Modeling the Spread of Cornavirus across U.S. County Borders

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

May 18, 2020

1. The Regression Model

The dependent variable for all regressions is the (natural) log of the total number of cases in county *i* at time *t*. The first set of models estimates the following equation,

$$\ln(cases_{it}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_n x_{ni} + \varepsilon_{it},$$

where x_{ji} is variable *j=1,…,n*, for county *i*. These control variables, at the county level, and their sources, are as follows:

- *Population* (in logs): American Community Survey, 2018. Table PEP_2018_PEPANNRES_with_ann.
- *Land area* (in logs): From U.S. County Shapefile from the Census Bureau: https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
- *Gross Domestic Product* (in logs): 2018, https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas
- *Number of days since first case*: (same as Coronavirus cases source below—USAFacts.org).
- *Race*: %Black, %Asian, %Hispanic, % Native American % Hawaiian or Pacific Islander, %Other Race, and %Two or More Races (note that %White is the omitted variable): American Community Survey 2018, 5-year average. Table ACSDT5Y2018.B02001_data_with_ov.
- *Enplanements Dummy Variable*: 1 if number of airplane passengers in county is >1 million, 0 otherwise. Created from enplanement data from https://www.arcgis.com/home/item.html?id=900d50de880644cdb90c4cab966d0e94
- *% of workers who take public transportation to work*: American Community Survey, 2018 5-year average. Table ACSST5Y2018.S0801_data_with_ove.
- *% of county workers 65 years of age or older*: American Community Survey, 2018, 5-year average. Table ACSST5Y2018.S0101_data_with_ove.
- *% of residents who are male*: American Community Survey, 2018, 5-year average. Table ACSST5Y2018.S0101_data_with_ove.
- *% below poverty line*. American Community Survey, 208, 5-year average. Table ACSST5Y2018.S1701_data_with_ove.

- Total COVID-19 cases at county level by each are from: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

2. Regression Results

The first set of regression results keeps the same right-hand side variables, but looks at number of cases in 10-day intervals starting on day 59 (March 20). The regression results are given in Table 1. A few things to note form the table. The population effect steadily increases over time. That is, once seeded denser counties see more rapid spread of the virus. The average county growth rate declined from about 6.4% to 3.4%, showing a systematic flattening of growth rates. The percent Black coefficient estimates rise over time, as does the %Hispanic estimates. The estimates from public transportation usage decline but remain statistically significant and positive.

	(1) lnCases59	(2) lnCases69	(3) lnCases79	(4) lnCases89	(5) lnCases99	(6) lnCases109
lnpop	0.234***	0.566***	0.745***	0.767***	0.859***	0.938***
	(2.80)	(7.81)	(9.79)	(12.59)	(14.51)	(17.09)
lnland	0.0289	-0.0924**	-0.0870**	-0.0736	-0.0716	-0.0878
	(0.45)	(-2.22)	(-2.19)	(-1.59)	(-1.39)	(-1.58)
lnRGDP	0.188**	0.166**	0.0586	0.0707*	0.0366	-0.00880
	(2.59)	(2.64)	(1.16)	(1.73)	(0.85)	(-0.21)
EmpGTMill	0.0880	0.00723	-0.0195	-0.0716	-0.119	-0.121
	(0.51)	(0.05)	(-0.16)	(-0.59)	(-0.98)	(-1.02)
tfirstcase	0.0639***	0.0684***	0.0568***	0.0490***	0.0400***	0.0341***
	(5.06)	(8.42)	(12.37)	(12.67)	(9.42)	(8.11)
pop_65+ _pct	0.0240**	0.0195***	0.00660	0.00375	-0.00186	-0.00766
	(2.15)	(2.91)	(0.99)	(0.45)	(-0.21)	(-0.83)
male_pct	-0.00237	-0.00947	-0.0280***	-0.0237**	-0.00960	-0.00332
	(-0.11)	(-1.00)	(-2.88)	(-2.19)	(-0.80)	(-0.25)
%Public Trans.	0.0472***	0.0371***	0.0308***	0.0299***	0.0227***	0.0171**
	(5.89)	(3.72)	(3.97)	(4.49)	(3.12)	(2.18)
black_pct	0.0176***	0.0131***	0.0220***	0.0260***	0.0299***	0.0327***
	(3.85)	(3.60)	(8.10)	(11.57)	(11.98)	(11.70)
natam_pct	0.0255**	0.0194***	0.0159**	0.0127*	0.0127	0.0142
	(2.39)	(2.94)	(2.45)	(1.69)	(1.47)	(1.58)
asian_pct	0.0529***	0.0255	0.0175	0.0160	0.0208	0.0231*
	(3.55)	(1.18)	(1.08)	(1.25)	(1.62)	(1.82)
hipi_pct	-0.0933	-0.256**	-0.183*	-0.202*	-0.246**	-0.231*
	(-0.40)	(-2.11)	(-1.93)	(-1.74)	(-2.05)	(-1.83)
otherrace_~t	0.409***	0.196*	0.204***	0.178***	0.109	0.0154
	(3.13)	(1.75)	(2.84)	(2.72)	(1.41)	(0.18)
twomore_pct	-0.0559	0.00305	0.0317	0.0216	-0.00185	-0.0191
	(-1.52)	(0.12)	(1.37)	(0.87)	(-0.06)	(-0.53)
hispanic_pct	0.0131**	0.0117***	0.0120***	0.0141***	0.0192**	0.0231**
	(2.32)	(3.47)	(3.53)	(2.72)	(2.42)	(2.31)
poverty_pct	-0.0402***	-0.0299***	-0.0327***	-0.0314***	-0.0313***	-0.0294***
	(-3.79)	(-4.23)	(-4.20)	(-4.61)	(-4.65)	(-4.73)
_cons	-7.699***	-6.650***	-4.654***	-4.823***	-5.359***	-4.983***
	(-4.59)	(-6.12)	(-4.95)	(-4.30)	(-4.38)	(-3.67)
N	711	1607	2046	2188	2187	2188
R-sq	0.756	0.816	0.839	0.833	0.815	0.803
adj. R-sq	0.732	0.808	0.834	0.827	0.809	0.797

Table 1: Regression Results for Number of Cases in 10-day intervals. t statistics in parentheses * p<0.10, ** p<0.05, *** p<0.01. All regressions include state fixed effects. Standard errors are clustered by state. Only counties with at least one reported case included in each regression.

3. Border Effects

To estimate county neighbor and border effects we took the following steps. From ArcGIS we obtained the neighbors of each county that share a border with it using the U.S. county shapefile. From this, we took the average number of cases for each bordering county on day 59 (March 20, 2020). We then took the log of the average and included as a control variable in the regression, where the dependent variable is *ln(cases)* on day 109 (May 9) for counties with at least one case. Given the 60-day lag, we likely need not be worried about endogeneity. We also include *ln(1+cases)* as of day 59 as a control variable. That is, we included a measure of the lag of cases in own county, to better estimate the border effects from day 59.

For the second variable, we first calculated the cartesian distance of the centroid of each county to every other one, in degrees. Then we created weights from county *i* to county *j*, for all *j* equal to one over the distance between each county. The weights were then normalized by dividing by the sum of weights from county *i* to all other counties. We then took a weighted average of all cases on day 59 based on the distances from county *i* (note $w_{ii}=0$). We repeated this for all counties in the United States.

Table 2 presents the regression results for four specification. Equation (1) is the same as Equation (6) in Table 1, given as a benchmark. Equation (2) is the same as (1) but replaces day since first case with *ln(1+cases day 59)*. Equation (3) includes the average of the cases from bordering counties. Equation (4) includes the weighted average of cases based on distance of each county to every other county. Equation (5) includes both *ln(border avg.)* and *ln(weighted average)*.

The results show that an increase in cases by 10% in border counties on day 59 is associated with an increase in cases on day 109 of about 1.8%, on average, *ceteris paribus*. An increase in the weighted average cases of 10% on day 59 is associated with 7.6% increase in cases on day 109, on average, *cet. par*.

Note the map given in the blog of cases originating outside the border was created by taking for each county: *exp(0.181*ln(average border cases))* and then adding up the cases for each state.

	(1) lnCases109	(2) lnCases109	(3) lnCases109	(4) lnCases99	(5) lnCases99
lnpop	0.932***	1.011***	0.953***	0.924***	0.883***
	(17.37)	(17.33)	(17.40)	(15.61)	(16.02)
lnland	-0.0869	-0.114**	-0.0715	-0.0569	-0.0288
	(-1.56)	(-2.14)	(-1.40)	(-1.28)	(-0.65)
lnRGDP	0.000867	-0.00550	0.0292	0.0408	0.0655*
	(0.02)	(-0.13)	(0.73)	(1.01)	(1.71)
EmpGTMill	-0.136	-0.308***	-0.183*	-0.170*	-0.116
	(-1.10)	(-3.18)	(-1.97)	(-1.99)	(-1.35)
tfirstcase	0.0341***				
	(8.12)				
medianage	0.0147	0.0140	-0.0000814	0.00769	-0.00126
	(0.89)	(0.92)	(-0.01)	(0.52)	(-0.08)
pop_65+_pct	-0.0225	-0.0299*	-0.0142	-0.0224	-0.0105
	(-1.38)	(-1.82)	(-0.86)	(-1.35)	(-0.62)
male_pct	-0.00391	0.00590	0.00771	0.00135	0.00254
	(-0.30)	(0.38)	(0.52)	(0.09)	(0.18)
Pub_trans_pct	0.0167**	-0.00180	-0.00967	-0.0172	-0.0179
	(2.06)	(-0.20)	(-1.05)	(-1.48)	(-1.55)
black_pct	0.0326***	0.0301***	0.0290***	0.0267***	0.0263***
	(11.92)	(11.65)	(11.02)	(11.38)	(11.10)
natam_pct	0.0142	0.0112	0.0108	0.0103	0.00913
	(1.58)	(1.39)	(1.31)	(1.28)	(1.12)
asian_pct	0.0235*	0.00187	0.00196	0.000610	0.00103
	(1.83)	(0.19)	(0.20)	(0.06)	(0.11)
hipi_pct	-0.228*	-0.137	-0.159	-0.147	-0.166
	(-1.80)	(-1.06)	(-1.17)	(-1.15)	(-1.23)
otherrace_pvt	0.0192	0.0540	0.0416	0.0599	0.0519
	(0.23)	(0.60)	(0.48)	(0.70)	(0.64)
twomore_pct	-0.0181	-0.0159	-0.00826	-0.0126	-0.00492
	(-0.50)	(-0.53)	(-0.27)	(-0.52)	(-0.20)
hispanic_pct	0.0234**	0.0188*	0.0170*	0.0132*	0.0127*
	(2.30)	(1.79)	(1.69)	(1.68)	(1.68)
poverty_pct	-0.0271***	-0.0219***	-0.0161**	-0.0211***	-0.0176**
	(-4.39)	(-3.16)	(-2.27)	(-3.20)	(-2.67)
lnCases59		0.290***	0.225***	0.306***	0.265***
		(6.97)	(5.49)	(8.09)	(6.74)
Lnavgbordercases_59			0.181***		0.147***
			(6.05)		(4.95)
Lnweightavgcases_59				0.760***	0.522**
				(3.56)	(2.14)
_cons	-5.415***	-5.142***	-5.892***	-7.098***	-7.319***
	(-3.97)	(-3.66)	(-4.42)	(-6.06)	(-6.46)
N	2188	2833	2821	2768	2758
R-sq	0.803	0.802	0.805	0.813	0.814
adj. R-sq	0.797	0.797	0.800	0.808	0.810

Table 2: Regression Results for Number of Cases on Day 109 (May 9). t statistics in parentheses * p<0.10, ** p<0.05, *** p<0.01. All regressions include state fixed effects. Standard errors are clustered by state. Only counties with at least one reported case included in each regression