**Prediction of New Confirmed COVID-19 Cases Using a Multiple Linear Regression Model**

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

aa**Introduction and Aims**

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/)

The aim of this research is to determine the cumulative amount of vaccine doses needed to be administered such that herd immunity can occur in the United States. We measure the existence of herd immunity with the rate of new confirmed COVID-19 cases.

**2. Materials and methods used**

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

**-test for a Portion of the a 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.

**-test**

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

At least one of

number of parameters (including )

number of samples

We will use

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 step.

**-tests**

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

Let

For each

the diagonal element of the matrix

number of samples

number of parameters (including )

We will use

Let

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

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

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

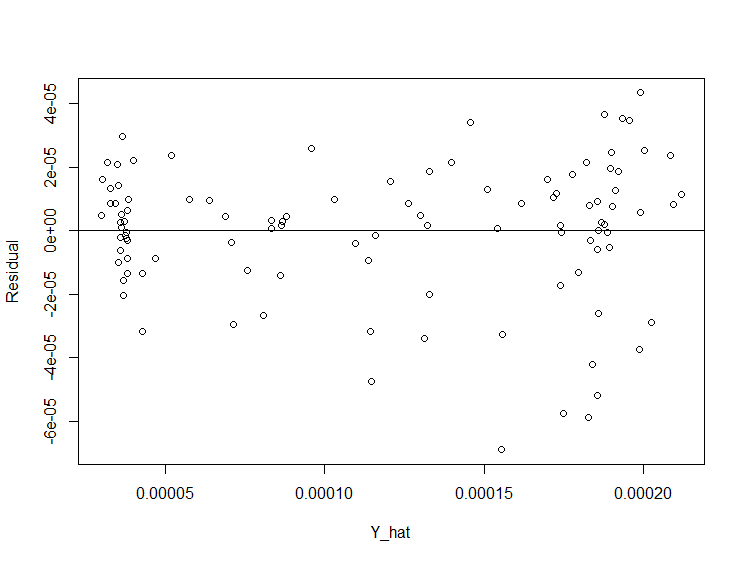
**Confidence Interval**

Our model falls within the 95% confidence interval:

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

**Residual Plots**

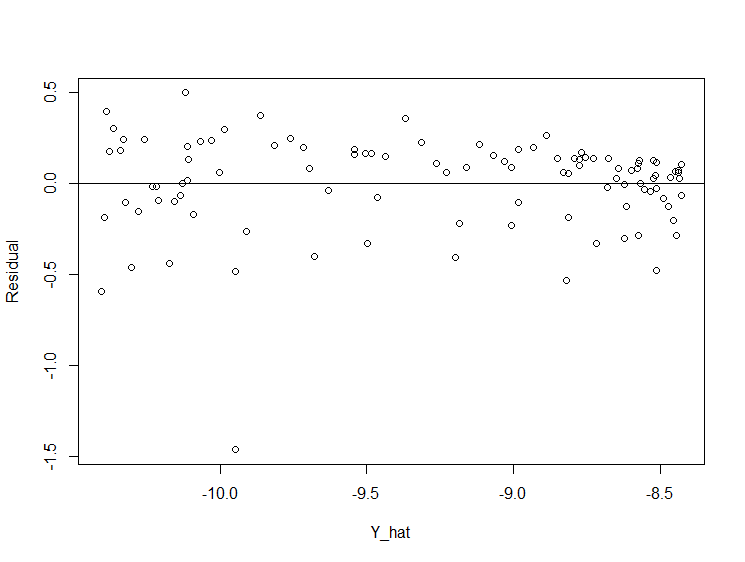
**Residual vs**



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.

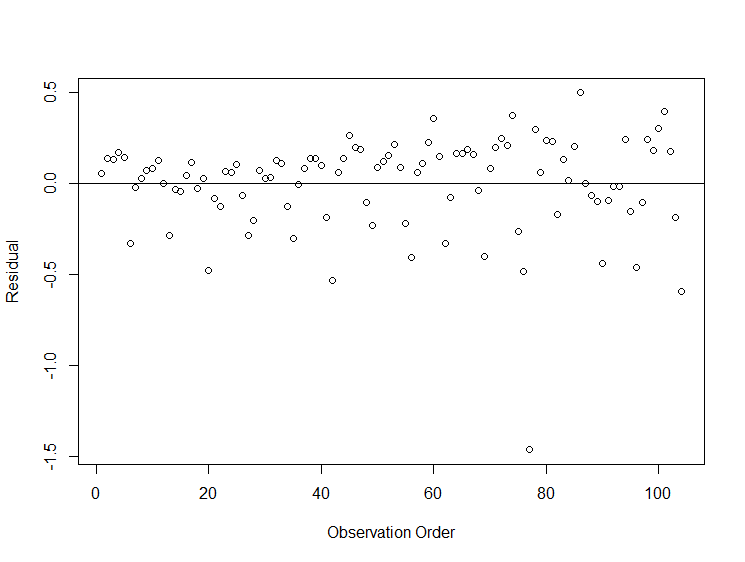
Then, we get the following graph:

**Residual vs**



Interpretation: Much better. No clear pattern (i.e. fan-in, fan-out, non-linear or double bow) exists, suggesting we have an appropriate model.

**Residual vs Time Order**



Interpretation: No clear pattern exists, suggesting we have an appropriate model.

FINISH OTHER TESTS THEN MOVE RESIDUAL PLOT TO AFTER INTERACTION TEST.

Prediction interval when forecasting

**Introduction and Aims**

**3. Results**

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

a**Introduction and Aims**

**4. Discussion**

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

Flaws:

* Vaccine does not protect 100%, the different cultural norms may influence the effectiveness of the model.

Derivations:

The average vaccine investment per person in USD to achieve a

**5. Conclusion**

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

aa**Introduction and Aims**

**Funding**

No funding.

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

**Declaration of competing interest**

None to declare.

**Acknowledgements**

The authors thank J.A. Professor for helpful comments on a draft of the manuscript.

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

**References**

1. https://www.census.gov/popclock/
2. https://www.muhealth.org/our-stories/covid-19-vaccine-key-reaching-herd-immunity
3. <https://www.cdc.gov/coronavirus/2019-ncov/if-you-are-sick/end-home-isolation.html>
4. <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. Chapter 10.3.2, The Truth about Linear Regression, Cosma Rohilla Shalizi, <http://www.stat.cmu.edu/~cshalizi/TALR/TALR.pdf>
8. https://testbook.com/question-answer/for-estimating-maximum-value-of-multiple-correlati--5faa49ea946b817c64f64b92
9. http://depts.washington.edu/lecturer/528-Sp05/Notes/Week%204.pdf