2nd Draft: 4/06/2020

The Geography of Covid-19 growth in the US: Counties and Metropolitan Areas

By

William C. Wheaton
Department of Economics
Center for Real Estate
MIT*
Cambridge, Mass 02139
wheaton@mit.edu

Anne Kinsella Thompson
Center for Real Estate
MIT*
Cambridge, Mass 02139
anniek@mit.edu

The authors are respectively Professor (Emeritus, Economics) and Lecturer at the MIT Center for Real Estate. They remain responsible for all results and conclusions derived there from.

JEL Code: R, I.
Keywords: Coronavirus, Health, Urban Economics
SSRN (ERN): Urban Economics & Regional Studies
(HEN): Health Economics

Abstract

It has been 70 days since the first case of Covid-19 was detected in the US. Since then it has spread and grown in all but 2 of 376 MSAs and all but 45 of the 636 counties that are contained in these MSA. In this paper we examine the determinants of how rapidly the virus grows once it has been seeded within a MSA or county. We find virus cases can be well predicted by area population, as well as days-since-onset. In the data, virus cases scale almost proportionately with population, and excluding population significantly changes the impact of days-since-onset. Growth is also related to residential density and per capita income, particularly at the county level. There are weaker relationships to MSA average household size, per capita income, and the fraction of the population that is over 65. These results come from parameterizing a simple power function model of cumulative infections since onset. This is shifted proportionately by the various MSA/County covariates. We also experiment with restricting the sample of areas so as to have a minimum number of cases – equal to .01% of the area's population. This effectively focuses on the more advanced part of the virus growth curve. Here we find a significant further decrease in the coefficient of days-since-onset. This is preliminary evidence that the virus growth is tapering. We intend to repeat our analysis as time progresses.

I. Introduction.

There is a standard epidemiological model describing the growth of a virus within a population once it has taken root with an initial "seed" from outside of that population. This is how such a model works.

Without further spread from external sources, once "seeded", the virus cases spread from those in the population who have the virus to those yet to experience it. This expands the pool of active cases. At the same time, the pool shrinks from recoveries and deaths. A sufficiently large spread rate implies the full population eventually experiences the virus.

Isolating those with the virus from those without reduces (or halts) contagion, and if sufficient, then the pool of those who have experienced the disease will be smaller. The differential equations that result from these processes depend on several parameters but even in their simplest form produce a "humped" shape distribution of active cases and a monotonically increasing cumulative function of total disease cases that is "S" shaped.

The determinants of these parameters have been speculated about, but with limited research to date. For example, little is known about the process of "seeding". Do larger populations experience the same absolute seeding as smaller ones or do larger populations inherently experience a higher inflow of initial contaminated members (a constant seeding rate)?

The spread rate is also complicated. Populations with greater social interaction, living close to each, other seemingly should have greater spread rates. The spread *rate* (from sick to healthy) is also not constant but must intrinsically decrease as those infected grow relative to those not infected.

Then there is the effectiveness of isolation. If those with the virus are easily identified, then isolation can be fully effective. On the other hand, if incomplete testing inhibits identification then isolation is far less effective at dampening contagion.

All these parameters come together to yield a solution to the model whose "humped" shaped active cases can be very peaked or flat. The "S" shaped cumulative number of cases will then eventually be a large or small share of the population. A survey of SSRN on 3/30/20 revealed 365 papers dealing with Covid-19 (jointly with Lancet). Many of these involve pure medical issues such as severity and treatment, but there are a number of papers that experiment with the model parameters discussed above, or use data to in some way parameterize them. Recent examples of such modeling include the Imperial College report (3/16/20), Harris (2020) and Wolfel et al. (2020).

II. Alternative Cross-Section Modeling of Covid-19

Our approach in this paper is not to directly model the process by which active infections spread, grow and eventually decline. Instead we focus on characterizing the function that represents the cumulative number of cases over time – the outcome. We do this across a wide range of geographies in the US (MSA, counties) and test whether there are systematic empirical patterns in how steeply sloped and how high this cumulative function has so far risen. This is estimated through 3/31/20. We use a class of power functions for the model where the cumulative number of cases (to date) can increase over time at either an increasing or decreasing rate – depending on the value of the power parameter.

Our Model is shown below in equation (1). If the parameter α exceeds one then the cumulative number of cases is not only growing over time, but it is growing at an increasing rate, so that total cases curve upward. Or course over the long run this is inconsistent with the "S" shaped epidemiological model as total cases eventually must asymptote to the total population (or less). Initially, during early stages of the disease parameter values greater than 1 may be seen – reflecting some degree of disease momentum. When α is less than one (but still greater than zero) the growth rate in cases slows over time, eventually approaching zero as total number cases peaks and flattens. This suggests perhaps that early on α may be greater or equal to one, but later on it drops to below 1.

In Model 1, the curve parameter α is assumed the same across geographies (i) – cases differ across areas purely from onset timing. Area characteristics (X_i), however, multiplicatively scale the curve, with the value of λ determining the magnitude of the scaling. The scaling that results from a unit increase in a X_i variable will shift the case curve the same regardless of the number of days that have passed (or cases contracted). Put differently, the ratio of cases between two areas with different X values will always be constant as the disease progresses. Figure 1 illustrates how the curve will change for positive and negative shifts in X (presuming that λ is estimated to be positive). We will also examine two versions of the curve. In versions A, the curve starts at day 1. In version B we examine the curve only after some initial number of days has passed since the first recorded contamination. The days are based upon when the number of cases at that date reaches some minimum number. Figure 2 shows how the curves look starting the curve at a later date (that differs depending on when the cases exceed some amount).

To statistically estimate the values of λ (and λ_0) as well as α we take natural logs and transform the equation – as in the second line of (1). If the original source of equation error is multiplicative and normal, then in the transformed equation it will be additive and log normal. The statistical standard error of the equation will likely be proportional to the number of cases, but this is taken care of by using Robust Standard Errors (clustered when using county data).

$$C_i = \beta_i T_i^{\alpha}, \quad \beta_i = \lambda_0 \prod X_i^{\lambda} \tag{1}$$

$$\ln(C_i) = \ln(\lambda_0) + \alpha \ln(T_i) + \sum \lambda \ln(X_i)$$

$where:$

C_i = Cases in area i at observation date

T_i = days since first case (version A : $C(T_i) \geq 1$), or minimun cases (version B : $C(T_i) \geq C_0$)

X_i = covariate values in area i

$\alpha, \lambda, \lambda_0$ = parameters to be estimated

Figure 1

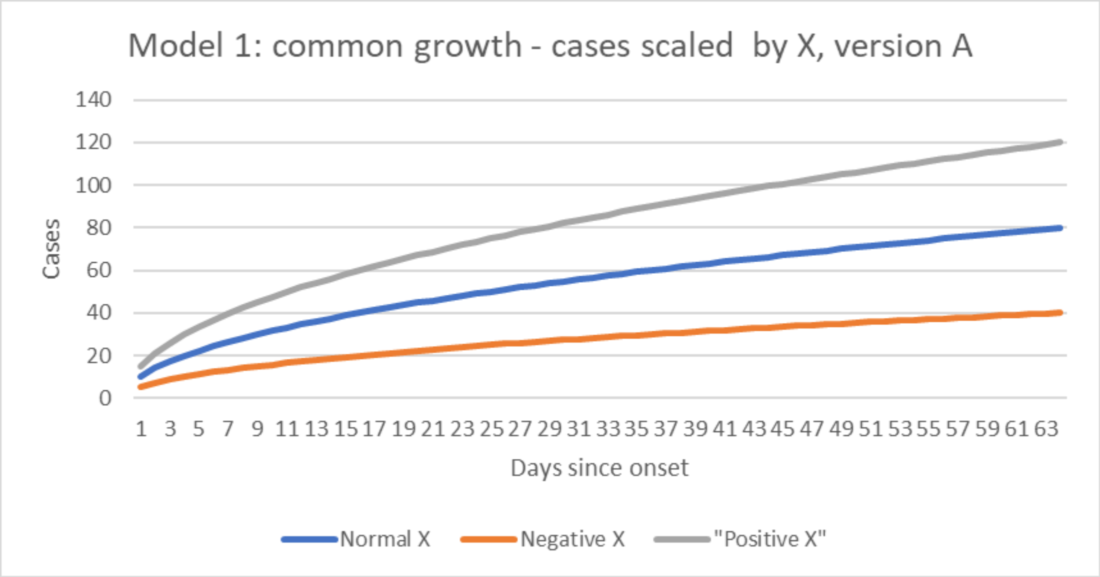
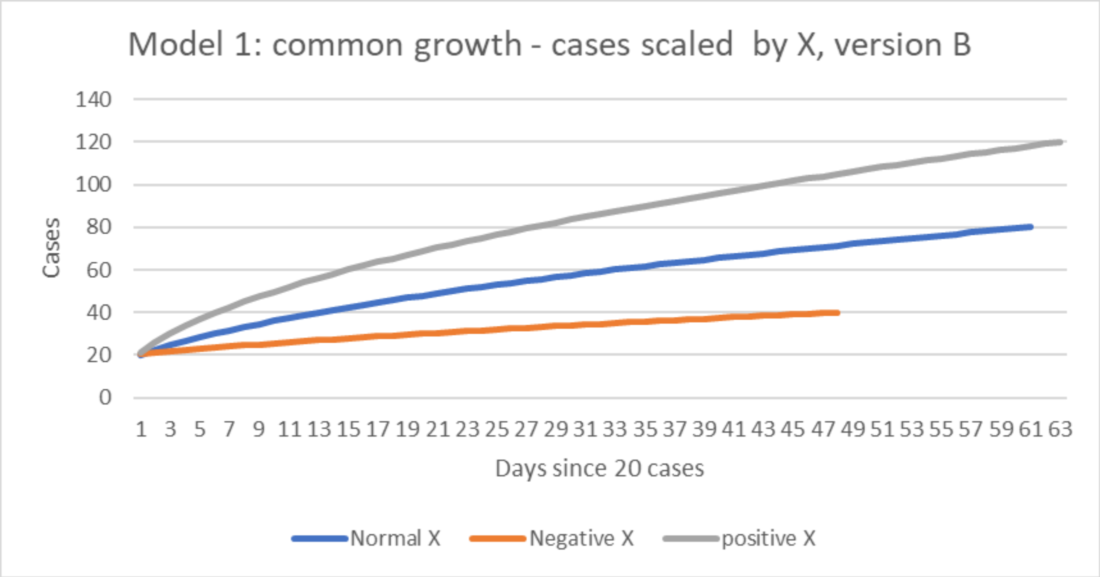


Figure 2



III. Data (as of 3/31/20)

Data collection began with identifying the counties that lie within each of 375 US Metropolitan Areas (CBSA). These counties fall into two groups identified by Census as "Primary" and "Outlying". The former is at least partially developed and collectively contain the large majority of MSA population. The latter are largely rural but with some connection to the MSA. We collected data for only 631 "central" counties.

We obtain current (2019) data for each county on its population, overall residential density (population/square mile), fraction of population over 65, average per capita income, and finally average size of household (population/household).

To this Census data we match data from 1/22/20 onward, recording the total number of (measured) Covid-19 cases to date. The data base grows "rightward" adding an additional column with each new date. Our initial analysis here is through 3/31/20. Finally, we identify the number of days that the virus has been active in each county since the first (recorded) case, (model version A) and then the number of days since cases began to exceed .01x population (one case per 10 thousand – model version B). Table 1 gives the initial county-level statistics, while Table2 aggregates this data up to the encompassing metropolitan areas. For most data the aggregation involves summation or weighted averaging. For the days variable it is the maximum of the exposure times for each county within that MSA.

Table 1: County Data

Series	Obs	Mean	Std Dev.	Minimum	Maximum
Covid 19 Cases	628	265.5	1,071.7	1	13,869
Population	631	403,879	658,828	62,607	10,105,518
Density	631	1,058.9	3,854.0	7.67	71,340
Per Capita Income	631	33,102	8,068.9	16,788	74,911
Household Size	631	2.38	0.29	0.93	3.46
Population >= 65	631	0.164	0.044	0.077	0.568
Days cases > 0	628	16.9	7.9	1	70
Days > .01% Pop	459	6.2	3.7	1	23

Table 2: MSA Aggregated Data

Series	Obs	Mean	Std Dev.	Minimum	Maximum
Covid 19 Cases	372	448.3	4,399.7	1	84,059
Population	375	679,594	1,590,459	64,265	19,247,875
Density	375	404.6	416.1	7.67	3,134
Per Capita Income	375	30,811	5,914.1	16,788	58,780
Household Size	375	2.33	0.28	0.93	3.46
Population >= 65	375	0.168	0.048	0.077	0.568
Days cases > 0	372	17.2	8.7	1	70
Days > .01% Pop	261	5.6	3.5	2	20

Some discussion on the appropriate geographic level for our analysis is in order. Counties cluster into MSA, and it is quite reasonable to expect that the growth and contagion of the virus between such counties is almost as great as between individuals within counties. For these counties, the growth of the disease then depends not only on the host county characteristics but also that of the counties it is clustered with. In any analysis of county-level data we must at a minimum use robust clustered standard errors. This is not a problem when the analysis compares MSA since they are created by Census presuming a significant degree of physical and economic unity within an MSA that is separate from other MSAs.

IV. Modeling results: MSA aggregations (372 CBSA)

In Table 3 we present the results of 4 regressions. The first runs log cases against the log of the independent variables without MSA population. The second includes population. The third is identical to the second but excludes the NY MSA, while the final regression is identical to the second but filters the sample to include only MSA that (as of 3/31/20) have cases that exceed 1-per-10 thousand people. This filter sheds 110 MSA, and has different values for Days. In all regressions the dependent variable is number cases (as of 3/31/20).

Table 3: MSA Regressions, dependent variable: log(cases)

Independent (logs)	No Pop	With Pop	No NYC	Case>.01%
R-squared	.6506	0.7946	0.7892	.9166
Constant	-23.6333	-20.6795	-19.9248	-8.5378
Population		0.9779***	0.9534***	1.0459***
		(0.067)	(0.062)	(0.040)
Density	-0.5675***	0.0646	0.0626	0.0628
	(0.079)	(0.065)	(0.064)	(0.043)
Per Capital Income	1.7551***	0.9102***	0.8779***	-0.1410
	(0.343)	(0.284)	(0.284)	(0.220)
Days	1.7793***	0.9713***	0.9958***	0.7058***
	(0.122)	(0.121)	(0.119)	(0.047)
Household Size	0.8983	-0.8948*	-0.4338**	-0.4643*
	(0.553)	(0.496)	(0.201)	(0.240)
% Population >=65	-0.3078	-0.2958	-0.3422	-0.0473
	(0.280)	(0.218)	(0.213)	(0.116)
Number of Observations	372	372	371	261

Notes: Standard errors in parentheses. *p 0.10, **p<.05, ***p<.01 (robust standard errors)

The first equation is a simple model of the raw # of cases. The Days coefficient is way above 1.0 implying a curve similar to an exponential (ever increasing) growth function. The second column, however, essentially shows that this is the wrong way to model the disease. The coefficient on population is very close to 1 (not statistically different). A log model with such a coefficient is effectively the same as a model where the dependent variable is cases/population. The fit of this model is substantially higher than the first model as unexplained variation drops by almost 45%.

The addition of population to the case equation also dramatically reduces the coefficient on number of days, from 1.78 down to .97. In this estimation, the impact of days is simply proportional to cases: cases at 40 days are twice that at 20… Medically speaking this is far more optimistic than the first model, although there still is no sign of "tapering".

Then in the final column, we limit the sample to just MSA with one or more cases per 10,000 residents. This filter sheds 110 observations, and furthermore focuses on MSAs where the disease is more advanced. This leads to a further drop in the coefficient on number of days, down to .70. At this value there is considerable tapering occurring and the curve will eventually reach a plateau. We experimented using a fixed minimum number of cases (20), but the point of the filter is to examine what the later part of the case curve looks like. If the basic model is cases/per capita then a common point on each area's curve requires a per/capita filter.

In column 3, eliminating the NY MSA from the sample has little if any impact on fit or on coefficient values. This is remarkable give that this MSA has 30% of the nation's total cases. This suggests NY simply lies at the end of a regression line from the other MSAs. Its high cases (per capita) likely result from the other covariates (longer exposure time, density…).

The other covariates have very mixed impacts. Density (at least at the average MSA level) is never significant, and the same result holds for the population share over 65. Household size has a negative impact that holds throughout, inferring greater contagion in a population of many single households relative to one with larger families. Finally, per capita income is very strong in the basic models suggesting the socializing is perhaps an income elastic activity. On the other hand, the final column implies this may just be spurious, and that MSAs where the disease is just starting happen to have lower per capita incomes

V. Modeling results: County Level Data (628 counties)

When the same models are run over a larger number of counties, the results are virtually identical to the MSA regressions. The impact of population again shows that a per capita model is superior to one with the raw number of cases, and once again this has a huge impact on the coefficient of days since onset. Also again, a filter to examine the curve only once the disease is more established continues to reduce the days coefficient down into the range where tapering is observed.

Household size and the over-65 population share continue to have weak impacts, and the impact of per capita income is exactly the same – being strongly positive until the sample is limited to those counties with a more advanced case load.

Interestingly, density differences, now between counties, have a significant positive effect. Perhaps this reflects the better ability to identify and compare downtown versus more suburban areas at the county level. The quantitative impact of this variable is quite strong as density ranges across these counties from 10 to 70,000. Two standard deviations around the mean generates a 70% increase in disease cases.

Table 4: County Regressions

Independent (logs)	No Pop	With Pop	No NYC	Case>.01%
R-squared	0.6500	0.7464	0.7375	0.8723
Constant	-10.3732	-18.4352	-19.4800	-7.7737
Population		0.8847***	0.8873***	0.9393***
		(0.072)	(0.073)	(0.052)
Density	0.5883***	0.2580**	0.2082**	0.2026**
	(0.096)	(0.109)	(0.084)	(0.089)
Per Capital Income	0.5685**	0.7197***	0.8437***	-0.187997
	(0.264)	(0.212)	(0.268)	(0.145)
Days	1.7185***	1.0159***	1.0137**	0.8216***
	(0.110)	(0.101)	(0.099)	(0.086)
Household Size	0.8573	-0.0573	-0.0663	0.5520
	(0.715)	(0.733)	(0.320)	(0.601)
% Population >=65	0.3444	0.2399	0.1716	0.3585
	(0.403)	(0.377)	(0.356)	(0.298)
Number of Observations	628	628	624	459

Notes: Standard errors in parentheses. *p 0.10, **p<.05, ***p<.01 (clustered robust errors)

VI. Discussion

The results of our estimated suggest that when comparing the spread of the disease across cities, regions or even countries, there really needs to be a focus on cases per capita. Larger areas (greater populations) seem to have higher infections early on and this carries over throughout the spread of the disease. Two population are at a similar stage of the disease (similar place along the cumulative curve) only when their *per capita* cases are the same. With more travel, more trade, it also is quite possible that larger populations have a greater number of external “seedings” so perhaps we should think of a seeding rate, rather than a single seeding.

Our results on the impact of time-since-initial exposure(s) are somewhat optimistic. If we examine the underlying growth of the disease in the US areas – beyond an initial level of 1 case per 10,000 – the growth rate is distinctively decreasing. An ebbing of case growth is certainly a prelude to a peak in the number of active cases and then subsequent recovery of the population - in an epidemiological model.

Of our demographic covariates, there were two surprises. The result that higher income areas have much larger cases per capita needs an explanation. It is tempting to suggest that perhaps dining out, entertaining, and socialization are all income elastic consumption items – items that also generate higher infection risk. But we need further direct research before drawing this conclusion. The other surprise was the anticipated strong impact of density occurs only when we begin to examine smaller areas like counties and not entire MSAs.

We will continue to redo these models as additional data become available. We also will explore using more flexible functions for the underlying relationship between exposure days and number of cases. The problem is that parameterizing them statistically can be quite complicated.

REFERENCES

Ferguson, Neil et al. (3/16/2020*) Impact of Non-Pharmaceutical Interventions to reduce Covid-19 Mortality and Health Care Demand*, Imperial College London, Covid-19 Response Team.

Harris, Jeffrey, *The Coronavirus Epidemic Curve is already flattening in NYC,* NBER Working Paper, (3/30/2020).

Wölfel, R., V.M. Corman, W. Guggemos, and et al. 2020. *Virological assessment of hospitalized cases of coronavirus disease 2019*
https://www.medrxiv.org/content/10.1101/2020.03.05.20030502v1.full.pdf: March 5, 2020.

https://www.nytimes.com/interactive/2020/04/03/world/coronavirus-flatten-the-curve-countries.html

https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

https://www.cnn.com/interactive/2020/health/coronavirus-us-maps-and-cases/