JSC370 Midterm Project

Impact of smoking, alcohol consumption, and happiness on life expectancy around the world

Christopher Matthew

2022-02-28

Contents

List of Figures

Introduction

How does smoking, alcohol consumption, and happiness effect life expectancy at birth of general world population? It is a common knowledge that smoking, drinking excessive amount of alcohol, or being sad and depressed will negatively impact your health but just how detrimental they will affect your life. This paper will emphasize how life expectancy get affected across various quantities of smoking, alcohol consumption, and happiness in the countries of the world and how I am going to generalize the results to the general population.

Even though smoking reduces someone's average life expectancy by 2.5 years in Canada (Manuel, D. G., 2012, p.1), alcohol reduces it by 1.5 years in Europe (Janssen et al., 2021, p.937), USA drink plenty of alcohol and smoke, yet they have a relatively high life expectancy. In addition, compared to very happy people, the risk of death over the follow-up period is 14% (95% CI 1.06 – 1.22) higher among those who are not happy (Lawrence et al., 2015). Interestingly, intake of both smoking and alcohol only contributes to -0.07 years in life expectancy of Nordic countries (Östergren et al., 2019, p.336). This piqued my interest in pursuing this topic because the trend of life expectancy as smoking, alcohol consumption, and happiness fluctuates is still unknown for general population.

Methods

Data Collection

I selected these variables based on some background research that I believe that the variables will have a high correlation with life expectancy in general. The initial regression model contains life expectancy at birth as the response variable while the predictor variables include smoking, alcohol consumption, and happiness.

Life expectancy at birth, denoted as life_expectancy_at_birth reflects the overall mortality level of a population. This represents the average number of years that a newborn is expected to live based on current mortality rates. This data is collected from https://data.worldbank.org/indicator/SP.DYN.LE00.IN

Smoking rates by country, denoted as smokepercentage, is defined as the proportion of population over the minimum smoking age that smokes regularly in the country is one of the well-known causes of reduction in life expectancy (Manuel, D. G., 2012, p.1 and Janssen et al., 2021). This data is collected from https://worldpopulationreview.com/country-rankings/smoking-rates-by-country

Alcohol consumption per capita, denoted as alcohol consumption, is defined as number of pure alcohol litres consumed every year per capita over legal drinking age also increases mortality and cancer risk in older adults (Kunzmann et al., 2018). This data is collected from https://data.worldbank.org/indicator/SH.ALC.PCAP.LI

World happiness index, denoted as happiness, is defined by average answers to the main life evaluation question known as Cantril ladder, which asks respondents to think of a ladder with the best possible life for them being a 10 and worst possible life for them being a 0 and rate themselves. This data is collected from https://www.kaggle.com/datasets/unsdsn/world-happiness

Each of these datasets are taken from year 2018 for consistency. We will also use a Country-Continent Dataset to group each of the countries based on their continents.

Data Wrangling

To begin with, I imported several libraries that I am going to use for the data analysis which includes tidyverse, dplyr, mgcv, and car ??????????????. After collecting the data from the sources containing similar time stamp, I used read.csv method to save the dataset into the environment data. The datasets are labeled as smoking, alcohol, happiness, continent, and life_expectancy. Then, for each of the dataset, I only selected the country column and the value index column for smokepercentage, alcohol_consumption, happiness and life_expectancy_at_birth respectively, all from year 2018 for consistency, and remove other unnecessary columns. This is done with the help of subset and colnames function from the base R package.

Next, for each of the datasets, I removed any missing values because imputing the values with median or mean is not suitable given that our information is limited. Next, I remove outliers for each of the datasets using my own defined outliers function which removes anything above or below the upper or lower quartile $\pm$ 1.5 Interquartile Range. In the end, I combined all of the datasets into a combined dataset grouped by each country using the group_by function and remove any countries that does not contain all three information. During each step, I also make sure that everything is going according to the plan by looking at the summaries and environment tab.

Linear Modelling

Afterwards, I am going to perform EDA by checking scatterplot between response and predictor and creating boxplots and histograms of the variables. I will also check for any nonlinear trend, clusters, discernible patterns, and normality of distribution of residuals to verify linear model assumptions.

First, I will take a look into the linearity of each pair of independent and dependent variables using the scatterplots to see if linearity seems to be present or not. Next, I am going to check normality using histogram, QQ-Plot, and a goodness of fit test. In case the data is not normally distributed, I will use boxcox power transformation to generate a better model regarding the normality assumption and see if the problem will be resolved or not. Subsequently, I will check for multicollinearity by looking at the VIF for the predictors and report any detrimental multicollinearity if they exist. Conclusively, I will check for homoscedasticity and also do appropriate adjustments if needed.

After fulfilling model assumptions, we will check all possible models (1, 2, or 3 predictors) to evaluate their significance and Adjusted R^2 and conduct partial F-test to find the best model. Note that since there are only 7
non-empty distinct subsets of the predictors, we will just manually check all three since using other selection methods like AIC or automated selection is inefficient.

In the end, we will proceed with fitting a linear model to see our result and performance of the linear model and compare it with advanced regression model with cubic regression splines on each non-empty distinct subsets of the predictor variables.

Machine Learning

Results

Linear Modelling

After the initial data cleaning, we have 112 observations in our combined dataset, none of which contains any missing values. Stepping into the Exploratory Data Analysis, Figure 1 shows that the distribution for smoking is like normal distribution while alcohol and life expectancy have a right skewed and left skewed non-normal distribution respectively. The value for each variable seems to have a big variance and there aren't any noticeable extreme observations.

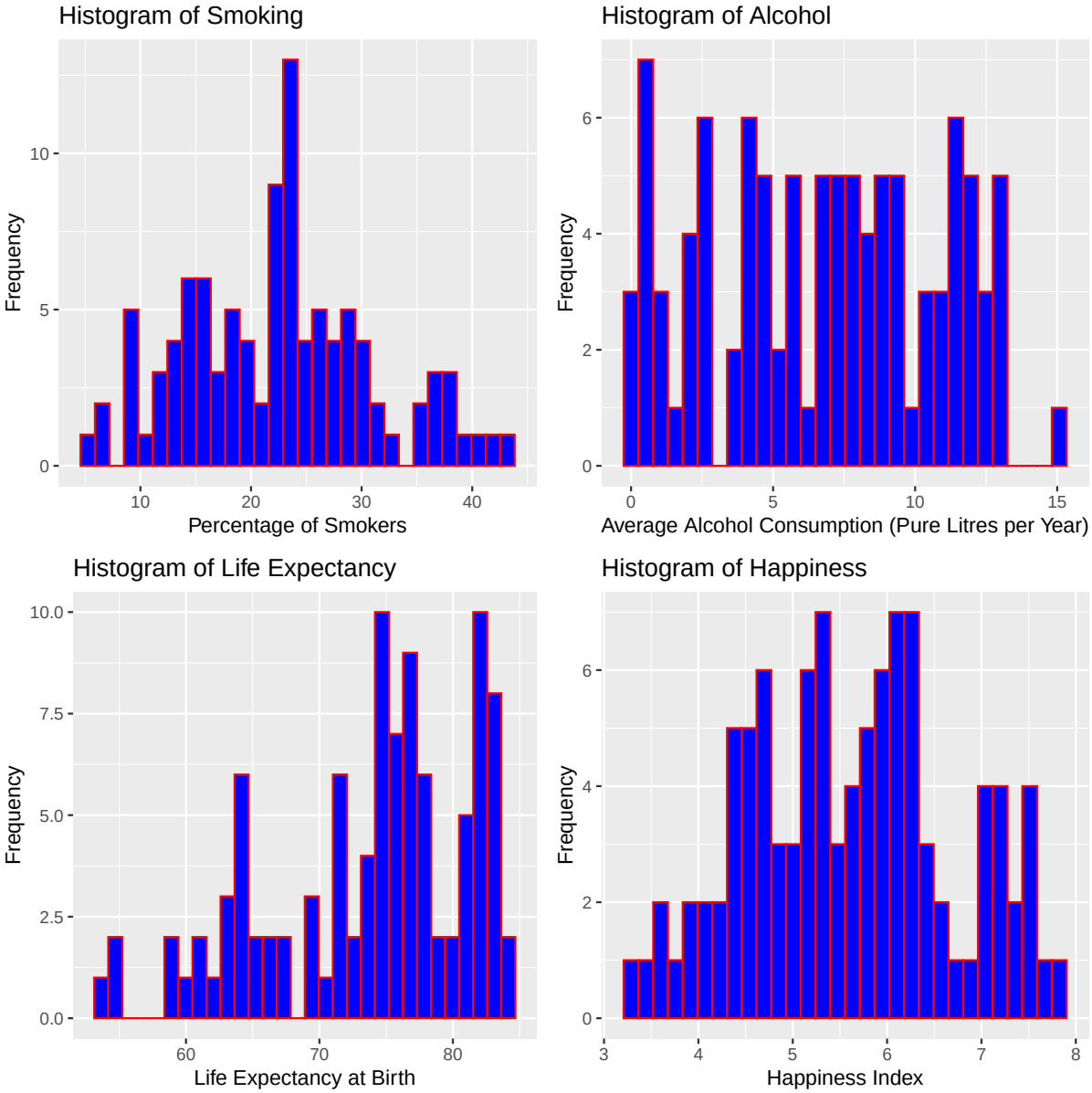

Figure 1: Boxplots and Histograms of Percentage of Smokers, Average Alcohol Consumption, and Life Expectancy at Birth

Next, Figure 2 shows that linearity assumption between each pair of independent variables and dependent variable other than Happiness Index is not very strong and we have to note this as our final result might not be accurate.

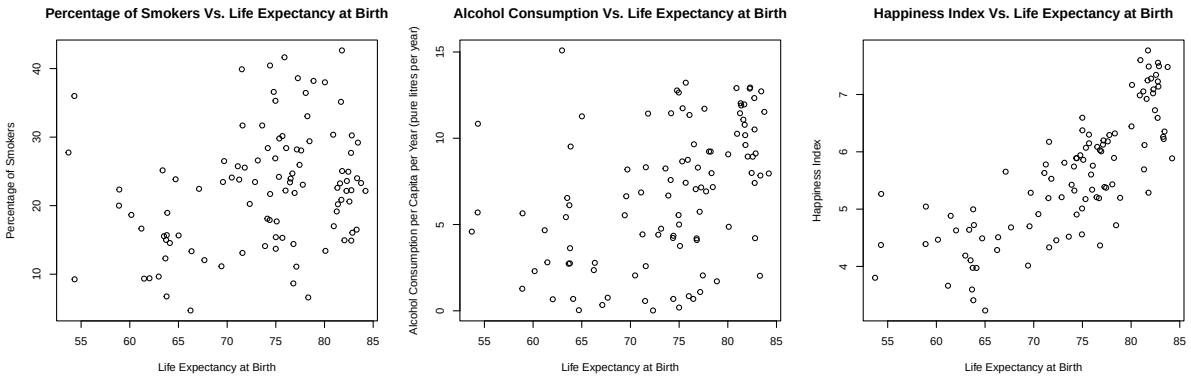

Figure 2: Scatterplots between Predictor Variables and Response Variable

Now, looking to Figure 3, we can see that the fitted against the real value seems to have a good linear relationship and the pattern in the residual plot have slight fanning which shows a problem in non-constant variance and there also exists some observations far from zero. Q-Q plot shows that normality seems to be fulfilled and it appears that smoking and alcohol also have a slight linear relationship which will be covered later in multicollinearity.

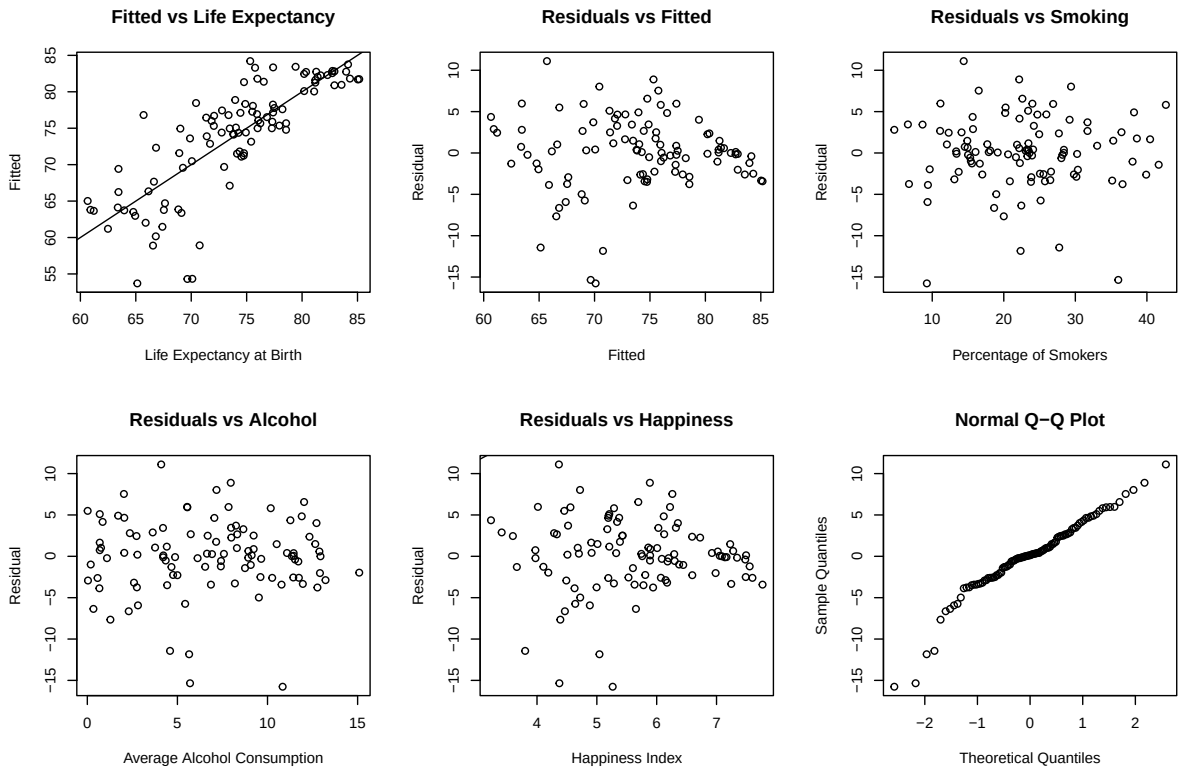

Figure 3: Fitted and Residual Plots, QQ-Plot, Predictors Plot Before Transformation

Based on our observation, we will do a power transformation to the variables, and it shows that we must increase the exponent of our response variable to the power of five. After transformation, we can see that in Figure 3, our residual plots seem to have a better pattern in homoscedasticity and a slightly better QQ-Plot. Our fitted value also has a better linear relationship with the actual value compared to the previous chart. Note that the VIF between the predictors are below 1.3 which is a very good indicator that there is no multicollinearity between the predictors.

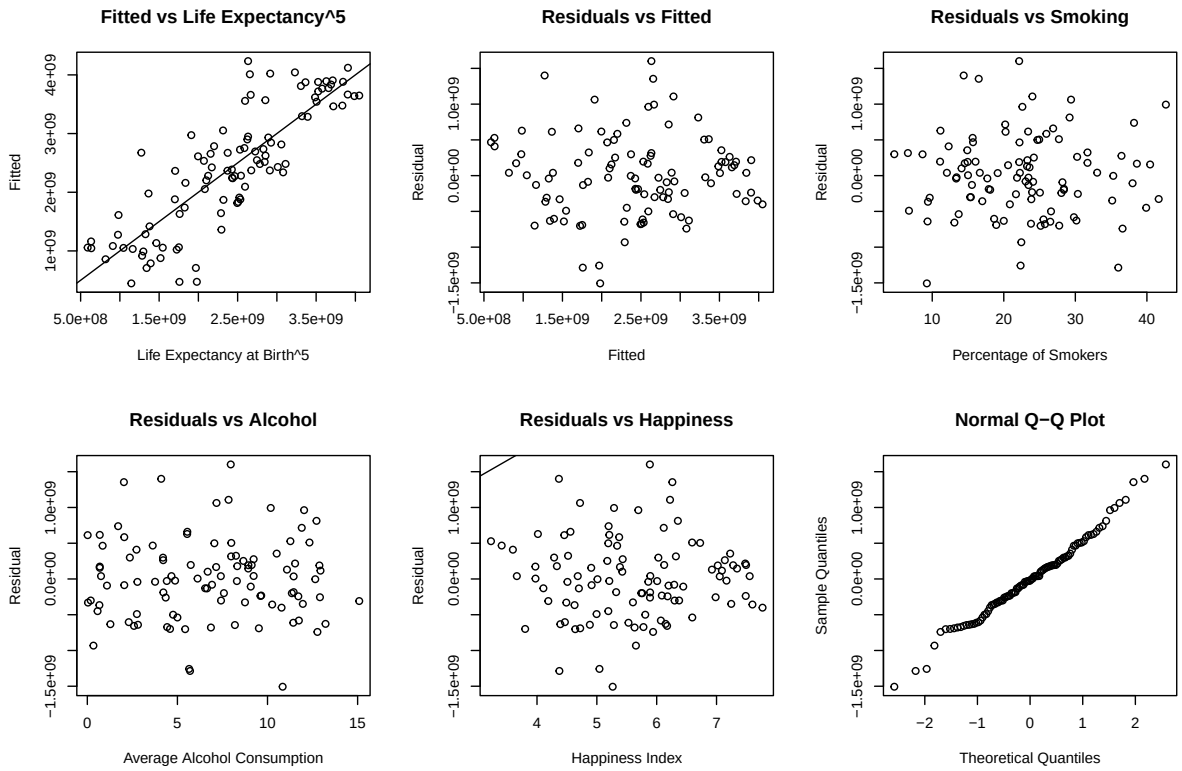

Figure 4: Fitted and Residual Plots, QQ-Plot, Predictors Plot After BoxCox Transformation

Now that our assumptions have been fulfilled, on Appendix 1, we run different models with all non-empty subsets of the predictors to find out that the model with both predictors have a low adjusted R^2 of 0.6964 which is better than other models as you can see from the table below. Hence, we are going to proceed with the model containing 3 predictors.

Table 1: Adjusted R^2 for Linear Models with Different Predictors

Model	Adjusted R-Squared
All Three Predictors	0.696
Percentage of Smokers and Average Alcohol Consumption	0.210
Average Alcohol Consumption and Happiness	0.672
Percentage of Smokers and Happiness	0.693
Percentage of Smokers only	0.063
Average Alcohol Consumption only	0.191
Happiness only	0.662

Following our model selection, we fit multiple models with cubic regression spline on the predictors in hope for a better model. However, based on Table 2 and Appendix 2, we can see that there isn't any significant jump in performance between the models. Each of the resulting model have an Adjusted R^2 of 0.696. Hence, we will stick with our original linear model with all the predictors without any splines.

Table 2: Adjusted R^2 for Linear Models With/Without Cubic Regression Splines on the Predictors

Model	Adjusted R-Squared
No Splines	0.6964
Percentage of Smokers only	0.6960
Average Alcohol Consumption only	0.6960
Happiness only	0.6960

Conclusion

Summary

From the seven possible models in Appendix 1, based on previous results and arguments, we will pick the original model with both smoking, alcohol, and happiness as the predictor variables and life expectancy as the response variable. The Adjusted R^2 for the selected model is 0.6964. Furthermore, we decided to not include any splines since including them does not result in significantly better model and it will increase the chance of overfitting given that our dataset size is very small. It's also critical to note that the model contains several limitations that we'll discuss soon.

From each of the linear models, smokepercentage have a slight strong linear relationship with life expectancy while alcohol does not show a significant linear relationship with life expectancy. In addition, happiness is the factor that have the strongest linear relationship with life expectancy. All predictors combined have a stronger linear relationship with life expectancy.

When other variable remains constant, an increase of one unit in smokepercentage on average will result in increase of 2 million units of Life Expectancy^5 and increase of one litre of alcohol consumption on average will result in increase of 2.3 million units of Life Expectancy^5 and increase of one in happiness index will result in increase of 7.3 million units of Life Expectancy^5.

Coefficients	Estimate	P-Value	Significance
Intercept	-2.3 * 10^9	10^-10	***
smokepercentage	2 * 10^7	0.003	**
alcohol_consumption	2.3 * 10^7	0.16	*
happiness	7.3 * 10^8	10^-16	***

Table 3: Model Summary

Limitation and Strengths

We also need to remember that there are a lot of limitations and problems along our research and methods. First, our dataset is small as it only contains 101 observations and we decided to not validate our data since dividing them into training and testing dataset will further reduce the sample size and the accuracy of the model. We also use country's general population data

which averages values over everyone which causes over-generalization as same weights are given to each country, despite huge difference in population count.

Our data source also included 4 different data collected independently which causes variance and inaccuracy of time in our data. In our model checking, we merely check model assumptions by observing the data and transforming using a single method which might not be suitable in every scenario. Since our data are small, the transformation might overfit into our data which cause our validation to fail by significant margin. Our model selection is very limited because we only collect three different predictors to begin with which limits us to seven different non-empty subsets. Finally, our final model has moderate Adjusted R^2 and very low p-value which means that our predictors explain life expectancy pretty well.

We can improve our research by collecting a smaller cluster of data (smaller groups instead of countries) and gathering scientifically proven predictors that affect life expectancy significantly to allow better model selection and reduce variance of data.

Appendix

Appendix 1

Original Model of 2 Predictors vs Three Models of Each Predictor

```
##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ smokepercentage + alcohol_consumption +
##     happiness, data = combined_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.507e+09 -3.484e+08 -2.346e+07  2.995e+08  1.602e+09
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.311e+09  3.303e+08  -6.997 3.39e-10 ***
## smokepercentage      2.052e+07  6.906e+06   2.971  0.00375 **
## alcohol_consumption  2.300e+07  1.624e+07   1.417  0.15978
## happiness            7.316e+08  5.820e+07  12.571  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 574300000 on 97 degrees of freedom
## Multiple R-squared:  0.7055, Adjusted R-squared:  0.6964
## F-statistic: 77.46 on 3 and 97 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ smokepercentage + happiness,
##     data = combined_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.374e+09 -3.785e+08 -4.307e+07  3.082e+08  1.620e+09
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.390e+09  3.271e+08  -7.307 7.41e-11 ***
## smokepercentage  2.268e+07  6.770e+06   3.350  0.00115 **
## happiness        7.650e+08  5.348e+07  14.305  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 577200000 on 98 degrees of freedom
## Multiple R-squared:  0.6994, Adjusted R-squared:  0.6933
## F-statistic:   114 on 2 and 98 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ smokepercentage + alcohol_consumption,
##     data = combined_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.096e+09 -6.489e+08 -6.316e+05  7.425e+08  2.232e+09
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.223e+09  2.797e+08   4.373 3.06e-05 ***
## smokepercentage     2.064e+07  1.114e+07   1.853   0.0669 .
## alcohol_consumption 1.057e+08  2.395e+07   4.414 2.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 926400000 on 98 degrees of freedom
## Multiple R-squared:  0.2257, Adjusted R-squared:  0.2099
## F-statistic: 14.28 on 2 and 98 DF,  p-value: 3.605e-06

##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ alcohol_consumption +
##     happiness, data = combined_dataset)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -1.822e+09 -3.949e+08 -2.761e+07  3.513e+08  1.581e+09
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.922e+09  3.152e+08  -6.100 2.12e-08 ***
## alcohol_consumption  3.367e+07  1.646e+07   2.046   0.0434 *
## happiness            7.319e+08  6.048e+07  12.102  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 596700000 on 98 degrees of freedom
## Multiple R-squared:  0.6787, Adjusted R-squared:  0.6722
## F-statistic: 103.5 on 2 and 98 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ smokepercentage, data = combined_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.373e+09 -8.072e+08 -2.400e+07  8.021e+08  1.840e+09
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.673e+09  2.837e+08   5.898 5.13e-08 ***
## smokepercentage 3.257e+07  1.177e+07   2.766  0.00677 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.009e+09 on 99 degrees of freedom
## Multiple R-squared:  0.07173,    Adjusted R-squared:  0.06236
## F-statistic:  7.65 on 1 and 99 DF,  p-value: 0.006773

##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ alcohol_consumption,
##     data = combined_dataset)
```

```
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.404e+09 -6.720e+08  1.130e+08  6.525e+08  2.159e+09
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.615e+09  1.852e+08   8.722 6.67e-14 ***
## alcohol_consumption 1.165e+08  2.352e+07   4.952 3.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 937700000 on 99 degrees of freedom
## Multiple R-squared:  0.1986, Adjusted R-squared:  0.1905
## F-statistic: 24.53 on 1 and 99 DF,  p-value: 3.019e-06

##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ happiness, data = combined_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.669e+09 -3.523e+08 -4.681e+07  3.674e+08  1.606e+09
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.982e+09  3.188e+08  -6.215 1.22e-08 ***
## happiness    7.833e+08  5.588e+07  14.018  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 606300000 on 99 degrees of freedom
## Multiple R-squared:  0.665,  Adjusted R-squared:  0.6616
## F-statistic: 196.5 on 1 and 99 DF,  p-value: < 2.2e-16
```

Appendix 2

Final Model of 3 Predictors With/Without cubic Regression Spline on Each Predictor

```
##
## Call:
## lm(formula = life_expectancy_at_birth^5 ~ smokepercentage + alcohol_consumption +
##     happiness, data = combined_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.507e+09 -3.484e+08 -2.346e+07  2.995e+08  1.602e+09
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.311e+09  3.303e+08  -6.997 3.39e-10 ***
## smokepercentage      2.052e+07  6.906e+06   2.971  0.00375 **
## alcohol_consumption  2.300e+07  1.624e+07   1.417  0.15978
## happiness            7.316e+08  5.820e+07  12.571  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 574300000 on 97 degrees of freedom
## Multiple R-squared:  0.7055, Adjusted R-squared:  0.6964
## F-statistic: 77.46 on 3 and 97 DF,  p-value: < 2.2e-16
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## life_expectancy_at_birth^5 ~ s(smokepercentage, bs = "cr") +
##     alcohol_consumption + happiness
##
## Parametric coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.848e+09  3.043e+08  -6.074 2.43e-08 ***
```

```
## alcohol_consumption  2.300e+07  1.624e+07   1.417     0.16
## happiness            7.316e+08  5.820e+07  12.571  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F p-value
## s(smokepercentage)   1      1 8.825 0.00375 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.696   Deviance explained = 70.6%
## GCV = 3.4337e+17  Scale est. = 3.2977e+17  n = 101

##
## Family: gaussian
## Link function: identity
##
## Formula:
## life_expectancy_at_birth^5 ~ smokepercentage + s(alcohol_consumption,
##     bs = "cr") + happiness
##
## Parametric coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.154e+09  3.656e+08  -5.893  5.5e-08 ***
## smokepercentage  2.052e+07  6.906e+06   2.971  0.00375 **
## happiness        7.316e+08  5.820e+07  12.571  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                         edf Ref.df     F p-value
## s(alcohol_consumption)   1      1 2.007    0.16
##
## R-sq.(adj) =  0.696   Deviance explained = 70.6%
## GCV = 3.4337e+17  Scale est. = 3.2977e+17  n = 101

##
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## life_expectancy_at_birth^5 ~ smokepercentage + alcohol_consumption +
##     s(happiness, bs = "cr")
##
## Parametric coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.788e+09  1.791e+08   9.985  < 2e-16 ***
## smokepercentage     2.052e+07  6.906e+06   2.971  0.00375 **
## alcohol_consumption 2.300e+07  1.624e+07   1.417  0.15978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df   F p-value
## s(happiness)   1      1 158  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.696   Deviance explained = 70.6%
## GCV = 3.4337e+17  Scale est. = 3.2977e+17  n = 101
```

References

1. Manuel, D. G. (2012). Seven more years the impact of smoking, alcohol, diet, physical activity and stress on health and life expectancy in Ontario: an ICES/PHO report. Institute for Clinical Evaluative Sciences.
2. Janssen, F., Trias-Llimós, S., & Kunst, A. E. (2021). The combined impact of smoking, obesity, and alcohol on life-expectancy trends in Europe. International Journal of Epidemiology, 50(3), 931–941. https://doi.org/10.1093/ije/dyaa273
3. Östergren, O., Martikainen, P., Tarkiainen, L., Elstad, J. I., & Brønnum-Hansen, H. (2019). Contribution of smoking and alcohol consumption to income differences in life expectancy: evidence using Danish, Finnish, Norwegian and Swedish register data. Journal of Epidemiology and Community Health (1979), 73(4), 334–339. https://doi.org/10.1136/jech2018-211640
4. Kunzmann AT, Coleman HG, Huang WY, Berndt SI (2018) The association of lifetime alcohol use with mortality and cancer risk in older adults: A cohort study. PLOS Medicine 15(6): e1002585. https://doi.org/10.1371/journal.pmed.1002585
5. Lawrence, E. M., Rogers, R. G., & Wadsworth, T. (2015). Happiness and longevity in the United States. Social science & medicine (1982), 145, 115–119. https://doi.org/10.1016/j.socscimed.2015.09.020
6. Smoking Rates by Country 2021: https://worldpopulationreview.com/country-rankings/smoking-rates-by-country
7. Alcohol consumption per capita per year (litres of pure alcohol): https://data.worldbank.org/indicator/SH.ALC.PCAP.LI
8. Life Expectancy at Birth (in years): https://data.worldbank.org/indicator/SP.DYN.LE00.IN
9. World Happiness Report 2019: https://www.kaggle.com/datasets/unsdsn/world-happiness
10. Country to Continent Dataset: https://raw.githubusercontent.com/dbouquin/IS_608/master/NanosatDB_munging/Countries-Continents.csv