Global Journal of Science Frontier Research: F
Mathematics & Decision Science
Volume 20 Issue 3 Version 1.0 Year 2020
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

A Regression Analysis on the Covid-19 Transmission

By Gregory L. Light
Department of Finance, Providence College

Abstract- This note applies least-squares regression to a cross-section comparison of the total infection numbers of COVID-19 as of a particular date among the fifty states of America to investigate any underlying factors; we also check the Gaussian normality of the time progression of the infection rate.

Keywords: *corona transmission causes, gaussian epidemic curve, epidemic statistical regression.*

GJSFR-F Classification: *MSC 2010: 62M10*

Strictly as per the compliance and regulations of:

A Regression Analysis on the Covid-19 Transmission

Gregory L. Light

Abstract- This note applies least-squares regression to a cross-section comparison of the total infection numbers of COVID-19 as of a particular date among the fifty states of America to investigate any underlying factors; we also check the Gaussian normality of the time progression of the infection rate.

Keywords: corona transmission causes, gaussian epidemic curve, epidemic statistical regression.

I. Introduction

This paper applies the least-squares regression to an analysis of the transmission of COVID-19 (cf. Kubiak, Arinaminpathy and McLean, 2010, for new infectious diseases). We first examine how the virus spread out (cf. Kraay, 2018, for movements of disease) by taking a sample of n =50 USA states' cumulative cases of infection Vas of March 26, 2020, against four independent variables: dummy variable B (=1) for a state containing a top-ten city in USA, dummy variable S (=1) for any state of more than 16.5% of its population of age 65 or higher, U = the number of universities/colleges in the state, and C = the number of (a popular national-chain) coffee shops in the state. We then examine the fitness of normal (Gaussian) distribution for the time progression of daily new cases as presented in the media (for a time-series cross-section treatment, see., e.g., Sharmin and Rayhan, 2012).

For the examination of the causes of the spread of the virus, we hypothesize: (1) if a state contains a big metropolitan city, then it generates more cases, (2) if a state has more than 16.5% of its residents of age sixty-five or higher, then due to seniors' vulnerability to diseases higher incidents occur, (3) the more universities/colleges there are in a state, the more the cases there are, and (4) the more "hub-coffee-shops" there are in a state, the more the transmission there is (cf. Romanescu and Deardon, 2017, Welch, 2011, for network models).

As the expression "flattening the curve" has become ubiquitous, we are motivated to see if the curve is a normal distribution (cf. e.g., Shao, 2020, for dynamic modeling). Our sample here is the daily new cases from March 1 to April 14in the State of Rhode Island of the USA, where the Author teaches, with the current students of Statistics contributing to the candidate independent variablesas well as their values for an explanation of V.

Author: Department of Finance, Providence College, Providence, Rhode Island 02918 USA. e-mail: glight@providence.edu

II. Analysis

a) Factors Contributing to the Transmission

We show the regression output below:

Regression Statistics						
Multiple R	0.74					
R Square	0.55					
Adjusted R Square	0.51					
Standard Error	3245					
Observations	50					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	4	584930295.6	1E+08	13.88707	1.8704E-07	
Residual	45	473855589.2	1E+07			
Total	49	1058785885				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	154	809.1	0.2	0.85	-1476	1784
U	4	2.9	1.5	0.15	-2	10
B	-4861	1910.5	-2.5	0.01	-8709	-1013
C	12	2.1	5.9	4E-07	8	16
S	-1477	936.2	-1.6	0.12	-3362	409

with the underlying input displayed as follows:

States:	V	U	B	C	S
Alabama	386	129	0	37	1
Alaska	59	35	0	2	0
Arizona	401	155	0	71	1
Arkansas	306	108	0	9	0
California	2982	1246	1	376	0
Colorado	1086	171	0	40	0
Connecticut	875	114	0	465	1
Delaware	119	23	0	4	1
Florida	1861	439	1	1037	1
Georgia	1441	210	0	208	0
Hawaii	76	43	0	17	1
Idaho	123	33	0	9	0
Illinois	1865	391	1	707	0
Indiana	477	175	0	16	0
Iowa	145	107	0	10	1
Kansas	98	99	0	21	0
Kentucky	198	165	1	20	0
Louisiana	1946	173	0	7	0
Maine	147	60	0	20	1
Maryland	580	148	0	20	0
Massachusetts	1838	261	1	1178	1
Michigan	2294	302	0	74	1

Notes

Minnesota	346	169	0	19	0
Mississippi	485	69	0	8	0
Missouri	356	242	0	36	1
Montana	65	54	0	5	1
Nebraska	64	68	0	24	0
Nevada	321	77	0	13	0
New Hampshire	137	43	0	19	1
New Jersey	4402	207	1	913	0
New Mexico	112	61	0	13	1
New York	32966	632	1	1562	0
North Carolina	636	188	0	345	0
North Dakota	45	30	0	22	0
Ohio	704	386	0	74	1
Oklahoma	223	158	0	21	0
Oregon	266	125	0	1	1
Pennsylvania	1128	544	1	517	1
Rhode Island	132	37	1	264	1
South Carolina	424	97	0	53	1
South Dakota	41	33	0	4	1
Tennessee	784	191	0	39	0
Texas	974	506	1	187	0
Utah	346	60	0	7	0
Vermont	123	32	0	34	1
Virginia	460	222	0	45	0
Washington	2580	164	0	2	0
West Virginia	51	99	0	5	1
Wisconsin	585	132	0	23	1
Wyoming	44	17	0	8	1

Thus, our hypotheses were moderately supported, but with surprises about the directions of the effect of B and S. A possible explanation of the negative effect of B on V may be because "big states" have better health infrastructure, and the negative coefficient of S may have the argument that seniors are less mobile. However, the most distinct observation is the p = 0.0000004 of C; we surmise the reason being people taking out food/beverages to their workplaces and spread out the virus.

b) The Fitness of Normal Distribution for Daily New Cases

We begin with a derivation of the least-squares linear regression equation (cf. e.g., Grassly and Fraser, 2008).

Let $E(x) \equiv$ the expected daily new cases; then

$$E(x) = \frac{N}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{t-\mu}{\sigma}\right)^2},$$

where $N \equiv$ the total infection number under the density curve, $t \equiv$ date, with $t = 1$ corresponding to March 1, 2020,

and $$y \equiv \ln E(x) = \ln\left(\frac{N}{\sigma\sqrt{2\pi}}\right) - 0.5\left(\frac{\mu}{\sigma}\right)^2 + \left(\frac{\mu}{\sigma^2}\right)t - \frac{0.5}{\sigma^2}t^2,$$

so that $\hat{y} = a + b_1 t + b_2 t^2$, with $-\left(\frac{b_1}{2b_2}\right) = \hat{\mu}$, $\hat{\sigma} = \sqrt{\frac{\hat{\mu}}{b_1}}$,

$$and\ \hat{N} = \sigma\sqrt{2\pi}\exp\left(a + 0.5\left(\frac{\hat{\mu}}{\hat{\sigma}}\right)^2\right).$$

Because prior to March 16 there had been the occurrences of x = 0, we select the input x from t = 16 until t = 43 (April 12) with n = 28 with the following input data:

x	lnx	t	t^2
1	0.0	16	256
2	0.7	17	289
10	2.3	18	324
11	2.4	19	361
10	2.3	20	400
12	2.5	21	441
17	2.8	22	484
23	3.1	23	529
18	2.9	24	576
8	2.1	25	625
33	3.5	26	676
38	3.6	27	729
36	3.6	28	784
55	4.0	29	841
108	4.7	30	900
87	4.5	31	961
77	4.3	32	1024
91	4.5	33	1089
52	4.0	34	1156
97	4.6	35	1225
116	4.8	36	1296
160	5.1	37	1369
148	5.0	38	1444
220	5.4	39	1521
277	5.6	40	1600
288	5.7	41	1681
334	5.8	42	1764
316	5.8	43	1849

The regression output is as follows:

Notes

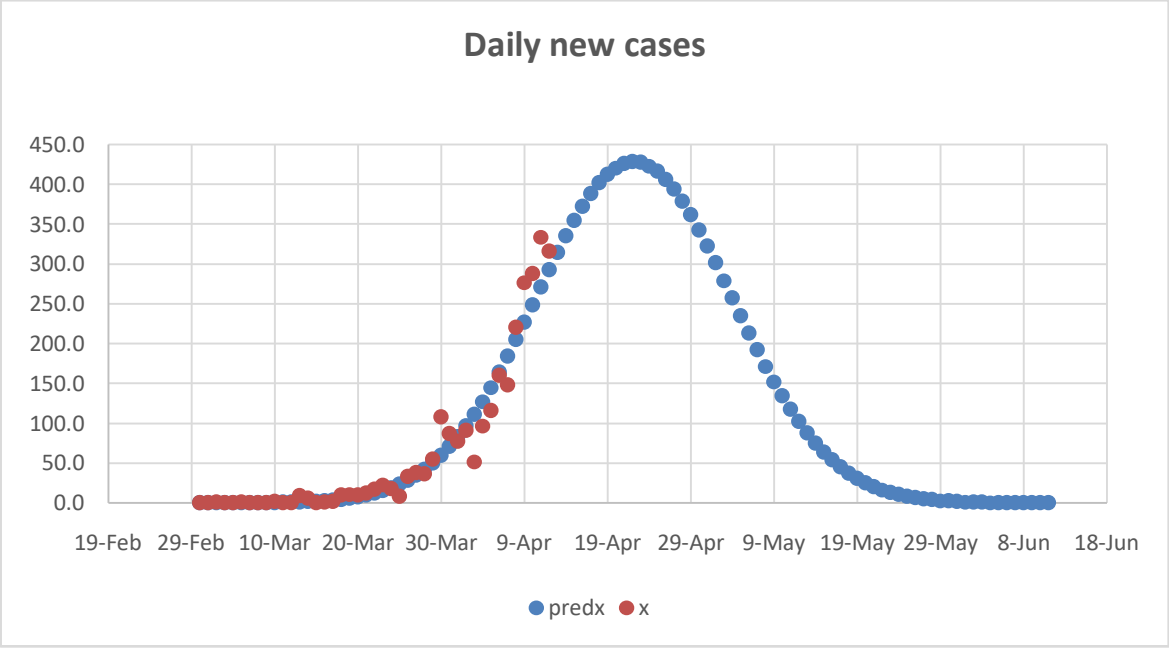
Notes

Regression Statistics						
Multiple R	**0.96**					
R Square	**0.92**					
Adjusted R Square	**0.91**					
Standard Error	**0.46**					
Observations	**28**					
ANOVA						
	df	***SS***	***MS***	***F***	***Significance F***	
Regression	**2**	**56.24430511**	**28.12215**	**135.4814**	**3.8352E-14**	
Residual	**25**	**5.189303157**	**0.207572**			
Total	**27**	**61.43360826**				
	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	**-4.298**	**1.234**	**-3.5**	**0.0018**	**-6.839**	**-1.757**
t	**0.389**	**0.088**	**4.4**	**0.0002**	**0.208**	**0.570**
t^2	**-0.004**	**0.001**	**-2.5**	**0.0204**	**-0.007**	**-0.001**

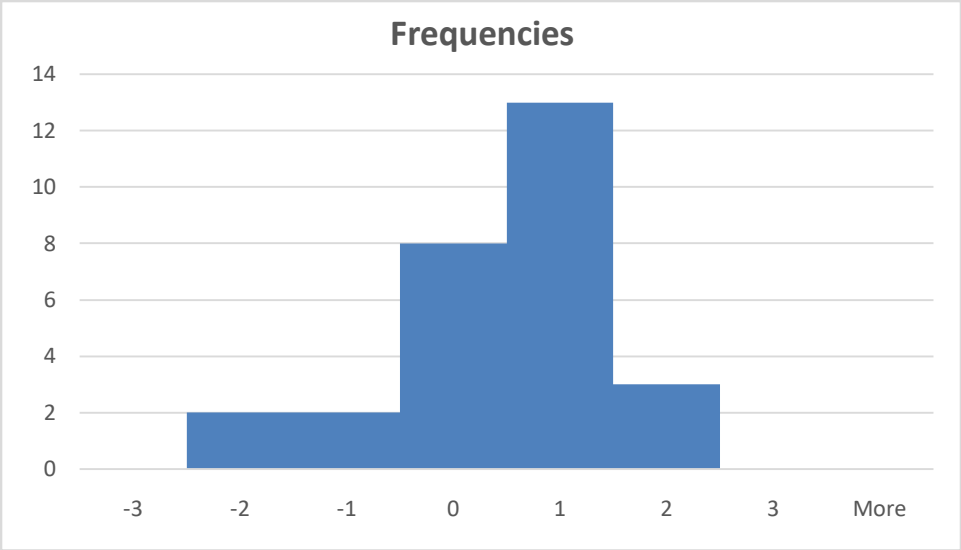
Then the above coefficients render the following estimates:

$$\hat{\mu} = 53.2 \text{ (= April 22, 2020)}, \hat{\sigma} = 11.7, \textit{and } \hat{N} = 12549,$$

with the following plot:



and the standardized residuals have the following histogram:



III. Summary

From the above analysis, we gather the following observations: (1) The medical/health/hygiene infrastructure is essential, which authorities need to improve. (2) Seniors were not the originators of the transmission; on the contrary they tend to slow down the spread. (3) College students in the USA during their Spring Break probably propelled the transmission. (4) Hubs of mass gathering, such as coffee shops, must be vigilant in observing health protocol. (5) The normal distribution of the daily new infection rates appears to be a good fit. We reason that the infection probability depends on two factors, personal immune degrees, and the external environment. The latter may have offsetting factors that tend to cancel one another (cf. Lafferty and Holt, 2003): the more rampant the cases, the more the pre-caution (as friction in mechanics). As such, individual immune degrees, which follow the normality of biology, take a normal distribution, with lower degrees of immunity occurring earlier and higher degrees later.

References Références Referencias

1. Kubiak, R. J., Arinaminpathy, N., and McLean. A. R., Insights into the evolution and emergence of a novel infectious disease, *PLoSComput. Bio.,* 6(9), 2010, 22364-22376.
2. Kraay, A. N. M., et al, Determinants of short-term movement in a developing region and implications for disease transmission, *Epidemiology,* 29(1), 2018, 117-125.
3. Sharmin, S. and Rayhan, M. I., Spatio-temporal modeling of infectious disease dynamics, *J. Appl. Stat.*, 39(4), 2012, 875-882.
4. Romanescu, R. G. and Deardon, R., Fast inference for network models of infectious disease spread, *Scand. J. Stat.*, 44(3), 2017, 666-683.
5. Welch, D., Is network clustering detectable in transmission trees? *Viruses,* 3(6), 2011, 659-676.
6. Shao, N., Dynamic models for Coronavirus Disease 2019 and data analysis, *Math. Meth. Appl. Sci.*, 43(7), 2020, 4943-4949.
7. Grassly, N. C. and Fraser, C., Mathematical models of infectious disease transmission, *Nat. Rev.Microbio.,* 6(6), 2008, 477-487.
8. Lafferty, K. D. and Holt, R. D., How should environmental stress affect the population dynamics of disease? *Ecology Lett.,* 6(7), 2003, 654-664.

Notes