Disease Spread: COVID-19 spread in the United States.

Marcus Stafford

GIS 470

Professor Bardin

April 28, 2022

**Introduction:**

The research in this paper will focus on the spread of COVID-19 in the United States in all 50 U.S. states in February 2022. The goal of this research is to answer the question of which and what kind of states had the highest COVID-19 infection rates in February 2022 per one hundred thousand people. The data in this analysis is based on the **COVID\_cases\_per\_100k.csv file.**

**Analysis:**

**I hypothesize that states with higher rates of health insurance coverage will have lower rates of COVID-19 and that it will be a negative correlation. My reasoning is that people who have health insurance will be more likely to receive the COVID-19 vaccine and to get tested. As the health insurance rate increases, the COVID-19 infection rate will decrease. Second, I believe that there will be a positive correlation between the poverty rate and the COVID-19 infection rate. My reasoning is that poverty will impact a person’s ability to get vaccinated or tested due to lack of transportation or other economic considerations. Finally, I hypothesize that there will be a negative correlation between the college graduation rate and the number of COVID-19 cases. As the college graduation rate increases, the number of COID-19 cases decreases. I base my reasoning on that a person with a college degree is likely to have health insurance and stable transportation.**

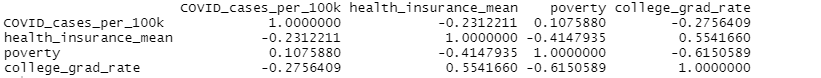
**Results:**

**Next, I will cover the mean and standard deviations of both my dependent and independent variables (except for the variable “region\_short”, which is a categorical variable).**

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Mean** | **Standard deviation** |
| **COVID\_cases\_per\_100k** | **33.13725** | **17.90533** |
| **health\_insurance\_mean** | **0.889356** | **0.04261372** |
| **poverty** | **0.09900143** | **0.03009081** |
| **college\_grad\_rate** | **0.359683** | **0.0752454** |

**Figure 1. This table shows the mean and standard deviation for all variables (excluding the one categorical variable).**

**Next, I analyzed the correlation coefficients of my continuous variables. The results are presented as follows.**



**Figure 2. The correlation coefficients show patterns among each variable. The COVID-19 case rate shows a weak negative correlation with both the health insurance mean and the college graduation rate variable. The poverty variable shows an extremely weak correlation that is positive.**

**Next, I wanted to create a scatterplot that looks at my dependent variable in relation to my three independent continuous variables. The scatterplots (See Figures 4, 5, and 6) also show the correlation coefficients (above) with the dependent variable and their relationships.**

Chart, scatter chart

Description automatically generated

**Figure 4.**

Chart, scatter chart

Description automatically generated

**Figure 5.**

Chart, scatter chart

Description automatically generated

**Figure 6.**

**For my next portion, I wished to look at the COVID-19 rate per state overlayed on a map.**

Map

Description automatically generated

**Figure 7. The states with the highest rates of COVID-19 are Idaho and Kentucky. The states with the lowest rates are Nebraska and Illinois. Generally, states in the Southeast have the highest rates of COVID-19 with the lowest rates located in the Midwest.**

**Lastly, I am now going to perform five regression models to test the impact that each of my independent variables has on my dependent variable. My dependent variable is “COVID\_cases\_per\_100K”, while my independent variables are “health\_insurance\_mean”, “poverty”, and “college\_grad\_rate” (X1, X2, and X3).**

Table

Description automatically generated

**Figure 6. The health insurance variable is not significant (P > 0.05), and it explains only 5% of the dependent variable. For every one-unit increase, the COVID-19 infection rate decreases by 4.15.**

Table

Description automatically generated

**Figure 7. The independent variable “poverty” is not significant with a p-value of 0.46 and explains 1% of the dependent variable. For every one-unit increase in the poverty rate, the COVID-19 infection rate increases by 1.91.**

Table

Description automatically generated

**Figure 8. The college grad rate explains 9% of the dependent variable. It’s significant with a p-value < 0.05. For every one-unit increase in the independent variable, the COVID-19 rate decreases by 5.35.**

A screenshot of a computer

Description automatically generated with low confidence

**Figure 9. The “region\_short” variable shows that the “West” region was most impacted and is the reference variable. Only the “Midwest” region has a p-value < 0.05.**

Table

Description automatically generated

**Figure 10.**

**In my multiple regression model, considering all variables, this model explains 27% of the dependent variable. The college graduation rate, and the Midwest region, are significant (P < 0.05). Areas that have a high college graduation rate are impacted by COVID-19. The region most impacted is the “West” region when only the “region\_short” variable is considered (see Figure 9). In the multiple regression model, the region most impacted by COVID-19 is the Northeast (4.52 more), the South (3.13 more), with the “Midwest” region being the least impacted (14.43 fewer). When accounting for all variables, this model explains more of the dependent variable than in the previous variable models (at 27%). The poverty and college grad rate numbers (per unit increase) all improve, while the health insurance variable becomes worse. However, as this model explains only 27% of the dependent variable, further examination of other independent variables is needed to better explain the COVID-19 infection rate.**

****