Analyzing COVID-19 by Hypothesis Tests and Linear Regression

Yi Lu[1, a†], Yifan Yang[2, *, †]

[1]Stony Brook University, College of Engineering & Applied Science, NY, United States
[2]University of Wisconsin-Madison, College of Letters & Science, WI, United States
[†]These authors contributed equally.

[a]yi.lu.2@stonybrook.edu
[*]yang677@wisc.edu

Abstract. The outbreak of COVID-19 has caused urgent global challenges due to its rapid contagious characteristics. Analyzing known data from the past is one way to effectively control the spread of the pandemic. The United States is a racially diverse country; therefore, the composition of social groups is relatively complex. This article selects the confirmed cases data from March to June 2020 in Chicago for analysis. The data was divided into three categories: age, gender, and race. Latinos and blacks are more worthy of attention in the racial category, and young and middle-aged in the age group are more significant. This paper analyzes the existing data set through basic data processing, two-sample t-test and linear regression. We propose a regression model with dummy variables to analyze the generic covid data. There was not much difference between men and women in the number and rate of diagnoses, so the effect of gender in subsequent tests was not considered. In terms of age, the number and rate of confirmed diagnoses are higher in the 18 to 49-year-old group; the Latino group is more prominent among different ethnic groups, followed by blacks and whites. Finally, we put forward targeted epidemic prevention suggestions for different groups of communities and companies.

Keywords: Two-sample t test, Linear Regression

1 Introduction

The outbreak of COVID-19 has caused an urgent global challenge due to its rapid contagiousness. This characteristic leads to significant influence on society, such as the implementation of lockdown strategies in public transportation, advanced needs for sanitation, and increased governmental input in public health [9]. It also impedes economic development in agriculture, industry, and tourism fields (Kotwal, 2020). However, based on the analysis of practical methods and collective databases, predictions through researchers' efforts worldwide now can helpfully ameliorate infection of

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 29–37, 2023.
https://doi.org/10.2991/978-94-6463-102-9_5

COVID-19 [3]. And all these changes notably impact people's daily life on physical and mental health.

Covid-19 can result in enduring physiological risks to citizens' health conditions. It caused acute reparatory distress syndrome (ARDS) for people with some initial symptoms including difficulty and shortness of breathing at the beginning of eight days [2]. Lower blood oxygen levels arising from such symptoms would further result in patients' dizziness, rapid heart rate, and sweating situations while receiving treatments [5]. After one year's discharge, 51.3% of patients who have ARDS in the hospital still have ARDS problem [2]. During the rehabilitation, a series of cognitive impairments yet occur remaining to the extent of executive function, mental processing speed, memory, and concentration loss.

In addition, COVID 19 has seriously affected people's mental health. The transmission of its risks through social media has caused psychological symptoms for the public. Hence, Castro and Singer have come up with a systematic review of Meta-Analyses guidelines to measure the general public's psychological status. According to their research, the symptoms of anxiety are from 6.33% to 18.7%, PTSD symptoms are approximately 7%, and depressive symptoms are from 14.6% to 32.8% [7]. Symptoms associated with anxiety, depression, and distress occurring from younger ages, female genders, and student populations shows up at a higher frequency [6]. Moreover, the impacts of COVID-19 on different characteristics of people are another area of study for researchers. When Covid-19 appeared, they studied the psychological situations of people of different genders, ages, and student groups. As reported by them, females are more likely to get distressed as they perceive covid 19 traumatic effects among all these groups. And those who are in various occupations such as retail, service industry, and healthcare had a high percentage to get anxiety and depression [6]. Also, there are more risks for the student population to infect with Covid-19.

Analysis of the data based on multiple aspects of race, gender identities, and age is a more cost-effective way to do the research referring to the prediction of the COVID-19. Since the outbreak of COVID-19 in late 2019, there have been many researchers analyzing data on infections, deaths, and transmission by mathematically modeling. Among them, a paper by Melodie Monod and other nineteen authors concluded that transmission is more efficient in the 20-49 age group [8]. Additional interventions are proposed for this age group to control the spread of the epidemic and avoid deaths. According to the CDC, there is a relationship between infection rates by race and ethnicity [10]. The U.S. Census database from April to December 2020 was analyzed for the race variables Latino, Asian, non-Hispanic Black, non-Hispanic White, and non-Hispanic all other races. The data shows that Hispanics/Latinos had the highest rate of inpatient admissions [9]. Smita Rath and other authors use multiple linear regression models by python to analyze the new cases, active cases, and deaths in India to predict the number of active cases per day for the coming week. According to the r-square value they get in results indicate that the model is a strong predictor model [11].

In terms of using similar methods, this paper will use regression models to analyze the Chicago COVID-19 dataset to find correlations between age, race, gender, and other factors and diagnosis rates [1]. The conclusions drawn will help target different groups for more efficient control of the epidemic transmission. For COVID-19, the above-

mentioned huge mental health risks are being studied by some psychologists, which will help them to develop their resilience and better cope with the epidemic situation through the new research that is based on analyzing relationships of factors. The impact of this pandemic on everyone is different, so the psychological resilience of individuals will be more complex when analyzing problems. In terms of this, clinical and public health institutions should also pay attention to and guide the adverse impact of COVID-19 on their mental health when paying attention to the impact on people's physical health.

2 Methodologies

2.1 Data selection and processing

Regression analysis for infectious population from these factors and backgrounds help contribute to the future measurement for planning and controlling the infection tendency of COVID-19. The prediction results made by these methods from the data set can be approximate to reality that exactly reflects the correlation among different groups. The data this paper chooses is based on a Chicago's data set, and relates to COVID-19 cases from aspects of local races, ages, and genders [1]. This paper also collects positive confirmed cases and weekdays' data to test the coefficients by using dummy variables.

2.2 Hypothesis tests

Two-sample t test will be used to determine whether there is a significant difference among ages groups.

$$t = \frac{\overline{X_1} - \overline{X_2}}{SE} \tag{1}$$

where $\overline{X_1}$and $\overline{X_2}$ are two chosen groups' mean values, and *SE* is the standard error for the difference between the two population means. And the two-sample t test uses for testing independent data from two population groups with the equal variances. Also, test in this paper is used for finding the significance from young, median, and old age population who are tested for COVID-19.

2.3 Regression Modeling Framework Using Dummy Variables

Dummy variables are used in the regression analysis, which can analyze groups by using a single regression equation as follows:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_6 x_{i,2} + \cdots + \beta_6 x_{i,6} + e_i, \tag{2}$$

where y_i is the target variable of interest;

$x_{i,1}$ is the dummy variable with value equal to 1 on Mondays and 0 for other days;

$x_{i,2}$ is the dummy variable with value equal to 1 on Tuesdays and 0 for other days; …

$x_{i,6}$ is the dummy variable with value equal to 1 on Saturdays and 0 for other days;

e_i is the residual.

Therefore, coefficients $\beta_1, \ldots, \beta_6$ represent the impact on the target variable on Mondays, …, Saturdays, and β_0 indicates the impact from Sundays. If a coefficient β_i for any dummy variable is 0, it shows that the dummy variable does not have an influence on the target variable. And this paper use dummy variables to get respective coefficients for the days from Monday to Saturday for any race such as Latino, Black, White, and Asian.

2.4 Linear Regression

Linear regression is the model that can use for predicting the chosen values from the data set in terms of the linear approximation. And it is linear model that can minimize the residual sum of squares between the actual value and predicted values.

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3}$$

In this formula, n is the number of samples. y_i is the value of samples.

P-value is to measure significance of results in null hypothesis. It can reflect the reliability of the coefficients between target variable and independent variable, and make measurement to find the strong or weak correlations. The higher P-value that is larger than 0.05 indicates the weaker correlation between the variables in the data set. And the lower P-value that is smaller than 0.05 shows that there is no correlation between the target variable and independent set of data.

MSE helps measure the model in the average squared difference between observed and predicted variables.

$$MSE = \frac{\sum(y_i - \bar{y_i})^2}{n} = \frac{SSE}{n} \tag{4}$$

R^2 can be measured to models' goodness of fit and it helps reflect the how good data is fitted in a regression line and how its predictions close to real data. The coefficient of determination R^2 is defined as

$$R^2 = 1 - \frac{SS_{residual}}{SS_{Total}} \quad SS_{res} \tag{5}$$

3 Results and discussion

3.1 Basic data analysis

The dataset provides tested data from March 1 to May 31, 2020. The data are divided into three categories: age, gender, and race. First, the daily proportion of confirmed cases are calculated in different ways - age groups, genders, and races, and then calculate the average of the percentages according to different groups.

Table 1. Daily average percentage of age

Age Groups	Age 0-17	Age 18-29	Age 30-39	Age 40-49	Age 50-59	Age 60-69	Age 70-79	Age 80+
Cases	1916	7967	7948	8214	7738	5469	2945	2288
%	4.10%	16.30%	16.80%	17.90%	17.40%	12%	7.90%	4.80%

From Table 1, it can be concluded that the group tested according to age, the largest proportion is the age group of 40-49 years old, accounting for 17.90%, followed by the 18-29 and 30-39 age groups, accounting for 16.30% and 16.80% of the total confirmed cases, respectively. In terms of gender, the proportion of males' and females' cases are almost the same, close to 50%. Therefore, it can be concluded that the diagnosis rate of males and females is roughly the same. In the racial data (Table 2), Latinos have the largest number of confirmed cases, reaching 30.10% of the total number of diagnoses; the second most are blacks, accounting for 29% of them. Through the above data, it can be known that the age group of 28-49 years old and the group of Latinos and blacks is the group with the largest number of confirmed cases.

Table 2. Daily average percentage of race

Name	Latino	Asian	Black	White	Other
Cases	16776	1007	10646	5133	1868
%	30.10%	2.90%	29%	18.50%	3.70%

3.2 Hypothesis tests

Hypothesis test was used to find out if there is an effective difference between different groups. In this paper, the two-sample t-tests were chosen, and the average percentage results showed that the highest percentage of cases was in the 40-49 age group, followed by the 50-59 and 30-39 age groups. The age group 4 and age group 1 were chosen first to test for differences between age groups.

Table 3. Results of hypothesis test of age groups

Age group	H_0	H_1	P-value	Decision
1&4	$\mu1=\mu4$	$\mu1<\mu4$	<0.001	Reject
4&6	$\mu4=\mu6$	$\mu4<\mu6$	<0.001	Reject
3&4	$\mu3=\mu4$	$\mu3<\mu4$	0.386	Accept
2&4	$\mu2=\mu4$	$\mu2<\mu4$	0.398	Accept
2&3	$\mu2=\mu3$	$\mu2<\mu3$	0.508	Accept
4&5	$\mu4=\mu5$	$\mu5<\mu4$	0.296	Accept
1&2	$\mu1=\mu2$	$\mu1<\mu2$	<0.001	Reject
5&6	$\mu5=\mu6$	$\mu5>\mu6$	<0.001	Reject

Because age group 4 has the highest percentage of confirmed cases of all age groups and is the middle-aged group, it will be tested against the youngest age group 1 and the older age group 6. The results of the hypothesis tests are presented in Table 3.

Among them, H_0 was rejected for the 40-49 age group and the 0-17 and 60-69 age groups, which also indicates that there is a difference between the three major age groups of young, middle-aged, and old. Then we conducted the hypothesis test for the age groups of young and middle-aged between 18-49 years old, and it can be seen from Table 3 that H_0 was accepted, and there was no significant difference between them.

Next question is which age group we can see significant difference. Thus, the group aged 0-17 years and the group aged 18-29 years were tested and H_0 was rejected; there was also a difference between the group aged 50-59 years and the group aged 60-69 years, with a p-value less than 0.001, so H_0 was rejected. In conclusion, 18 and 60 years old groups were the dividing lines for a significant difference in the number of confirmed cases. Young children and teenagers aged 0-18 years are less susceptible to infection and the number of cases decreases significantly from age 60 years onwards.

Next, we consider racial impact. From Table 2, it is assumed that Latinos are more likely to be diagnosed than blacks. After two sample t tests, H_0 was rejected, so there was a difference between Latinos and Blacks. Then we continue with the testing of blacks and whites, by assuming that whites have a lower proportion of confirmed cases, i.e., were less likely to be diagnosed as positive. The result of the test is p-value less than 0.001, so we reject H_0. Table 4 shows the results for whites and Asians, and the results show that H_0 is rejected as well. Therefore, there are indeed differences between the races.

Table 4. Results of hypothesis test of race

Race	H_0	H_1	P-value	Decision
L&B	$\mu L=\mu B$	$\mu L>\mu B$	<0.001	Reject
B&W	$\mu B=\mu W$	$\mu W<\mu B$	<0.001	Reject
W&A	$\mu W=\mu A$	$\mu W>\mu A$	<0.001	Reject

3.3 Dummy variables and linear regression

The dataset selected in this paper collects information on the number of cases in different days of a week. Therefore, the data can also be sorted by Monday to Sunday.

Table 5. Total cases in different weekday

Date	Latino	Asian	Black	White
Monday	3167	180	1801	885
Tuesday	2772	149	1700	842
Wednesday	3137	166	1858	880
Thursday	2251	135	1532	748
Friday	2734	178	1550	827
Saturday	1424	102	972	412
Sunday	1132	106	1102	504

Table 5 lists the total number of confirmed cases for each race on different seven days of the week for the full dataset. It can be seen from the table that Latino, Black, White, and Asian have more total confirmed cases on weekdays (Monday to Friday)

than on weekends (Saturday and Sunday). Then, we want to try to find a weekly predictive model, or to better parse the current data through the results of linear regression. So dummy variables were created to do multiple linear regression of this dataset. The Dummy variables are Monday to Saturday (Sunday is the intercept), using only 0 and 1 for presence and absence. Dependent Variable Y is set to the proportion of daily confirmed cases of whites, blacks, and Latinos, respectively.

Table 6. Linear regression results (White)

Date	coefficient	p-value
Mon	-0.029	0.676
Tue	0.001	0.99
Wed	-0.031	0.661
Thu	0.011	0.877
Fri	0.055	0.439
Sat	-0.010	0.884
Intercept	0.189	<0.001

Table 6, 7, and 8 list the coefficients and p-values of the three races after linear regression. It can be seen from the numerical size of p-values that the impact from Monday to Saturday are not significant, but the impact from Sunday is significantly different from zero. In the linear regression of Latinos, the intercept is 0.299, and the coefficients on Thursday, Friday and Saturday are negative, so the number of cases will be less than (but not in significant way) Sunday, that is, the intercept. The intercept of linear regression of blacks is equal to 0.284, and the cases on Monday, Tuesday, and Friday will be less than Sunday, and Whites had fewer cases on Monday, Wednesday, Thursday, and Saturday than on Sunday.

Table 7. Linear regression results (Black)

Date	coefficient	p-value
Mon	-0.015	0.821
Tue	-0.033	0.612
Wed	0.006	0.928
Thu	0.050	0.446
Fri	-0.034	0.604
Sat	0.059	0.382
Intercept	0.284	<0.001

Table 8. Linear regression results (Latino)

Date	coefficient	p-value
Mon	0.028	0.635
Tue	0.015	0.794
Wed	0.031	0.604
Thu	-0.019	0.746

Fri	-0.004	0.948
Sat	-0.031	0.612
Intercept	0.299	<0.001

4 Conclusion

Through the analysis of the data set [1] selected in this paper, the government and companies can reduce the number of confirmed cases by targeted epidemic prevention. From the results of the daily number of confirmed cases, Latinos and blacks account for a higher proportion of all races, and people between the ages of 18-49 years old also make up most of the confirmed cases. Most of the 18 to 49-year-old groups are already working people and a small number of college students, so people who need to deal with customers and colleagues in companies should pay more attention to disinfection and wearing masks. And among those who work, the higher rates of Latinos and blacks may be due to the type of work they are doing. According to an article by the U.S. Bureau of Labor Statistics, it is Black and Latino who do more service jobs than other races [4]. The service industry requires more exposure to different people than other types of work, so they have higher risk. From the Hypothesis test, it can be concluded that infants and teenagers under the age of 18 are less likely to be diagnosed with the COVID-19. There is not much difference between the 18- to 49-year-olds, and the number of confirmed diagnoses will be significantly different after the age of 59 years old. We also propose a generic linear regression modeling framework with dummy variables to analyze the data. From the results, there are generally more people (although not in a significant way) who are confirmed on Mondays to Fridays. On the way to work, people tend to go through popular places, such as subways and buses. Compared with weekends, when people can choose travel location and time at will, people do not have much flexibility during weekdays. Therefore, people should be more cautious in daily work and commuting.

References

1. A. Jain, “Understanding regression using COVID-19 dataset---detailed analysis” Towards Data Science, 2020.
2. B. Peach, S. Cooney, S. Richards, “Prominent cognitive impairment sequelae in adult survivors of acute respiratory distress syndrome” Rehabilitation Nursing Journal. *47*(2), pp.72-81, 2022.
3. E. Ortiz-Ospina, M. Roser, “Global health” Our World in Data. 2016.
4. H. L. Solis, J. M. Galvin, “Labor force characteristics by race and ethnicity” US Department of Labor and US Bureau of Labor Statistics. pp.1036, 2012.
5. I. Rahimi, F. Chen, H. A. Gandomi. “A review on COVID-19 forecasting models” Neural Computing and Applications. pp.1-11, 2021.
6. J. Xiong, O. Lipsitz, F. Nasri, L. Lui, H. Gill, L. Phan, R. S. McIntyre, “Impact of COVID-19 pandemic on mental health in the general population: A systematic review” Journal of Affective Disorders. pp. 55-64, 2020.

7. M. C. Castro, B. Singer, “Prioritizing COVID-19 vaccination by age” Proceedings of the National Academy of Sciences. 118(15), 2021.
8. M. Monod, A. Blenkinsop, X. Xi, D. Hebert, S. Bershan, “Age groups that sustain resurging COVID-19 epidemics in the United States” Science, 371(6536), 2021.
9. S. D. Romano, A. J. Blackstock, E. V. Taylor, S. E. B. Felix, S. Adjei, C. M. Singleton, T. K. Boehmer, “Trends in racial and ethnic disparities in COVID-19 hospitalizations, by region—United States” Morbidity and Mortality Weekly Report. *70*(15), pp.560, 2021.
10. S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara, M. R. Edinburgh, “Analysis and prediction of COVID-19 using regression models and time series forecasting” International Conference on Cloud Computing. pp. 989-995, 2021.
11. S. Rath, A. Tripathy, A. R. Tripathy, “Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model” Diabetes and Metabolic Syndrome: Clinical Research and Reviews. *14*(5), pp.1467-1474, 2020.