**Prediction of Daily, New, Confirmed COVID-19 Cases with Multiple Linear Regression**

Author: XiaoTong (Jack) Wu

Source Code: [github.com/JackOfSpade/COVID\_Regression\_Analysis](https://github.com/JackOfSpade/COVID_Regression_Analysis) (rename repo to better name later)

## Abstract

The aim of this research is to create a model that can predict the amount of daily, new, confirmed COVID-19 cases as a percentage of the population with respect to the cumulative amount of vaccine doses administered. Thus, the model is most accurate when used for a country with a similar healthcare infrastructure as the United States. Our final model is as follows:

where for a specific sample:

Note: time order

coefficient from Cochrane-Orcutt procedure

Response variable is the predicted daily new confirmed COVID-19 cases as a percentage of the population.

Explanatory variables are the cumulative vaccine doses administered by Pfizer, Moderna and Johnson&Johnson, respectively

The COVID-19 pandemic originated from the city of Wuhan of China has highly affected the health, socio-economic and financial matters of the different countries of the world. India is one of the countries which is affected by the disease and thousands of people on daily basis are getting infected. In this paper, an analysis of daily statistics of people affected by the disease are taken into account to predict the next days trend in the active cases in Odisha as well as India.

**Material and methods**

A valid global data set is collected from the WHO daily statistics and correlation among the total confirmed, active, deceased, positive cases are stated in this paper. Regression model such as Linear and Multiple Linear Regression techniques are applied to the data set to visualize the trend of the affected cases.

**Results**

**1. Introduction**

[Go to:](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395225/)

According to University of Missouri Health Care, “herd immunity would require around 90% of the population to have COVID-19 immunity, either through prior infection or vaccination.”2 The effectiveness of vaccines on each person varies due to the different response of individual immune systems. Some people may not generate an adequate response to the vaccine to acquire an effective protection.4 Thus, we cannot say when 90% of the population is vaccinated, we will have herd immunity. However, if herd immunity exists, we should expect no more than 10% of the population with active COVID-19 cases at any given time. According to Centers for Disease Control and Prevention, the average duration of COVID-19 cases is about 2 weeks (14 days).3 Thus, we arrive at our target for daily, new, confirmed COVID-19 cases at of the total population.

We will create a model that will alert us of such an event by first starting with an intuitive base model and then improving it from there through various tests.

**2. Materials and methods used**

[Go to:](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395225/)

To find our desired model, we will take data from Google’s COVID-19 public dataset program (<https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-open-data>) with the following BigQuery standard SQL:

SELECT new\_confirmed, population, cumulative\_vaccine\_doses\_administered\_pfizer,cumulative\_vaccine\_doses\_administered\_moderna,cumulative\_vaccine\_doses\_administered\_janssen

FROM `bigquery-public-data.covid19\_open\_data.covid19\_open\_data`

where new\_confirmed >= 0

AND cumulative\_vaccine\_doses\_administered\_pfizer is not NULL

AND cumulative\_vaccine\_doses\_administered\_moderna is not NULL

AND cumulative\_vaccine\_doses\_administered\_janssen is not NULL

We begin with the following multiple linear regression equation which we will improve on later in our analysis:

Response variable is the the daily new confirmed COVID-19 cases as a percentage of the population.

Explanatory variables are the cumulative vaccine doses administered by Pfizer, Moderna and Johnson&Johnson, respectively.

**Interaction Terms**

We will first check for interactions between the explanatory variables. Following the convention suggested by Cohen5 and popularized by Aiken and West6, we will use the mean value of the moderating variable (moderator) as well as one standard deviation above and below the mean to plot the effect of the moderator on an explanatory variable. Then, we have the following graphs:

Chart, line chart

Description automatically generated

Interpretation: There is no significate interaction effects between and .

Chart, line chart

Description automatically generated

Interpretation: There exists an interaction effect between and .

Chart, line chart

Description automatically generated

Interpretation: There exists an interaction effect between and .

In order to account for these interaction effects, we add interaction terms to our regression equation. This gives us 4 possible models for consideration:

Model #1:

Model #2:

Model #3:

Model #4:

In this case, the best model is tentatively model :

This model outperformed all three other models in most of our tests. In areas where it did fall short, the difference was too insignificant to matter. The adjusted of this model is with a multiple correlation coefficient of 0.9472, meaning there is a high predictability of the response variable from the explanatory variables.8 However, we do understand that all forms of has flaws as a goodness of fit measurement.7 As such, we must conduct additional tests.

**-test for a Portion of the Model**

Although there exists an interaction effect between and , we chose not to add the interaction term to the regression equation on the basis that it was too insignificant. We will prove this using an -test for a portion of a model:

Let be the complete model

Let be the reduced model

number of explanatory variables dropped

number of samples

number of parameters (including ) in the complete model

We will use

Thus, we cannot reject in favor of .

In addition, the partial coefficient of determination is:

The partial coefficient of determination tells us the portion of the unexplained variation in the reduced model that is explained by the extra explanatory variable, , in the completed model As you can see, it is very low.

By the -test for a portion of a model and the partial coefficient of determination, there is no evidence that is a significant explanatory variable. As such, we conclude that dropping from model to arrive at our chosen model (model #2) was the right choice.

**Residual Plots**

**Residual vs**

Chart, scatter chart

Description automatically generated

**Interpretation:** We can see a slight fan-out pattern on the data points. The residual variance increases as the criterion increases. The homoscedasticity assumption is violated. To fix this, we will transform our linear model into a log-linear model.

Then, we get the following graph:

**Residual vs**

Chart, scatter chart

Description automatically generated

**Interpretation:** Much better. No clear pattern (i.e. fan-in, fan-out, non-linear or double bow) exists. The residuals bounce randomly around the residual = 0 line. The residuals roughly form a "horizontal band" around the residual = 0 line. This suggests that the variances of the error terms are roughly equal. In addition, no one residual stands out from the basic random pattern of the residuals. This suggests that there are no outliers.

**Residual vs Time Order**

Chart, scatter chart

Description automatically generated

**Interpretation:** The pattern is not random enough. We suspect there is positive autocorrelation. We will confirm the existence of first-order positive autocorrelation through the **Durbin-Watson test**:

The error terms are not autocorrelated.

The error terms are positively autocorrelated.

We choose

Then we have:

p-value

Thus, we reject the null hypothesis and accept the alternative hypothesis that the model is positively autocorrelated in the first order.

We will attempt to fix this using the **Cochrane-Orcutt** procedure.

We determine

We then transform each of our sample data to:

Note: time order

The above transformation suggested by Cochrane and Orcutt disregards the first observation (we must start from ), causing a loss of efficiency that can be substantial in small sample sizes such as this one. A superior transformation, which retains the first observation with a weight of was suggested first by Prais and Winsten10 and later independently by Kadilaya11, which we will apply here.

Our new tentative model becomes:

We perform the **Durbin-Watson test** again on our new model.

The error terms are not autocorrelated.

The error terms are positively autocorrelated.

We choose

Then we have:

p-value

Thus, cannot be rejected, so we don’t have significant evidence that the model is positively autocorrelated

We do note here that the resulting p-value is barely above our p-value threshold of 0.01, so we might consider modifying our model to a time series model or some non-linear model in the future, but for the purpose this paper we consider it satisfactory for a multiple linear regression model.

To save space, we will just state here that after a few -tests, we also had to remove the explanatory variable due to finding it insignificant in our new model, having a -test p-value of which is above our acceptable threshold of . Finally, we arrive at our final model of:

It maintains a Durbin-Watson p-value of , which is still above our threshold of

The rest of this paper will be dedicated to proving that our chosen model is adequate.

**-test**

To answer the question, “Does our chosen model have significant explanatory power overall?”9, we will conduct an -test:

At least one of

statistic

number of parameters (including )

number of samples

We will use

statistic

Thus, we reject and accept . This model does in fact have significant explanatory power overall. That is a good sign. We can move on to the next test.

**-tests**

We will now test the significance of each explanatory variable through -tests:

Let

For each

statistic

the diagonal element of the matrix

number of samples

number of parameters (including )

We will use

statistic

To save space, we omit the detailed calculations of the statistic for each coefficient and present the following results:

|  |  |  |
| --- | --- | --- |
| **Coefficient** | **Statistic** |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

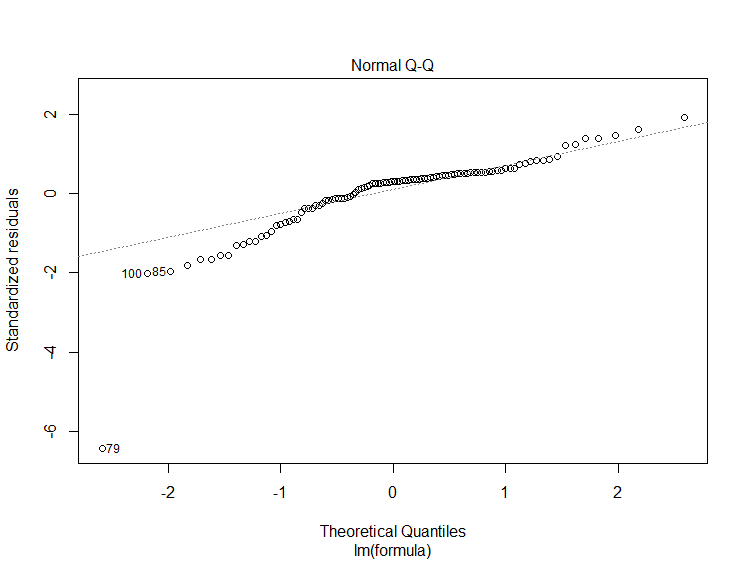
Because every , we reject every for all the coefficients and accept every Thus, every coefficient in our chosen model is significant.

**Confidence Interval**

Our model falls within the 95% confidence interval:

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficient** | **Lower Bound (2.5%)** | **Our Estimate** | **Upper Bound (97.5%)** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Q-Q PlotIntroduction and Aims**



**Interpretation:** The plot is roughly a straight line, so the normality assumption holds.

**3. Results**

[Go to:](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395225/)**Introduction and Aims**

Our final model is as follows:

where for a specific sample:

Note: time order

coefficient from Cochrane-Orcutt procedure

Response variable is the predicted daily new confirmed COVID-19 cases as a percentage of the population.

Explanatory variables are the cumulative vaccine doses administered by Pfizer, Moderna and Johnson&Johnson, respectively

For a specific sample, we can derive from as follows:

Let

**4. Conclusion**

[Go to:](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395225/)

Our final model can be used to predict daily, new, confirmed COVID-19 cases when given the cumulative doses of Pfizer, Moderna, and Johnson&Johnson administered separately.

Because we have more than 1 explanatory variable, predicting the combination of cumulative vaccines administered when given a response variable equivalent to of the total population (herd immunity) can have infinite solutions and be very complicated to solve. As such, it is beyond the scope of this research. However, if new data of explanatory variables arrive such that herd immunity can be achieved, this model will alert us of such a prediction. In order words, we can easily predict given new s, but much more difficult to predict given new .

**References**

1. *U.S. and world Population clock*. United States Census Bureau. (2021, September 12). Retrieved September 12, 2021, from https://www.census.gov/popclock/.
2. *Covid-19 vaccine key to reaching 'herd immunity'*. University of Missouri Health Care. (2021). Retrieved September 13, 2021, from https://www.muhealth.org/our-stories/covid-19-vaccine-key-reaching-herd-immunity.
3. Centers for Disease Control and Prevention. (2021, July 9). *COVID-19 quarantine and isolation*. Centers for Disease Control and Prevention. Retrieved September 13, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/your-health/quarantine-isolation.html.
4. *Top 20 questions about vaccination*. History of Vaccines. (2018, January 25). Retrieved September 13, 2021, from https://ftp.historyofvaccines.org/index.php/content/articles/top-20-questions-about-vaccination.
5. Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Thousand Oaks, CA: Sage.
6. Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426 – 443.
7. Shalizi, C. R. (2019). *The Truth about Linear Regression*. Carnegie Mellon University.
8. *Estimating maximum value of multiple correlation*. Testbook. (n.d.). Retrieved September 13, 2021, from https://testbook.com/question-answer/for-estimating-maximum-value-of-multiple-correlati--5faa49ea946b817c64f64b92.
9. *PBAF 528 Week 4*. University of Washington. (n.d.). Retrieved September 13, 2021, from http://depts.washington.edu/lecturer/528-Sp05/Notes/Week%204.pdf.
10. Prais, S. J.; Winsten, C. B. (1954). "Trend Estimators and Serial Correlation". Cowles Commission Discussion Paper No. 383. Chicago.
11. Kadiyala, Koteswara Rao (1968). "A Transformation Used to Circumvent the Problem of Autocorrelation". Econometrica. **36** (1): 93–96. JSTOR 1909605.