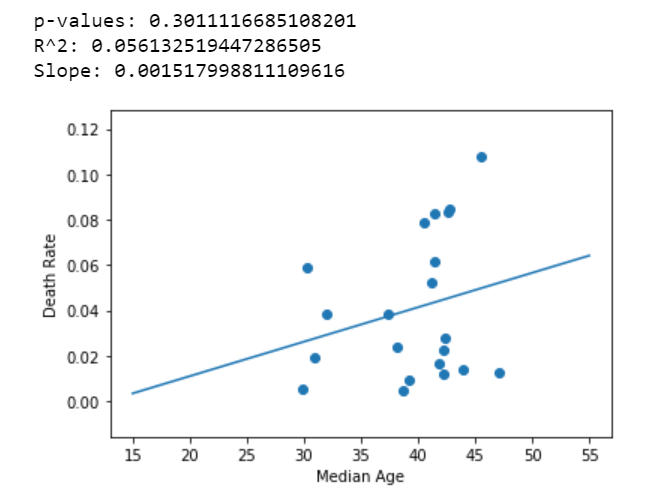
MATH 189Z Homework 1

Jeremy Tsai

**Task 1:**



I first filtered the data to only include countries that had a minimum of 5000 cases in order to reduce biases regarding inflated death rates due to small sample sizes. After cleaning the data and running a regression, I received the above graph. As shown, the slope is now positive, showing a general increase in death rate with an increase in median age. Notice that the R^2 value is still relatively low at 0.056, but it is still 10 times greater than the R^2 value of 0.0053 for the data that was not cleaned for a minimum sample size. The p-value is also 0.30, which is greater than 0.05, so this is statistically significant. It seems that there exists a positive correlation between median age of a country and the country’s death rate.

This would make sense since a greater median age means that there are older people residing in that country versus others. Since the immune system weakens over the span of human life, the Coronavirus will be more deadly to these people. Thus, countries with a greater median age will experience a greater death rate. In countries with younger people and a lower median age, the individuals’ immune systems are on average stronger and they will be less susceptible to contracting the Coronavirus and perishing from it, leading to a lower death rate.

**Task 2:**

Research question: I wanted to research whether there was a correlation between economic prosperity/state of development and death rate by country.

Data sources:

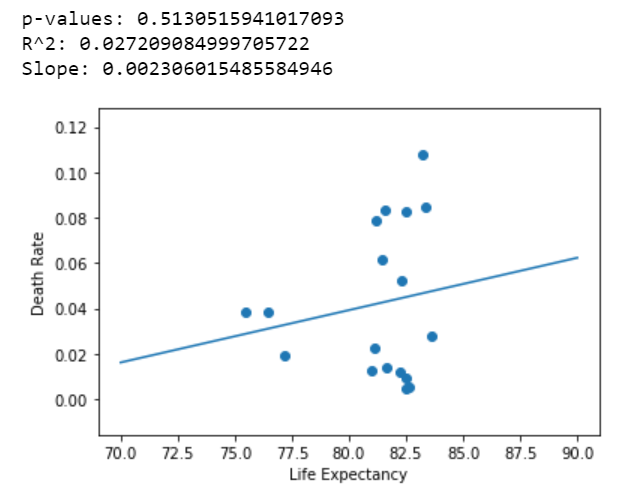
<https://data.worldbank.org/indicator/NY.GDP.MKTP.CD> (GDP by country). Provides csv file with the GDP of countries around the world by year. For the regression, I used the 2017 GDP since most countries had that as a non-null entry.

<https://data.worldbank.org/indicator/SP.DYN.LE00.IN> (Life expectancy by country). Provides csv file with the life expectancy of countries around the world. For the regression, I used the 2017 life expectancy at birth in years.

First regression: Testing for correlation between life expectancy and death rate within a country.

Prediction: My prediction is that there will be a negative correlation between life expectancy and death rate within a country. This is because, if the life expectancy of a country is higher, than it usually signifies they have more advanced medical infrastructure and stability. Thus, they should be better equipped to treat patients with the Coronavirus and the death rate due to Coronavirus should be lower.

Methods/Results:



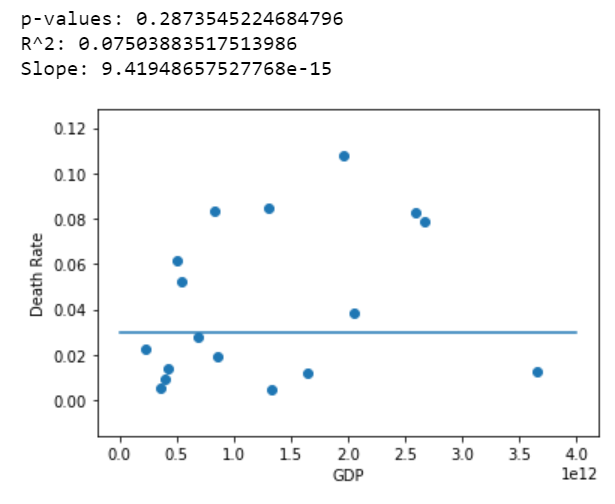
For this regression, I first opened the life expectancy csv file and set the country name as the index of the Data Frame. I then took the column with the life expectancy at birth values for each country and only kept the rows where the country was in the list of countries that had greater than 5000 confirmed cases. I set the minimum to be 5000 confirmed cases to reduce biases regarding small sample size. I then merged the two data frames of confirmed cases for each country and life expectancy for each country. Lastly, I did the linear regression and plotted the results as shown above.

The slope is positive, showing a positive correlation between life expectancy and death rate. There is an R^2 value of 0.27 and the p-value is .51, meaning this is statistically significant. This conflicts with my initial hypothesis of there being a negative correlation. However, upon further thought, it could make sense that there is a negative correlation. If a country has a higher life expectancy, it means they are more developed. This typically also means the country is more urban and populations are more bunched together. Additionally, there is also greater access of transportation. All of these are perfect for the spread of a virus. Thus, more-developed countries can be more susceptible to rapid spread of the Coronavirus and thus have a higher death rate. To be more sure of this claim, I would have to perform a linear regression on other variables that indicate the development of a country and see if there is still a negative correlation.

Second Regression:

Prediction: My prediction is that there will be a negative correlation between GDP and death rate within a country. Similar to the previous prediction, if the GDP of a country is higher, this typically means they are more developed and have a better medical infrastructure to deal with sick patients. Therefore, the country should be more effective at dealing with the Coronavirus and have a lower death rate.

Methods/Results:



For this regression, I first opened the GDP csv file and set the country name as the index of the Data Frame. I then took the column with the 2017 GDP values for each country and only kept the rows where the country was in the list of countries that had greater than 5000 confirmed cases. I set the minimum to be 5000 confirmed cases to reduce biases regarding small sample size. I then merged the two data frames of confirmed cases for each country and GDP for each country. Lastly, I did the linear regression and plotted the results as shown above.

The slope is near 0, showing very little correlation between GDP and death rate. There is an R^2 value of 0.075 and the p-value is .29, meaning this is statistically significant. This conflicts with my initial hypothesis of there being a negative correlation. This also goes against my aforementioned theory that more-developed countries are more susceptible to the spread of a virus and thus will have a higher death rate. Perhaps there are other variables besides GDP that better support my theory regarding well-developed, urban countries, such as population density and accessibility to busses/trains. To be more sure of this claim, I would have to perform a linear regression on these other variables see if there are negative correlations for them.

**Task 3:**

I wanted to take this course because I have always been interested in data analytics and the methods that people use in order to extract important information from data. After all, in this age where data is so prevalently shared and with our developments in technology, there are lots of insights to be discovered that are buried within the masses of data we have stored. I also believe that applying these data analytics methods to real-world examples such as the Coronavirus will be impactful and serve to increase my satisfaction in learning. I hope to learn the basic methods that are prevalently used in the data science industry and be able to use them to uncover information regarding the Coronavirus.

It took me 6 hours to do this assignment.