Group069_fa20

December 16, 2020

1 Group069_FA20

1.1 Permissions

We, as a group, is giving permission to make our project to be made available to the public. (Note that we do not give permission to have our PID made public).

- () Yes - Make Available
- (X) No - Keep Private

1.2 Names & PID

Names	PID	Year
William Hwong	A14436409	Computer Science
Yu Chieh Lin	A15051297	Computer Science
Juan Roa	A15541981	Cognitive Science Spec. ML
Hong-An Vu	A15472509	Cognitive Science Spec. ML

1.3 Overview

In this project, our team will be performing linear regression to find the relationship between the number of covid cases a county in California has against certain characteristics. The team has sourced data from several government agencies such as the U.S. Census Bureau and the U.S. open data website and merged it into one data frame for ease of analysis. Several different visualization techniques such as choropleths and charts will be included for ease of understanding.

1.4 Research Question

The question this project is exploring is does where you live in California affect your chances of contracting COVID-19? More specifically, we are assessing how factors in an individual's environment and location correlate with the number of overall cases in that individual's county. Some specific factors we are interested in observing are covid-19 related citations, population size, poverty rate, and air quality in the area.

1.5 Background and Prior Work

SARS-CoV-2 or more commonly known as Covid-19 has spawned a global pandemic sending the world's economies into shutdown as governments around the world attempt to slow the progression of the highly contagious and infectious disease. While researchers rush to understand the mechanisms of the disease, what is already known is that the disease has disproportionately ethnic minority groups such as Latinos, African Americans, and Native Americans within the United States and they experience a higher death rate compared to other ethnicities (1). There are societal and biological determinants that contribute to the disparity experienced by these high risk groups such as a lower access to healthcare which can exacerbate the symptoms of the disease due to lack of treatment as well as having more comorbidities such as heart disease, obesity, and many others. Furthermore, many of these same individuals are deemed "essential workers" such as cashiers, cooks, child care takers, and healthcare workers and up to 61% of these individuals are at an increased risk of contracting severe Covid-19 (2). It is clear that socio-economic factors are a strong factor in who contracts the disease and how severely it will affect the patient.

While the socio-economic factors of the pandemic are being thoroughly investigated, there are plenty of biological factors that also factor into the hospitalization rates and death rates of Covid-19 patients with age being one of the most prominent factors. According to the CDC the eldery aged 75-84 years old have a hospitalization rate of almost 8 times higher than the rate for an 18-29 year old and a death rate that is 220 times higher (3). Common sense would suggest that this is to be expected as the eldery are generally more susceptible to infectious diseases but there are also other compounding factors that contribute to such a disappropriate effect. The eldery are also typically found in nursing homes, hospice care facilities, and other long term care facilities which at the most recent estimate, account for 42% of all covid-19 related fatalities (4). The problem with such environments is that these facilities are enclosed spaces where a high risk group of individuals congregate for extended periods of time which goes against all guidelines provided by the CDC to slow the spread of the disease (3).

It is clear that there are a wide variety of socio-economic and political factors that lead to this pandemic affecting more than others but it is our desire in this report and study to gain a little more information on the virus and exactly how certain factors contribute to an individual's chance of contracting the disease. It is unlikely that the world will fully understand the long term effects and consequences of such a novel disease any time soon but it

1.5.1 References

1. Tai, Don, et al. "The Disproportionate Impact of COVID-19 on Racial and Ethnic Minorities in the United States." Oxford Academic, Clinical Infectious Diseases, 20 June 2020
2. Reinberg, Steve. "74 Million U.S. Workers at High Risk for COVID." WebMD, WebMD, 9 Nov. 2020
3. "COVID-19 Hospitalization and Death by Age." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 18 Aug. 2020,
4. Girvan, Gregg. "42% Of COVID-19 Deaths in Nursing Homes & Assisted Living Facilities." Medium, FREOPP.org, 17 Sept. 2020.

1.6 Hypothesis

We hypothesize that there will be a correlation with respect to one's location and their chance of contracting covid-19 as the environmental factors such as a higher number of COVID-related citations, population size, poverty rates, and air quality all contribute to the spread of the disease. We believe this as our hypothesis because rule violations mean that people in the county are not abiding to regulations aimed at preventing the spread of COVID-19, and therefore people in these communities would be more susceptible to contracting it. Counties with a higher population size usually represent larger cities, and cities tend to have more covid cases because it is more difficult to socially distance from others. We also believe that poverty rates of a county will impact the number of covid-19 cases as this indicates either a lack of resources available to combat covid-19 or that these individuals will hold jobs such as service workers who will be in higher risk groups. The air quality of a county may be related to the overall number of covid-19 cases as poor air quality may lead to inflammation of lungs and impact those with high susceptibility to respiratory infections including covid-19.

1.7 Dataset(s)

df_population has the following columns (due to excessive amount of columns we are showing the ones we are using):

Header	Meaning
CTYNAME	Name of city
POPESTIMATE2019	Population estimate

df_poverty has the following columns (due to excessive amount of columns we are showing the ones we are using):

Header	Meaning
Unnamed: 3	County Name
Unnamed: 7	Pverty Percentage

df_air has the following columns:

Header	Meaning
FIPSCode	Codes assigned for county
LocationType	Type of location
Location	County Name
TimeFrame	Year
DataFormat	Type of Date
Data	Count of Air Violation

df_citations has the following columns:

Header	Meaning
DATE	Full Date
DBA	DBA Name
Street Address	Street Address
City	City Name
Violation/Cited	Type of Citations
License Type	Type of License
License #	License Number
County	County name

df_covid has the following columns:

Header	Meaning
County	Full Date
totalcountconfirmed	total confirmed covid cases
totalcountdeaths	total count of covid deaths
newcountconfirmed	confirmed covid cases
newcountdeaths	Covid related deaths
date	Full Date

1.7.1 Size of Dataset:

DataFrame Name	Number of Rows	Number of Columns
df_population	3193	164
df_poverty	3197	31
df_air	59	6
df_citations	145	8
df_covid	16,205	6

1.7.2 Dataset Link:

1. https://data.ca.gov/dataset/covid-19-citations
2. https://data.ca.gov/dataset/covid-19-hospital-data
3. https://data.ca.gov/dataset/covid-19-cases
4. https://www2.census.gov/programs-surveys/popest/tables/2010-2019/counties/totals/co-est2019-annres.xlsx
5. https://data.ers.usda.gov/reports.aspx?ID=17828

1.7.3 Ethical Consideration

With our research question concerning where an individual lives and how that affects their chance of contracting Covid-19 and the environmental factors that contribute to it there are potentially

dangerous implications with the misuse or misunderstanding of our study and data. Counties that possess a higher number of Covid-19 violations citations may be deemed to be negligent and unfairly blamed for a lax response to the pandemic when in reality it may be due to other compounding factors such as simply having a higher population or a lack of resources to properly follow guidelines.

Since the study also analyzes how poverty rates affect the cases of Covid-19 in a county it is possible that the data's findings can be weaponized politically against the very people it is trying to help. For instance, if poverty stricken individuals are demonized or seen as carriers of the disease then it could realistically lead to a social outcasting and avoidance which only compounds the issue where they fundamentally lack the resources to properly combat the disease and stay healthy.

In fact, throughout the initial exploration of the factors this study focuses on, it was clear to see that this study will be entering into a complex socio-economic and political discussion on how to best combat the virus. Taking this into consideration our group tried its best to be as transparent and clear with our visualizations, methodologies, and to explain the conclusions so that it would be as difficult as possible to misuse the study for malicious purposes.

Covid-19 is speculated to have prolonging health effects, even after the individuals no longer have the virus. Our analysis outlines how some aspects of a county create an environment where the virus is more likely to be contracted, which may mean that citizens will have to deal with the long-term health consequences. Health insurance companies may use this to profile people in counties and base premiums off this. Taking this into consideration, our group emphasises that the relationship between our variables of analysis and Covid-19 cases as correlational, not casual. We are not able to say that areas with certain variables do not cause higher proportions of the population to have the virus.

1.7.4 Privacy Consideration

The data sets utilized in this study were only from government agencies that were open for public use and did not usually contain any personal identifiable information since the focus was on a California county level. One data set in particular did include personal information which was the Covid-19 related citation database that included the address of the violation and the name of the offender if it was possible. To ensure the privacy and confidentiality of the businesses the columns including addresses and names of the violators were intentionally removed from the data frame. This was also an ethical consideration because we do not intend to "expose" these businesses for the violations. Our group has fully believed that this study and report has followed all privacy guidelines and considerations.

1.7.5 Conclusion

After conducting analysis of the datasets and visualizing the models of linear regression, the team has agreed that the hypothesis was indeed correct: a person's location in California affects their chances of contracting covid-19. The air quality, population size, and number of citations in California county has all strongly correlated to the total number of confirmed cases in that county. The only factor that seemed to not affect the total number of cases was the poverty rate of the county which was surprising to discover. It was hypothesized that counties with fewer resources would have a difficult time dealing with the pandemic and inhabitants would've been in financial situations that would've put them at higher risk as they could not afford to not work. Extrapolating

upon our discoveries, it appears that the findings correspond with realistic expectations. Larger counties have more people, which means there is likely to be more pollution, explaining the bad air quality index, and more people to break social distancing rules leading to higher violations. This understanding of why denser and more populated areas appeared to be "more infected" is why the team chose to plot 2 choreopleths, one showing a sheer number of confirmed cases, and one show the number of covid cases relative to population size. The goal was to show that although counties like Los Angeles and the Bay Area were depicted as "heavily infected", in reality, they performed fairly well on a per capita basis.

2 Code Begins

2.0.1 Importing Packages and Pips

[101]:
```
# !pip install plotly
# !pip install plotly-geo
# !pip install geopandas==0.3.0
# !pip install pyshp==1.2.10
# !pip install shapely==1.6.3
```

[102]:
```
%matplotlib inline
import pandas as pd
import numpy as np
from functools import reduce
import matplotlib.pyplot as plt
import seaborn as sns
import os
import patsy
import scipy.stats as stats
from scipy.stats import ttest_ind, chisquare, normaltest
import statsmodels.api as sm
from sklearn.linear_model import LinearRegression
import plotly.figure_factory as ff
from IPython.display import Image
import warnings
warnings.filterwarnings("ignore")
```

2.0.2 Converting .csv & .xls to Pandas Dataframe

[103]:
```
# upload AQI CSV file
df_air = pd.
 ↪read_csv('Kidsdata-environment-airquality-ozone-Air-Quality-Days-wit.csv')

# upload Population Estimate CSV file
df_population = pd.read_csv('co-est2019-alldata.csv',encoding = "ISO-8859-1")
```

```
# upload Citation CSV file
df_citation = pd.read_csv('8ea0217f-b19c-4fc1-974a-6823b51e4c35.csv')

# upload CA covid cases CSV file
df_covid = pd.read_csv('statewide_cases.csv')


# upload Poverty Level CSV file
df_poverty = pd.read_excel('est18all.xls')
```

2.0.3 Helper Methods

[87]:
```
def Reverse(lst):
    return [ele for ele in reversed(lst)]

def find_linear_regression(X,Y):
    A1 = np.vstack([np.ones(len(X)), X]).T
    w1 = np.linalg.lstsq(A1,Y, rcond= None)[0]
    x_smooth = np.linspace(min(X), max(X), 100)
    y_M1 = w1[0] + (w1[1] * x_smooth)
```

2.0.4 Dataframe Merging

[88]:
```
# clean up the data by removing unneeded columns, renaming columns
df_citation = df_citation[['COUNTY']]
df_citation.columns = ['County']
df_citation['violation_count'] = '1'

# set index to county name
#df_citation = df_citation.set_index('County')

# format data so that dataframe displays the count of violations for each county
df_citation = df_citation.groupby(['County'],as_index = False).count()

df_citation.head(5)


# clean up the data by removing unneeded columns, renaming columns
df_population = df_population.loc[df_population['STNAME'] == 'California']
df_population = df_population[['CTYNAME','POPESTIMATE2019']]
df_population.columns = ['County','Population_Estimate']

# remove "county" from end of county string
```

```
df_population['County'] = df_population['County'].map(lambda x: x.lstrip('+-').
 ↪rstrip('County'))

# set index to county name
#df_population = df_population.set_index('County')

df_population.head(5)


# clean up the data by removing unneeded columns, renaming columns
df_poverty = df_poverty.loc[df_poverty['Unnamed: 2'] == 'CA']
df_poverty = df_poverty[['Unnamed: 3','Unnamed: 7']]
df_poverty.columns = ['County','Poverty_Percent']

# remove "county" from end of county string
df_poverty ['County'] = df_poverty['County'].map(lambda x: x.lstrip('+-').
 ↪rstrip('County'))

# set index to county name
#df_poverty = df_poverty.set_index('County')

df_poverty.head(5)



# clean up the data by removing unneeded columns, renaming columns
df_air=df_air[['Location','Data']]
df_air.columns = ['County','Bad_Ozone']

# remove "county" from end of county string
df_air ['County'] = df_air['County'].map(lambda x: x.lstrip('+-').
 ↪rstrip('County'))

df_air.head(5)

# Cleaning df_covid table

df_covid = df_covid.loc[df_covid['date'] == '2020-12-13'].reset_index()
df_covid = df_covid.rename(columns={'county':'County'})
df_covid = df_covid[df_covid.County != 'Out Of Country']
df_covid.County = df_covid.County.str.replace(' ', '')
df_covid = df_covid.drop(['newcountconfirmed', 'newcountdeaths', 'date'], 1)


# concat all data into a single data table
df = [df_population, df_poverty, df_air]
```

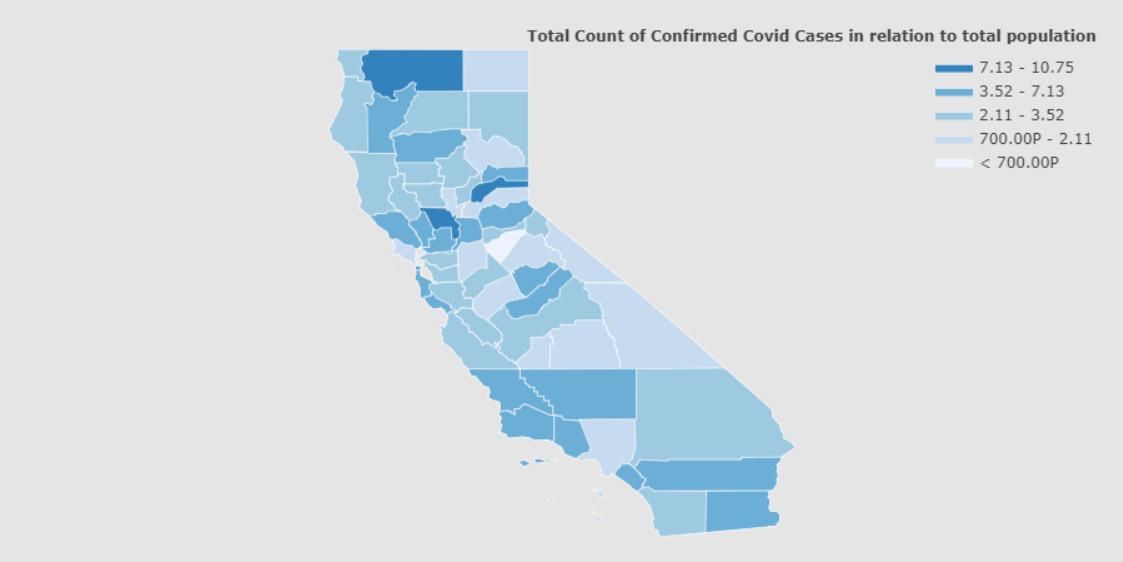
```
df = reduce(lambda left,right: pd.merge(left,right,on='County'), df)

df.County = df.County.str.replace(' ', '')

df = pd.merge(df_covid, df, on = 'County')

df['Bad_Ozone'] = df['Bad_Ozone'].str.replace('N_A', '0')
df['Bad_Ozone'] = df['Bad_Ozone'].astype(int)

df = df.merge(df_citation, how='left')
df = df.drop('index', 1)
df['violation_count'].fillna(0, inplace=True)
```

2.1 Choropleth map

[89]:
```
df_sample = pd.read_csv('https://raw.githubusercontent.com/plotly/datasets/
↪master/minoritymajority.csv')
df_sample_r = df_sample[df_sample['STNAME'] == 'California']
#values = df_sample_r['TOT_POP'].tolist()
values = df['totalcountconfirmed'].tolist()
fips = df_sample_r['FIPS'].tolist()



lower_endpt = min(df.totalcountconfirmed)
higher_endpt = max(df.totalcountconfirmed)
middle = np.mean(df.totalcountconfirmed)

lower_middle = (lower_endpt + middle) / 2
higher_middle = (higher_endpt + middle) / 2

endpts = [lower_endpt, lower_middle, middle, higher_middle, higher_endpt]

colorscale = ['#eff3ff' ,'#c6dbef','#9ecae1',
              '#6baed6','#3182bd','#08519c']

colorscale = Reverse(colorscale)

fig = ff.create_choropleth(
    fips=fips, values=values, scope=['California'], show_state_data=True,
    colorscale=colorscale, binning_endpoints=endpts, round_legend_values=True,
    plot_bgcolor='rgb(229,229,229)',
    paper_bgcolor='rgb(229,229,229)',
    legend_title='Total Count of Confirmed Covid Cases by County',
    county_outline={'color': 'rgb(255,255,255)', 'width': 0.5},
    exponent_format=True,
```

```
)
fig.layout.template = None
Image(filename='newplot (2).png')
```

[89]:

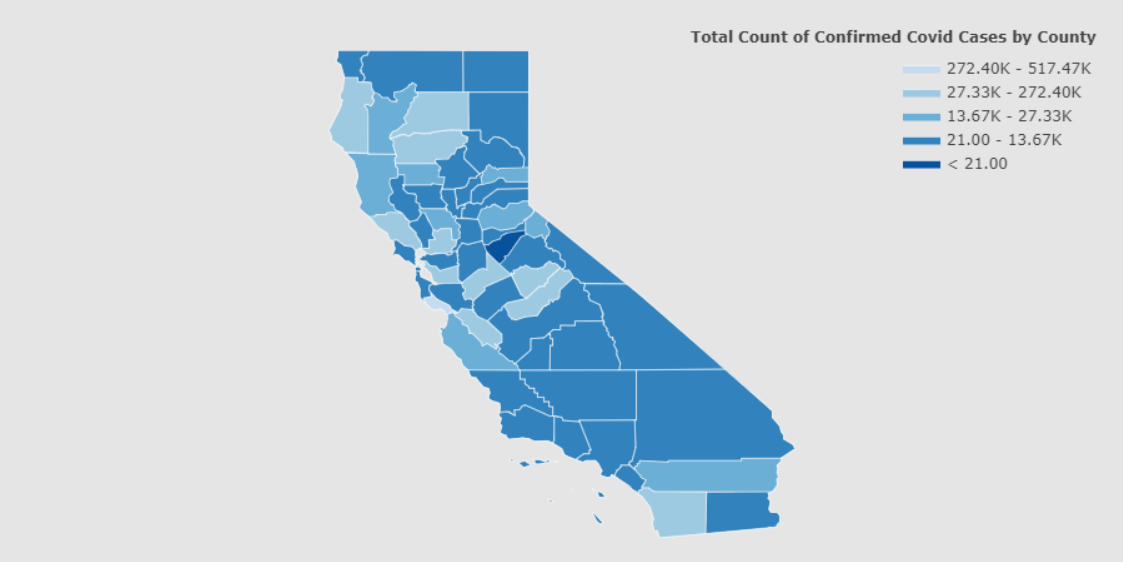


[90]:

```
df_sample = pd.read_csv('https://raw.githubusercontent.com/plotly/datasets/
 ↪master/minoritymajority.csv')
df_sample_r = df_sample[df_sample['STNAME'] == 'California']
#values = df_sample_r['TOT_POP'].tolist()
values = (df['totalcountconfirmed'] / df['Population_Estimate']) * 100
fips = df_sample_r['FIPS'].tolist()


lower_endpt = min(values)
higher_endpt = max(values)
middle = np.mean(values)

lower_middle = (lower_endpt + middle) / 2
higher_middle = (higher_endpt + middle) / 2

endpts = [lower_endpt, lower_middle, middle, higher_middle, higher_endpt]


colorscale = ['#eff3ff' ,'#c6dbef','#9ecae1',
              '#6baed6','#3182bd','#08519c']



fig = ff.create_choropleth(
```

```
    fips=fips, values=values, scope=['California'], show_state_data=True,
    colorscale=colorscale, binning_endpoints=endpts, round_legend_values=False,
    plot_bgcolor='rgb(229,229,229)',
    paper_bgcolor='rgb(229,229,229)',
    legend_title='Total Count of Confirmed Covid Cases in relation to total␣
 ↪population',
    county_outline={'color': 'rgb(255,255,255)', 'width': 0.5},
    exponent_format=True,
)
fig.layout.template = None
Image(filename='newplot (3).png')
```

[90]:



2.2 Linear Regression

2.2.1 Predicted Covid Cases = w0 + (w1 * features)

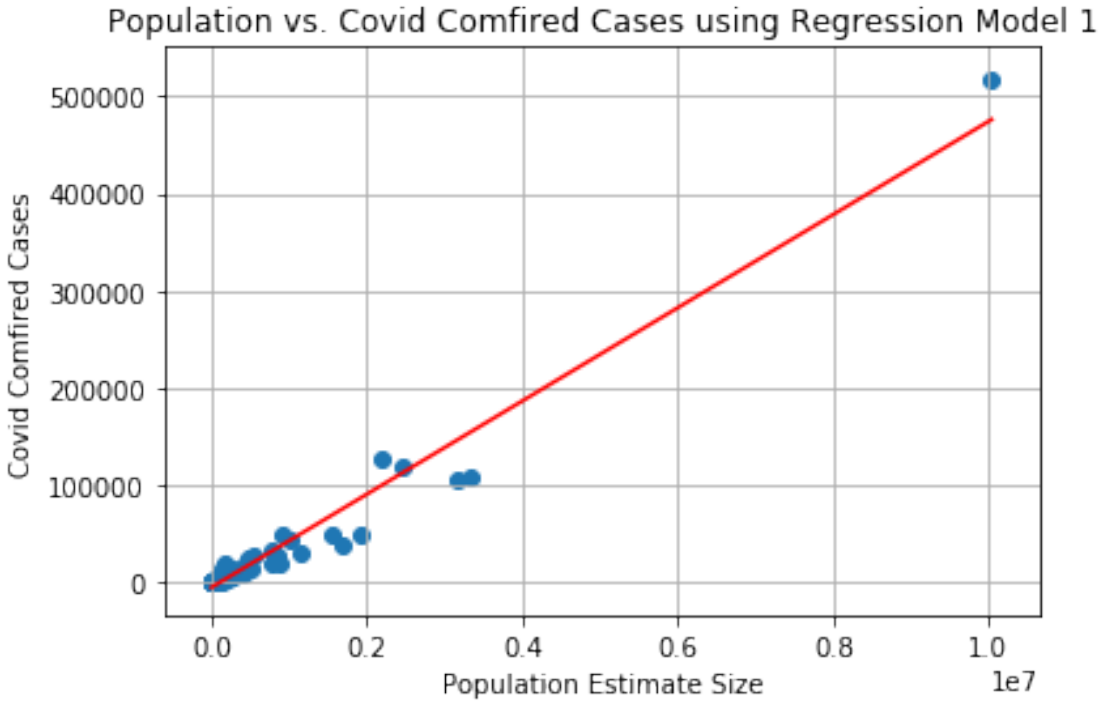
Features include: 1. Population Size 2. Poverty Percentage 3. Bad Ozone Citations Count 4. Violations Citations Count

2.2.2 Model 1

Confirmed Covid Cases = w0 + (w1 * Population Size)

[91]:
```
X = df.Population_Estimate
Y = df.totalcountconfirmed

A1 = np.vstack([np.ones(len(X)), X]).T
```

```
w1 = np.linalg.lstsq(A1,Y, rcond= None)[0]
x_smooth = np.linspace(min(X), max(X), 100)
y_M1 = w1[0] + (w1[1] * x_smooth)

plt.plot(x_smooth, y_M1, color='red')
plt.scatter(X,Y)
plt.xlabel('Population Estimate Size')
plt.ylabel('Covid Confirmed Cases')
plt.title('Population vs. Covid Comfired Cases using Regression Model 1')
plt.grid()
```

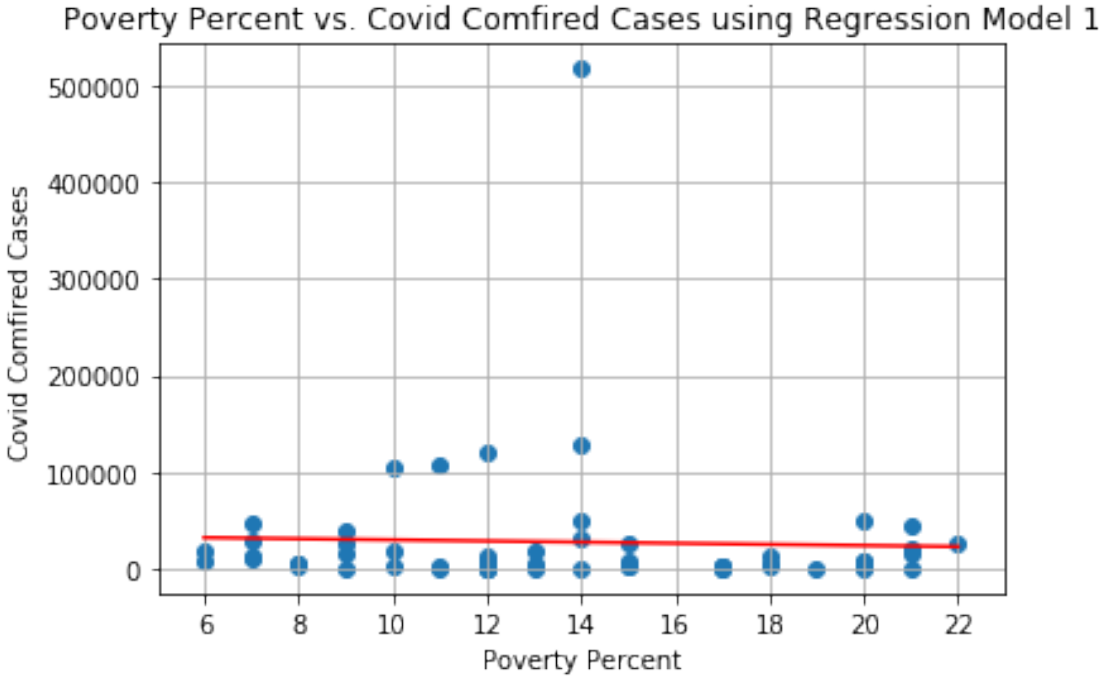


2.2.3 Model 2

Confirmed Covid Case = w0 + (w1 * Poverty Percent)

[92]:
```
X2 = df.Poverty_Percent.astype(int)
Y2 = df.totalcountconfirmed

A2 = np.vstack([np.ones(len(X2)), X2]).T
w2 = np.linalg.lstsq(A2,Y2, rcond= None)[0]
x_smooth2 = np.linspace(min(X2), max(X2), 100)
y_M1 = w2[0] + (w2[1] * x_smooth2)
```

```
plt.plot(x_smooth2, y_M1, color='red')
plt.scatter(X2,Y2)
plt.xlabel('Poverty Percent')
plt.ylabel('Covid Confirmed Cases')
plt.title('Poverty Percent vs. Covid Comfired Cases using Regression Model 1')
plt.grid()
```

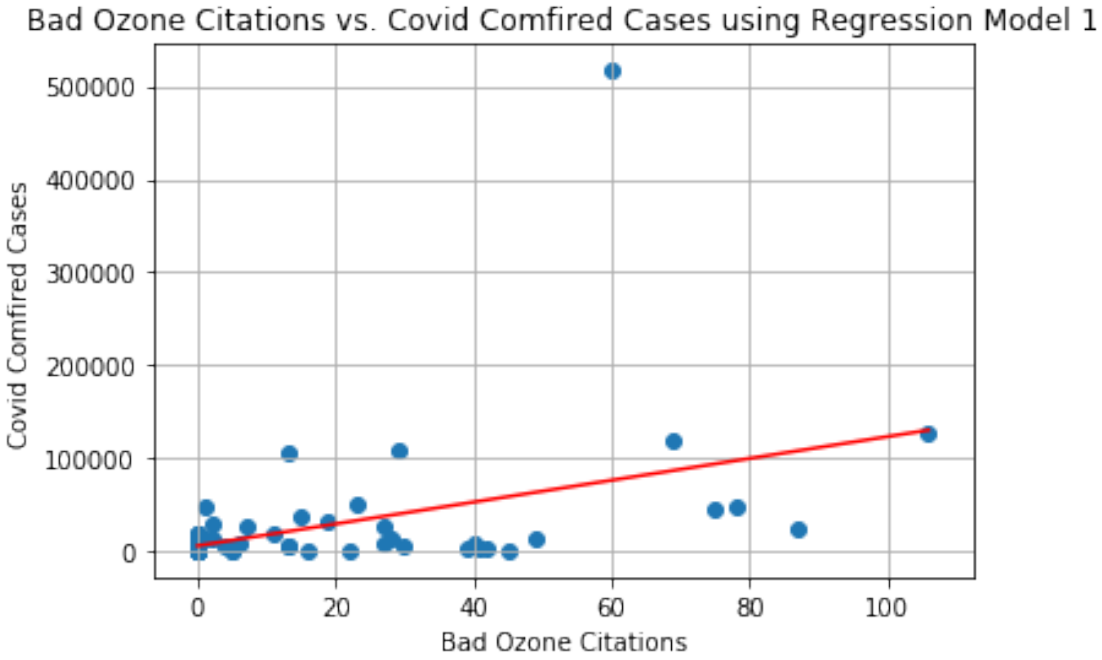


2.2.4 Model 3

Confirmed Covid Case = w0 + (w1 * Bad Ozone Citations)

[93]:
```
X3 = df.Bad_Ozone
Y3 = df.totalcountconfirmed

A3 = np.vstack([np.ones(len(X3)), X3]).T
w3 = np.linalg.lstsq(A3,Y3, rcond= None)[0]
x_smooth3 = np.linspace(min(X3), max(X3), 100)
y_M1 = w3[0] + (w3[1] * x_smooth3)

plt.plot(x_smooth3, y_M1, color='red')
plt.scatter(X3,Y3)
plt.xlabel('Bad Ozone Citations')
plt.ylabel('Covid Confirmed Cases')
```

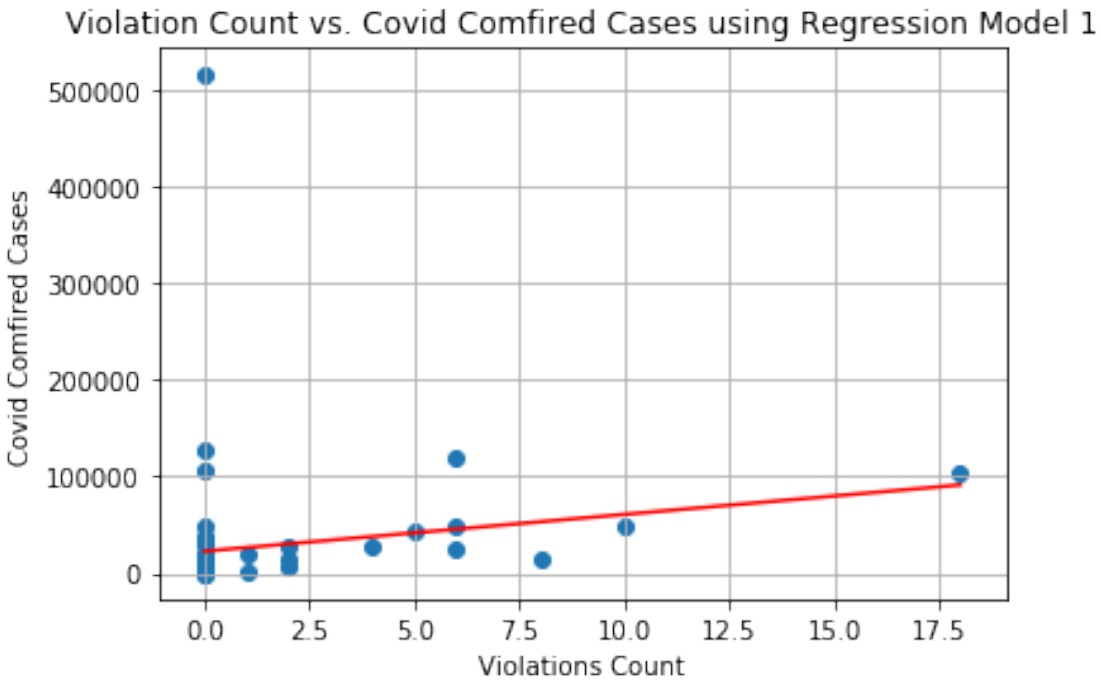
```
plt.title('Bad Ozone Citations vs. Covid Confirmed Cases using Regression Model␣
 ↪1')
plt.grid()
```



2.2.5 Model 4

Confirmed Covid Case = w0 + (w1 * Violation Count)

[100]:
```
X4 = df.violation_count
Y4 = df.totalcountconfirmed

A4 = np.vstack([np.ones(len(X4)), X4]).T
w4 = np.linalg.lstsq(A4,Y4, rcond= None)[0]
x_smooth4 = np.linspace(min(X4), max(X4), 100)
y_M1 = w4[0] + (w4[1] * x_smooth4)

plt.plot(x_smooth4, y_M1, color='red')
plt.scatter(X4,Y4)
plt.xlabel('Violations Count')
plt.ylabel('Covid Confirmed Cases')
plt.title('Violation Count vs. Covid Comfired Cases using Regression Model 1')
plt.grid()
```


Violation Count vs. Covid Comfired Cases using Regression Model 1
Covid Comfired Cases
500000
400000
300000
200000
100000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Violations Count