rate and number of medical doctors) we found that UHC could be the most significant predictor. This led into our second question, where we continued to explore the global distribution of healthcare coverage over time and into the future. Not surprisingly, we found that more developed countries typically provide more healthcare coverage, which can lead to higher healthy average life expectancies for the general population and women in particular. We also found in the future, UHC is predicted to increase overall.

A significant limitation in our analysis was the poor data availability from the WHO. While the WHO does amass large amounts of health data, they do not necessarily collect data on variables that might be more useful in

assessing women's health (such as contraceptive availability or domestic abuse rates). However, though our model for the first question only includes three variables, we believe it can still play a vital role in the real world. Member State governments or other health organizations can use this model as a foundation on which to consider how to improve females' well-being. Besides, how can countries enhance females' status? Which element should countries care the most about in order to reduce the differential treatment between men and women? What are the most significant factors that lead to the gap of HALE between men and women between each country? After including more variables into this model, researchers could use train and test data to estimate and solve for these questions.

We looked into what variables might be more useful in future analysis. One of our findings was that the proportion of women in the labor market has increased year by year. In the report called Women's Trends from ILO (International Labour Organization), we can find that the gap in employment rates between men and women in developing countries is the smallest, followed by developed countries and third in emerging countries. Most developed countries and developing countries have similar HALE values. Some countries in Africa or Asia have smaller HALE values than other countries. This brings the question whether the employment rate of females could be a significant variable, whether national strength influences female HALE, or whether the country's Gross Domestic Product, education level, or population has a significant impact on female healthy average life expectancy. To achieve several goals mentioned before, researchers require more dataset with variables such as Gross Domestic Product, the length of education, and employment rate in each country all over the world. Therefore, there are more variables researchers could investigate into to build a better model.

There are some ways that our model can be improved for future analyses. One key way we have already discussed is to consider a greater number of relevant variables. Besides this, future modeling processes might also want to add interaction terms between variables (for example, the number of medical doctors and the UHC index likely has some interaction because they are related concepts). We may also want to consider using a nonlinear model by squaring or taking the log of certain variables. This may produce a more significant and accurate predictive model. For the predictive model for UHC in the future, future models should definitely investigate a greater number of factors that affect decisions around providing UHC, including government stability, funding, and political leaning. An interesting question to look into would be, how has COVID affected this? This might make these estimates less reliable and disrupt the current linear pattern.

Women's health remains an important matter as countries develop. Our topic can be extended further by looking at changes in female HALE and universal healthcare index on a more granular level (ex. by state or province instead of by nation) in order to identify hotspots of public health concern. Universal healthcare poses a particularly interesting possibility, as the provision of healthcare services is a matter of controversy especially in the United States. Investigating the spatial distribution of UHC or looking at the relationship between UHC and other key metrics of health besides life expectancy could provide grounds for more compelling arguments to fund universal healthcare.

GitHub Link to Final Paper:https://github.com/JessB2/Global-Female-Life-Expectancy/blob/main/Final_Report_Template%20(1).Rmd (https://github.com/JessB2/Global-Female-Life-Expectancy/blob/main/Final_Report_Template%20(1).Rmd)