STAT 306 Group Project:
Effects of Health Expenditure, Income and Alcohol Consumption on Life Expectancy

James Shaw, Sophia Zhang, Eric Gu, Zach Vavasour

Introduction

Although average human life expectancy has been increasing over time, there is still significant variance from country to country. What is it that makes one country's life expectancy greater than another? With greater knowledge of the metrics that most affect the overall health of a population, perhaps we can deploy more efficient policy to extend lifespan.

Data

To explore this topic, we analyzed life expectancy - and other metrics - across 140 countries from the World Health Organization's (WHO) publicly available data repository. The original dataset that we proposed to use for analysis was found to have a significant number of errors (mislabeled categorical variables, inaccurate GDPs, percentage values in the 1000s). This set had been collected by members of the Kaggle community, where the original data was pulled from the public WHO and United Nations (UN) websites.

As a result of these errors, we compiled our own dataset drawing from the same sources to create a more accurate version of what was outlined in the original proposal. The major difference between these two sets is that the categorical variable of “Developing Status” was omitted and instead replaced with the more descriptive variable “Adjusted net national income per capita” which provides a sense of the country’s level of industrialization. Overall, the new dataset is more reliable and traceable.

Description of Variables

The following variables were collected from The World Bank’s *Data Bank,* an open source global metric aggregator. More information about the aggregation methods and statistical concepts related to the selected measurements can be found in Appendix A.

Life expectancy at birth

Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. This metric was selected as the response variable.

(1) United Nations Population Division. World Population Prospects: 2019 Revision, or derived from male and female life expectancy at birth from sources such as: (2) Census reports and other statistical publications from national statistical offices, (3) Eurostat: Demographic Statistics, (4) United Nations Statistical Division. Population and Vital Statistics Report (various years), (5) U.S. Census Bureau: International Database, and (6) Secretariat of the Pacific Community: Statistics and Demography Programme.

Domestic general government health expenditure (% of GDP)

Public expenditure on health from domestic sources as a share of the economy as measured by GDP. This estimate includes healthcare goods and services consumed on a yearly basis without capital expenditures such as building, machinery, IT and stocks of vaccines for emergency outbreaks (WHO, 2018). This metric was selected as a covariate to help understand if a country's health and spending policy is correlated to life span.

World Health Organization Global Health Expenditure database (http://apps.who.int/nha/database).

Adjusted net national income per capita

Adjusted net national income is Gross National Income (GNI) minus consumption of fixed capital and natural resources depletion. This variable is measured in 2010 US dollars. This covariate was added to determine if the wealth of a country release to our selected response.

World Bank staff estimates based on sources and methods in World Bank's "The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium" (2011).

Total alcohol consumption per capita

Total alcohol per capita consumption is defined as the total (sum of recorded and unrecorded alcohol) amount of alcohol consumed per person (15 years of age or older) over a calendar year, in litres of pure alcohol, adjusted for tourist consumption. Due to its prevalence and variance in consumption levels across the globe, we selected alcohol

consumption as a covariate.

World Health Organization, Global Health Observatory Data Repository (http://apps.who.int/ghodata/).

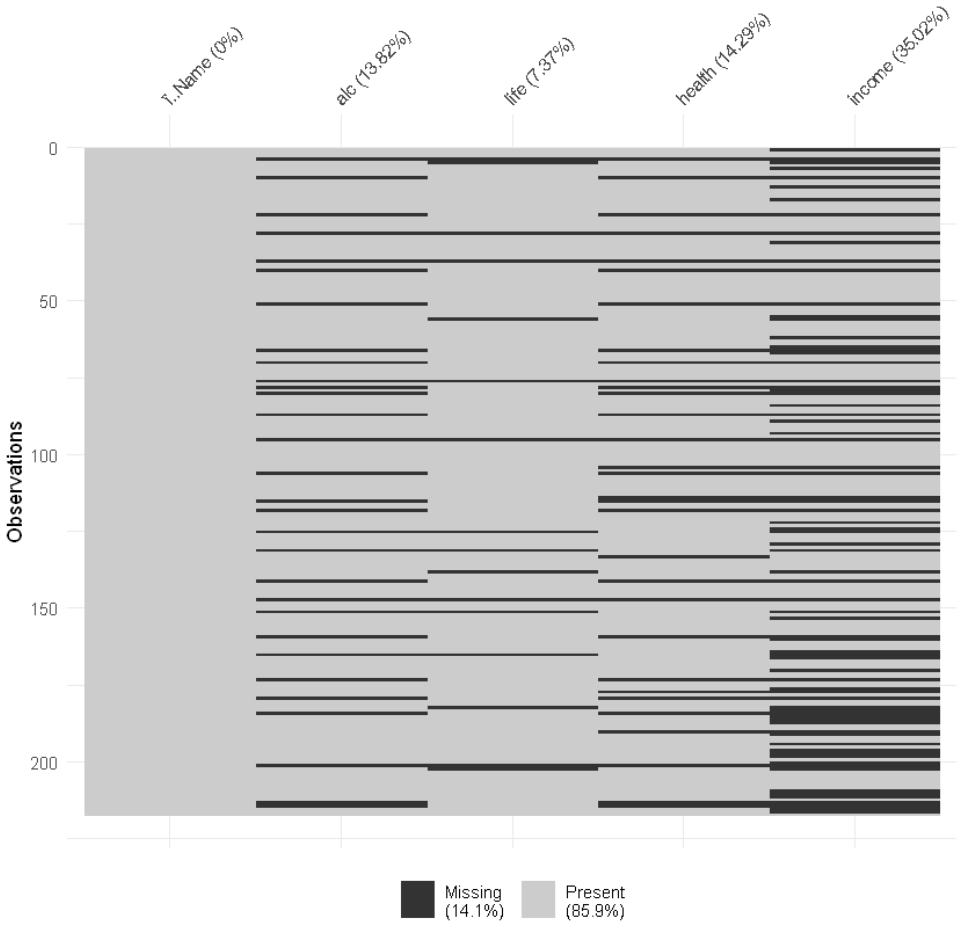
Research Question

We explored which of the selected metrics affect life expectancy in a country. To do this, we created predictive models using the above variables with life expectancy as our response. Our analysis answers which explanatory variables have a significant effect on the overall life expectancy of a country.

Analysis

Preprocessing

Before beginning our analysis, we removed 77 countries from the dataset due to missing data in one of the variables described above. Many of these omissions were from lack of income per capita measurements. See the table below for a breakdown of which variables were missing.

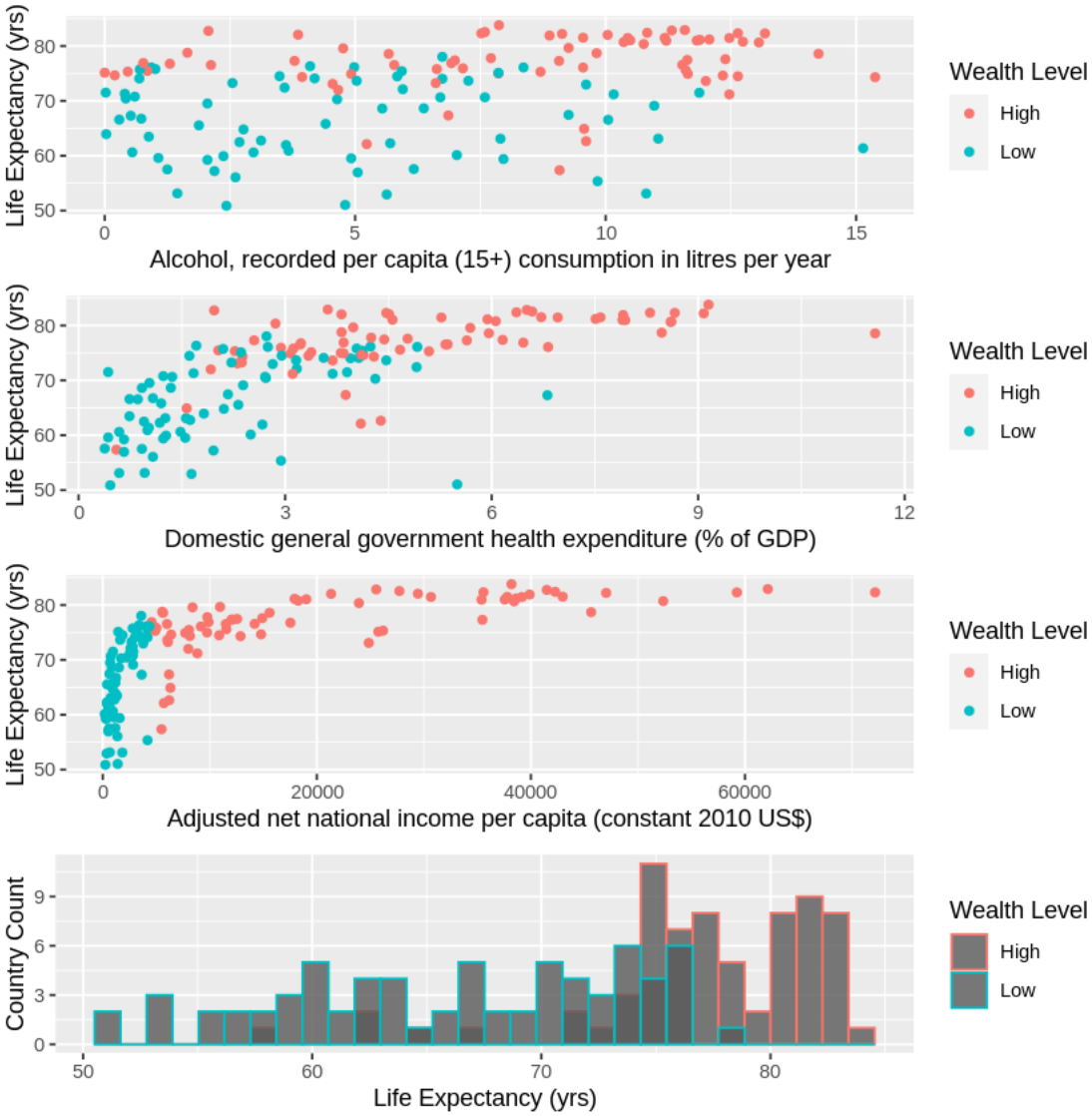


Figs 1: Summary of missing data in the World Bank Data

Secondly, we created a new categorical variable from the income level, to differentiate between countries with net income per capita higher or lower than the median. Though this variable is not modelled later in the regression, it provides a general sense of the level of wealth among the countries of interest.

Visualizations

After the two preprocessing steps described above, we began by visualizing the relationships between each of the individual variables and life expectancy.



Figs 2,3,4,5: Life expectancy vs selected covariates and a histogram of the life expectancy for the countries of interest

From the first plot, there does not seem to be an apparent relationship between alcohol consumption and life expectancy; the data points are scattered evenly. However, it's obvious to tell that countries with a 'High' income level tend to have higher life expectancy.

The second scatter plot clearly suggests a relationship between health expenditure and life expectancy. Similar to the first plot, countries with a 'High' income level inclines to have higher life expectancy.

The third plot also clearly displays a relationship between income per capita and life expectancy. Furthermore, it seems that the life expectancy levels off at slightly above the median income per capita value. Similar to the second plot, countries with a 'High' income level inclines to have

higher life expectancy. Nevertheless, this plot appears to have a smaller variance for life expectancy between points that belong to the same wealth level.

The final histogram visualizes the relationship between a country's income status (High/Low) and its life expectancy. We can see that all the countries with a high income status are clearly at the higher end of life expectancy.

Creating the Model

To explore and predict life expectancy, we created a series of linear models and used Adjusted R^2, the C_p statistic, and the AIC value to select the best one. We chose to consider quadratic terms for each of the continuous variables, as well as interaction terms between a country's income per capita and other terms.

Using 'regsubsets' in R we performed exhaustive model selection search with the following result:

Fig 6: Output of the regsubsets model selection function for the covariates described

	(Intercept)	alc	health	income	alc_sq	health_sq	income_sq	alc_health	alc_income	health_income
1	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
3	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
4	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
5	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
6	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE
7	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
8	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE

To select which model is best, we first examined the Mallows C_p statistic for each model. Models 4, 5, and 6 have the lowest three C_p statistics at 4.44, 3.34, and 4.30 respectively.

Model 4: Health + Income + Health^2 + Income^2
Model 5: Health + Income + Health^2 + Income^2 + Health:Income
Model 6: Health + Income + Health^2 + Income^2 + Alcohol^2 + Health:Income

To further narrow down our model selection, we then turned to Adjusted R^2, and AIC. The table below shows the model selection statistics for our three competing models. Moreover, we examined the standardized residual plot for each model to examine their normality.

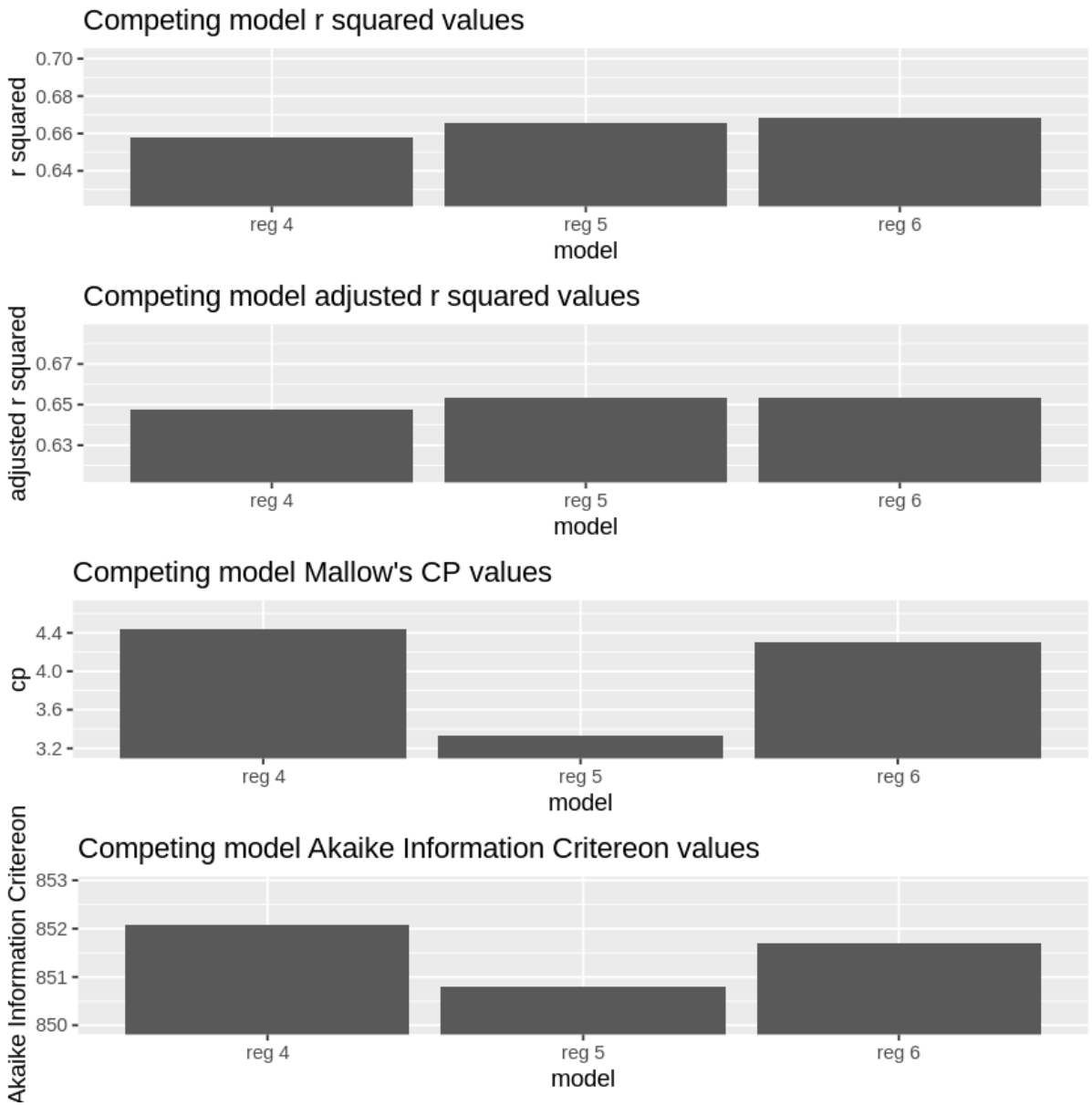
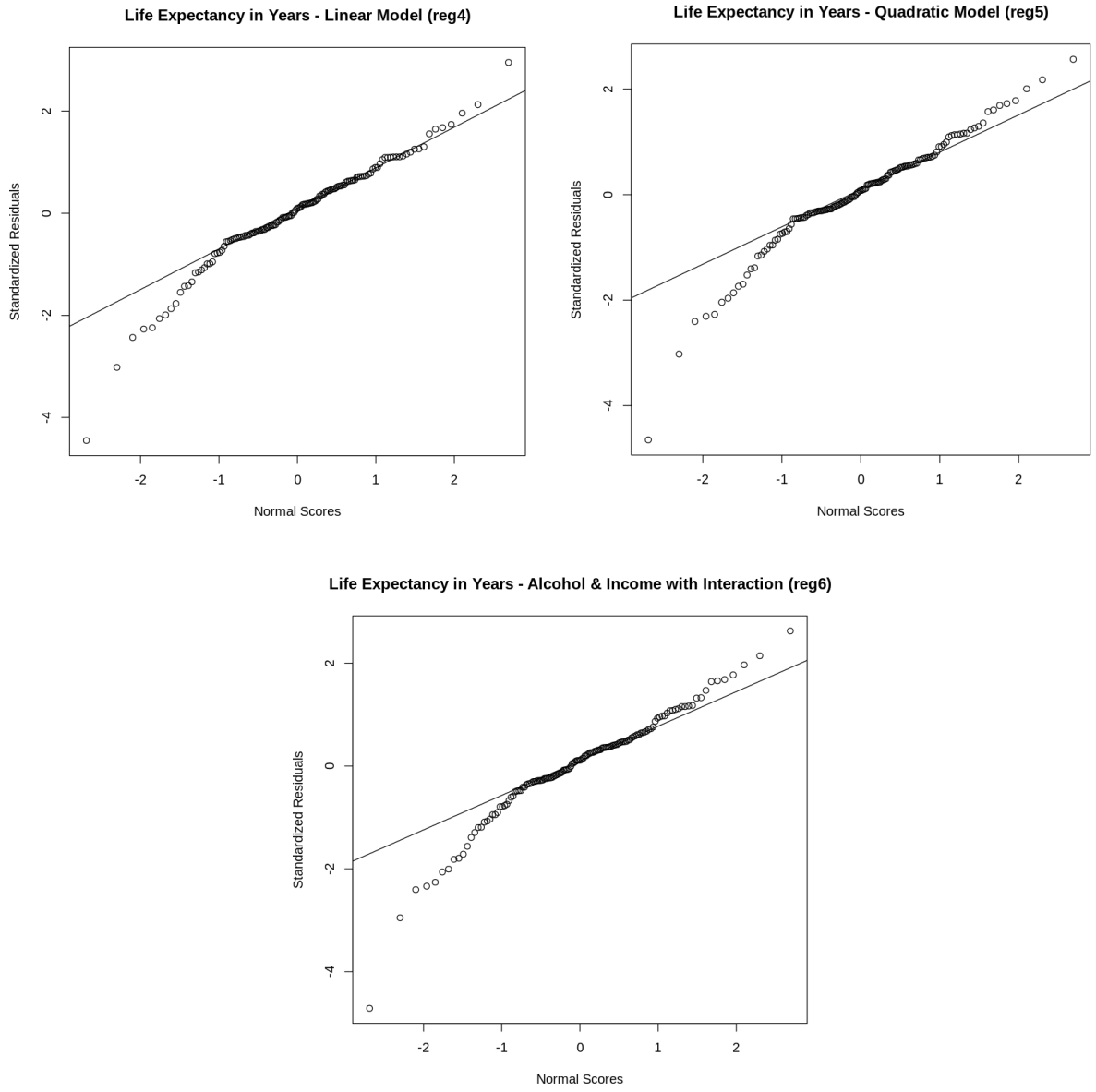


Fig 7: Comparison of the model selection statistics for models 4,5 and 6, as outlined above



Figs 8, 9 and 10: Quantile-Quantile plots for the three competing models outlines above

Each of the above standardized residual plots show a pattern suggestive of residuals deviating away from normality, indicating heavy-tailed data. These tables suggest that the life expectancy of countries does not follow a normal distribution and instead has data potentially indicating a t-distribution. We chose **model 5** since it has the lowest AIC, lowest C_p statistic, and highest adjusted R^2.

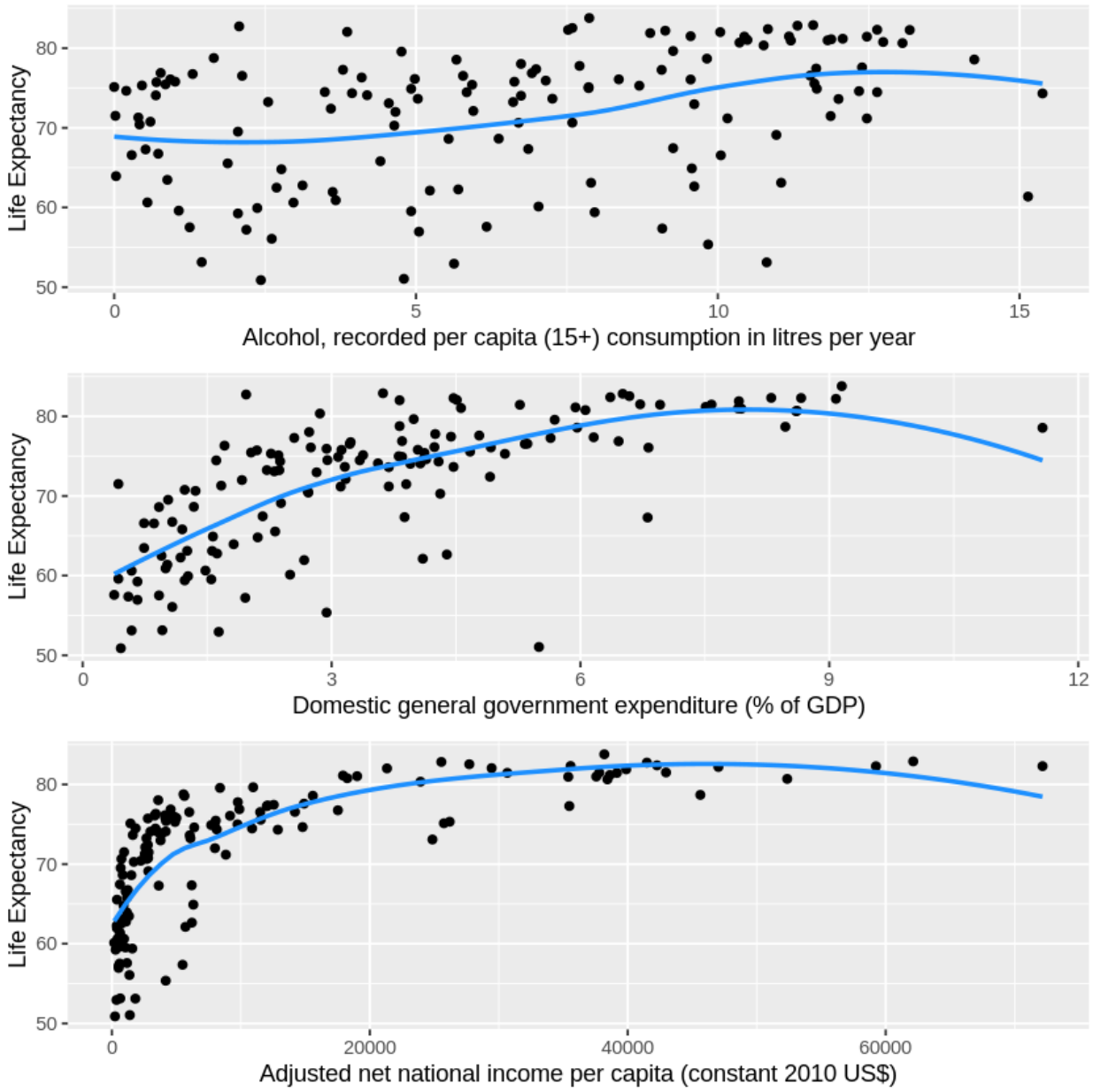
The final model is as follows:

$$\mathbf{Y = 58.8 + 3.77x_1 + 0.000612x_2 + -0.232x_1^2 + -0.00000000482x_2^2 + -0.0000261x_1x_2}$$

Where:
Y: Life Expectancy at Birth
x_1: Domestic general government health expenditure (% of GDP)
x_2: Adjusted net national income per capita (2010 USD)

RSE: 4.911 on 134 degrees of freedom
Adjusted R-squared: 0.6531

Coefficient p-values:
Health: 5.87e-08
Income: 6.51e-08
Health^2: 0.00194
Income^2: 0.00970
Health:Income: 0.07749



Figs 11, 12 and 13: Smoothed model (reg5) response overlaid over the countries of interest

Above we see the smoothed fit of model five against the three covariates of interest. The line is shown is smoothed for ease of interpretation, since the life expectancy is also related to the other two covariates which are not plotted, leading to an erratic line. Visually, the model fits the data with reasonable accuracy.

Conclusion

With our best model we achieve an R^2 of 0.6531, implying that ~65% of the variance in a country's life expectancy is explained by our model. Furthermore, the low p-values for the Health and Income terms imply that these terms have a significant effect on the overall life expectancy of a country. We can reject the null hypothesis that life expectancy does not depend on health and income. Since alcohol consumption was not significant enough to be included in our model, we can not reject the null hypothesis, and thus can't conclude that it has an effect on life expectancy.

An important decision made in the analysis of life expectancies was the removal of countries who were missing any of the covariates. Approximately ⅓ of the available countries were removed from analysis and were therefore not represented in our model. This will have incorporated bias in our model, leading to a data sample which is likely not to be an accurate representation of the global population as a whole. Also, one potential reason that some of these countries are missing data is due to lack of adequate funding for health surveying, indicating that low-income and low-GDP countries will be underrepresented in the data set. Regardless, since ⅔ of the countries are being represented in the data, the conclusions can still provide insight.

Overall, our results are not particularly surprising. It should be expected that richer countries tend to have greater life expectancy; however, one thing we can conclude is that a country's expenditure on healthcare is the greatest factor in its citizens' life expectancies (as the coefficient is much greater) among the covariates that we explored. One possible interpretation of this relationship is that countries whose health policies allocate a larger amount of funds into healthcare would have more effective treatments for its citizens, leading to longer lifespans. It is important to note that more research and analysis would be required to make such a claim. Healthcare spending could for example, be a consequence of a culture's value for health. Increased lifespan could therefore be explained by a cultural inclination to take better care of the elderly or eat a more healthy diet. The relationships that are seen in model 5 are only a small number of the factors which may affect a country's life expectancy at birth .

Moreover, the amount of money spent on healthcare is not necessarily equivalent to how accessible healthcare is to an average citizen. The United States has the 7th highest health expenditure, yet you will often hear of people avoiding hospital visits out of fear of the costs that they might incur (Leonhardt, 2020). In reality, there is much more nuance to the state of healthcare in a country than how much is being spent per country, per capita. Further research could explore which areas of healthcare contribute the most to life expectancy. Perhaps examining the accessibility of healthcare or the socialized/privatized split of healthcare infrastructure would provide valuable insights into metrics like life expectancy.

Citations

Baffes, John, et al. “World Bank Open Data.” *Data*, 8 Apr. 2021, data.worldbank.org/.

Leonhardt, Megan. “Nearly 1 in 4 Americans Are Skipping Medical Care Because of the Cost.” *CNBC*, CNBC, 12 Mar. 2020, www.cnbc.com/2020/03/11/nearly-1-in-4-americans-are-skipping-medical-care-because-of-the-cost.html.

“Current Health Expenditure (% of GDP).” *Data*, data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS.

Rajarshi, Kumar. “Life Expectancy (WHO).” *Kaggle*, 10 Feb. 2018, www.kaggle.com/kumarajarshi/life-expectancy-who.

Appendix A - Covariate and Response Provenance (World Bank, 2015)

Indicator Name	Long definition	Source	Statistical concept and methodology	Development relevance	General comments
Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)	Total alcohol per capita consumption is defined as the total (sum of recorded and unrecorded alcohol) amount of alcohol consumed per person (15 years of age or older) over a calendar year, in litres of pure alcohol, adjusted for tourist consumption.	World Health Organization, Global Health Observatory Data Repository (http://apps.who.int/ghodata/).	The estimates for the total alcohol consumption are produced by summing up the 3-year average per capita (15+) recorded alcohol consumption and an estimate of per capita (15+) unrecorded alcohol consumption for a calendar year. Tourist consumption takes into account tourists visiting the country and inhabitants visiting other countries.	According to the World Health Organization, alcohol consumption is a causal factor in more than 200 disease and injury conditions. In the world, an estimated 3 million deaths are from harmful use of alcohols every year. Drinking alcohol is associated with a risk of developing health problems such as mental and behavioural disorders, including alcohol dependence, major noncommunicable diseases such as liver cirrhosis, some cancers and cardiovascular diseases, as well as injuries resulting from violence and road clashes and collisions.	This is the Sustainable Development Goal indicator 3.5.2[https://unstats.un.org/sdgs/metadata/].
Life expectancy at birth, total (years)	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.	(1) United Nations Population Division. World Population Prospects: 2019 Revision, or derived from male and female life expectancy at birth from sources such as: (2) Census reports and other statistical publications from national statistical offices, (3) Eurostat: Demographic Statistics, (4) United Nations Statistical Division. Population and Vital Statistics Report (various years), (5) U.S. Census Bureau: International Database, and (6) Secretariat of the Pacific Community: Statistics and Demography Programme.	Life expectancy at birth used here is the average number of years a newborn is expected to live if mortality patterns at the time of its birth remain constant in the future. It reflects the overall mortality level of a population, and summarizes the mortality pattern that prevails across all age groups in a given year. It is calculated in a period life table which provides a snapshot of a population's mortality pattern at a given time. It therefore does not reflect the mortality pattern that a person actually experiences during his/her life, which can be calculated in a cohort life table.	Mortality rates for different age groups (infants, children, and adults) and overall mortality indicators (life expectancy at birth or survival to a given age) are important indicators of health status in a country. Because data on the incidence and prevalence of diseases are frequently unavailable, mortality rates are often used to identify vulnerable populations. And they are among the indicators most frequently used to compare socioeconomic development across countries.	Annual data series from United Nations Population Division's World Population Prospects are interpolated data from 5-year period data. Therefore they may not reflect real events as much as observed data.

Domestic general government health expenditure (% of GDP)	**Public expenditure on health from domestic sources as a share of the economy as measured by GDP.**	**World Health Organization Global Health Expenditure database (http://apps.who.int/nha/database).**	**The health expenditure estimates have been prepared by the World Health Organization under the framework of the System of Health Accounts 2011 (SHA 2011). The Health SHA 2011 tracks all health spending in a given country over a defined period of time regardless of the entity or institution that financed and managed that spending. It generates consistent and comprehensive data on health spending in a country, which in turn can contribute to evidence-based policy-making.**	**Strengthening health financing is one objective of Sustainable Development Goal 3 (SDG target 3.c). The levels and trends of health expenditure data identify key issues such as weaknesses and strengths and areas that need investment, for instance additional health facilities, better health information systems, or better trained human resources. Health financing is also critical for reaching universal health coverage (UHC) defined as all people obtaining the quality health services they need without suffering financial hardship (SDG 3.8). The data on out-of-pocket spending is a key indicator with regard to financial protection and hence of progress towards UHC.**	
Adjusted net national income per capita (constant 2010 US$)	**Adjusted net national income is GNI minus consumption of fixed capital and natural resources depletion.**	**World Bank staff estimates based on sources and methods in World Bank's "The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium" (2011).**	**Weighted Average**		