Bill Huang

April 29th, 2022

Stat 410 Project

Stat 410 Final Project

Life expectancy is defined as the average period that a person is expected to live. As humans, we all want to live longer, so we all want to have a higher life expectancy. Although there have been reports about certain people living a longer life with unhealthy lifestyles like excessive smoking and drinking, I believe the life expectancy around the world does follow a general pattern based on different factors in different countries. Recently I found a dataset[1] that contains data on life expectancy, adult mortality, infant mortality, alcohol consumption, percent expenditure on health, GDP, etc.(22 variables in total) on 183 countries from Year 2000 to 2015. The purpose of this research project is to first investigate the relationship between life expectancy and alcohol consumption (SLR), and then examine the relationship between life expectancy and some of the other factors that may potentially impact life expectancy (MLR).

SLR

I first decided to run a simple linear regression on life expectancy as the response variable and alcohol consumption (per capita consumption in liters of pure alcohol) as the covariate for all the data points within the dataset. It turns out that there is a positive relationship between life expectancy and alcohol consumption, with a coefficient of around 0.95 for the covariate.

[1] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

I then decided to run the same thing, but separating the data into different years, so I would have the data on each country's life expectancy and alcohol consumption in each year (So there would be 15[2] subdatasets). After grouping the data and running a linear regression on the variables, it turns out that there has been a consistent pattern: there is a positive correlation between life expectancy and alcohol consumption (the coefficient is significant for all the linear models) each year, and the coefficient for the covariate stays roughly consistent, oscillating around 1 throughout the years. Some of the plots are shown below: the 4 plots (Figure 1) included are the SLR models that used life expectancy as the response variable and alcohol consumption as the covariate in 2000, 2001, 2012, and 2013 respectively. Here we can see almost an identical pattern and slope for the fitted lines, which means that the relationship between life expectancy and alcohol consumption had stayed consistent throughout the years. We also see that this relationship is positive, which means that an increase in alcohol consumption is expected to increase one's life expectancy. This result is somewhat surprising, so I decided to split the data in groups to see if there is a different trend between the variables.

[2] There is not enough data on alcohol consumption for Year 2015, so we have to exclude Year 2015, which makes it 15 subdatasets, from Year 2000 to Year 2014.

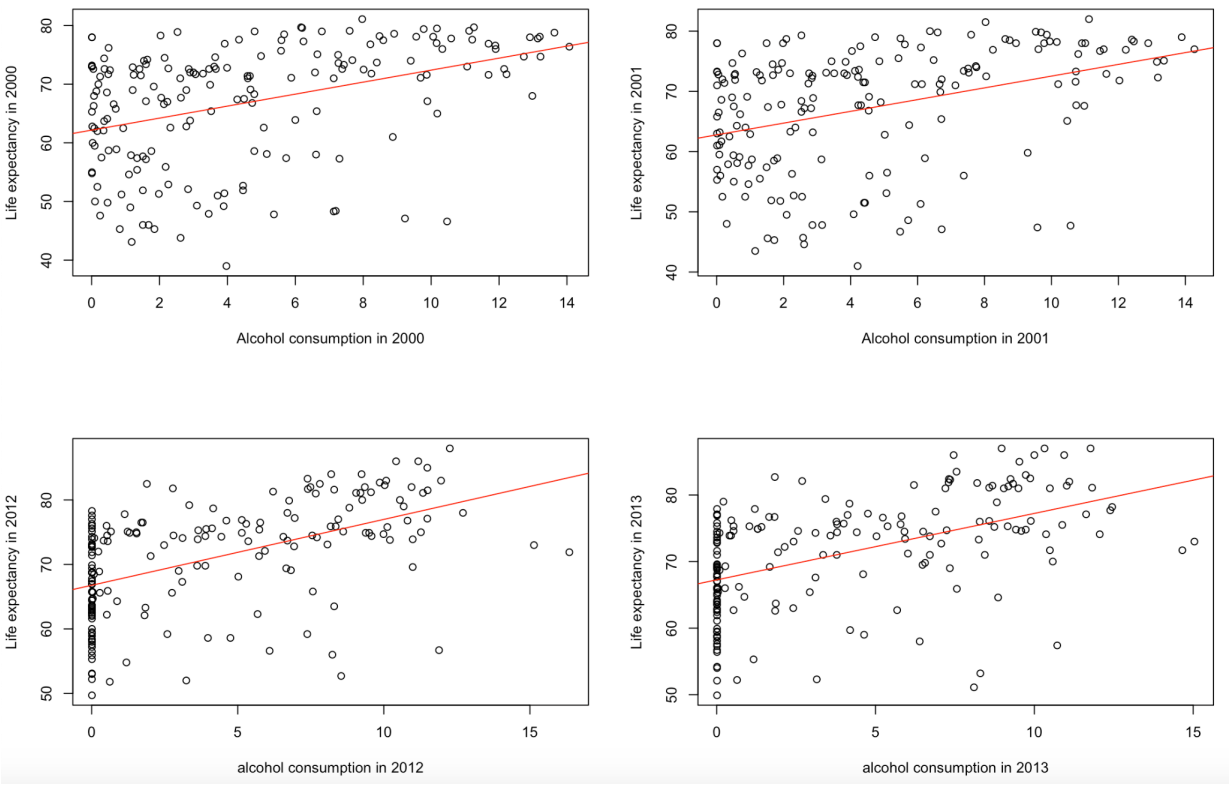

Figure 1

I decided to separate the dataset into two groups: one with all the developing countries, and the other one with all the developed countries. After running a SLR on the two datasets, it appears that there still exists a positive association between life expectancy and alcohol consumption for developing countries. However, this relationship becomes negative when it comes to developed countries. From the dataset, there appears to be 32 developed countries and 151 developing countries. This makes me think about the possibility that the positive correlation between life expectancy and alcohol consumption may be largely influenced by the developing countries —- that the positive relationship is largely offset by the data from developing countries. After running regressions, the trend does turn out to be true: developed countries tend to have a negative relationship between life expectancy and alcohol consumption, while

developing countries tend to have a positive relationship between life expectancy and alcohol consumption throughout the years.

I also ran the “gam” function between life expectancy and alcohol consumption, and it turns out that the plot displays a nonlinear relationship (Figure 2), which means that SLR may not be the best model. Although this is the case, there is still a consistent trend throughout the years based on the linear models, which means that we can make a general conclusion on the relationship between life expectancy and alcohol consumption.

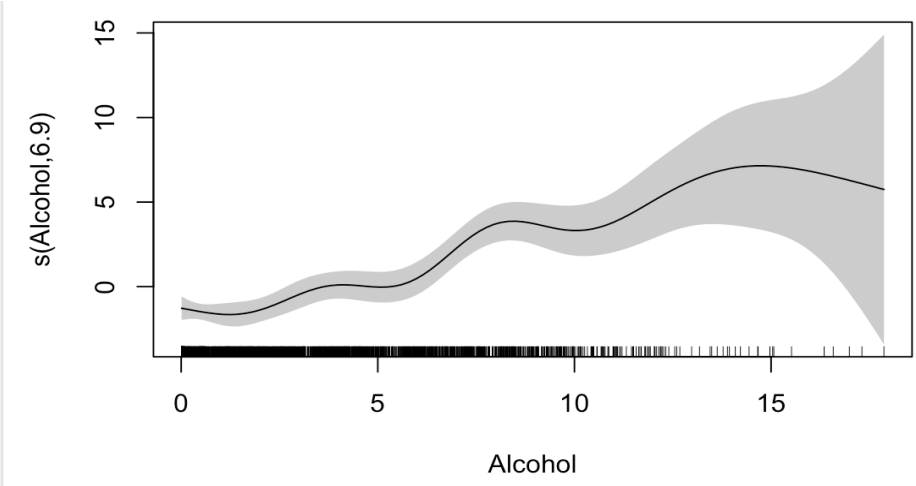

Figure 2

MLR

Since there are a huge amount of variables within this dataset, it is important to select the relevant variables as the covariates in order to have an appropriate multiple linear regression. Before conducting the variable selection, I decided to add a dummy variable for a country’s status (call it “Status2”): 1 for developed countries and 0 for developing countries. Then I decided to run an AIC for all the variables. After conducting

both the forward method and the backward method of AIC, it turns out that both models include the same variables, which helps me determine which variables would potentially have an impact on life expectancy. The variables that are included after the AIC are: “Schooling”, “HIV.AIDS”, “Adult.Mortality”, “Income composition of resources”, “percentage expenditure”, “BMI”, “Diphtheria”, “under five deaths”, “infant deaths”, “thinness for 5 to 9 years”, “Total expenditure”, “Alcohol”, and “Status2”. On the other hand, the variables that are not included are “Hepatitis.B", "Measles", "Polio", "GDP", "Population”, and "thinness 10 to 19 years."

After the variable selection process, I decided to run an MLR on the 13 variables selected, using life expectancy as the response variable. It turns out that the coefficients for all but one variables are statistically significant (the one that isn’t significant has a p-value of around 0.065, which is significant at 0.1 level). This means that a linear model is potentially an appropriate model. I also did an ANOVA test on the MLR model, the F-statistic turns out to be large enough for us to reject the null hypothesis. This also shows that all the coefficients for the variables in our MLR model are statistically significant.

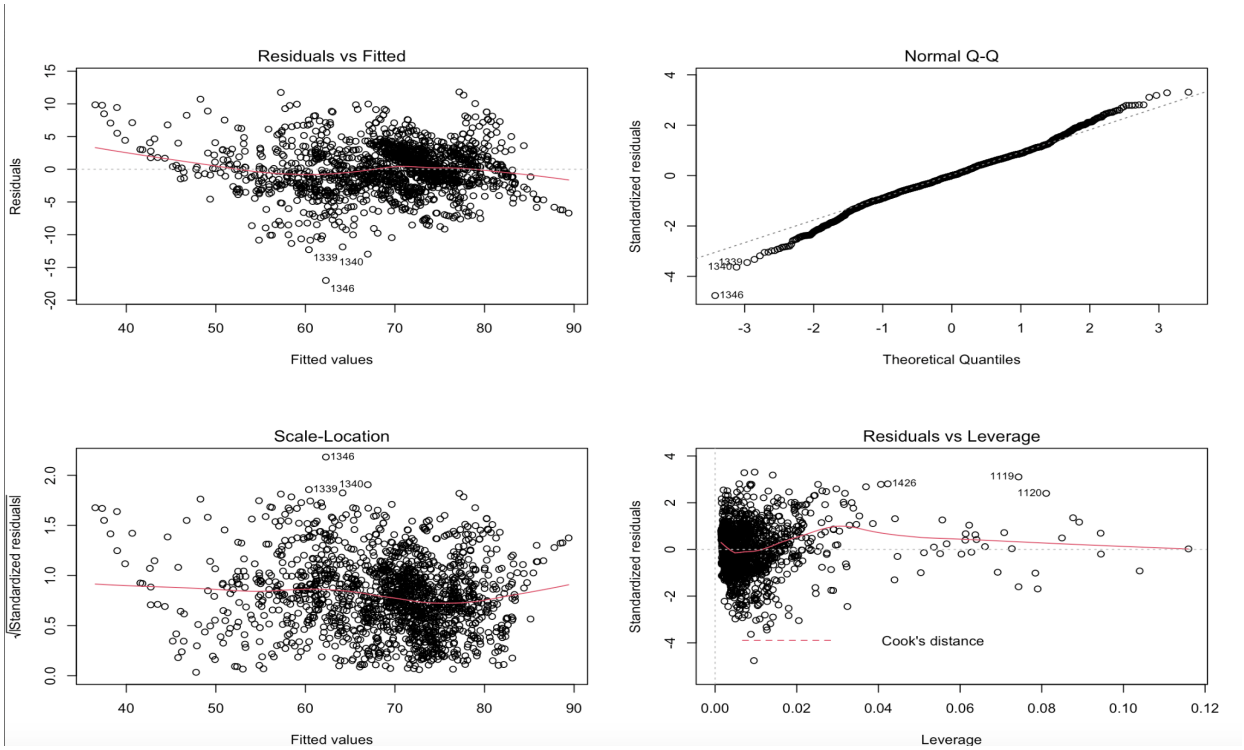

Figure 3

Looking at the 4 diagnostic plots (Figure 3): from the residual plot, I don't see a particular pattern which means that the linear model is appropriate; from the Normal QQ plot, although there is some deviation at the ends of the plot, the points do follow a linear pattern, so it seems appropriate for us to have the normal Gaussian assumption; from the scale-location plot, the red line is roughly horizontal in the plot, we also see that the residuals are pretty randomly scattered around the graph, which means that the distance from the residuals to each fitted values are roughly equal; last but not least, from the residuals vs leverage plot, there appears to be a few leverage points and outliers, but majority of the data points are within the Cook's distance lines. From the diagnostics plot, it appears that the linear model (MLR) is appropriate for our selected variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.349e+01	7.088e-01	75.462	< 2e-16
X(Intercept)	NA	NA	NA	NA
XAdult.Mortality	-1.672e-02	9.465e-04	-17.669	< 2e-16
Xinfant.deaths	9.186e-02	9.953e-03	9.229	< 2e-16
XAlcohol	-8.747e-02	3.295e-02	-2.654	0.00802
Xpercentage.expenditure	4.259e-04	5.918e-05	7.196	9.43e-13
XBMI	3.368e-02	5.949e-03	5.662	1.76e-08
Xunder.five.deaths	-6.951e-02	7.394e-03	-9.402	< 2e-16
XTotal.expenditure	7.488e-02	4.055e-02	1.847	0.06499
XDiphtheria	1.488e-02	4.522e-03	3.291	0.00102
XHIV.AIDS	-4.359e-01	1.781e-02	-24.479	< 2e-16
Xthinness.5.9.years	-5.695e-02	2.639e-02	-2.158	0.03108
XIncome.composition.of.resources	9.871e+00	8.288e-01	11.910	< 2e-16
XSchooling	8.743e-01	5.869e-02	14.897	< 2e-16
XStatus2	9.461e-01	3.359e-01	2.817	0.00491

Figure 4

Figure 4 shows the values of the coefficients for the MLR: here the alcohol consumption actually has a negative coefficient, which is different from our SLR model. Some other notable things to look at is the magnitude of each value: it appears that a one-unit difference in “income composition of resources” would have the largest expected change in average life expectancy; and a one-unit difference in “percentage expenditure” would result in the smallest expected change in average life expectancy. The sign of the coefficients for most variables are consistent with their relationship with life expectancy based on our own understanding.

Another thing worth looking at is how well the response variable can be predicted given a subdataset. I decided to split the dataset[3] into training data (70%) and testing data (30%). I first ran an MLR on the training data. After that, I used the “predict” function to predict the response variable using the MLR from the training data and the “newdata” using the testing data with a level of 90%. This would tell me how well we can

[3] The dataset here only contains the variables selected after AIC (so the 13 variables mentioned above)

predict the response variable based on the training data, using testing data as a standard for the prediction. Aftering getting the prediction intervals, I looked at the maximum value and mean value of the absolute value of the difference between the predicted value and the actual values from the testing data for life expectancy, which corresponds to 17.7 and 5.00 respectively. The MSE for the difference between the predicted value and the actual value is about 36.4. I also plotted the histogram for the absolute difference between the predicted value and the actual values from the testing data which looks like this:

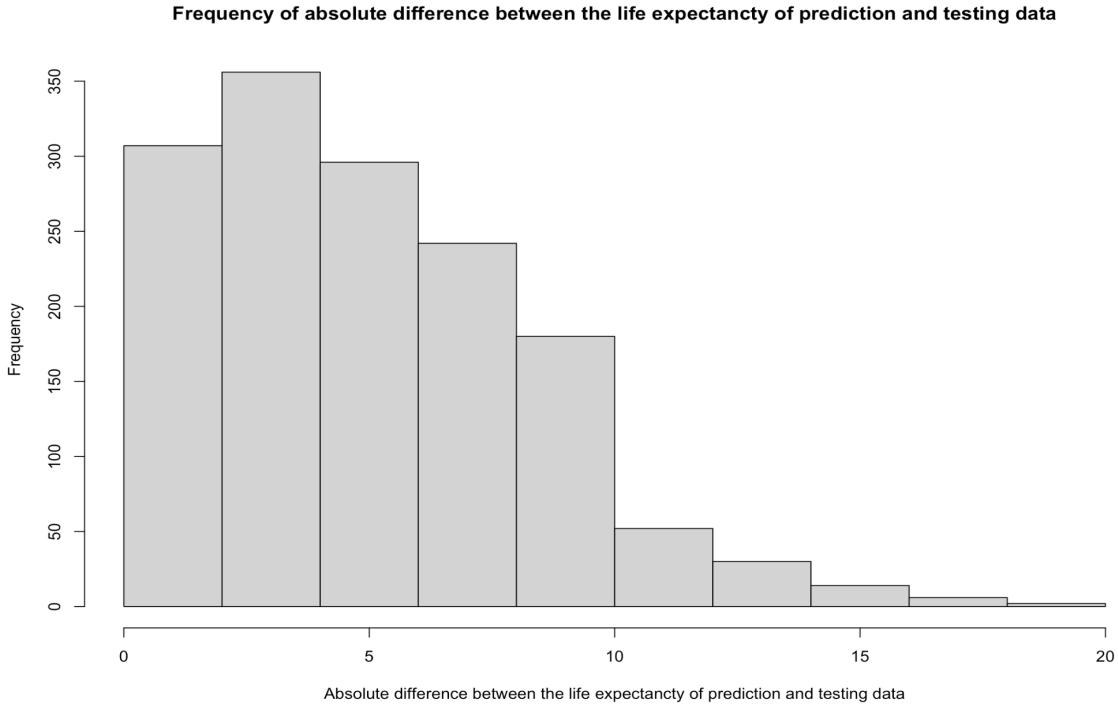

Figure 5

The histogram (Figure 5) shows how the different absolute differences are distributed. It looks like most of the values of the differences lie between 0 and 5, which means that our prediction is roughly accurate. This means that if we are given a portion

of the data from the original dataset, then we would be able to predict life expectancy pretty effectively using the given data.

I decided to run an additive model using the gam function. After plotting each function corresponding to each variable, it does appear most of the variables have a linear function with respect to the response variable. I did apply the smooth function in "gam" on two variables, "Total expenditures" and "thinness for 5 to 9 years", and both appear to have a nonlinear relationship from the plot. This means that depending on the different values of the covariates ("Total expenditures" and "thinness for 5 to 9 years"), the relationship between the response variable and the covariates would vary.

Conclusion

The purpose of this research is to use SLR to examine the relationship between life expectancy and alcohol consumption, and to use MLR to examine the relationship between life expectancy and other variables. After looking at all the regressions and diagnostics, there does seem to be a clearer picture of how life expectancy is related to all the variables within the dataset.

From the SLR models, it appears that there has been a consistent trend of a positive relationship between life expectancy and alcohol consumption for all countries. On the other hand, this positive relationship is still present in developing countries, while the relationship becomes negative for developed countries. This shows the overall positive relationship is likely to be offset by the trend in developing countries.

I then ran some MLR models. There are 22 variables within the original dataset, so I decided to run an AIC to have a variable selection process which ended up giving

me 13 of them. I then ran an MLR on the selected variables using life expectancy as the response variable. The diagnostic plots show that the linear model seems appropriate, despite a few leverage points and outliers. We looked at the coefficients for each variable. We also split the original dataset into training data and testing data, so that we could predict the data using the training data. After comparing the predicted data and the testing data, it turns out that the difference between the two were mostly distributed between 0 and 5, which means our prediction is rather accurate. We also ran an additive model in the end, which shows that there does appear to be nonlinear relationships between life expectancy and a few covariates.

There certainly exist some flaws within this research, but we do get a better understanding of how life expectancy is related to some of the factors. A more comprehensive study in the future will give us an even clearer picture on this topic.