# Contents

# 1 Abstract

# 2 Introduction

For the past 5 months, the coronavirus (Covid-19) pandemic has been raging all over the world. According to the World Health Organization, there are more than 9.44 million confirmed cases worldwide, and the total number of death cases is as high as 483,000 (World Health Organization (2020)). Therefore, we are interested in finding the factors that can potentially explain the Covid-19 mortality rate, as well as the degree to which those factors impact the mortality rate. Since the US has a large number of confirmed cases, we would analyze the Covid-19 cases from US from a statistical standpoint in the below report. By fitting and comparing different models, we hope to gain some insights into the causes of mortality rate of this pandemic.

# 3 Dataset Description

The first dataset we found was the US Covid-19 data, which recorded the mortality rate under each county (USAFacts (2020)). Another machine-readable dataset (Killeen et al. (2020)) we have used contains socioeconomic, demographic, health care, education and transit data for each county in the 50 states in US. In total, there are 347 different factors in this dataset, such as the population estimate, migration rate, number of females, number of hospitals. Furthermore, both datasets include a key called Federal Information Processing Standard Publication (FIPS) code, which is a five-digit code uniquely identifying each area. Thus, by joining these two data sets together with Excel, we obtained our combined data called alldata.

# 4 investigation: single model fitting -> significant

# 5 Model Selection

```
##   (Intercept) ElderlyRate RegionNortheast RegionSouth RegionWest Specialist
## 1           1           1               0           0          0          0
## 1           1           0               0           0          0          0
## 2           1           1               0           0          1          0
## 2           1           1               0           0          0          0
## 3           1           1               0           1          1          0
## 3           1           1               0           0          0          1
## 4           1           1               0           1          1          0
## 4           1           1               0           0          1          1
## 5           1           1               0           1          1          1
## 5           1           1               0           1          1          0
## 6           1           1               0           1          1          1
## 6           1           1               0           1          1          0
## 7           1           1               1           1          1          1
## 7           1           1               0           1          1          1
##   populationSmall Physician_rate Unemployment_rate_2018 Poverty Migration Nurse
## 1               0              0                      0       0         0     0
## 1               0              0                      0       0         1     0
## 2               0              0                      0       0         0     0
## 2               0              1                      0       0         0     0
## 3               0              0                      0       0         0     0
## 3               0              0                      0       0         1     0
## 4               0              0                      1       0         0     0
```

```
## 4               0             0                     0       0         1     0
## 5               0             0                     0       0         1     0
## 5               0             0                     0       0         1     1
## 6               0             0                     1       0         1     0
## 6               0             0                     1       0         1     1
## 7               0             0                     1       0         1     0
## 7               1             0                     1       0         1     0
##          cp       adjr2
## 1 16.752134 0.014242392
## 1 33.849500 0.004577108
## 2 14.784746 0.015916025
## 2 15.187326 0.015688311
## 3 11.194281 0.018510199
## 3 12.616925 0.017705035
## 4 10.038394 0.019728699
## 4 10.152792 0.019663916
## 5  7.373784 0.021803483
## 5  8.298368 0.021279596
## 6  6.506799 0.022861515
## 6  7.351198 0.022382784
## 7  7.242004 0.023011889
## 7  7.661512 0.022773912
```

# 6 Assumption Checking

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate + factor(population) + factor(Region))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8736 -0.4927  0.1368  0.6701  3.5247
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.66345    0.10047 -36.464  < 2e-16 ***
## ElderlyRate               2.37758    0.51978   4.574 5.12e-06 ***
## factor(population)Small   0.04445    0.06052   0.734  0.46284
## factor(Region)Northeast -0.07284    0.09176  -0.794  0.42741
## factor(Region)South     -0.15243    0.05933  -2.569  0.01027 *
## factor(Region)West      -0.25399    0.09083  -2.796  0.00523 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.037 on 1725 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.02105,    Adjusted R-squared:  0.01821
## F-statistic: 7.417 on 5 and 1725 DF,  p-value: 6.804e-07
```

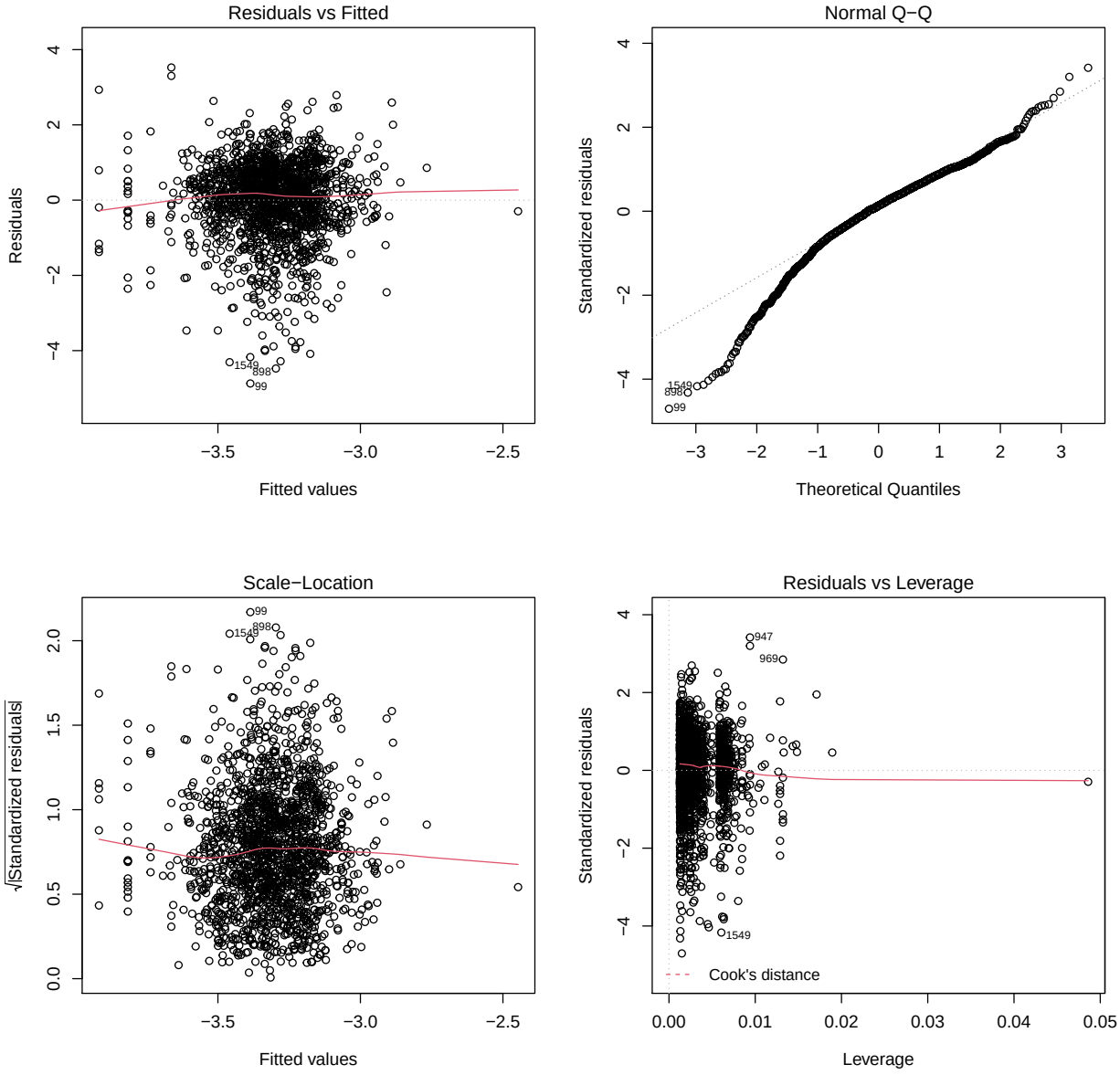


Figure 1: Plot for Model Checking

### 6.1 Mean of 0:

Looking at the plot of Residuals vs. Fitted values the residuals do appear randomly scatter around the value 0. As such, it appears to satisfy the assumption of mean of 0.

### 6.2 Independence:

Again the plot of Residuals vs. Fitted values shows the residuals randomly scattered with no apparent trend and thus satisfied.

### 6.3 Constant Variance:

The Scale location plot suggests that the majority of residuals are randomly scattered evenly within an upper and lower band around the value 0, which indicates a constant variance. However, since there is a few larger residuals (such as observation 31, 90 and 98) on the top, the red line bulges in the middle of the majority data point, causing a slightly increase and decrease in variability. Thus, it is considered fairly satisfied.

### 6.4 Normality:

The Q-Q plot overall appears good, since most of the residuals are lying on a straight line. However, the normality assumption can only be considered as fairly satisfied, since there are some point at the tails are off the line, suggesting some potential outliers.

### 6.5 Outliers:

It could be seen that there are some potential outliers, as we could find that there are some points off the straight at the tails in Q-Q plot, which appears to be outliers.

In terms of the leverage and Standard Residual plot, it could seen that there is a point on the very right side and a few other points slightly upper and lower compared to the rest of majority of observations. As such, those points are considered to have a high leverage, whereas they are all within the Cook's Distance. Thus, these points are considered not to be influential.

## 7 Results

## 8 Limitation of Study

## 9 Conclusion

# Appendix

# Reference

# R Code

Killeen, Benjamin D., Jie Ying Wu, Kinjal Shah, Anna Zapaishchykova, Philipp Nikutta, Aniruddha Tamhane, Shreya Chakraborty, et al. 2020. "A County-Level Dataset for Informing the United States' Response to COVID-19," April.

USAFacts. 2020. "Coronavirus Locations: COVID-19 Map by County and State," March. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

World Health Organization. 2020. "Coronavirus Disease (Covid-19) Outbreak Situation." https://www.who.int/emergencies/diseases/novel-coronavirus-2019.