Contents

1 Abstract

The purpose of this study is to identify the factors that prominently affect the U.S. COVID-19 mortality rate. We used a dataset comprising various information of 1739 different US counties, which contains the variable of interest- the mortality rate of each county, as well as over 300 potential explanatory variables, including but not limited to information on demographicss and healthcare. The question of interest is to identify the most appropriate explanatory variables and to find the most suitable and statistically significant linear model to reflect these variables' effects on the mortality rate. We started by identifying the best 9 explanatory variables that were the most reasonable and statistically significant in their respective single models. Then, using the automatic model selection method, we identified the best 2 models for each model size. Next, by comparing their R^2_{adj}, AIC values, BIC values, C_p values, and taking multicollinearity into consideration, we selected the final model consisting of 4 explanatory variables, including the percentage of the population above 65 years old in the county, the total number of specialist in the county, the incoming number of International immigrants in the county for the past year, and the large-scale region to which the county belongs, which is a categorical variable of 4 large region bounds. Based on our dataset, the linear model consisting of these 4 variables best represents the trend and variation of the mortality rate. Further studies on the US COVID-19 mortality rate is recommended where more suitable explanatory variables are included in order to increase the explanatory power of the model.

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 U.S. has a large number of confirmed cases (USAFacts (2020)), we would use the COVID-19 cases from the U.S. and the corresponding counties' data (Killeen et al. (2020)) to analyze the causes of the mortality rate for this pandemic.

From our previous report (Ye, Wang, and Liu (2020)), we have found some statistically significant variables to the mortality rate of COVID-19, which are the factors of region, percentage of elder population, specialists and physicians. Thus, in this report, we would first extend our investigation, specifically discovering reasonable factors that are statistically significant to the COVID-19 mortality rate. Based on all these results, we would perform statistical analysis in comparing and fitting different models. Ultimately, we wish to gain some important insights from the mortality rate of COVID-19 by finding the best statistically significant model.

3 Dataset Description

The first dataset we found was the U.S. COVID-19 data, which recorded the mortality rate under each county (USAFacts (2020)). Another machine-readable dataset (Killeen et al. (2020)) we used contains socioeconomic, demographic, health care, education and transit data for each county in the U.S. 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 the Federal Information Processing Standard Publication (FIPS) code, which is a five-digit code that uniquely identifies each county. Thus, by joining these two data sets together in Excel, we obtain our master data.

Furthermore, we are interested in the mortality rate and have set it to be our response variable. To satisfy the normality assumption, we created a column called Mortality, which is the natural logarithm of the mortality rate of each county. As such, we could use Mortality to be our response variable when computing the statistic models.

4 Investigation

In the previous report, we have identified 6 individually significant explanatory variables, including Specialist, Physician_rate, ElderlyRate, Region, Population, Unemployment_rate_2018. Due to the low R^2_{adj} value in those models, we searched in our dataset for additional explanatory variables that reasonably affect the COVID-19 mortality rate. Among the variables that are reasonable to be included, we selected 3 of them whose corresponding single-variable models were tested significantly at $\alpha = 0.05$ level of significance. These additional variables are Poverty, Migration, and Nurse, making a total of 9 variables selected.

The p-values of the global F-tests for the corresponding 9 single-model fittings are listed below. Full single-variable model summaries are included in the appendix.

Table 1: P-values for Selected Explanatory Variables

Variable	p_value
ElderlyRate	0.0000007
Region	0.0247100
Specialist	0.0274500
population	0.0144000
Physician_rate	0.0460600
Unemployment_rate_2018	0.0484100
Poverty	0.0250100
Migration	0.0027500
Nurse	0.0140400

As shown above, all 9 of those explanatory variables are statistically significant; we then select the best model based on those 9 variables. Firstly, we will observe the correlation matrix and VIF of those variables to spot red flags for multicollinearity.

Table 2: Correlation Matrix

	Elderly	Specilist	Physician	Unemployment	Poverty	Migration	Nurse
ElderlyRate	1.000	-0.389	0.026	0.130	-0.371	-0.266	-0.391
Specialist	-0.389	1.000	0.127	-0.049	0.958	0.855	0.983
Physician_rate	0.026	0.127	1.000	0.028	0.041	0.080	0.106
Unemployment_rate_2018	0.130	-0.049	0.028	1.000	-0.026	-0.045	-0.055
Poverty	-0.371	0.958	0.041	-0.026	1.000	0.835	0.939
Migration	-0.266	0.855	0.080	-0.045	0.835	1.000	0.883
Nurse	-0.391	0.983	0.106	-0.055	0.939	0.883	1.000

Table 3: VIF Values

	GVIF	Df	GVIF^(1/(2*Df))
ElderlyRate	1.357739	1	1.165221
factor(Region)	1.924274	3	1.115264
Specialist	50.404208	1	7.099592
factor(population)	1.311527	1	1.145219
Physician_rate	1.986869	1	1.409563
Unemployment_rate_2018	1.091201	1	1.044606
Poverty	13.701774	1	3.701591
Migration	5.028916	1	2.242524
Nurse	41.407467	1	6.434863

In the correlation matrix, we see signs of multicollinearity issues in the following pairs of variables: Specialist & Poverty, Specialist & Migration, Specialist & Nurse, Poverty & Migration, Poverty & Nurse, Migration & Nurse, using 0.8 as a cutoff.

In terms of the GVIF values with a value 5 as a cutoff, we see signs of multicollinearity issues in Specialist, Poverty, Migration, and Nurse.

5 Model Selection

5.1 Automatic Model Selection Method

To start our model selection process, we use the stepwise selection method first. Starting with the full model of all 9 variables, we begin the stepwise algorithm to find the best model.

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate + factor(Region) + Specialist +
##     Physician_rate + Unemployment_rate_2018 + Migration)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8599 -0.5010  0.1324  0.6692  3.5356
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.051e+00  2.057e-01 -19.690  < 2e-16 ***
## ElderlyRate               2.519e+00  5.377e-01   4.684 3.04e-06 ***
## factor(Region)Northeast -1.827e-01  1.056e-01  -1.730  0.08386 .
## factor(Region)South     -1.338e-01  6.226e-02  -2.149  0.03176 *
## factor(Region)West      -2.880e-01  9.211e-02  -3.127  0.00180 **
## Specialist                3.348e-05  1.599e-05   2.095  0.03636 *
## Physician_rate            9.720e-04  6.566e-04   1.480  0.13894
## Unemployment_rate_2018    3.337e-02  1.993e-02   1.674  0.09422 .
## Migration                -2.206e-05  8.203e-06  -2.690  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.034 on 1722 degrees of freedom
```

```
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.0282, Adjusted R-squared:  0.02369
## F-statistic: 6.247 on 8 and 1722 DF,  p-value: 5.454e-08
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Mortality ~ ElderlyRate + factor(Region) + Specialist + factor(population) +
##     Physician_rate + Unemployment_rate_2018 + Poverty + Migration +
##     Nurse
##
## Final Model:
## Mortality ~ ElderlyRate + factor(Region) + Specialist + Physician_rate +
##     Unemployment_rate_2018 + Migration
##
##
##                    Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                                         1719   1839.327 129.0726
## 2             - Poverty  1 0.03228038      1720   1839.359 127.1030
## 3               - Nurse  1 0.06151220      1721   1839.421 125.1609
## 4 - factor(population)  1 1.03274628      1722   1840.454 124.1325
```

The stepwise method gives the model: Mortality ~ ElderlyRate + factor(Region) + Specialist + Physician_rate + Unemployment_rate_2018 + Migration) as the final model. However, we noticed the p-value for Physician_rate = 0.13894 $> \alpha = 0.05$ and the p-value for Unemployment_rate_2018 = 0.09422 $> \alpha = 0.05$, which do not indicate statistical significance. Thus, the model is not the most suitable model for our dataset.

5.2 Model Selection Process with All Possible Combinations

In order to find the best fitting model, we performed another model selection process by fitting all possible combinations.

Table 4: Variable Reference

Num	Variable	Num.1	Variable.1	Num.2	Variable.2
0	Intercept	4	RegionWest	8	Unemployment_Rate_2018
1	ElderlyRate	5	Specialist	9	Poverty
2	RegionNorthEast	6	populationSmall	10	Migration
3	RegionSouth	7	PhysicianRate	11	Nurse

Table 5: Model Selection with CP values

	0	1	2	3	4	5	6	7	8	9	10	11	cp	adjr2
1	1	1	0	0	0	0	0	0	0	0	0	0	16.752134	0.0142424
1	1	0	0	0	0	0	0	0	0	0	1	0	33.849500	0.0045771
2	1	1	0	0	1	0	0	0	0	0	0	0	14.784746	0.0159160
2	1	1	0	0	0	0	0	1	0	0	0	0	15.187326	0.0156883
3	1	1	0	1	1	0	0	0	0	0	0	0	11.194281	0.0185102
3	1	1	0	0	0	1	0	0	0	0	1	0	12.616925	0.0177050
4	1	1	0	1	1	0	0	0	1	0	0	0	10.038394	0.0197287
4	1	1	0	0	1	1	0	0	0	0	1	0	10.152792	0.0196639
5	1	1	0	1	1	1	0	0	0	0	1	0	7.373784	0.0218035
5	1	1	0	1	1	0	0	0	0	0	1	1	8.298368	0.0212796
6	1	1	0	1	1	1	0	0	1	0	1	0	6.506799	0.0228615
6	1	1	0	1	1	0	0	0	1	0	1	1	7.351198	0.0223828
7	1	1	1	1	1	1	0	0	1	0	1	0	7.242004	0.0230119
7	1	1	0	1	1	1	1	0	1	0	1	0	7.661512	0.0227739
8	1	1	1	1	1	1	0	1	1	0	1	0	7.052842	0.0236871
8	1	1	1	1	1	0	0	1	1	0	1	1	7.399981	0.0234901
9	1	1	1	1	1	1	1	1	1	0	1	0	8.087657	0.0236680
9	1	1	1	1	1	0	1	1	1	0	1	1	8.363984	0.0235110
10	1	1	1	1	1	1	1	1	1	0	1	1	10.030169	0.0231330
10	1	1	1	1	1	1	1	1	1	1	1	0	10.069118	0.0231109
11	1	1	1	1	1	1	1	1	1	1	1	1	12.000000	0.0225819

From the result, we noticed that models containing at least 4 variables have fair Cp values, whereas models containing 3 variables or less have Cp values of at least 11, which is way above the acceptable range. In addition, the R^2_{adj} value increases with the number of variables in the model. Hence, we only consider models of 4 variables and above, as listed below.

Table 6: Automatic Selection Model Summary

Model	Regression_Equation
1	Log of Mortality Rate vs. ElderlyRate + Specialist + Migration + Region
2	Log of Mortality Rate vs. ElderlyRate + Nurse + Migration + Region
3	Log of Mortality Rate vs. ElderlyRate + Specialist + Migration + Unemployment_rate_2018 + Region
4	Log of Mortality Rate vs. ElderlyRate + Nurse + Migration + Unemployment_rate_2018 + Region
5	Log of Mortality Rate vs. ElderlyRate + Specialist + Migration + Unemployment_rate_2018 + Region + population
6	Log of Mortality Rate vs. ElderlyRate + Specialist + Migration + Unemployment_rate_2018 + Physician_rate + Region
7	Log of Mortality Rate vs. ElderlyRate + Nurse + Migration + Unemployment_rate_2018 + Physician_rate + Region
8	Log of Mortality Rate vs. ElderlyRate + Specialist + Migration + Unemployment_rate_2018 + Physician_rate + Region + population
9	Log of Mortality Rate vs. ElderlyRate + Nurse + Migration + Unemployment_rate_2018 + Physician_rate + Region + population
10	Log of Mortality Rate vs. ElderlyRate + Specialist + Nurse + Migration + Unemployment_rate_2018 + Physician_rate + Region + population
11	Log of Mortality Rate vs. ElderlyRate + Specialist + Migration + Unemployment_rate_2018 + Physician_rate + Poverty + Region + population
12	Log of Mortality Rate vs. ElderlyRate + Specialist + Nurse + Migration + Unemployment_rate_2018 + Physician_rate + Poverty + Region + population

Table 7: AIC and BIC values

	df	AIC	BIC
model1	8	5039.856	5083.508
model2	8	5040.714	5084.366
model3	9	5038.699	5087.807
model4	9	5039.464	5088.573
model5	10	5040.188	5094.752
model6	10	5038.498	5093.062
model7	10	5038.847	5093.412
model8	11	5039.526	5099.547
model9	11	5039.804	5099.825
model10	12	5041.468	5106.946
model11	12	5041.507	5106.985
model12	13	5043.438	5114.372

By comparing this result, we notice that model1, model2, model3, and model4 generally have the lowest AIC and BIC values. Since model1 is nested in model3, and model2 is nested in model4, it is proper to use backward elimination procedure here. We start with a comparison between model1 and model3.

5.3 Backward elimination procedure

Step 1: start by fitting Model3: Mortality~ElderlyRate + Specialist + Migration + Unemployment_rate_2018 + factor(Region), and consider the partial t tests for each explanatory variable.

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate + Specialist + Migration +
##     Unemployment_rate_2018 + factor(Region))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8914 -0.4922  0.1263  0.6632  3.5146
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.810e+00  1.263e-01 -30.179  < 2e-16 ***
## ElderlyRate              2.566e+00  5.370e-01   4.779 1.92e-06 ***
## Specialist               3.561e-05  1.593e-05   2.236  0.02549 *
## Migration               -2.253e-05  8.200e-06  -2.747  0.00607 **
## Unemployment_rate_2018   3.528e-02  1.990e-02   1.773  0.07634 .
## factor(Region)Northeast -1.016e-01  9.036e-02  -1.125  0.26080
## factor(Region)South     -1.588e-01  5.995e-02  -2.648  0.00816 **
## factor(Region)West      -2.942e-01  9.205e-02  -3.196  0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.034 on 1723 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.02697,    Adjusted R-squared:  0.02301
## F-statistic: 6.821 on 7 and 1723 DF,  p-value: 5.179e-08
```

According to the result, the p-value for Unemployment_rate_2018 $= 0.07634 > \alpha = 0.05$ which means Unemployment_rate_2018 is not statistically significant. Then, we omit the variable Unemployment_rate_2018 from the model.

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate + Specialist + Migration +
##     factor(Region))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8565 -0.5003  0.1307  0.6712  3.5013
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.700e+00  1.099e-01 -33.666  < 2e-16 ***
## ElderlyRate              2.699e+00  5.320e-01   5.073 4.33e-07 ***
## Specialist               3.641e-05  1.593e-05   2.285  0.02241 *
## Migration               -2.297e-05  8.201e-06  -2.801  0.00516 **
## factor(Region)Northeast -8.959e-02  9.016e-02  -0.994  0.32053
## factor(Region)South     -1.422e-01  5.926e-02  -2.400  0.01650 *
## factor(Region)West      -2.641e-01  9.054e-02  -2.917  0.00358 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.035 on 1724 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.02519,    Adjusted R-squared:  0.0218
## F-statistic: 7.425 on 6 and 1724 DF,  p-value: 7.282e-08
```

Step 2: Repeat the procedure in step 1 with a simplified model excluding the Unemployment_rate_2018, which is the our model1. The p-value for ElderlyRate = 4.33e-07 $< \alpha = 0.05$,the p-value for Specialist = 0.02241 $< \alpha = 0.05$, p-value for Migration = 0.00516 $< \alpha = 0.05$, and the p-values for most categories under Region are under $\alpha = 0.05$. This suggests that all variables inside this model are statistically significant. Thus we cannot further simplify this model and this model would be one of our best models to consider.

5.4 Extra Sum of Squares Principle

Next, we use the extra sum of squares principle to compare between model2 and model4 since they have a nested relationship. So, we performed a hypothesis test on the variable Unemployment_rate_2018.

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate + Nurse + Migration + factor(Region))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8555 -0.5041  0.1376  0.6711  3.4571
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.698e+00  1.109e-01 -33.352  < 2e-16 ***
## ElderlyRate              2.695e+00  5.371e-01   5.017 5.79e-07 ***
## Nurse                    2.056e-04  9.837e-05   2.090  0.03680 *
## Migration               -2.390e-05  9.150e-06  -2.612  0.00907 **
## factor(Region)Northeast -9.286e-02  9.037e-02  -1.028  0.30427
## factor(Region)South     -1.446e-01  5.924e-02  -2.441  0.01475 *
## factor(Region)West      -2.611e-01  9.052e-02  -2.885  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.035 on 1724 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.02471,    Adjusted R-squared:  0.02131
## F-statistic: 7.279 on 6 and 1724 DF,  p-value: 1.076e-07
```

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate + Nurse + Migration + Unemployment_rate_2018 +
##     factor(Region))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8908 -0.4927  0.1311  0.6602  3.4712
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.811e+00  1.273e-01 -29.942  < 2e-16 ***
```

```
## ElderlyRate              2.562e+00  5.418e-01   4.730 2.43e-06 ***
## Nurse                    2.023e-04  9.832e-05   2.058  0.03975 *
## Migration               -2.355e-05  9.146e-06  -2.575  0.01012 *
## Unemployment_rate_2018   3.580e-02  1.990e-02   1.799  0.07214 .
## factor(Region)Northeast -1.051e-01  9.056e-02  -1.161  0.24582
## factor(Region)South     -1.613e-01  5.993e-02  -2.692  0.00717 **
## factor(Region)West      -2.917e-01  9.204e-02  -3.169  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.034 on 1723 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.02653,    Adjusted R-squared:  0.02258
## F-statistic: 6.709 on 7 and 1723 DF,  p-value: 7.293e-08
```

Hypothesis test:

(1) H_0: $\beta_{unemployment} = 0$ vs. H_a: $\beta_{unemployment} \neq 0$

(2) $f_0 = \frac{(1.035-1.034)/(1724-1723)}{1.034/1723} = 1.66634$

```
## [1] 0.1969227
```

(3) p-value $= P(F_{1,1723} > 1.66634) = 0.197$

(4) Since p-value $> \alpha = 0.05$, then we do not reject H_0 at a 5% level of significance, which suggests the addition variable Unemployment_rate_2018 is not useful.

Thus, we choose model2 above model4.

5.5 Log of Mortality Rate with Interaction Term

Then, we compare the results from model1 and model2. Notice that model1 and model2 have the same standard error, but model1 has a lower p-value and higher R^2_{adj} compared to model2. Thus, we choose model1 to further investigate on. Moreover, we want to investigate on the collinearity between Specialist and Migration since they have a correlation coefficient of 0.855.

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate + Specialist + Migration +
##     factor(Region) + Specialist * Migration)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8525 -0.5057  0.1349  0.6702  3.4808
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.715e+00  1.113e-01 -33.373  < 2e-16 ***
## ElderlyRate               2.768e+00  5.384e-01   5.141 3.04e-07 ***
## Specialist                3.605e-05  1.594e-05   2.262   0.0238 *
## Migration                -1.429e-05  1.322e-05  -1.081   0.2797
## factor(Region)Northeast -9.857e-02  9.081e-02  -1.086   0.2778
## factor(Region)South     -1.428e-01  5.927e-02  -2.410   0.0160 *
## factor(Region)West      -2.676e-01  9.064e-02  -2.952   0.0032 **
## Specialist:Migration    -1.801e-10  2.151e-10  -0.837   0.4025
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.035 on 1723 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.02559,    Adjusted R-squared:  0.02163
## F-statistic: 6.463 on 7 and 1723 DF,  p-value: 1.548e-07
```

```
## Analysis of Variance Table
##
## Response: Mortality
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## ElderlyRate              1   28.05 28.0523 26.1915 3.437e-07 ***
## Specialist               1    0.06  0.0583  0.0545  0.815509
## Migration                1    8.65  8.6463  8.0728  0.004546 **
## factor(Region)           3   10.95  3.6492  3.4071  0.016987 *
## Specialist:Migration     1    0.75  0.7511  0.7013  0.402473
## Residuals             1723 1845.41  1.0710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the summary with the addition of interaction term between Specialist and Migration, the p-value for the interaction term itself is $0.4025 > \alpha = 0.05$, proving the term to be not significant. Besides, the p-values for Migration increases to $0.2797 > \alpha = 0.05$ which do not indicate statistically significant. The overall p-value = 1.548e-07 $< \alpha = 0.05$, which is higher than that of model1. The adjusted $R^2 = 0.02163$ is lower than 0.0218 from model1, meaning that the model with the interaction term explains less variability of the response variable. Moreover, the global F-statistic = 6.463, which is lower than 7.425 from model1. Thus, adding the interaction term does not contribute to our study, and we decide to keep model1.

We notice from the previous analysis that Specialist and Migration have a correlation of 0.855, which shows they are highly correlated. Also, Specialist has a VIF value of more than 50, and Migration has a VIF value of 5 indicating multicollinearity issue between these two variables. However, for the current dataset, we choose to stay with model1 as our best model and will need to investigate on this multicollinearity issue using a larger dataset.

Based on the method we used and the models we discussed, we select model1: Mortality ~ ElderlyRate + Specialist + Migration + factor(Region) as the most suitable model for the given COVID-19 dataset.

6 Assumption Checking

6.1 Mean of Zero

Looking at the plot of Residuals vs. Fitted values, the residuals do appear to randomly scatter around the value 0, whereas there are some points far off (such as observation 99, 898 & 1114) the majority of the data points. However, the model overall appears to fairly satisfy the assumption of mean of zero.

6.2 Independence

In the plot of Residuals vs. Fitted values, most of the residuals are concentrated. Therefore, no apparent trend is detected and the assumption of Independence appears to be satisfied.

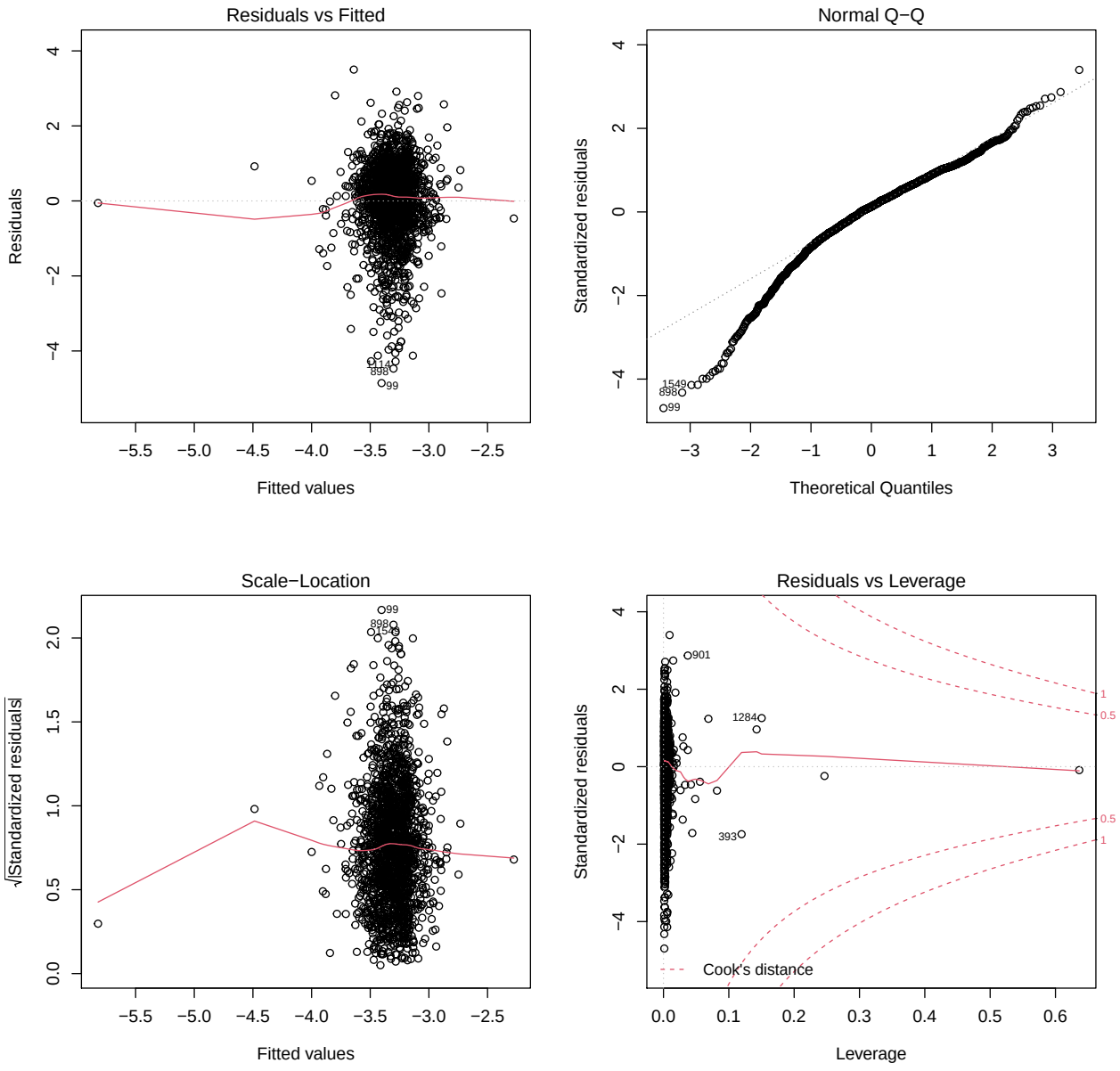

Figure 1: Plot for Model Checking

6.3 Constant Variance

The scale location plot suggests that the majority of residuals are randomly scattered within an upper and lower band around the value 0, which suggests a constant variance. However, since there is one data point near the bottom left that has a smaller residual, the red line bulges up and down in the middle of the plot, causing a slight increase and decrease in variability. However, the assumption of constant variance is still considered fairly satisfied.

6.4 Normality

The Q-Q plot overall appears adequate, as most of the residual points lay on a straight line. However, the normality assumption can only be considered as fairly satisfied, since there are points on the left tail that are below the line (such as observations 99, 898 & 1549), suggesting some potential outliers.

6.5 Outliers

As we discussed in the Q-Q plot above, there are some points off the straight on the left tail of the curve that appear to be outliers.

In the leverage and Standard Residual plot, there is a point on the right end of the plot and a few other points, including data point 393 and 1284, that are far away from the majority of observations (observations that are greater than the leverage 0.1). As such, those points are considered to have a high leverage, whereas they are all within the Cook's Distance. Hence, these points are not considered to be influential.

7 Results

We found that Elderly rate, number of Specialists, number of International migrations, and Region are four variables that have major impacts on the log of COVID-19 mortality rate given the dataset. This result matches our initial speculation. Based on the above discussion, the final model is selected to be Mortality ~ ElderlyRate + Specialist + Migration + factor(Region) as it has high R^2_{adj} compared to other models and is overall statistically significant. Besides, the hypothesis tests for individual variables indicate statistical significance. Furthermore, the best linear model fitting for these variables is $log(MortalityRate) = -3.7 + (2.699)x_{ElderlyRate} + (3.641 \times 10^{-5})x_{Specialist} - (2.297 \times 10^{-5})x_{Migration} - (8.959 \times 10^{-2})x_{Northeast} - (0.1422)x_{South} - (0.2641x_{West}$. However, there can be several limitations of this study and will be discussed below.

8 Limitation and Conclusion

Even though we have considered model 1: Mortality ~ ElderlyRate + Specialist + Migration + factor(Region) as our best model, there is still a great degree of limitation from using this model to analyze the mortality rate of COVID-19. First, model 1 has a low explanatory power. Specifically, $R^2_{adj} = 0.0218$, which means model 1 could only explain 2.18% of variability of the COVID-19 mortality rate. This indicates that there is a large room for model improvement. Thus, further investigation on additional explanatory variables is desired.

Secondly, multicollinearity issue is present, which not only undermines the statistical significance of an independent variable, but also seems to violate our model independence assumption, even though the diagnosis plots does not present an obvious trend. For instance, the correlation between Specialist & Migration is as high as 0.855, highlighting a strong positive linear association in this pair variable. In addition, Specialist has a very high VIF value (50.404208), suggesting that the association will affect the standard errors. Since the multicollinearity issue is considered very severe in this case, we would need to do further

investigation and study on the cause of this issue. The solution for this problem varies by situation. If it is structural multicollinearity, centering the variables could be helpful. (Frost (2017)). Another potential solution could be linearly combining the independent variables together into higher power variables. Any of these solutions discussed above would require further study since they can be difficult and time-consuming to implement.

In conclusion, the selected model consists of 4 explanatory variables: ElderlyRate, Specialist, Migration, Region, all of which have a statistically significant effect on Morality. This model does the best job of representing the U.S. mortality rate of COVID-19 at this stage of the study.

Appendix

Below lists the 9 single-model summaries for the selected explanatory variables.

```
##
## Call:
## lm(formula = Mortality ~ ElderlyRate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8993 -0.4917  0.1336  0.6605  3.6109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.74966    0.09149 -40.982  < 2e-16 ***
## ElderlyRate  2.40160    0.48247   4.978 7.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 1737 degrees of freedom
## Multiple R-squared:  0.01406,    Adjusted R-squared:  0.0135
## F-statistic: 24.78 on 1 and 1737 DF,  p-value: 7.072e-07
```

```
##
## Call:
## lm(formula = Mortality ~ Region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9109 -0.4914  0.1316  0.6785  3.0620
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.20073    0.04848 -66.017  < 2e-16 ***
## RegionNortheast -0.07649    0.09046  -0.846  0.39790
## RegionSouth     -0.14826    0.05972  -2.482  0.01314 *
## RegionWest      -0.24041    0.09116  -2.637  0.00843 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 1727 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.005412,   Adjusted R-squared:  0.003684
## F-statistic: 3.132 on 3 and 1727 DF,  p-value: 0.02471
```

```
##
## Call:
## lm(formula = Mortality ~ Specialist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9596 -0.4855  0.1351  0.6562  3.2403
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
```

```
## (Intercept) -3.299e+00  2.565e-02 -128.642   <2e-16 ***
## Specialist  -1.728e-05  7.828e-06   -2.207   0.0274 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.045 on 1737 degrees of freedom
## Multiple R-squared:  0.002796,   Adjusted R-squared:  0.002222
## F-statistic: 4.871 on 1 and 1737 DF,  p-value: 0.02745

##
## Call:
## lm(formula = Mortality ~ population)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9877 -0.4918  0.1333  0.6649  3.2686
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.40734    0.04648  -73.31   <2e-16 ***
## populationSmall  0.13515    0.05518    2.45   0.0144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 1737 degrees of freedom
## Multiple R-squared:  0.003442,   Adjusted R-squared:  0.002869
## F-statistic:     6 on 1 and 1737 DF,  p-value: 0.0144

##
## Call:
## lm(formula = Mortality ~ Physician_rate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8957 -0.4926  0.1193  0.6658  3.1942
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.5697674  0.1318121 -27.082   <2e-16 ***
## Physician_rate  0.0009903  0.0004961   1.996   0.0461 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.045 on 1737 degrees of freedom
## Multiple R-squared:  0.002289,   Adjusted R-squared:  0.001715
## F-statistic: 3.985 on 1 and 1737 DF,  p-value: 0.04606

##
## Call:
## lm(formula = Mortality ~ Unemployment_rate_2018)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9849 -0.4643  0.1295  0.6592  3.2247
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -3.46670    0.08251 -42.016   <2e-16 ***
## Unemployment_rate_2018  0.03688    0.01867   1.975   0.0484 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.045 on 1737 degrees of freedom
## Multiple R-squared:  0.002241,   Adjusted R-squared:  0.001666
## F-statistic: 3.901 on 1 and 1737 DF,  p-value: 0.04841

##
## Call:
## lm(formula = Mortality ~ Poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9570 -0.4868  0.1355  0.6550  3.2323
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -3.300e+00  2.555e-02 -129.168   <2e-16 ***
## Poverty     -3.448e-07  1.537e-07   -2.243    0.025 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.045 on 1737 degrees of freedom
## Multiple R-squared:  0.002888,   Adjusted R-squared:  0.002314
## F-statistic: 5.032 on 1 and 1737 DF,  p-value: 0.02501

##
## Call:
## lm(formula = Mortality ~ Migration)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9585 -0.4795  0.1358  0.6560  3.2221
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -3.301e+00  2.526e-02 -130.691  < 2e-16 ***
## Migration   -1.269e-05  4.232e-06   -2.999  0.00275 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 1737 degrees of freedom
## Multiple R-squared:  0.00515,    Adjusted R-squared:  0.004577
## F-statistic: 8.992 on 1 and 1737 DF,  p-value: 0.00275

##
## Call:
## lm(formula = Mortality ~ Nurse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.9610 -0.4833  0.1338  0.6569  3.2986
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -3.298e+00  2.566e-02 -128.519   <2e-16 ***
## Nurse       -1.063e-04  4.324e-05   -2.459    0.014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 1737 degrees of freedom
## Multiple R-squared:  0.003468,   Adjusted R-squared:  0.002895
## F-statistic: 6.046 on 1 and 1737 DF,  p-value: 0.01404
```

R Code

```
knitr::opts_chunk$set(echo = FALSE, tidy = TRUE,tidy.opts=list(width.cutoff= 70),
                      fig.pos = "!h")

# Input Data and library
library(tidyverse)
library("here")
library(formattable)
library(leaps)
library(car)
library(MPV)
library(leaps)
library(MASS)
library(knitr)
library(kableExtra)
alldata <- read.csv(here::here("data","counties.csv"), header=T)
attach(alldata)
# Rename the variable
Physician_rate <- Active.Physicians.per.100000.Population.2018..AAMC.
Specialist <- All.Specialties..AAMC.
ElderlyRate <- Pop_Above_65
Poverty <- POVALL_2018
Migration <- INTERNATIONAL_MIG_2018
Nurse <- Total.nurse.practitioners..2019.
Covid19 <- data.frame(Mortality, ElderlyRate, Region, Specialist, population,
                      Physician_rate, Unemployment_rate_2018, Poverty, Migration, Nurse)

# Summary of fitting individual model with its corresponding p-value
d1=data.frame(Variable= c("ElderlyRate", "Region", "Specialist", "population",
                          "Physician_rate", "Unemployment_rate_2018","Poverty",
                          "Migration","Nurse"),
              p_value = c(7.072*10^(-7), 0.02471,0.02745, 0.01440, 0.04606, 0.04841,
                          0.02501, 0.00275, 0.01404))
kable(d1,align = "l",booktabs = T, caption = "P-values for Selected Explanatory Variables",
      linesep = "")%>%
  kable_styling(position = "center", full_width = FALSE,latex_options =
                  "HOLD_position")%>%
  column_spec(1, width = "20em")

# Find the correlation matrix and put it in the table
x<-data.frame(ElderlyRate, Specialist, Physician_rate, Unemployment_rate_2018, Poverty,
              Migration, Nurse)
cor_matrix <- cor(x)
output_matrix <- round(cor_matrix,3)
kable(output_matrix,align = "l",booktabs = T, caption = "Correlation Matrix",
      col.names = c("Elderly","Specilist","Physician","Unemployment", "Poverty","Migration",
                    "Nurse"))%>%
  kable_styling(position = "center", full_width = TRUE, latex_options = "HOLD_position",
                font_size = 10)%>%
  column_spec(1, width = "12em")%>%
  column_spec(5, width = "6em")%>%
  column_spec(7, width = "5em")%>%
```

```
  column_spec(4, width = "4em")
# Fit the model with all the variable we found its statistically significant
fullmodel<-lm(Mortality ~ ElderlyRate + factor(Region) + Specialist + factor(population)
              + Physician_rate + Unemployment_rate_2018 + Poverty + Migration + Nurse)
# Compute full model's corresponding VIF value and put it in the table
vif_value <- vif(fullmodel)
kable(vif_value,align = "l",booktabs = T, caption = "VIF Values")%>%
  kable_styling(position = "center", full_width = FALSE, latex_options = "HOLD_position")%>%
  column_spec(1, width = "15em")

# Stepwise Automatic Model Selection Method
step.model <- stepAIC(fullmodel, direction = "both",  trace = FALSE)
summary(step.model)
step.model$anova

# Form a Table of Details about Models Used
f <- regsubsets(Mortality~., data=Covid19,nbest=2, nvmax=11)
e <- summary(f)
attach(e)
explain1 <- data.frame(
  Num = c("0","1","2","3"),
  Variable = c("Intercept","ElderlyRate", "RegionNorthEast","RegionSouth"),
  Num = c("4","5", "6","7"),
  Variable = c("RegionWest","Specialist", "populationSmall","PhysicianRate"),
  Num = c("8","9","10","11"),
  Variable = c("Unemployment_Rate_2018","Poverty","Migration","Nurse"))
kable(explain1,align = "l",booktabs = T, caption = "Variable Reference")%>%
  kable_styling(position = "center", full_width = FALSE,latex_options =
                  "HOLD_position")%>%
  column_spec(1, width = "5em")

# All Possible Combinations by Automatic Model Selection Process
combination <- cbind(which,cp,adjr2)
kable(combination,align = "l",booktabs = T, caption = "Model Selection with CP values",
      col.names = c("0", "1","2","3","4","5","6","7","8","9","10","11", "cp",
                    "adjr2"))%>%
  kable_styling(position = "center", full_width = FALSE, latex_options =
                  "HOLD_position")%>%
  column_spec(1, width = "5em")

# Suitable Models from Above Model Selection Process
model1 <- lm(Mortality~ElderlyRate + Specialist + Migration + factor(Region))
model2 <- lm(Mortality~ElderlyRate + Nurse + Migration + factor(Region))
model3 <- lm(Mortality~ElderlyRate + Specialist + Migration + Unemployment_rate_2018 +
               factor(Region))
model4 <- lm(Mortality~ElderlyRate + Nurse + Migration + Unemployment_rate_2018 +
               factor(Region))
model5 <- lm(Mortality~ElderlyRate + Specialist + Migration + Unemployment_rate_2018 +
               factor(Region) + factor(population))
model6 <- lm(Mortality~ElderlyRate + Specialist + Migration + Unemployment_rate_2018 +
               Physician_rate +  factor(Region))
model7 <- lm(Mortality~ElderlyRate + Nurse + Migration + Unemployment_rate_2018 +
               Physician_rate + factor(Region))
```

```
model8 <- lm(Mortality~ElderlyRate + Specialist + Migration + Unemployment_rate_2018 +
               Physician_rate + factor(Region) + factor(population))
model9 <- lm(Mortality~ElderlyRate + Nurse + Migration + Unemployment_rate_2018 +
               Physician_rate + factor(Region) + factor(population))
model10 <- lm(Mortality~ElderlyRate + Specialist + Nurse + Migration +
                Unemployment_rate_2018 + Physician_rate + factor(Region)
              + factor(population))
model11 <- lm(Mortality~ElderlyRate + Specialist + Migration + Unemployment_rate_2018
              + Physician_rate + Poverty + factor(Region) + factor(population))
model12 <- lm(Mortality~ElderlyRate + Specialist + Nurse + Migration +
                Unemployment_rate_2018 + Physician_rate + Poverty + factor(Region)
              + factor(population))

# AIC Values for Above Models
A <- AIC(model1,model2,model3,model4,model5,model6,model7,model8,model9,
         model10,model11,model12, k=2)

# BIC Values for Above Models
B <- BIC(model1,model2,model3,model4,model5,model6,model7,model8,model9,
         model10,model11,model12)

# Prepare a Table for Available Models
d2=data.frame(Model= c(1, 2, 3, 4, 5, 6,7, 8,9,10,11,12), Regression_Equation =
                c("Log of Mortality Rate vs. ElderlyRate + Specialist + Migration +
                  Region","Log of Mortality Rate vs. ElderlyRate + Nurse + Migration
                  + Region", "Log of Mortality Rate vs. ElderlyRate + Specialist +
                  Migration + Unemployment_rate_2018 + Region","Log of Mortality Rate
                  vs. ElderlyRate + Nurse + Migration + Unemployment_rate_2018 +
                  Region","Log of Mortality Rate vs. ElderlyRate + Specialist +
                  Migration + Unemployment_rate_2018 + Region + population","Log of
                  Mortality Rate vs. ElderlyRate + Specialist + Migration +
                  Unemployment_rate_2018 + Physician_rate +  Region",
                  "Log of Mortality Rate vs. ElderlyRate + Nurse + Migration +
                  Unemployment_rate_2018 + Physician_rate + Region","Log of
                  Mortality Rate vs. ElderlyRate + Specialist + Migration +
                  Unemployment_rate_2018 + Physician_rate + Region + population",
                  "Log of Mortality Rate vs. ElderlyRate + Nurse + Migration +
                  Unemployment_rate_2018 + Physician_rate + Region + population",
                  "Log of Mortality Rate vs. ElderlyRate + Specialist + Nurse +
                  Migration + Unemployment_rate_2018 + Physician_rate + Region +
                  population","Log of Mortality Rate vs. ElderlyRate + Specialist +
                  Migration + Unemployment_rate_2018 + Physician_rate + Poverty +
                  Region + population","Log of Mortality Rate vs. ElderlyRate +
                  Specialist + Nurse + Migration + Unemployment_rate_2018 +
                  Physician_rate + Poverty + Region + population"))

# Generate Table including Available Models
kable(d2,align = "cl",booktabs = T, caption = "Automatic Selection Model Summary")%>%
  kable_styling(position = "center", full_width = TRUE,latex_options =
                  "HOLD_position")%>%
  column_spec(1, width = "5em")%>%
  row_spec(1:12, hline_after = TRUE)
```

```r
# Generate Table including AIC & BIC values
output_table <- data.frame(A, B[2])
kable(output_table,align = "l",booktabs = T, caption = "AIC and BIC values")%>%
  kable_styling(position = "center", full_width = FALSE,latex_options =
                  "HOLD_position")%>%
  column_spec(1, width = "5em")

# Model 3: Log of Mortality Rate vs. ElderlyRate + Specialist + Migration +
#           Unemployment_rate_2018 + Region
summary(model3)

# Model 1: Log of Mortality Rate vs. ElderlyRate + Specialist + Migration + Region
summary(model1)

# Model 2: Log of Mortality Rate vs. ElderlyRate + Nurse + Migration + Region
summary(model2)

# Model 4: Log of Mortality Rate vs. ElderlyRate + Nurse + Migration +
#           Unemployment_rate_2018 + Region
summary(model4)

# Calculating p-value for Unemployment_rate_2018
(pval4 <- pf(1.66634,1, 1723, lower.tail = FALSE))

# Model with Interaction term
model_int <- lm(Mortality~ElderlyRate + Specialist + Migration + factor(Region) +
                  Specialist * Migration)
summary(model_int)
anova(model_int)

# Plot the 2x2 diagnosis plot for modeling checking
par(mfrow = c(2, 2))
plot(model1,cex.lab=1.1, cex.axis=1.1, cex.main=1, cex.sub=0.8)

# Single Model Fitting
modela <- lm(Mortality~ ElderlyRate)
modelb <- lm(Mortality~ Region)
modelc <- lm(Mortality~ Specialist)
modeld <- lm(Mortality~ population)
modele <- lm(Mortality~ Physician_rate)
modelf <- lm(Mortality~ Unemployment_rate_2018)
modelg <- lm(Mortality~ Poverty)
modelh <- lm(Mortality~ Migration)
modeli <- lm(Mortality~ Nurse)

summary(modela)
summary(modelb)
summary(modelc)
summary(modeld)
summary(modele)
summary(modelf)
summary(modelg)
summary(modelh)
```

```r
summary(modeli)
```

Reference

Frost, Jim. 2017. "Multicollinearity in Regression Analysis: Problems, Detection, and Solutions," September. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/.

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.

Ye, Juezhao, Pei Wang, and Liu. 2020. "Investigation on Covid-19 Mortality Rate." https://app.crowdmark.com/student/assessments/project-report-submission-1.