Modeling the Impact of Density on the Spread of the Corona Virus

Jason M. Barr, Rutgers University-Newark
Troy Tassier, Fordham University

April 3, 2020

I. A Simple Theoretical Model

1. New Cases and the Reproduction Number

If we take the reproduction number, *R*, then the number of cases at time *t* is given time is given by:

$$cases_t = a_0 R^t \qquad (1)$$

where a_0 is number of initial cases, and *R* is the number of infections that an infected individual will pass along to others. For example, let's say *R=2*. If one person initially gets infected (say the virus fully forms within or jumps to that person at time *t=0)*, then on *t=1* (say the next day), then there are two new cases. On day two, those two people pass each to two more, for a total of four new cases, and so on.

2. The Reproduction Rate

The formula for the reproduction rate is given by:

$$R = C * T * S * D, \qquad (2)$$

where *C* is the contact rate, the number of interactions that people have per day. *T* is the transmission rate, which is the probability that any one infected individual will pass it to another if contact is made. *D* is the duration, the time from infection to recovery (or death). *S* is the susceptibility rate, the fraction of the people who are still able to be infected at any given time. Presumably, the susceptibility rate falls as people get sick and recover, since they develop immunity (or they die).

Given that we are dealing only with COVID-19 in the United States over a three-month period, we can assume *T* and *D* are constant across the country. We can also assume that *S* is constant or approximately constant since the spread of the virus is in its initial stages and is probably still very close to 100%. Finally, we assume that $a_0 = 1$.

3. The Econometric Specification

Plugging (2) into (1) then gives

$$cases_t = a_0 (C * T * S * D)^t \equiv (\beta_0 C)^t, \qquad (3)$$

where $\beta_0 = (T * S * D) \approx (0.05 * 30 * 1) = 1.5$, on the assumption that $a_0 = 1$, *D* = 30 days, the approximate time from infection to recovery (though duration probably runs from 12 to 30 days; Santilli, 2020), *T=.05* (5% transmission rate; Otto, 2020) and *S=1* (100% susceptibility).

Thus, taking the log of Equation (3) gives:

$$ln(cases_t) = ln\beta_0 t + tlnC \qquad (4)$$

Our interest is diving more deeply into the factors that impact *C*. We assume that *lnC* is a linear function of log population density and other controls:

$$ln(C) = \gamma_0 + \gamma_1 ln(Pop. Density) + \boldsymbol{\gamma_1}'\boldsymbol{X} + \varepsilon, \qquad (5)$$

where ε is the random error term (with mean zero) and $\boldsymbol{X}$ is a set of control variables.

Plugging Equation (5) into Equation (4) gives a relationship between the number of cases on day *t* and population density that is linear in the coefficients.

$$ln(cases_t) = (ln\beta_0 + \gamma_0)t + t[\gamma_1 ln(Pop. Density) + \boldsymbol{\gamma_1}'\boldsymbol{X} + \varepsilon_t] \qquad (6)$$

Equation (6) suggests that some form of ordinary least squares is appropriate (we discuss how we account for the product of *t* below).

Our econometric specification is Equation (6) but for each county in the United States. That is,

$$ln\left(cases_{it_i}\right) = (ln\beta_0 + \gamma_0)t_i + t_i\left[\gamma_1 ln(Pop. Density_i) + \boldsymbol{\gamma_1}'\boldsymbol{X_i} + \varepsilon_{i,t_i}\right],$$

Where *i={1,…,N}* is the county and t_i is the number of days since county *i* observed its first case. Population density and other control variables are assumed fixed over the three month period of analysis.

In summary, we do not know the reproduction numbers across counites. But we hypothesize that it’s tied to population density. Our aim is to estimate γ_1 and to see to what degree denser counties have higher reproduction rates. Thus, our model suggests we can regress the log of the number of COVID-19 cases on density and control for time since first case. The question becomes how to properly account for the multiplicative factor of t_i. For this analysis we experiment with two methods.

1) Ignore the multiplicative factor of *t*, but still include it as a right hand side variable and estimate the equations via OLS, which controls for the number of cases since day first day a case was observed.
2) Use weighted least squares, where we weight all the variables by number of days by first case.

II. Data and Regression Results

1. Data Sources and Preparation

Confirmed COVID-19 Cases: Our dependent variable is the number of confirmed cases of COVID-19 on March 27, 2019 (day 66) for nearly all counties in the United States. The source of the data set is https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/. From this data set we also extracted the first day of a confirmed case in each county (if there was at least one case). As additional controls we also look at number of cases on day 40 (March 1, 2020) and day 50 (March 11, 2020). In specifications that include only counties with at least one case we work with *ln(#cases)*. In specifications with all counties, we work with *ln(1+#cases on day 66).*

County Population: We take total county population as of 2018 and from the U.S. census. Table PEP_2018_PEPANNRES_with_ann.

County GDP as of 2018: https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas

Airport Traffic: For each airport in the United States we add up the total enplanements for 2019. We then sum them for the county. We created a dummy variable =1 if enplanements is greater than one million people; zero otherwise. Source: https://www.arcgis.com/home/item.html?id=900d50de880644cdb90c4cab966d0e94.

County Land Area: From US County Shapefile from the Census Bureau: https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

States: From the county shapefile we also take the states of the counties.

All data was merged using the county-level FIPS.

2. Regressions Results

Table 1 gives results of several specifications using ordinary least squares.

Table 1: Ordinary Least Squares Results: Dep. Var.: Ln(# Cases on 3/27)

	(1)	(2)	(3)	(4)	(5)
Ln(Population)	0.761***	0.433***	0.556***	0.516***	0.550***
	(13.44)	(5.68)	(6.25)	(6.10)	(5.90)
Ln(Land Area)	-0.235***	-0.224**	-0.265**	-0.185***	-0.241**
	(-2.97)	(-2.36)	(-2.58)	(-2.74)	(-2.51)
#Days 1st Case	0.0780***	0.0736***			
	(7.55)	(6.94)			
Ln(RGDP)		0.301***	0.373***	0.333***	0.388***
		(5.06)	(5.22)	(4.59)	(5.09)
Enplanements Mill+		0.269*	0.629***	0.493***	0.577***
		(1.75)	(4.10)	(5.05)	(4.65)
Ln(1+#Day40)			0.641***		
			(3.41)		
Ln(1+#Day50)				0.661***	
				(6.66)	
At Least 1 Day 40					0.636**
					(2.38)
Constant	-2.505*	-3.500*	-4.370**	-5.060***	-4.960***
	(-1.76)	(-1.92)	(-2.35)	(-3.64)	(-2.83)
N	1375	1356	1356	1702	1702
R-sq	0.779	0.786	0.740	0.759	0.733
adj. R-sq	0.770	0.777	0.729	0.751	0.724
AIC	3219.8	3141.0	3406.9	4165.1	4341.3
BIC	3235.5	3167.1	3433.0	4192.3	4368.5

t-statistics in parentheses * p<0.10, ** p<0.05, *** p<0.01. All regressions include state fixed effects. Standard errors clustered by state. Note only counties with at least one case is included.

Table 1 presents five specifications. Equation (1) regressions *ln(# cases on day 66)* for each county on *lnPopulation*, *lnLand Area*, number of days since first case. Equation (2) adds *lnRGPD* and the enplanements dummy. Equation (3) adds *ln(1+# cases on day 40)*. Equation (4) uses *ln(1+#cases on day 50)* instead of *ln(1+#cases on day 40)*. Finally, Equation (5) uses a dummy variable that takes on value of 1 if a county had at least one confirmed case, 0 otherwise. In short, across specifications, the results suggest that the elasticity of population (holding land area constant) is just between 0.44 and 0.56 on average. A doubling of city size increases the number of cases by about 50%, on average. This means a reduction in per capita cases as city size increases since cases don't increase one-for-one with city size. The days since first case coefficient in Table 1 is about 0.074, suggesting a 7.4% growth rate, on average (before systematically interventions began).

Table 2 presents the same regressions but this time using weighted least squares (i.e. aweights=days since first case, in Stata).

Table 2: Weighted Least Squares Results. Dep. Var.: Ln(# Cases on 3/27)

	(1)	(2)	(3)	(4)	(5)
Ln(Population)	0.958***	0.577***	0.636***	0.621***	0.623***
	(14.62)	(4.35)	(4.58)	(4.67)	(4.36)
Ln(Land Area)	-0.270***	-0.241**	-0.264**	-0.209***	-0.254**
	(-2.91)	(-2.19)	(-2.45)	(-2.71)	(-2.31)
#Days 1st Case	0.0415***	0.0383***			
	(4.99)	(4.46)			
Ln(RGDP)		0.348***	0.382***	0.300***	0.396***
		(4.01)	(4.16)	(3.04)	(4.14)
Enplanements Mill+		.0372	0.288	0.227*	0.322*
		(0.20)	(1.54)	(1.71)	(1.76)
Ln(1+#Day40)			0.511***		
			(3.85)		
Ln(1+#Day50)				0.500***	
				(5.59)	
At least 1 Day 40					0.493*
					(1.78)
Constant	-3.480**	-4.972**	-5.186***	-5.099***	-5.468***
	(-2.13)	(-2.41)	(-2.74)	(-3.16)	(-2.79)
N	1375	1356	1356	1356	1356
R-sq	0.828	0.833	0.819	0.838	0.818
adj. R-sq	0.821	0.826	0.811	0.832	0.810
AIC	3371.9	3296.7	3405.6	3251.7	3413.4
BIC	3387.6	3322.7	3431.7	3277.8	3439.5

t-statistics in parentheses * p<0.10, ** p<0.05, *** p<0.01. All regressions include state fixed effects. Standard errors clustered by state. Note only counties with at least one case is included.

The results are broadly similar but give higher estimates for the population coefficient since counties with earlier cases were weighted more heavily (and were likely more dense). In all cases the elasticity of number of cases with respect to population, holding land area constant, is much less than one. This suggests that a doubling of city size increases the number of cases by less than double, suggesting, on average, larger cities have lower per capita counts than smaller cities.

Finally, we estimated a third model:

$$\frac{ln\left(cases_{it_i}\right)}{t_i} = (ln\beta_0 + \gamma_0) + \gamma_1 ln(Pop.Density_i) + \boldsymbol{\gamma_1}'\boldsymbol{X_i} + \varepsilon_{i,t_i},$$

where we divided the number of cases on March 27 by the number of days since the first case. Table 3 presents the results. The problem with this model is that it changes the interpretation of γ_1, since now the dependent variable is a measure of average of log cases per day.

The results, however, are consistent with the OLS and WLS models above. The elasticity of *ln(#cases day 66)/days since first case* is far below one with respect to population. Larger cities have higher "average" number of cases but at a decreasing rate. The fact that the enplanements dummy is negative suggests that places with large airports are lower averages in their cases loads. Why this might be is left for future research. It could be tied to the fact that counties with large airports were first to get hit by the virus and were therefore more likely to implement social distancing measures, thus causing their average case rates to decline relative to other places.

Table 3: OLS. Dep. Var.: Ln(# Cases on 3/27)/day since first case

	(1)	(2)	(3)	(4)
Ln(Population)	0.0471***	0.0469***	0.0495***	0.0473***
	(5.85)	(5.73)	(6.17)	(5.94)
Ln(Land Area)	-0.00195	-0.00159	-0.00614	-0.00221
	(-0.39)	(-0.32)	(-1.18)	(-0.44)
Ln(RGDP)	0.0109	0.0121*	0.0148**	0.0118*
	(1.58)	(1.70)	(2.10)	(1.68)
Enplanements 1M+	-0.0583***	-0.0359***	-0.0393***	-0.0344***
	(-3.24)	(-3.00)	(-3.26)	(-2.92)
Ln(1+#Day40)		-0.149***		
		(-5.92)		
Ln(1+#Day50)			-0.0507***	
			(-5.89)	
At Least 1 Day 40				-0.186***
				(-9.39)
Constant	-0.478***	-0.500***	-0.466***	-0.487***
	(-4.37)	(-4.58)	(-3.94)	(-4.43)
N	1356	1356	1356	1356
R-sq	0.328	0.339	0.349	0.341
adj. R-sq	0.301	0.312	0.322	0.314
AIC	-2071.8	-2092.1	-2111.8	-2095.6
BIC	-2050.9	-2066.0	-2085.8	-2069.6

t-statistics in parentheses * p<0.10, ** p<0.05, *** p<0.01. All regressions include state fixed effects. Standard errors clustered by state. Note only counties with at least one case is included.