Linear Regression

Matthew Taylor, Xiantong Zhao, Nick Tufte

November 3, 2022

1 Introduction

Although covid-19 is not as before, this major health security event will continue to affect our life for a while if not for the rest of our lifetime. It is worthwhile to do some research on it. From the data perspective, for most people, who do not have time to analyze some raw data, the growth curve is the chart for most people to catch up on the updates of covid-19. As for the people who know some learning machine methods and are interested in data analysis, covid-19 becomes the appropriate target for us to analyze. Due to its impact and recency, there is a lot of related data for us to obtain. Besides that, there are many related topics that are also worthy to analyze, for example, we could analyze the vaccine penetration and hospital bed rates. These two data are highly related to covid-19, and people are likely wanting to see them along with the covid-19 grow rate. What we can do to analyze these data is visualizing them; Make a chart or map to make those data easier to understand. We would like to implement regression techniques to explore the best model to use and the significance of said model. We will start by using simple linear regression to analyze some of the covid-19 data and adjust the modeling accordingly.

2 Data Description

There is so much data out there on covid-19 worldwide. However, for our purposes, we decided to stay within the United States. Then again, we restricted the area simply to Chicago. Our data set comes from here. This dataset contains data from several months in Chicago giving information such as the date, the number of tests conducted, the number of positive cases, and many of those numbers broken down by age, gender, and race/ethnicity. The data collected spans from March 1st of 2020 and goes until May 31st, 2020. The data contains 90 observations in that span. The most tests in a day ended up being 5,971 while the most cases were 1,285. These, slightly surprisingly, occurred in different days. The maximum cases number occurred at the end of April, while the maximum tests occurred towards the end of May. The minimum value for tests and cases were 1 and 0 respectively, which was of course on the earliest date in the dataset. The mean for tests was ~2,165.5 while the median was 1645. The disparity between the two alludes to a large spike in number of tests taken which can be accounted for by the increasing accessibility of testing. A similar mean/median disparity can be seen in the cases, with values 500 for the mean and 460 for the median. Another interesting category of the data to look at is the race of individuals tested. This could inform us in things such as races more or less likely to be tested. The mean tests for people categorized as black was 375.16, for white was 284.92, and for Latinx was 388.25. There was, on average, more Latinx individuals that were tested even though they are a minority in Chicago relative to black and white. This could have to do to the testing location or the makeup of unknown race tests which has a significant mean of 996.39. For the linear regression example, we will be using the test numbers and the cases numbers from this data set.

3 Theory

The theory of linear regression is fairly simple. Linear regression is an algorithm that is used to predict the future by providing a linear relationship between an independent variable and a dependent variable. The formula for linear regression looks like this:

$$y^0(x) = p_0 + p_1x_1 + p_2x_2 + ... + p_nx_n$$

where p is the coefficients of the inputs x. p also has the same dimensionality as x. If we have just one input and output, it is called simple linear regression, but if we have more than one input, it would be called multiple linear regression. With either of these types of linear regression, the goal is to try to find the line that would minimize the sum of squared errors:

$$sse = \sum_{n}^{i}(y^i - y^{0i})$$

where y is the actual y-yalue and y^0 is the predicted y-value. We remember that y^0 is equal to the above equation for linear regression.

From here, we need to find the coefficients for the best fit line. This can be done a number of ways, but one way is the least squares method. A formula for calculating the regression line is:

$$\theta_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

$$\theta_0 = \bar{y} - \theta_1\bar{x}$$

We can then use gradient decent as an optimization algorithm. This algorithm determines a minimal cost function by generating various values for θ_1 and θ_0. The starting values are generated at random. We then plug those values into the cost function to calculate the cost. Then the derivative of the cost function will then give us the slop letting us know which direction the values of θ_1 and θ_0 need to be changed.

$$\frac{\partial}{\partial\theta_j}MSE(\theta) = \frac{2}{m_i}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x_j^{(i)}$$

In order to calculate them all at the same time, we can use the gradient vector containing all the partial derivatives for each of the parameters.

$$\nabla_\theta MSE(\theta) = \begin{bmatrix} \frac{\partial}{\partial\theta_0}MSE(\theta) \\ \frac{\partial}{\partial\theta_1}MSE(\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_n}MSE(\theta) \end{bmatrix} = \frac{2}{m}X^T(X\theta - y)$$

4 Python Implementation of Linear Regression

First, we will import all the necessary libraries.

```
import pandas as pd import numpy as np from sklearn.linear_model import

LinearRegression from sklearn.preprocessing import PolynomialFeatures from

sklearn.model_selection import train_test_split from sklearn.metrics import

mean_squared_error from sklearn.metrics import r2_score import seaborn as

sns import matplotlib.pyplot as plt
```

Then we need to import our data set that we talked about earlier.

```
import requests url = ’https://raw.githubusercontent.com/Ayushijain09/Regression-on-COVID-dataset/

                                                    master/COVID-19_Daily_Testing.csv’

res = requests.get(url, allow_redirects=True) with open(’COVID-

19_Daily_Testing.csv’,’wb’) as file: file.write(res.content) sales_team =

pd.read_csv(’COVID-19_Daily_Testing.csv’)


data = pd.read_csv("COVID-19_Daily_Testing.csv") data.head()
```

Next, we need to clean up the data a little bit.

```
data[’Cases’] = data[’Cases’].str.replace(’,’, ’’) data[’Tests’] =

data[’Tests’].str.replace(’,’, ’’) data[’Cases’] = pd.to_numeric(data[’Cases’])

data[’Tests’] = pd.to_numeric(data[’Tests’])
```

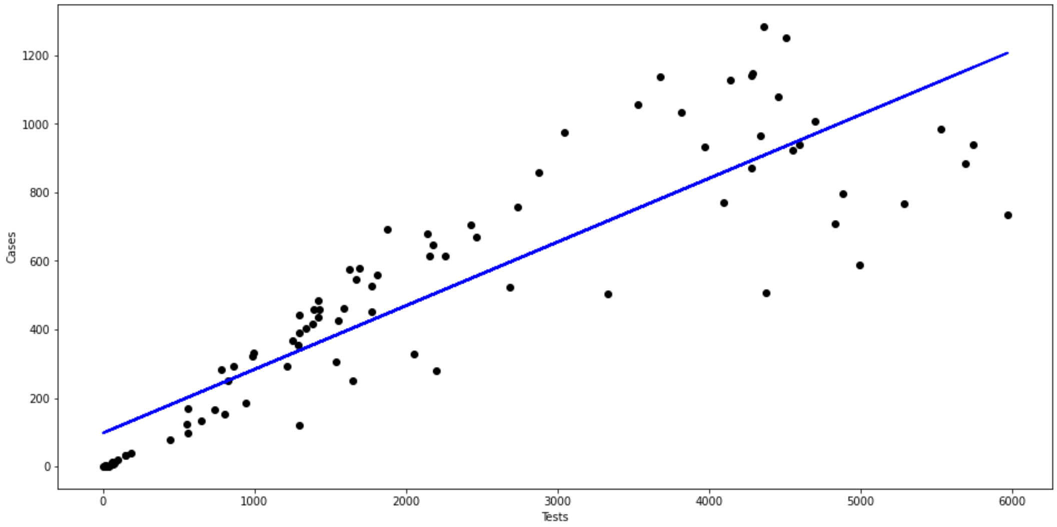
From here, we are going to scale the data set so that the visual representation is more accurate.

```
X = data[’Tests’].values.reshape(-1,1) y =

data[’Cases’].values.reshape(-1,1)
```

Next, we apply the linear regression.

```
reg = LinearRegression() reg.fit(X, y) predictions = reg.predict(X) print("The linear model is: Y = {:.5} + {:.5}X".format(reg.intercept_[0],

reg.coef_[0]

                                                [0]))

plt.figure(figsize=(16, 8)) plt.scatter( X, y,

c='black'

) plt.plot( X, predictions,

c='blue', linewidth=2

) plt.xlabel("Tests")
plt.ylabel("Cases") plt.show()
```

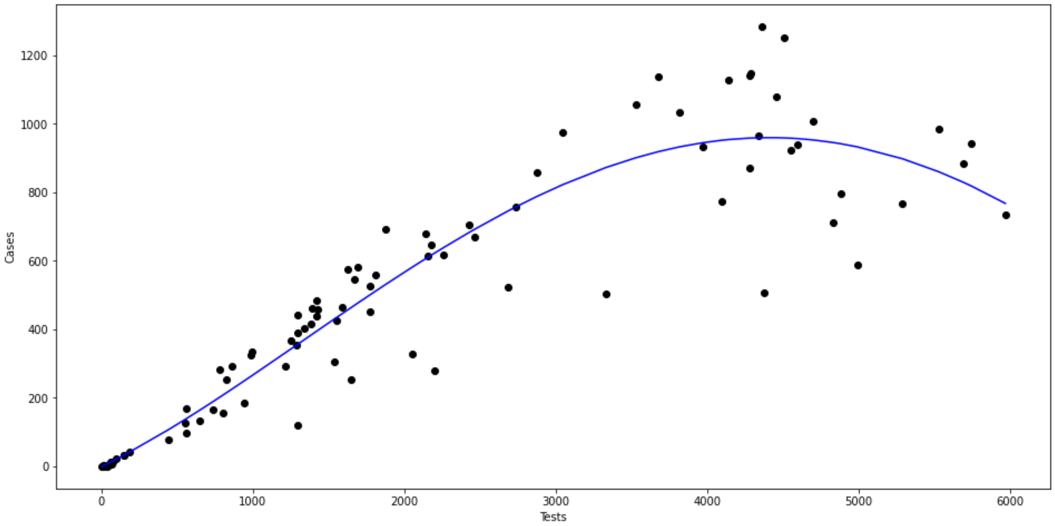
The linear model is: $Y = 97.777 + 0.18572X$



The Root Mean Square Error for Linear Regression = 171.8

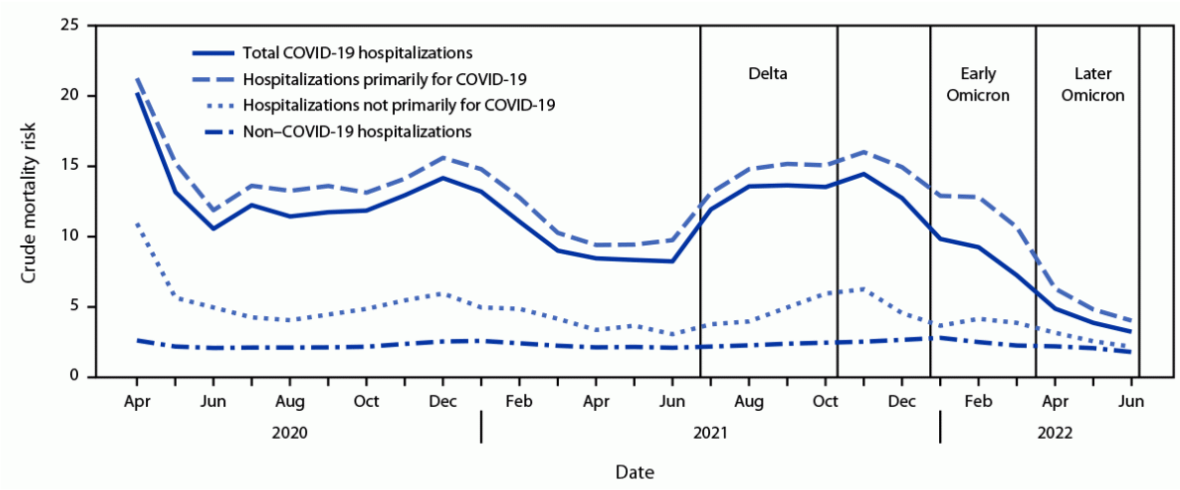
There are many outliers in this case, so the regression isn't fitting the data very well. For some cases like these, it would be better to use a slightly different approach and use a polynomial regression rather than a linear regression.

```
poly = PolynomialFeatures(degree =4) X_poly =

poly.fit_transform(X)


poly.fit(X_poly, y) lin2 = LinearRegression() lin2.fit(X_poly, y)

pred = lin2.predict(X_poly) new_X, new_y = zip(*sorted(zip(X,

pred))) plt.figure(figsize=(16, 8)) plt.scatter( X, y, c='black'

) plt.plot( new_X, new_y,

c='blue'

) plt.xlabel("Tests")
plt.ylabel("Cases") plt.show()
```



5. Model Prediction.

With applying the linear regression model on the data, we get some results on the data. First of all, according to below image(Centers for Disease Control and Prevention, 2022), for the hospitalized perspective, it had some ups and downs before 2021 November, after that, the situation becomes much better and total covid-19 hospitalizations keep decreasing.



With the hospitalization data analysis, we could infer about the trends of the total infected people since those two data are positive correlation. And the results are same as our prediction. We also have the data from CDC and the chart below(Centers for Disease Control and Prevention, 2022). With the time and corresponding data, those two charts fit most part. However, I noticed that when it comes to November 2021, hospitalization reaches its peak, but weekly increased case reaches its peak around January 2022, that does not fit out expected, according to our expectation, the hospitalization should reach its peak around January 2022 too, but there is a significant difference between them. So, we assume there are some factors we did not consider that will affect the hospitalizations and cases. There could be some, for example, the rate of severe condition caused by COVID-19 is decreasing, some not as many patients need to go to hospital to receive remedy.

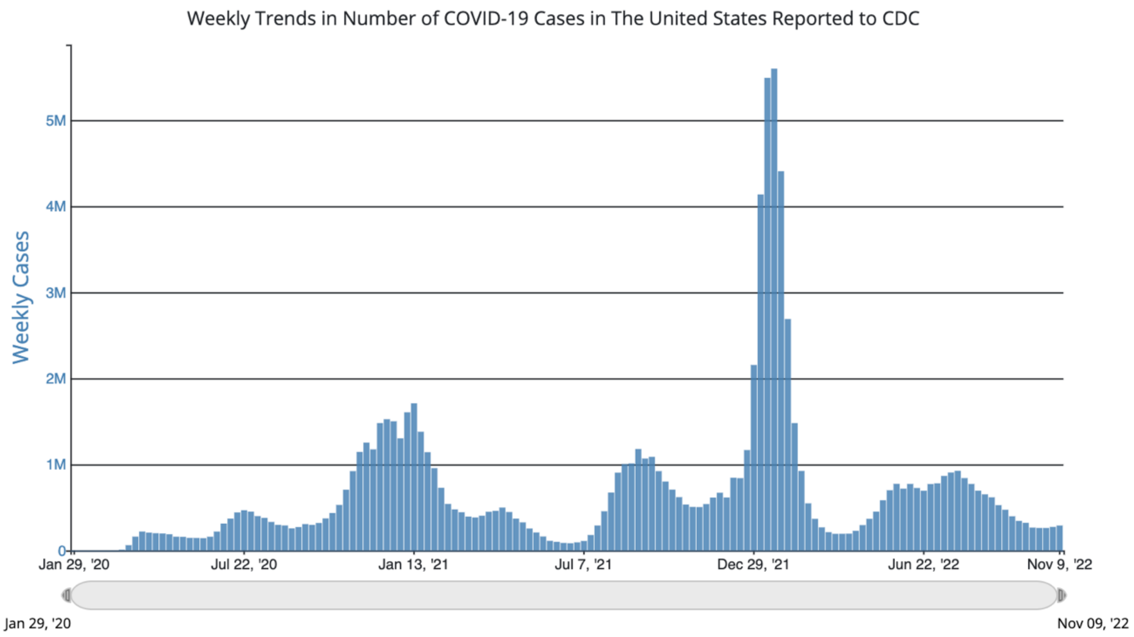


Figure 1(Centers for Disease Control and Prevention, 2022)

6. Discussion

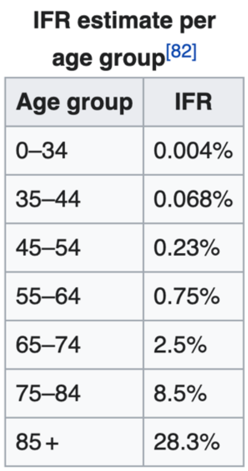
Using a simple linear regression model, we saw the equation $\hat{Y} = 97.78 + 0.19x$. If we were to interpret this, we would say for every additional test taken, there was, on average, a 0.19 increase in covid cases. While this does not see outlandish, if we view some other characteristics of the linear regression, we can see some issues. The first is that the intercept is 97.78, meaning when there are no tests, the model assumes there are already 97.78 cases. This is not a realistic value there would be no knowledge of cases if there were no tests. The second place where simple linear regression falls apart is the variance increase as we increase the number of tests. The fit of the data points on to the estimated line is increasingly worse. This is poor for the significance of the model as it becomes less reliable the more tests that are done. These two obvious issues with the simple linear regression led us to the need for further exploration which is where the polynomial fit comes in.

Since the data points begin quite linear, but change behavior as the number of tests increase, we need to use a polynomial model. As we can see in the plot of the fitted polynomial and the data points, the model appears to be significantly better. This leads us to believe this model would be appropriate. As far as the contextual, "why?", we can look at several possibilities. The first of which is the increased convenience of testing. When there were less cases, people were less aware of covid and the ways to get tested for covid. This meant that people were only getting tested if they had strong enough evidence to seek out testing. For example, loss of taste/smell and the combination of other symptoms that would convince someone they had covid. If someone did not have these telling symptoms, they were far less likely to be tested when less people were being tested. Then, as the amount of testing increased, so did the rapidness of the spread of covid-19 which allowed the slope, $\frac{\text{cases}}{\text{tests}}$, to maintain steady. Once testing became widely available and people were informed of the ways to get tested much more people began testing with less obvious symptoms. This was likely the reason for the decrease in the slope of the model. Eventually the data points and corresponding polynomial graph begin to start decreasing. This is likely due to the decrease in infectiousness (as a result of natural immunity), but the persistence of the commonality of testing.

While the new polynomial fit can map out, and provide clarity on the topics discussed above, it also has limitations. The most concerning of which is the nonconstant variation seen with larger test numbers. While it is remedied a bit with the new fit there is still some clear increase in variance. This is likely due to the fact that there are other variables at play in which the model is not accounting for. The first of which is new variants. When new covid-19 variants pop up, some more contagious than others, the amount of infected can experience a surge which the number of tests cannot keep up with. This is a likely explanation of the data points that show up much higher than the predict value according to the model. Another explanation for the variance is the variable of other diseases. For example, flu season may take over, which increases the number of sick people and, in turn, increases the number of tests. Since this new influx of tests are a disproportionate amount of people with the flu and not covid-19, there could be a much smaller number of cases compared to the predicted value of the model. It is important we understand both how the model can help us and how it may not tell the full story.

According to the results and data, we can make a conclusion that the covid situation is much better now and the death rate is not high for age below 54, according to below chart (Wikimedia Foundation, 2022), but the variants are still updating, so we still need to pay some attention to it, because with the variants, the death rate and spread speed could get improve. So, the relative analysis on this is still important, we need to keep track on the new variants.

An interesting point worth to mention: the highest peak of weekly increase in cases did not appear at the beginning state of covid but instead appeared at the middle/end state, there could be some interesting data to be analyzed, but with the data we collected this time, we could not find the reason for that.



IFR estimate per age group[82]

Age group	IFR
0–34	0.004%
35–44	0.068%
45–54	0.23%
55–64	0.75%
65–74	2.5%
75–84	8.5%
85 +	28.3%

In future, there are still many works could be done to make this research more all-round, the data we collect and analysis we done are not deep enough to find all the connection between increasing case and hospitalization (as we mentioned in last part).

7. References

Jane, A. (n.d.). Understanding regression using COVID-19 dataset — detailed analysis. Retrieved November 3, 2022, from https://towardsdatascience.com/understandingregression-using-covid-19-dataset-detailed-analysis-be7e319e3a50

Centers for Disease Control and Prevention. (2022, September 15). Mortality risk among patients hospitalized primarily for covid-19 during the Omicron and delta variant pandemic periods - United States, April 2020–June 2022. Centers for Disease Control and Prevention. Retrieved November 12, 2022, from https://www.cdc.gov/mmwr/volumes/71/wr/mm7137a4.htm

Centers for Disease Control and Prevention. (2022). CDC Covid Data tracker. Centers for Disease Control and Prevention. Retrieved November 12, 2022, from https://covid.cdc.gov/covid-data-tracker/#trends_weeklycases_select_00

Wikimedia Foundation. (2022, November 12). Covid-19 pandemic. Wikipedia. Retrieved November 12, 2022, from https://en.wikipedia.org/wiki/COVID-19_pandemic#Deaths