Florida International University

Predicting COVID-19 Death Rate per Million in the United States: A Linear Regression Approach

Carlos Figueredo Simoza

PID: 6162319

Oral Presentations in Statistics

Dr. BM Golam Kibria

June 2022

TABLE OF CONTENTS

ABSTRACT 3

INTODUCTION 3

DATA ANALYSIS 5

Preliminary Assessment of the Data 5

Initial Regression Model Output and Interpretation 6

Logarithmic Transformation: Second Model Output and Interpretation 8

Inverse Transformation: Third Model Output and Interpretation 10

Square Root Transformation: Fourth Model Output and Interpretation 12

Further Examination of the Original (First) Model 14

Reduced Original Model by Backward Elimination: Fifth Model Output and Interpretation 15

Reduced Original Model by Best Subset Regression: Sixth Model Output and Interpretation 17

Evaluation of the Models 19

Cross Validation of the Model 20

DISCUSSION AND CONCLUDING REMARKS 22

REFERENCES 23

DATA TABLE 24

ABSTRACT

COVID-19 has found its way into every aspect of the lives of individuals in the world. There is a particular assumption that the disease has hit the hardest in the USA. This project explores the factors that affect the COVID-19 number of deaths per 1 million in 51 states of America. A total of six models were developed in order to determine which factors influence the death rate. Preliminary findings show that vaccinations, hospitalization rate per 1 million, percentage of smokers, AQI, and prevalence of diabetics are the factors that affect the most. The R-squared value for the model selected as the most precise is 0.5899. There needs to be more research conducted to develop more accurate models.

INTODUCTION

Background Information

The novel coronavirus disease, also known as COVID-19, has notoriously gained its popularity among the global arena since the end of 2019. This disease has caused major disruptions to life in every single country in the world; from mass lockdowns in cities to reaching maximum capacity at hospitals, there is no denying that the coronavirus disease has had a significant potential to wreak havoc among populations of positively all countries. Most concerning of all is its capability to increase the excess deaths per year at any given country. It is for this reason, that this project is centered on predicting and analyzing data pertained to the total number of deaths caused by the coronavirus disease in the United States of America. In this sense, more knowledge, and a brighter insight into the disease's ability to take inhabitants' lives can help individuals across the world to better understand the nature of COVID-19.

Sample

A list containing a total of 51 territorial states of the US was selected for this project. The data obtained from public sources goes in line with the period marked by the beginning of the pandemic (August 2020) to June 2022 (or closest available date). The states in question are: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, District Of Columbia, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, South

Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, and Wyoming.

The ultimate goals of this project are to efficiently analyze the data to determine which of the regressors considered initially are significant to the number of deaths; as well as to attain an efficient model that would be accurate at the time of predicting future number of deaths in the different states of USA.

Dependent Variable

- Death rate per one million (y): The coronavirus death rate per 1 million people.

Regressors

- Confirmed Cases per one million (x_1): The amount of COVID positive people reported by a country since 2020 until April 2022
- Vaccination Rate (x_2): Percentage of the total population of a state that have received a full protocol of the vaccine. (Ex: one dose of the Johnson and Johnson vaccine, or two doses of Pfizer, Moderna, Coronavac, etc.)
- Hospitalizations per one million(x_3): The number of covid-positive adults that were admitted to the hospital divided by the population and multiplied by a million.
- Population density (x_4): The number of residents for the state per square mile.
- Median Age (x_5): the combined median age for both men and women.
- GDP per Capita (x_6): Measures the overall economic output (and wellbeing) of a state per person inside its population.
- Percentage of Smokers (x_7): The fraction of the population of a state that report smoke tobacco-products.
- Air Quality Index (x_8): Environmental index related to the quality of breathable air in the state.
- Diabetes Prevalence (x_{19}): The fraction of people from a state that are reportedly diabetic.
- Obesity Prevalence (x_{10}): The fraction of residents from a state that are reportedly obese.
- Average Temperature (x_{11}): The combined 24 month (2020-2022) mean temperature of a state.
- Hypertension Rate (x_{12}): The fraction of adult population who are reportedly suffering from high blood pressure.
- Education Quality ($_{X13}$): Metric designed to qualify the overall student success and school quality factors from the states.

The selection of the variables was based upon a logical process regarding the possible factors that are often associated with the mobility of the coronavirus diseases within the country.

DATA ANALYSIS

Preliminary Assessment of the Data

Descriptive Statistics			
	Mean	Std. Deviation	N
Y	2965.6274510	818.50013954	51
X1	263023.5532098	37679.58327363	51
X2	.6664157	.10074825	51
X3	13397.0374944	5341.70864888	51
X4	15819.7058824	64683.25064723	51
X5	38.9098039	2.33394558	51
X6	67951.2745098	25534.10435561	51
X7	16.5921569	3.30665046	51
X8	42.3039216	5.24823631	51
X9	.1088039	.02089882	51
X10	.3205882	.04099911	51
X11	52.0882353	8.68040660	51
X12	.3318039	.04329481	51
X13	66.7641667	7.87897388	51

Table 1.1: Summary of the descriptive statistics for the initial variables of the model.

We will first fit the initial model with all the variables discussed. Once the model is fitted, we will check for the regular assumptions, normality, and constant variance for the residuals of the model. If needed, the model will be transformed and through methods of subset testing and stepwise forward elimination we will select the best possible variables. We will also compare the latest models through metrics for model evaluation; namely, mean absolute error (MAE), root mean squared error (RMSE), range normalized RMSE (NRMSE), and mean absolute percentage error (MAPE). After the initial assessment of the data is performed, a cross validation process will take place in order to compare the best models.

Initial Regression Model Output and Interpretation

```
Residuals:
     Min       1Q   Median       3Q      Max
-1241.79  -280.25    -1.47   274.33  1229.82

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.211e+02  2.735e+03   0.081  0.93598
x1           8.374e-04  2.548e-03   0.329  0.74427
x2          -1.955e+03  2.287e+03  -0.855  0.39820
x3           1.433e-02  1.874e-02   0.764  0.44944
x4           1.950e-03  1.553e-03   1.255  0.21722
x5          -1.054e+01  6.477e+01  -0.163  0.87163
x6          -5.834e-04  6.134e-03  -0.095  0.92474
x7           6.961e+01  6.856e+01   1.015  0.31654
x8           5.380e+01  1.853e+01   2.904  0.00618 **
x9           2.001e+04  1.183e+04   1.692  0.09897 .
x10         -4.288e+03  4.614e+03  -0.929  0.35881
x11         -1.011e-01  1.961e+01  -0.005  0.99591
x12         -2.132e+03  4.055e+03  -0.526  0.60219
x13          7.980e+00  1.593e+01   0.501  0.61941
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

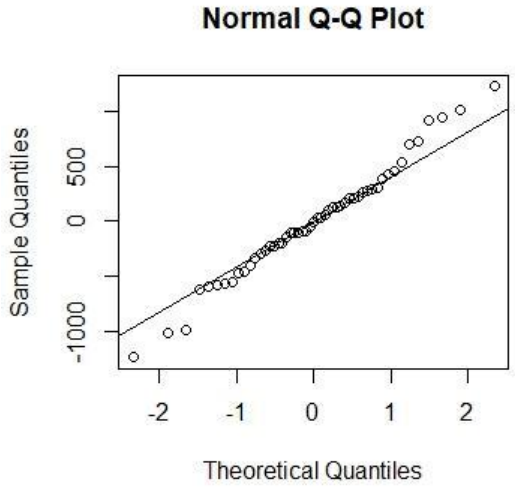
Residual standard error: 595.5 on 37 degrees of freedom
Multiple R-squared:  0.6084,    Adjusted R-squared:  0.4708
F-statistic: 4.421 on 13 and 37 DF,  p-value: 0.0001798
```

Obtaining our first model:

Y = 2.211e+02 + 8.374e-04X1 -1.955e+03X2 + 1.433e-02X3 + 1.950e-03X4 - 1.054e+01X5 -5.834e-04 X6 + 6.961e+01X7 5.380e+01X8 2.001e+04X9 -4.288e+03X10 -1.011e-01X11 -2.132e+03X12 + 7.980e+00X13

Testing for significance of the regression we can conclude that the regression is significant with an F statistic equaling 4.421 and a p-value of 0.0001798, considerably less than 0.05. Checking for significance of individual coefficients (β), we notice that only X9, and X10 seem to be significant with p-values that are less than 0.10.

Furthermore, the coefficient of determination R-squared is equal to 0.6084; this means that around 61% of the total variability in y can be accounted for by the thirteen regressors. However, when comparing R-squared to the adjusted R-squared (0.4708), we notice that their difference cannot be ignored. Since adjusted R-squared accounts for the variables that have a real effect on the regression, we have to explore ways to improve ideally both values, or adjusted R-squared.



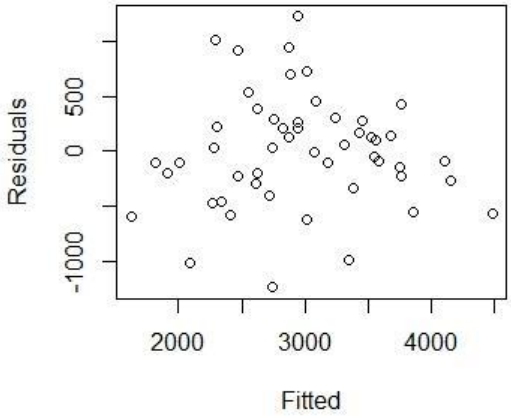


Figure 1.1: Normal Probability Plot.

Figure 1.2: Residuals vs Fitted Plot

Both figures above help us determine whether the residuals meet the assumptions in order to consider the model for future examination. **Figure 1.1** has some residual values deviating from the straight line, yet the vast majority are tightly close and strongly following the line, we can conclude that the residuals follow a distribution that is close to normal. **Figure 1.2** shows how the variance of the error terms, var(e), is approximately a constant. Though it might not have the perfectly looking constant shape, it does not resemble any pattern-like shape. Hence, the variance assumption is met. Altogether, this model meets both assumptions regarding the error terms. This model can be selected for future examination.

Logarithmic Transformation: Second Model Output and Interpretation

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.260880 -0.051104  0.003166  0.049633  0.199721

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.968e+00  4.958e-01   5.987 6.53e-07 ***
x1           3.329e-08  4.620e-07   0.072  0.94295
x2          -3.804e-01  4.147e-01  -0.917  0.36493
x3           1.509e-06  3.398e-06   0.444  0.65944
x4           3.928e-07  2.817e-07   1.394  0.17150
x5          -1.645e-03  1.174e-02  -0.140  0.88937
x6           3.345e-07  1.112e-06   0.301  0.76528
x7           1.001e-02  1.243e-02   0.805  0.42602
x8           1.052e-02  3.359e-03   3.132  0.00339 **
x9           3.608e+00  2.144e+00   1.683  0.10079
x10         -3.373e-01  8.366e-01  -0.403  0.68913
x11         -1.404e-03  3.555e-03  -0.395  0.69508
x12         -6.197e-01  7.353e-01  -0.843  0.40473
x13          1.886e-03  2.889e-03   0.653  0.51778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

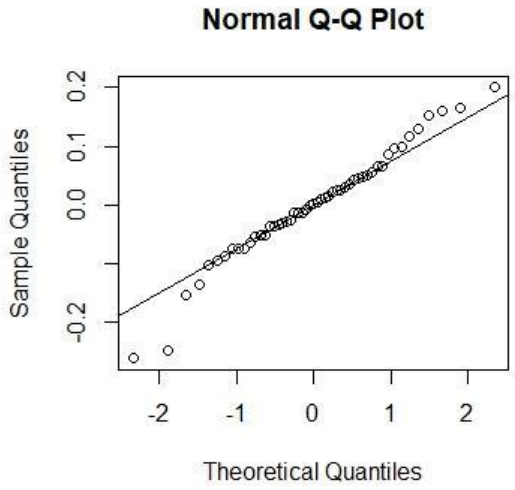
Residual standard error: 0.108 on 37 degrees of freedom
Multiple R-squared:  0.5882,    Adjusted R-squared:  0.4436
F-statistic: 4.066 on 13 and 37 DF,  p-value: 0.0003841
```

Obtaining the second model:

Y= 2.968e+00 + 3.329e-08X1 - 3.804e-01X2 + 1.509e-06X3 + 3.928e-07X4 -1.645e-03X5 + 3.345e-07X6 + 1.001e-02X7 + 1.052e-02X8 + 3.608e+00X9 - 3.373e-01X10 -1.404e-03X11 -6.197e-01X12 + 1.886e-03X13

Where Y = Log10(Y)

In this model it is determined that the R-squared value is minimally lower than the first model, 0.5882. Thus, 59% of the total variability in y can be accounted for by the regression. The adjusted R-squared value is also minimally lower than the original model, hinting in initial assessment of this model that the transformation did not significantly improve the model. Checking for significance in regression for the β’s, we notice that only x8 carry real significance. This latter fact makes the model more challenging to work with, since it could be counterproductive trying to attain an accurate number of deaths per million while only having AQI as the most significant variable.



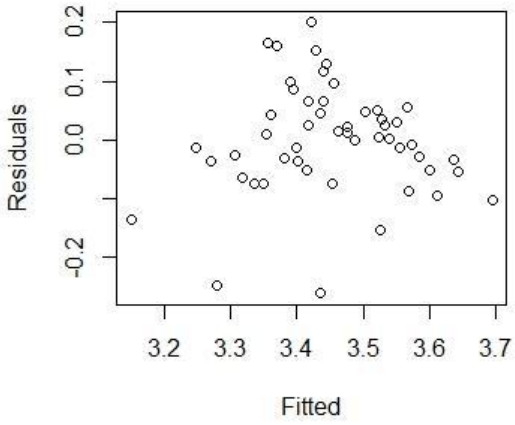


Figure 2.1: Normal Probability Plot

Figure 2.2: Residuals vs Fitted Plot

Figure 2.2 does not look like an improved version of the residuals plotted against the fitted values of the model, rather than looking like a constant, it resembles a shape that is indicative of y being a proportion between 0 and 1. **Figure 2.1** for this model clearly shows that still most of the residuals are following the straight line, hinting that the residuals follow a normal distribution. We can conclude that the normality assumption can be accepted, it is not as clear the case for the residual vs fitted plot. This failure to meet the constant variance assumption, alongside he fact that both R-squared and adjusted R-squared decreased after applying the log transformation allows us to discard the model from future consideration.

Inverse Transformation: Third Model Output and Interpretation

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.958e-04 -5.251e-05 -3.310e-06  3.730e-05  3.553e-04

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.600e-04  5.547e-04   1.550  0.12955
x1           1.119e-10  5.169e-10   0.216  0.82983
x2           4.250e-04  4.640e-04   0.916  0.36563
x3          -4.448e-10  3.801e-09  -0.117  0.90748
x4          -4.714e-10  3.151e-10  -1.496  0.14317
x5           1.873e-06  1.314e-05   0.143  0.88744
x6          -9.017e-10  1.244e-09  -0.725  0.47318
x7          -6.242e-06  1.391e-05  -0.449  0.65617
x8          -1.240e-05  3.758e-06  -3.298  0.00216 **
x9          -3.836e-03  2.399e-03  -1.599  0.11828
x10         -1.675e-04  9.360e-04  -0.179  0.85896
x11          3.643e-06  3.978e-06   0.916  0.36570
x12          7.830e-04  8.226e-04   0.952  0.34740
x13         -2.171e-06  3.232e-06  -0.672  0.50582
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0001208 on 37 degrees of freedom
Multiple R-squared:  0.5615,    Adjusted R-squared:  0.4074
F-statistic: 3.644 on 13 and 37 DF,  p-value: 0.0009778
```

Thus, we have our third model:

Y= 8.600e-04 + 1.119e-10X1 + 4.250e-04X2 -4.448e-10X3 -4.714e-10X4 + 1.873e-06X5 -9.017e-10X6 -6.242e-06X7 -1.240e-05X8 -3.836e-03X9 -1.675e-04X10 + 3.643e-06X11 + 7.830e-04X12 -2.171e-06X13

Where Y = 1/Y

The model is significant with a p-value that is less than 0.05. The coefficient of determination is even lower that the second model and the adjusted R-squared is also lower, at 0.5615 and 0.4074, respectively. Once again, only β8 is significant to the regression.

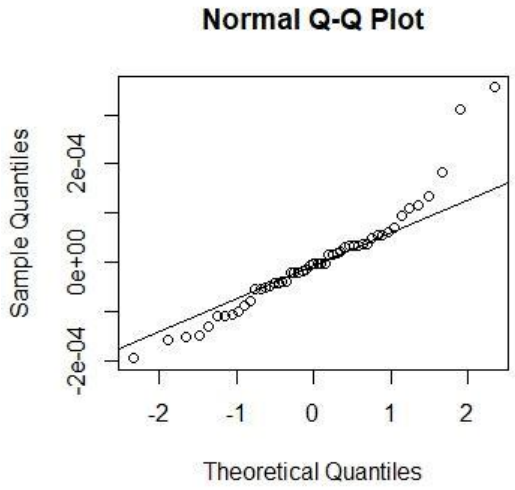


Figure 3.1: Normal Probability Plot

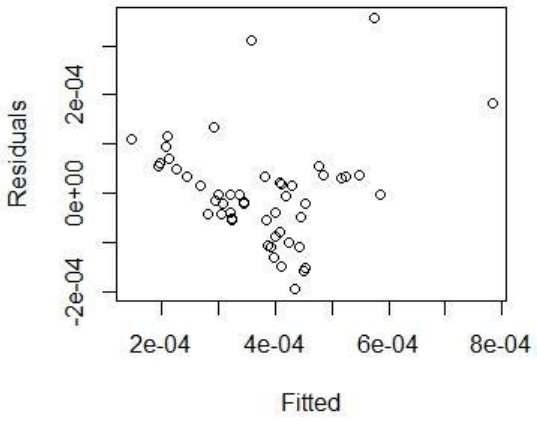


Figure 3.2: Residuals vs Fitted Plot

From the plots above, we are inclined to conclude that none of the assumptions are met. **Figure 3.2** shows that the constant variance assumption cannot be accepted since the residuals vs fitted plot can arguably be seen taking the form of a fennel-like pattern, possibly implying that the variance of the error terms is an increasing function of y. Whereas **Figure 2.1** presents many more deviated values than the previous models. In order to determine whether or not the error terms follow a normal distribution we shall conduct a test for normality; namely, the Shapiro-Wilk test for normality.

```
        Shapiro-Wilk normality test

data:  residuals(t2model)
W = 0.9221, p-value = 0.002511
```

Since p-value is less than0.05 we reject the null hypothesis and conclude that the residual terms for this model do not follow a normal distribution. This model is then discarded from future consideration.

Square Root Transformation: Fourth Model Output and Interpretation

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.5380  -2.7404   0.3198   2.8860  11.7187

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.649e+01  2.747e+01   0.964  0.34106
x1           5.310e-06  2.559e-05   0.207  0.83678
x2          -2.050e+01  2.297e+01  -0.892  0.37803
x3           1.147e-04  1.882e-04   0.609  0.54613
x4           2.073e-05  1.560e-05   1.328  0.19216
x5          -9.747e-02  6.506e-01  -0.150  0.88172
x6           5.760e-06  6.161e-05   0.093  0.92602
x7           6.410e-01  6.887e-01   0.931  0.35797
x8           5.622e-01  1.861e-01   3.021  0.00455 **
x9           2.015e+02  1.188e+02   1.697  0.09817 .
x10         -3.153e+01  4.635e+01  -0.680  0.50053
x11         -3.544e-02  1.970e-01  -0.180  0.85819
x12         -2.875e+01  4.073e+01  -0.706  0.48475
x13          9.467e-02  1.600e-01   0.592  0.55770
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.981 on 37 degrees of freedom
Multiple R-squared:  0.5993,    Adjusted R-squared:  0.4585
F-statistic: 4.256 on 13 and 37 DF,  p-value: 0.0002551
```

Obtaining the following model:

oefficients:

Y= 2.649e+01 + 5.310e-06X1 -2.050e+01X2 + 1.147e-04X3 + 2.073e-05X4 -9.747e-02X5 + 5.760e-06X6 + 6.410e-01X7 + 5.622e-01X8 + 2.015e+02X9 -3.153e+01X10 -3.544e-02X11 -2.875e+01X12 + 9.467e-02X13.

Where Y = $\sqrt{y}$

Checking for significance of regression we conclude that the regression is significant with a p-value that is less than 0.05. The coefficient of determination R-squared equals to 0.5993, letting us know that around 60% of the total variability in y is owed to the 13 regressors. Adjusted R-squared equals to 0.4684. Altogether this model still fails to show initial improvement from the original model, in which both values of R-squared and adjusted R-squared were larger. Lastly, only β8, and β9 are shown to be significant to the regression.

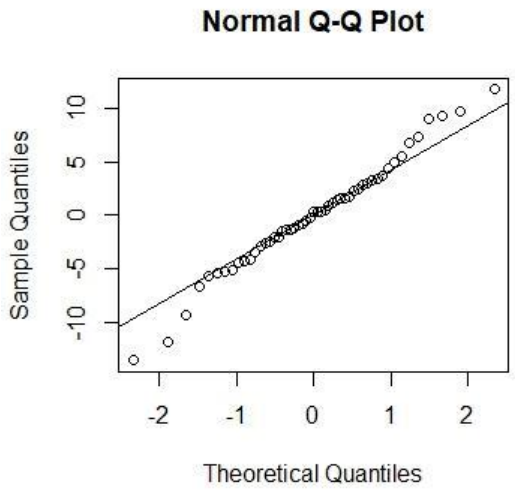


Figure 4.1: Normal Probability Plot

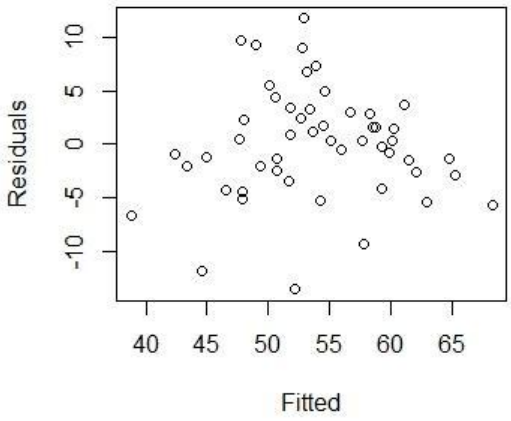


Figure 4.2: Residuals vs. Predicted Plot

From both graphs above, it is not challenging to conclude that the variance and normality assumptions for this model are met. **Figure 4.1** shows a steady course of the points following along the line, some of them are deviating, but overall, it does look like the residuals are following a normal distribution. **Figure 4.2** also presents a clear (but not perfect) case in which we can conclude that the variance of the errors is a constant. Thus, the two assumptions are met, and this model can be considered for future examination.

Further Examination of the Original (First) Model

Previous assessment of the data showed that the models with the original regression, log transformation, and square root transformation were fit for future examination. However, out of those 3 models, the original had the biggest value for R-squared and adjusted R-squared. Hence, selecting the original model for further examination/analysis seems like the appropriate course of action.

Multicollinearity

To determine whether the variables are closely related to each other, we should check for their variance inflation factor. In this case, if any of the 13 variables used in the model has a VIF>10, we must conclude that we are in presence of severe multicollinearity. Having severe collinearity could detrimentally impact the estimates of the coefficient of the model. Using R, we determined that the VIF for the variables were:

```
> vif(model)
      x1       x2       x3       x4       x5       x6       x7       x8       x9      x10      x11
1.299866 7.488877 1.413009 1.423951 3.222915 3.459275 7.248220 1.333242 8.612416 5.047372 4.085481
     x12      x13
4.347424 2.221997
```

Since none of the thirteen variable exhibits a VIF larger than 10, we can conclude that multicollinearity will not be a problem.

Error Terms Normality Check

Since we did not initially test whether the original model's error terms were following a normal distribution, it now becomes imperative that we do so. To determine whether or not the terms are normally distributed, we will conduct a Shapiro-Wilk test for normality. The results are given by:

```
        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.98619, p-value = 0.8127
```

Since we obtained a p-value of 0.8127 for the test, we failed to reject the null hypothesis that the residuals are normally distributed and concluded that the terms are normally distributed. This

latest finding together with **Figure 1.2** concisely and definitively confirm that the assumptions needed to proceed with the model are met.

Reduced Original Model by Backward Elimination: Fifth Model Output and Interpretation

There are several procedures we can perform in order to obtain a reduced form of the original model. For the fifth model, we decided to employ a backward elimination procedure so we can obtain the variables that affect the variability the most by eliminating the least significant one by one.

We obtained the following model:

```
Call:
lm(formula = y1 ~ x2 + x3 + x4 + x8 + x9)

Residuals:
     Min       1Q   Median       3Q      Max
-1473.53  -308.93   -13.87   258.49  1122.30

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.451e+02  1.101e+03   0.313  0.75545
x2          -2.114e+03  9.130e+02  -2.316  0.02519 *
x3           1.089e-02  1.523e-02   0.715  0.47814
x4           1.659e-03  1.265e-03   1.311  0.19638
x8           5.092e+01  1.581e+01   3.220  0.00238 **
x9           1.565e+04  4.616e+03   3.391  0.00146 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 558.9 on 45 degrees of freedom
Multiple R-squared:  0.5803,    Adjusted R-squared:  0.5337
F-statistic: 12.45 on 5 and 45 DF,  p-value: 1.319e-07
```

Y = 3.451e+02 -2.114e+03X2 + 1.089e-02X3 + 1.659e-03X4 + 5.092e+01X8 + 1.565e+04X9

We can see that the coefficient of determination, R-squared, is 0.5803, meaning that 58% of the total variability in y is accounted for by the 5 regressors. The value of the adjusted R-squared is 0.5337, which places it much closer to R-squared than in the original model, from this perspective we see an improvement. However, the R-squared for the original model remains

higher and adjusted R-squared from the original model is significantly lower. Checking for significance of the regression we can conclude that the regression is significant with a p-value that is much lower than 0.05. Checking for significance of βs we notice that β2, β8, and β9 are all significant for the regression. I have decided to keep β3, and β**4** because they were the last 2 coefficients that prevailed in the model after eliminating the other 8. Furthermore, both variables represent logically important elements when considering the number of deaths due to covid.

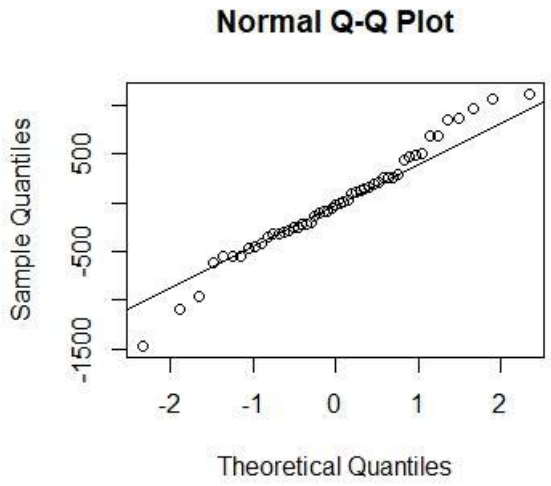


Figure 5.1: Normal Probability Plot.

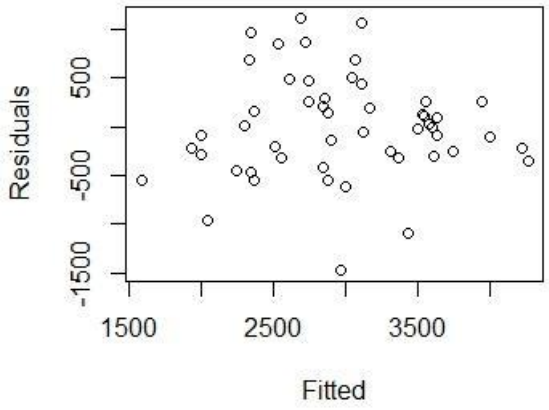


Figure 5.2: Residuals vs. Fitted Plot.

From **Figure 5.2** we can deduce that the variance of the error terms is in between the realm of being a constant or having y as proportion between 0 and 1. However, these two assumptions might change, and we could see an improvement if we decided to do an influential analysis and then remove the influential points or outliers. For examination purposes of this project, we will conclude that the variance of the error terms is close to a constant. **Figure 5.1** can be deceiving, since we have a noticeable number of points deviating from the straight line. This shows a need for normality test:

```
        Shapiro-Wilk normality test

data:  residuals(redmodel1)
W = 0.97993, p-value = 0.5359
```

Our p-value is 0.5359, that means we do not reject the null hypothesis and conclude that the residual terms are normally distributed. The two assumptions for the residuals are met.

Reduced Original Model by Best Subset Regression: Sixth Model Output and Interpretation

Another method for reducing a model is using in R for OLS (ordinary least squares) stepwise function for best subsets in the model. We applied this method to the original model, and we obtained the following output:

Best Subsets Regression

Model Index	Predictors
1	x9
2	x8 x9
3	x2 x8 x9
4	x2 x4 x8 x9
5	x2 x4 x7 x8 x9
6	x2 x3 x4 x7 x8 x9
7	x2 x3 x4 x7 x8 x9 x10
8	x2 x3 x4 x7 x8 x9 x10 x13
9	x2 x3 x4 x7 x8 x9 x10 x12 x13
10	x1 x2 x3 x4 x7 x8 x9 x10 x12 x13
11	x1 x2 x3 x4 x5 x7 x8 x9 x10 x12 x13
12	x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x12 x13
13	x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.4218	0.4100	0.3813	7.6214	805.9411	660.9354	811.7366	20157885.6964	410739.6054	8234.1666	0.6254
2	0.5167	0.4966	0.4181	0.6563	798.7976	654.7407	806.5249	17207817.3860	357089.2246	7175.5518	0.5437
3	0.5621	0.5342	0.462	-1.6309	795.7687	652.7910	805.4278	15930949.9758	336561.8711	6784.4488	0.5124
4	0.5756	0.5387	0.478	-0.9028	796.1760	654.0155	807.7669	15784264.6450	339365.4606	6868.1105	0.5167
5	0.5825	0.5362	0.4483	0.4388	797.3317	655.9458	810.8545	15877948.9471	347304.1685	7062.4053	0.5288
6	0.5899	0.5339	0.4315	1.7482	798.4307	657.9257	813.8853	15962694.2138	355099.8945	7261.4654	0.5406
7	0.5995	0.5343	0.3895	2.8356	799.2152	659.8083	816.6016	15957857.8716	360917.5624	7428.0854	0.5495
8	0.6046	0.5292	0.3505	4.3593	800.5690	662.1429	819.8873	16141262.8420	371042.7094	7692.3489	0.5649
9	0.6070	0.5207	0.3346	6.1288	802.2534	664.7145	823.5035	16442729.5234	384043.5352	8027.1395	0.5847
10	0.6081	0.5101	0.3021	8.0273	804.1139	667.3993	827.2958	16818261.5682	399004.6750	8415.7314	0.6075
11	0.6083	0.4978	0.2519	10.0096	806.0894	670.1444	831.2031	17252552.5118	415636.3031	8854.4075	0.6328
12	0.6084	0.4847	0.142	12.0000	808.0762	672.8960	835.1218	17714276.6422	433233.5690	9330.6216	0.6596
13	0.6084	0.4708	0.0342	14.0000	810.0762	675.6527	839.0536	18206326.8003	451894.4876	9848.9824	0.6880

There needs to be a major human logic input towards selecting the model to be used. There are many metrics in these tables, and one must choose which makes the most sense overall. For this model, I selected model #6. Overall, it has the most sense regarding the variables selected, and metrics provided.

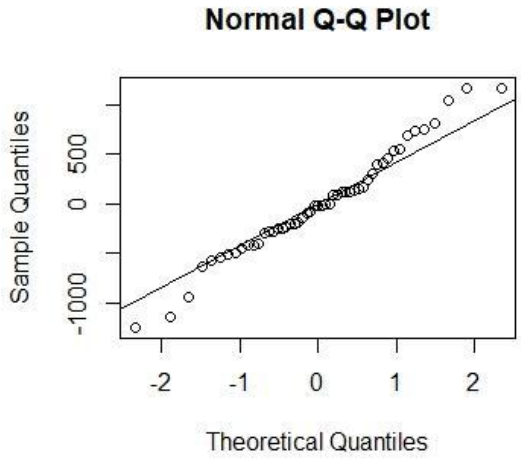
Then, we have the model:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.079e+02  1.450e+03  -0.419  0.67712
x2          -1.456e+03  1.122e+03  -1.298  0.20101
x3           1.370e-02  1.547e-02   0.886  0.38067
x4           1.897e-03  1.286e-03   1.475  0.14741
x7           3.791e+01  3.754e+01   1.010  0.31805
x8           5.304e+01  1.595e+01   3.325  0.00179 **
x9           1.340e+04  5.128e+03   2.612  0.01227 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 558.8 on 44 degrees of freedom
Multiple R-squared:  0.5899,    Adjusted R-squared:  0.5339
F-statistic: 10.55 on 6 and 44 DF,  p-value: 3.113e-07
```

Y= -6.079e+02 -1.456e+03X2 + 1.370e-02X3 + 1.897e-03X4 + 3.791e+01X7 + 5.304e+01X8 + 1.340e+04X9

The regression is significant with a p-value that is less than 0.05. R-squared is 0.5899, so 59% of the total variability in y is owed by the 5 regressors used in this model. Adjusted R-squared is 0.5362, the difference between R-squared and adjusted R-squared is minimal like the fifth model. Checking for significance of the variables we find that X8 and X9 are both significant with the other variables not being as significant, but still carrying some weight towards the regression.



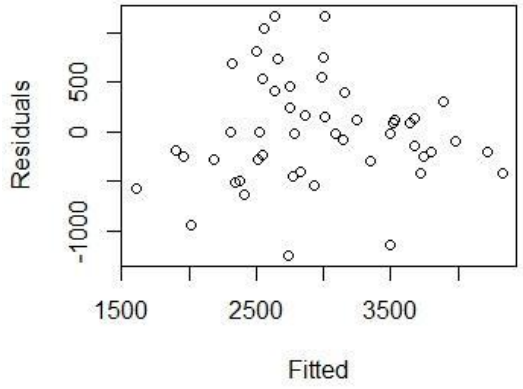


Figure 6.1: Normal Probability Plot. **Figure 6.2: Residuals vs Fitted Plot.**

From the plots above we can make the same conclusions as we did from the sixth model. Thus, we are establishing that the two assumptions for the residuals are met.

```
        Shapiro-Wilk normality test

data:  residuals(rmodel2)
W = 0.98113, p-value = 0.5879
```

We confirm that the residuals are normally distributed.

Evaluation of the Models

To proceed with our data analysis, we want to select a reduced model and determine which has the most powerful predictive power. To do so, we are going to compare them by applicating metric of model evaluation; namely, mean absolute error (MAE), root mean squared error (RMSE), range normalized RMSE (NRMSE), mean absolute percentage error (MAPE), and the predicted residual error sum of squares (PRESS) statistics. In this sense, we will be decisively comparing the performance of the models obtained.

Model	MAE	RMSE	NRMSE	MAPE	PRESS
Reduced Original Model by Backward Elimination (Fifth)	**403.82**	525.0056	0.166	0.1662	0.4572
Reduced Original Model by Best Subset Function (sixth)	403.87	**519.0242**	**0.1641**	**0.1646**	**0.4314**

Thus, we arrive to the conclusion that the model with the better predictive power is the sixth model, The Re Original Model by Best Subset Function. Said model has the lowest overall values of metrics for model evaluation.

Cross Validation of the Model

To cross validate the model that was previously selected, we are going to fit another model, using the same variables, but testing 80% of the original data, that is 41 states. The 20% (11 states) left of the data will be used to evaluate how close are to the actual values for those states. The 11 states that will be used to test the predictive power are: New Mexico, Maryland, Pennsylvania, Missouri, Wyoming, Mississippi, Louisiana, New York, Massachusetts, New Hampshire, and Wisconsin. Obtaining the following model.

```
Residual standard error: 571.7 on 33 degrees of freedom
Multiple R-squared:  0.6152,    Adjusted R-squared:  0.5453
F-statistic: 8.794 on 6 and 33 DF,  p-value: 9.418e-06

> print(cmodel)

Call:
lm(formula = ny ~ nx2 + nx3 + nx4 + nx7 + nx8 + nx9)

Coefficients:
(Intercept)          nx2          nx3          nx4          nx7          nx8          nx9
 -3.864e+02   -1.697e+03    6.439e-03    2.214e-03    4.757e+01    5.554e+01    1.058e+04
```

Y= **-3.864e+02 -1.697e+03X2 + 6.439e-03X3 + 2.214e-03X4 + 4.757e+01X7 + 5.554e+01X8 1.058e+04X9**

The model is significant with a p-value that is less than 0.05. with an R-squared of 0.6152. In short, around 62% of the total variability in y is owed by X2, X3, X4, X7, X8, and X9.

State	Actual Number of Deaths	Predicted Number of Deaths
Louisiana	3,735	3,507
Maryland	2,430	2670.23
Massachusetts	3,016	2233.002
Mississippi	4,196	3980.496
Missouri	3,363	3208.705
New Hampshire	1,886	2321.613
New Mexico	3,751	2846.86
New York	3,590	2473.866
Pennsylvania	3,551	2991.993
Wisconsin	2,525	2414.589
Wyoming	3,152	3019.677

Overall, the predicted values and the actual values are similar enough to notice the capability of the model to predict the number of deaths per 1 million. Differences between actual values and predicted values can arguably be considered not substantial. Thus, we can conclude that the sixth model of the project.

DISCUSSION AND CONCLUDING REMARKS

The project in its entirety was designed around factors that have been, overtime, related to the severity of the coronavirus disease. The data analyzed was intended to shed light towards the unpredictable capacity that COVID-19 has to take infected peoples' lives. Though the final model did not present a perfect coefficient of determination (R-squared) it did show some strength at the time of predicting the number of deaths per million.

The variables selected for the last model go in hand with what previous research has shown regarding the factors that affect the severity of the disease and ultimately the risk of dying from it. To elaborate about the selection of the variables, we must recall what the variables are. X2 is the vaccination rate of the state, X3 is the hospitalization rate per 1 million, X4 is the population density, X7 is the percentage of smokers in the state, X8 is the average Air Quality Index of the state, and X9 is the prevalence of diabetics in the state. Hence, the variables used to predict the number of deaths per million have a solid foundation about research on COVID-19.

The relatively low values of R-squared and adjusted R-squared for the selected model, 0.5899 and 0.5339 could be the result of the limited sample size used for the project (N=51). To polish the data for a better analysis a larger sample size and ideally with less variability. Although these could a couple of reasons why the model's coefficient of determination was not as high as expected, they might not be the only ones. There are factors/variables that were not taken into consideration for the project. For instance, the rate of COVID-19 positive patients that were admitted to the ICU, the amount of personnel available at hospitals, the percentage of people over the age of 65 in the state, etc.

To conclude, this model could represent a foundation for future models in which more and different variables are used. The findings in this project are based on the rationale surrounding

COVID-19 and its severity, and the basic principles for linear regression models. More research needs to be conducted in order to obtain more accurate models and results.

REFERENCES

"Adult Obesity Rates." *The State of Childhood Obesity*, 19 May 2022, stateofchildhoodobesity.org/adult-obesity.

"Average Annual Temperatures by USA State." *Current Results*, www.currentresults.com/Weather/US/average-annual-state-temperatures.php. Accessed 1 June 2022.

CDC. "COVID Data Tracker." *Centers for Disease Control and Prevention*, 28 Mar. 2020, covid.cdc.gov/covid-data-tracker/#new-hospital-admissions.

"COVID-19 United States Cases by Sate." *Johns Hopkins Coronavirus Resource Center*, coronavirus.jhu.edu/us-map. Accessed 1 June 2022.

"GDP by State." *U.S. Bureau of Economic Analysis (BEA)*, www.bea.gov/data/gdp/gdp-state. Accessed 1 June 2022.

Guzman, C. I and Kibria, B. M. G. (2019). Developing Multiple Linear Regression Models forthe Number of Citations: A Case Study of Florida International University Professors. International Journal of Statistics and Reliability Engineering, 6(2), 75-81.

"Hypertension in the United States." *The State of Childhood Obesity*, 21 Sept. 2021, stateofchildhoodobesity.org/hypertension.

"Median Age by State 2022." *World Population Review*, worldpopulationreview.com/state-rankings/median-age-by-state. Accessed 1 June 2022.

Montgomery, Douglas, et al. *Introduction to Linear Regression Analysis*. 5th ed., Wiley, 2012.

Rosas, Estrella. “Best States for Education (2022 Rankings).” *Scholaroo*, 30 May 2022, scholaroo.com/state-education-rankings.

Siegrist, M and Kibria, B. M. G. (2020). Predicting Total Death using COVID-19 Data: Regression Model Approach. International Journal of Industrial and Operations Research.

Statista. “Number of COVID-19 Deaths in the United States as of June 1, 2022, by State.” *Statista*, 1 June 2022, www.statista.com/statistics/1103688/coronavirus-covid19-deaths-us-by-state.

“U.S. Air Quality Index State Rank.” *USA.Com*, www.usa.com/rank/us--air-quality-index--state-rank.htm. Accessed 1 June 2022.

USAFacts. “US Coronavirus Vaccine Tracker.” *USAFacts*, 1 June 2022, usafacts.org/visualizations/covid-vaccine-tracker-states.

DATA TABLE

State	YTotDeath	X2Vaccination	X3HospitalPerM	X4Pop	X5MedianAge	X6GDPcap
Alabama	4,017	0.518	9384.248406	94407	39.5	49027
Alaska	1,711	0.6234	16996.26459	1	35.3	75027
Arizona	4,173	0.7153	17938.20937	67	38.5	56511
Arkansas	3,820	0.5522	10873.31887	58	38.6	47770
California	2,329	0.7316	13044.00827	255	37.3	85546
Colorado	2,318	0.7257	11795.40385	58	37.3	72597
Connecticut	3,090	0.7972	15312.74518	732	41.2	82233
Delaware	3,050	0.7114	15968.85012	512	41.4	80446
District Of Columbia	1,910	0.9908	11650.28463	11535	34.4	226861
Florida	3,490	0.6935	7778.245491	429454	42.7	56301
Georgia	3,606	0.5634	11627.445	190	37.3	63271
Hawaii	1,035	0.7724	5142.039175	218	40	62474
Idaho	2,770	0.5798	16473.02323	23	37.2	49616
Illinois	3,026	0.6845	13250.10222	225	38.8	74052
Indiana	3,535	0.5589	15516.00686	191	38	61760
Iowa	3,047	0.6262	29707.18761	57	38.6	68849
Kansas	3,071	0.625	15533.70616	36	37.3	65530
Kentucky	3,597	0.5799	7573.954108	114	39.2	52002
Louisiana	3,735	0.5364	11961.76443	107	37.8	55213
Maine	1,789	0.8079	9714.758195	44	45	55425
Maryland	2,430	0.7664	10167.33007	626	39.2	71083

Massachusetts	3,016	0.8027	15107.44254	887	39.7	91129
Michigan	3,659	0.6069	16111.79797	149032	40.1	56554
Minnesota	2,312	0.7056	12330.38738	72	38.5	72187
Mississippi	4,196	0.5202	17060.82419	63	38.3	42411
Missouri	3,363	0.5783	9373.315359	90	39.1	58356
Montana	3,213	0.5816	13818.21375	24171	40.2	53703
Nebraska	2,238	0.646	14627.80227	26	36.9	76584
Nevada	3,550	0.6355	13491.87432	30	38.5	61375
New Hampshire	1,886	0.719	12146.73719	154	43.1	70729
New Jersey	3,814	0.7729	15399.00953	1,206	40.2	72524
New Mexico	3,751	0.7197	17053.32318	17	38.6	51481
New York	3,590	0.7712	23659.51128	408	39.4	93463
North Carolina	2,351	0.6363	7258.394727	222	39.2	62077
North Dakota	2,998	0.563	14537.27229	11	35.4	81795
Ohio	3,307	0.5887	8840.234821	287	39.6	62517
Oklahoma	3,654	0.5814	15207.69569	58	37.1	51861
Oregon	1,828	0.716	16723.99607	45	39.9	62867
Pennsylvania	3,551	0.6935	11162.34211	286	40.9	64751
Rhode Island	3,393	0.8369	6692.425668	1,028	40.3	60185
South Carolina	3,495	0.5898	10373.71014	73877	40.1	52031
South Dakota	3,310	0.6332	6831.612736	15006	37.6	68357
Tennessee	3,887	0.5586	17515.30427	170	39.1	59969
Texas	3,067	0.6367	17321.62185	115	35.2	67235
Utah	1,495	0.6689	16359.14085	41	31.5	66011
Vermont	1,074	0.8177	19254.21945	68	43	56028
Virginia	2,395	0.7504	19995.30956	219	38.7	68483
Washington	1,715	0.7512	14922.81305	119	37.9	86265
West Virginia	3,913	0.5717	22615.54785	73	43	49017
Wisconsin	2,525	0.6634	34.57173966	108	40	62065
Wyoming	3,152	0.5103	13.5645486	6	38.7	71911

State	X7SmokerPre	X8AQI	X9DiabetesPre	X10ObesityPre	X11AvgTemp	X12RateHyper
Alabama	19.2	46.6	0.15	0.39	62.8	0.425
Alaska	19.1	29.1	0.079	0.319	26.6	0.328
Arizona	14	45.4	0.113	0.309	60.3	0.325
Arkansas	22.7	43.1	0.132	0.364	60.4	0.41

California	11.2	46	0.102	0.303	59.4	0.278
Colorado	14.5	47.1	0.075	0.242	45.1	0.258
Connecticut	12.2	45	0.095	0.292	49	0.309
Delaware	16.5	46.4	0.127	0.365	55.3	0.272
District Of Columbia	13.8	46.8	0.078	0.243	59.3	0.364
Florida	14.5	38.9	0.118	0.284	70.7	0.335
Georgia	16.1	48.2	0.118	0.343	63.5	0.348
Hawaii	13.4	21.2	0.111	0.245	70	0.307
Idaho	14.7	44.3	0.086	0.311	44.4	0.306
Illinois	15.5	43.6	0.103	0.324	51.8	0.322
Indiana	21.1	47.5	0.12	0.368	51.7	0.348
Iowa	16.6	37.6	0.101	0.365	47.8	0.318
Kansas	17.2	42.8	0.112	0.353	54.3	0.335
Kentucky	23.4	46.1	0.131	0.366	55.6	0.409
Louisiana	20.5	40.4	0.143	0.381	66.4	0.397
Maine	17.8	36.5	0.105	0.31	41	0.362
Maryland	12.5	47	0.103	0.31	54.2	0.343
Massachusetts	13.4	41.4	0.09	0.244	47.9	0.281
Michigan	18.9	42.5	0.121	0.352	44.4	0.351
Minnesota	15.1	38.3	0.087	0.307	41.2	0.287
Mississippi	20.5	43.7	0.146	0.397	63.4	0.436
Missouri	19.4	44	0.109	0.34	54.5	0.309
Montana	18	39.6	0.091	0.285	42.7	0.295
Nebraska	16	37	0.098	0.34	48.8	0.31
Nevada	15.7	42.1	0.112	0.287	49.9	0.328
New Hampshire	15.6	38.9	0.09	0.299	43.8	0.315
New Jersey	13.1	44.1	0.1	0.277	52.7	0.302
New Mexico	15.2	42.1	0.122	0.309	53.4	0.316
New York	12.8	40.4	0.108	0.263	45.4	0.296
North Carolina	17.4	46.5	0.127	0.336	59	0.351
North Dakota	19.1	37	0.099	0.331	40.4	0.298
Ohio	20.5	48.2	0.125	0.355	50.7	0.345
Oklahoma	19.7	43.5	0.13	0.364	59.6	0.378
Oregon	15.6	36.1	0.097	0.281	48.4	0.306
Pennsylvania	17	45.6	0.114	0.315	48.8	0.333
Rhode Island	14.6	43.7	0.106	0.301	50.1	0.33

South Carolina	18	44.8	0.136	0.362	62.4	0.383
South Dakota	19	39.6	0.078	0.332	45.2	0.309
Tennessee	20.7	47.5	0.142	0.356	57.6	0.393
Texas	14.4	41	0.13	0.358	64.8	0.317
Utah	9	51.2	0.08	0.286	48.6	0.258
Vermont	13.7	38.5	0.08	0.263	42.9	0.302
Virginia	14.9	45	0.111	0.322	55.1	0.336
Washington	12	33.5	0.088	0.28	48.3	0.303
West Virginia	25.2	47.6	0.157	0.391	51.8	0.438
Wisconsin	16.4	39.5	0.09	0.323	43.1	0.31
Wyoming	18.8	45	0.083	0.307	42	0.307

State	X13EdQua
Alabama	59.25
Alaska	53.98
Arizona	51.56
Arkansas	64.17
California	58.55
Colorado	68.01
Connecticut	81.44
Delaware	74.07
District Of Columbia	76.8525
Florida	64.58
Georgia	63.27
Hawaii	61.78
Idaho	60.71
Illinois	70.03
Indiana	64.89
Iowa	70.81
Kansas	68.15
Kentucky	68.34
Louisiana	51.95
Maine	70.45
Maryland	74.33
Massachusetts	86.12
Michigan	61.94
Minnesota	72.61

Mississippi	59.61
Missouri	68.65
Montana	61.21
Nebraska	71.93
Nevada	56.73
New Hampshire	74.91
New Jersey	85.51
New Mexico	54.3
New York	77.38
North Carolina	62.14
North Dakota	65.65
Ohio	68.02
Oklahoma	56.92
Oregon	62.68
Pennsylvania	72.24
Rhode Island	72.16
South Carolina	60.48
South Dakota	64.88
Tennessee	64.39
Texas	65.82
Utah	66.09
Vermont	74.82
Virginia	75.59
Washington	67.54
West Virginia	59.38
Wisconsin	71.94
Wyoming	66.16