Stat 512 Group 4 Project Cover Page

1. Project Topic: COVID-19 Infection and Death Cases In Indiana

2. Group number: 4

3. List of group members: Xiaoke He, Yuan Gao, Hao Yin, Baoxuan Tao

4. Project YouTube link: https://www.youtube.com/watch?v=UHBjoHa8wno&t=588s

5. Project background introduction:

COVID-19 has created a global pandemic, and it has dramatically impacted our daily life. There have been many measures to reduce the spread of diseases, such as vaccines and masks. We are interested in how well each of them works compared to each other. Therefore, we have raised research questions to investigate effectiveness of each measure as well as the collective effectiveness.h6. Project result highlight (what are the major findings of your project, what do you consider the most contribution of this project):

1) A full series of Covid-19 vaccines can restrict the infection rate, which may indirectly prove that a full series vaccination is able to increase the immunity to Covid-19.
2) The complete series and the booster do not have the same effects in terms of reducing the number of infections.
3) The mask mandate plays a more important role in determining the total number of death cases. When the mask mandate is lifted, the number of death cases is higher than it used to be in Indiana.

7. Project data introduction (the exact data resource, a table summarizes variable notation and definition, such as the one on the first page in the homework).

Source of data:

1. https://covid.cdc.gov/covid-data-tracker/#trends_weeklycases_select_00

2.https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2

TotalVac	The total number of administered Covid-19 vaccines in Indiana.
Doseone	The total number of administered first doses of Covid-19 vaccine in Indiana.
Series	The total number of administered Second dose of Covid-19 vaccine in Indiana.
Booster	The total number of administered Booster shots of Covid-19 vaccine in Indiana.
infection	The total number of Covid-19 infection cases in Indiana.
death	The total number of Covid-19 death cases in Indiana.
Mask	Mask Mandate Policy in IN. 0 means no Mask Mandate Policy.

8. Briefly describe what you learn from the project and what is the most challenging part.

The most challenging part is our models didn’t reach our expectations. Our expectations are very ideal. For example, we assume that vaccines take effect within a week and we can see a sudden decrease in the number of infections. Also, we assume that Indiana will follow the pattern of the general environment and didn’t consider the geographical restrictions in Indiana may lead to a different pattern.

Also, in the process of coding, we realize that real data has more severe violations than the database we used in homeworks, and these violations can't always be completely solved by remedial methods covered in our course. It’s hard for me to make the decision.

From this project, we learn a lot. First, when we are deciding research questions, we should consider whether the database we will use can handle the scale of our research questions. Besides, the standard of diagnostic treatment is not fixed. In study, we should combine the objective output we get and our human intuition to make decisions.

9. In one sentence, what is your advice for the future student to deliver a high-quality project in the course.

When your database is clearly inadequate to support your research questions in the course of your project, scaling up the database is necessary or a must.

Introduction & Background

Since the end of 2019, the spread of COVID-19, a highly infectious disease caused by the novel coronavirus, has quickly become a global pandemic. The symptoms range from mild to severe, and can lead to serious illness or even death, particularly in those with underlying health conditions. The pandemic has brought widespread disruptions to daily life and has greatly impacted the economy in the whole world.

Scientists and health officials have been working very hard to finally develop vaccines to fight the disease in December 2020. Currently, several vaccines have been authorized to vaccinate the population. In the state of Indiana, there are three types of vaccines that are mainly used, Pfizer, Moderna, and Johnson & Johnson. Taking the most vaccinated brand, Pfizer, as an example, it has, evidenced from clinical trials, a high level of efficacy. In a trial involving 43,548 participants, it has been reported that 2 doses of the vaccine confer 95% protection against Covid-19 (F. polack, et. al., 2020). There are similar results from medical trials for other vaccines as well. These results report the efficacy of a vaccine, which is the medical term to represent the effect of a vaccine in a controlled environment. However, when it comes to vaccinating the population, there are many other factors that could impact the result of the vaccines on individuals. Common factors that should be considered are the age of the individual, the family size, the frequency of travel, etc. To understand how the vaccines perform in such a complicated environment, non-medical study, or statistical study to be more specific, is then important to understand the effectiveness of vaccines. In a study called Real-world effectiveness of COVID-19 vaccines: a literature review and meta-analysis (C Zheng, et. al., 2021), the authors have studied the effectiveness of vaccines against many factors, and the results show that two-dosage vaccines are highly protective against COVID-19.

Most medical companies had developed two-dosage vaccines in the first place. However, in 2021, the variation in COVID-19 virus had restrained the effectiveness of the original two-dosage vaccines. Therefore, a third dose of booster is introduced to the public. Besides, there was another administrative decision to have people wear masks when staying away from home. It is believed that the vaccines and masks all have an impact on the number of infection cases. Moreover, the number of deaths during COVID-19 is also believed to be impacted by these factors. Therefore, our interest focuses on investigating how these factors influence each other. Since there have been studies about the effectiveness of each of these factors, our project is a confirmatory observational study. In this project, we have raised 5 research questions and come up with hypotheses accordingly. In this report, we will be using statistical knowledge in multiple linear regression to answer 4 of these questions. The result from our investigation will be reported in the end.

Measurements

All data comes from the Centers for Disease Control and Prevention. 6 continuous variables and 1 categorical variable are in the dataset. The continuous variables are Total Vaccine, First Dose, Complete Series, Booster, Infection, and Death. The categorical variable is Mask. All 6 continuous variables only record data in Indiana. The original vaccine data is the total number of administered vaccines per day. To match the unit of infection and death variables, we calculate the weekly total vaccine number by summing up every day's vaccine in one week.

Infection records the weekly infection cases in Indiana from December 31 to March 29. As the vaccine takes 1-2 weeks to take effect, we postpone the infection cases two weeks to correspond to the effective vaccine.

Death records the weekly death cases in Indiana from December 31 to March 29. Similar to Infection variables, we postpone the death cases by two weeks to correspond to the effective vaccine.

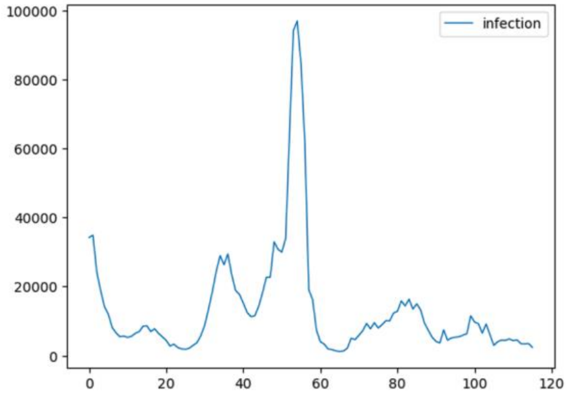
Total Vaccine is the total number of administered vaccines per week in Indiana, which includes the first dose, second dose, and both first and second booster. It records all vaccines from different providers like Pfizer-BioNTech, Moderna, Novavax, and Johnson & Johnson. However, the same people could receive multiple vaccine doses in different jurisdictions or from different providers. Thus, CDC's data may overestimate the total number of vaccines.

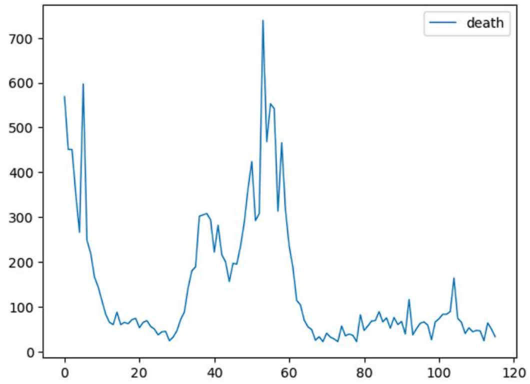
The Doseone records the total number of the first-dose vaccine in Indiana per week from different providers. The same people could receive multiple vaccine doses in different jurisdictions or from different providers. Thus, CDC's data may overestimate the total number of first dose vaccines.

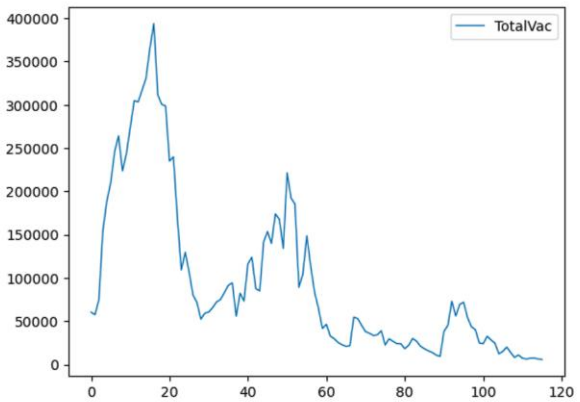
Complete Series records the total number of the second-dose vaccine in Indiana per week from different providers. The same people could receive multiple vaccine doses in different jurisdictions or from different providers. Some second doses might be recorded as the first dose, so CDC's data may underestimate the total number of second dose vaccines.

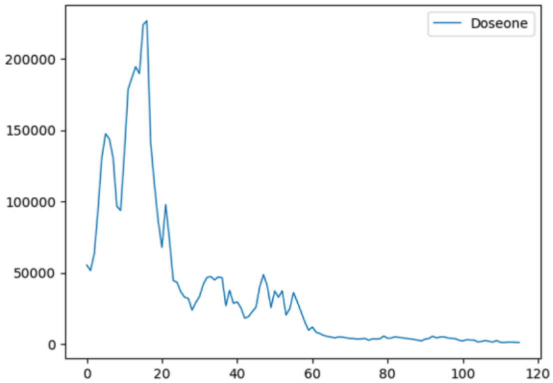
Booster records the total number of first booster dose vaccines in Indiana per week from different providers.

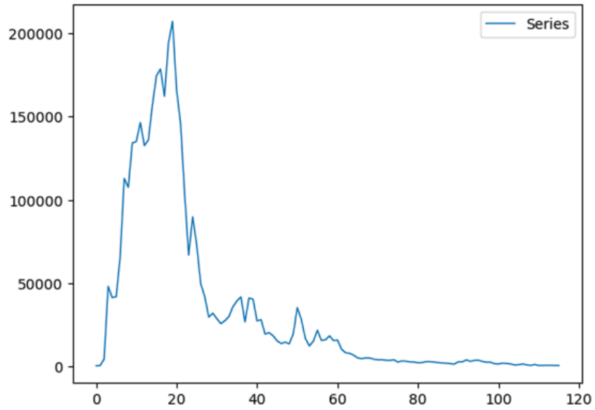
Mask is the categorical variable which includes 0 and 1. If Mask is 0, it means the mask mandate policy ended in Indiana. If Mask is 1, it means the mask mandate policy is still activated. The mask mandate policy required everyone over the age of 8 to wear a mask both indoors and outside wherever social distancing is not possible since July 27, 2020. The mask mandate ended on April 6, 2020. (IN gov, July 27)

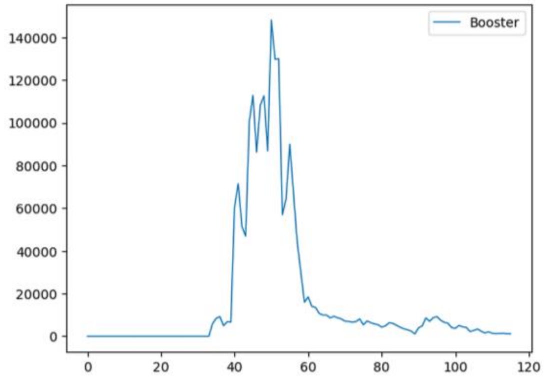


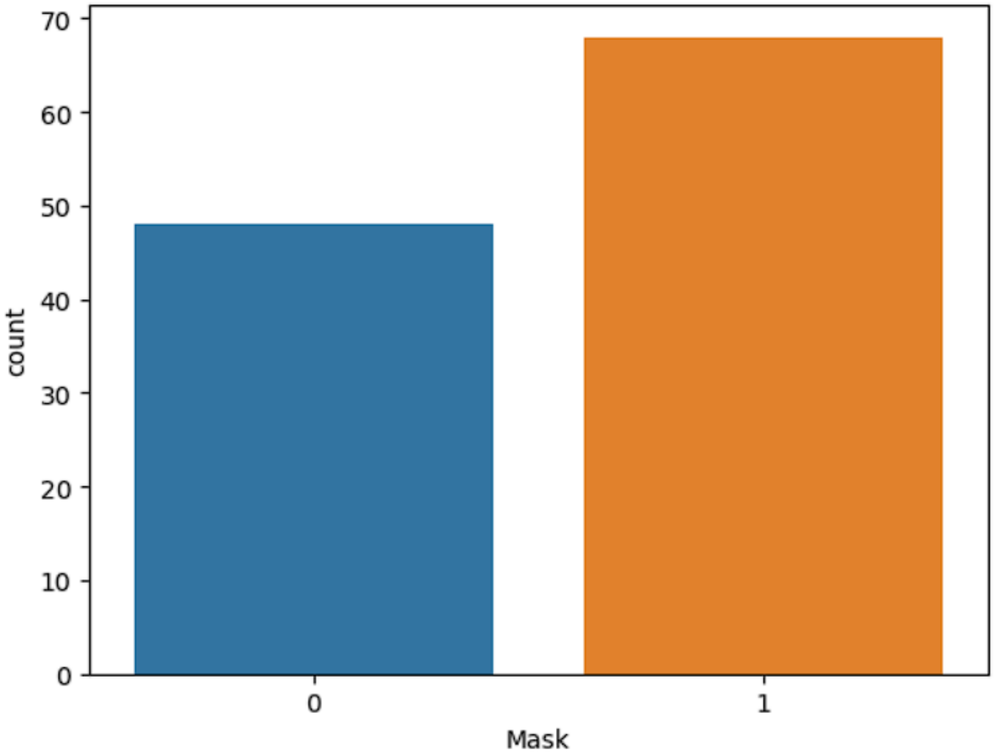



TotalVac
400000
350000
300000
250000
200000
150000
100000
50000
0
0
20
40
60
80
100
120


Doseone
200000
150000
100000
50000
0
0
20
40
60
80
100
120


Series
200000
150000
100000
50000
0
0
20
40
60
80
100
120


Booster
140000
120000
100000
80000
60000
40000
20000
0
0
20
40
60
80
100
120


70
60
50
40
30
20
10
0
count
0
1
Mask

Research Questions & Methodology

Variables(Y - Infection rate, X1- Total number of vaccine, X2 – First dose, X3 – Complete series, X4 – Booster, X5 – Death Cases, X6 – Mask)

1. Do complete series have significant impact on infected cases, given death cases and first dose?

H0: beta3 = 0 given beta2 and beta5 Ha: beta3 is not 0 given beta2 and beta5
Reduced model: Y = beta0 + beta5(X5) + beta2(X2)
Full model: Y = beta0 + beta5(X5) + beta2(X2) + beta3(X3)

2. Do complete series and first dose have the same impact on infected cases, given death cases?

H0: beta2 = beta3 = betanew Ha: beta2 is not equal to beta3
Reduced model: Y = beta0 + betanew(X2 + X3) + beta5(X5)
Full model: Y = beta0 + beta2(X2) + beta3(X3) + beta5(X5)

3. Do complete series have significant impact on the death cases given first dose, infected cases in Indiana?(Waived)

H0: beta3 = 0 given beta1, beta2 and beta4 Ha: beta3 is not 0 given beta1, beta2 and beta4
Reduced model: Y = beta0 + beta1(X1) + beta2(X2) + beta4(X4)
Full model: Y = beta0 + beta1(X1) + beta2(X2) + beta4(X4) + beta3(X3)

4. Does booster shot have the same impact on the infected cases as the complete series with first dose and death cases in the model?

H0: beta3 = beta4 = betanew Ha: beta3 is not equal to beta4
Reduced model: Y = beta0 + beta2(X2) + betanew(X3 + X4) + beta5(X5)
Full model: Y = beta0 + beta2(X2) + beta3(X3) + beta4(X4) + beta5(X5)

5. Does the number of vaccines have the same impact on the number of death cases before and after the mask mandate was lifted in Indiana?

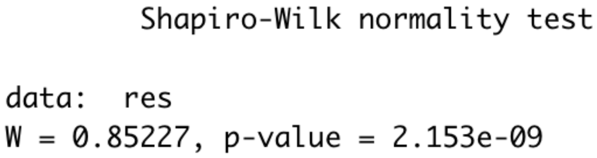
H0: beta3 = 0 Ha: beta3 is not 0
Reduced model: Y = Beta0 + beta1(X1) + beta2(X2)
Full model: Y = beta0 + beta1(X1) + beta2(X2) + beta3(X1 * X2)

Question 1 & 2(sharing data and model):

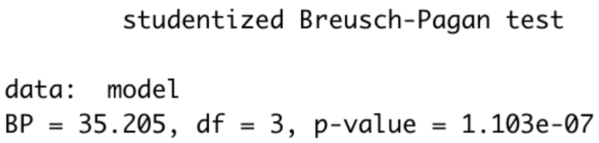
First, diagnostic tests are performed to check the model assumptions, which include Breusch-pagan test for constant variance, and Shapiro-Wilk test and normalized Q-Q plot for normality. We discover that the original data violate both the constant-variance and normality assumption. To resolve this issue, we perform transformation on both X and Y. For transformation on the independent explanatory variables, we took the log of the x variables. For transformation on the response variable, we perform Box-cox transformation and find that a lambda of 0.0303 is optimal. After the transformation, the violation on constant variance has been fixed. Though the transformed data passes the normality test, there are still some data points bending the normal Q-Q plot on the two tails, which later will be resolved using bootstrapping.

The added-variable plot shows that all three predictors each have their own marginal contribution to the model. Then, we use Studentized Residual to check for Y outliers, and discover there are 0 outliers on the Y scale. The Hat matrix is used to check for X outliers and a cutoff of 2 * p / n is used since the dataset is relatively small, and there are 3 outliers on the X scale. However, these outliers are not influential points, thus do not pose a rather significant problem: to check for influential points, we use DFBETAS to check if there are influential points affecting the linear impacts, DFFITS to check if there are influential points affecting single fitted values, and Cook's Distance to check if there are influential points affecting all fitted values as a whole, and fortunately, all three tests give 0. After this, we use the Variance Inflation Factor to check if there exists a multicollinearity issue. Though none of the VIFs exceed 10(suggesting there is no excessive multicollinearity issue), we still perform Ridge Regression to make the predictors more distinguishable. The K we use for Ridge Regression is 0.14, as VIFs for three parameters are closer to 1 when K is equal to 0.14. Building on top of the Ridge Regression, we use bootstrapping to obtain a more precise estimate of the confidence intervals of the predictors when the normal assumption may not hold.

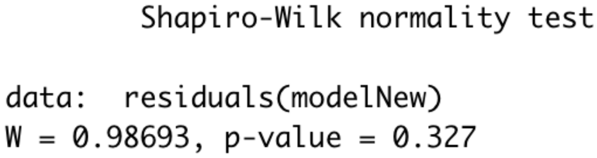
To test the predictive ability of the model, we use K-fold cross validation and obtain a Root Mean-squared Error of 0.0272. The low RMSE suggests the model behaves stably across different data and has good predictive power.



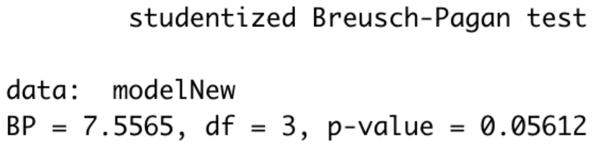
```
	Shapiro-Wilk normality test

data:  res
W = 0.85227, p-value = 2.153e-09
```



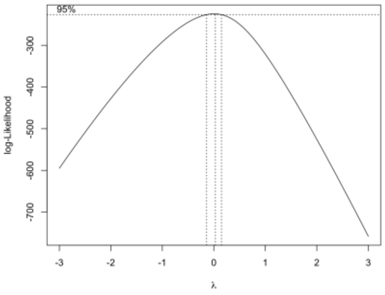
```
	studentized Breusch-Pagan test

data:  model
BP = 35.205, df = 3, p-value = 1.103e-07
```

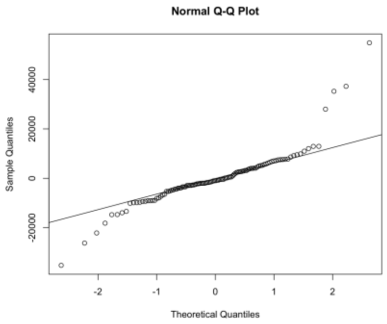


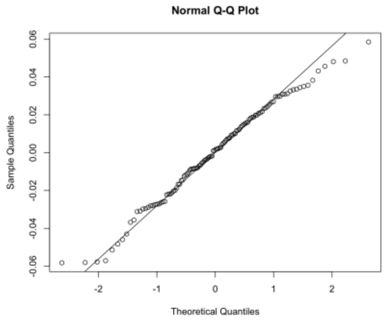
```
	Shapiro-Wilk normality test

data:  residuals(modelNew)
W = 0.98693, p-value = 0.327
```

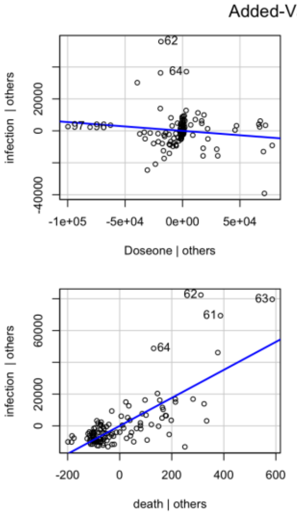


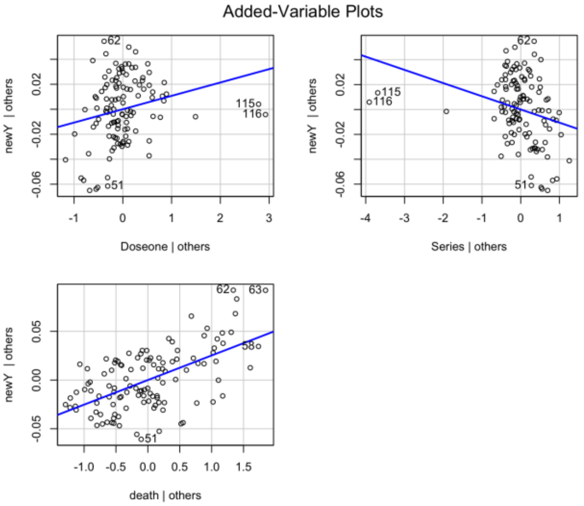
```
	studentized Breusch-Pagan test

data:  modelNew
BP = 7.5565, df = 3, p-value = 0.05612
```

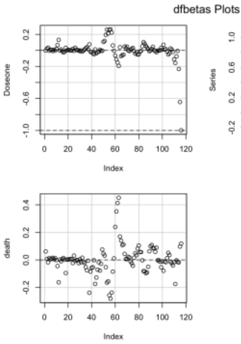


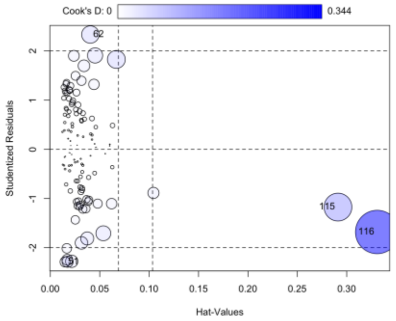


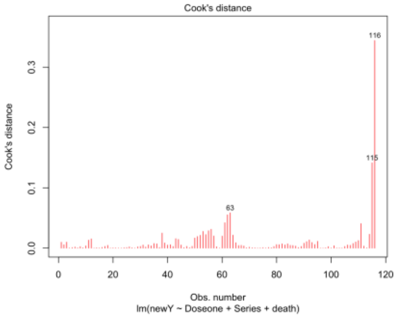


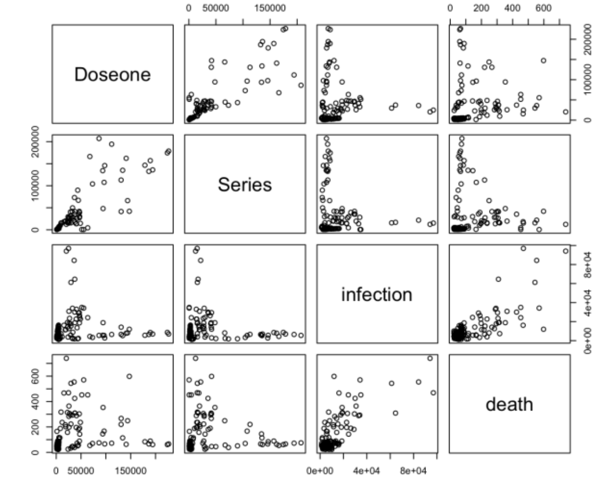

Added-Variable Plots
infection | others
Doseone | others
-1e+05
-5e+04
0e+00
5e+04
62
64
97
96
Series | others
-50000
0
50000
100000
death | others
-200
0
200
400
600
61
63
20000
0
-40000
60000

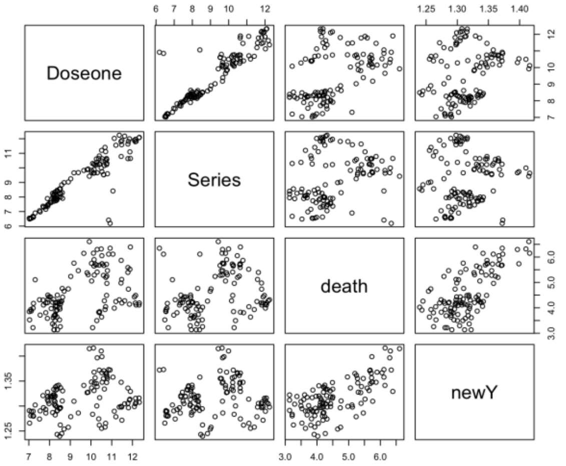

Added-Variable Plots
newY | others
Doseone | others
Series | others
death | others
62
51
115
116
63
58
0.02
-0.02
-0.06
0.05
0.00
-0.05
-1
0
1
2
3
-4
-3
-2
-1.0
-0.5
0.0
0.5
1.0
1.5

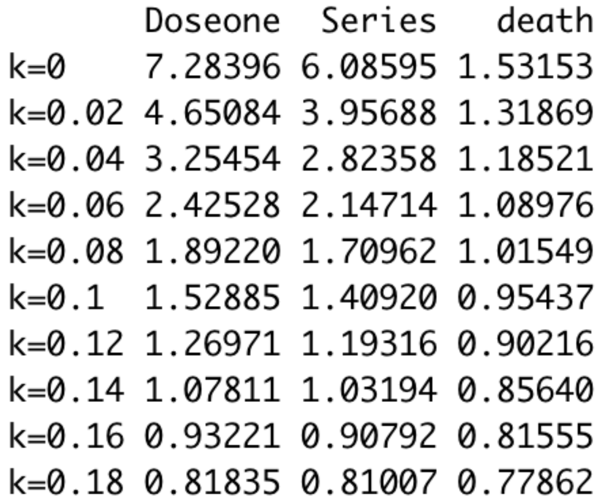

dfbetas Plots
Doseone
Series
death
Index
0
20
40
60
80
100
120


Cook's D: 0
0.344
Studentized Residuals
Hat-Values
62
115
116
0.00
0.05
0.10
0.15
0.20
0.25
0.30

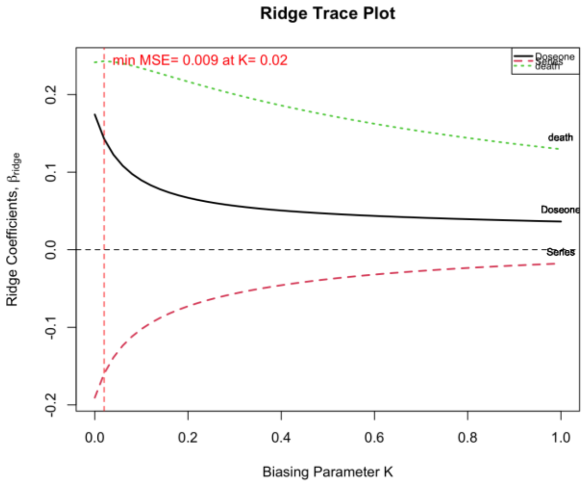

Cook's distance
Cook's distance
Obs. number
lm(newY ~ Doseone + Series + death)
116
115
63
0.0
0.1
0.2
0.3
0
20
40
60
80
100
120

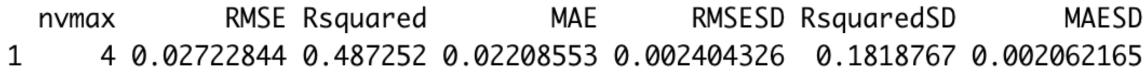

Doseone
Series
infection
death
0
50000
150000
200
400
600
100000
200000
0e+00
4e+04
8e+04


Doseone
Series
death
newY
6
7
8
9
10
11
12
1.25
1.30
1.35
1.40
3.0
4.0
5.0
6.0



```
        Doseone  Series   death
k=0     7.28396 6.08595 1.53153
k=0.02 4.65084 3.95688 1.31869
k=0.04 3.25454 2.82358 1.18521
k=0.06 2.42528 2.14714 1.08976
k=0.08 1.89220 1.70962 1.01549
k=0.1  1.52885 1.40920 0.95437
k=0.12 1.26971 1.19316 0.90216
k=0.14 1.07811 1.03194 0.85640
k=0.16 0.93221 0.90792 0.81555
k=0.18 0.81835 0.81007 0.77862
```





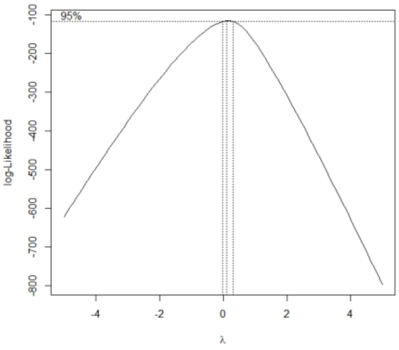
```
  nvmax        RMSE Rsquared         MAE      RMSESD RsquaredSD       MAESD
1     4 0.02722844 0.487252 0.02208553 0.002404326  0.1818767 0.002062165
```

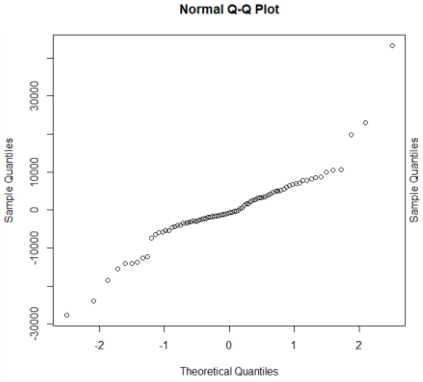
Question 4:

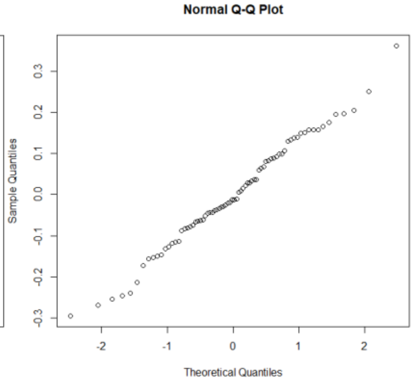
The full model is tested for the linear regression model assumptions: constance of error terms, normality of error variance, and the existence of outliers. The first test is the Breuch-pagan test for constancy of error, and the result shows a p-value much smaller than 0.05. Therefore, the full model has the problem of non-constant error. The second test is the Shapiro test to verify the normality of the error variance. The p-value from the Shapiro test is also smaller than 0.05. Therefore, the distribution of residuals is far away from normal. A Q-Q plot is also used to verify the result from the Shapiro test. The points are deviated away from a straight line. The upper tail and the lower tail shows that there are more residuals towards the extreme values than there should be if the distribution is normal, which verifies the conclusion from the Shapiro test. The number of outliers were counted. If the absolute value of studentized residuals are larger than 2, the data point is considered an Y outlier. There are 6 Y outliers in total. To mitigate the problem of assumption violation, a transformation on X is performed first. The outliers are removed beforehand. Based on the new predictor variables, a new model is built, and Box-Cox transformation procedure is followed to transform y to better eliminate the problem of non-constant and non-normal error in the model. The optimal lambda value found is 0.1. The tests are performed again on the transformed model to check the existence of assumption violation: non-constant variance and non-normal residuals. The result of both tests show a p-value larger than 0.05. Thus, the transformed model has successfully overcome the problems of non-constant variance and non-normal residuals. A Q-Q plot is also used to observe the normality in error distribution. It can be seen that the residuals form a line, which means normality in residual distribution is achieved.

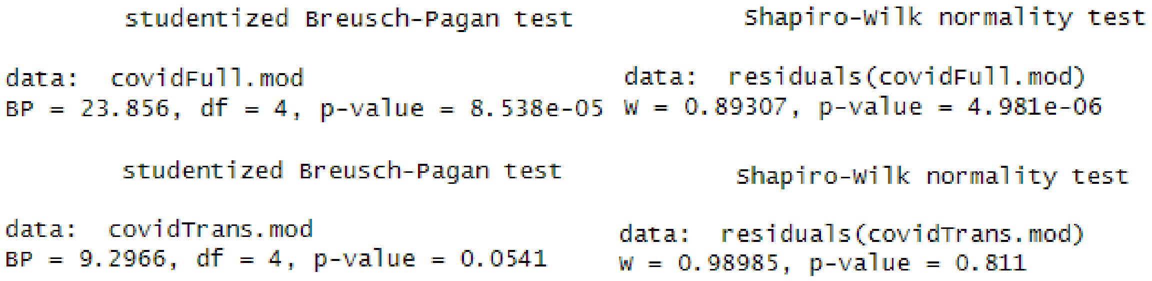
More advanced diagnostic tests are performed to evaluate the transformed model. The added-variable plots of the transformed model are used to evaluate the marginal effect when each predictor is added to the model when all other predictors are already in the model. If any predictor shows a 0 marginal effect, it can be removed from the model to reduce complexity while preserving the predictability of the model. It turns out that none of the predictors has 0 marginal effect, so they are all necessary to be kept in the model. The number of influential points in the model is checked. DFBETAS, DFFITS, and Cook's Distance are performed separately. The results show that there are no influential points according to DFBETAS, 1 according to DFFITS,

0 minor influential points and 0 major influential points according to Cook’s Distance. From the Cook’s Distance graph, it can also be seen that the influence of individual points is very limited. Multicollinearity is also checked in this model. A graph plotting untransformed variables is used to visually determine whether there exists multicollinearity issue, then the VIF values are calculated to numerically check for multicollinearity. The graph does not show a strong linear relationship between any pair of predictors, and the VIF values are all smaller than 10. However, one of the combinations shows a VIF value as high as 9, so ridge regression is then used to study the 0.95 confidence interval of the coefficients with minimal impact from the issue of multicollinearity. Ridge regression has shown that K=0.12 is a good fit for the ridge model.
A K-fold cross correlation is performed to evaluate the predictability of the model. The Root Mean-squared error in the original full model is 9634, and 0.1352 in the transformed model. Therefore, the transformation performed reduces a large amount of error variance and improves the model’s ability to make new predictions.







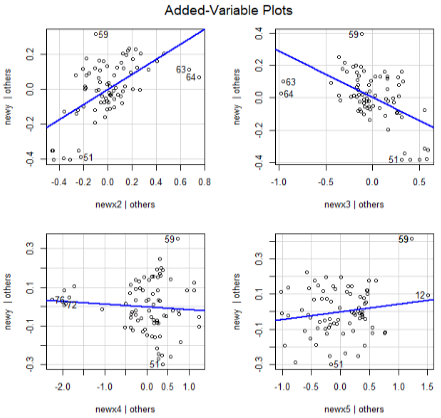


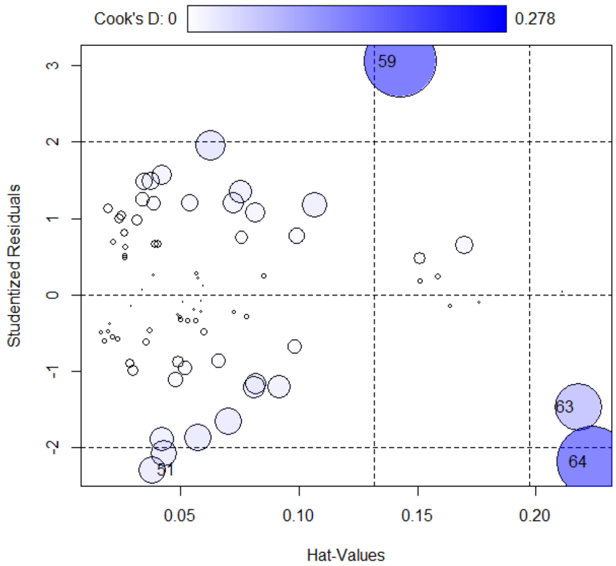
```
	studentized Breusch-Pagan test

data:  covidFull.mod
BP = 23.856, df = 4, p-value = 8.538e-05
```

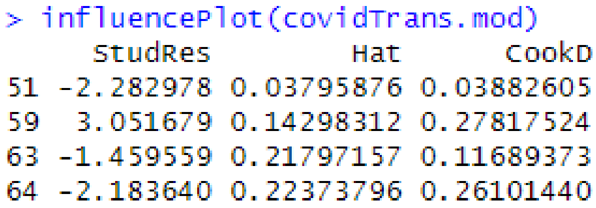
```
	Shapiro-Wilk normality test

data:  residuals(covidFull.mod)
W = 0.89307, p-value = 4.981e-06
```

```
	studentized Breusch-Pagan test

data:  covidTrans.mod
BP = 9.2966, df = 4, p-value = 0.0541
```

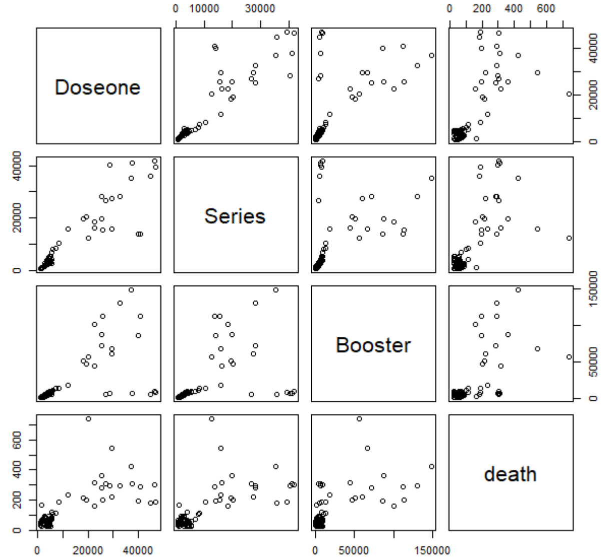
```
	Shapiro-Wilk normality test

data:  residuals(covidTrans.mod)
W = 0.98985, p-value = 0.811
```

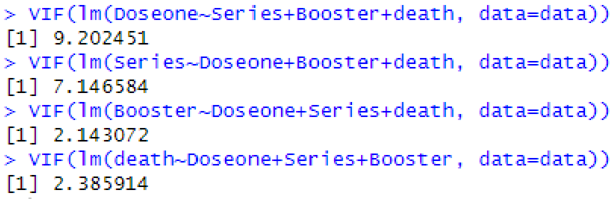




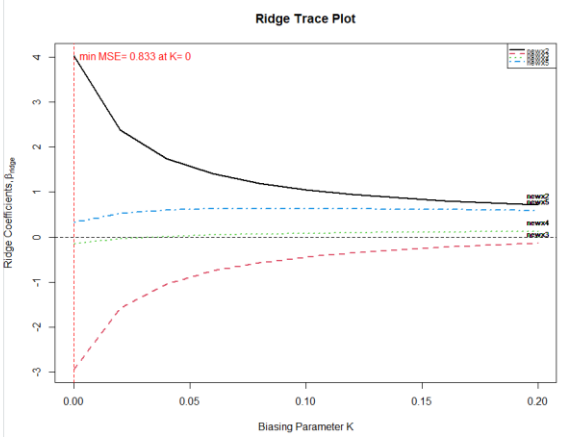


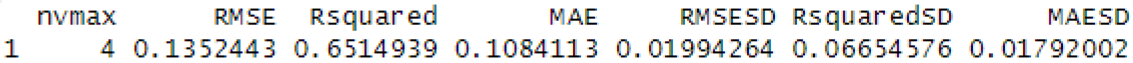
```
> influencePlot(covidTrans.mod)
     StudRes        Hat      CookD
51 -2.282978 0.03795876 0.03882605
59  3.051679 0.14298312 0.27817524
63 -1.459559 0.21797157 0.11689373
64 -2.183640 0.22373796 0.26101440
```





```
> VIF(lm(Doseone~Series+Booster+death, data=data))
[1] 9.202451
> VIF(lm(Series~Doseone+Booster+death, data=data))
[1] 7.146584
> VIF(lm(Booster~Doseone+Series+death, data=data))
[1] 2.143072
> VIF(lm(death~Doseone+Series+Booster, data=data))
[1] 2.385914
```





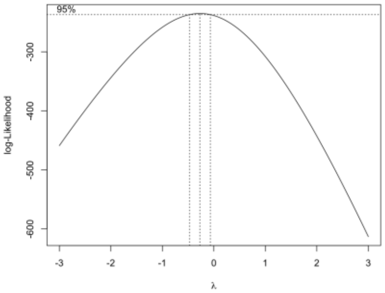
```
  nvmax      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
1     4 0.1352443 0.6514939 0.1084113 0.01994264 0.06654576 0.01792002
```

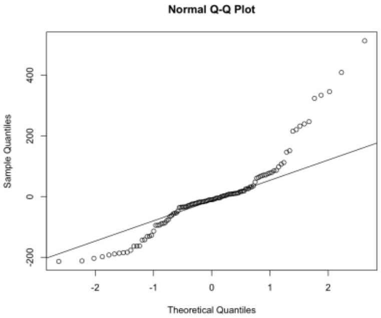
Question 5:

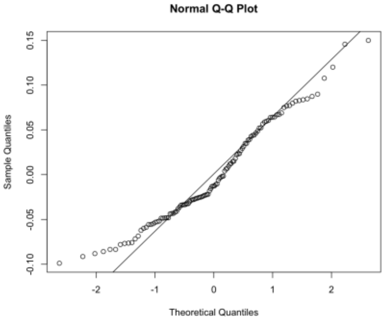
The first step is to check the assumptions of the model. We perform Breusch-pagan test for constant variance check, and the Shapiro-Wilk test and the normalized Q-Q plot for normality check. Both constant-variance and normality assumptions are violated, and we need transformation on both X and Y .We took the log of the x variables as the X transformation. We perform Box-cox transformation and select the optimal value of -0.27272 as lambda as the Y transformation. After the transformation, both issues show an improvement, but the problem is still not resolved by transformation.

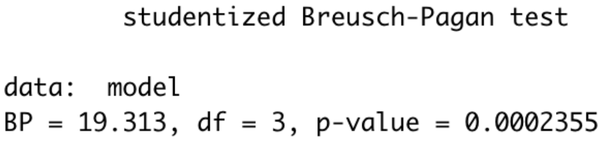
From added-variable plots, we find that mask mandate has a relatively clear added-on effect, but the total number of vaccines and the interaction of the number of vaccines and mask mandate does not provide much additional information. According to the Hat matrix, the data has 13 outliers on the X scale. According to the Studentized deleted residual, there are no outliers on the Y scale. Through checking DFBETAS, DFFITS, and Cook's distance, no influential points are found. However, the influential plot shows there are 4 data points which are very likely to be influential points, and it seems that the X outliers are not that different from the rest of the data points to be posing potential problems. To reduce the effect of multicollinearity caused by the interaction effect, we choose to use Ridge Regression. According to VIF, the model performs optimally when K = 0.06..

We perform bootstrapping to account for unresolved nonnormality issues. By using K-fold cross validation, the Root Mean-squared Error of the model is 0.05637, indicating a good predicting ability.

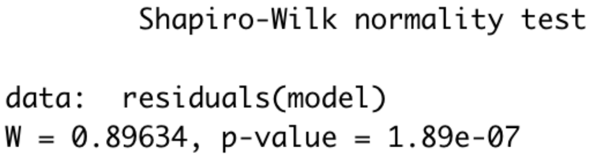




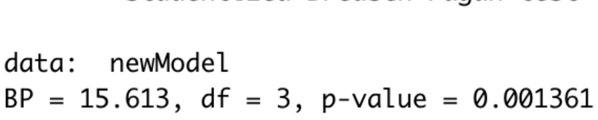




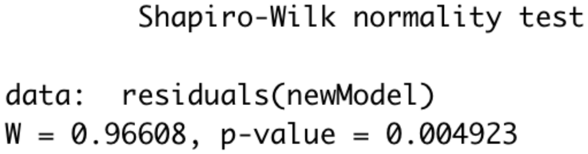
```
        studentized Breusch-Pagan test

data:  model
BP = 19.313, df = 3, p-value = 0.0002355
```



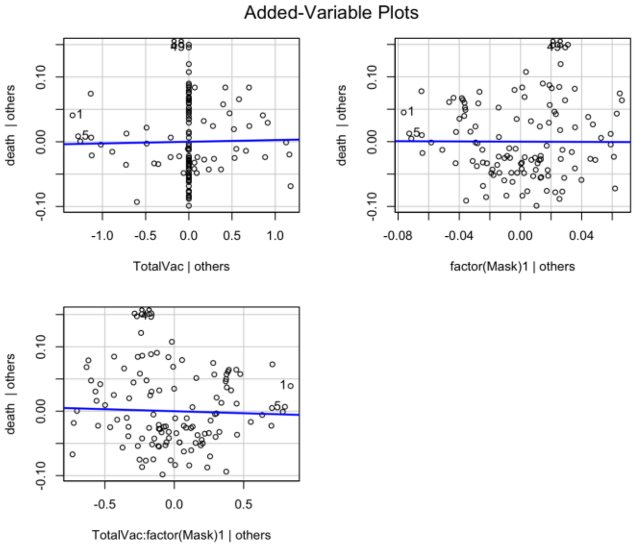
```
        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.89634, p-value = 1.89e-07
```

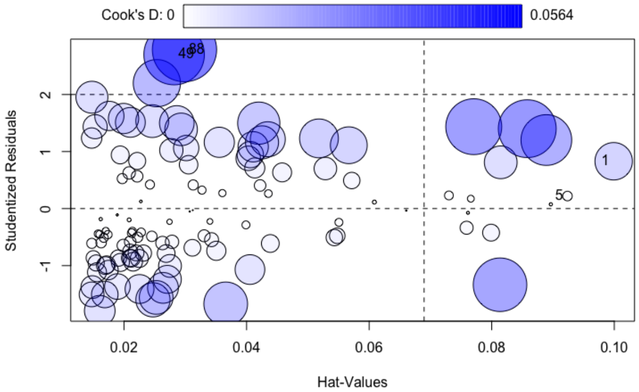


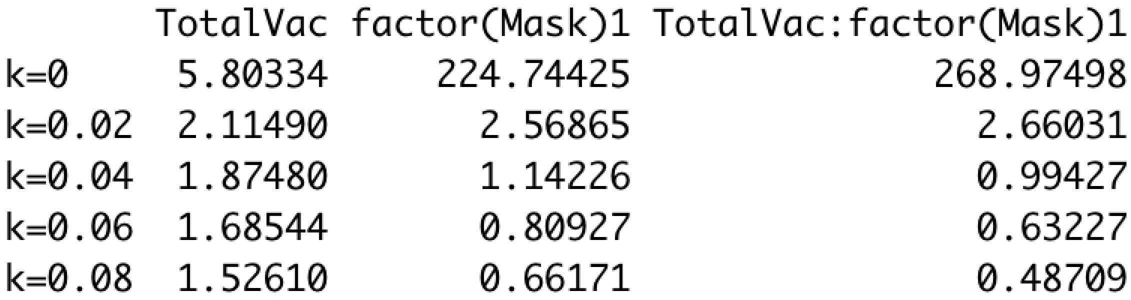
```
        studentized Breusch-Pagan test

data:  newModel
BP = 15.613, df = 3, p-value = 0.001361
```

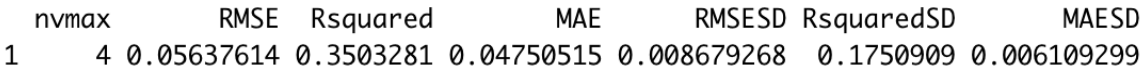


```
        Shapiro-Wilk normality test

data:  residuals(newModel)
W = 0.96608, p-value = 0.004923
```






```
        TotalVac factor(Mask)1 TotalVac:factor(Mask)1
k=0      5.80334     224.74425              268.97498
k=0.02   2.11490       2.56865                2.66031
k=0.04   1.87480       1.14226                0.99427
k=0.06   1.68544       0.80927                0.63227
k=0.08   1.52610       0.66171                0.48709
```


```
  nvmax       RMSE  Rsquared        MAE      RMSESD RsquaredSD       MAESD
1     4 0.05637614 0.3503281 0.04750515 0.008679268  0.1750909 0.006109299
```

Discussions

Question 1 & 2(sharing data and model):

The full model for research question 1 and 2 both contains three predictors, first dose, complete series, and death cases, each with a different linear impact. In this model, the linear impact of the first dose, surprisingly, is positive(the linear impact obtained by OLS and Ridge Regression and the confidence interval obtained by both bootstrapping is strictly above 0), suggesting that the first dose does not help in lowering infection rate. This may be due to several reasons. First of all, the amount of data is insufficient and limited to the state of Indiana. The scope needs to be expanded to all parts of the world to draw further conclusions. The second point is that it is uncertain whether a two-week delay is enough for the first dose to manifest. In addition, Covid-19 virus is notorious for its fast mutation, so it may be the case that the virus has already mutated by the time that dose one is taken, but this is beyond the scope of our study.

Our first research question is interested in the linear impact of a complete series on infection cases after having the first dose of vaccine. The null hypothesis is that the complete series has no significant linear impact on the number of infected cases. The reduced model contains two predictors, death cases and first dose.

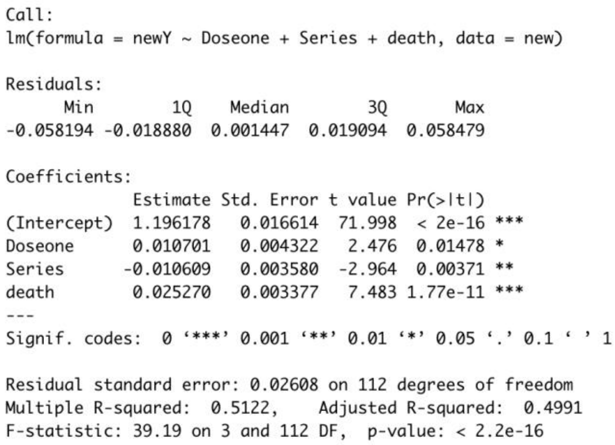
For the F-test of OLS, we obtained a test statistic of 8.854, which is much larger than the critical value of 3.926. The p-value is 0.00358, much smaller than our predetermined alpha of 0.05. Based on these statistics, we reject the null hypothesis, as we have sufficient evidence that the complete series has a significant impact on infected cases.

The second research question focuses on analyzing the difference of the linear impact between one dose and a complete series, given that the model contains the death case as a predictor. The null hypothesis is that the first dose and complete series have the same impact on the number of infected cases. The full model contains three predictors, first dose, complete series, and death cases, each with a different linear impact. The reduced model contains two predictors, death cases, and a predictor that combines first dose and complete series together to check if they have the same effect by checking if they have the same linear impacts.

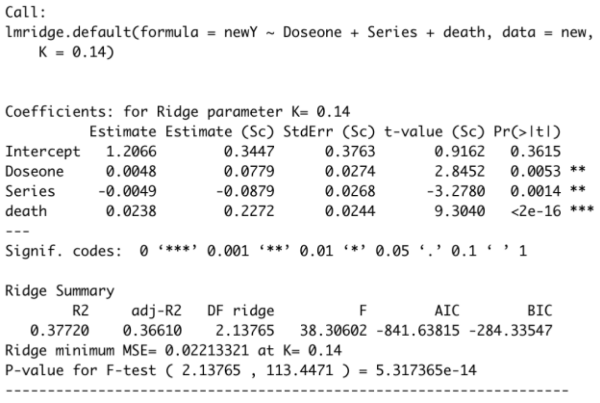
For the F-test of OLS, we obtained a test statistic of 7.627, which is much larger than the critical value of 3.926. The p-value is 0.00672, much smaller than our predetermined alpha of 0.05. Based on these statistics, we reject the null hypothesis, as we have sufficient evidence that the first dose and complete series have different impacts on infected cases.

The linear impact of the complete series is negative(the linear impact obtained by OLS and Ridge Regression and the confidence intervals obtained by both bootstrapping is strictly below 0), which indicates that taking a complete series is more effective than taking only a single dose for COVID-19 prevention.

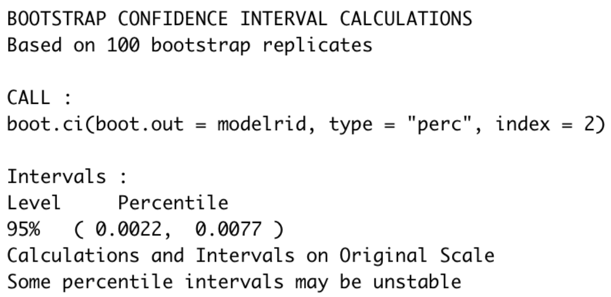
Taking together research question 1 and 2, we can conclude that a full series of Covid-19 vaccines can restrict the infection rate, which may indirectly prove that a full series vaccination is able to increase the immunity to Covid-19.



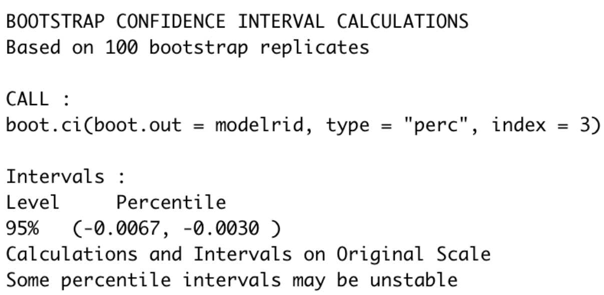
```
Call:
lm(formula = newY ~ Doseone + Series + death, data = new)

Residuals:
      Min        1Q    Median        3Q       Max
-0.058194 -0.018880  0.001447  0.019094  0.058479

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.196178   0.016614  71.998  < 2e-16 ***
Doseone      0.010701   0.004322   2.476  0.01478 *
Series      -0.010609   0.003580  -2.964  0.00371 **
death        0.025270   0.003377   7.483 1.77e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02608 on 112 degrees of freedom
Multiple R-squared:  0.5122,	Adjusted R-squared:  0.4991
F-statistic: 39.19 on 3 and 112 DF,  p-value: < 2.2e-16
```



```
Call:
lmridge.default(formula = newY ~ Doseone + Series + death, data = new,
    K = 0.14)


Coefficients: for Ridge parameter K= 0.14
          Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept   1.2066        0.3447      0.3763       0.9162   0.3615
Doseone     0.0048        0.0779      0.0274       2.8452   0.0053 **
Series     -0.0049       -0.0879      0.0268      -3.2780   0.0014 **
death       0.0238        0.2272      0.0244       9.3040   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary
        R2     adj-R2   DF ridge           F        AIC        BIC
   0.37720    0.36610    2.13765    38.30602 -841.63815 -284.33547
Ridge minimum MSE= 0.02213321 at K= 0.14
P-value for F-test ( 2.13765 , 113.4471 ) = 5.317365e-14
-------------------------------------------------------------------
```



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates

CALL :
boot.ci(boot.out = modelrid, type = "perc", index = 2)

Intervals :
Level     Percentile
95%   ( 0.0022,  0.0077 )
Calculations and Intervals on Original Scale
Some percentile intervals may be unstable
```



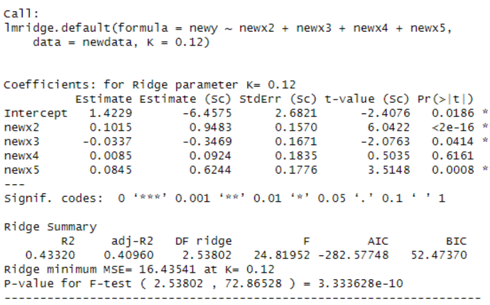
```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates

CALL :
boot.ci(boot.out = modelrid, type = "perc", index = 3)

Intervals :
Level     Percentile
95%   (-0.0067, -0.0030 )
Calculations and Intervals on Original Scale
Some percentile intervals may be unstable
```

Question 4:

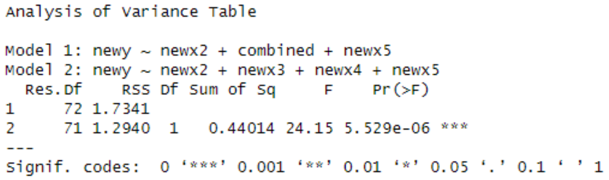
The fourth research question investigates the difference between the effect of the booster shot and a complete series in terms of reducing the number of infections. Since there are cases of booster shot variable being 0, these cases are removed for model building to compare the effect of booster shot and the completed series. The full model contains four predictors, the first dose, the booster shot, the complete series, and the number of death cases. The reduced model contains three predictors, the first dose, the number of death cases, and a combined predictor of complete series and booster shot to check if they share a linear impact.

The result of the ridge model shows a scaled confidence interval of (-0.6801, -0.0137) for the coefficient of the complete series and (-0.3574, 0.3744) for the coefficients of the booster. From ridge regression, we can conclude that the impact of the booster is very small, but the complete series has a negative impact on the number of infections. Although the 0.95 confidence intervals for the two coefficients are overlapped, we can not conclude that the booster has similar effectiveness as the complete series.

An F-test is performed to compare the reduced model to the full model. An ANOVA table for both models is built, and the test statistic is calculated to be 24.14986. The critical value is calculated to be 3.97581 on a 0.95 confidence level. Since the test statistic is larger than the critical value and the p-value is close to 0, the null hypothesis is rejected, and we can conclude that there is large discrepancy in the linear impact between the complete series and the booster. We come to the conclusion that the complete series and the booster do not have the same effects in terms of reducing the number of infections.



```
Call:
lmridge.default(formula = newy ~ newx2 + newx3 + newx4 + newx5,
    data = newdata, K = 0.12)


Coefficients: for Ridge parameter K= 0.12
          Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept   1.4229       -6.4575      2.6821      -2.4076   0.0186 *
newx2       0.1015        0.9483      0.1570       6.0422   <2e-16 ***
newx3      -0.0337       -0.3469      0.1671      -2.0763   0.0414 *
newx4       0.0085        0.0924      0.1835       0.5035   0.6161
newx5       0.0845        0.6244      0.1776       3.5148   0.0008 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary
        R2      adj-R2    DF ridge           F          AIC         BIC
   0.43320     0.40960     2.53802    24.81952 -282.57748    52.47370
Ridge minimum MSE= 16.43541 at K= 0.12
P-value for F-test ( 2.53802 , 72.86528 ) = 3.333628e-10
-------------------------------------------------------------------
```

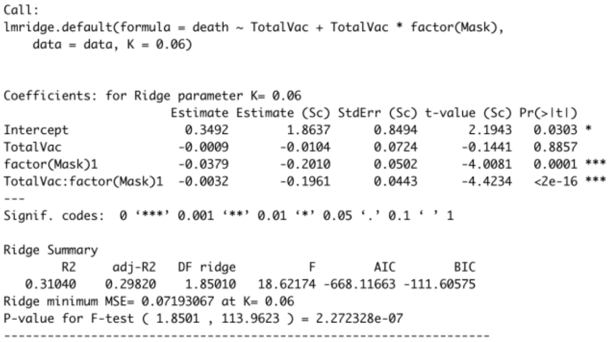


```
Analysis of Variance Table

Model 1: newy ~ newx2 + combined + newx5
Model 2: newy ~ newx2 + newx3 + newx4 + newx5
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     72 1.7341
2     71 1.2940  1   0.44014 24.15 5.529e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

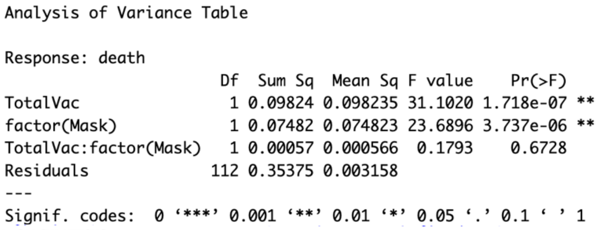
Question 5:

The fifth research question looks into the difference of the linear impact the number of vaccines has on the number of death cases before and after the mask mandate was lifted in Indiana. The null hypothesis is that the linear impact of the number of vaccines has on the number of death cases before the mask mandate was lifted in Indiana is the same as after. The full model contains 4 predictors, the number of vaccines, the indicator variables of mask mandate, and the interaction terms between the former two.

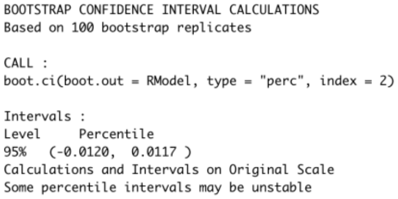
From the summary of the Ridge Regression and OLS, we have a very interesting finding: both the mask mandate and the interaction factor of the number of vaccines and mask mandate are significant, but the number of vaccines is not significant. Combining the result we got from the added-variable plot that the number of vaccines does not provide any additional information to the model, we can see its insignificance for explaining the number of death cases. The mask mandate plays a more important role in determining the total number of death cases. When the mask mandate is lifted, the number of death cases is higher than it used to be in Indiana. The confidence interval of the number of vaccines provided by bootstrapping includes 0, which confirms its insignificance. The confidence interval of the mask mandate and the interaction term between the number of vaccines and mask mandate are both strictly below 0, indicating both predictors contribute to the reduction of the number of death cases.



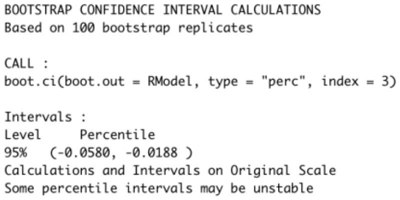
```
Call:
lmridge.default(formula = death ~ TotalVac + TotalVac * factor(Mask),
    data = data, K = 0.06)


Coefficients: for Ridge parameter K= 0.06
                     Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept              0.3492        1.8637      0.8494       2.1943   0.0303 *
TotalVac              -0.0009       -0.0104      0.0724      -0.1441   0.8857
factor(Mask)1         -0.0379       -0.2010      0.0502      -4.0081   0.0001 ***
TotalVac:factor(Mask)1 -0.0032      -0.1961      0.0443      -4.4234   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary
        R2      adj-R2    DF ridge           F          AIC         BIC
   0.31040     0.29820     1.85010    18.62174 -668.11663 -111.60575
Ridge minimum MSE= 0.07193067 at K= 0.06
P-value for F-test ( 1.8501 , 113.9623 ) = 2.272328e-07
-------------------------------------------------------------------
```



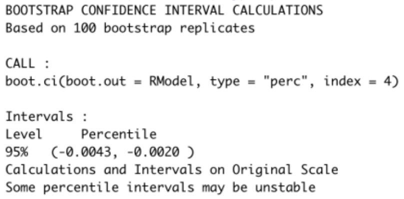
```
Analysis of Variance Table

Response: death
                        Df  Sum Sq  Mean Sq F value    Pr(>F)
TotalVac                 1 0.09824 0.098235 31.1020 1.718e-07 ***
factor(Mask)             1 0.07482 0.074823 23.6896 3.737e-06 ***
TotalVac:factor(Mask)    1 0.00057 0.000566  0.1793    0.6728
Residuals              112 0.35375 0.003158
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates

CALL :
boot.ci(boot.out = RModel, type = "perc", index = 2)

Intervals :
Level     Percentile
95%   (-0.0120,  0.0117 )
Calculations and Intervals on Original Scale
Some percentile intervals may be unstable
```



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates

CALL :
boot.ci(boot.out = RModel, type = "perc", index = 3)

Intervals :
Level     Percentile
95%   (-0.0580, -0.0188 )
Calculations and Intervals on Original Scale
Some percentile intervals may be unstable
```



```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates

CALL :
boot.ci(boot.out = RModel, type = "perc", index = 4)

Intervals :
Level     Percentile
95%   (-0.0043, -0.0020 )
Calculations and Intervals on Original Scale
Some percentile intervals may be unstable
```

Conclusion

In this project, we use linear regression models to determine how factors including the first dose of vaccine, the complete series of vaccines (second dose), the booster, the total number of vaccines people take, and the categorical factor like whether the mask mandate was lifted in Indiana affects the number of infections and the number of deaths.

Our study is a confirmatory observational study. One past study showed that the vaccine effectiveness against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, COVID-19-related hospitalization, admission to the intensive care unit, and death was over 85% (Zheng et al. 2022). Also, there was a study that has already proved that mask wearing reduces transmissibility per contact by reducing transmission of infected respiratory particles in both laboratory and clinical contexts (Howard et al., 2021).

According to conclusions of past studies, our group expected to see a reduction in the number of infections after people take the first dose of COVID-19 vaccine, and continuous reduction when people take the second dose. As we include booster as a factor, we also expected booster to further reduce the number of infections in Indiana. Besides, since COVID-19 VE against death is also very high, we expected to see the same effect of vaccines before and after the mask mandate was lifted which can further confirm that vaccines do have a strong effectiveness in preventing COVID-19 even when people didn't wear a mask. Also, through wearing masks and getting vaccinated, we also expected to see a much lower number of death cases before the mask mandate was lifted.

The obtained results showed that all questions all have a very small value of RMSE indicating that our models have good predictive power. But, just as we mentioned in the discussion part, not all the conclusions reached our expectations. First, there is no reduction in the number of infections after taking the first dose. Second, the booster has a smaller influence on reducing the number of infections than the first two doses. Third, when adding the factor of mask mandate into the model which includes the number of vaccines, vaccines didn't play any significant role in reducing the number of death cases. Instead, the model confirmed the effectiveness of wearing a mask. The number of death cases was lower before the mask mandate was lifted.

After our group discussion, our members believe that the reason is that our current database was unable to take into account all the influencing factors such as variability of virus, regional limitations, the fact that vaccines will not take effect immediately. Variability of virus was very quick especially for COVID-19, so it's impossible to consider it by mere statistical models or at least by linear regression model. To improve the other two problems , we can scale up our database in two directions. The first one is expanding our research from Indiana to the United States which is a "vertical extension" to remove the restrictions of only one state, and the second one is extending the time of data from several weeks to years which is a "horizontal extension" that gives vaccines plenty of time to take effect.

Reference

Executive order 20-37: Face covering requirement - Indiana. (n.d.). Retrieved April 23, 2023, from https://www.in.gov/gov/files/Executive%20Order%2020-37%20Face%20Covering%20Requirement.pdf

Novel Coronavirus Vaccine Provider Resources. (n.d.). Retrieved April 25, 2023, from https://www.coronavirus.in.gov/vaccine/vaccine-provider-resources/

Zheng, C., Shao, W., Chen, X., Zhang, B., Wang, G., & Zhang, W. (2022). Real-world effectiveness of covid-19 vaccines: A literature review and meta-analysis. International Journal of Infectious Diseases, 114, 252–260. https://doi.org/10.1016/j.ijid.2021.11.009

Howard, J., Huang, A., Li, Z., Tufekci, Z., Zdimal, V., van der Westhuizen, H.-M., von Delft, A., Price, A., Fridman, L., Tang, L.-H., Tang, V., Watson, G. L., Bax, C. E., Shaikh, R., Questier, F., Hernandez, D., Chu, L. F., Ramirez, C. M., &amp; Rimoin, A. W. (2021). An evidence review of face masks against covid-19. Proceedings of the National Academy of Sciences, 118(4). https://doi.org/10.1073/pnas.2014564118

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., … Gruber, W. C. (2020). Safety and efficacy of the BNT162B2 mrna covid-19 vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615. https://doi.org/10.1056/nejmoa2034577