Project Assignment 6

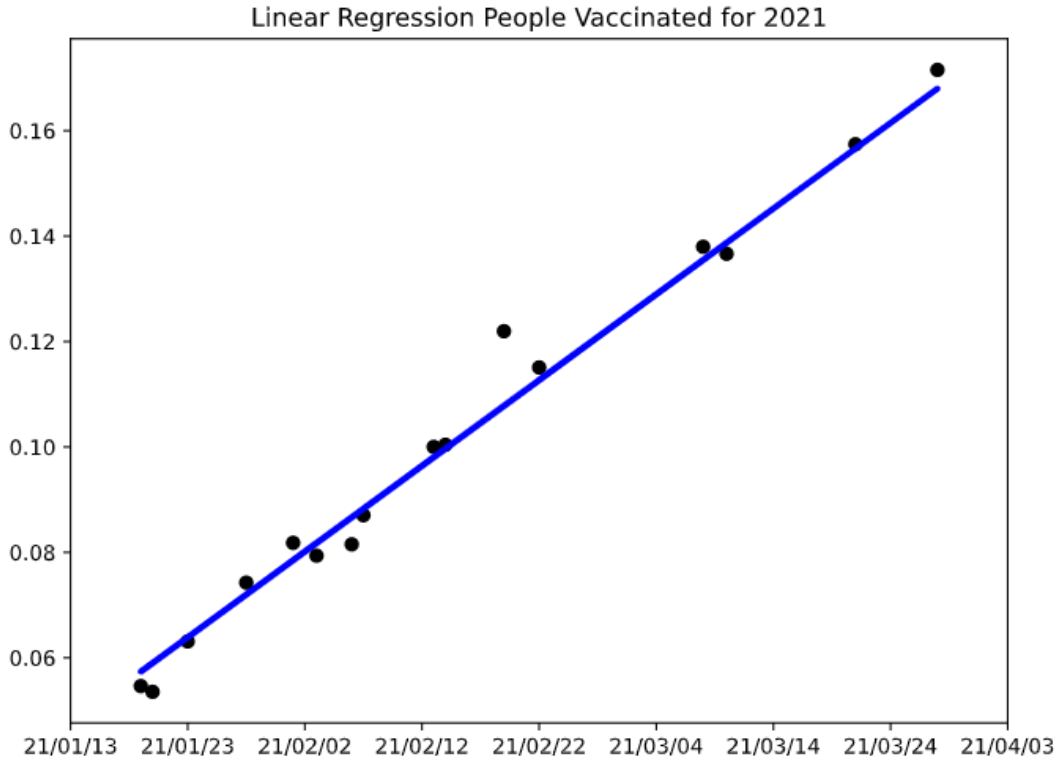
by Alisiya Balayan, Bishoy Abdelmalik, Arian Dehghani, Ariel Kohanim, Sergio Ramirez, Natalie Weingart

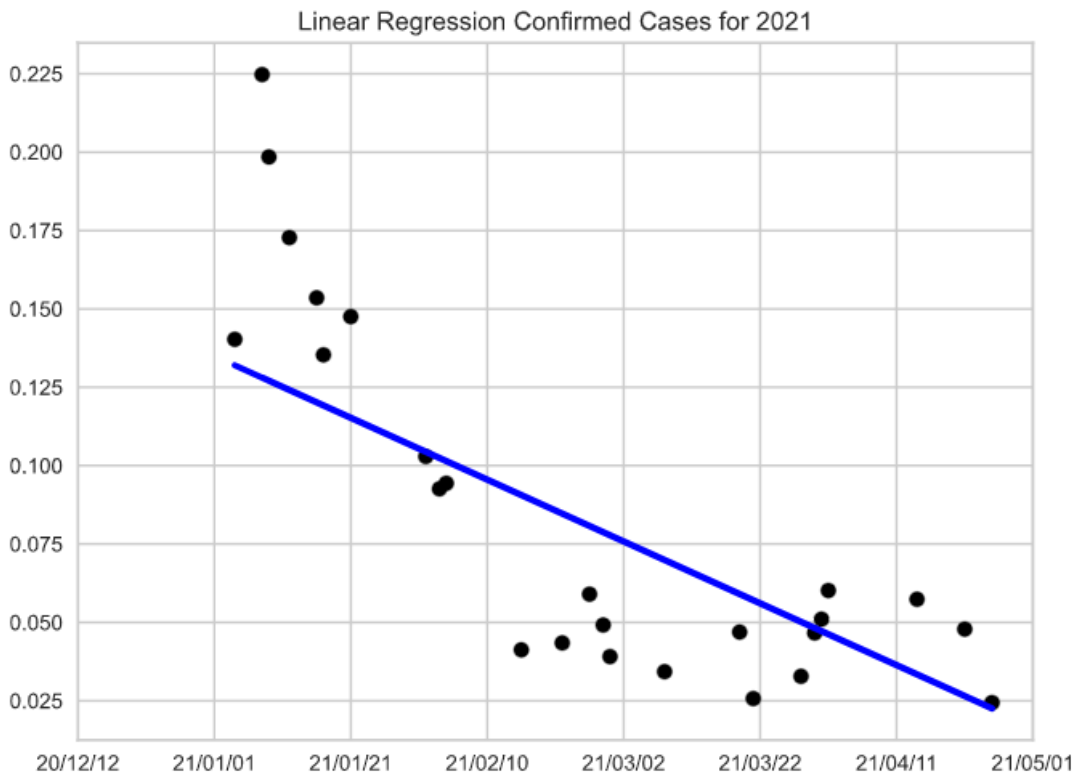
1. Modeling technique

a. Linear Regression

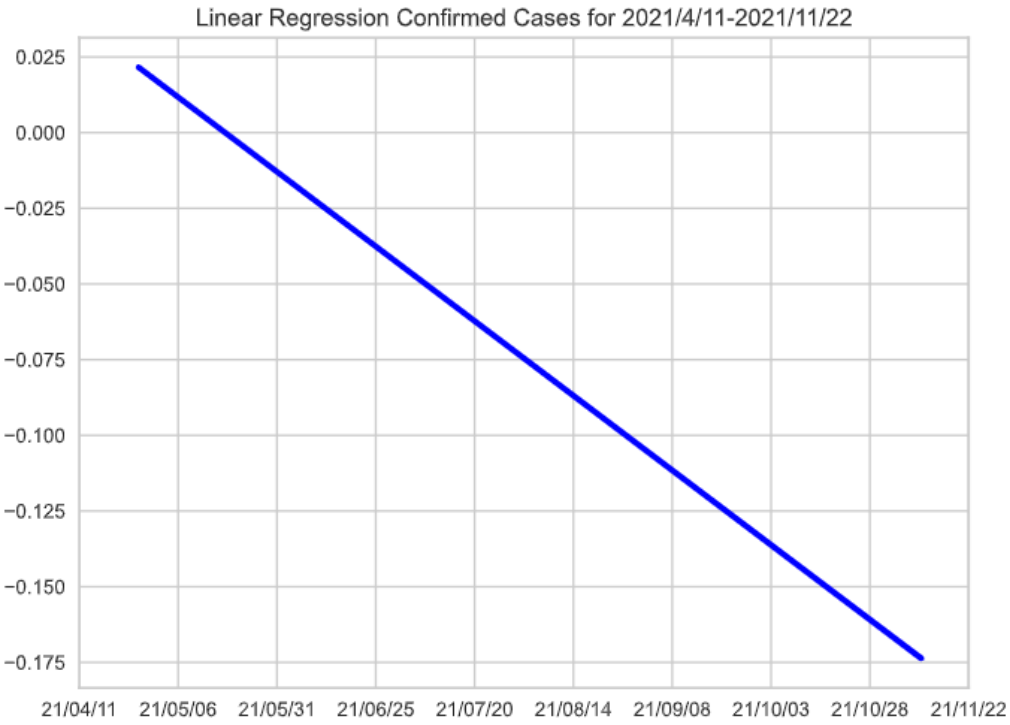
We used linear regression as a modeling technique to analyze the relationship between people getting fully vaccinated and confirmed COVID-19 cases. Linear regression is a linear approach to modelling the relationship between a scalar response and independent and dependent variables. It is used to predict the value of a certain variable depending on the value of the other variable (hence, independent and dependent). Our independent variable is time, and dependent variables are people fully vaccinated, and confirmed COVID-19 cases. After independent and dependent variables have been identified, we attempted to make a model that fits linear equation relationship between the data.

First, we made linear regression models to see the relationship between time and vaccinations, and time and confirmed cases.





b. <u>Linear Regression to Predict</u>
The following linear regression shows us that by November of 2022, confirmed COVID-19 cases will significantly drop and vaccination rates will stabilize which means that we can safely come back in person. For example, CSUN plans to safely open more in-person classes by Spring 2022 which supports our results.



2. Evaluate modeling accuracy and validity

a. Mean Squared Error

MSE tells us how close a regression line is to a set of points. It takes the distance from the points to the regression line and squaring them. The distances are the errors and the squaring is necessary to remove any negative signs. The fact that MSE is almost always strictly positive and not zero is because of the randomness or because it doesn't account for information that can produce more accurate estimates. As you can see from the results, we got mse close to 0 which means that the model is almost perfect.

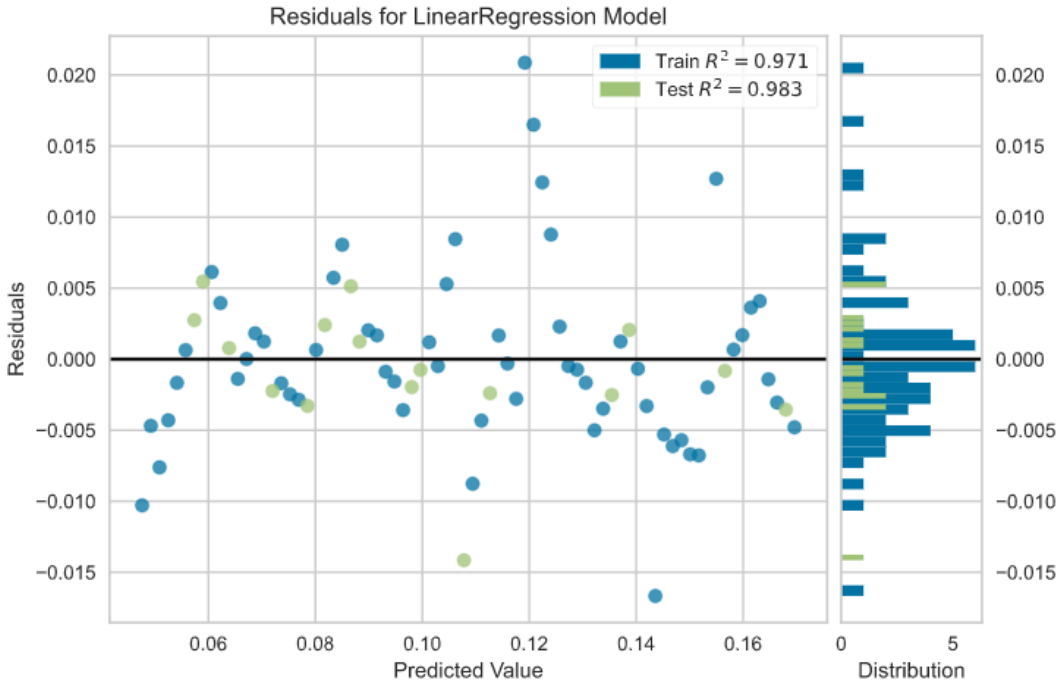
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

```
mse = metrics.mean_squared_error(test_y, pred_y)
mse = f"{mse:.9f}"
print("mean squared error:"+mse)
```

```
mean squared error:0.001214634
```

b. Residual Plot for Linear Regression Model

A residual plot is a plot that shows residuals (errors) on the y-axis and the independent variable on the x-axis. If the points on the residual plot are randomly dispersed around the horizontal x-axis which means that the linear regression model is appropriate for the data, otherwise the nonlinear model will fit better. For example, the screenshot below shows residual plot and scatters are distributed along the horizontal axis which signifies that the linear model fits our dataset.



c. Correlation Coefficients

```
from scipy import stats
print("Pearson's correlation coefficient:"+str(stats.pearsonr(covid_19_data["Confirmed_diff"],daily_vaccines["daily_vaccinations"] )
[0]))
```

```
Pearson's correlation coefficient:-0.6829676521183081
```

Pearson's correlation coefficient is -0.68 which means that confirmed cases and vaccinations go different directions (inverse correlation).

d. Outlier Analysis

We performed outlier analysis to see if there are any outliers in the data before making models. Once we saw where outliers were in our datasets, we performed IQR where we removed outliers from Q1 or Q3 depending on the situation.

```
#find Q1, Q3, and interquartile range for each column
Q1 =  cleaned["Confirmed"].quantile(q=.25)
Q3 =  cleaned["Confirmed"].quantile(q=.75)
IQR = cleaned["Confirmed"].apply(iqr)
print(Q1)
print(Q3)
#only keep rows in dataframe that have values within 1.5*IQR of Q1 and Q3
cleaned = cleaned[~((cleaned["Confirmed"] > (Q3+1.5*IQR)))]
```

3. Parameters and their chosen values, along with the rationale for the choice of parameter settings

a. Observation dates

We selected dates because we need to evaluate the COVID-19 spread and vaccinations over a period of time in order to accurately predict when we will reach herd immunity and safely open up the country.

```
us_daily_vaccines['Day'] = pd.to_datetime(us_daily_vaccines.Day)

groupedByMonth=us_daily_vaccines.groupby(pd.Grouper(key='Day', freq='1M')).sum().reset_index()
groupedByMonth.head()
```

b. Number of people fully vaccinated

We selected the number of people fully vaccinated to show us the progress of vaccines as time progressed. This was used in regression as well as in calculations to calculate results.

```
usa_values=usa_values[['date','total_vaccinations','people_vaccinated','people_fully_vaccinated']]
usa_values.head()
```

c. Number of confirmed COVID-19 cases

We selected a total number of confirmed COVID-19 cases because we needed to measure the correlation between vaccines and confirmed cases to predict when the total confirmed cases will drop.

d. Location

We narrowed down our models to only the U.S. for easier data cleaning and preprocessing as certain countries didn't have enough data or it was inconsistent and we couldn't get updates.

```
data = pd.read_csv("datasets/COVID-19 World Vaccination Progress/country_vaccinations.csv")
usa_values=data.loc[data['iso_code']=='USA']
```

4. Model 1: Linear Regression

a. We ran linear regression in jupyter notebook, notebook file attached.

```
Model took 0.14335846900939941 seconds
```

5. Overview of results

Results from the linear regression model show that by November 2022, we will not have confirmed COVID-19 cases given that the rate of decline of new cases remains the same. In addition we can also find out that we will reach 70% of the population vaccinated in 13.7 months.

6. Interpret models according to domain knowledge

a. Data mining success criteria

i. Our linear regression model aligns with the business goal and confirms our hypothesis. We wanted to see when we will reach 70% of immunity and our model predicted 13.7 months from today (May 10th). Data sources we selected for this project match the goals we have set as in our datasets containing valuable information about COVID-19 cases and vaccinations. The datasets had complete data, all missing values were filled by the nearest neighbor, null and NaN values were removed.

b. Test result

i. The result of the model predicted that in 13.7 months, the U.S. population will reach 70% immunity and according to the experts that is an acceptable percentage to get back safely in person. Also, CSUN's plans to open campus next year align with the results of our model as it confirms that we also got the same date.

7. Rank the models (accuracy, performance, and generality of the model)

a. Linear Regression
Performance: fast (0.14 sec)
Accuracy: mse and residuals showed our model is almost perfect.
Generality: our model can be applied to other diseases as well (predict flu outbreak, etc.)

8. Assess the degree to which the model meets the business objectives (determine any deficiencies)

The model addresses the business objectives as it is able to successfully determine the rate of people getting vaccinated in the feature and determine how long until we reach certain levels of vaccinations. The deficiencies that the model might have is that it relies on the assumption that no new variants of the virus are discovered.