Response to Reviewers

Antonio Paez

5/11/2020

Reviewer 1

Thanks for the revisions. I think they make the paper stronger.

> Thank you for your positive assessment of the paper and comments.

I have a question on the rationale for using daily new cases (log transformed) as a dependent variable. Given the large differences in population size, urban areas and densities across states, wouldn't using percentage change in new cases day to day be a more appropriate dependent variable (or some moving average of it)? Such a setup removes the population size effect… I would suggest the authors at least test the percent variable and see if it allows for a simpler specification. Or alternately, explain why log(new cases) is a valid measure to answer their question given differences across states that are not accounted for.

> Thank you for this comment. I understand percentage change in new cases for state i and day t to mean the following:
>
> $$pct\_change_{i,t} = \frac{new\_cases_{i,t} - new\_cases_{i,t-1}}{new\_cases_{i,t-1}}$$
>
> Alas, as seen in Table 1 below, percent change correlates poorly with the independent variables, and the low correlations translate into a poor fit with non-significant coefficients (see Table 2). This is because the daily percent change in new cases is very noisy and not a good indication of a trend: even small changes (say from +1 to -1 case change from day to day) change the sign of the variable.
>
> Instead of day-to-day percent change in new cases, we could define a variable that measures the percentage increase in total cases:
>
> $$pct\_increase_{i,t} = \frac{cases_{i,t} - cases_{i,t-1}}{cases_{i,t-1}}$$

Table 1: Simple correlation between day to day percent change in new Cases and the mobility indicators

	pct_change	retail	groceries	parks	transit	work	residential
pct_change	1.00	0.04	0.04	0.03	0.03	0.03	-0.03
retail	0.04	1.00	0.88	0.38	0.91	0.96	-0.98
groceries	0.04	0.88	1.00	0.40	0.88	0.88	-0.88
parks	0.03	0.38	0.40	1.00	0.47	0.33	-0.37
transit	0.03	0.91	0.88	0.47	1.00	0.91	-0.93
work	0.03	0.96	0.88	0.33	0.91	1.00	-0.98
residential	-0.03	-0.98	-0.88	-0.37	-0.93	-0.98	1.00

Note:
All mobility indicators are lagged 11-day moving averages

Table 2: Dependent variable is day-to-day percent change in new cases.

Variable	Coefficient Estimate	p-value
date	-0.0081	0.8750
date^2	0.0022	0.3161
parks^2	-0.9679	0.2919
parks	1.8082	0.4856
parks x work	1.7504	0.7383
work	-3.5551	0.4027
work^2	1.5319	0.8299
parks x date	0.0326	0.5388
parks x date^2	-0.0013	0.4423
work x day^2	-0.0415	0.6918
work x date^2	-0.0015	0.5182
NY	-0.0809	0.8857
NY x date	-0.0174	0.5536

Note:
Coefficient of Determination R^2= 0.012
Adjusted Coefficient of Determination R^2= 0.007
Standard Error σ= 4.441

Table 3: Simple correlation between daily percent increase in total cases and the mobility indicators

	pct_increase	retail	groceries	parks	transit	work	residential
pct_increase	1.00	0.48	0.39	0.10	0.42	0.50	-0.47
retail	0.48	1.00	0.88	0.38	0.91	0.96	-0.98
groceries	0.39	0.88	1.00	0.41	0.88	0.87	-0.88
parks	0.10	0.38	0.41	1.00	0.46	0.33	-0.36
transit	0.42	0.91	0.88	0.46	1.00	0.92	-0.93
work	0.50	0.96	0.87	0.33	0.92	1.00	-0.98
residential	-0.47	-0.98	-0.88	-0.36	-0.93	-0.98	1.00

Note:
All mobility indicators are lagged 11-day moving averages

Again, a variable like this would normalize for the caseload in each state. The correlations for percent increase are shown in Table 3, were we see that they are considerably higher than for day-to-day percent change in cases, but still lower than for log of new cases. If we use daily percent increase in total cases as the dependent variable, the resulting model is somewhat more parsimonious but the goodness of fit is also quite a bit lower than the model with log of new cases (see Table 4).

The examples above show that using the day-to-day percent daily change in new cases and percent daily increase in total cases does not result in better models. The point about controlling for population still stands. This made me think of a different dependent variable, namely incidence per 100,000:

$$incidence_{i,t} = \frac{Cases_{i,t}}{population_i/100,000}$$

Incidence normalizes the number of cases and accounts for the size of the population at risk. This variable is log-transformed to scale it (it has a long tail, as you would expect). The correlations with the mobility indicators are shown in Table 5. Using the log of incidence as the dependent variable results in a somewhat more parsimonious model with better goodness of fit (see Table 6). Based on these results, this is the model that I report in the paper now. It is worthwhile noting that qualitatively the results are very similar, but with a slightly smaller model that also performs better.

I am also concerned that almost all the variables in the model have a quadratic form and by the number of

Table 4: Dependent variable is daily percent increase in total cases.

Variable	Coefficient Estimate	p-value
parks	0.2246	0.0071
parks x work	-0.3642	0.0066
work	-0.2737	0.0452
work^2	0.7591	<0.001
parks x date	-0.0047	<0.001
parks x date^2	0.0002	0.0011
work x date^2	-0.0002	0.0183
NY	0.0941	0.0037
NY x date	-0.0070	<0.001

Note:
Coefficient of Determination R^2= 0.453
Adjusted Coefficient of Determination R^2= 0.452
Standard Error σ= 0.26

Table 5: Simple correlation between log(incidence) and the mobility indicators

	log_incidence	retail	groceries	parks	transit	work	residential
log_incidence	1.00	-0.86	-0.73	-0.20	-0.82	-0.91	0.89
retail	-0.86	1.00	0.88	0.38	0.91	0.96	-0.98
groceries	-0.73	0.88	1.00	0.41	0.88	0.87	-0.88
parks	-0.20	0.38	0.41	1.00	0.46	0.33	-0.36
transit	-0.82	0.91	0.88	0.46	1.00	0.92	-0.93
work	-0.91	0.96	0.87	0.33	0.92	1.00	-0.98
residential	0.89	-0.98	-0.88	-0.36	-0.93	-0.98	1.00

Note:
All mobility indicators are lagged 11-day moving averages

Table 6: Dependent variable is log(Incidence).

Variable	Coefficient Estimate	p-value
date	0.1618	<0.001
date^2	-0.0008	0.0093
parks^2	0.2518	0.0562
parks	4.1307	<0.001
parks x work	-8.1480	<0.001
work	9.4568	<0.001
work^2	-3.4421	<0.001
parks x date	-0.0948	<0.001
parks x date^2	0.0036	<0.001
work x date^2	-0.0058	<0.001
NY	1.8627	<0.001

Note:
Coefficient of Determination R^2= 0.972
Adjusted Coefficient of Determination R^2= 0.972
Standard Error σ= 0.689

interactions in the model. Is this an artifact of not controlling for the population size effect?

> The quadratic form is to capture trends in attribute space. As the performance of the model shows, these terms are effective at this task. And seeing how the terms are significant even after controlling for population size (i.e., modelling the incidence per 100,000 instead of new cases), I hope that your concerns will be allayed.
>
> Finally, I have taken the opportunity of this revision to update the analysis to May 5, 2020.