The George Washington University

November 20, 2021

# PROJECT

# Inference on Covid-19 Case Rate by Linear Regression Analysis

Mu Hu, Zibo Hong, Yuanzhe He

STAT 6214

Professor Hu

## Summary

With Covid-19 virus accompanying human beings for over two years, many research papers have been published on developing powerful models to predict Covid cases in certain regions. There are also many researchers focusing on the causal relationship among Covid and a series of explanatory variables. In this Project, we decomposed factors that may be statistically significant (at 5 percent level) in explaining Covid case rate. Modelling on inference of Covid case rate by traditional linear regression method can be expected.

The dataset used in this project is an aggregation of County-based Covid-19 case counts (as of 11/6/2021) from USAFACTS and United States Census Bureau's Planning Database (2021 Tract Data). A subjective variable selection was performed first based on existing research to help reduce the number of predictors from the original dataset. A Lasso regression was then run to help further reduce the number of predictors before the linear regression analysis. Log transformation of Covid case rate was used as the response variable. After a series of  exploratory data analysis (EDA), a reference linear model was established with R-squared value 33.5%. In the linear regression analysis, interactions and quadratic terms were added to the model according to the diagnostic plots of the reference model. "Stepwise" method was then used to further select variables. Model assumptions, data anomalies (e.g. Influential points, outliers, influence, collinearity) and model fit statistics are checked continuously throughout the whole linear regression analysis. The final model successfully raised the R-squared value to 39%, an acceptable level using real-life data. The conclusion drawn from the final model is that Covid case rate is significantly correlated with vaccination, region, education, age, health insurance, race, hygiene and whether a household has computers.

## 1. Dataset Preparation

Two datasets were used in this project. One is the USAFACTS website which has the latest covid counts data as well as the vaccination progress in the United States (US covid-19 cases and deaths by State 2021). The other is the United States Census Bureau's Planning Database 2021 Tract Data (US Census Bureau, *Planning database* 2021). Because our goal was to find out factors that can well interpret the Covid case rate, we wanted to combine the demographic dataset to the Covid-related dataset.

To combine the two datasets, we first aggregated the demographic data from tract level to county level. Then an innerjoin function with merged variables 'State' and 'County' was used to merge the two datasets.

## 2.Creating categorical and response variable

A categorical explanatory variable "region" was created where all the 50 states plus a Washington DC were divided into four region levels according to the US Census Bureau(US Census Bureau, *Census Regions and Divisions of the United States*). This is the only categorical predictor in our dataset.

Covid case rate was assumed in the first place to be the response variable. It was calculated as $Covid\ case\ rate\ =\ 100\ \times\ Covid\ case\ counts/total\ population$[1]. However, an histogram of the response variable looked heavily right skewed, suggesting extreme outliers exist. Investigation and removal of these extreme outliers and influential points was required to see if we should keep using Covid case rate as the response variable.

A base model was established whose diagnostic plots were shown in Figure 1. The "Residuals vs Leverage" plot gave us information on influential points if certain points were very close to or have exceeded the Cook's distance of 0.5 or 1. "Normal QQ" plot helped us determine if the residuals followed a normal distribution. The "Residuals vs Fitted" and "Scale-Location" plots gave us information on whether the residuals' mean were generally scattered around zero, and whether the residuals' variance was constant respectively. After an investigation of these outliers, some observations were taken out from the dataset. Then a refitted model was established where another round of diagnostic plots should be examined. Followed by several trials, a final refitted model "lmod_base_r" was formed. The principal in removing extreme outliers and influential points is to keep a balance in overall model fitting and the completeness of data.

[1] The denominator *total population* used in USAFACTS website is from the 2019 Vintage US Census Bureau Annual Estimates of the Resident Population for the United States (Reporting covid-19 vaccinations in the United States).

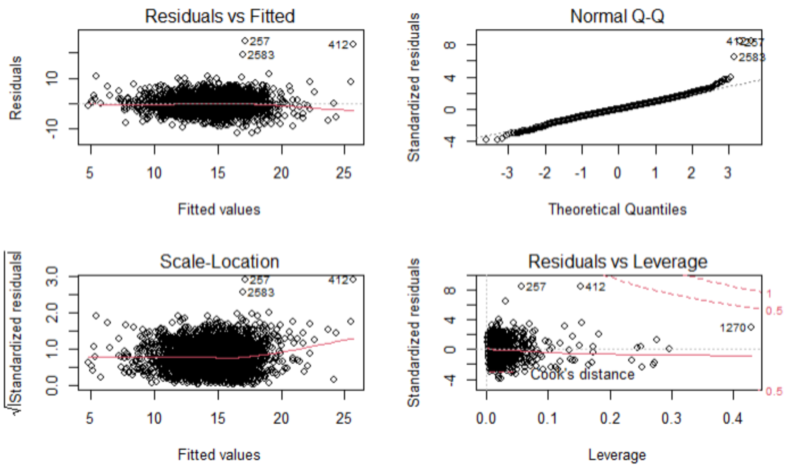


Figure 1. Diagnostic plots of base model

A histogram overlaid with the density function was drawn again on the Covid case rate after extreme outliers have been removed. The heavily right skewed phenomenon still existed. It naturally occurred to us to try a log-transformation. In Figure 2, a comparison was made between the two histograms of Covid case rate and log(Covid case rate). The right part looked better, suggesting the log form of Covid case rate was closer to meet the normal distribution assumption. Therefore, our response variable was replaced by log(Covid case rate). In the dataset, some counties had Covid case rate equalling zero. In order to avoid error in future calculation, we added a tiny number 0.000001 to the Covid case rate and then took a log transformation whose formula was shown below:

$$Log(Covid\ Case\ Rate) = log(Covid\ Case\ Rate + 0.000001)$$

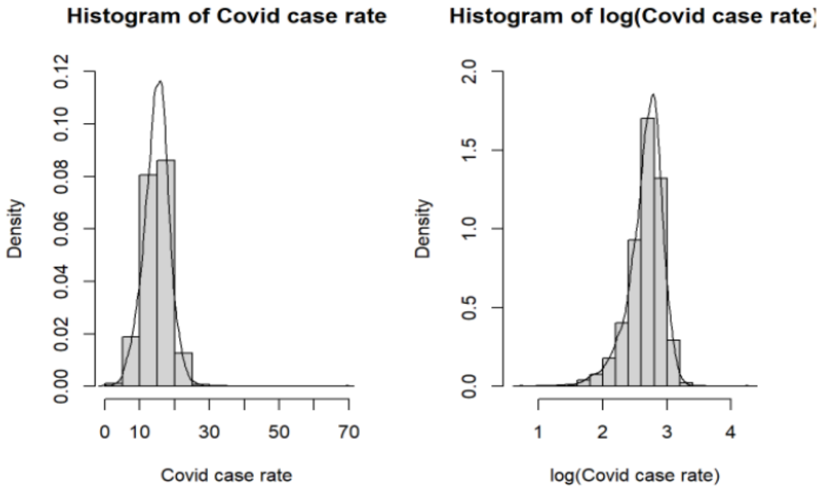


Figure 2. Comparison in histogram of Covid case rate and its log form

### 3.Reducing predictors

There were over 500 predictors in the demographic dataset. To better perform a linear regression, reducing the number of predictors was necessary.

We first subjectively selected variables based on existing research.  According to the massive research in recent two years, political, demographic, income and health related factors are found to play a key role in interpreting Covid case rate (Wylezinski et al.,2021; Richmond et al.,2020; Mueller and Eric et al.,2021). This reduced the number of predictors to around 50. Then we used a Lasso Regression Analysis to further help us reduce the number of predictors to 18. A full description of the 18 predictors selected is shown in Table 3 in the Appendix.

### 4.Exploratory Data Analysis(EDA)

EDA performed in this project included checking dimension of data, types of variables, missing values and correlations. The analysis plots were shown in Figure 4-6 in the Appendix. Some findings from the EDA were listed as followed:

1. The dataset had 3138 observations and 18 predictors
2. There were no missing values in the dataset
3. Among all the 18 predictors, '*pct_NH_NHOPI_alone_ACS_15_19*', '*pct_Pop_45_64_ACS_15_19*', *'pct_Pop5_17_ACS_15_19*' and '*% of POP. FULLY VACCINATED*' had the highest absolute correlation value with the response variable.
4. Different region levels differed much in log(Covid case rate). Region Northeast had the lowest median value in log(Covid case rate)

### 5.Fitting Linear Regression Reference Model and Investigate outliers again

Because the response variable changed from Covid case rate to log(Covid case rate), a new reference model was needed and the investigation and removal procedure performed in step two needed to be done again. The influential points and extreme outliers removed above should be added back and reconsidered. The diagnostic plots of the new reference model after removing certain extreme outliers and influential points were shown in Figure 7. We could say the residuals were generally scattered around zero according to the "Residuals vs Fitted' plot. The residuals also followed a close-to-normal distribution from the 'Normal Q-Q' plot. The 'Scale-Location' plot gave us information that the residual variance was not constant enough, as was later confirmed by the Breusch-Pagan Test. The p-value in the bp-test was extremely small, rejecting the null hypothesis that the residual variance were homoscedastic. The

‘Residuals vs Leverage’ plot indicated there were no influential points as all the observations were far from close to Cook’s distance of 0.5. A quadratic term was observed in both the ‘Residuals vs Fitted’ plot and ‘Scale-Location’ plot, meaning some variation of the response variable not explained by the current model. After running a summary analysis, the new reference model had an R-squared value of 33.5%.

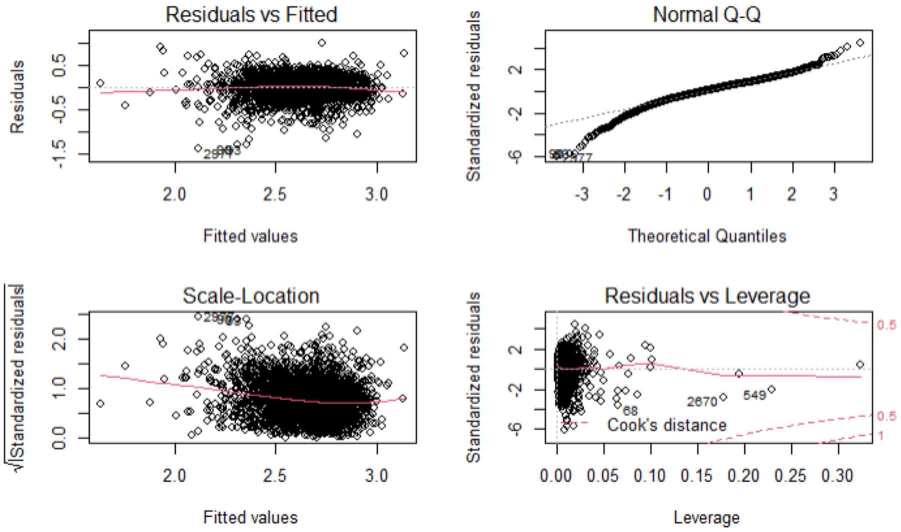


Figure 7. Diagnostic plots of the new reference model

## 6.Exploring interaction and polynomial terms

Because the diagnostic plots in Figure 7 indicated that there existed some unexplained patterns in log(Covid case rate), interaction and polynomial terms would be considered to further help explain the response variable.

For interaction, because there was only one categorical predictor ‘*region*’ with four levels, interaction would be tried between ‘*region*’ and every other predictor. Trails between ‘*region*’ and each continuous predictor were performed one by one. The criteria to determine whether an interaction existed was to look at the slopes. For example, Figure 8 showed there were some interactions on region and percentage of fully vaccinated population as the slopes of different regions were different.

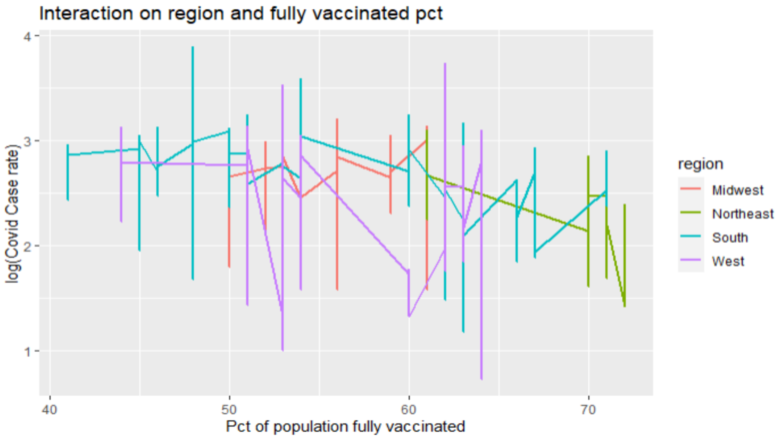
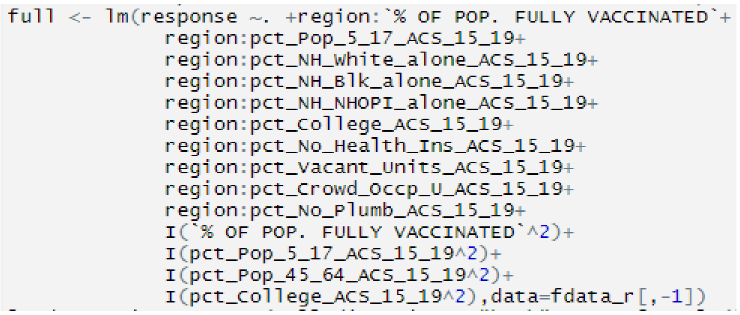


Figure 8. Interaction on region and percentage of fully vaccinated population

For polynomial terms, we would try quadratic terms in this case as it was the only shape of pattern observed in Figure 7. Trails between the response variable and the quadratic term of each predictor were also performed one by one. The threshold to decide to include the quadratic term was to see if the model with only response variable and the quadratic term had an R-squared value larger than 4% and meanwhile the quadratic term was statistically significant at 5 percent significance level.

## 7.Variable Selection in Linear Regression Model

Stepwise method was used in variable selection. Figure 9 displayed the scope of full formula with all the original 18 predictors plus the interaction and polynomial terms we decided to put in based on step 6. The basic algorithm behind the stepwise method is that it starts from an intercept-only model. It then starts to add a predictor that the computer finds most significant within the scope of full formula. It will also drop from the most insignificant predictor if adding a new predictor makes an existing one not significant any more. In another word, the stepwise method has both advantages from forward and backward selection, and it uses AIC criteria to automatically select models.

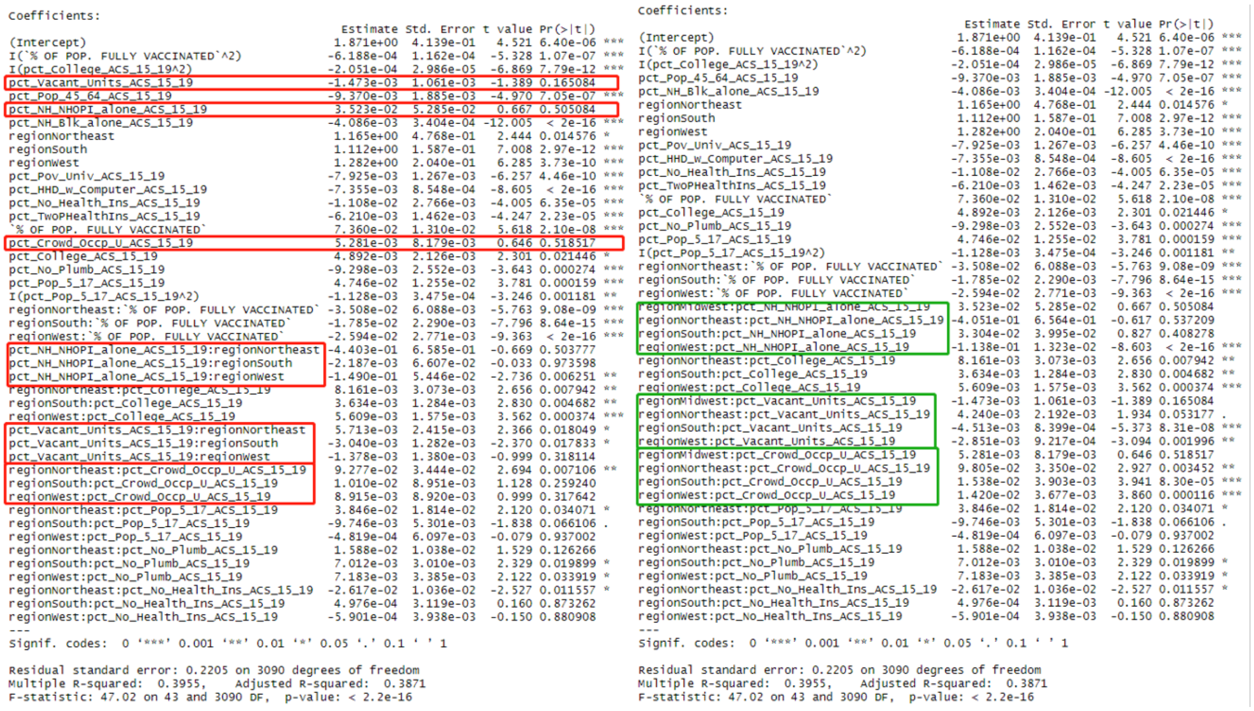


```
full <- lm(response ~. +region:`% OF POP. FULLY VACCINATED`+
             region:pct_Pop_5_17_ACS_15_19+
             region:pct_NH_White_alone_ACS_15_19+
             region:pct_NH_Blk_alone_ACS_15_19+
             region:pct_NH_NHOPI_alone_ACS_15_19+
             region:pct_College_ACS_15_19+
             region:pct_No_Health_Ins_ACS_15_19+
             region:pct_Vacant_Units_ACS_15_19+
             region:pct_Crowd_Occp_U_ACS_15_19+
             region:pct_No_Plumb_ACS_15_19+
             I(`% OF POP. FULLY VACCINATED`^2)+
             I(pct_Pop_5_17_ACS_15_19^2)+
             I(pct_Pop_45_64_ACS_15_19^2)+
             I(pct_College_ACS_15_19^2),data=fdata_r[,-1])
```

Figure 9. Full scope of predictors put in the stepwise model

**8.Further adjustment and collinearity checking**

After the stepwise selection, a summary analysis of the stepwise model was run for further adjustment. Three lower-term predictors had a very large P-value, indicating not significant. After removing them one by one, we found an interesting phenomenon that the model didn't have any change(in  terms of adjusted R-squared value and the coefficients of other predictors) without these three predictors. Normally, removing an insignificant predictor would result in a slight drop in the new model's adjusted R-squared value. It is often regarded as a tradeoff between the overall model fit and the model simplicity. In our case, however, not only the model didn't change, the interaction term with the *region*'s reference level showed up in the output. It seemed like the computer automatically added a lower order term of the removed predictor in the model as can be seen in Figure 10 in the Appendix. In figure 10, the left model output was the stepwise model the computer picked out. The right part was the model without the three insignificant predictors. *Region midwest* was the default reference level where it should not be included as a predictor like the left part did. In the right part,  *Region midwest* entered the model in the form of an interaction term together with other three region levels. After massive research, we found out that if a lower order term was removed in the linear regression model whereas its higher order term was still kept, a scale of change in the predictor could add back the lower order term (Faraway, 2005). And in order to maintain the overall same number of degrees of freedom, the computer languages would automatically add an additional degree of freedom to the interaction term with the reference level of the categorical variable (WHAT HAPPENS IF YOU OMIT THE MAIN EFFECT IN A REGRESSION MODEL WITH AN INTERACTION? | STATA FAQ). This explained why the model didn't change but the interaction order got flipped and the region reference level got displayed in the summary result. Because the interaction term for the highlighted predictors in Figure 10 were statistically significant at 5 percent significance level, there was no need to remove the individual lower order term.



```
Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                 1.871e+00  4.139e-01   4.521 6.40e-06 ***
I(`% OF POP. FULLY VACCINATED`^2)          -6.188e-04  1.162e-04  -5.328 1.07e-07 ***
I(pct_College_ACS_15_19^2)                 -2.051e-04  2.986e-05  -6.869 7.79e-12 ***
pct_Vacant_Units_ACS_15_19                 -1.473e-03  1.061e-03  -1.389 0.165084    
pct_Pop_45_64_ACS_15_19                    -9.370e-03  1.885e-03  -4.970 7.05e-07 ***
pct_NH_NHOPI_alone_ACS_15_19                3.523e-02  5.285e-02   0.667 0.505084    
pct_NH_Blk_alone_ACS_15_19                 -4.086e-03  3.404e-04 -12.005  < 2e-16 ***
regionNortheast                             1.165e+00  4.768e-01   2.444 0.014576 *  
regionSouth                                 1.112e+00  1.587e-01   7.008 2.97e-12 ***
regionWest                                  1.282e+00  2.040e-01   6.285 3.73e-10 ***
pct_Pov_Univ_ACS_15_19                     -7.925e-03  1.267e-03  -6.257 4.46e-10 ***
pct_HHD_w_Computer_ACS_15_19               -7.355e-03  8.548e-04  -8.605  < 2e-16 ***
pct_No_Health_Ins_ACS_15_19                -1.108e-02  2.766e-03  -4.005 6.35e-05 ***
pct_TwoPHealthIns_ACS_15_19                -6.210e-03  1.462e-03  -4.247 2.23e-05 ***
`% OF POP. FULLY VACCINATED`                7.360e-02  1.310e-02   5.618 2.10e-08 ***
pct_Crowd_Occp_U_ACS_15_19                  5.281e-03  8.179e-03   0.646 0.518517    
pct_College_ACS_15_19                       4.892e-03  2.126e-03   2.301 0.021446 *  
pct_No_Plumb_ACS_15_19                     -9.298e-03  2.552e-03  -3.643 0.000274 ***
pct_Pop_5_17_ACS_15_19                      4.746e-02  1.255e-02   3.781 0.000159 ***
I(pct_Pop_5_17_ACS_15_19^2)                -1.128e-03  3.475e-04  -3.246 0.001181 ** 
regionNortheast:`% OF POP. FULLY VACCINATED` -3.508e-02  6.088e-03  -5.763 9.08e-09 ***
regionSouth:`% OF POP. FULLY VACCINATED`   -1.785e-02  2.290e-03  -7.796 8.64e-15 ***
regionWest:`% OF POP. FULLY VACCINATED`    -2.594e-02  2.771e-03  -9.363  < 2e-16 ***
pct_NH_NHOPI_alone_ACS_15_19:regionNortheast -4.403e-01  6.585e-01  -0.669 0.503777    
pct_NH_NHOPI_alone_ACS_15_19:regionSouth   -2.187e-03  6.607e-02  -0.033 0.973598    
pct_NH_NHOPI_alone_ACS_15_19:regionWest    -1.490e-01  5.446e-02  -2.736 0.006251 ** 
regionNortheast:pct_College_ACS_15_19       8.161e-03  3.073e-03   2.656 0.007942 ** 
regionSouth:pct_College_ACS_15_19           3.634e-03  1.284e-03   2.830 0.004682 ** 
regionWest:pct_College_ACS_15_19            5.609e-03  1.575e-03   3.562 0.000374 ***
pct_Vacant_Units_ACS_15_19:regionNortheast  5.713e-03  2.415e-03   2.366 0.018049 *  
pct_Vacant_Units_ACS_15_19:regionSouth     -3.040e-03  1.282e-03  -2.370 0.017833 *  
pct_Vacant_Units_ACS_15_19:regionWest      -1.378e-03  1.380e-03  -0.999 0.318114    
regionNortheast:pct_Crowd_Occp_U_ACS_15_19  9.277e-02  3.444e-02   2.694 0.007106 ** 
regionSouth:pct_Crowd_Occp_U_ACS_15_19      1.010e-02  8.951e-03   1.128 0.259240    
regionWest:pct_Crowd_Occp_U_ACS_15_19       8.915e-03  8.920e-03   0.999 0.317642    
regionNortheast:pct_Pop_5_17_ACS_15_19      3.846e-02  1.814e-02   2.120 0.034071 *  
regionSouth:pct_Pop_5_17_ACS_15_19         -9.746e-03  5.301e-03  -1.838 0.066106 .  
regionWest:pct_Pop_5_17_ACS_15_19          -4.819e-04  6.097e-03  -0.079 0.937002    
regionNortheast:pct_No_Plumb_ACS_15_19      1.588e-02  1.038e-02   1.529 0.126266    
regionSouth:pct_No_Plumb_ACS_15_19          7.012e-03  3.010e-03   2.329 0.019899 *  
regionWest:pct_No_Plumb_ACS_15_19           7.183e-03  3.385e-03   2.122 0.033919 *  
regionNortheast:pct_No_Health_Ins_ACS_15_19 -2.617e-02  1.036e-02  -2.527 0.011557 *  
regionSouth:pct_No_Health_Ins_ACS_15_19     4.976e-04  3.119e-03   0.160 0.873262    
regionWest:pct_No_Health_Ins_ACS_15_19     -5.901e-04  3.938e-03  -0.150 0.880908    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2205 on 3090 degrees of freedom
Multiple R-squared:  0.3955,	Adjusted R-squared:  0.3871 
F-statistic: 47.02 on 43 and 3090 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                 1.871e+00  4.139e-01   4.521 6.40e-06 ***
I(`% OF POP. FULLY VACCINATED`^2)          -6.188e-04  1.162e-04  -5.328 1.07e-07 ***
I(pct_College_ACS_15_19^2)                 -2.051e-04  2.986e-05  -6.869 7.79e-12 ***
pct_Pop_45_64_ACS_15_19                    -9.370e-03  1.885e-03  -4.970 7.05e-07 ***
pct_NH_Blk_alone_ACS_15_19                 -4.086e-03  3.404e-04 -12.005  < 2e-16 ***
regionNortheast                             1.165e+00  4.768e-01   2.444 0.014576 *  
regionSouth                                 1.112e+00  1.587e-01   7.008 2.97e-12 ***
regionWest                                  1.282e+00  2.040e-01   6.285 3.73e-10 ***
pct_Pov_Univ_ACS_15_19                     -7.925e-03  1.267e-03  -6.257 4.46e-10 ***
pct_HHD_w_Computer_ACS_15_19               -7.355e-03  8.548e-04  -8.605  < 2e-16 ***
pct_No_Health_Ins_ACS_15_19                -1.108e-02  2.766e-03  -4.005 6.35e-05 ***
pct_TwoPHealthIns_ACS_15_19                -6.210e-03  1.462e-03  -4.247 2.23e-05 ***
`% OF POP. FULLY VACCINATED`                7.360e-02  1.310e-02   5.618 2.10e-08 ***
pct_College_ACS_15_19                       4.892e-03  2.126e-03   2.301 0.021446 *  
pct_No_Plumb_ACS_15_19                     -9.298e-03  2.552e-03  -3.643 0.000274 ***
pct_Pop_5_17_ACS_15_19                      4.746e-02  1.255e-02   3.781 0.000159 ***
I(pct_Pop_5_17_ACS_15_19^2)                -1.128e-03  3.475e-04  -3.246 0.001181 ** 
regionNortheast:`% OF POP. FULLY VACCINATED` -3.508e-02  6.088e-03  -5.763 9.08e-09 ***
regionSouth:`% OF POP. FULLY VACCINATED`   -1.785e-02  2.290e-03  -7.796 8.64e-15 ***
regionWest:`% OF POP. FULLY VACCINATED`    -2.594e-02  2.771e-03  -9.363  < 2e-16 ***
regionMidwest:pct_NH_NHOPI_alone_ACS_15_19  3.523e-02  5.285e-02   0.667 0.505084    
regionNortheast:pct_NH_NHOPI_alone_ACS_15_19 -4.051e-01  6.564e-01  -0.617 0.537209    
regionSouth:pct_NH_NHOPI_alone_ACS_15_19    3.304e-02  3.995e-02   0.827 0.408278    
regionWest:pct_NH_NHOPI_alone_ACS_15_19    -1.138e-01  1.323e-02  -8.603  < 2e-16 ***
regionNortheast:pct_College_ACS_15_19       8.161e-03  3.073e-03   2.656 0.007942 ** 
regionSouth:pct_College_ACS_15_19           3.634e-03  1.284e-03   2.830 0.004682 ** 
regionWest:pct_College_ACS_15_19            5.609e-03  1.575e-03   3.562 0.000374 ***
regionMidwest:pct_Vacant_Units_ACS_15_19   -1.473e-03  1.061e-03  -1.389 0.165084    
regionNortheast:pct_Vacant_Units_ACS_15_19  4.240e-03  2.192e-03   1.934 0.053177 .  
regionSouth:pct_Vacant_Units_ACS_15_19     -4.513e-03  8.399e-04  -5.373 8.31e-08 ***
regionWest:pct_Vacant_Units_ACS_15_19      -2.851e-03  9.217e-04  -3.094 0.001996 ** 
regionMidwest:pct_Crowd_Occp_U_ACS_15_19    5.281e-03  8.179e-03   0.646 0.518517    
regionNortheast:pct_Crowd_Occp_U_ACS_15_19  9.805e-02  3.350e-02   2.927 0.003452 ** 
regionSouth:pct_Crowd_Occp_U_ACS_15_19      1.538e-02  3.903e-03   3.941 8.30e-05 ***
regionWest:pct_Crowd_Occp_U_ACS_15_19       1.420e-02  3.677e-03   3.860 0.000116 ***
regionNortheast:pct_Pop_5_17_ACS_15_19      3.846e-02  1.814e-02   2.120 0.034071 *  
regionSouth:pct_Pop_5_17_ACS_15_19         -9.746e-03  5.301e-03  -1.838 0.066106 .  
regionWest:pct_Pop_5_17_ACS_15_19          -4.819e-04  6.097e-03  -0.079 0.937002    
regionNortheast:pct_No_Plumb_ACS_15_19      1.588e-02  1.038e-02   1.529 0.126266    
regionSouth:pct_No_Plumb_ACS_15_19          7.012e-03  3.010e-03   2.329 0.019899 *  
regionWest:pct_No_Plumb_ACS_15_19           7.183e-03  3.385e-03   2.122 0.033919 *  
regionNortheast:pct_No_Health_Ins_ACS_15_19 -2.617e-02  1.036e-02  -2.527 0.011557 *  
regionSouth:pct_No_Health_Ins_ACS_15_19     4.976e-04  3.119e-03   0.160 0.873262    
regionWest:pct_No_Health_Ins_ACS_15_19     -5.901e-04  3.938e-03  -0.150 0.880908    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2205 on 3090 degrees of freedom
Multiple R-squared:  0.3955,	Adjusted R-squared:  0.3871 
F-statistic: 47.02 on 43 and 3090 DF,  p-value: < 2.2e-16
```

Figure 10. An interesting finding for hierarchical terms in linear regression model

Collinearity was then checked on the stepwise model. The threshold to determine if there was a serious collinearity issue was to see if the VIF value was larger than 10 (Hair et al., 1995). *I(POP. FULLY VACCINATED ^ 2)* was removed based on its high VIF value. Then another round of summary and VIF test were performed on the new model. Noticeably, it was expected that some of the predictors naturally had high Variance Inflation Factor(VIF) values like *region* because that was the only categorical predictor in the model. Even the VIF of '*region'* was high, it cannot be removed. And removing the interaction term '*region:'% of POP. FULLY VACCINATED*' would result in a large drop in model explanation power, we decided to keep it as well.

## 9. Model Comparison

The final model had an R-squared value of  39% compared with the 33.5% in the initial reference model. And it also had better diagnostic plots compared to the reference model as are shown in Figure 11-12. We were satisfied with the result.

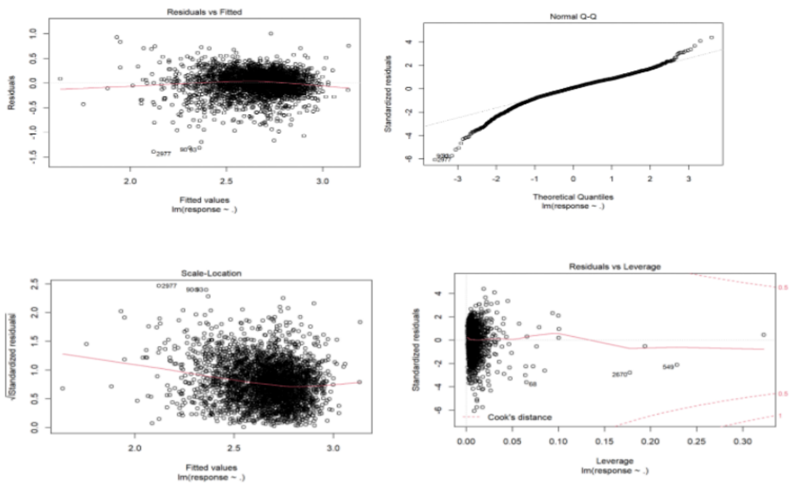


Figure 11. Diagnostic plots of the reference model

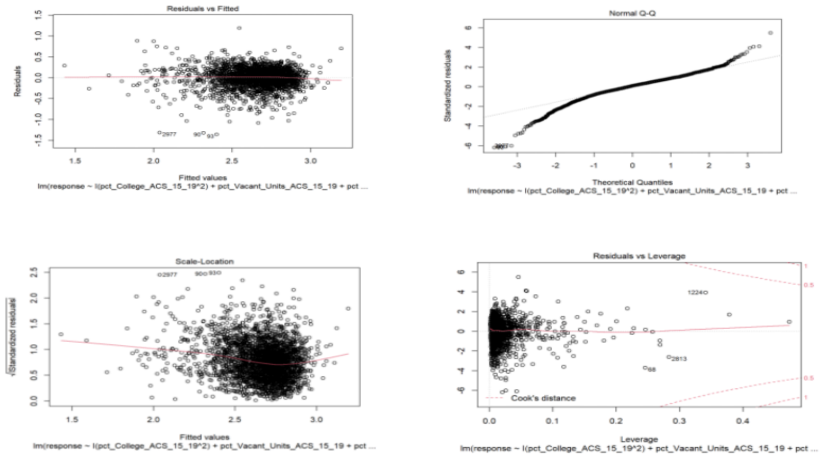
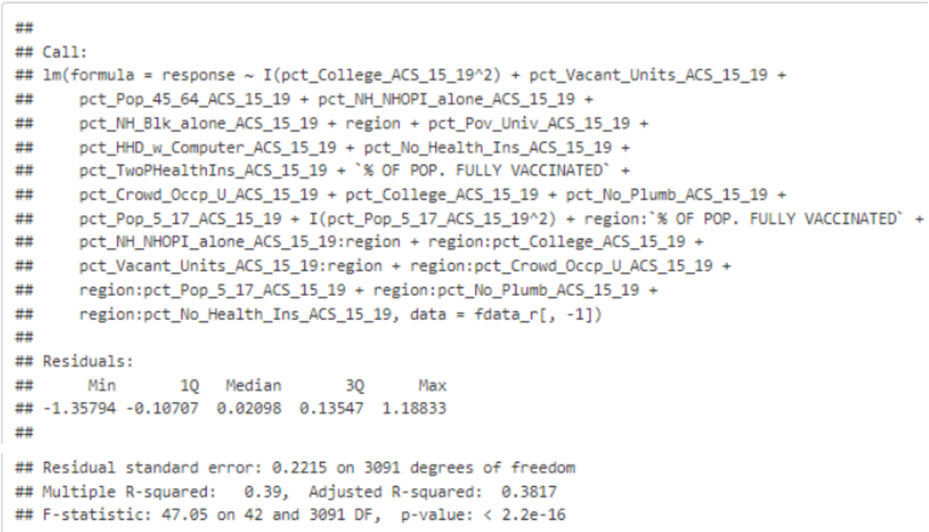


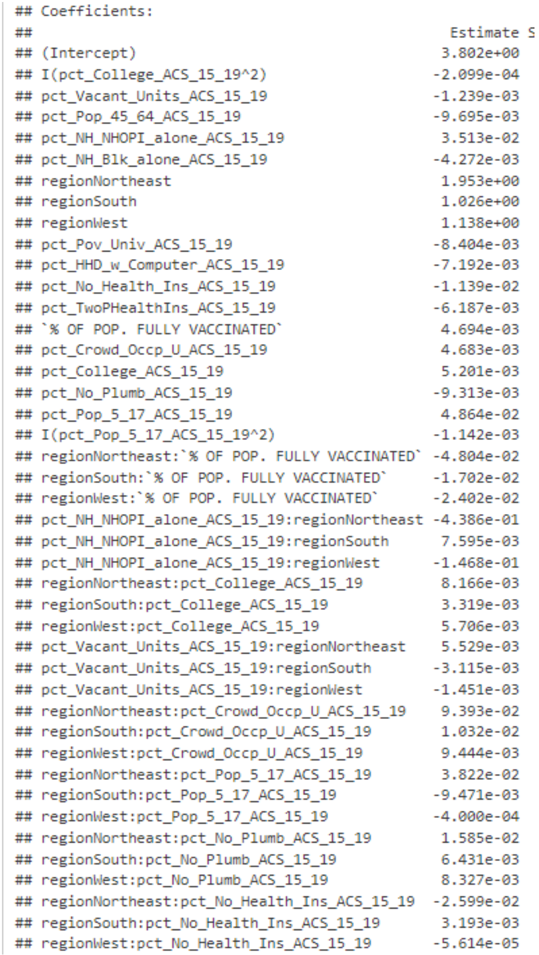
Figure 12. Diagnostic plots of the final model

## 10.Conclusion

The final model was displayed in Figure 13. From the coefficients, we could find predictors in ethnicity( *pct_NH_NHOPI_alone_ACS_15_19*), education(*pct_College_ACS_15_19*) , vaccination(*% of POP. FULLY VACCINATED*) and living condition(*pct_Crowd_Occp_U_ACS_15_19*) were key factors in interpreting log(Covid case rate). For example, the coefficient of *pct_College_ACS_15_19* equaled 0.005201. In reality, this meant while holding other covariates constant, one unit of change in percentage of population who went to college could bring 1.0052(e^0.005201) times change in Covid case rate. And from the coefficients of interactions with different region levels, we could infer that region South experienced the highest Covid case rate compared with other three levels.



```
##
## Call:
## lm(formula = response ~ I(pct_College_ACS_15_19^2) + pct_Vacant_Units_ACS_15_19 +
##     pct_Pop_45_64_ACS_15_19 + pct_NH_NHOPI_alone_ACS_15_19 +
##     pct_NH_Blk_alone_ACS_15_19 + region + pct_Pov_Univ_ACS_15_19 +
##     pct_HHD_w_Computer_ACS_15_19 + pct_No_Health_Ins_ACS_15_19 +
##     pct_TwoPHealthIns_ACS_15_19 + `% OF POP. FULLY VACCINATED` +
##     pct_Crowd_Occp_U_ACS_15_19 + pct_College_ACS_15_19 + pct_No_Plumb_ACS_15_19 +
##     pct_Pop_5_17_ACS_15_19 + I(pct_Pop_5_17_ACS_15_19^2) + region:`% OF POP. FULLY VACCINATED` +
##     pct_NH_NHOPI_alone_ACS_15_19:region + region:pct_College_ACS_15_19 +
##     pct_Vacant_Units_ACS_15_19:region + region:pct_Crowd_Occp_U_ACS_15_19 +
##     region:pct_Pop_5_17_ACS_15_19 + region:pct_No_Plumb_ACS_15_19 +
##     region:pct_No_Health_Ins_ACS_15_19, data = fdata_r[, -1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35794 -0.10707  0.02098  0.13547  1.18833
##
## Residual standard error: 0.2215 on 3091 degrees of freedom
## Multiple R-squared:   0.39,  Adjusted R-squared:  0.3817
## F-statistic: 47.05 on 42 and 3091 DF,  p-value: < 2.2e-16
```



```
## Coefficients:
##                                              Estimate S
## (Intercept)                                 3.802e+00
## I(pct_College_ACS_15_19^2)                 -2.099e-04
## pct_Vacant_Units_ACS_15_19                 -1.239e-03
## pct_Pop_45_64_ACS_15_19                    -9.695e-03
## pct_NH_NHOPI_alone_ACS_15_19                3.513e-02
## pct_NH_Blk_alone_ACS_15_19                 -4.272e-03
## regionNortheast                             1.953e+00
## regionSouth                                 1.026e+00
## regionWest                                  1.138e+00
## pct_Pov_Univ_ACS_15_19                     -8.404e-03
## pct_HHD_w_Computer_ACS_15_19               -7.192e-03
## pct_No_Health_Ins_ACS_15_19                -1.139e-02
## pct_TwoPHealthIns_ACS_15_19                -6.187e-03
## `% OF POP. FULLY VACCINATED`                4.694e-03
## pct_Crowd_Occp_U_ACS_15_19                  4.683e-03
## pct_College_ACS_15_19                       5.201e-03
## pct_No_Plumb_ACS_15_19                     -9.313e-03
## pct_Pop_5_17_ACS_15_19                      4.864e-02
## I(pct_Pop_5_17_ACS_15_19^2)                -1.142e-03
## regionNortheast:`% OF POP. FULLY VACCINATED` -4.804e-02
## regionSouth:`% OF POP. FULLY VACCINATED`     -1.702e-02
## regionWest:`% OF POP. FULLY VACCINATED`      -2.402e-02
## pct_NH_NHOPI_alone_ACS_15_19:regionNortheast -4.386e-01
## pct_NH_NHOPI_alone_ACS_15_19:regionSouth      7.595e-03
## pct_NH_NHOPI_alone_ACS_15_19:regionWest      -1.468e-01
## regionNortheast:pct_College_ACS_15_19         8.166e-03
## regionSouth:pct_College_ACS_15_19             3.319e-03
## regionWest:pct_College_ACS_15_19              5.706e-03
## pct_Vacant_Units_ACS_15_19:regionNortheast    5.529e-03
## pct_Vacant_Units_ACS_15_19:regionSouth       -3.115e-03
## pct_Vacant_Units_ACS_15_19:regionWest        -1.451e-03
## regionNortheast:pct_Crowd_Occp_U_ACS_15_19    9.393e-02
## regionSouth:pct_Crowd_Occp_U_ACS_15_19        1.032e-02
## regionWest:pct_Crowd_Occp_U_ACS_15_19         9.444e-03
## regionNortheast:pct_Pop_5_17_ACS_15_19        3.822e-02
## regionSouth:pct_Pop_5_17_ACS_15_19           -9.471e-03
## regionWest:pct_Pop_5_17_ACS_15_19            -4.000e-04
## regionNortheast:pct_No_Plumb_ACS_15_19        1.585e-02
## regionSouth:pct_No_Plumb_ACS_15_19            6.431e-03
## regionWest:pct_No_Plumb_ACS_15_19             8.327e-03
## regionNortheast:pct_No_Health_Ins_ACS_15_19  -2.599e-02
## regionSouth:pct_No_Health_Ins_ACS_15_19       3.193e-03
## regionWest:pct_No_Health_Ins_ACS_15_19       -5.614e-05
```

Figure 13. Final model

## Reference

"US Covid-19 Cases and Deaths by State." USAFacts.org, November 6, 2021. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

"Planning Database." US Census Bureau.gov, October 8, 2021. https://www.census.gov/topics/research/guidance/planning-databases.html.

Wylezinski, Lukasz S., Coleman R. Harris, Cody N. Heiser, Jamieson D. Gray, and Charles F. Spurlock. "Influence of Social Determinants of Health and County Vaccination Rates on Machine Learning Models to Predict COVID-19 Case Growth in Tennessee," 2021. https://doi.org/10.1101/2021.07.28.21260814.

Richmond, Holly L., Joana Tome, Haresh Rochani, Isaac Chun-Hai Fung, Gulzar H. Shah, and Jessica S. Schwind. "The Use of Penalized Regression Analysis to Identify County-Level Demographic and Socioeconomic Variables Predictive of Increased COVID-19 Cumulative Case Rates in the State of Georgia." International Journal of Environmental Research and Public Health 17, no. 21 (2020): 8036. https://doi.org/10.3390/ijerph17218036.

Mueller, Klaus, and Eric Papenhausen. "Using Demographic Pattern Analysis to Predict COVID-19 Fatalities on the US County Level." *Digital Government: Research and Practice* 2, no. 1 (2021): 1–11. https://doi.org/10.1145/3430196.

"Census Regions and Divisions of the United States" US Census Bureau.gov, 2010. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

"Reporting Covid-19 Vaccinations in the United States." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. Accessed November 6, 2021. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/reporting-vaccinations.html.

Faraway, Julian James. Essay. In Linear Models with R, 130–131. Boca Raton: Chapman &amp; Hall/CRC, 2005.

Introduction to SAS. UCLA: Statistical Consulting Group. From https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/ (accessed November 9, 2021).

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan.

# Appendix

## Tables and Figures

Table 3. Description of the selected 18 predictors

| Variable Name | Unit of Measure | Definition |
|---|---|---|
| % of POP. FULLY VACCINATED | Calculated Percentages | The percentage of the total population in the United States that gets full vaccination |
| pct_Pop_under_5_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that is under five years old |
| pct_Pop5_17_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that is between 5 and 17 years old |
| pct_Pop_45_64_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that is between 45 and 64 years old |
| pct_NH_White_alone_15_19 | Calculated Percentages | The percentage of the ACS population that indicate no Hispanic origin, and their only race as "White" or report entries such as Irish, German, Italian, Lebanese, Arab, Moroccan, or Caucasian |
| pct_NH_Blk_alone_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that indicate no Hispanic origin, and their only race as "Black,African Am., or Negro" or report entries such as African American, Kenyan, Nigerian, or Haitian |
| pct_NH_NHOPI_alone_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that indicate no Hispanic origin, and their only race as "Native Hawaiian", "Guamanian or Chamorro", "Samoan" or "Other Pacific Islander" |
| pct_College_ACS_15_19 | Calculated Percentages | The percentage of the ACS population aged 25 years and over that have a college degree or higher |
| pct_Pov_Univ_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that are not institutional people, people in military group quarters, people in college dormitories, and unrelated individuals under 15 years old |

| | | |
|---|---|---|
| pct_TwoPHealthIns_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that have two or more types of health insurance |
| pct_No_Health_Ins_ACS_15_19 | Calculated Percentages | The percentage of the ACS population that have no health insurance, public or private |
| pct_Vacant_Units_ACS_15_19 | Calculated Percentages | The percentage of all ACS housing units where no one is living regularly at the time of interview; units occupied at the time of interview entirely by persons who are staying two months or less and who have a more permanent residence elsewhere are classified as vacant |
| pct_Renter_Occup_HU_ACS_15_19 | Calculated Percentages | The percentage of ACS occupied housing units that are not owner occupied, whether they are rented or occupied without payment of rent |
| pct_Crowd_Occp_U_ACS_15_19 | Calculated Percentages | The percentage of ACS occupied housing units that have more than 1.01 persons per room |
| pct_NO_PH_SRVC_ACS_15_19 | Calculated Percentages | The percentage of ACS occupied housing units that do not have a working telephone and available service |
| pct_No_Plumb__ACS_15_19 | Calculated Percentages | The percentage of all ACS housing units that do not have complete plumbing facilities |
| pct_HHD_w_Computer_ACS_15_19 | Calculated Percentages | The percentage of ACS households that contain a laptop or desktop computer |
| Region | Categorical Variable | See https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf |

Source: https://www2.census.gov/adrm/PDB/2021/2021_Tract_PDB_Documentation.pdf

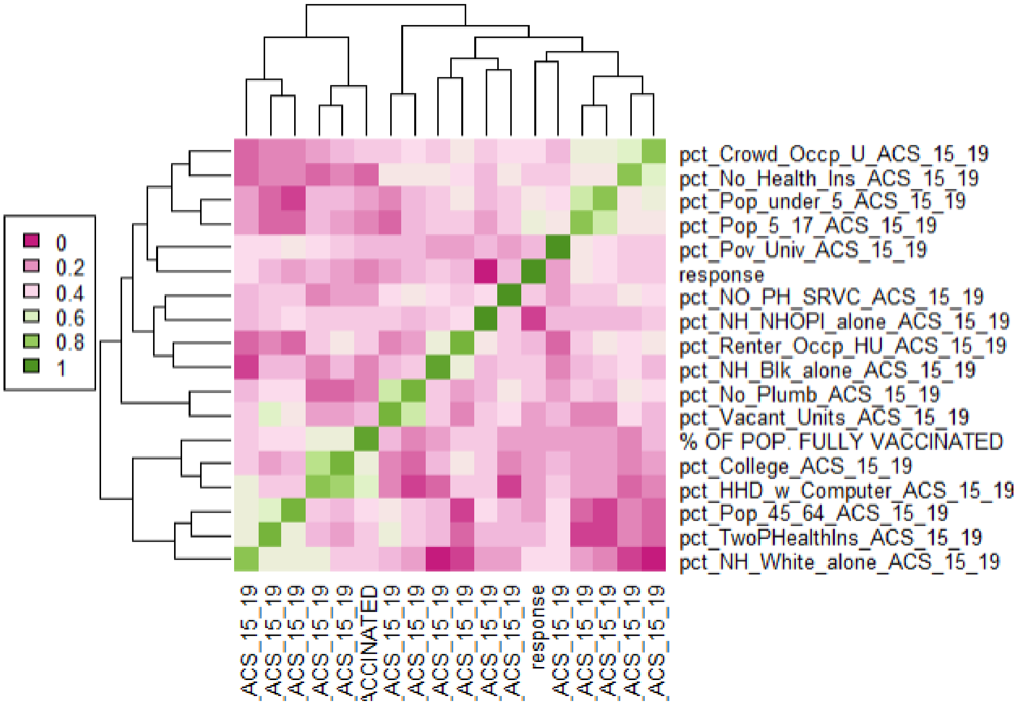


Figure 4. Heatmap correlation analysis of the variables in the reference model

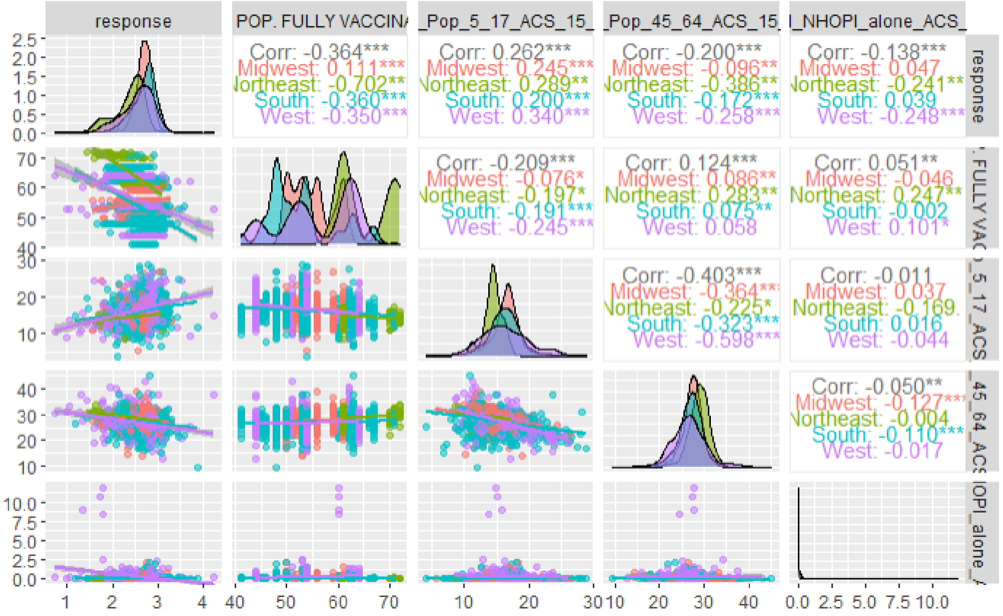


Figure 5. Pairs plot of correlation among response variable and high correlated predictors(correlation > 0.2)

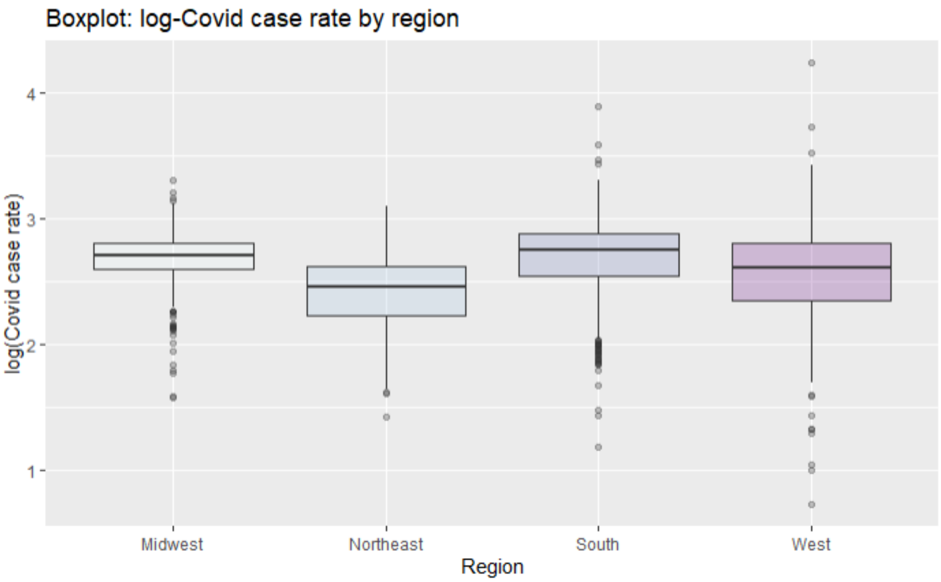


Figure 6. Boxplot of response variable by region

## R Codes

~~~
```{r}
library(xml2)
library(rvest)
library(stringr)
library(dplyr)
~~~
```{r}
# Extract covid case and death counts by state and county
setwd("C:/Users/Mu/Desktop/STAT6214 Applied linear")
url <- read_html("https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/")
str(html_table(url))
state <- html_table(url)[[2]]$State
state_r <- gsub(" ","-",tolower(state))
count_case <- NULL
for (i in state_r) {
  url_s <- paste0("https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/",i)
  html <- read_html(url_s)
  county <- as.data.frame(html_table(html)[[2]])
  county_state <- cbind(str_to_title(i),county)
  count_case <- rbind(count_case,county_state)
}
colnames(count_case)[1] <- "State"
count_case$Cases <- as.integer(gsub(",","",count_case$Cases))
count_case$Deaths <- as.integer(gsub(",","",count_case$Deaths))
str(count_case)

# Extract covid vaccine rate by state
url_vaccine <- read_html("https://usafacts.org/visualizations/covid-vaccine-tracker-states")
str(html_table(url_vaccine))
count_vaccine <- html_table(url_vaccine)[[1]][,1:3]
```

~~~
head(count_vaccine)
count_vaccine$State <- gsub(" ", "-", str_to_title(count_vaccine$STATE))
count_vaccine$`% OF POP. WITH AT LEAST ONE DOSE` <- as.numeric(gsub("%","",count_vaccine$`% OF
POP. WITH AT LEAST ONE DOSE`))
count_vaccine$`% OF POP. FULLY VACCINATED` <- as.numeric(gsub("%","",count_vaccine$`% OF POP.
FULLY VACCINATED`))

# Merge the two tables
count <- dplyr::inner_join(count_case,count_vaccine[,-1])
head(count)
save(count,file = "count.RData")
```
~~~

**Step 1: Dataset Preparation**

~~~
```{r}
#1st round of reducing predictors based on existing research
rm(list=ls())
library(dplyr)
setwd("C:/Users/Mu/Desktop/STAT6214 Applied linear")
myvars <-
c("State_name","County_name","pct_Males_ACS_15_19","pct_Pop_under_5_ACS_15_19","pct_Pop_5_17_ACS_
15_19","pct_Pop_18_24_ACS_15_19","pct_Pop_25_44_ACS_15_19","pct_Pop_45_64_ACS_15_19","pct_Pop_65
plus_ACS_15_19","pct_Hispanic_ACS_15_19","pct_NH_White_alone_ACS_15_19","pct_NH_Blk_alone_ACS_1
5_19","pct_NH_AIAN_alone_ACS_15_19","pct_NH_Asian_alone_ACS_15_19","pct_NH_NHOPI_alone_ACS_1
5_19","pct_NH_SOR_alone_ACS_15_19","pct_Not_HS_Grad_ACS_15_19","pct_College_ACS_15_19","pct_Pov
_Univ_ACS_15_19","pct_Prs_Blw_Pov_Lev_ACS_15_19","pct_One_Health_Ins_ACS_15_19","pct_TwoPHealthI
ns_ACS_15_19","pct_No_Health_Ins_ACS_15_19","pct_Civ_unemp_16p_ACS_15_19","pct_Civ_unemp_16_24_
ACS_15_19","pct_Civ_unemp_25_44_ACS_15_19","pct_Civ_unemp_45_64_ACS_15_19","pct_Pop_Disabled_A
CS_15_19","pct_Children_in_Pov_ACS_15_19","pct_NoHealthIns_U19_ACS_15_19","pct_NoHealthIns1964_AC
S_15_19","pct_NoHealthIns_65P_ACS_15_19","avg_Tot_Prns_in_HHD_ACS_15_19","avg_Agg_HH_INC_ACS_
15_19","pct_Vacant_Units_ACS_15_19","pct_Renter_Occp_HU_ACS_15_19","pct_Crowd_Occp_U_ACS_15_19"
,"pct_NO_PH_SRVC_ACS_15_19","pct_No_Plumb_ACS_15_19","avg_Agg_House_Value_ACS_15_19","pct_H
HD_w_Computer_ACS_15_19","pct_HHD_No_Internet_ACS_15_19","pct_Vacants_CEN_2010","pct_Schl_Enroll
_3_4_ACS_15_19")

#Aggregate 'demography' dataset from tract level to county level
demography = read.csv(file="pdb2021trv3_us.csv",header=TRUE)
demography_sub = demography[,myvars]
demography_sub$avg_Agg_HH_INC_ACS_15_19 <-
as.numeric(gsub("\\$|,","",demography_sub$avg_Agg_HH_INC_ACS_15_19))
demography_sub$avg_Agg_House_Value_ACS_15_19 <-
as.numeric(gsub("\\$|,","",demography_sub$avg_Agg_House_Value_ACS_15_19))
data1 <-
aggregate(demography_sub[,-c(1:2)],by=list(demography_sub$State_name,demography_sub$County_name),mean,
na.rm=TRUE)
data2 <-
aggregate(demography$Tot_Population_ACS_15_19,by=list(demography$State_name,demography$County_name),
sum,na.rm=TRUE)
colnames(data2)[3] <- "total_population"
data <- cbind(data1, data2$total_population)
colnames(data)[1:2] <- c("State","County")
data$State <- gsub(" ", "-", stringr::str_to_title(data$State))
head(data)

#Create categorical variable: region
data$region = ifelse(data$State %in% c("Connecticut","Maine","Massachusetts","New Hampshire","Rhode
Island","Vermont","New Jersey","New York","Pennsylvania"),"Northeast",
~~~

```
        ifelse(data$State %in%
c("Indiana","Illinois","Michigan","Ohio","Wisconsin","Iowa","Kansas","Minnesota","Missouri","Nebraska","North
Dakota","South Dakota"), "Midwest",
        ifelse(data$State %in% c("Arizona","Colorado","Idaho","New
Mexico","Montana","Utah","Nevada","Wyoming","Alaska","California","Hawaii","Oregon","Washington"),"West",
            "South")))
table(data$region, data$State)

#Merge two datasets(covid count & demographic)
load("count.RData")
str(count)
count$County <- gsub(",| of","",count$County) %>% gsub("City","city",.)
Covid <- dplyr::inner_join(count,data,by=c("State","County"))
```
**Step 2: determine response variable**
```{r}
Covid_case_rate <- 100*(Covid$Cases / Covid$`data2$total_population`)
newdata <- Covid[,-c(1,2,3,4,5,6,51)]
Covid_new <- cbind(Covid_case_rate,newdata)
lmod_base <- lm(Covid_case_rate~., Covid_new, na.action=na.exclude)
plot(lmod_base)

#Investigate outliers
Covid_new[c(257,412,1270,2583),]
#Remove influential points and extreme outliers
Covid_a <- Covid_new[-c(257,412,1270,2583),]
lmod_a <- update(lmod_base,data=Covid_a)
plot(lmod_a)
Covid_r <- Covid_new[-c(257,412,1270,2583,2509,2797,448,2670,2650),]
lmod_r <- update(lmod_base,data=Covid_r)
plot(lmod_r)

#Log-transform and compare
par(mfrow=c(1,2))
hist(Covid_r$Covid_case_rate,freq=FALSE,main="Histogram of Covid case rate",xlab="Covid case rate",ylim =
c(0,0.12))
lines(density(Covid_r$Covid_case_rate))
hist(log(Covid_r$Covid_case_rate), freq = FALSE, main = "Histogram of log(Covid case rate)", xlab = "log(Covid
case rate)",ylim = c(0,2.0))
lines(density(log(Covid_r$Covid_case_rate)))

#Create response variable
Covid_case_rate_r <- ifelse(Covid_new$Covid_case_rate == 0,
Covid_new$Covid_case_rate+0.000001,Covid_new$Covid_case_rate)
```
**Step3: Further reduce predictors**
```{r}
#2nd round of reducing predictors by Lasso Regression
library(glmnet)
set.seed(123)
options(na.action = "na.pass")
x <- model.matrix(Covid_case_rate~., Covid_new)[,-1]
y <- log(Covid_case_rate_r)
cv <- cv.glmnet(x,y, alpha = 1)
cv$lambda.min
model <- glmnet(x, y, alpha = 1, lambda = cv$lambda.min)
```

~~~
coef(model)

#Determine predictors for linear regression model
myvars2 <- c("% OF POP. FULLY VACCINATED","pct_Pop_under_5_ACS_15_19","pct_Pop_5_17_ACS_15_19","pct_Pop_45_64_ACS_15_19","pct_NH_White_alone_ACS_15_19","pct_NH_Blk_alone_ACS_15_19","pct_NH_NHOPI_alone_ACS_15_19","pct_College_ACS_15_19","pct_Pov_Univ_ACS_15_19","pct_TwoPHealthIns_ACS_15_19","pct_No_Health_Ins_ACS_15_19","pct_Vacant_Units_ACS_15_19","pct_Renter_Occp_HU_ACS_15_19","pct_Crowd_Occp_U_ACS_15_19","pct_NO_PH_SRVC_ACS_15_19","pct_No_Plumb_ACS_15_19","pct_HHD_w_Computer_ACS_15_19","region")
new_data <- Covid_new[,myvars2]
response <- log(Covid_case_rate_r)
fdata <- cbind(response,new_data)
```
~~~

**Step 4: Exploratory Data Analysis(EDA)**

~~~
```{r}
dim(fdata)
str(fdata)
fdata$region <- as.factor(fdata$region)
summary(fdata)

#Check missing value
sort(Covid$pct_Pop_5_17_ACS_15_19)
#sort(Covid$pct_NH_Blk_alone_ACS_15_19)
#sort(Covid$pct_NH_NHOPI_alone_ACS_15_19)
#sort(Covid$pct_No_Health_Ins_ACS_15_19)
#sort(Covid$pct_Crowd_Occp_U_ACS_15_19)
#sort(Covid$pct_NO_PH_SRVC_ACS_15_19)
#sort(Covid$pct_No_Plumb_ACS_15_19)

#Check correlation
corr_score <- cor(fdata[,-19])
library(ggplot2)
library(RColorBrewer)
library(GGally)
heatmap(corr_score,scale="column",col=colorRampPalette(brewer.pal(8,"PiYG"))(18))
legend(x="left",legend=c(0,0.2,0.4,0.6,0.8,1.0),cex=0.8,
       fill=colorRampPalette(brewer.pal(8,"PiYG"))(6))
cor(as.matrix(fdata[,-19]))
ggpairs(fdata[fdata$response >= 0,],columns=c(1,2,4,5,8),aes(color=region,alpha=0.5),
        lower=list(continuous="smooth"))

#Understand categorical feature(region)
ggplot(fdata[fdata$response >= 0,],aes(x=region,y=response,fill=region),)+
  geom_boxplot(alpha=0.3)+
  theme(legend.position = "none")+
  scale_fill_brewer(palette="BuPu")+
  labs(title = "Boxplot: log-Covid case rate by region",x="Region",y="log(Covid case rate)")
```
~~~

**Step 5: Fit linear regression reference model**

~~~
```{r}
lmod_ref <- lm(response~.,fdata)
summary(lmod_ref)
#Assumption checking
plot(lmod_ref)
require(lmtest)
~~~

```
bptest(lmod_ref)
```

**Step 6: Investigate outliers again(response variable change compared to the last one)**

```{r}
Covid_new[c(548,550,77),]
fdata_r <- Covid_new[-c(548,550,77,72),c("Covid_case_rate",myvars2)] #this is my final dataset
#Recreate response variable
fdata_r$response <- log(ifelse(fdata_r$Covid_case_rate == 0,
fdata_r$Covid_case_rate+0.000001,fdata_r$Covid_case_rate))
fdata_r$region <- as.factor(fdata_r$region)
lmod_ref_r <- update(lmod_ref,data=fdata_r[,-1])
plot(lmod_ref_r)
```

```{r}
summary(lmod_ref_r) #this is my reference model for comparison in the future
summary(lmod_ref)
```

**Step 7: Explore interaction and polynomial terms**

```{r}
#Explore interaction terms
ggplot(data=fdata_r, aes(x=`% OF POP. FULLY VACCINATED`, y=response, group=region))+
   geom_line(size=1, aes(color=region))+
   ylab("log(Covid Case rate)")+
   xlab("Pct of population fully vaccinated")+
   ggtitle("Interaction on region and fully vaccinated pct")
ggpairs(fdata_r,columns=c(20,4),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,6),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,7),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,8),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,9),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,12),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,13),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,15),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,17),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
ggpairs(fdata_r,columns=c(20,18),aes(color=region,alpha=0.5),
    lower=list(continuous="smooth"))
```

Above are the continuous explanatory variables I selected that may have interactions with region.

```{r}
#Explore polynomial terms
#Both the residual vs fitted plot and scale-location plot suggest a quadratic term
summary(lm(response~I(`% OF POP. FULLY VACCINATED`^2),fdata_r[,-1]))
#summary(lm(response~I(pct_Pop_5_17_ACS_15_19^2),fdata_r[,-1]))
#summary(lm(response~I(pct_Pop_45_64_ACS_15_19^2),fdata_r[,-1]))
#summary(lm(response~I(pct_College_ACS_15_19^2),fdata_r[,-1]))
```

The selective criterion is p-value < 0.05 and R squared value > 4%

**Step8: Variable Selection in Linear Regression Model**

~~~
```{r}
#Stepwise
null <- lm(response~1,data=fdata_r[,-1]) # define intercept_only model
full <- lm(response ~. +region:`% OF POP. FULLY VACCINATED`+
        region:pct_Pop_5_17_ACS_15_19+
        region:pct_NH_White_alone_ACS_15_19+
        region:pct_NH_Blk_alone_ACS_15_19+
        region:pct_NH_NHOPI_alone_ACS_15_19+
        region:pct_College_ACS_15_19+
        region:pct_No_Health_Ins_ACS_15_19+
        region:pct_Vacant_Units_ACS_15_19+
        region:pct_Crowd_Occp_U_ACS_15_19+
        region:pct_No_Plumb_ACS_15_19+
        I(`% OF POP. FULLY VACCINATED`^2)+
        I(pct_Pop_5_17_ACS_15_19^2)+
        I(pct_Pop_45_64_ACS_15_19^2)+
        I(pct_College_ACS_15_19^2),data=fdata_r[,-1])
lmod_stepwise <- step(null,direction = "both",scope=formula(full),trace=0)
summary(lmod_stepwise)
```
~~~

**Step 9: Further adjustment**

~~~
```{r}
#Remove statistically insignificant(at 5% level) predictors
lmod_f <- update(lmod_stepwise,.~. -pct_Crowd_Occp_U_ACS_15_19-
            pct_NH_NHOPI_alone_ACS_15_19-
            pct_Vacant_Units_ACS_15_19)
summary(lmod_f)
library(car)
vif(lmod_stepwise)
lmod_f2 <- update(lmod_stepwise,.~. -I(`% OF POP. FULLY VACCINATED`^2)
           )
summary(lmod_f2)
vif(lmod_f2)
lmod_f3 <- update(lmod_f2,.~. -region:`% OF POP. FULLY VACCINATED`
           )
summary(lmod_f3)
vif(lmod_f3)
```
~~~

**Step10: Model Comparison**

~~~
```{r}
plot(lmod_f2)
plot(lmod_ref_r)
summary(lmod_f2) #final model
summary(lmod_ref_r)
```
~~~