 ISE 5435 Data Analytics

**COVID-19 Case Rate Model**

Fall 2021

Instructor: Dr. Nasim Nezamoddini

**Submitted By Group 01:**

Ashraf Fakhouri

Josh Greib

Ra’fat Alrawashdeh

**Submitted On:** Dec 7, 2021

**Abstract**

The ongoing COVID-19 pandemic has resulted in the death of millions of people globally within two years. In an effort to better predict which U.S. states are more likely to suffer from high case rates in future pandemics, a logistic regression model was developed to determine which socio-economic factors have an influence on COVID-19 cases rates. The model was also developed to help predict which states would have low, medium, or high case rates. The regression model that was developed indicated that higher levels of population density, poverty levels, and percent of population with at least some college all contributed to a statistically significant increase in COVID-19 case rates. While, higher percentages of the population with the age of 65 or greater resulted in statistically significant lower case rates of COVID-19. The regression model was able to correctly predict the level of case rate (low, medium, or high) of each state, based on these socio-economic factors, with a 84% success rate using the training data set and with a 92% success rate in the validation data set.

**Introduction**

The ongoing Coronavirus pandemic (COVID-19) starting from 2019 is the first global pandemic of the 21st century. This pandemic killed millions of people within a time span of two years. The COVID-19 pandemic was unique in that the data regarding the number of cases around the world was tracked with higher accuracy than previous pandemics. This is due to the ease at which data can be recorded with the use of computers. The COVID-19 case rate data from around the world is easily and freely accessible to anyone that wishes to access it. Due to the easy accessibility of this data and our team’s interest in this ongoing pandemic we decided to conduct our research into the COVID-19 pandemic.

**Problem and Data Description**

The question we wanted to answer is, are there any pre-existing socio-economic factors that influence the COVID-19 case rate? If yes, can we develop a model to predict which states or countries will be affected by higher cases using these socio-economic factors? In order to simplify this assessment, our area of study was limited in scope to the 50 U.S. states.

The data for COVID-19 cases rates was accessed via github.com. This data is collected and published by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The COVID-19 cases are broken down into cases per state and the data is updated everyday based on the number of cases reported by each state’s health department.

The team brainstormed socio-economic factors that we hypothesised would have an impact on the COVID-19 case rate. The list of factors we chose to include in the model are as follows: population, geographical area, population density (calculated from total population and geographical area), gross domestic product (GDP), number of COVID-19 cases, percent of population 18 years old or younger, percent of population between the ages of 19 and 64, percent of population 65 years or older, average temperature within the state, median age, gender ratio, percent of population below poverty level, unemployment percent, percent of adults with less than a high school diploma, percent of adults with a high school diploma, percent of adults completing at least some college, and percent of adults with at least a bachelor’s degree.

In order to assess how these factors influenced the spread of COVID-19 within each state, the data was collected from the most recently recorded data from before the start of the COVID-19 pandemic. In most cases the data is from 2019 or 2020. Most of the socio-economic data was taken from reports published by U.S. federal government organizations including the Census Bureau, the Bureau of Economic Analysis, and the U.S. Department of Agriculture Economic Research Service. The data from these various sources was collected for each U.S. state and tabulated by the team for use in our COVID-19 case rate model.

**Methodology**

Our dataset, containing fourteen features expecting that might be a significant and one target variable (dependent) which is the “Case Rate”, was obtained from US government websites as listed in the references’ section as shown in table 1. Logistic regression was used to describe our model knowing that there are two types of logistic regression: Nominal and ordinal. The case rate was categorized as three classes:

* 0: low ”case rate < 0.05”
* 1: medium “0.05≤case rate <0.10”
* 2: high “case rate ≥ 0.10”)

Hence, our target to predict the spread rate of COVID-19, the ordinal logistic regression will be used because the dependent variable has a meaningful value.

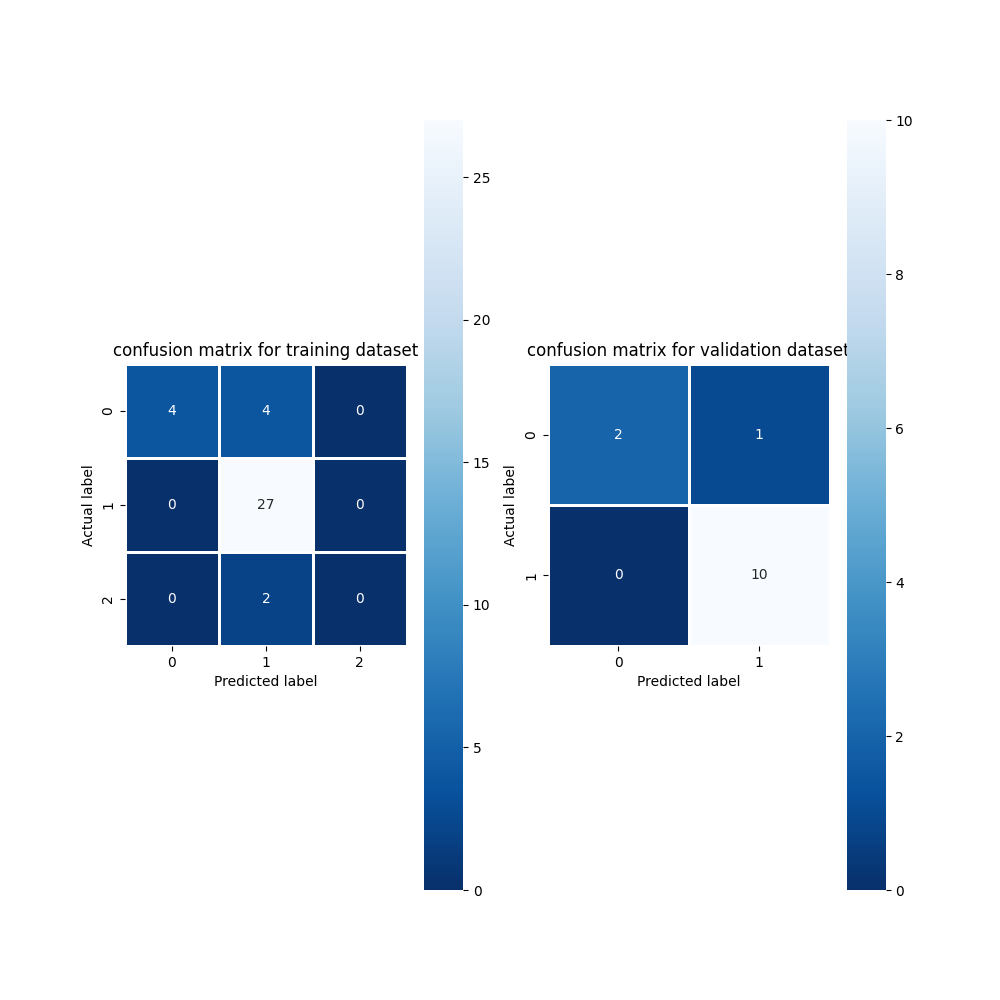
*Table 1: Variable’s name and full description*

|  |  |
| --- | --- |
| Column name | Full name |
| Population Density | Population Density P/Sq. MI. |
| GDP\_2020 | TOTAL GDP (Millions) |
| Case Rate | Case Rate (Cases per Total Population) |
| Age≤18 | Age ≤ 18 |
| 19≤Age≤64 | 19 ≤ Age ≤ 64 |
| Age≥65 | Age ≥ 65 |
| Weather | Weather |
| Median Age | Median Age |
| Gender Ratio | Gender Ratio (Male to Female) |
| Poverty Level | Poverty Level (Percent of People) |
| Unemployment % | Unemployment Level (Percentage of Workforce) |
| % adults < high school | Percent of adults with less than a high school diploma, 2015-19 |
| % adults = high school | Percent of adults with a high school diploma only, 2015-19 |
| % adults = some college or associate degree | Percent of adults completing some college or associate degree, 2015-19 |
| % adults = bachelor's degree or higher | Percent of adults with a bachelor's degree or higher, 2015-19 |

Python was used (See Appendix A1 and A2) to find the required outcomes following these steps as below:

* Required package was imported
* Data was loaded
* Preprocessing the data defining two variables: X and y. X was scaled using standard scaler and y was categorized based on the classification that mentioned previously.
* Define function named “Best\_subset\_search” returning the best score, best subset variable names and best subset dataset using the cross validation with number of folds =4 and sequential features selectors as shown in Figure 1
* Find the Theta and the coefficients of our Model as presented in Figure 1
* Split the data to training and validation datasets and evaluate our model using the validation dataset as well as finding the model performance for both datasets as introduced in Figure 2
* Figure 3 visualizing the difference in prediction performance between training and validation datasets using Confusion matrix.

|  |  |
| --- | --- |
| *Figure 1: Best subsets, scores and the Model parameters* | *Figure 2: The classification reports for training and validation datasets* |



*Figure 3: Confusion matrix for training and validation datasets*

**Results and Discussion**

The model based on the training dataset has a good accuracy with 0.84 and it is expected because we don’t have enough data in a high range of case rate to train it, but it did well on the prediction with accuracy 0.92. On the other hand, the coefficient value of *Age≥65* has a negative impact on our model compared with the rest of features that have a positive one.

The following socio-economic factors were found to have statistically significant influence on COVID-19 cases rates within each US state:

Population density

Distancing has been considered as one of the most effective ways to prevent the spread of COVID-19 prior to the availability of vaccination and therapeutic drugs. The distance people can be separated from each other could be considered an individual choice, however it also could be affected by population density. Our model shows that population density is one of the most effective predictors of COVID-19 cases in the U.S. at the state level. The higher the population density - The higher the COVID-19 infection rate

Percent of population whose age is equal or higher than 65

More than 81% of COVID-19 deaths occur in people over age 65. The number of deaths among people over age 65 is 80 times higher than the number of deaths among people aged 18-29. Even though older adults are more likely to get severely ill from the virus, our study shows that age is inversely proportional with the infection rate. The higher number of people over the age of 65 - the lower COVID-19 infection rate.

This surprising finding could be explained as follows:

* older adults having earlier access to vaccines
* older adults being more cautious due to the severe ramifications of infection
* In general, people often grow less physically active as they age, thus, we can argue that the older one gets - the less physically active one is - the less socially active one becomes - The lower one’s chances of infection

Poverty level

Our analysis shows that the higher poverty levels in a state - the higher COVID-19 infection rate. This finding could be related to the following:

* Workers in high poverty areas are less likely to work from home, suggesting that they therefore may also be less able to distance themselves physically than those living in less poverty areas.
* Individuals may be faced with the choice between staying home and not getting paid, or going to work and increasing their risk of becoming infected with the virus.
* People living in high poverty areas have less access to virus-spread-prevention tools such as (Sanitizers, gloves and masks)

Percent of population with some college or associate degree

Our study has shown that the higher the number of people with some college or associate degree in one area - the higher COVID-19 infection rate in comparison to the percentage of population with a bachelor's degree or higher. This could be due to the fact that people with a bachelor’s degree or higher had the chance to work from home and reduce their physical presence and close contact with their colleagues. In addition, people with some college or associate degree could be less likely aware or cautious in their contact with others.

**Conclusion**

This team developed a model that found four socio-economic factors that were statistically significant factors in the COVID-19 case rate within each state. These factors being, poverty level, population density, percent of population with an age of 65 or greater, and percentage of population with some college or an associates degree. The model was able to correctly predict the COVID-19 case rate with a 92% accuracy level using the validation data set. However, we did notice that the model is tuned better towards predicting the “low” and “medium” states and is not as well tuned to predicting the states with a “high” case rate. This is due to the fact that only two states were classified as having a “high” case rate. To improve the model in the future, we could modify the buckets of “low”, “medium”, and ‘high” case rates to spread the case rate data more evenly across the buckets.

**Reference**

* [State-by-State Visualizations of Key Demographic Trends From the 2020 Census](https://www.census.gov/library/stories/state-by-state.html)
* [State Area Measurements and Internal Point Coordinates](https://www.census.gov/geographies/reference-files/2010/geo/state-area.html)
* [Historical Population Density Data (1910-2020)](https://www.census.gov/data/tables/time-series/dec/density-data-text.html)
* [Gross Domestic Product by State](https://apps.bea.gov/regional/histdata/releases/0621gdpstate/index.cfm)
* [COVID-19/12-31-2020.csv at master · CSSEGISandData/COVID-19 (github.com)](https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_daily_reports_us/12-31-2020.csv)
* [Population Distribution by Age | KFF](https://www.kff.org/other/state-indicator/distribution-by-age/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D)
* <http://www.usa.com/rank/us--average-temperature--state-rank.htm>
* <https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_median_age>
* <https://worldpopulationreview.com/state-rankings/male-to-female-ratio-by-state>
* <https://data.census.gov/cedsci/table?q=Official%20Poverty%20Measure&g=0100000US%240400000&tid=ACSST1Y2019.S1701&hidePreview=true&moe=false&tp=true>
* <https://www.bls.gov/news.release/archives/laus_01262021.pdf>
* <https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>

[](https://docs.google.com/spreadsheets/d/1wkjObq1tXtiM46POiLdmQG5bbTtxEnJJeeqkyIAAGes/edit#gid=0) [](https://docs.google.com/presentation/d/1--L6YnxYwyBdTc_WLwWjoLWmiBMbCFLMsF4xhMBvqV8/edit#slide=id.p1) [ISE5435 Group Project Data](https://docs.google.com/spreadsheets/d/1wkjObq1tXtiM46POiLdmQG5bbTtxEnJJeeqkyIAAGes/edit#gid=0) [ISE 5435 Group Project Presentation](https://docs.google.com/presentation/d/1--L6YnxYwyBdTc_WLwWjoLWmiBMbCFLMsF4xhMBvqV8/edit#slide=id.p1)

**Appendices**

Appendix A1

1. # Import required packages
2. import pandas as pd
3. from sklearn.model\_selection import KFold
4. from mlxtend.feature\_selection import SequentialFeatureSelector as SFS
5. from sklearn.model\_selection import train\_test\_split
6. from sklearn.metrics import confusion\_matrix, classification\_report
7. import matplotlib.pylab as plt
8. import seaborn as sns
9. from sklearn.preprocessing import StandardScaler
10. from mord import LogisticIT
12. # Load data
13. data\_df = pd.read\_excel(r'C:\Users\rafat\Desktop\ISE5435 Group Project.xlsx', sheet\_name=data)

16. # define function to get cleaned and transformed data
17. def prepared\_data(df, dependent):
18. df.columns = [s.strip().replace(' ', '\_') for s in df.columns]
19. df.set\_index(df.State, inplace=True)
20. df.drop(columns=['State'], inplace=True)
21. # to define the case rank category of each state based its value
22. df['Case\_Rank'] = pd.cut(df['Case\_Rate'], [0, 0.05, 0.10, 0.15], labels=[0, 1, 2]).astype(int)
23. X = pd.DataFrame(df, columns=df.columns)
24. exclude\_col = ['Cases', 'Death', 'Case\_Rate', 'Death\_Rate', 'Case\_Rank']
25. X.drop(columns=exclude\_col, inplace=True)
26. scaler = StandardScaler()
27. X\_new = pd.DataFrame(scaler.fit\_transform(X), columns=X.columns, index=X.index)
28. y = df[dependent]
29. return X\_new, y

32. # get X and y from our dataset
33. dependent = 'Case\_Rank'
34. X, y = prepared\_data(data\_df, dependent)

37. # define the estimator model and selector process which is stepwise and number of folds cv
38. def best\_subset\_research(random\_seed):
39. # build the Ordinal logistic model
40. logit = LogisticIT(alpha=0)
41. # define number of folds to divide our dataset
42. cv = KFold(n\_splits=4, random\_state=random\_seed, shuffle=True)
43. # define the sequential features selector tool using the logistic model
44. sfs1 = SFS(logit, k\_features=(1, 14), forward=True, floating=True, scoring='accuracy', cv=cv)
45. sfs1.fit(X, y)
46. best\_score = sfs1.k\_score\_
47. best\_subset\_var\_name = sfs1.k\_feature\_names\_
48. best\_subset = pd.DataFrame(sfs1.fit\_transform(X, y), columns=best\_subset\_var\_name)
49. return best\_score,best\_subset\_var\_name, best\_subset
51. # get the score, selected features and new X
52. random\_seed = 7
53. score, new\_col,X\_new = best\_subset\_research(random\_seed)
54. print(pd.DataFrame({'Selected Features': new\_col}))
55. print('------------------------------------')
56. print('Score for the cross validation = {:.2f}'.format(score))
57. print('------------------------------------')
59. # split data to train and valid dataset based on new X
60. train\_X, valid\_X, train\_y, valid\_y = train\_test\_split(X\_new, y, test\_size=0.25, random\_state=random\_seed)
62. # build the Ordinal logistic model for the new X and get the theta and the coefficients of the model
63. model = LogisticIT(alpha=0)
64. model.fit(train\_X, train\_y)
65. print(' theta', model.theta\_)
66. print(' coefficients', model.coef\_)
68. # find the probability of each state (0, 1, 2) for the validation dataset
69. valid\_probs = model.predict\_proba(valid\_X)
70. # find the prediction for the validation dataset
71. valid\_prediction = model.predict(valid\_X)
72. train\_prediction = model.predict(train\_X)
73. # find the prediction for the validation dataset
74. valid\_prediction = model.predict(valid\_X)
75. # evaluate our model
76. results = pd.DataFrame({'actual': valid\_y, 'predicted': valid\_prediction, 'P(0)': [p[0] for p in valid\_probs],'P(1)': [p[1] for p in valid\_probs], 'P(2)': [p[2] for p in valid\_probs] })
77. print(results)
79. # Plot confusion matrix graph for training and validation dataset
80. confusion\_matrix\_train = pd.DataFrame(confusion\_matrix(train\_y, train\_prediction))
81. confusion\_matrix\_valid = pd.DataFrame(confusion\_matrix(valid\_y, valid\_prediction))
82. fig = plt.figure(figsize=(10, 10)) # create figure
83. ax0 = fig.add\_subplot(1, 2, 1) # add subplot 1 (1 row, 2 columns, first plot)
84. ax1 = fig.add\_subplot(1, 2, 2) # add subplot 2 (1 row, 2 columns, second plot). See tip below\*\*
85. sns.heatmap(confusion\_matrix\_train, annot=True, fmt=".0f", linewidths=.8, square=True, cmap='Blues\_r', ax=ax0)
86. ax0.set\_title('confusion matrix for training dataset')
87. ax0.set\_xlabel('Predicted label')
88. ax0.set\_ylabel('Actual label')
89. sns.heatmap(confusion\_matrix\_valid, annot=True, fmt=".0f", linewidths=.8, square=True, cmap='Blues\_r', ax=ax1)
90. ax1.set\_title('confusion matrix for validation dataset')
91. ax1.set\_xlabel('Predicted label')
92. ax1.set\_ylabel('Actual label')
93. plt.show()
95. # Classification report for training and validation dataset
96. print()
97. print('Training Classification report is as:')
98. print(classification\_report(train\_y, train\_prediction))
99. print()
100. print('Validation Classification report is as:')
101. print(classification\_report(valid\_y, valid\_prediction))

Appendix A2

