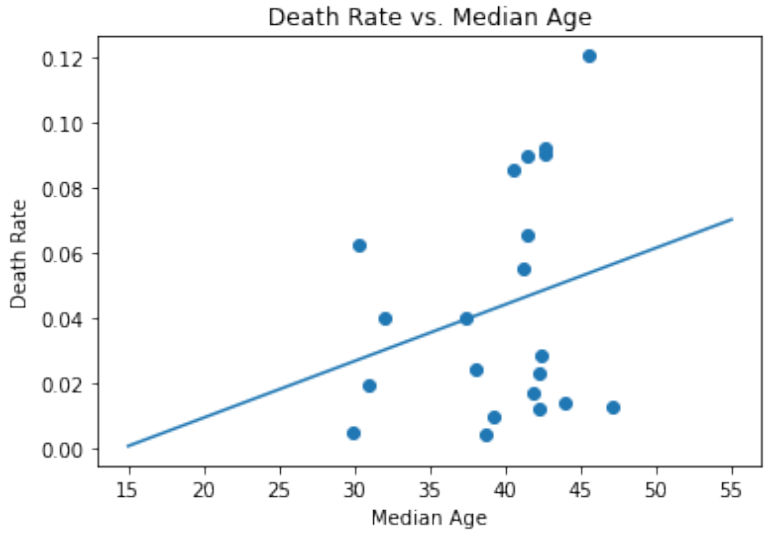
Eve Kazarian
April 10th, 2020

Math 189Z

Homework 1: Analyzing COVID-19 Data with Regression

Task 1

Graph of new Linear Regression:



Important statistics:
p-values: 0.286
R^2: 0.0597
Slope: 0.00174

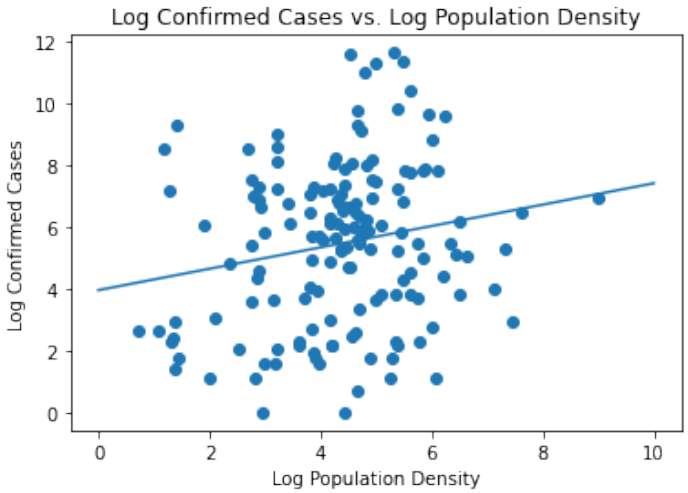
Results, Interpretation, and Discussion:
I decided to filter the data such that only countries with over 5000 confirmed cases were considered. This decision narrowed the dataset from 181 to 21 countries. Plotting death rate versus median age for these countries, I noticed the slope was positive (0.00174), indicating the possibility that death rate increases with increasing median age. Additionally, the R^2 value indicates that median age explained 5.97% of variation in the death rate, a small percentage. Since the p-value is greater than a significance level of 0.05, we cannot conclude our results are significant; death rate does not necessarily increase with increasing median age across these countries.

Task 2

Data source and description of data:
My data came from The World Bank Group, which has many datasets concerning COVID-19. The data set I chose was formatted as two columns of "Country Name" and "Population Density," with population density ranging from 0 to 21,000.

Research question #1:
How does the number confirmed cases change with increasing population density?

Methods and results with accompanying graphs:



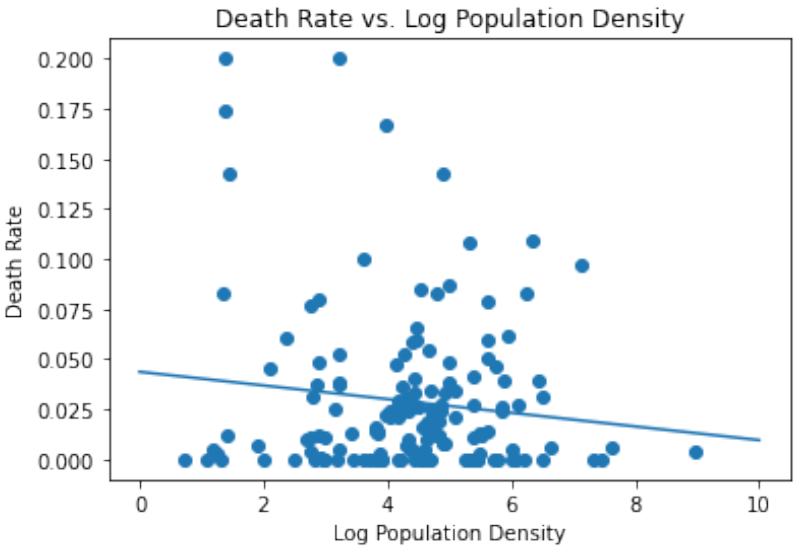
Important statistics:
p-values: 0.0191
R^2: 0.0358
Slope: 0.346

I used our provided data set of confirmed cases per country and merged it with my chosen population density data set, effectively selecting the confirmed cases and population densities of all countries common to both data sets (157 countries). Then, I performed linear regression on confirmed cases versus population density. The range of confirmed cases and population densities was very broad, and the data seemed it might fit a power or exponential curve. As such, I log transformed both population density and number of confirmed cases.

I expected the number of confirmed cases to increase with increasing population density, since COVID-19 spreads more easily in populated areas. My results confirm this relationship because the slope of the regression of log confirmed cases on log of population density is positive. This result is significant because the p-value is less than a significance level of 0.05. However, the R^2 value of 0.0358 indicates that the linear regression is a weak fit for the data.

Research question #2:
How does the death rate change with increasing population density?

Methods and results with accompanying graphs:



Important statistics:
p-values: 0.1246
R^2: 0.01555
Slope: -0.003405

Again, I used our previously generated data set of death rate per country and merged it with my chosen population density data set, effectively selecting the death rates and population densities of all countries common to both data sets. Then, I performed linear regression on death rates versus population density. The range of population densities was very broad, and the data seemed it might fit a power or exponential curve. As such, I log transformed population density.

I expected there to be no significant relationship between death rate and population density, since those with COVID-19 do not necessarily die from it. My results confirm this expectation because, although there is a negative slope, the p-value is less than a significance level of 0.05.

Task 3

I wanted to take this course because I was hoping to learn which machine learning techniques to apply to which problems. Also, I wanted to gain the skills of working with libraries such as pandas and numpy.

This assignment took me 4 hours.