

Predicting the Total number of COVID-19 Deaths by County for the State of Florida: Ridge Regression Approach

B. M. GOLAM KIBRIA, MANUEL URBISTONDO

Department of Mathematics and Statistics

Florida International University

Miami, FL 33199, USA

Abstract: - Coronavirus disease 2019 (COVID-19) is a novel infectious disease that was detected in Wuhan, China at the end of 2019. The virus quickly spread worldwide and caused a global pandemic. This paper investigates if there are any regressors that could help impact the number of deaths due to COVID-19. The variables that were used in the models were total deaths, hospitalizations, total cases, population, minimum temperature, average temperature, maximum temperature, precipitation, mobility index, median age, adults age 65 or older, PM2.5 average, ozone average, and positive non-residents. After fitting six different regression models, we found that the most significant regressors were hospitalizations per county, total cases per county, population per county, median age per county, positive adults 65 or older per county, and positive non-residents per county. The COVID-19 data of this paper will be an excellent source for illustrating the multicollinear linear regression models.

Key-Words: - Cross Validation; Linear Regression Model; MAPE; Multicollinearity; OLS; Ridge Regression; RMSE.

Received: June 22, 2021. Revised: November 21, 2021. Accepted: December 13, 2021. Published: December 31, 2021.

1 Introduction

Coronavirus Disease 19 (COVID-19) is the disease caused by a new coronavirus called SARS-CoV-2. It first appeared at the end of 2019 in Wuhan, China. It rapidly spread out throughout Asia and eventually in the world, leading the World Health Organization to declare a Public Health Emergency of International Concern and a global pandemic. The virus reached the West Coast of the United States in January 2020. On January 31st President Donald Trump declared a public health emergency under the Public Health Service Act. These actions were taken to mitigate the potential impact in the United States. To date, coronavirus has infected about 66 million and killed about 872 thousand people in the United States. At present, in Florida, the total number of positive cases is above 5 million and about 63 thousand deaths (January 15, 2022).

The infectious disease COVID-19 may not have a high death ratio, but the potential long-term risks are still being studied. The most common symptoms of the disease are fever, cough, shortness of breath, fatigue, body aches, headache, loss of taste or smell, nausea, and other infrequent but possible ailments. Mild or moderate COVID-19 will last roughly two weeks for most people, but others suffer lingering health problems even when they have recovered

from the illness. While it's clear that people with certain risk factors are more likely to have complications of COVID-19, there isn't a clear link between risk factors and long-term problems. This means COVID-19 can have prolonged effects, even on those who only suffered mild symptoms. The complete severity of coronavirus is still an unknown, for this reason, this disease is still the most prevalent topic in the world.

The purpose of this study is to fit different linear regression models and find possible variables that could contribute to the coronavirus death count in Florida. This paper will study all 67 counties in the state of Florida to determine if there are specific factors that could influence deaths due to COVID-19. The data of this paper was collected in April 2021. The variables we used in the models are:

- Total Deaths (Y): Deaths due to coronavirus is our dependent variable. Each observation is tracked by deaths per county.
- Total Hospitalizations (X1): Hospitalizations could be related to more deaths due to severity of those that visited the hospital.
- Total Cases (X2): Coronavirus may have a high survival rate, but the amount of total

cases could help predict the number of deaths.

- Population (X3): Considering how this disease is transmitted, areas with a higher population may influence the number of infected individuals in a particular area.
- Minimum Temperature (X4): We want to consider if areas with lower atmospheric temperature could have a relation to the infectious rate of the coronavirus.
- Average Temperature (X5): Average temperature could give us an idea of which counties have prevailing weather that could cause infection of the disease.
- Maximum Temperature (X6): COVID-19 is known to be less infectious in areas with warm weather. The maximum temperature could have a relation to areas that have less infections.
- Precipitation (X7): This is another environmental factors that we will consider that could influence COVID-19 deaths.
- Mobility Index (X8): Since COVID-19 is transmitted from person to person, counties with higher Mobility Index could have more deaths due to corona virus.
- Median Age (X9): COVID-19 affects ages differently. A higher median age could anticipate a larger number of deaths in a county.
- Age 65 or Older (X10): The death rate of the disease is increased dependent on the cases for people who are 65 or older.
- PM2.5 Average (X11): COVID-19 is a respiratory disease, higher levels of PM2.5 could affect the severity of the disease in a particular area.
- Ozone Average (X12): Areas where the ozone concentration is less could affect the spread of COVID-19.
- Positive Non-residents (X13): Areas where there are more positive non-residents could lead to a higher number of deaths due to COVID-19.

For detailed on linear regression models with applications, we refer to Draper and Smith (1998) and Montgomery (2013), and Abdullah et al. (2016) and recently Saleh et al. (2019) among others. There are limited number of applications in literature for fitting regression mode with COVID-19 data. First, Ghosal et al. (2020) consider SARS-CoV-2 at 6 weeks from day 0 data to predict the number of deaths in India. Qin et al. (2020) consider COVID-

19 data to predict the number of cases using social media search index data. Ogundokun et al. (2020) consider the regression model to predict the COVID-19 confirmed cases in Nigeria. However, none of the aforementioned researchers have addressed the issue of multicollinearity, ie, when the regressors in the linear regression model are correlated. There were many papers those handle the problem of multicollinearity. To mention a few, Gibon (1978), Kibria (2003), Gruber (2010), Muniz and Kibria (2009), Muniz et al (2012), Kibria and Banik (2016), Lukman et al. (2019a, 2019b), Kibria and Lukman (2020), Suhail et al. (2020), and very recently Qasim et al. (2021) and Amin et al. (2021) among others.

The objective of this paper is two folds: (i) Find a suitable model that provides the better prediction of deaths (ii) The data of this paper will be an important source for illustrating multicollinear linear regression model. The organization of the paper is as follows: The sources of data and descriptive statistics are given in section 2. The statistical models and data analysis are given in section 3. Cross validation about the models are given in section 4. Finally, some concluding remarks are provided in section 5.

2 Data Collection and Descriptive Statistics

2.1 Data Collection

The focus when collecting data was to see which factors can help predict COVID-19 deaths in each Florida county. The initial data collection was challenging due to the information being limited for some particular counties around the state. At the beginning of the project, the timeline of the data was going to only be from January 2020 until January 2021. We believed this would be the most accurate time for our variables to be significant. However, the statistics we collected are either the total until April 5, 2021 or it was recorded on a yearly basis. The most complete data set we were able to find was from Florida Department of Health, Division of Disease Control and Health Protection. Their Florida's COVID-19 Data and Surveillance Dashboard presents a variety of statistics for each Florida county. Their county cases map gives the information for total number of deaths, hospitalizations, population, positive residents, and positive non-residents. It also states the median age, and total number of people age 65 or older for each

county. These variables were the totals until April 5, 2021.

We also wanted to consider if there were any environmental factors that could help us predict the number of deaths due to coronavirus. The National Oceanic and Atmospheric Administration (NOAA) provides a county map called Climate at a Glance. The map offers the statistics for minimum temperature, average temperature, maximum temperature, and precipitation for each county and at the time of your choosing. We recorded each of the previous variables for the 67 counties and selected the data to be dated from January 2020 to January 2021.

There were different environmental factors that we considered to use. Since COVID-19 is a respiratory disease, we wanted to find any air pollutants that could exaggerate the infection of coronavirus. The Florida Department of Environmental Protection reports data on the carbon Monoxide (CO), Nitrogen Dioxide (NO₂), Ozone (O₃), Particular Matter Trend (PM₁₀), Particular Matter Pollution (PM_{2.5}) and Sulfur Dioxide (SO₂) averages for a number of Florida counties. However, only (PM_{2.5}) and (O₃) averages had data for at least 32 of the 67 counties. To complete the values for each of the 67 counties, we estimated the missing data by averaging the

values from the neighboring counties. If there was any data value that was not given for the PM_{2.5} or Ozone average, it was estimated by averaging the value of the closest 2 counties. The PM_{2.5} and Ozone averages were recorded from the yearly 2020 report. The last data variable that was collected was for Mobility Index by county. It is the measure the daily average distance traveled. The Center for Disease Control and Prevention has a COVID Data Tracker that contributes the Mobility Index for each county. The data was collected as the average Mobility Index for the past year. The complete data set is presented in the Appendix of this paper.

2.2 Descriptive Statistics

The following Table 1 contains the descriptive statistics for each of the 14 variables that will be used for the multiple linear regression models in the following section. It provides the minimum, maximum, median, mean, and the standard deviation of the data set value for each variable (1 dependent variable and the 13 independent variables).

All analyses of this paper are done by free software R 4.1.1, which is available from online. However, one can use the books by Cohen and Cohen (2008, 2021) among others.

Table 1. Summary statistics of the variables

	Variables	Min	Max	Median	Mean	Standard Deviation
Y	Deaths	16	5900	259	502.3	871.061
X1	Hospitalizations	22	12624	610	1279	2124.816
X2	Total Cases	926	449891	8852	30947	64400.11
X3	Population	8354	2716940	132420	320563	501304.4
X4	Min. Temperature	57.5	70.4	62.9	62.57	3.4107
X5	Avg. Temperature	68.4	78.5	72.9	72.68	2.746
X6	Max Temperature	78.7	86.6	82.7	82.78	2.242
X7	Precipitation	47.47	78.94	54.82	57.61	7.331
X8	Mobility Index	3.8	4.4	4.1	4.131	0.169
X9	Median Age	27	65	43	42.82	5.447
X10	Age 65 or More	116	55621	2433	4593	7925.297
X11	PM _{2.5} Average	15.28	32.86	20.5	21.95	4.769
X12	Ozone Average	50.4	63	57.5	57.3	2.311
X13	Positive Non-Residents	0	5773	103	539.7	984.75

3 Statistical Analysis

We will consider the following linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \varepsilon \quad (1)$$

where y = total deaths because of COVID-19, x_1 = hospitalizations per county, x_2 = total cases per county, x_3 = population per county, x_4 = minimum temperature per county, x_5 = average temperature per county, x_6 = max temperature per county, x_7 = precipitation per county, x_8 = mobility index per county, x_9 = median age per county, x_{10} = age 65 or more per county, x_{11} = PM2.5 average per county, x_{12} = ozone average per county, and x_{13} = positive non-residents per county. In order to fit the model, we will assume that all regressors are independent and that the residuals are normally distributed with mean 0 and variance σ^2 . In this section we will fit different types of regression models and try to find a possible best model.

3.1 Full Model (Model 1)

Using R software, we obtain the following regression analysis for model 1.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.708e+02  8.740e+02   0.195  0.845801
X1           7.762e-02  2.189e-02   3.545  0.000829 ***
X2          -1.991e-03  1.333e-03  -1.494  0.141205
X3           1.237e-04  8.949e-05   1.382  0.172621
X4          -1.532e-02  1.816e-02  -0.844  0.402597
X5           3.021e+02  3.610e+02   0.837  0.406428
X6          -1.524e+02  1.797e+02  -0.848  0.400123
X7           2.779e+00  1.947e+00   1.427  0.159442
X8           4.456e+01  1.010e+02   0.441  0.660902
X9          -3.041e+00  2.688e+00  -1.131  0.263062
X10          1.060e-01  1.268e-02   8.360  2.99e-11 ***
X11          1.985e-01  3.231e+00   0.061  0.951245
X12          -3.007e+00  6.736e+00  -0.446  0.657146
X13          -6.889e-02  2.849e-02  -2.418  0.019089 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.17 on 53 degrees of freedom
Multiple R-squared:  0.9912,    Adjusted R-squared:  0.989
F-statistic: 459.3 on 13 and 53 DF,  p-value: < 2.2e-16
```

The first fitted model (model 1) is,

$$\hat{y} = 0.01708 + 0.07762x_1 - 0.001991x_2 + 0.0001237x_3 - 0.01532x_4 + 0.03021x_5 - 0.01524x_6 + 2.779x_7 + 0.4456x_8 - 3.041x_9 + 0.106x_{10} + 0.1985x_{11} - 3.007x_{12} - 0.06889x_{13}$$

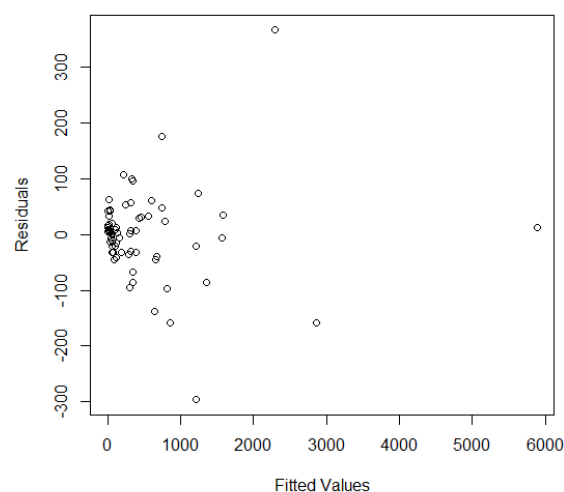


Fig. 1: Residual vs fitted plot for model 1

From (Model 1) we can see that the R Square is 0.9912, which means that about 99% of total variation of deaths due to COVID-19 can be explained by the 13 variables. There are only three regressors statistically significant. However, the p-value in (Model 1) is significantly less than 0.05; then, we can reject the null hypothesis that the regressors are not significant. Hence, we can assume that at least one of the 13 variables are significant to the model. To check the model diagnostics, the residuals vs. fitted plot and normal Q-Q plot are shown in Figures 1 and 2 respectively.

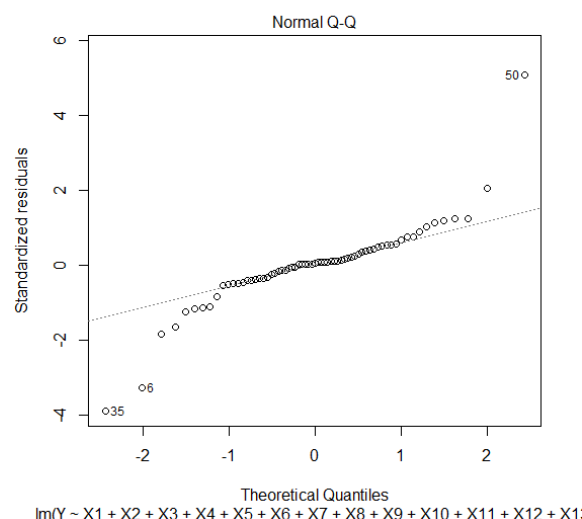


Fig. 2: Q-Q plot of the residuals for model 1

The Residuals vs Fitted plot (Figure 1) demonstrates how the data is clustered to the left, suggesting that the variance is not constant. Since the residuals

seem to follow a heavy-tailed distribution within the Normal Q-Q plot (Figure 2), we cannot conclude that the distribution of these residuals is completely normal. Although the residuals at the extremities are not close to the line, the center of the graph may indicate that the distribution is approximately normal. A Shapiro Wilk test of normality (p-value=0.0000013) will test the null hypothesis that the residuals are normally distributed. The p-value is significantly less than 0.05, we reject the null hypothesis and conclude the model is not normally distributed. Thus the full model is not adequate to predict the total number of deaths.

Since Model 1 is not an adequate to predict the number of deaths per county, we tried several different transformations on the dependent variable.

3.2 Log Transformation Model (Model 2)

First we attempted the most popular log (y) transformation model:

$$\log(\hat{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \varepsilon \quad (2)$$

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.058e+01  5.812e+00   3.542 0.000839 ***
X1           7.661e-05  1.456e-04   0.526 0.600962
X2          -4.409e-05  8.866e-06  -4.973 7.29e-06 ***
X3           3.239e-06  5.952e-07   5.441 1.38e-06 ***
X4          -1.231e+00  1.208e+00  -1.019 0.312661
X5           2.174e+00  2.401e+00   0.906 0.369270
X6          -9.932e-01  1.195e+00  -0.831 0.409608
X7           5.784e-03  1.295e-02   0.447 0.656966
X8          -3.731e+00  6.718e-01  -5.553 9.21e-07 ***
X9           2.711e-03  1.788e-02   0.152 0.880067
X10          2.799e-04  8.436e-05   3.318 0.001642 **
X11          1.828e-02  2.149e-02   0.851 0.398832
X12          -9.485e-03  4.480e-02  -0.212 0.833130
X13          -1.344e-04  1.895e-04  -0.709 0.481224
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6064 on 53 degrees of freedom
Multiple R-squared:  0.8573,    Adjusted R-squared:  0.8223
F-statistic: 24.49 on 13 and 53 DF,  p-value: < 2.2e-16

```

The fitted regression model (model 2) is

$$\log(\hat{y}) = 20.58 + 0.00007661x_1 - 0.00004409x_2 + 0.000003239x_3 - 1.231x_4 + 2.174x_5 - 0.9932x_6 + 0.005784x_7 - 3.731x_8 + 0.002711x_9 + 0.0002799x_{10} + 0.01828x_{11} - 0.009485x_{12} - 0.000134x_{13}$$

Model 2 has R square equal to 0.8573, which represents that almost 86% of total variation of deaths due to COVID-19 can be explained by the 13 variables. The R square for (Model 2) is less than the R square for (Model 1). The p-value in (Model 2) is still significantly less than 0.05; we reject the

null hypothesis that the regressors are not significant. Hence, we can assume that at least one of the 13 variables are significant in (Model 2). To check the model diagnostics, residuals vs. fitted plot and the normal Q-Q plot are shown in Figures 3 and 4 respectively.

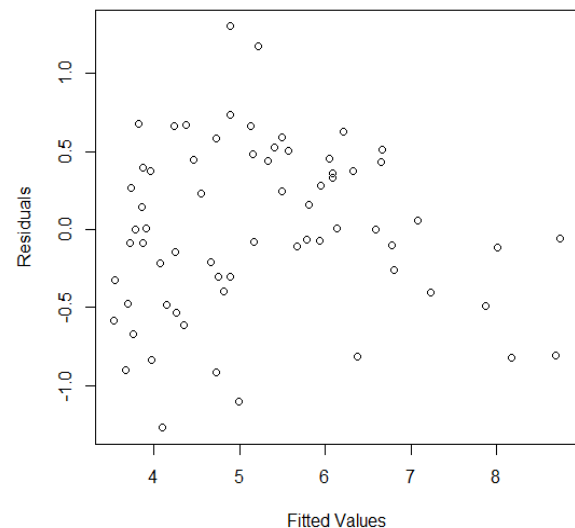


Fig. 3: Residual vs fitted plot for model 2

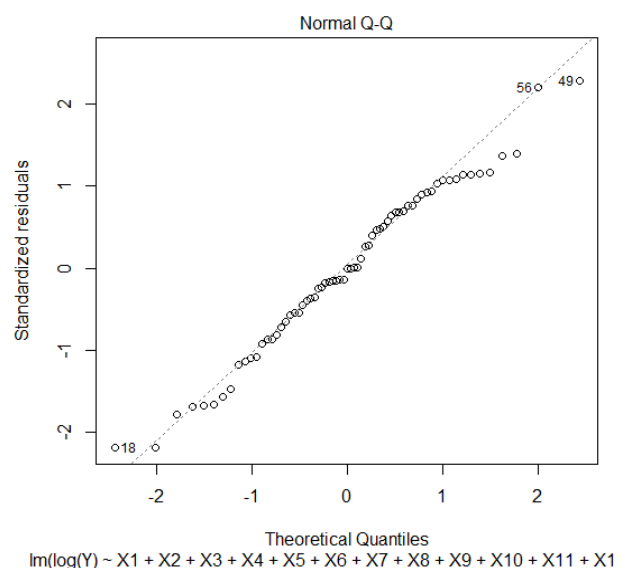


Fig. 4: Q-Q plot of the residuals for model 2

The log transformation significantly improved the residuals vs fitted plot and normal Q-Q plot. Figure 3 follows a more scattered pattern around 0 than that of Figure 1, yet the data is still slightly scattered to the left side of the graph. Figure 4 shows an improvement on the normal probability plot of the residuals. The normality line is straighter and more of the residuals lie on the line. The Shapiro-Wilk

test on the residuals has a p-value = 0.5147, we cannot reject the null hypothesis that the residuals are normally distributed. This model meets the normality of residuals assumption. We can carry on using the full log transformation.

3.3 Reduced Log Transformation (Model 3)

$$y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_8 X_8 + \beta_{10} X_{10} + \varepsilon \quad (3)$$

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.345e+01  3.992e+00  5.874 1.90e-07 ***
X2           -4.153e-05  7.007e-06 -5.926 1.55e-07 ***
X3            3.190e-06  5.017e-07  6.358 2.90e-08 ***
X4           -8.671e-02  3.254e-02 -2.665 0.00985 **
X8           -3.318e+00  6.113e-01 -5.428 1.04e-06 ***
X10          2.642e-04  5.314e-05  4.972 5.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5887 on 61 degrees of freedom
Multiple R-squared:  0.8452,    Adjusted R-squared:  0.8325
F-statistic: 66.6 on 5 and 61 DF,  p-value: < 2.2e-16

```

The fitted reduced model (model 3) is
 $\log(\hat{y}) = 23.45 - 0.00004153x_2 + 0.000003190x_3 - 0.08671x_4 - 3.318x_8 + 0.0002642x_{10}$

The results show that R square is 0.8452, which represents that almost 85% of total variation of deaths due to COVID-19 can be explained by these 5 variables. The R square for (Model 3) is slightly less than the R square for (Model 2). To check the model diagnostics, residuals vs. fitted plot and the normal Q-Q plot are shown in Figures 5 and 6 respectively.

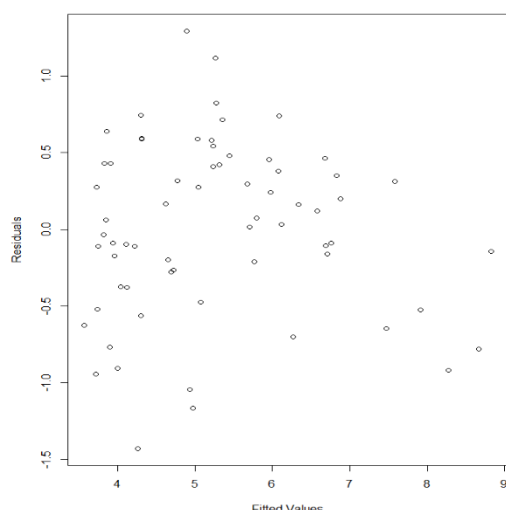


Fig. 5: Residual vs fitted plot for model 3

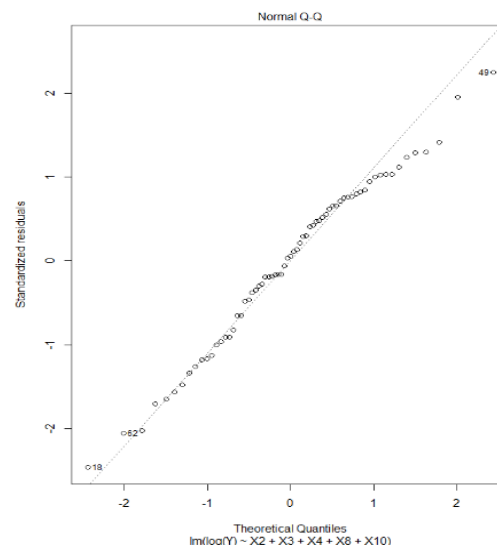


Fig. 6: Q-Q plot of the residuals for model 3

Compared to Model 2, the reduced model did not have a significant effect on the plots. The reduced log transformation model meets the required assumptions. We have also fitted both square root and inverse transformations models and found that constant variance assumption have not met, so we have decided not to report in this paper.

3.4 Full Beta Regression (Model 4)

We divided the dependent variable (y) by x_3 , ie. $y^* = y/x_3$, which is the total population by county, to get a proportion for the beta model. To see whether y^* follow beta distribution or not, we have done the following analysis in R. Using R we obtain the values of the parameters of Beta distribution as, $\alpha=6.725$ and $\beta=3674.04$. One sample Kolmogorov-Smirnov test is obtained as $ks=0.0693$ with P-value=0.882, which indicates that the transformed data follow a beta distribution with $\alpha=6.72$ and $\beta=3674.035$. For more on Beta regression modelling, see Ferrari and Cribari-Neto (2004) and Cribari-Neto and Zeileis (2010) among others. The mean of the response variable y^* can be written as,

$$E(y^*) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} \quad (4)$$


```

Coefficients (mean model with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.151e+01 2.604e+00 -4.418 9.95e-06 ***
X1           6.090e-05 6.736e-05  0.904 0.36597
X2          -3.634e-06 3.959e-06 -0.918 0.35876
X3          -4.322e-07 2.961e-07 -1.459 0.14444
X4          -6.268e-01 5.193e-01 -1.207 0.22741
X5           1.096e+00 1.034e+00  1.060 0.28902
X6          -4.588e-01 5.153e-01 -0.890 0.37322
X7           9.891e-03 5.460e-03  1.811 0.07009 .
X8           4.709e-01 2.975e-01  1.583 0.11350
X9           2.037e-02 7.533e-03  2.704 0.00685 **
X10          5.293e-05 3.771e-05  1.404 0.16045
X11          4.564e-04 9.350e-03  0.049 0.96107
X12          -1.323e-02 1.909e-02 -0.693 0.48838
X13          -5.628e-05 8.254e-05 -0.682 0.49534

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)      8063      1402  5.752 8.82e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

```

The fitted full Beta regression model (Model 4) is
 $\hat{y}^* = -11.51 + 0.00006090x_1 - 0.000003634x_2 - 0.0000004322x_3 - 0.6268x_4 + 1.096x_5 - 0.04588x_6 + 0.009891x_7 + 0.4709x_8 + 0.02037x_9 + 0.00005293x_{10} + 0.0004564x_{11} - 0.01323x_{12} - 0.00005628x_{13}$

The observed vs fitted plot of the full model is presented in Figure 7, which indicates that the Beta regression model fit the data well.

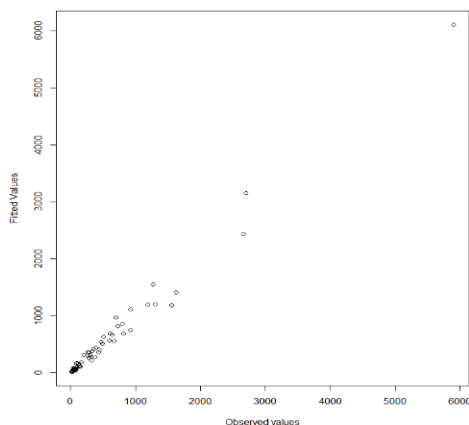


Fig. 7: Observed vs fitted plot for model 4

3.5 Reduced Beta Model (Model 5)

$$E(y^*) = \beta_0 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_7x_7 + \beta_8x_8 + \beta_9x_9 \quad (5)$$

```

Coefficients (mean model with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.183e+01 2.458e+00 -4.812 1.49e-06 ***
X1           1.073e-04 5.687e-05  1.886 0.05929 .
X3          -3.602e-07 2.517e-07 -1.431 0.15244
X4          -1.669e-01 6.104e-02 -2.734 0.00625 **
X5           1.713e-01 7.241e-02  2.366 0.01799 *
X7           7.902e-03 5.178e-03  1.526 0.12701
X8           4.659e-01 2.768e-01  1.683 0.09238 .
X9           2.534e-02 6.240e-03  4.061 4.89e-05 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)       7701      1339  5.75 8.9e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

```

The reduced fitted Beta regression model (Model 5) is

$$\hat{y}^* = -11.83 + 0.0001073x_1 - 0.0000003602x_3 - 0.1669x_4 + 0.1713x_5 + 0.007902x_7 + 0.4659x_8 + 0.02534x_9$$

The fitted vs observed plot is presented in Figure 8, which indicates that the model fit the data well.

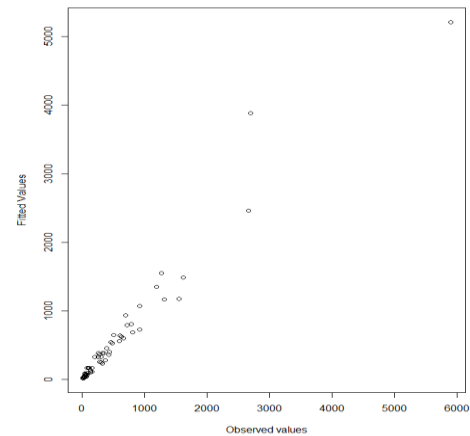


Fig. 8: Observed vs fitted plot for model 5

3.6 Full Ridge Regression Model (Model 6)

In linear regression model (3.1), we assumed that the regressors in the model are independent. However, the following correlation matrix (Table 2) exhibit linear relationship among the regressors, which causes the multicollinearity problem. About the correlation analysis, readers are referred to Chutiman et al. (2019) among others. The parameters in the linear regression model (3.1) are usually estimated using the ordinary least squares (OLS) estimator. In the presence of multicollinearity, least square estimators possess high standard error, and occasionally, the estimated regression coefficients exhibit wrong signs. As such, it becomes impossible to make valid statistical inference about the regression parameters (Kibria 2003). Hoerl and Kennard (1970) developed the ridge regression estimator as a notable alternative to the OLS estimator in the linear regression model (LRM).

Table 2. Correlation Matrix

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
X1	1.00	0.94	0.94	0.55	0.52	0.45	0.33	-0.38	-0.17	0.95	-0.30	0.39	0.84
X2	0.94	1.00	0.95	0.48	0.45	0.39	0.30	-0.32	-0.22	0.98	-0.20	0.32	0.87
X3	0.94	0.95	1.00	0.56	0.53	0.45	0.24	-0.43	-0.24	0.94	-0.31	0.39	0.84
X4	0.55	0.48	0.56	1.00	0.98	0.89	0.00	-0.69	0.04	0.53	-0.59	0.60	0.42
X5	0.52	0.45	0.53	0.98	1.00	0.96	-0.05	-0.65	0.06	0.50	-0.59	0.62	0.38
X6	0.45	0.39	0.45	0.89	0.96	1.00	-0.13	-0.55	0.09	0.43	-0.57	0.61	0.29
X7	0.33	0.30	0.24	0.00	-0.05	-0.13	1.00	-0.02	-0.12	0.27	0.15	-0.13	0.34
X8	-0.38	-0.32	-0.43	-0.69	-0.65	-0.55	-0.02	1.00	0.01	-0.37	0.52	-0.41	-0.40
X9	-0.17	-0.22	-0.24	0.04	0.06	0.09	-0.12	0.01	1.00	-0.12	0.00	0.04	-0.20
X10	0.95	0.98	0.94	0.53	0.50	0.43	0.27	-0.37	-0.12	1.00	-0.26	0.38	0.89
X11	-0.30	-0.20	-0.31	-0.59	-0.59	-0.57	0.15	0.52	0.00	-0.26	1.00	-0.48	-0.20
X12	0.39	0.32	0.39	0.60	0.62	0.61	-0.13	-0.41	0.04	0.38	-0.48	1.00	0.26
X13	0.84	0.87	0.84	0.42	0.38	0.29	0.34	-0.40	-0.20	0.89	-0.20	0.26	1.00

Table 3. VIF for the predictors

Variable	VIF
X1	17.2
X2	58.5
X3	16.0
X4	3044.9
X5	7804.2
X6	1287.8
X7	1.6
X8	2.3
X9	1.7
X10	80.2
X11	1.9
X12	1.9
X13	6.3

From Table 3, we can see that 7 out of the 13 regression coefficients have a $VIF_j > 10$, which certainly exhibit the existing of multicollinearity in the data. There are many ways to estimate ridge regression parameter k exist in literature (see Kibria 2003, Kibria and Banik 2012). However, for our convenient, we consider the ridge trace to estimate the parameter k . The Ridge Trace (Hoerl and Kennard, 1970) plot in (Figure 9), shows $k=0.1$ is the biased estimate which stabilizes the regression coefficients the best.

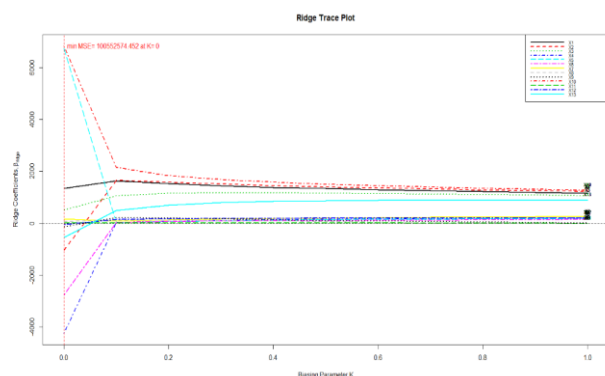


Fig. 9: Ridge trace plot for the coefficients of the predictors

By adjusting to $k=0.1$ the VIF for the Full Model is corrected and none of the VIF values are above 10. The results from (Model 6) show that R square is 0.9375, which represents about 94% of total variation of deaths due to COVID-19 could be explained by the 13 variables. The p-value in (Model 6) is approximately 0. From the lmrige calculation the ridge parameter $k=0.1$ improves the model the most. We can see from the following output that all of the coefficients have positive signs, while some of the coefficients in models (3.1) to (3.5) have wrong negative signs.

```

Coefficients: for Ridge parameter K= 0.1
              Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept -1.2975e+03 -4.0039e+08 5.2404e+07 -7.6404 <2e-16 ***
X1 9.4600e-02 1.6321e+03 1.6406e+02 9.9479 <2e-16 ***
X2 3.2000e-03 1.6471e+03 1.3130e+02 12.5440 <2e-16 ***
X3 3.0000e-04 1.0516e+03 1.6297e+02 6.4527 <2e-16 ***
X4 1.0682e+00 2.9598e+01 1.3814e+02 0.2143 0.8311
X5 8.1650e-01 1.8216e+01 6.4913e+01 0.2806 0.7800
X6 1.3786e+00 2.5105e+01 1.4982e+02 0.1676 0.8675
X7 1.4635e+00 8.7167e+01 1.1855e+02 0.7353 0.4652
X8 7.7285e+01 1.0609e+02 1.3651e+02 0.7771 0.4403
X9 5.0143e+00 2.2187e+02 1.1279e+02 1.9672 0.0540 .
X10 3.3600e-02 2.1609e+03 1.1724e+02 18.4314 <2e-16 ***
X11 3.2350e-01 1.2533e+01 1.3254e+02 0.0946 0.9250
X12 7.7767e+00 1.4601e+02 1.3332e+02 1.0952 0.2780
X13 6.1500e-02 4.9192e+02 1.6541e+02 2.9739 0.0043 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ridge Summary
      R2    adj-R2    DF ridge      F      AIC      BIC
0.93750 0.92360 8.08271 263.05913 647.71361 947.24790
Ridge minimum MSE= 102194027 at K= 0.1
P-value for F-test ( 8.08271 , 57.2754 ) = 2.537327e-42

```

The fitted Full Ridge Regression Model (Model 6) is

$$\hat{y} = -1297.5 + 0.0946x_1 + 0.0032x_2 + 0.0003x_3 + 1.0682x_4 + 0.8165x_5 + 1.3786x_6 + 1.4635x_7 + 77.285x_8 + 5.0143x_9 + 0.0336x_{10} + 0.3235x_{11} + 7.7767x_{12} + 0.0615x_{13}$$

4 Cross Validation

Since we have several well fitted models for the same data, a cross validation is needed to find the possible best model. To decide which model we will give the best fit, we will use five different randomly selected variations of the 67 observation. We will fit the models with 80% (50 counties) of each variation and then predict the other 20% (17 counties) to evaluate the adequacy of the model. Since the following models, namely, the Log Transformation (Model 2), Reduced Log Transformation (Model 3), Beta Transformation (Model 4), Reduced Beta Transformation (Model 5), Full Ridge Model (Model 6) met the model assumption, we consider them only for the cross validation. To evaluate the models, following Bai et al. (2019), we will calculate the following metrics, namely, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Range-normalized RMSE (NRMSE), and Mean Absolute Percentage Error (MAPE) of each variation and average them. The estimated value of the metrics is presented in Table 4.1. From Table 4.1, we can see that the MAE, RMSE, and NRMSE are the lowest for the full ridge regression model. The model with the best MAPE value is the reduced beta transformation. Therefore, we consider the full ridge regression model as the possible best model.

The final fitted model is

$$\hat{y} = -1297.5 + 0.0946x_1 + 0.0032x_2 + 0.0003x_3 + 1.0682x_4 + 0.8165x_5 + 1.3786x_6 + 1.4635x_7 + 77.285x_8 + 5.0143x_9 + 0.0336x_{10} + 0.3235x_{11} + 7.7767x_{12} + 0.0615x_{13} \quad (6)$$

Table 4. Evaluation metrics for different models

Model	MAE	RMSE	NRMSE	MAPE
Log Transformation	289.8	585.5	0.3469	1.07
Reduced Log Transformation	234.47	413.63	0.2411	0.989
Beta regression	118.62	247.995	0.1169	0.308
Reduced Beta regression	112.997	228.31	0.1105	0.304
Full Ridge	67.0491	114.0379	0.05417	0.3662

As an application of the final fitted model (4.1), we predicted the total number of deaths due to COVID-19 for the first five Florida counties and provided them in Table 5.

Table 5. Observed and predicted values for some selected counties

County	Observed	LT	RLT	BR	RBR	FR
Alachua	259	289.2	319.6	299.5	322.6	264.8
Baker	61	70.6	68.2	66.2	66.5	41.9
Bay	375	221.8	231.6	271.3	277.8	345.2
Bradford	55	47.5	60.7	58.5	64.7	82.6
Bevard	814	559.8	722.5	690.7	687.6	757.0

From Table 5, we can see that the full ridge model predicted the observed number of deaths better than the rest of the fitted models.

5 Conclusion

The paper considers some regression models to fit the number of deaths due to COVID-19 for Florida counties. Thirteen (13) regressors, namely, hospitalizations, total cases, population, minimum temperature, average temperature, max temperature, precipitation, mobility index, median age, adults age 65 or more, PM2.5 average, ozone average, and positive non-residents were used to fit the models. Since VIFs indicated the problem of multicollinearity, we used ridge regression method to address the multicollinearity problem. Cross validation results indicated that the full ridge regression model performed the best among five fitted models. Of the 13 regressors, the following six variables: hospitalizations per county (x_1), total cases per county (x_2), population per county (x_3), median age per county (x_9), positive adults 65 or older per county (x_{10}), and positive non-residents per county (x_{13}) were significant for the final model.

For our convenient, we have estimated the ridge parameter k from ridge trace (Hoerl and Kennard 1970). However, since there are many methods exist to estimate the ridge parameter k , as a future research, this data can be applied for estimation of k for different models. Besides ridge regression approach, Liu (1993) and Kibria-Lukman (Kibria and Lukman 2020) estimators and Stein type (1956) estimator among other methods can be used to solve the multicollinearity problem. We strongly believe that the data of this paper will be an excellent source for illustrating multicollinear linear regression models.

Acknowledgements:

The authors are thankful to the editor and reviewers for their valuable comments/suggestions, which certainly improved the quality and presentation of the article.

References:

- [1] Abdullah, W. G., Rianse, U., Iswandi, R. M., Taridala, S. A. A., Yunus, L., Fausayana, I., Dirgantoro, M. A. Rianese, I. S. and Batoa, H. (2016) Factors Affecting the Preferences of Agribusiness Actors of Aren Sugar toward Marketing Risk, WSEAS Transactions on Business and Economics. 13, 408-413.
- [2] Amin, M., Akram, M.M. and Kibria, B. M. G. (2021). A new adjusted Liu estimator for the Poisson regression model. Concurrency Computer Pract Exper. DOI: 10.1002/cpe.6340
- [3] Bai, X. and Zhang, F. and Lee, I. (2010). Predicting the citations of scholarly paper. Journal of Informetrics, 13, 407-418.
- [4] Climate at a Glance. National Oceanic and Atmospheric Administration (NOAA).
- [5] <https://www.ncdc.noaa.gov/cag/county/mapping/8/pcp/202101/18/value>
- [6] Chutiman, N., Busabarodhin, P., Senapeng, T. P. and Chitangpradit, M. (2019). Correlation analysis between extreme weather event and agricultural damage value in Northeast of Thailand. WSEAS Transactions on Environmental and development. 15, 257-260.

- [7] Chen, D-G and Chen, J. K. (2021). Statistical regression analysis modelling with R. Springer, Cham, ISBN: 978-3-030-67585-1.
- [8] Cohen, Y. and Cohen, J. Y. (2008). Statistics and Data with R: An applied approach through examples, Wiley, New York.
- [9] Covid-19 Data Tracker. Center for Disease Control and Prevention. <https://www.cdc.gov/coronavirus/2019-ncov/index.html>
- [10] Cribari-Neto F, Zeileis A (2010). Beta Regression in R. Journal of Statistical Software, 34(2), 1–24. URL <http://www.jstatsoft.org/v34/i02/>.
- [11] Draper, N. R. and Smith, H.. (1998). Applied Regression Analysis, Third Edition. Wiley, New York 1998.
- [12] Ferrari, S. I. O and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics. 31(7), 799-815.
- [13] Florida coronavirus cases and deaths. <https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/florida>
- [14] Florida's COVID-19 Data and Surveillance Dashboard. Site Map. <https://experience.arcgis.com/experience/96dd742462124fa0b38ddedb9b25e429>
- [15] Gibbons, D.G. (1981). A simulation study of some ridge estimators. Journal of the American Statistical Association, vol. 76, pp.131-139.
- [16] Gruber, M. H. J. (2010). Regression estimators, Second Edition, University Press, Baltimore.
- [17] Ghosal, S., Sengupta, S., Majumder, M. and Sinha, B. (2020). Linear regression analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from 0 (100 cases-Marcg 14th 2020). Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 12, 311-315.
- [18] Guzman, C. I and Kibria, B. M. G. (2019). Developing Multiple Linear Regression Models for the Number of Citations: A Case Study of Florida International University Professors. International Journal of Statistics and Reliability Engineering, 6(2), 75-81.
- [19] Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55-67.
- [20] Kibria, B.M.G. (2003). Performance of some new ridge regression estimators. Communication in Statistics-Simulation and Computation, vol.32, pp.419-435.
- [21] Kibria, B. M. G. and S. Banik (2016). Some Ridge Regression Estimators and Their Performances. Journal of Modern Applied Statistical Methods. 15 (1), 206-238.
- [22] Kibria, B. M. G. and Lukman, A. F. (2020). A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications. Scientifica, Volume 2020, Article ID 9758378, 16 pages.
- [23] Liu, K.J., 1993. A new class of biased estimate in linear regression. Commun. Statist. Theory Meth. 22 (2), 393–402.
- [24] Lukman, A. F. K. Ayinde, S. Binuomote, and O. A. Clement, (2019), “Modified ridge-type estimator to combat multicollinearity: application to chemical data,” Journal of Chemometrics, vol. 33, no. 5, p. e3125, 2019.
- [25] Lukman, A. F., K. Ayinde, S. K. Sek, and E. Adewuyi, (2019b), “A modified new two-parameter estimator in a linear regression model,” Modelling and Simulation in Engineering, vol. 2019, Article ID 6342702, 10 pages, 2019.
- [26] Montgomery, D. C., Peck, E. A. and Vining, G. G. (2013). Introduction to Linear Regression Analysis. Wiley-Blackwell, 2013.
- [27] Muniz, G. and Kibria, B.M.G. (2009). On some ridge regression estimators: An empirical comparison. Communication in Statistics-Simulation and Computation, vol.38, pp.621-630.
- [28] Muniz, G., Kibria, B.M.G. Mansson, K and Shukur, G. (2012). On developing ridge regression parameters: A graphical investigation. SORT, vol.36, pp.115-138.
- [29] Ogundokun, R. O., Lukman, A. F., Kibria, B. M. G., Awotunde, J. B. and Aladetian, B. B. (2020). Predictive modeling of COVID-10 confirmed cases in Nigeria. Infectious Disease Modelling. 5, 543-548.
- [30] Qasim, M., Mansson, K. and Kibria, B. M. G. (2021). On some beta ridge regression estimators: methods, simulation and application. Journal of Statistical Computation and Simulation. 91(9), 1699-1712.

- [31] Qin, L., Sun, Q., Wang, Y., Wu, K-F, Chen M. Shia, B-C, and Wu, S-Y (2020). Predicting of number of cases of 2019 Novel Coronavirus (COVID-19) using social media search index. *International Journal of Environmental Research and Public Health*. 17, 2365.
- [32] 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*. 2020, 3:008
- [33] Saleh, A. K. Md. E., Arashi, M., and Kibria, B. M. G. (2019). *Theory of Ridge Regression Estimation with Applications*. Wiley, New York.
- [34] Suhail, M, Chand, S and Kibria, B. M. G. (2020). Quantile-based robust ridge m-estimator for linear regression model in presence of multicollinearity and outliers. *Communications in Statistics-Simulation and Computation*. 50 (11), 3194-3206.
- [35] Stein, C. (1956), "Inadmissibility of the usual estimator for the mean of a multivariate distribution", *Proc. Third Berkeley Symp. Math. Statist. Prob.*, vol. 1, pp. 197–206, MR 0084922, Zbl 0073.35602

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

B M Golam Kibria designed the statistical methods. Manuel Urbistondo collected and edited the data and analyzed them. Both authors discussed the results and contributed to the final manuscript. B M Goalm Kibria revised the whole paper and typed in the journal format.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US

Appendix: Covid data for the state of Florida, USA

Counties	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
Alachua County	259	1132	23559	269043	60.6	71.6	82.7	53.89	3.9	27	2407	16.34	58.5	279
Baker County	61	177	3371	29210	58.1	69.6	81.1	56.06	4.3	42	488	23.66	54.5	15
Bay County	375	800	19462	174705	61.2	70.7	80.2	64.84	4	43	3136	28.04	57.3	765
Bradford County	55	146	2925	28201	59.8	70.9	82.1	55.62	4.3	45	548	16	58	13
Brevard County	814	2176	37563	601942	66.4	74.4	82.4	53.35	4	43	6835	17.06	57.8	674
Broward County	2699	9502	217060	1952778	68.4	77.4	86.3	78.94	4	38	23192	19.26	59.5	2604
Calhoun County	42	57	1662	14105	59.3	70.1	80.8	64.57	4.3	45	392	28	57	3
Charlotte County	392	861	11714	188910	65.7	75.8	85.9	56.9	4	57	4090	19	61	203
Citrus County	432	675	10389	149657	63	73.5	83.9	49.54	4.1	55	3515	23.45	60	91
Clay County	329	901	17447	219252	61.8	72.1	82.4	54.32	4.1	43	2859	16.5	58	48
Collier County	505	1611	32836	384902	66.7	76.5	86.4	56.87	4	44	6534	21.27	58.3	1254
Columbia County	163	456	7877	71686	57.7	69.3	80.8	55.84	4.2	43	1389	32.23	50.4	37
DeSoto County	86	288	4101	38001	64.3	74.9	85.6	55.24	4	42	643	21	58	26
Dixie County	22	112	1556	16826	60.6	71	81.4	53.4	4.3	44	317	17	56.3	3
Duval County	1311	2046	92993	957755	61.9	71.6	81.3	54.82	4.1	38	11952	25.31	57	1530
Escambia County	669	2103	34950	318316	59	69.1	79.1	69.05	4	41	6513	20.01	55.7	3636
Flagler County	99	365	6751	115081	64	73.2	82.4	51.69	4	46	1364	17	54	94
Franklin County	17	29	1296	12125	61.1	70.7	80.3	58.31	4.2	44	246	24.6	55.3	33
Gadsden County	90	417	5544	45660	58.7	69.4	80.1	58.46	4.4	41	710	28.43	56	16
Gilchrist County	38	75	1495	18582	59.6	70.9	82.1	52.47	4.4	40	271	19.76	57.8	4
Glades County	19	69	926	13811	65.1	75.4	85.6	53.56	4.3	38	119	16	58	4
Gulf County	42	96	1920	13639	61	70.8	80.5	61	4.2	50	436	27	56	23
Hamilton County	23	80	1593	14428	57.5	68.7	79.8	55.25	4.4	41	232	27.06	53	19
Hardee County	39	218	2915	26937	63.9	74.7	85.4	51.65	4.2	37	388	20.39	59.68	14
Hendry County	77	335	4479	42022	65.7	75.9	86.1	49.2	4.2	38	576	25	59	18
Hernando County	443	977	12543	193920	63.8	74.1	84.3	49.12	4.1	50	3265	24.28	58.73	96
Highlands County	323	610	7732	106221	64.6	74.9	85.2	54.4	4	52	2433	18.5	58.3	88
Hillsborough County	1619	2971	123169	1471968	65.6	75.1	84.6	53.32	4	37	14565	21.32	63	1059
Holmes County	47	89	2216	19617	57.5	68.9	80.3	71.82	4.4	44	414	28.12	53.9	32
Indian River County	282	813	11790	159923	65.6	74.6	83.6	64.54	4	45	2753	17	58	104
Jackson County	155	352	6068	46414	58.6	69.4	80.2	65.94	4.3	44	1155	28	54	65
Jefferson County	25	76	1432	14246	59.6	69.9	80.2	54.51	4.4	47	279	32	56	22
Lafayette County	25	48	1606	8422	59	70.1	81.2	54.14	4.4	39	186	18.38	57.05	0
Lake County	610	1386	26924	367118	64	74.2	84.3	48.46	4.1	45	6098	21	60.1	460
Lee County	923	2143	63334	770577	66.9	76.3	85.8	55.91	4	43	12389	18.53	61.2	2180
Leon County	306	657	30406	293582	59.2	69.8	80.4	56.94	3.9	28	2499	32.44	55.5	895
Levy County	44	207	3139	41503	62	72.4	82.8	49.22	4.3	43	578	20.26	59.53	8
Liberty County	16	22	1056	8354	59.3	70.1	80.9	61.36	4.4	39	116	27.5	57.5	1
Madison County	44	88	2034	18493	58.8	69.5	80.1	53.47	4.4	43	371	29.53	55	26
Manatee County	635	1434	35161	403253	64.8	74.7	84.6	54.51	4	42	6410	18.59	54.7	820
Marion County	926	1971	29225	365579	62.3	72.8	83.3	51.79	4.1	45	6420	20.48	60	147
Martin County	309	745	11465	161000	68	76.7	85.3	73.33	3.9	45	2509	19.47	56.6	214
Miami-Dade County	5813	12624	449891	2716940	69.5	78	86.6	70.64	4	40	55621	20.49	60.4	5773
Monroe County	49	253	6504	74228	70.4	78.5	86.5	62.81	3.8	42	899	21	59	373
Nassau County	120	296	7996	88625	60.2	70.6	81	54.13	4.2	45	1432	26.21	55.75	436
Okaloosa County	353	692	19799	210738	58	68.8	79.5	68.63	4	40	3019	21	55.8	211
Okeechobee County	86	387	3768	42168	65.4	75.1	84.8	61.75	4	41	655	19	58	52
Orange County	1190	2606	125108	1393452	65.1	74.7	84.3	50.77	4	36	11750	17.55	54.2	1898
Osceola County	487	1364	40024	375751	65	74.8	84.5	52.06	4.1	38	4363	20.76	60.1	704
Palm Beach County	2661	6427	132961	1496770	68	76.9	85.9	71.65	4	40	19434	15.28	57.6	2417
Pasco County	697	2420	36787	553947	64.4	74.4	84.4	52.01	4	44	7112	16	57.5	322
Pinellas County	1557	4581	73080	974996	67.6	75.5	83.3	51.42	3.9	43	13443	16.14	62.1	1813
Polk County	1267	4820	61747	724777	64.6	74.9	85.1	54.14	4.1	41	10182	19.13	60.5	453
Putnam County	135	528	6060	74521	63.2	73	82.7	52.01	4.2	42	1128	18.17	57	21
St. Johns County	203	742	21294	264672	64.1	72.9	81.8	52.48	4.1	41	3008	22.5	55.3	298
St. Lucie County	594	1657	24197	328297	66.6	75.5	84.5	72.69	4.1	43	4388	19	57	176
Santa Rosa County	278	858	17805	184313	58.1	68.4	78.7	69.28	4.2	42	2706	20.64	58.6	231
Sarasota County	791	1298	29679	433742	64.6	74.7	84.8	54.96	4	49	7716	20.5	57.7	1478
Seminole County	468	1199	30410	471826	65	74.5	83.9	50.96	3.9	38	3797	16.97	58.9	236
Sumter County	261	558	8852	132420	62.9	73.7	84.6	47.47	3.9	65	4394	19.3	58	103
Suwannee County	135	311	5224	44417	58.2	69.5	80.7	54.76	4.3	44	957	26	54.73	28
Taylor County	45	96	2695	21569	60.7	70.7	80.8	53.67	4	41	394	19	56	5
Union County	71	82	1776	15237	58.4	70	81.6	55.98	4.4	45	282	24	54	1
Volusia County	726	1980	38819	553284	64.4	73.7	83	52.33	4	44	7382	16.87	57.5	689
Wakulla County	55	157	3256	33739	60.2	70.4	80.6	58.09	4.4	43	504	32.86	54.4	12
Walton County	83	290	7360	74071	58.2	69.2	80.1	69.52	4.2	42	1139	28	57	788
Washington County	50	135	2678	25473	58.7	69.6	80.5	68.2	4.4	44	461	29.4	56.5	20