ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

Advances in Decision Sciences

Volume 25
Issue 2
June 2021

Michael McAleer

Editor-in-Chief
University Chair Professor
Asia University, Taiwan

Published by Asia University, Taiwan

ADS@ASIAUNIVERSITY

Predicting COVID-19 Cases and Deaths in the USA from Tests and State Populations*

David E. Allen[a], and Michael McAleer[b,*]

[a] *School of Mathematics and Statistics, University of Sydney, Department of Finance, Asia University, Taiwan, and School of Business and Law, Edith Cowan University, Australia*
[b] *Department of Finance, College of Management, Asia University, Taiwan, Discipline of Business Analytics, University of Sydney Business School, Australia, Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands, Department of Economic Analysis and ICAE, Complutense University of Madrid, Spain, Department of Mathematics and Statistics, University of Canterbury, New Zealand.*

Keywords: Risk management, Curve projection, Live data, Global pandemic, COVID 19, Lockdown, CFR.
JEL: C22, C53, C88.

Acknowledgements: The authors are most grateful for very helpful comments and suggestions from two reviewers. For financial support, the first author acknowledges the Australian Research Council, and the second author is most grateful to the Australian Research Council, Ministry of Science and Technology (MOST), Taiwan, and the Japan Society for the Promotion of Science.

*Corresponding author
Email address: michael.mcaleer@gmail.com

Abstract

The paper presents a novel analysis of the US spread of the SARS-CoV-2 causes the COVID-19 disease across 50 States and 2 Territories. Simple cross-sectional regressions are able to predict quite accurately both the total number of cases and deaths, which cast doubt on measures aimed at controlling the disease via lockdowns. Population density appears to play a significant role in transmission. This throws in sharp relief the relative effectiveness of the attempts to risk manage the spread of the virus by 'flattening the curve' (aka planking the curve) of the speed of transmission, and the efficacy of lockdowns in terms of the spread of the disease and death rates. The algorithmic techniques, results and analysis presented in the paper should prove useful to the medical and health professions, science advisers, and risk management and decision making of healthcare by state, regional and national governments in all countries.

1. Introduction

The outbreak of the SARS-CoV-2 virus that causes the COVID-19 disease was first detected in Wuhan, the capital city of Hubei Province, China, and reported to the World Health Organization (WHO) office in Wuhan on 31 December 2019. The WHO declared a "Public Health Emergency of International Concern" on 30 January 2020, and gave the name COVID-19 to the novel coronavirus disease on 11 February 2020. The virus has spread to all continents, except Antartica, and has dominated the daily news as governments struggle to contain the spread of the virus. In China, the outbreak effectively confined well over one billion people to their apartments and homes since the end of January 2020, and continues to disrupt healthcare, well-being, and the economy, while much of the rest of the world has rapidly followed suit.

Gostin, Hodge Jr., and Wiley (2020) analyzed Presidential powers and the response to COVID-19, which led to the challenging Comment that "With Great Power Comes Great Responsibility". The authors suggest a comprehensive balance is required between individual rights and liberty, and public health concerns, with self-isolation, quarantining, social distancing, and international travel restrictions being essential to curb the spread of the disease.

This raises the issues relating to the importance of either the banning, or restrictions on the size, of gatherings of more than 2 persons in public. The duration of any time frame of government lockdowns beyond 2 weeks, and the imposition of domestic travel restrictions. In the US this involves the US President using the 'soft' powers associated with federal leadership in working with the pro-active State Governors.

McAleer (2020) refers to the GHS Index (see Chang and McAleer (2020)), which is a comprehensive assessment of global health security capabilities in 195 countries. The GHS Index suggests that international preparedness for epidemics and pandemics is weak.

Atalan (2020) analyzed the effect of lockdown days on the spread of coronavirus in 49 countries. COVID-19 cases and lockdown days data were collected for countries that implemented the lockdown between certain dates (without interruption). The analysis was undertaken on 5 May 2020 and the results suggested that there was a significant negative relationship between lockdowns and the spread of the virus.

Mandel and Veetil (2020) developed a multi-sector disequilibrium model with buyer-seller relations between agents located in different countries to assess the

economic costs of lockdowns. Their estimate of the total impact amounts to 9% of global GDP. Guan et al. (2020) modelled four different sets of pandemic scenarios, three of which (36 scenarios in total) represent different spread extents and containment responses to the COVID-19 pandemic. They reported that economic effects (losses) are relatively less sensitive to the strictness of lockdown measures compared to the extent of pandemic or duration of the lockdown.

In this paper we use cross-sectional regression analysis and quantile regression to analyse the relationship in US States and Territories among population size and population density, and the incidence of the number cases of and related deaths from COVID-19. We do not analyse the efficacy of lockdowns per se, or the economic impacts of COVID-19. We udertake the analysis at two points in time; 28 September 2020 and 21 March 202, thus are sample is roughly at a six month interval.

The paper is divided into four sections: the introduction is followed in section 2 by a description of the the sample and methods of analysis. The striking results are presented in section 3, and a brief conclusion follows in section 4.

2. Sample and research method

The paper examines prediction of the number of cases of COVID-19 and the number of related deaths using cross-sectional regressions and quantile regression applied to US State level data sourced from the Johns Hopkins University website on 28 September 2020 and six months later on 21 March 2021. The sample totals 52 States and Territories, given that Washington DC and Puerto Rico are included in the analysis.

The population density figures are reported as population per square mile, and population density is defined as the population per (divided by) land area. Resident population is from the United States Census Bureau estimates for July 1, 2015, (for the 50 states, Washington DC and Puerto Rico, as sourced from Wikipedia, see: https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population_density#2015_density_(states,_territories_and_DC)), accessed on 28 September 2020.

Apart from the application of Ordinary Least Squares regression (OLS), we also use quantile regression which has the advantage of providing an analysis of the relationships between variables across the quantiles of their respective distributions.

Koenker and Hallock (2001, p.145) provide an introduction to quantile regression, and note that quantiles seem inseparably linked to the operations of ordering and sorting the sample observations used to define them: 'The symmetry of the piecewise linear absolute value function implies that the minimization of the sum of absolute residuals must equate the number of positive and negative residuals, thus assuring that there are the same number of observations above and below the median'.

What about the other quantiles? As the symmetry of the absolute value yields the median, it follows that minimizing the sum of asymmetrically weighted absolute residuals by simply giving differing weights to positive and negative residuals provides the other quantiles. The solution to:

$$\xi \in RMin \sum \rho_\tau(y_i - \xi), \tag{1}$$

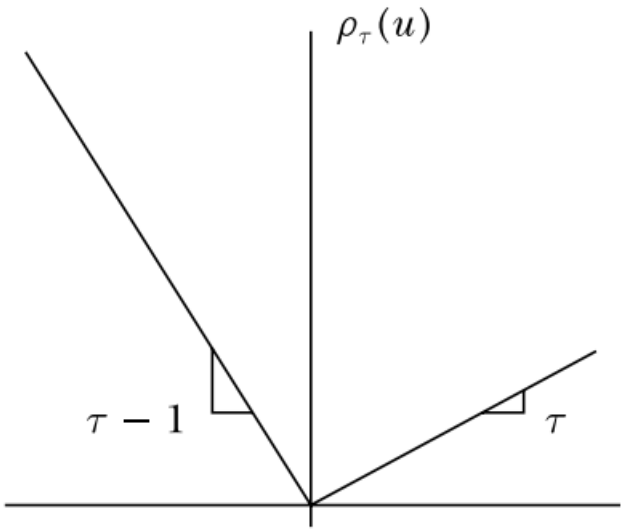
where the function $\rho_\tau(\cdot)$ is the titled absolute value function, as shown in Figure (1), gives the τth sample quantile function.

An estimate of the conditional median function can be obtained by replacing the scalar, ξ, in equation (2), by the parametric function $\xi(x_i, \beta)$, and setting τ to 1/2. Estimates of the other conditional quantile functions can be obtained by replacing absolute values by $\rho_\tau(\cdot)$, and solving expression (11) by linear programming:

$$\beta \in Rpmin \sum \rho_\tau(y_i - \xi(x_i, \beta). \tag{2}$$

We use quantile regression to analyse the relationships among the variables across the five quantiles of their distributions to assess whether there is any change in their relationships in the tails of their respective distributions.

Figure 1: Quantile regression ρ function



3. Results of the analysis

The analysis concentrates on the number of cases, tests, deaths, population, and population density at the state level. Summary statistics for these five variables are shown in Table 1. It is apparent that there is a large variation across different states in the values of these five variables. The State with the most cases on 28 September 2020 was California at 809,890, and this remained so on 21 March 2021 when it had 3,637,700 cases. Vermont had the lowest number of cases at 1,742, on 28 September 2020, and still had the lowest number on 21 March 2021 with 17,393 cases, when these figures were accessed on the Johns Hopkins website. New York conducted the most tests on 28 September 2020, with a total of 10,508,000, and had the most deaths at 33,131 on that date. On 21 March 2021, California had the maximum number of deaths of 57,350, and also had conducted the most tests with a figure of 51,812,000. California had the largest population at 39.51 million, but the densest population was recorded by the District of Columbia, closely followed by New Jersey and Puerto Rico.

3.1. Regression analysis of cases

We regressed the number of cases, c_i, for each state i, on the total population of that state, p_i, as shown in equation (3):

$$c_i = a + bp_i + e_i. \tag{3}$$

This approach ignores the time dimension to the spread of the data, apart from the fact that the regression is undertake using data taken from two different points in time, namely; 28 September 2020 and 21 March 2021. The results are shown in Table 2.

The results of the regression, as shown in Table 2, are statistically significant. There is a significant positive relationship between the size of the population and the number of cases in both time periods. On 28 September 2020 the adjusted R squared had a high value of 88 percent, whilst on 21 March 2021 the adjusted R squared had increased to 96 percent. This suggests that 88 and 96 percent of the variations in the number of cases at the two different points in time respectively, can be accounted for by population alone.

A plot of the actual cases versus the fitted cases from the regression is shown in Figure 2.

Table 1: Summary statistics

Variable	Mean	Median	Minimum	Maximum	Standard Deviation
As on 28 September 2020					
Cases	136750	85301	1742.0	809890	178730
Tests	1671100	1135100	15606	10508000	1.887400
Deaths	3936.4	1813.0	50.000	33131	5920
As on 21 March 2021					
Cases	571520	389340	17393	3637700	679700
Tests	7300900	4062200	1057	51812000	9.6817000
Deaths	10403	5880	217	57350	12625
Constant in both periods					
Population (Millions)	6.2413	4.0875	0.57876	39.510	7.2859
Population Density	424.25	106.50	1.0000	11011	1524

Table 2: OLS regression of cases on population

28 September 2020

OLS, using observations 1–52
Dependent variable: cases

	Coefficient	Std. Error	t-ratio	p-value
const	−7035.13	11369.1	−0.6188	0.5389
popmil	23038.3	1191.68	19.33	0.0000

Mean dependent var	136753.5	S.D. dependent var	178731.1
Sum squared resid	1.92e+11	S.E. of regression	62005.83
R^2	0.882005	Adjusted R^2	0.879645
$F(1, 50)$	373.7460	P-value(F)	7.47e–25
Log-likelihood	−646.5842	Akaike criterion	1297.168
Schwarz criterion	1301.071	Hannan–Quinn	1298.665

21 March 2021

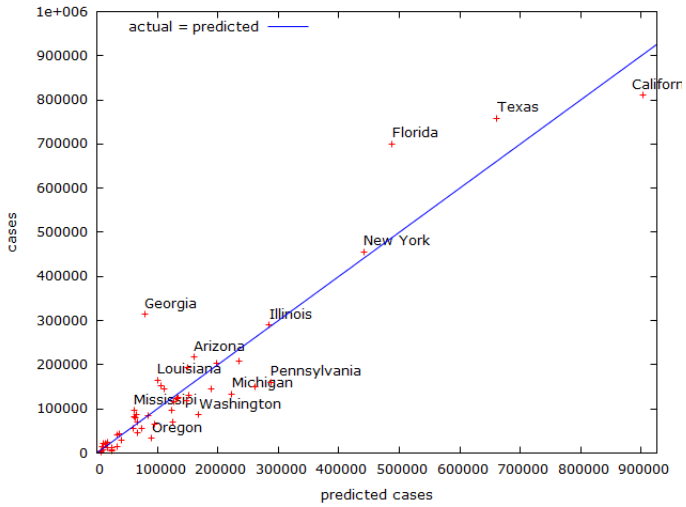
OLS, using observations 1–52
Dependent variable: cases

	Coefficient	Std. Error	t-ratio	p-value
const	1744.87	25906.8	0.06735	0.9466
popmil	91291.4	2715.50	33.62	0.0000

Mean dependent var	571522.0	S.D. dependent var	679698.2
Sum squared resid	9.98e+11	S.E. of regression	141293.0
R^2	0.957635	Adjusted R^2	0.956788
$F(1, 50)$	1130.215	P-value(F)	5.43e–36
Log-likelihood	−689.4118	Akaike criterion	1382.824
Schwarz criterion	1386.726	Hannan–Quinn	1384.320

Figure 1: **Fit of cases regressed on population**

28 September 2020



21 March 2021

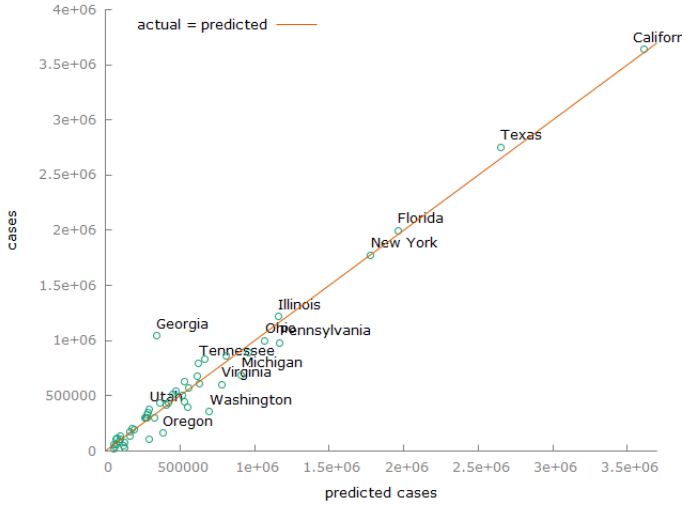


Table 3: Regression of cases on population and tests

28 September 2020

OLS, using observations 1–52
Dependent variable: cases

	Coefficient	Std. Error	t-ratio	p-value
const	−15917.5	11699.6	−1.361	0.1799
popmil	21327.1	1392.25	15.32	0.0000
tests	0.0117064	0.00537457	2.178	0.0342

Mean dependent var	136753.5	S.D. dependent var	178731.1
Sum squared resid	1.75e+11	S.E. of regression	59806.99
R^2	0.892421	Adjusted R^2	0.888030
$F(2, 49)$	203.2387	P-value(F)	1.89e–24
Log-likelihood	−644.1814	Akaike criterion	1294.363
Schwarz criterion	1300.217	Hannan–Quinn	1296.607

21 March 2021

OLS, using observations 1–52
Dependent variable: cases

	Coefficient	Std. Error	t-ratio	p-value
const	2793.50	25958.6	0.1076	0.9147
popmil	85736.4	6480.48	13.23	0.0000
tests	0.00460518	0.00487686	0.9443	0.3497

Mean dependent var	571522.0	S.D. dependent var	679698.2
Sum squared resid	9.80e+11	S.E. of regression	141446.3
R^2	0.958392	Adjusted R^2	0.956694
$F(2, 49)$	564.3291	P-value(F)	1.48e–34
Log-likelihood	−688.9429	Akaike criterion	1383.886
Schwarz criterion	1389.740	Hannan–Quinn	1386.130

The number of cases was also regressed on population density, pd_i by State, but there was no significant relationship. The number of tests per State, t_i was then added to the list of explanatory variables, as shown in equation (4):

$$c_i = a + bpd_i + t_i + e_i. \tag{4}$$

which produced the results shown in Table 3.

The results of the regression in Table 3 show that both population and tests have positive and statistically significant coefficients in the first period, but in the second period the number of tests undertaken becomes insigificant, and in both periods the adjusted R squared increases by less than one per cent.

3.2. Regression analysis of deaths

We then focused on the analysis of deaths, and explored whether we could predict the number of deaths using cross-sectional regression based on 52 States and Territories. We regressed the number of deaths in each State, d_i, on the number of cases, c_i, as shown in equation (5):

$$d_i = a_i + bc_i + e_{i.} \tag{5}$$

The results of the regression are shown in Table 4. The regression is statistically significant in both time periods, and has positive and significant coefficients on cases. The adjusted R squared has a value of 58 per cent on data sampled on 28 September 2020 but by 21 March 2021 the adjusted R squared had increased to 91 per cent.

The World Health Organisation notes that an important feature of a novel pathogen is the estimation of fatality rates, which helps to evaluate the severity of a disease, identify at-risk populations, and evaluate quality of healthcare (see https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19). One metric is the case fatality ratio (CFR), which estimates the proportion of deaths among identified confirmed cases. The coefficient on ci in equation (5) provides an estimate of CFR across the 52 States and Territories at the time of estimation. The value of this coefficient of 0.025, on 28 September 2020, suggests a CFR of 2.5 per cent. This is a high death rate, particularly when contrasted with the typical death rate of seasonal flu, which has variously been suggested to be a fraction of 1 per cent. However, when the regression equation was re-estimated using data taken on 21 March 2021, the CFR had

Table 4: OLS regression of deaths on cases

28 September 2020

OLS, using observations 1–52
Dependent variable: deaths

	Coefficient	Std. Error	t-ratio	p-value
const	453.522	669.891	0.6770	0.5015
cases	0.0254683	0.00299489	8.504	0.0000

Mean dependent var	3936.404	S.D. dependent var	5920.028
Sum squared resid	7.31e+08	S.E. of regression	3822.658
R^2	0.591226	Adjusted R^2	0.583050
$F(1, 50)$	72.31696	P-value(F)	2.79e–11
Log-likelihood	−501.6975	Akaike criterion	1007.395
Schwarz criterion	1011.298	Hannan–Quinn	1008.891

21 March 2021

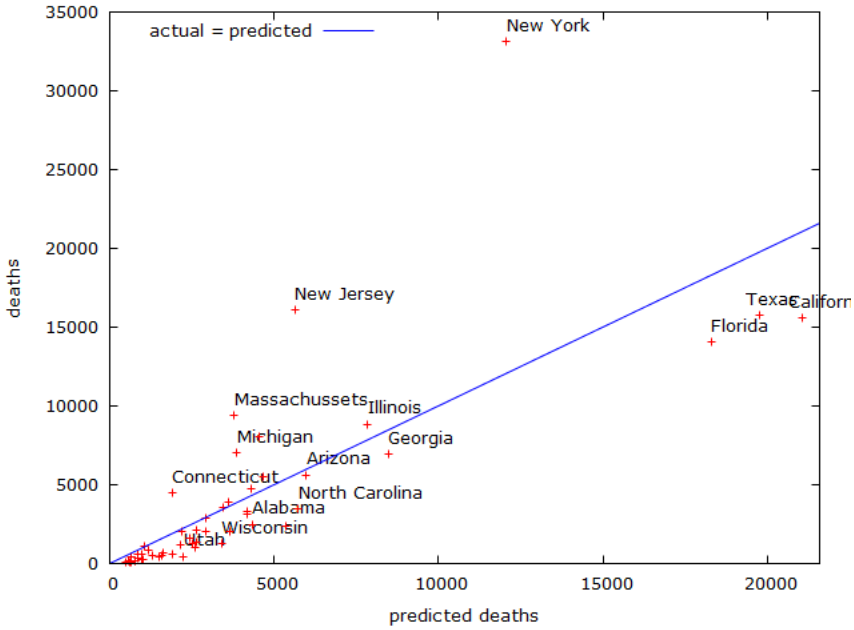
OLS, using observations 1–52
Dependent variable: deaths

	Coefficient	Std. Error	t-ratio	p-value
const	243.123	672.148	0.3617	0.7191
cases	0.0177777	0.000761183	23.36	0.0000

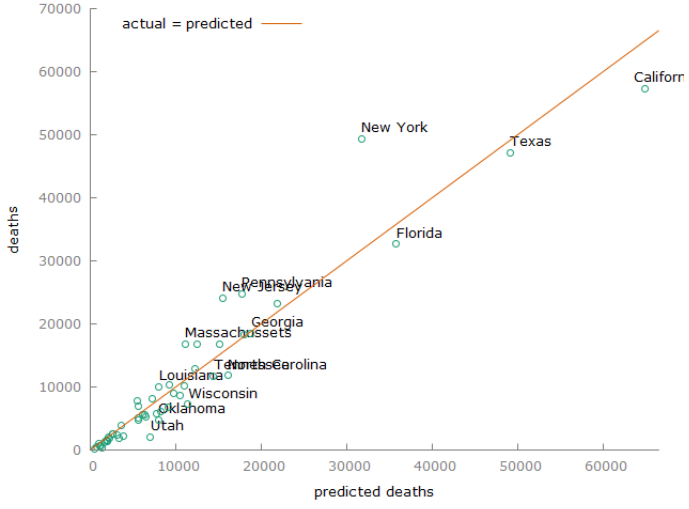
Mean dependent var	10403.44	S.D. dependent var	12625.11
Sum squared resid	6.83e+08	S.E. of regression	3694.797
R^2	0.916033	Adjusted R^2	0.914353
$F(1, 50)$	545.4691	P-value(F)	1.48e–28
Log-likelihood	−499.9285	Akaike criterion	1003.857
Schwarz criterion	1007.759	Hannan–Quinn	1005.353

Figure 2: **OLS regression of deaths on cases**

28 September 2020



21 March 2021



reduced to 1.8 per cent. This is still a high value but is likely to be more accurate, as far more tests had been undertaken by this date.

Two plots of the fitted deaths on actual deaths from the regression in equation (5) are shown in Figure 3. Its is clear that New York is a significant outlier, in both plots, which was probably not helped by Governor Cuomo's policy of sending COVID-19 afflicted patients back into nursing homes. In the second plot which reflects data from 21 March 2021 New York, New Jersey, Pennsylvania, Massachusetts and Georgia, remain above the line. This suggests they experienced more deaths per case than the average.

Ramsey's (1969) functional form RESET test suggested a non-linear specification. Applying a logarithmic transformation led to the model in equation (6):

$$ld_i = a + \beta lc_i + e_i, \tag{6}$$

where ld_i and lc_i are the logarithmic transformations of the number of deaths and cases per State, respectively.

The interpretation of logarithmic regressions is slightly different from standard regressions. The interpretation of the above relationship is given as an expected percentage change in d_i when c_i increases by one percent. Such relationships, where both d_i and c_i are log-transformed, are commonly referred to as elasticities in economics, and the coefficient of log c_i is referred to as an elasticity. In terms of the effects of changes in c_i on d_i :

- multiplying c_i by e will multiply expected value of d_i by e^{β}.
- to obtain the proportional change in d_i associated with a p percent increase in c_i, calculate a = log([100 + p]/100) and take $e^{a\beta}$.

The results of the log-transformed regression of deaths per State on the number of cases per State are shown in Table 5. This regression is statistically significant, and is an improvement on the previous untransformed regression. The adjusted R squared has improved from 59 per cent to 84 percent for the first period data set representing 28 September 2020. Both the coefficients are significant at the 1 percent significance level, whereas previously the coefficient on the constant term was not significant.

Table 5: Regression of log transformed d_i on log transformed c_i

28 September 2020

OLS, using observations 1–52
Dependent variable: l_deaths

	Coefficient	Std. Error	t-ratio	p-value
const	−4.78157	0.719680	−6.644	0.0000
l_cases	1.08707	0.0642908	16.91	0.0000

Mean dependent var	7.302752	S.D. dependent var	1.566246
Sum squared resid	18.62304	S.E. of regression	0.610296
R^2	0.851146	Adjusted R^2	0.848169
$F(1, 50)$	285.8998	P-value(F)	2.53e–22
Log-likelihood	−47.08685	Akaike criterion	98.17371
Schwarz criterion	102.0762	Hannan–Quinn	99.66983

21 March 2021

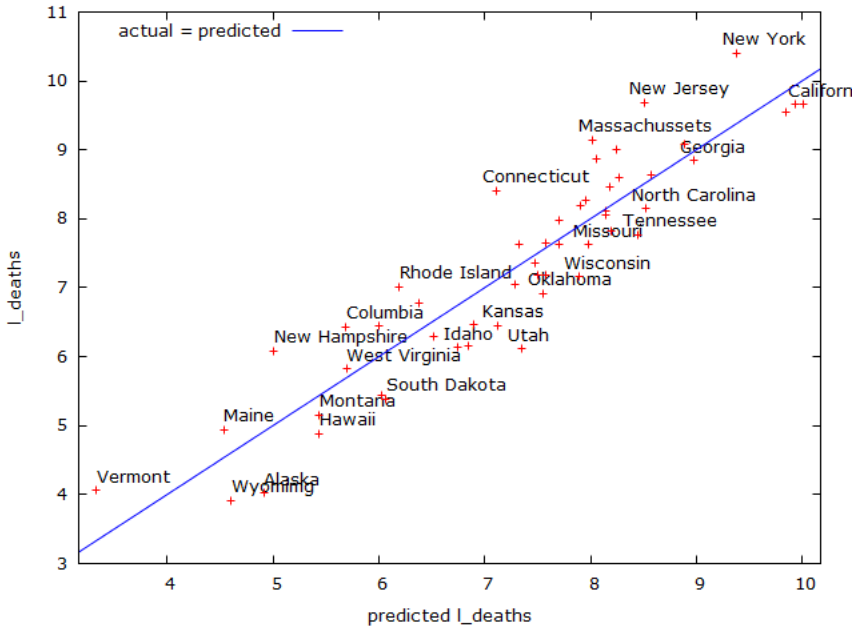
OLS, using observations 1–52
Dependent variable: l_deaths

	Coefficient	Std. Error	t-ratio	p-value
const	−5.10065	0.502934	−10.14	0.0000
l_cases	1.07777	0.0395011	27.28	0.0000

Mean dependent var	8.565309	S.D. dependent var	1.296347
Sum squared resid	5.394077	S.E. of regression	0.328453
R^2	0.937063	Adjusted R^2	0.935804
$F(1, 50)$	744.4478	P-value(F)	1.09e–31
Log-likelihood	−14.87031	Akaike criterion	33.74061
Schwarz criterion	37.64310	Hannan–Quinn	35.23673

Figure 3: **Actual versus fitted deaths on cases logarithmic transformation**

28 September 2020



21 March 2021

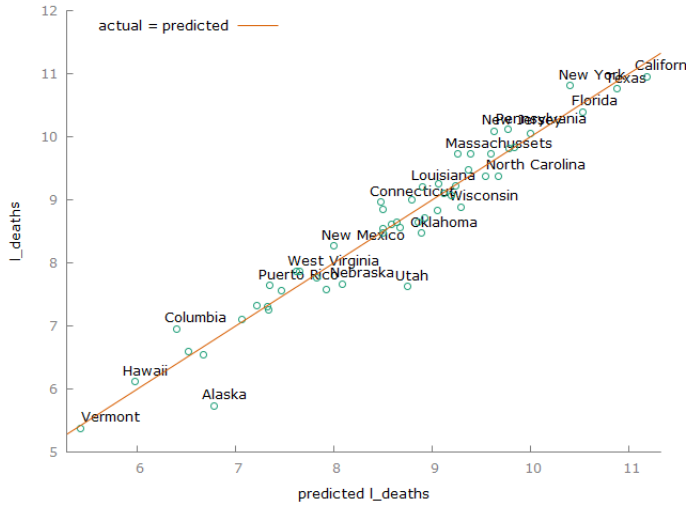


Table 6: OLS regression of deaths per state on cases, tests, population and Ppopulation densities per state

28 September 2020

OLS, using observations 1–52
Dependent variable: deaths

	Coefficient	Std. Error	t-ratio	p-value
const	−1474.64	496.671	−2.969	0.0047
cases	0.00189047	0.00577799	0.3272	0.7450
tests	0.00196253	0.000227652	8.621	0.0000
popmil	286.650	135.568	2.114	0.0398
PopDens	0.197797	0.223011	0.8869	0.3796

Mean dependent var	3936.404	S.D. dependent var	5920.028
Sum squared resid	2.75e+08	S.E. of regression	2418.844
R^2	0.846151	Adjusted R^2	0.833057
$F(4, 47)$	64.62336	P-value(F)	1.65e–18
Log-likelihood	−476.2906	Akaike criterion	962.5813
Schwarz criterion	972.3375	Hannan–Quinn	966.3216

RESET Specification Test –
Null hypothesis: specification is adequate
Test statistic: $F(2, 45) = 8.0197$
with p-value = $P(F(2, 45) > 8.0197) = 0.00104967$

21 March 2021

OLS, using observations 1–52
Dependent variable: deaths

	Coefficient	Std. Error	t-ratio	p-value
const	75.9757	708.788	0.1072	0.9151
cases	0.0148429	0.00374107	3.968	0.0002
popmil	283.224	348.775	0.8121	0.4208
PopDens	0.180819	0.344587	0.5247	0.6022

Mean dependent var	10403.44	S.D. dependent var	12625.11
Sum squared resid	6.69e+08	S.E. of regression	3733.886
R^2	0.917677	Adjusted R^2	0.912531
$F(3, 48)$	178.3555	P-value(F)	5.06e–26
Log-likelihood	−499.4143	Akaike criterion	1006.829
Schwarz criterion	1014.634	Hannan–Quinn	1009.821

Table 7: Logarithmically transformed regression of deaths per state on cases, tests, population and population densities

28 September 2020

OLS, using observations 1–52
Dependent variable: l_deaths

	Coefficient	Std. Error	t-ratio	p-value
const	−4.26766	1.26142	−3.383	0.0015
l_cases	0.692404	0.117335	5.901	0.0000
l_tests	0.173625	0.0668887	2.596	0.0126
l_popmil	0.286841	0.150370	1.908	0.0626
l_PopDens	0.235721	0.0424921	5.547	0.0000

Mean dependent var	7.302752	S.D. dependent var	1.566246
Sum squared resid	9.031175	S.E. of regression	0.438352
R^2	0.927814	Adjusted R^2	0.921670
$F(4, 47)$	151.0235	P-value(F)	3.40e–26
Log-likelihood	−28.27021	Akaike criterion	66.54043
Schwarz criterion	76.29664	Hannan–Quinn	70.28073

21 March 2021

OLS, using observations 1–52
Dependent variable: l_deaths

	Coefficient	Std. Error	t-ratio	p-value
const	−5.37392	1.06930	−5.026	0.0000
l_cases	1.05136	0.0957669	10.98	0.0000
l_popmil	−0.0290042	0.108053	−0.2684	0.7895
l_PopDens	0.133960	0.0264819	5.059	0.0000
l_tests	0.00152883	0.0325650	0.04695	0.9628

Mean dependent var	8.565309	S.D. dependent var	1.296347
Sum squared resid	3.365012	S.E. of regression	0.267574
R^2	0.960738	Adjusted R^2	0.957396
$F(4, 47)$	287.5204	P-value(F)	2.14e–32
Log-likelihood	−2.601688	Akaike criterion	15.20338
Schwarz criterion	24.95959	Hannan–Quinn	18.94368

If we use the formula above, and analyse the effect of a 1 percent increase in the number of cases, then the slope coefficient on c_i of 1.08707, suggests that a 1 percent increase in the number of cases will increase the number of deaths by 0.01 percent. The same calculation for the second period, representing 21 March 2021 and using a c_i of 1.077, also suggests that a 1 percent increase in cases will lead to a 0.01 percent increase in deaths.

A plot of the actual versus fitted values of the logarithmically transformed regression of deaths per state on cases per state is shown in Figure 4. It can be seen that the logarithmically transformed version of the regression provides a much better fit. New York is still an outlier, but not to the same degree as in Figure 3. However, in both periods, as reflected in the two plots, New York, New Jersey, Massachussets, Connecticut and Columbia, plot above the line, and therefore experience above-average deaths per case. The adjusted R squared values show a large increase to 85 per cent in the first period and 94 per cent in the second.

We now add further variables to the regression model and explore the extent to which tests, t_i, population, p_i, and population density, pd_i, contribute to the explanation of deaths by estimating the model in equation (7):

$$d_i = a + bc_i + ct_i + dp_i + fdp_i + e_i. \tag{7}$$

The results of the regression are shown in Table 6.

Table 6 shows that this regression is highly significant, with significant coefficients on tests and population, and an adjusted R squared of 83 per cent in the first period and 91 per cent in the second one. However, the RESET test suggests a non-linear specification, so the regression is re-estimated with all the variables transformed to logarithms. The results of the regression are shown in Table 7.

The regression results in Table 7 show a strong improvement. The results for the data from 28 September 2020 reveals that all the explanatory variables are statistically significant at the 5 per cent level, with the exception of population, which is significant at the 10 per cent level. The adjusted R squared is now a remarkable 92 per cent, which suggests that only 8 percent of the variation in the dependent variable remains unaccounted in the regression. The results for 21 March 2021 show a marked improvement in the adjusted R squared to 96 per cent. However, the significance of the coefficients on the explanatory variables has changed and only the coefficients on the logarithm of cases and the logarithm

of population density remain significant at the 1 per cent level. The cumulative number of tests undertaken and the actual size of the population have become insignificant. This suggests that virtually all the variation in deaths across the various US states can be explained by the number of cases and the density of the population.

Figure 5 plots the actual and fitted cases from the regression. The regression fit is more than adequate, as would be expected from the value of the adjusted R squared. In the first graph, which reflects the position on 28 September 2020, New York sits at the top of the figure, and remains something of an outlier above the regression line, while Hawaii sits well below, possibly reflecting the advantage of being an island, far removed from the contiguous States. The situation is not greatly changed in the second period graph representing data from 21 March 2021. New York remains above the regression line, though California has moved down below it, reflecting a relative improvement in its position. Massachussets, Michigan, Lousiana, Mississipi, Nevada, New Mexico, Maine and Wyoming remain above the line, but all states sit very close to it, reflecting the high R squared of the regression.

In the next section, we analyse the relationship between deaths and cases using quantile regression analysis.

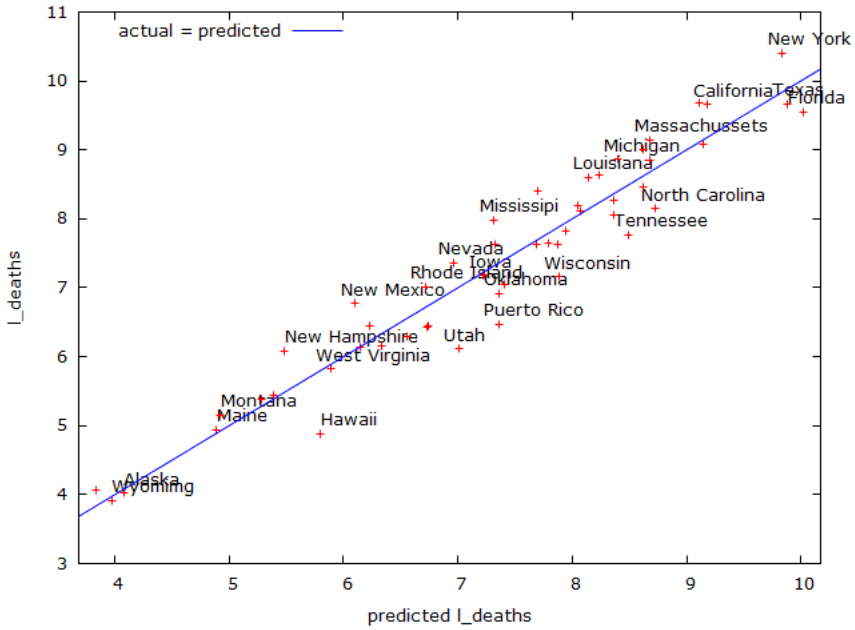
4. Quantile regression analysis

Table 8 reports the results of a quantile regression of deaths, d_i regressed on cases, c_i in the two periods. Table 8 reveals that on 28 September, the constant is significant at the 0.05 and 0.25 tau values, but not in the others, whereas the coefficient on cases is significant across the quantiles. In the second period, l as at 21 March 2021, the constants and the coefficient on cases for all tau values are significant. The coefficient on cases in the first period varies across the quantiles, from a maximum of 0.072 at the 0.95 quantile, to a minimum of 0.019 at the 0.25 quantile. This suggests that the CFR case fatality rate varies across the quantiles, which can be seen clearly in Figure 6.

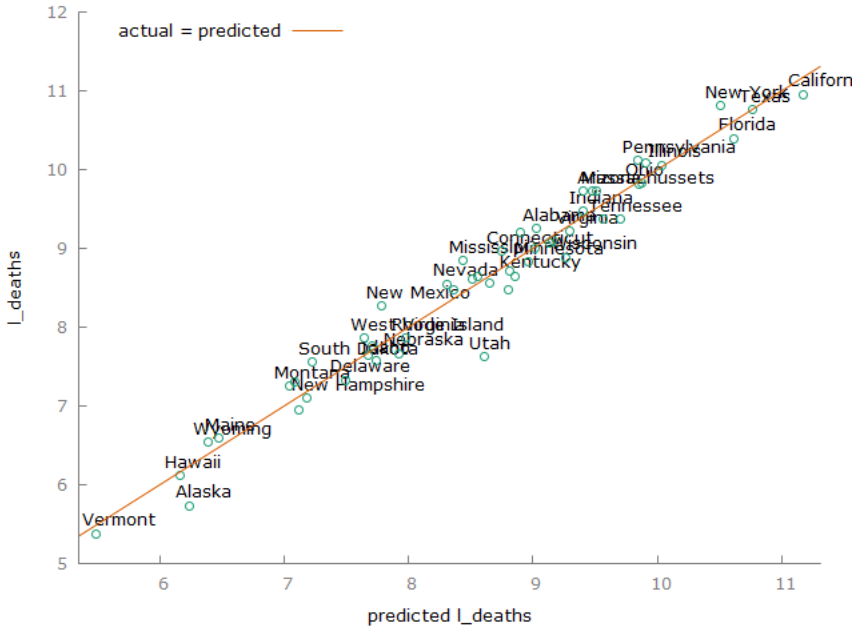
The blue line in Figure 6 shows the OLS regression coefficient, and the dotted blue line shows the 95 percent confidence bands around the regression. The confidence lines around the quantile regression estimates are much tighter, and are shown in grey.

Figure 4: **Plot of actual and fitted deaths with explanatory variables logarithmically Ttransformed**

28 September 2020



21 March 2021



It is clear that most of the quantile regression estimates are within the limits of the OLS regression estimates, with the exception of the 0.95 quantilein the first period and both the 0.75 and 0.95 quantiles in the second one. The estimate of the slope coefficient of 0.072 at the 0.95 quantile is statistically different, and represents a CFR of over 7.2 per cent in this extreme quantile in the first period, but in the second period the estimate for this quantile had dropped down to 1.7 per cent.

5. Conclusion

The implication of the success of the cross-sectional regressions, which accurately predicts the number of deaths across US States and Territories on the basis of two variables, is of some concern. This simple regression omitted all references to demographics, policies vis-a-vis lockdowns, and other possible factors of influence, yet was remarkably successful. The interpretation of the regression model is that public policies ultimately make little difference to outcomes in terms of the number of deaths. Therefore, policies that destroy economies and lead to other adverse medical, healthcare, and social effects should be viewed with caution and skepticism. The authors reach this conclusion because this simple cross-section regression specification explains between 85 and 94 percent of the variation in the dependent variable which is the number of deaths. If different public policies in different US states had a large impact, the adjusted R squares of these regressions would not be so high.

It was found that the most successful regression specifications were non-linear, involving a logarithmic transformation of the variables. The logarithmic transformation of the regression of deaths on cases led to an adjusted R squared of 85 percent and 94 percent for the two periods, and similar estimates that a 1 percent increase in COVID-19 cases would lead to a 0.01 percent increase in deaths. These regressions reflect a cross-sectional regression analysis of US State statistics at two single points in time, namely as of 28 September 2020 and 21 March 2021.

It is hoped that the algorithmic techniques, results and analysis presented in the paper will prove useful to the medical and health professions, science advisers, and risk management and decision making of healthcare by state, regional and national governments in all countries.

Table 8: Quantile Regression of Deaths on Cases using tau of 0.05, 0.25, 0.5, 0.75, and 0.95

28 September 2020

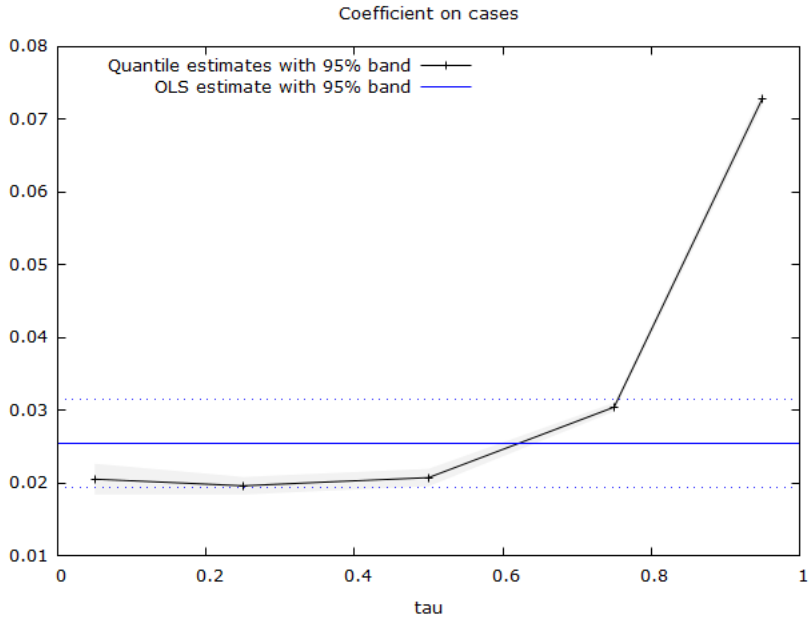
	tau	coefficient	t-ratio
constant	0.05	-994.403	-4.20783
	0.25	-263.299	-3.15842
	0.5	21.9019	0.165446
	0.75	17.0536	0.261376
	0.95	-68.9337	-0.633332
cases	0.05	0.0204971	19.4005
	0.25	0.0195944	32.6666
	0.50	0.0207222	35.0134
	0.75	0.0304021	104.226
	0.95	0.0728666	149.745

21 March 2021 2020

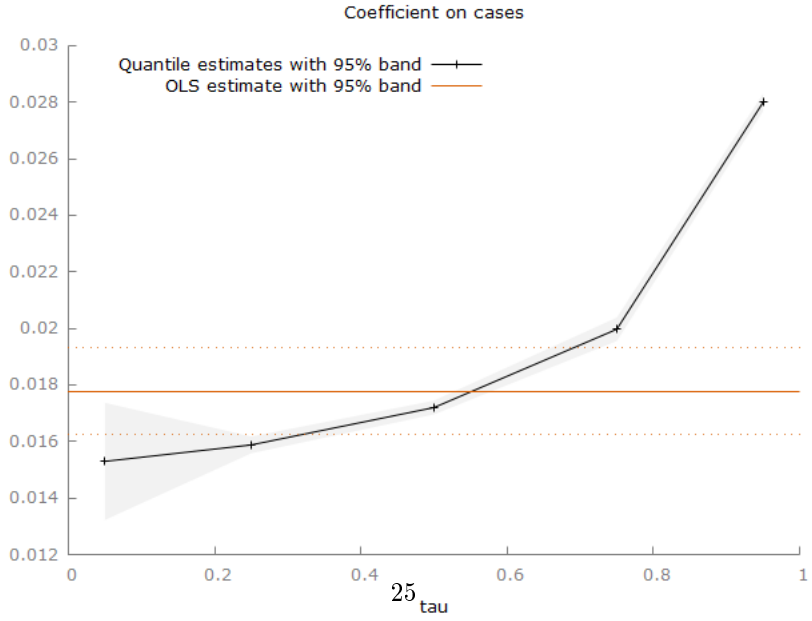
	tau	coefficient	t-ratio
constant	0.05	-1858.11	-2.04438
	0.25	-407.854	-3.15842
	0.50	-98.8797	-0.929199
	0.75	17.4678	0.0963835
	0.95	-269.793	-2.01910
cases	0.05	0.0152963	14.8611
	0.25	0.0158775	108.573
	0.50	0.0172012	142.737
	0.75	0.0199676	97.2895
	0.95	0.0279879	184.957

Figure 6: Quantile Regression Coefficients on Cases at tau 0.05, 0.25, 0.50, 0.75, and 0.95

28 September 2020



21 March 2020



Data Appendix

As stated in the text, all the data are in the public domain as they have been downloaded from the Johns Hopkins University website, with the population details obtained from Wikipedia.

The paper examines prediction of the number of cases of COVID-19 and the number of related deaths using cross-sectional regressions and quantile regression applied to US State level data sourced from the Johns Hopkins University website on 28 September 2020 and on 21 March 2021. The sample totals 52 States and Territories, given that Washington DC and Puerto Rico are included in the analysis.

The population density figures are reported as population per square mile, and population density is defined as the population per (divided by) land area. Resident population is from the United States Census Bureau estimates for July 1, 2015, (for the 50 states, Washington DC and Puerto Rico, as sourced from Wikipedia, see: https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population_density#2015_density_(states,_territories_and_DC)), accessed on 28 September 2020.

References

[1] Atalan, A. (2020), Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective, *Annals of Medicine and Surgery,* 56, 38-42

[2] Chang, C-L. and M. McAleer (2020), Alternative global health security indexes for risk analysis of COVID-19, *International Journal of Environmental Research and Public Health*, 17(9:3161) 1-18.

[3] Gostin, L.O., J.G. Hodge Jr., and L.F. Wiley (2020), Presidential powers and response to COVID-19, *Journal of the American Medical Association (JAMA) Network.* Published online March 18, 2020, doi:10.1001/jama.2020.4335

[4] Guan. D., D. Wang, S. Hallegatte, S.J. Davis, J. Huo, S. Li, Y. Bai, T. Lei, Q. Xue, D'Maris Coman , D. Cheng, P. Chen, X. Liang, B. Xu1, X. Lu, S. Wang, K. Hubacek, and P. Gong, (2020), Global supply-chain effects of COVID-19 control measures, *Nature Human Behaviour*, 4, 577-587

[5] Koenker, R., and K.F. Halloch, (2001), Quantile regression, *Journal of Economic Perspectives,* 15(4), 1434-156.

[6] Mandel, A., and V. Veetil (2020), The economic cost of COVID lockdowns: An out-of-equilibrium analysis, *Economics of Disasters and Climate Change*, 4, 431-451

[7] McAleer, M. (2020), Prevention Is Better Than the Cure: Risk Management of COVID-19, *Journal of Risk and Financial Management*, 13(3:46), 1-5.

[8] Ramsey, J. B. (1969), Tests for Specification Errors in Classical Linear Least Squares Regression Analysis, J*ournal of the Royal Statistical Society, Series B.* 31 (2), 350–371