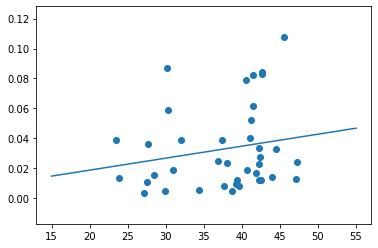
Lauren West

Math 189Z

10 April 2020

**Homework 1: Analyzing COVID 19 Data with Regression**

Graph of Linear Regression:



p-values: 0.2511999715361453

R^2: 0.03542419131914649

Slope: 0.0007994940475298446

Results:

We see a trend of increasing median age and increasing death rate, when performing a linear regression. We had a p-value of 0.2511999715361453, and R^2 of 0.03542419131914649, and a positive Slope of 0.0007994940475298446.

Interpretation:

Our results, when excluding small sample sizes, still has a high p-value. So, even though we see a trend of death rate increasing with median age, we cannot pull from this with statistical significance those results.

Discussion:

My method of excluding small sample size was to exclude countries with <1000 confirmed cases before the date of 28 March 2020. I did this because small same sizes seem to be what was causing the ‘garbage in garbage out’ on the previous graph from the same data source provided in the homework.

Additionally, our R^2 is far from 1, a perfect fit for the data. Having an R^2 = 0.03542419131914649, we see that our line is a poor fit for the data and not a good model, as our data is not very linear.

**Task 2:**

I was researching:

Does state population affect the total number of coronavirus in a state?

The Data Source used was from covidtracking.com/api, using the States Current Values found at [/api/v1/states/current.json](https://covidtracking.com/api/v1/states/current.json) | [CSV](https://covidtracking.com/api/v1/states/current.csv).

Methods:

In my code, I dropped the unnecessary columns, I incorporated states populations (wiki) into a dictionary I created, then appended those populations to a population list and included that list in a column in my dataframe.

Statistical function used to plot a linear regression of the dataframe, with population in millions on the x-axis and total cases on the y-axis.

Results:

Its resulting p-values was 0.00310911689560331, the R^2 was 0.1649103558857895, and its Slope was 0.0011841271654768673.

The Graph, Total Cases vs Population, shown below has population in millions on the x-axis and total cases on the y-axis.

A screenshot of a cell phone

Description automatically generated

Interpretation + Discussion of Results:

With resulting p-values was 0.00310911689560331, it seems that the results are statistically significant, with p< 0.05. Since R^2 was 0.1649103558857895, the relationship is not very linear. The Slope was 0.0011841271654768673, with shows that as the state size increased, the total number of cases also increased. Additionally, another factor implicating the credibility of my results is the source from where I got my state population. In hindsight, Wikipedia may not be the most up to date and reliable place to find current state populations. Then again, these populations are in the millions, and small variations would likely not impact our results very much at all.

Since our R^2 is not near 1, we cannot say that a linear regression models the data well. However, with a p< 0.05, we could take away that as the state size increased, the total number of cases also increased.

**Task 3:**

s

I have zero experience, this class may be a bit more difficult for me personally, and my coding skills are not strong. But that’s kind of why I’m taking this class. I’m interested in learning how to use GitHub and I want to get experience with data analysis and linear regression. I think these tools are super powerful, and knowing how to use them can help me in the future in doing things that are important to me.

It took me 6 hours.