Modeling Report

Introduction:

Our research question investigates the potential association between major economic indexes and the COVID-19 death rate. Specifically, we seek to understand how economic and social conditions may impact individuals' health outcomes, particularly during a large-scale pandemic period. To achieve this goal, we employ OLS as our model, which will allow us to estimate and interpret the effect of our explanatory variable and evaluate the model's goodness of fit. The importance of our analysis lies in its ability to shed light on the relationship between COVID-19 and social indexes. By doing so, we can better understand the disease, track its spread, inform public health policies, and develop effective treatments and vaccines. Ultimately, we hope our findings will contribute to future efforts to combat pandemics and ensure public health and safety.

Additionally, we would like to mention that we are using a different data frame than the previous one for EDA. Therefore, we will introduce and specify the meaning for each column in the data frame:

"**emp**": This column represents the Employment level for all workers.

"**res**": This column represents Residential.

"**ts**": This column represents Transit Station.

"**rev_retail**": This column represents the percent change in retail businesses, specifically using NAICS 2-digit codes 44-45.

"**init_claim_rate**": This column represents the number of initial claims per 100 people in the 2019 labor force, combining Regular and PUA (Pandemic Unemployment Assistance) claims.

"**if_good_healthcare**": This column indicates whether the state is among the top 10 most popular states in the US for healthcare.

"**if_large_population**": This column indicates whether the state is among the top 10 states in terms of population size.

"**log_y**": This column represents the logarithmic transformation of y.

Modeling Tool Motivations:

ANOVA provides insight into the statistical significance of the overall model and individual predictor variables. With the addition of two category features, we aim to determine if these additions are necessary for our final model.

Model selection involves ensuring that all variables in our model are necessary, selecting a model that includes only significant features, and validating our previous model choice as the best option.

Model diagnostics play a role in identifying outliers, assessing their influence on coefficients and predictions, and verifying if the assumptions for using OLS hold true.

Modeling Tool Assumptions:

ANOVA requires four assumptions for unbiased and accurate results: (1) Normality, where the dependent variable should be normally distributed within each group; (2) Homogeneity of variance, indicating that the variance of the dependent variable should be equal across groups;

(3) Independence, ensuring that the observations are independent of each other; and (4) Random sampling, meaning the observations are randomly sampled from the population of interest.
Model selection using AIC relies on several assumptions: (1) The models being compared are nested, meaning one model is a special case of another obtained by imposing constraints; (2) The models fit the same data; (3) The likelihood function is correctly specified; (4) There are no missing data in the models; and (5) The sample size is large.
Model diagnostics are conducted to assess whether these assumptions hold true.

death_rate <dbl>	emp <dbl>	res <dbl>	ts <dbl>	rev_retail <dbl>	init_claim_rate <dbl>	if_good_healthcare <lgl>	if_large_population <lgl>	log_y <dbl>	log_icr <dbl>
14.6	-0.19150	0.14426667	-0.29210000	-0.18850	3.66250	FALSE	FALSE	2.681022	1.298145974
63.7	-0.17420	0.09455806	-0.10364839	0.21024	1.63200	FALSE	FALSE	4.154185	0.489806257
109.0	-0.13250	0.06871333	0.01790267	0.41850	0.99375	FALSE	FALSE	4.691348	-0.006269613
165.0	-0.14300	0.07170645	0.03944774	0.38275	1.04150	FALSE	FALSE	5.105945	0.040661982
267.0	-0.13975	0.06174839	-0.01212645	0.25900	0.53260	FALSE	FALSE	5.587249	-0.629984606

Presentation and interpretation of results
Tool 1: Categorical variables / ANOVA

```
Call:
lm(formula = log_y ~ emp + res + ts + rev_retail + if_good_healthcare +
    if_large_population + log_icr, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4366 -0.5280  0.0428  0.6039  2.9367

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.05699    0.12791  47.354  < 2e-16 ***
emp                     7.19637    0.63683  11.300  < 2e-16 ***
res                    -6.96258    1.54893  -4.495 7.98e-06 ***
ts                     -2.53264    0.25792  -9.820  < 2e-16 ***
rev_retail              0.72483    0.11706   6.192 9.50e-10 ***
if_good_healthcareTRUE  0.05462    0.09497   0.575  0.56536
if_large_populationTRUE 0.26117    0.07958   3.282  0.00108 **
log_icr                -0.53654    0.03981 -13.478  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.856 on 800 degrees of freedom
Multiple R-squared:  0.5969,    Adjusted R-squared:  0.5934
F-statistic: 169.3 on 7 and 800 DF,  p-value: < 2.2e-16
```

The figure presented above depicts our full model, which not only includes all relevant economic indexes but also incorporates two additional features that we believed may have a correlation with death rate. To determine the necessity of these two categorical features, we utilized ANOVA.

```
Analysis of Variance Table

Response: log_y
                     Df Sum Sq Mean Sq  F value    Pr(>F)
emp                   1 480.31  480.31 655.5542 < 2.2e-16 ***
res                   1   6.57    6.57   8.9700  0.002829 **
ts                    1 215.38  215.38 293.9665 < 2.2e-16 ***
rev_retail            1  24.71   24.71  33.7268 9.147e-09 ***
if_good_healthcare    1   0.33    0.33   0.4466  0.504135
if_large_population   1   7.71    7.71  10.5163  0.001232 **
log_icr               1 133.10  133.10 181.6647 < 2.2e-16 ***
Residuals           800 586.14    0.73
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table provides valuable information about the statistical significance of each feature in our model. Upon analyzing the table, we observed that the p-value for "**if_good_healthcare**" was 0.504, which is much larger than the significance level of 0.05. This large p-value indicates that the effect of "**if_good_healthcare**" is insignificant and that it is not a correlative feature for predicting death rate. Based on this analysis, we decided to exclude "**if_good_healthcare**" from our full model.

After excluding "**if_good_healthcare**" from our full model, we experimented with different combinations of **interactive terms** using ANOVA. In our exploration of different interactive terms, we tried various combinations and found that the model which only included an interaction between "**emp**" and "**if_large_population**" provided the best fit for our data. This result suggests that the relationship between employment and population size is particularly important for predicting variations in COVID-19 death rate.

```
Analysis of Variance Table

Model 1: log_y ~ emp + res + ts + rev_retail + if_large_population + log_icr +
    emp * if_large_population
Model 2: log_y ~ emp + res + ts + rev_retail + if_large_population + log_icr
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    800 575.17
2    801 586.38 -1   -11.215 15.599 8.522e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 2036.364
```

The above figure provides important insights into the significance of the interaction term we included in our model. From the ANOVA table, we can see that the p-value associated with adding the "**emp * if_large_population**" interaction term is extremely small, indicating that this term is highly significant in predicting variations in COVID-19 death rate. This result is significant because it demonstrates that the relationship between employment and population size is not simply additive, but rather interacts in complex ways to influence the spread and severity of COVID-19.

Furthermore, the inclusion of this interaction term also resulted in the smallest **AIC** value for our model, with a value of 2036.364. This result suggests that our model is well-specified and accurately captures the relationships between economic indexes and COVID-19 death rate. Taken together, we confirmed that adding the categorical term "**if_large_population**" and the interactive term "**emp * if_large_population**" is indeed necessary.

Tool 2: Model selection

To ensure that all features in our model are significant, we utilized the analysis of variance (ANOVA) tool, which examines the F-statistics of each feature. However, simply identifying significant features is not enough to confirm the accuracy of the model. Therefore, we proceeded to verify our model using forward selection based on the Akaike information criterion (AIC).

Our forward selection process began with a null model, which contained only an intercept term, and subsequently added features with the lowest AIC until the best-fit model was achieved. After completing the forward selection process, we compared the resulting model to our previously defined model and found that they were exactly the same. This suggests that our main model is indeed the best-fit model for the data we are analyzing.

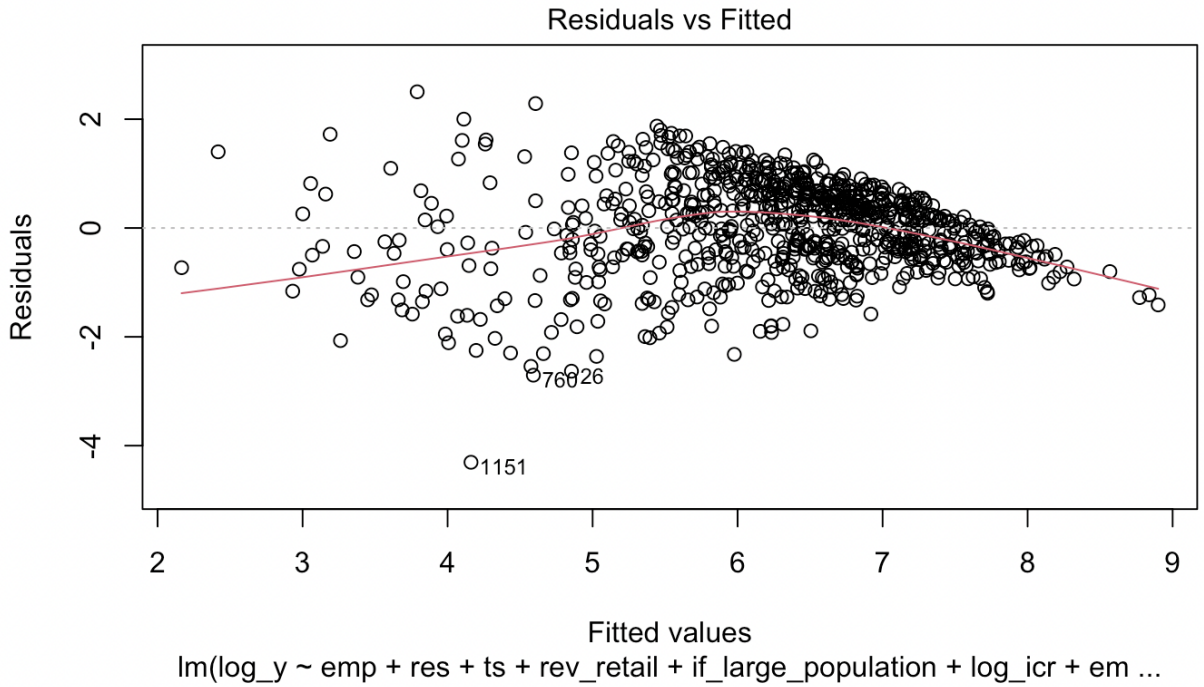
```
Call:
lm(formula = log_y ~ log_icr + rev_retail + emp + ts + res +
    if_large_population + emp:if_large_population, data = train_data)

Coefficients:
                (Intercept)                      log_icr                   rev_retail
                    6.12572                     -0.55852                      0.74270
                        emp                           ts                          res
                    7.54160                     -2.74918                     -8.28474
    if_large_populationTRUE  emp:if_large_populationTRUE
                   -0.08709                     -4.63062
```

By using AIC-based forward selection, we were able to systematically add variables to our model while minimizing the risk of overfitting. Furthermore, the use of AIC allowed us to compare different models and select the one that was most appropriate for our analysis.

In conclusion, through the use of ANOVA and AIC-based forward selection, we have ensured that our model contains only significant features and is the best fit model for our data. These steps are critical to ensuring the accuracy and reliability of our analysis, and ultimately, to drawing meaningful conclusions from our research.

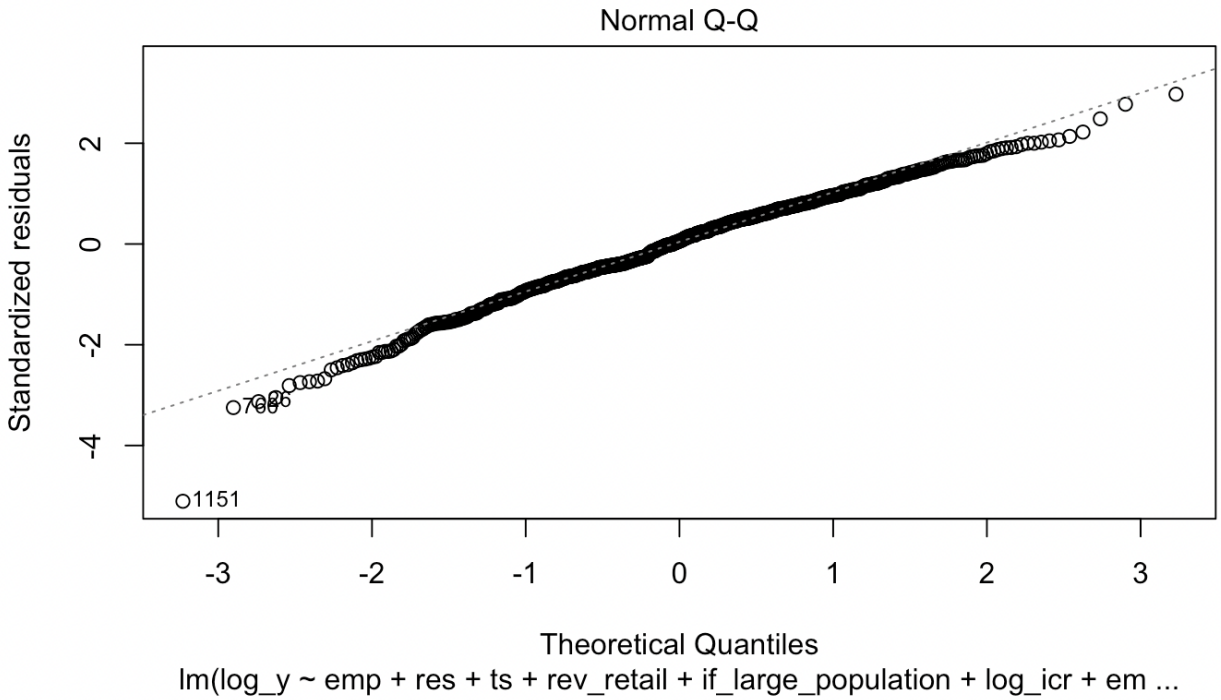
Tool 3: Model diagnostics and Influential observations



In model diagnostics, we first did a residual plot. It can be seen from the figure that the residuals don't look totally randomly distributed around 0. This implies that either (1) the linear relationship or (2) the common variance assumption are not completely satisfied.

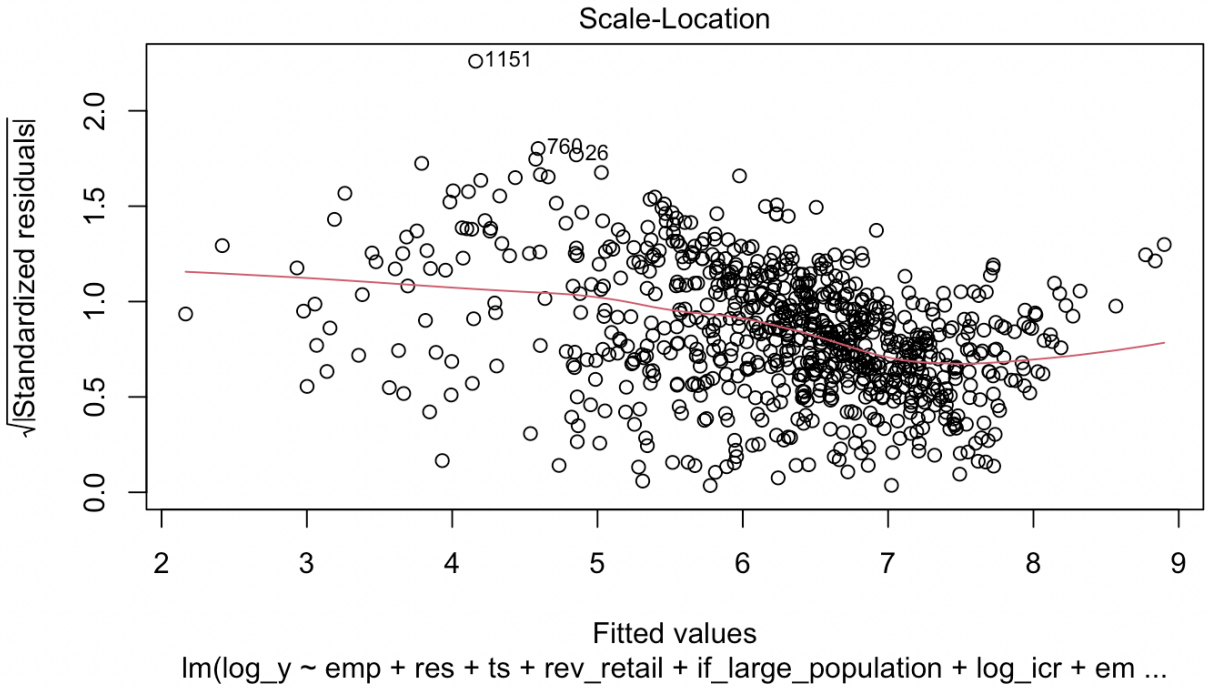
If the linear relationship assumption is not satisfied, it suggests that the relationship between the dependent variable and one or more independent variables may not be accurately modeled using a linear regression model. This could be due to non-linear relationships between the variables or the presence of influential outliers in the data.

Alternatively, if the assumption of common variance is not satisfied, it suggests that the variance of the residuals is not constant across the range of predicted values. This is known as heteroscedasticity and can lead to biased parameter estimates, incorrect standard errors, and unreliable hypothesis testing.

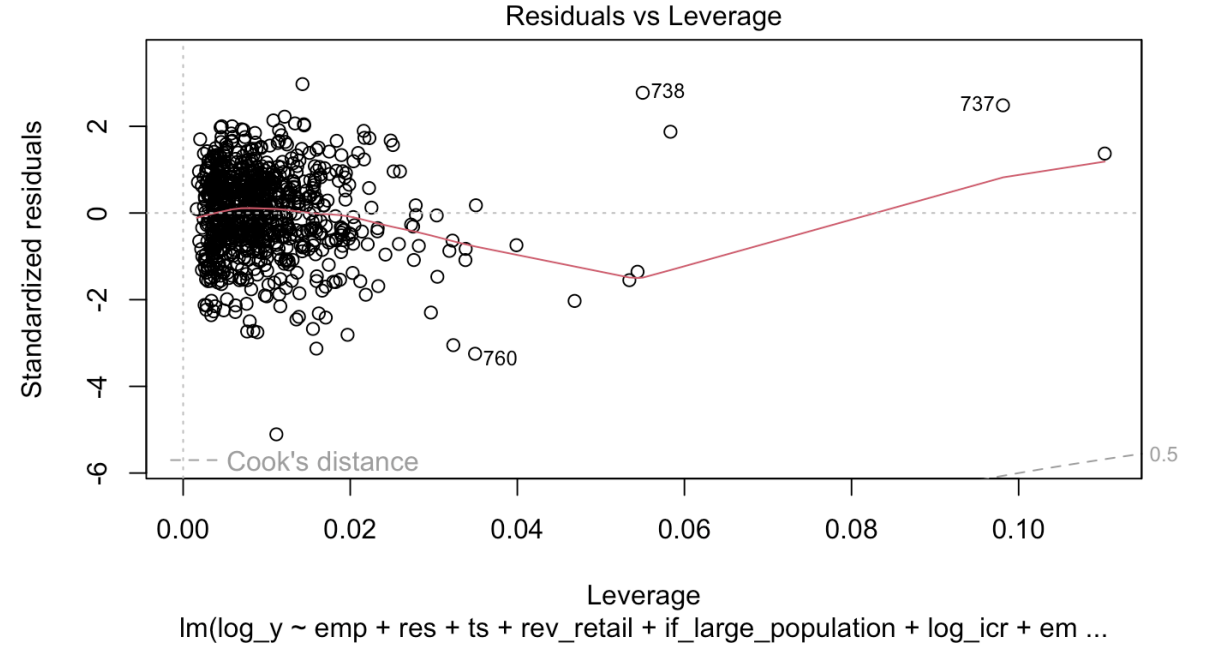


After examining the residual plot in our linear regression analysis, we moved on to checking the normality assumption using a QQ-plot. The QQ-plot allows us to compare the distribution of the standardized residuals to a normal distribution.

Upon examination of the QQ-plot, we observed that the points were not distributed roughly along the diagonal line, indicating that the normal distribution assumption is not fully satisfied. Specifically, the plot showed a left-skewed distribution of residuals, indicating that there are more residuals with lower values than expected in a normal distribution.



Next, we moved on to check the assumption of homoscedasticity. Upon examination of the homoscedasticity plot, we observed that there were some points that were not randomly scattered around the horizontal line, indicating that the homoscedasticity assumption may not be fully valid. Specifically, the plot showed a pattern of increasing or decreasing variability in the residuals as the predicted values increased, suggesting the presence of heteroscedasticity.



As part of our model diagnostic process, we also examined whether there were any influential observations in our regression analysis. Influential observations refer to individual

data points that can have a significant impact on the regression model's parameter estimates and overall fit.

We used Cook's distance to identify any influential observations in our model. Cook's distance measures the effect of deleting a particular observation on the regression coefficient estimates and can be used to identify data points that have a disproportionate influence on the model.

Upon examination of Cook's distance plot, we observed that no points were beyond the critical value, suggesting that there may be no highly influential observations in our model. This is a positive indication, as influential observations can lead to biased parameter estimates and affect the validity and reliability of our regression analysis.

Conclusion

```
Call:
lm(formula = log_y ~ emp + res + ts + rev_retail + if_large_population +
    log_icr + emp * if_large_population, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3058 -0.5245  0.0509  0.5989  2.5033

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.12572    0.12688  48.278  < 2e-16 ***
emp                            7.54160    0.63650  11.849  < 2e-16 ***
res                           -8.28474    1.55104  -5.341 1.20e-07 ***
ts                            -2.74918    0.25079 -10.962  < 2e-16 ***
rev_retail                     0.74270    0.11532   6.440 2.06e-10 ***
if_large_populationTRUE       -0.08709    0.11349  -0.767    0.443
log_icr                       -0.55852    0.03966 -14.084  < 2e-16 ***
emp:if_large_populationTRUE   -4.63062    1.17246  -3.950 8.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8479 on 800 degrees of freedom
Multiple R-squared:  0.6045,    Adjusted R-squared:  0.601
F-statistic: 174.7 on 7 and 800 DF,  p-value: < 2.2e-16
```

The summary of our final model is as above. We discovered the COVID death rate correlated with **emp** (employment rate), **res** (residential), **ts** (transit time), **rev_retial** (retail revenue), **if_large_population**, **log_icr** (), and **emp * if_large_population**. These factors all have an influential effect on COVID death rate.

Upon further investigation through model diagnostics, we found that many assumptions were not held, suggesting a possible violation of the OLS assumptions. One potential reason for this violation could be the use of states as granularity, which may result in some rows of data

being similar or following specific patterns due to the dependence of the current state on the previous period.

Another limitation of our model is the lack of sufficient features. Although we have included many major economic and social factors, there may still be many variables that can have a significant impact on COVID death rate, which were not considered in our analysis.

To address these limitations, in future analyses, we could consider adding more features to our model to improve its predictive power. Additionally, we could explore other modeling approaches such as Generalized Least Squares (GLS), which can help alleviate the violation of OLS assumptions and potentially improve the performance of our model.

Overall, despite the limitations, our analysis provides valuable insights into the factors that influence COVID death rate and highlights the importance of conducting thorough model diagnostics and considering potential limitations in our analyses to ensure the validity and reliability of our findings.