607 Final project report

Zeren li

12/06/2024

Abstract

This Project essentially investigate the questions like Was vaccination rate effectively prevent the pandemic? Are there other factors? It begins with the data analysis and findings of the project which investigates the effectiveness of vaccination rates in preventing the pandemic. It also explores other influencing factors, such as the emergence of variants

Introduction

What are the predominant factor that affect the confirmed cases in the Covid-19 dataset? We firstly assume that vaccination play a part it predicting the confirmed cases and then included other factors like variant of the virus. We firstly investigate the potential factor before diving into the individual states and performing regression.

Data Loading, Cleaning, and preparing

For United state country data
We set the confirmed, deaths, people_vaccinated, people_fully_vaccinated to be 0 if they are null.
Then, extract the date from 2020-03-01 to 2022-03-01.
Add a new column of vaccination rate.
Add a new column called variant to account for the number of variants of the virus that appeared.

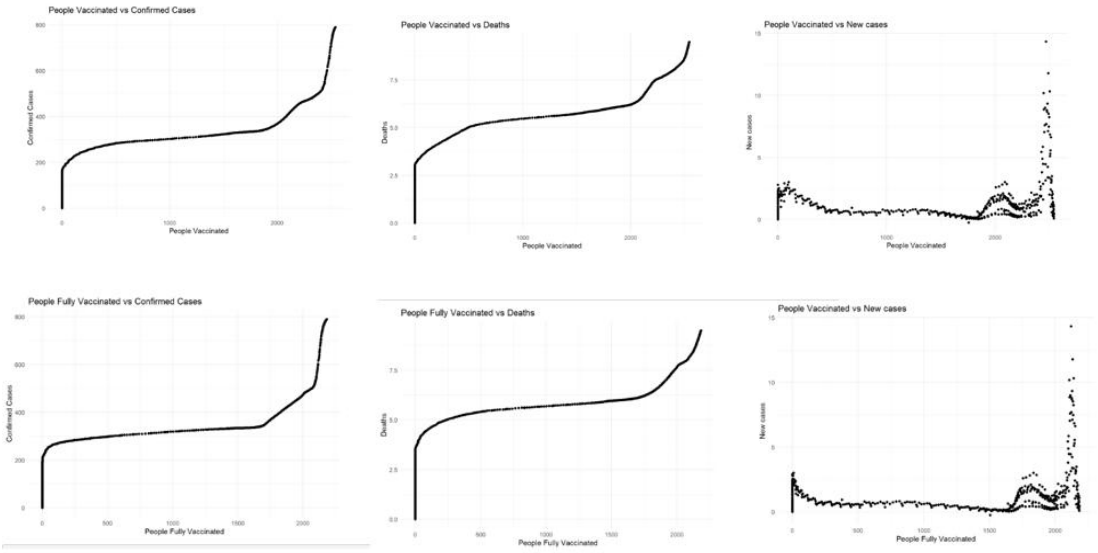
For the state data,
We set the confirmed, deaths, people_vaccinated, people_fully_vaccinated to be 0 if they are null.
Then, extract the date from 2020-03-01 to 2022-03-01
Add a new column of vaccination rate.
Add a new column called variant to account for the number of variants of the virus that appeared.

Refer to the code C1-3 in tables & code sections to see the code.

Analysis Methods and steps

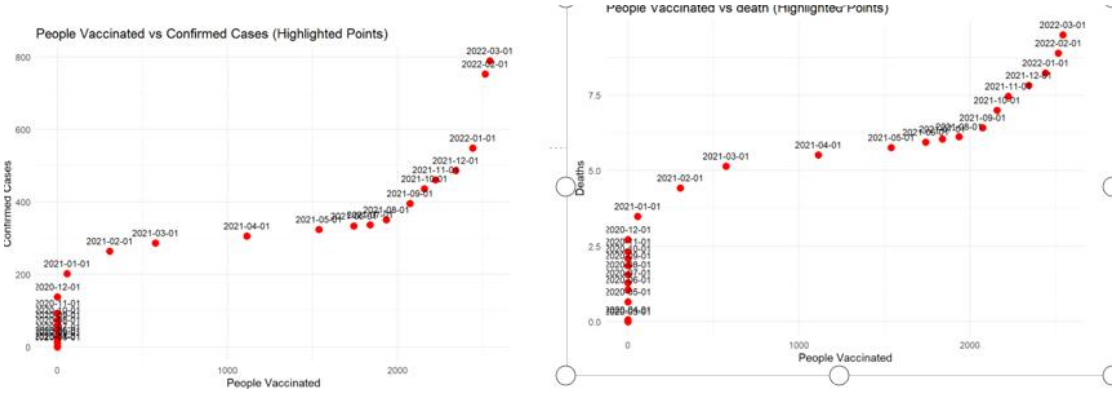
Scatterplots

Scatterplots were utilized to examine relationships between vaccination rates and outcomes such as confirmed cases, deaths, and new cases. Both 'people fully vaccinated' and 'people vaccinated' were considered as dependent variables.



we used both people_fully_vaccinated, and people_vaccinated as the dependent variable. The graph shows that no matter which we choose as the dependent variable, the confirmed cases and death graph shows a “ladder”. That is the cases will quickly go up first, and then stay steady at some point, and then go up again at the end.

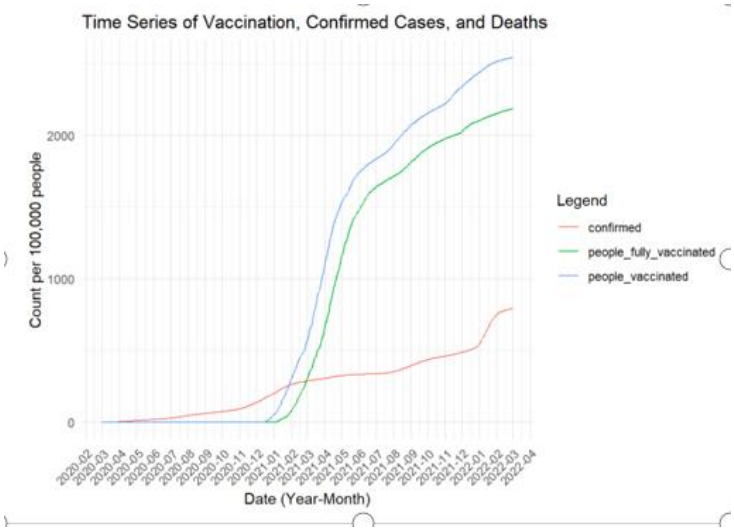
To investigate this further, we can highlight those points for start of each month.



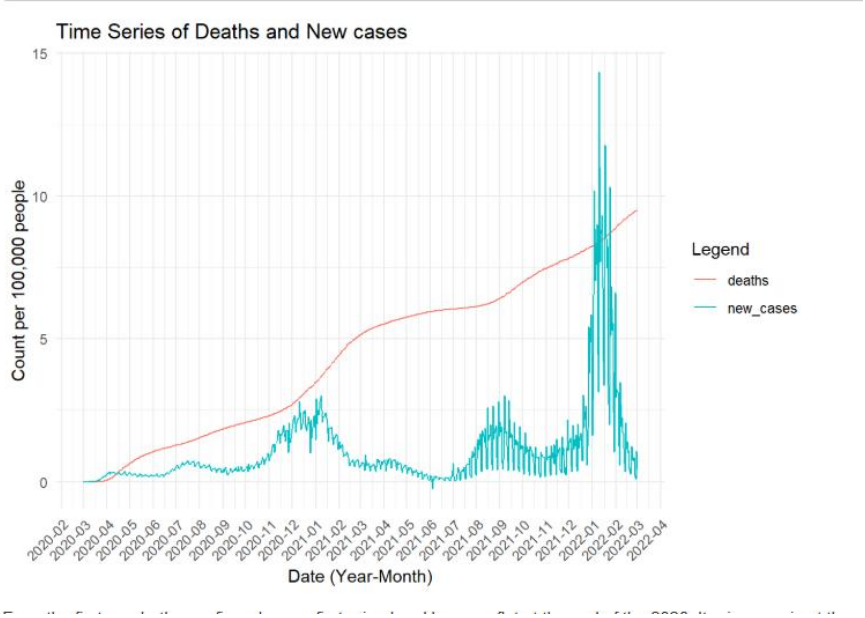
From the cases of confirmed and cases of death, we saw even clearer that both cases go up first, but become flat at the end of the 2020 till the end of the 2021, after which the cases go up again. But the question still remains: What makes it stopped accelerating the end of the 2020? And what makes it go up again at the end of the 2021?

Time series plot

To see the trend more clearly. we can make time series plot for people vaccinated, fully vaccinated, confirmed, and deaths, new cases.



From the first graph, we see the vaccination starts off in late dec 2020, and the number of people vaccinated quickly catch up the number confirmed. The Vaccination goes to about 0.8 at the end of the 2022, despite the number of confirmed still rising at the time.

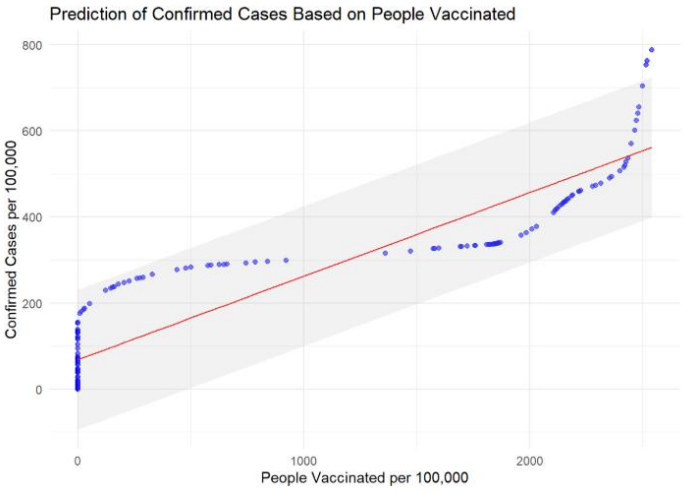


Recall that from the history Original Strain: March 2020. Vaccination begins: Dec 2020 Alpha: Widespread in Spring 2021. Delta: June-July 2021. Omicron: December 2021-January 2022. Omicron Sub variants: Throughout 2022-2023.

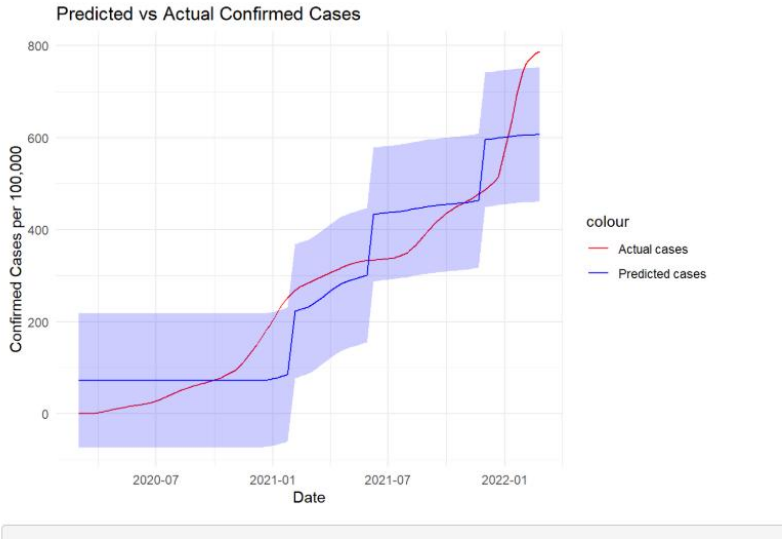
Let's take into account of the variant outbreak and go over it again.

In March 2020, pandemic outbreaks. At the end of the 2020, vaccination was adopted. Despite the appearance of the alpha variants, the number of cases dropped, showing the effectiveness of the vaccination. However, we experience a sudden spike during each variants outbreak, namely from variants of Alpha, Delta, Omicron.

Then, we can use vaccination rates or combination of vaccination rates and variants number to predict the confirmed cases. See which model is more reliable from below.



with RMSE: 79.18228. From the prediction interval you saw that by a 95% of confidence, our model is actually not a bad model. It followed the general trend of the new cases well. However, we also wants to incorporate the variants as a variable together with vaccination.



From model: P-value is less than 2.2e-16, and so is statistically significant. RMSE is 73.679 which is lower than RMSE of using just vaccination which is 79.18228.

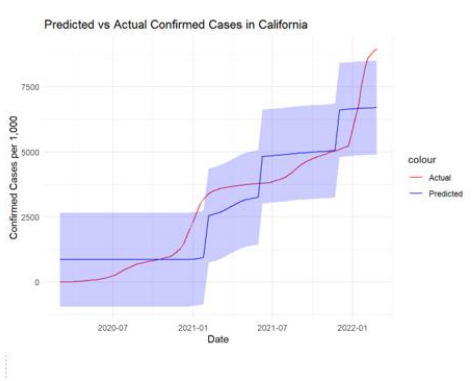
From the interval you can also see the prediction is much closer to the trend of the actual data.

So, we will be using the vaccination rates and variants number as predictor for the cases in Sample state.

Prediction within state and across the states

We will first conduct a regression model on California, using variant and vaccination rate as

predictor to predict the confirmed.




```
# Print model summary and RMSE for California
summary(model_ca)
```

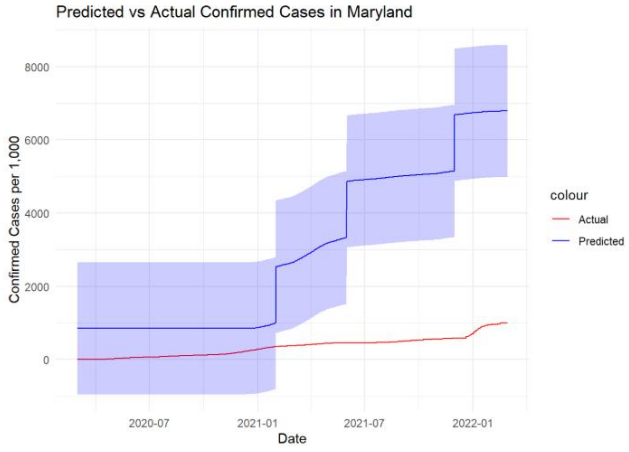
```
##
## Call:
## lm(formula = confirmed ~ variant + vaccination_rate, data = train_data_ca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1503.3  -768.5  -141.9   609.1  2338.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       855.08      52.89  16.168  < 2e-16 ***
## variant          1535.56     130.07  11.806  < 2e-16 ***
## vaccination_rate 1539.61     439.30   3.505 0.000492 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 914.7 on 584 degrees of freedom
## Multiple R-squared:  0.8506, Adjusted R-squared:  0.8501
## F-statistic:  1662 on 2 and 584 DF,  p-value: < 2.2e-16
```

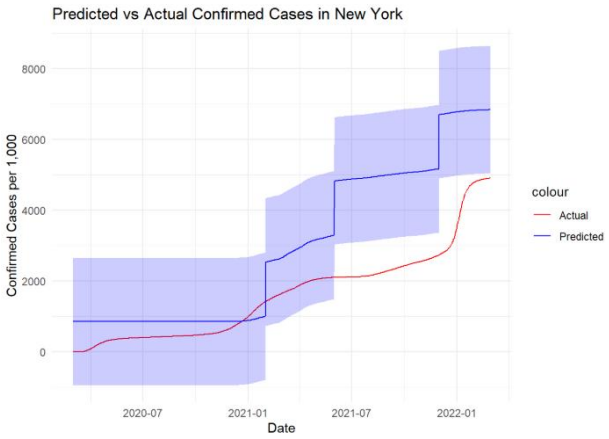
```
cat("RMSE for California: ", rmse_ca, "\n")
```

```
## RMSE for California:  900.6865
```

Again, P-value is less than 2.2e-16(small). And the RMSE is around 900. As we saw from the graph, Prediction follows well with the actual data.

Next, we will use the prediction model by california to predict the confirmed cases in different states like Maryland, New York to see if the prediction can be used accross the states.





For comparison, The RMSE of Maryland is 3278.314, while the RMSE of New York is 1761.337. Although the prediction of New York based on the situation of California is much better than the prediction of Maryland, both predictions deviate a lot from the actual data. This means that although variant and vaccination rate can be combined to well predict the confirmed cases, but

this is only true for a specific region, like for a state or for a country as a whole. We cannot use the model trained by one state to predict the other, because of the difference that lies within the population base and/or policies enforced within that region.

Conclusion

This project highlights the complexity of pandemic control, emphasizing the multifaceted influences beyond vaccination rates.

the vaccination is effective in controlling the pandemic.

However, there are also other factors like variants of the virus that play a crucial roles in determining the future confirmed cases. When combined, vaccination rate and variants can be a good predictor of the confirmed cases.

Last but not least, the situation is different from state to state. Thus, we cannot use the model train by one state to predict other states without the assumption that ensures the high similarity between the state that was used to predict and the state that was actually predicted.

Reference

Guidotti, E., (2022), “A worldwide epidemiological database for COVID-19 at fine-grained spatial resolution”, Sci Data 9(1):112, doi: 10.1038/s41597-022-01245-1

Link to the Zoom recording

https://us06web.zoom.us/rec/share/QJT6rCZfiWn6s0zxZO2Jk5gwbV6_mpdGueU6tndmkLqCub-tnyXtgrCLl1K2mhek.G_Csjaiw1G6OwJPg?startTime=1733524536000
Passcode: 2bJ9=Y6n

Tables and Code

```
cleaned_df_1 <- df_1 |>
  select(id, date, confirmed, deaths, people_vaccinated, people_fully_vaccinated, population,
         administrative_area_level_1) |>
  filter(
    administrative_area_level_1 == 'United States',
    !is.na(date),
    between(as.Date(date), as.Date("2020-03-01"), as.Date("2022-03-01"))
  ) |>
  mutate(
    confirmed = ifelse(is.na(confirmed), 0.0, confirmed),
    deaths = ifelse(is.na(deaths), 0.0, deaths),
    people_vaccinated = ifelse(is.na(people_vaccinated), 0.0, people_vaccinated),
    people_fully_vaccinated = ifelse(is.na(people_fully_vaccinated), 0.0, people_fully_vaccinated)
  )

# Convert counts to per 100,000 population
# and add two new columns

cleaned_df_1 <- cleaned_df_1 |>
  mutate(
    vaccination_rate = people_vaccinated / population,
    confirmed = confirmed / 100000,
    deaths = deaths / 100000,
    people_vaccinated = people_vaccinated / 100000,
    people_fully_vaccinated = people_fully_vaccinated / 100000,
    population = population / 100000,
    new_cases = c(0, diff(confirmed))
  )

# Add the 'variant' column based on historical data
cleaned_df_1 <- cleaned_df_1 |>
  mutate(
    variant = case_when(
      date >= as.Date("2020-03-01") & date <= as.Date("2021-01-31") ~ 0,  # Original Strain
      date >= as.Date("2021-02-01") & date <= as.Date("2021-05-31") ~ 1,  # Alpha
      date >= as.Date("2021-06-01") & date <= as.Date("2021-11-30") ~ 2,  # Delta
      date >= as.Date("2021-12-01") & date <= as.Date("2022-03-01") ~ 3   # Omicron
    )
  )
```

C1

```
cleaned_df_2 <- df_2 |>
  select(id, date, confirmed, deaths, people_vaccinated, people_fully_vaccinated, population,
         administrative_area_level_1, administrative_area_level_2) |>
  filter(
    administrative_area_level_1 == 'United States',
    !is.na(date),
    between(as.Date(date), as.Date("2020-03-01"), as.Date("2022-03-01"))
  ) |>
  mutate(
    confirmed = ifelse(is.na(confirmed), 0.0, confirmed),
    deaths = ifelse(is.na(deaths), 0.0, deaths),
    people_vaccinated = ifelse(is.na(people_vaccinated), 0.0, people_vaccinated),
    people_fully_vaccinated = ifelse(is.na(people_fully_vaccinated), 0.0, people_fully_vaccinated)
  )

# Convert counts to per 1,000 population  for level 2 dataset
# and add two new columns
cleaned_df_2 <- cleaned_df_2 |>
  group_by(administrative_area_level_2) |>
  mutate(
    vaccination_rate = people_vaccinated / population,
    confirmed = confirmed / 1000,
    deaths = deaths / 1000,
    people_vaccinated = people_vaccinated / 1000,
    people_fully_vaccinated = people_fully_vaccinated / 1000,
    new_cases = c(0, diff(confirmed)),
    population = population / 1000
  ) |>
  ungroup()

# Add the 'variant' column based on historical data
cleaned_df_2 <- cleaned_df_2 |>
  mutate(
    variant = case_when(
      date >= as.Date("2020-03-01") & date <= as.Date("2021-01-31") ~ 0,  # Original Strain
      date >= as.Date("2021-02-01") & date <= as.Date("2021-05-31") ~ 1,  # Alpha
      date >= as.Date("2021-06-01") & date <= as.Date("2021-11-30") ~ 2,  # Delta
      date >= as.Date("2021-12-01") & date <= as.Date("2022-03-01") ~ 3   # Omicron
    )
  )
```

C2

```
# Conduct regression model on California
california_data <- cleaned_df_2 |>
   filter(administrative_area_level_2 == 'California')

# Split data into training and testing sets for California
set.seed(123)
train_index_ca <- createDataPartition(california_data$confirmed, p = 0.8, list = FALSE)
train_data_ca <- california_data[train_index_ca, ]
test_data_ca <- california_data[-train_index_ca, ]

# Train a linear model for California
model_ca <- lm(confirmed ~ variant + vaccination_rate, data = train_data_ca)
summary(model_ca)
# Predict on test data for California
predictions_ca <- predict(model_ca, newdata = test_data_ca, interval = "prediction")

# Calculate RMSE for California
rmse_ca <- sqrt(mean((test_data_ca$confirmed - predictions_ca[, "fit"])^2))
cat("RMSE for California: ", rmse_ca, "\n")
# Combine predictions with test data for visualization
predicted_df_ca <- cbind(test_data_ca, predictions_ca)

# Plot predictions vs actual values for California
ggplot(predicted_df_ca, aes(x = date)) +
   geom_line(aes(y = confirmed, color = "Actual")) +
   geom_line(aes(y = fit, color = "Predicted")) +
   geom_ribbon(aes(ymin = lwr, ymax = upr), alpha = 0.2, fill = "blue") +
   labs(title = "Predicted vs Actual Confirmed Cases in California", y = "Confirmed Cases per 1,000", x =
"Date") +
   theme_minimal() +
   scale_color_manual(values = c("Actual" = "red", "Predicted" = "blue"))
```

C3