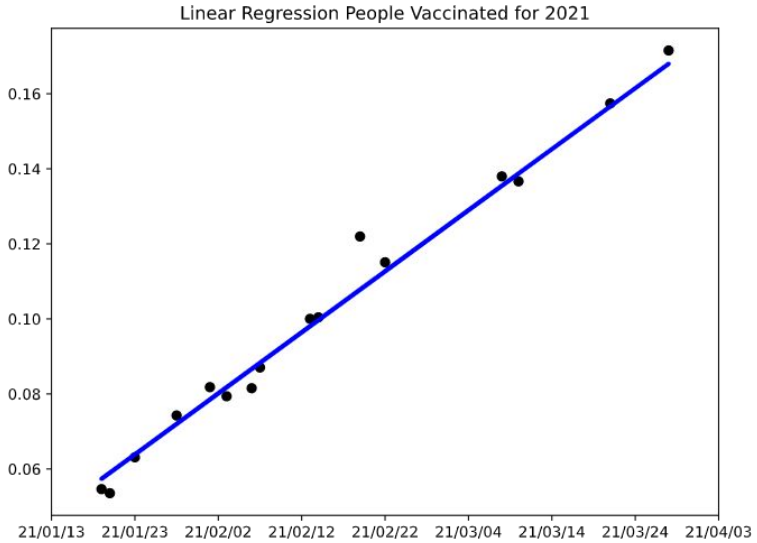
Assignment 6

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

TABLE OF CONTENTS

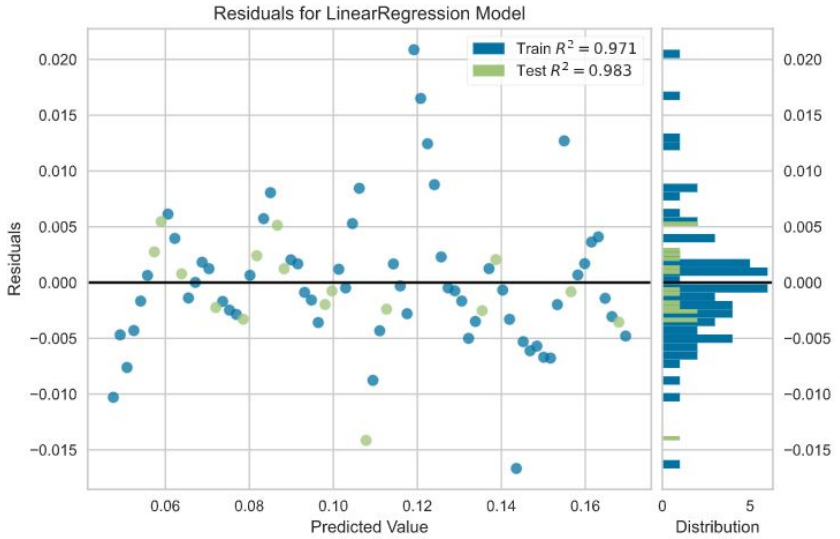
Modeling Technique

- Linear Regression
  - We used linear regression as a modeling technique to analyze the relationship between people getting fully vaccinated and confirmed COVID-19 cases.
  - we made linear regression models to see the relationship between time and vaccinations, and time and confirmed cases.



Model Accuracy and Validity

- Mean Squared Error
  - 0.0012
- Residual Plot
- Pearson's Correlation
  - -0.68
- Outlier Analysis
  - IQR test

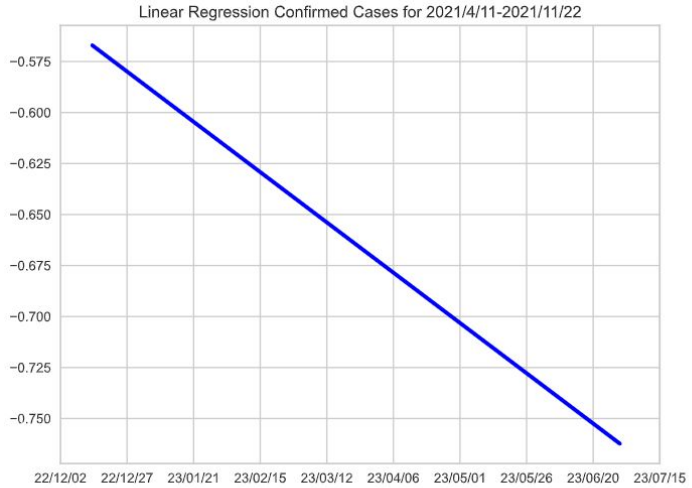


Chosen Parameters

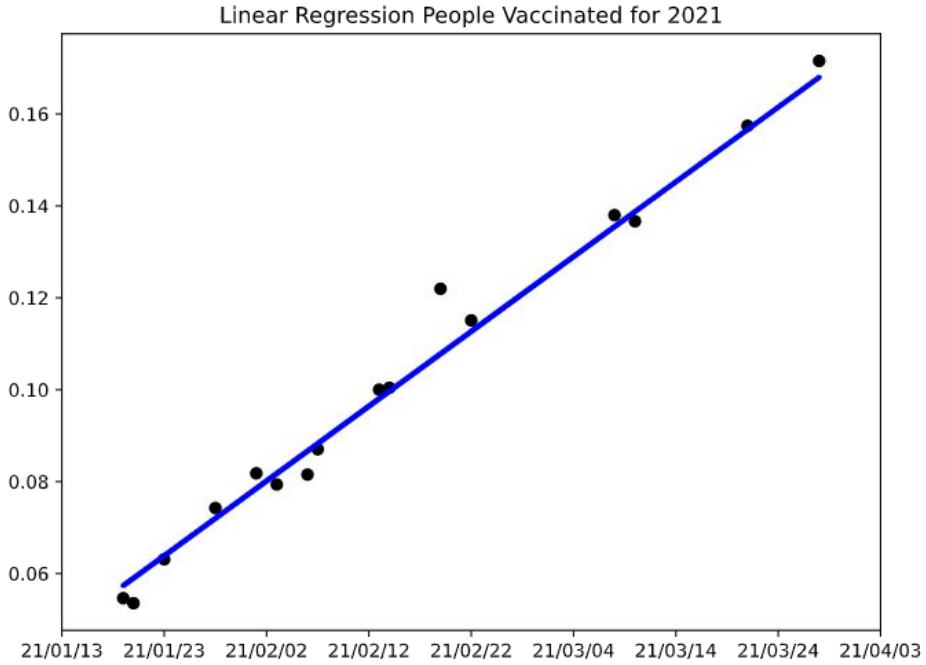
- Observation dates
- Number of people fully vaccinated
- Number of confirmed COVID-19 cases
- Location
  - U.S. only

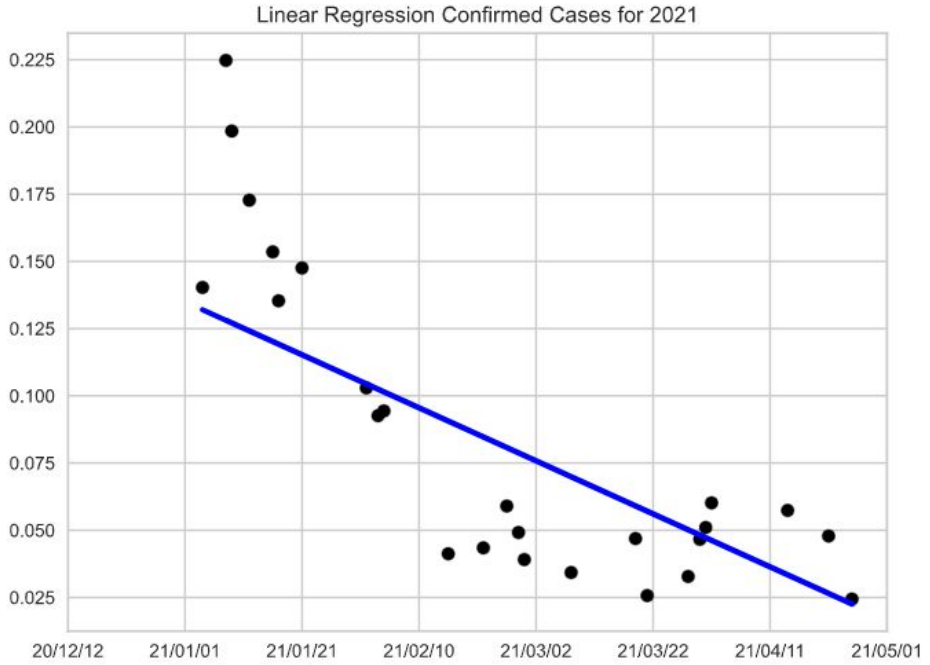
Linear Regression

```
import time
start_time = time.time()

max_x=test_x.max()
test_x=np.array([[x] for x in range(int(max_x)+1,int(max_x)+200)])

pred_y = lr.predict(test_x)


# Plot outputs
plt.figure(figsize=(8, 6))

plt.plot(test_x, pred_y, color='blue', linewidth=3)

xticks=plt.xticks()[0]
xticks_str=[]
first=covid_19_data['ObservationDate'].min()
for tick in xticks:
    xticks_str.append((first+ timedelta(days=int(tick))).strftime("%y/%m/%d"))

plt.xticks(xticks,xticks_str)
plt.title("Linear Regression Confirmed Cases for 2021/4/11-2021/11/22")

plt.show()


print("Model took %s seconds " % (time.time() - start_time))
```



Overview of Results





Interpretation of Domain Knowledge

- Data Mining Success Criteria:
  - Aligns with business model
  - Confirms the goal set
  - Valuable datasets
- Test Results:
  - 70% in 13.7 months
  - Most schools and businesses plan to open in about a year.

Model Ranking

- Performance:
  - 0.14 execution
- Accuracy:
  - Mse, residual plot and Pearson’s correlation
- Generality:
  - Model can be applied to many other diseases

Business Objectives Fit

- 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.