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

XiaoTong (Jack) Wu

Source Code

Abstract

The aim of this research is to create a model that can predict the amount of daily, new COVID-19 cases as a percentage of the population with respect to the cumulative amount of vaccine doses administered. The model uses data from the United States, but it can also be used in prediction for any country that has a similar social culture. Our final model is as follows:

$$\log(Y^*) = \beta_0 + \beta_2 X_2^* + \beta_3 X_3^* + \beta_4 (X_1 X_3)^*$$

where for a specific sample:

$y_t^* = y_t - \rho y_{t-1}$

$x_{1,t}^* = x_{1,t} - \rho x_{1,t-1}$

$x_{2,t}^* = x_{2,t} - \rho x_{2,t-1}$

$\left(x_{1,t} x_{3,t}\right)^* = x_{1,t} x_{3,t} - \rho x_{1,t-1} x_{3,t-1}$

$t =$ time order

$\rho =$ coefficient from the Cochrane-Orcutt procedure

Response variable (Y) is the predicted daily, new COVID-19 cases as a percentage of the population.

Explanatory variables (X_1, X_2, X_3) are the cumulative vaccine doses administered by Pfizer, Moderna and Johnson&Johnson, respectively.

For a specific prediction, we can reverse the log-transformation and derive $\hat{y}_t$ as follows:

Consider $\log(\hat{y}_t^*) = \hat{\beta}_0 + \hat{\beta}_2 x_{2,t}^* + \hat{\beta}_3 x_{3,t}^* + \hat{\beta}_4 \left(x_{1,t} x_{3,t}\right)^*$

Let $\hat{\beta}_0 + \hat{\beta}_2 x_{2,t}^* + \hat{\beta}_3 x_{3,t}^* + \hat{\beta}_4 \left(x_{1,t} x_{3,t}\right)^* = C$

$\log(\hat{y}_t{}^*) = C$

$\hat{y}_t{}^* = e^C$

$\hat{y}_t - \rho y_{t-1} = e^C$

$\hat{y}_t = e^C + \rho y_{t-1}$

1. Introduction

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 upper limit target for daily, new COVID-19 cases at $\frac{10\%}{14\ days} = \frac{0.1}{14} = \frac{1}{140}$ of the total population.

We will create a model that will alert us when we are most likely to achieve such an event, saving us the costs of manually surveying the number of daily, new COVID-19 cases. We first start with an intuitive base model and then improving it from there through various tests.

2. Materials and methods used

To find our desired model, we import data from Google's COVID-19 public dataset program.[12]

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

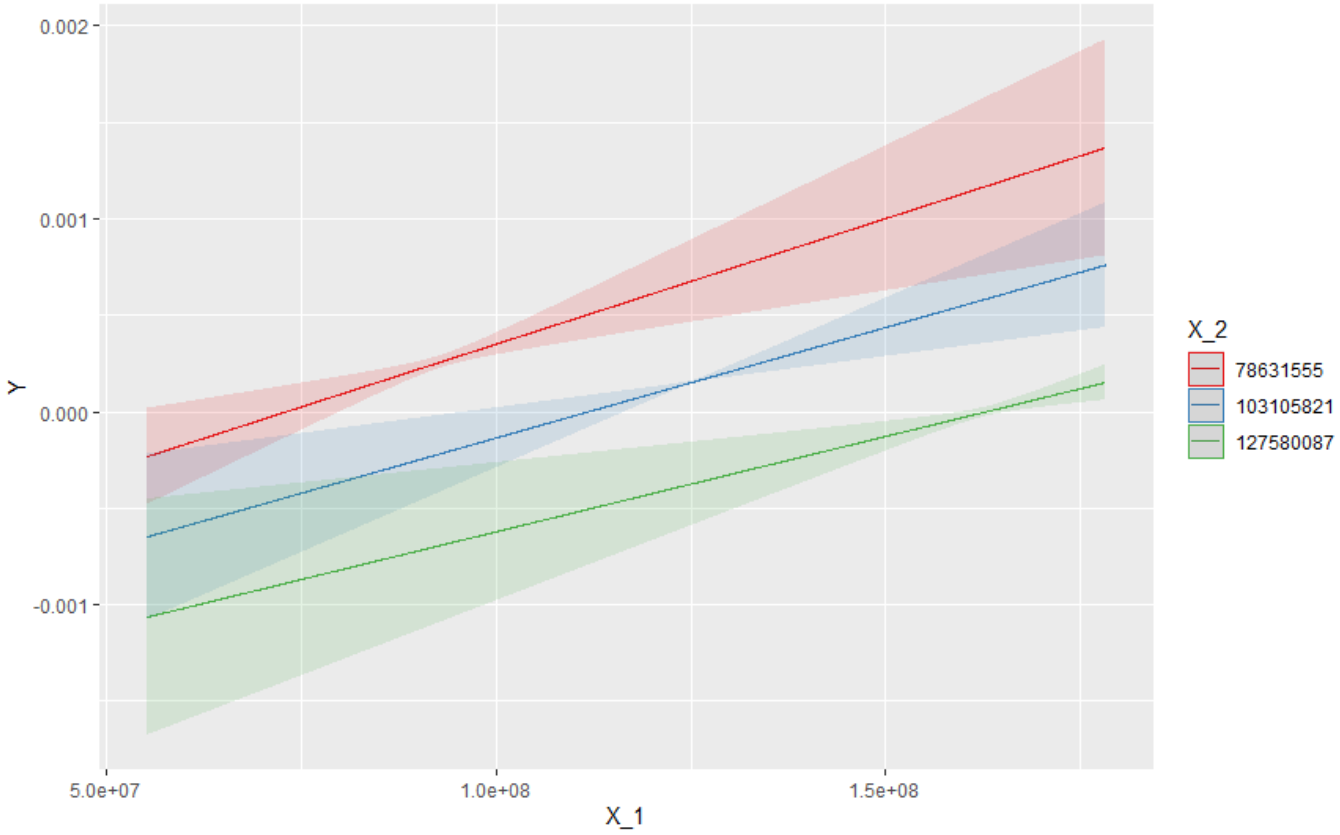
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Response variable (Y) is the predicted daily, new COVID-19 cases as a percentage of the population.

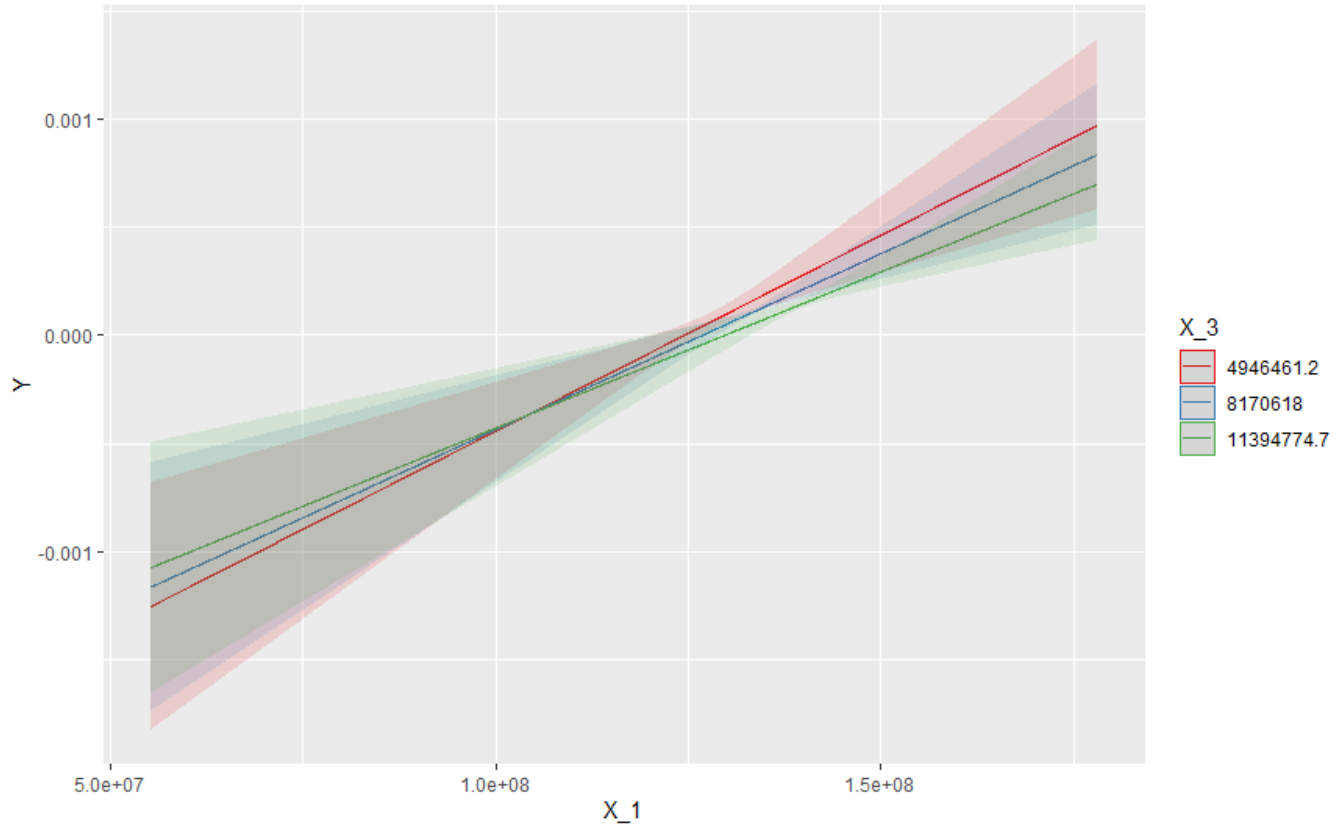
Explanatory variables (X_1, X_2, X_3) are the cumulative vaccine doses administered by Pfizer, Moderna and Johnson&Johnson, respectively.

Interaction Terms

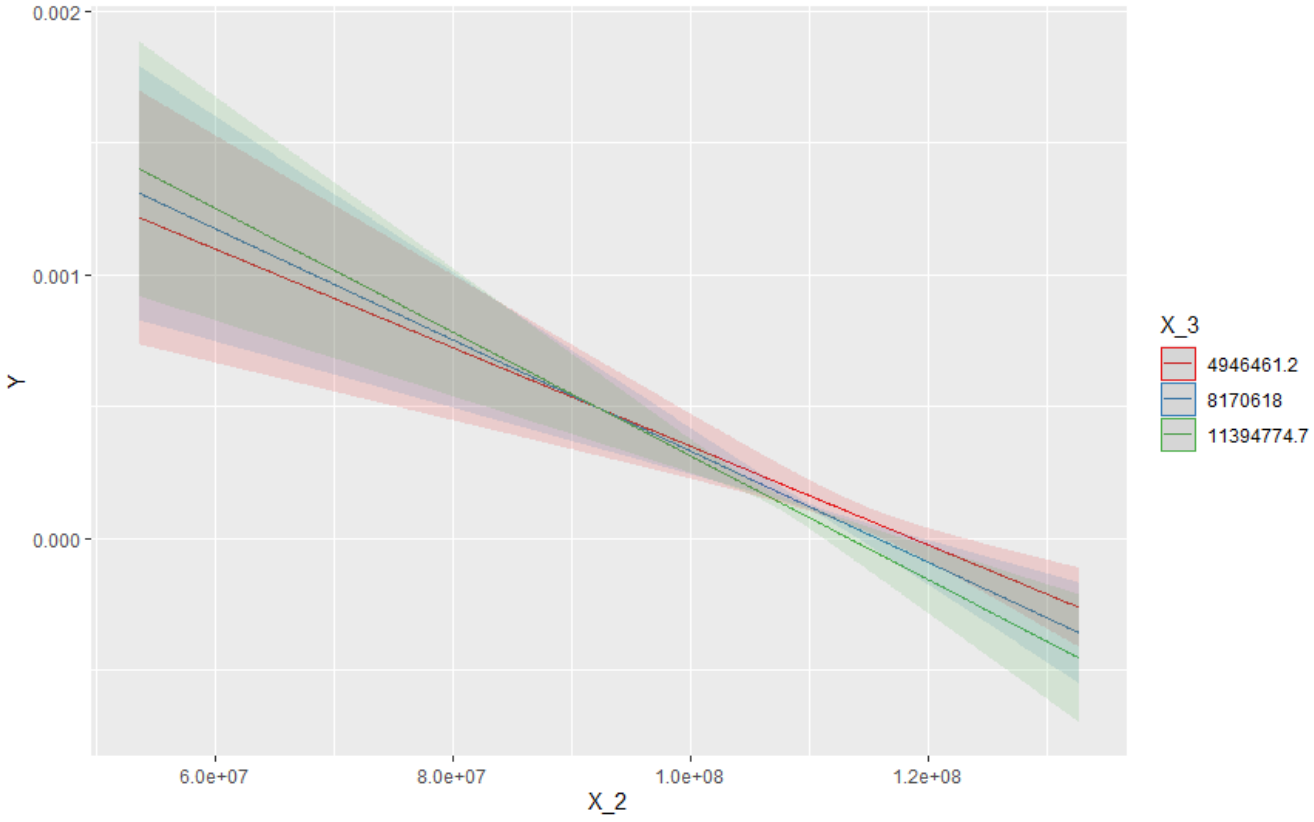
We will first check for interactions between the explanatory variables. Following the convention suggested by Cohen[5] and popularized by Aiken and West[6], we will use the mean value of the moderating variable (moderator) as well as one standard deviation above and below the mean to plot the effect of the moderator on an explanatory variable. Then, we have the following graphs:



Interpretation: There is no significate interaction effects between X_1 and X_2.



Interpretation: There exists an interaction effect between X_1 and X_3.



Interpretation: There exists an interaction effect between X_2 and X_3.

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: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Model #2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3$

Model #3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_2 X_3$

Model #4: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3$

In this case, the best model is tentatively model #2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3$$

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 R^2 of this model is 0.8972 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 R^2 has flaws as a goodness of fit measurement.[7] As such, we must conduct additional tests.

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

Although there exists an interaction effect between X_2 and X_3, we chose not to add the interaction term X_2X_3 to the regression equation on the basis that it was too insignificant. We will prove this using an F-test for a portion of a model:

Let C be the complete model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_3 + \beta_5X_2X_3$

Let R be the reduced model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_3$

$H_0: \beta_5 = 0$

$H_a: \beta_5 \neq 0$

$$F(x_2x_3|x_1, x_2, x_3, x_1x_3) = \frac{MS_{drop}}{MS_{Res}(C)} \sim F((p-g), (n-k))$$

$p - g =$ number of explanatory variables dropped

$n =$ number of samples

$k =$ number of parameters (including β_0) in the complete model C

$$F(x_2x_3|x_1, x_2, x_3, x_1x_3) = \frac{MS_{drop}}{MS_{Res}(C)}$$

$$= \frac{\frac{SS_{drop}}{p-g}}{4.8113 \cdot 10^{-10}}$$

$$= \frac{\frac{SS_{drop}}{1}}{4.8113 \cdot 10^{-10}}$$

$$= \frac{SS_{drop}}{4.8113 \cdot 10^{-10}}$$

$$= \frac{SS_{Res}(R) - SS_{Res}(C)}{4.8113 \cdot 10^{-10}}$$

$$= \frac{4.7216 \cdot 10^{-8} - 4.7151 \cdot 10^{-8}}{4.8113 \cdot 10^{-10}}$$

$$= \frac{6.4512 \cdot 10^{-11}}{4.8113 \cdot 10^{-10}}$$

$$= 0.1341$$

$F((p-g), (n-k))$

$\sim F\big(1,(104-6)\big)$

$\sim F(1,98)$

We will use $\alpha = 0.01$

$F_{\alpha}^{1,98} = F_{0.01}^{1,98} = 6.9008$

$0.1341 < 6.9008$

$F(x_2x_3|x_1,x_2,x_3,x_1x_3) < F_{0.01}^{1,98}$

Thus, we cannot reject H_0 in favor of H_a.

In addition, the partial coefficient of determination is:

$$R^2(x_2x_3|x_1,x_2,x_3,x_1x_3) = \frac{SS_{drop}}{SS_{Res}(R)}$$

$$= \frac{SS_{Res}(R) - SS_{Res}(C)}{4.7216 \cdot 10^{-8}}$$

$$= \frac{6.4512 \cdot 10^{-11}}{4.7216 \cdot 10^{-8}}$$

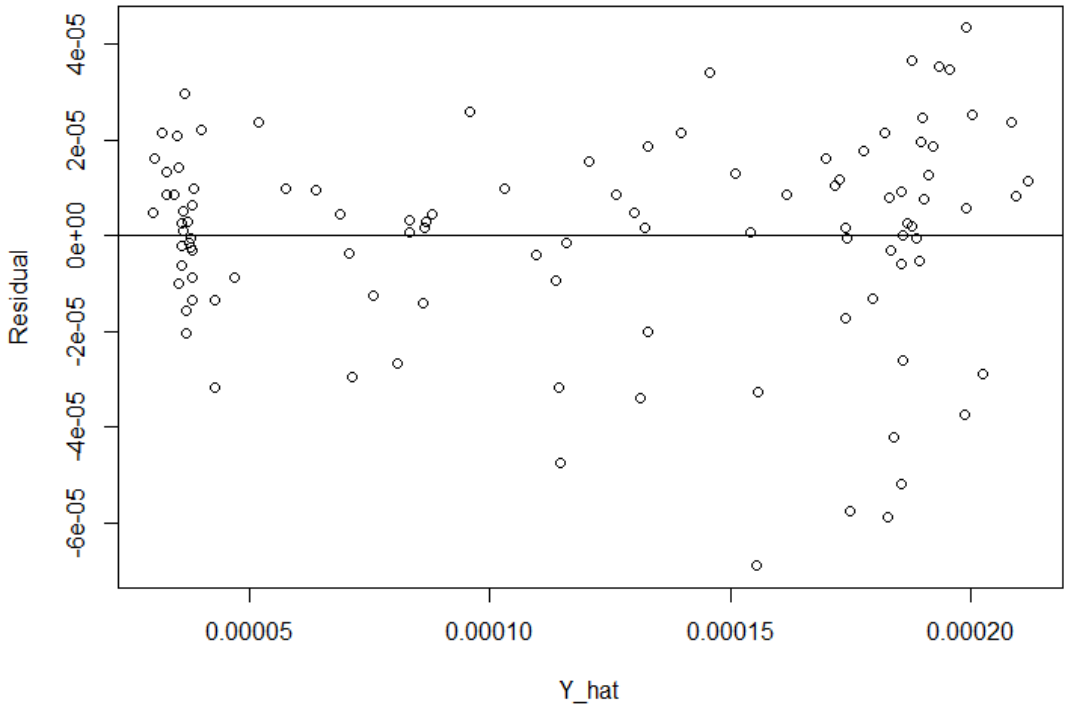
$$= 0.001466$$

The partial coefficient of determination tells us the portion of the unexplained variation in the reduced model R that is explained by the extra explanatory variable, X_2X_3, in the completed model C. As you can see, it is very low.

By the F-test for a portion of a model and the partial coefficient of determination, there is no evidence that X_2X_3 is a significant explanatory variable. As such, we conclude that dropping X_2X_3 from model C to arrive at our chosen model R (model #2) was the right choice.

Residual Plots

Residual vs $\widehat{Y}$

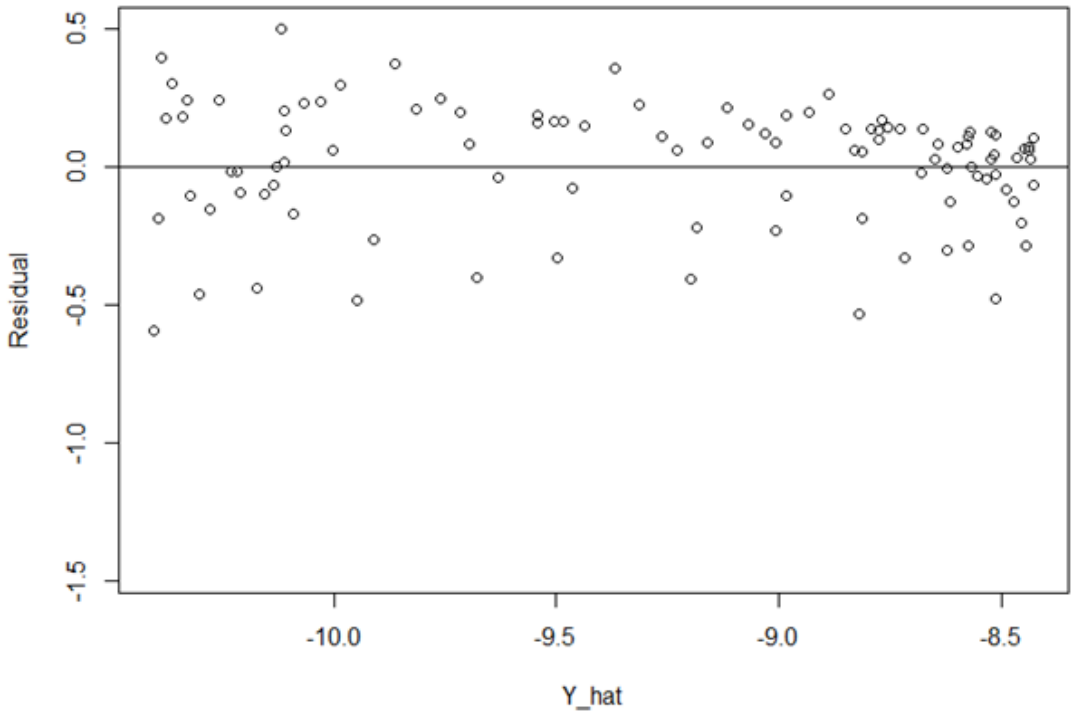


Interpretation: We can see a slight fan-out pattern on the data points. The residual variance increases as the criterion $\hat{Y}$ increases. The homoscedasticity assumption is violated. To fix this, we will transform our linear model into a log-linear model.

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3$$

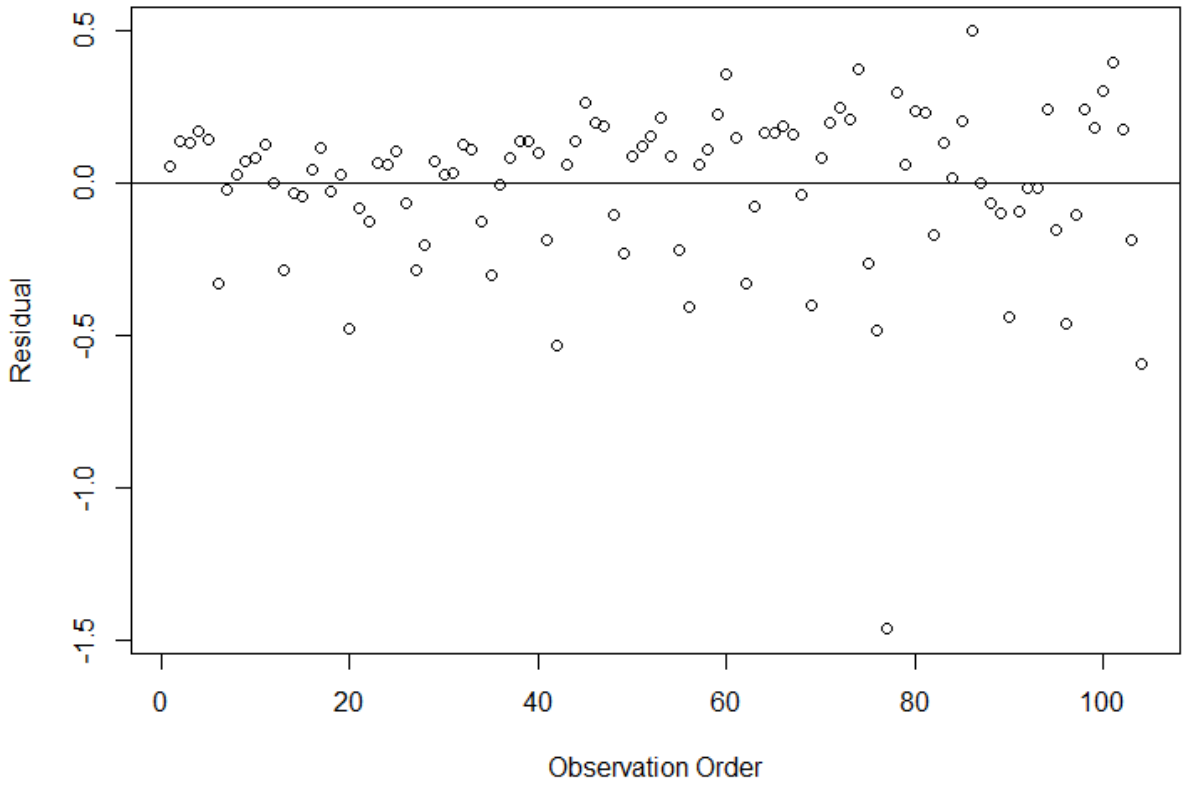
Then, we get the following graph:

Residual vs $\hat{Y}$



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



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**:

H_0: The error terms are not autocorrelated.

H_a: The error terms are positively autocorrelated.

We choose $\alpha = 0.01$

Then we have:

$d = 1.4104$

p-value $= 0.0002763$

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.

$e_t = \rho e_{t-1} + v_t$

We determine $\rho = 0.2844$

We then transform each of our sample data to:

$y_t^* = y_t - \rho y_{t-1}$

$x_{1,t}^* = x_{1,t} - \rho x_{1,t-1}$

$x_{2,t}^{*} = x_{2,t} - \rho x_{2,t-1}$

$\left(x_{1,t}x_{3,t}\right)^{*} = x_{1,t}x_{3,t} - \rho x_{1,t-1}x_{3,t-1}$

Note: $t =$ time order

The above transformation suggested by Cochrane and Orcutt disregards the first observation (we must start from $t = 2$), 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 $\sqrt{1-\rho^2}$ was suggested first by Prais and Winsten[10] and later independently by Kadilaya[11], which we will apply here.

$y_1^{*} = \sqrt{1-\rho^2} \cdot y_1$

Our new tentative model becomes:

$$\log\left(\hat{Y}^{*}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1^{*} + \hat{\beta}_2 X_2^{*} + \hat{\beta}_3 X_3^{*} + \hat{\beta}_4 (X_1 X_3)^{*}$$

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

H_0: The error terms are not autocorrelated.

H_a: The error terms are positively autocorrelated.

We choose $\alpha = 0.01$

Then we have:

$d = 1.6435$

p-value $= 0.03648$

$0.03648 > 0.01$

Thus, H_0 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 t-tests, we also had to remove the explanatory variable X_1 due to finding it insignificant in our new model, having a t-test p-value of $2P(T > |t|) = 0.4793$ which is above our acceptable threshold of 0.05. Finally, we arrive at our final model of:

$$\log(Y^{*}) = \beta_0 + \beta_2 X_2^{*} + \beta_3 X_3^{*} + \beta_4 X_1 X_3^{*}$$

It maintains a Durbin-Watson p-value of 0.02227, which is still above our threshold of 0.01.

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

***F*-test**

To answer the question, "Does our chosen model have significant explanatory power overall?"[9], we will conduct an F-test:

H_0: $\beta_2 = \beta_3 = \beta_4 = 0$

H_a: At least one of $\beta_2, \beta_3, \beta_4 \neq 0$

F statistic$= \frac{MS_R}{MS_{Res}} \sim F\big((k-1), (n-k)\big)$

$k =$ number of parameters (including β_0)

$n =$ number of samples

$$F \text{ statistic} = \frac{MS_R}{MS_{Res}}$$

$$= \frac{3236.914}{0.1723}$$

$$= 18782.85$$

$$F\big((k-1), (n-k)\big)$$

$$\sim F\big((4-1), (104-4)\big)$$

$$\sim F(3,100)$$

We will use $\alpha = 0.01$

$$F_{\alpha}^{3,100} = F_{0.01}^{3,100} = 3.9837$$

$$18782.85 > 3.9837$$

F statistic$> F_{0.01}^{3,100}$

Thus, we reject H_0 and accept H_a. This model does in fact have significant explanatory power overall. That is a good sign. We can move on to the next test.

***t*-tests**

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

Let $j = \{0, 2, 3, 4\}$

For each β_j :

$H_0: \beta_j = 0$

$H_a: \beta_j \neq 0$

t statistic= $\frac{\hat{\beta}_j - 0}{\hat{\sigma} - \sqrt{c_{jj}}} \sim t(n-k)$

$\hat{\sigma} = \sqrt{MS_{Res}}$

c_{jj} = the diagonal element of the matrix $\left(\tilde{X}'\,\tilde{X}\right)^{-1}$

$$\tilde{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} & x_{1,1}x_{1,3} \\ 1 & x_{2,1} & x_{2,,2} & x_{2,3} & x_{2,1}x_{2,3} \\ 1 & x_{3,1} & x_{3,2} & x_{3,3} & x_{3,1}x_{3,3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{104,1} & x_{104,2} & x_{104,3} & x_{104,1}x_{104,3} \end{bmatrix}$$

n = number of samples

k = number of parameters (including β_0)

We will use $\alpha = 0.05$

t statistic$\sim t(n-k)$

$\sim t(104-4)$

$\sim t(100)$

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

Coefficient	$\boldsymbol{t}$ **Statistic**	$\boldsymbol{2P(T > \lvert t \rvert)}$
$\hat{\beta}_0$	-30.374	$< 2 \cdot 10^{-16}$
$\hat{\beta}_2$	4.263	$4.58 \cdot 10^{-5}$
$\hat{\beta}_3$	4.168	$6.54 \cdot 10^{-5}$
$\hat{\beta}_4$	-8.958	$1.9 \cdot 10^{-14}$

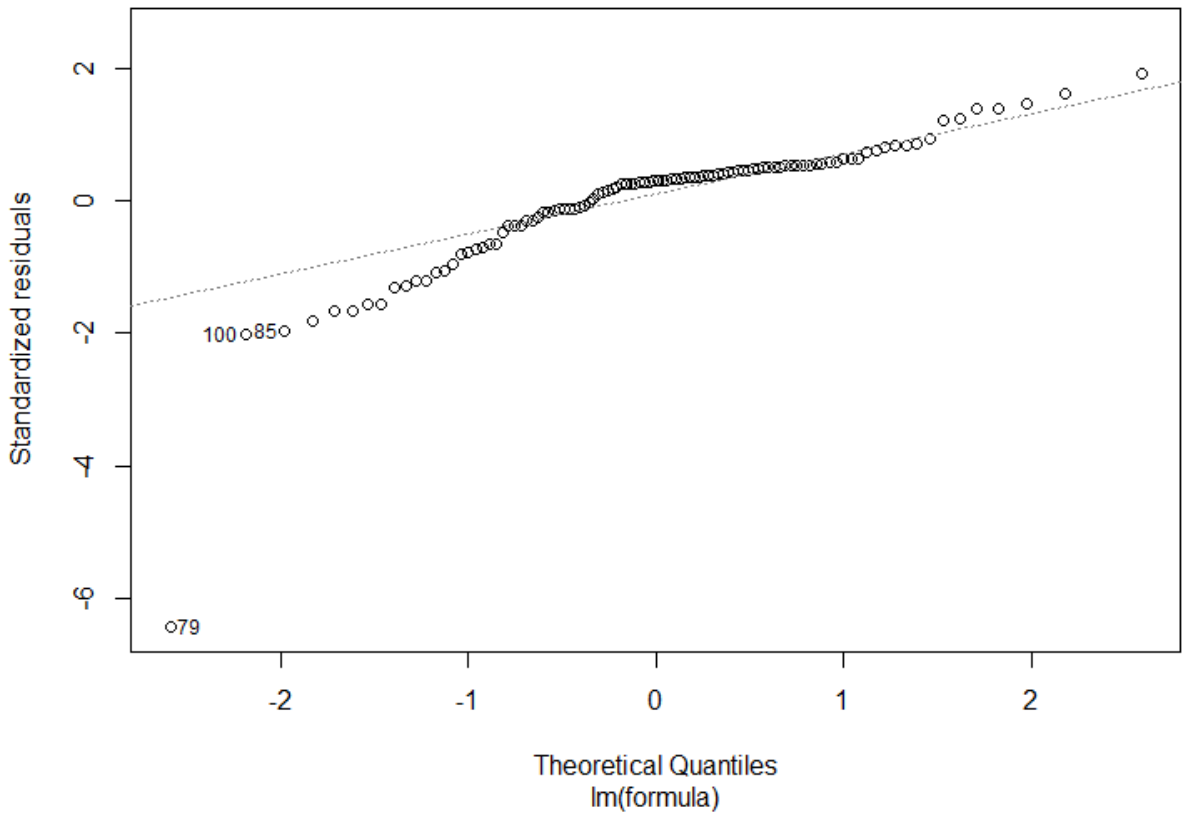
Because every $2P(T > |t|) < 0.05$, we reject every H_0 for all the coefficients and accept every H_a. Thus, every coefficient in our chosen model is significant.

Confidence Interval

Our model falls within the 95% confidence interval:

Coefficient	Lower Bound (2.5%)	Our Estimate	Upper Bound (97.5%)
$\hat{\beta}_0$	$-1.108 \cdot 10^{1}$	$-1.040 \cdot 10^{1}$	-9.720
$\hat{\beta}_2$	$1.528 \cdot 10^{-8}$	$2.857 \cdot 10^{-8}$	$4.187 \cdot 10^{-8}$
$\hat{\beta}_3$	$1.884 \cdot 10^{-7}$	$3.596 \cdot 10^{-7}$	$5.308 \cdot 10^{-7}$
$\hat{\beta}_4$	$-7.095 \cdot 10^{-15}$	$-5.809 \cdot 10^{-15}$	$-4.522 \cdot 10^{-15}$

Q-Q Plot



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

3. Results

Our final model is as follows:

$$\log(Y^*) = \beta_0 + \beta_2 X_2^* + \beta_3 X_3^* + \beta_4 (X_1 X_3)^*$$

where for a specific sample:

$y_t^* = y_t - \rho y_{t-1}$

$x_{1,t}^* = x_{1,t} - \rho x_{1,t-1}$

$x_{2,t}^* = x_{2,t} - \rho x_{2,t-1}$

$(x_{1,t} x_{3,t})^* = x_{1,t} x_{3,t} - \rho x_{1,t-1} x_{3,t-1}$

$t =$ time order

$\rho =$ coefficient from the Cochrane-Orcutt procedure

Response variable (Y) is the predicted daily, new COVID-19 cases as a percentage of the population.

Explanatory variables (X_1, X_2, X_3) are the cumulative vaccine doses administered by Pfizer, Moderna and Johnson&Johnson, respectively.

For a specific prediction, we can reverse the log-transformation and derive $\hat{y}_t$ as follows:

Consider $\log(\hat{y}_t^*) = \hat{\beta}_0 + \hat{\beta}_2 x_{2,t}^* + \hat{\beta}_3 x_{3,t}^* + \hat{\beta}_4 (x_{1,t} x_{3,t})^*$

Let $\hat{\beta}_0 + \hat{\beta}_2 x_{2,t}^* + \hat{\beta}_3 x_{3,t}^* + \hat{\beta}_4 (x_{1,t} x_{3,t})^* = C$

$\log(\hat{y}_t{}^*) = C$

$\hat{y}_t{}^* = e^C$

$\hat{y}_t - \rho y_{t-1} = e^C$

$\hat{y}_t = e^C + \rho y_{t-1}$

4. Conclusion

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

Because we have more than 1 explanatory variable, predicting the combination of cumulative vaccines administered that will give us a response variable equivalent to $\frac{1}{140}$ 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 before manual surveying commences. In simple terms, we can use this model to predict Y given new Xs, but further work is needed to predict Xs given new Y.

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.*
12. Google. (2021, September 13). *COVID-19 Open Data*. Google cloud platform. Retrieved September 13, 2021, from https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-open-data.