COVID-19 Spread and Temperature: a Statistical Analysis

Sean Eli
Rice University

Introduction: idea and impact

It is important to understand the relationship between disease spread and temperature: this knowledge can influence how we allocate resources across the country, and can be useful in predicting *how the disease will affect us in the future*. The purpose of this project is **to analyze the relationship between COVID-19 rate of spread (i.e. the rate of growth of number of cases) and local temperature, across the United States.** We briefly summarize our methods for estimating the rate of spread, and describe temperature considerations; we then describe a multiple testing setup/evaluation for correlations between growth rates and temperature, and include a basic linear model. **Our result (as of April 16),** based on 27 U.S. cities, suggests the rate of spread of COVID-19 may be correlated with temperature, in a *mild*, but interesting way.

Quantifying the growth rate

Early-stage disease spread tends to follow a roughly exponential model (every day, the # cases is multiplied by some approximately constant factor). However, the rate factor is usually variable: the growth rate changes with time. Thus, a realistic functional approximation $N(t)$ to the number of cases per day is of the form

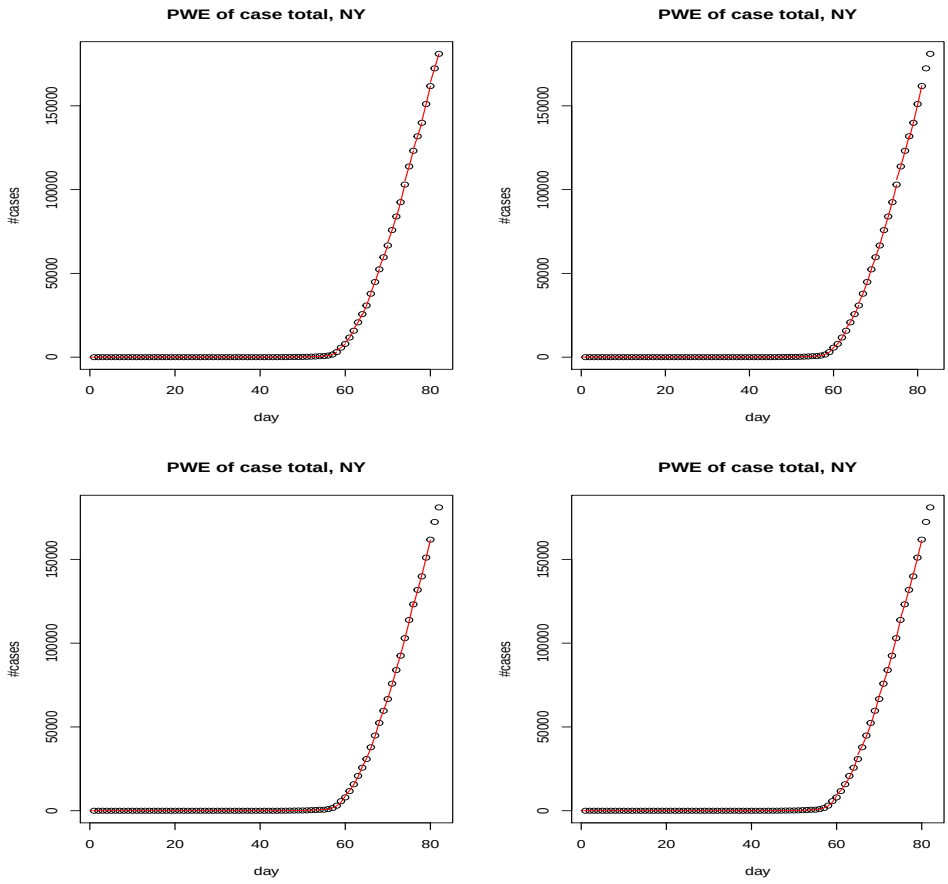
$$N(t) \propto e^{t \cdot c(t) + b(t)},$$

where $c(t)$ and $b(t)$ are time-dependent growth parameters. Keeping this in mind, here are the three ways in which we quantify the COVID-19 rate of spread in a given community:

1. **(First Order Linear Difference)** In a given region, let $N(t)$ be the total number of cases at day t. Compute directly $(N(t+k)-N(t))/k$ for each t, and various k (usually $k=1$). This is an approximation to the first derivative of $N(t)$.

2. (**Multiplicative Factors**) Given case totals $N(t)$, compute $MF(t) = N(t+1)/N(t)$, starting at the first t for which $N(t) > 0$. This is the approximate daily multiplicative factor.

3. (**Piecewise Exponential (PWE)**) Given case totals $N(t)$, partition time (days) into groups of k, and fit each group of k points with an exponential curve. This gives a PWE approximation to $N(t)$, which can be used to obtain rate information. Below are PWE approximations to the NY state case total (04/12/20), for $k = 2, 3, 4, 5$.



We assume the **linear difference is a good approximation to rates obtained through the PWE approximation**, and captures the changing rate well, for its computational ease.

Correlation testing and analysis

Now that we can quantify COVID-19 spread rates, let's look at the relationship between these rates and temperature. Our data consists of pairs $(R(t), T(t))$ for (the various) rate measurements $R(t)$ and (various) temperature measurements $T(t)$. We face a difficulty in this analysis since **COVID-19 has an incubation period between 0 days and 2 weeks.** Moreover, persons showing symptoms may not be tested right away. We address this by offsetting temperature, i.e. using data pairs such as

$$(R(t), T(T-i)), \qquad \text{for } i = 0, ..., 14.$$

Our analysis takes the form of **a series of correlation tests**, where we run one test for each set of paired data $(R(t), T(t-i))$, for:

1. each choice of $R(t)$ (linear difference, multiplicative factors)
2. each temperature time offset $i = 1, ..., 14$ (and possibly higher than 2 weeks)
3. each choice of temperature $T(t)$ (daily max, daily min, daily average, etc.)
4. **each choice of correlation test** (R provides Spearman's ρ, Kendall's τ, and Pearson's test. The first two are appropriate when data is not necessarily $N(0, 1)$.)

Given a choice of $R(t)$, $T(t)$, and correlation test, we evaluate the resulting series of tests using both the standard rule of $p < .05$ and also the stricter **Benjamini-Hochberg procedure**, with false discovery rate control at .05.

For each choice of $R(t)$, $T(t)$, and offset i, we perform **linear regression** of $R(t)$ on $T(t-i)$, and include the least-squares line in our results. This may contain some useful information.

Preliminary Results: April 16

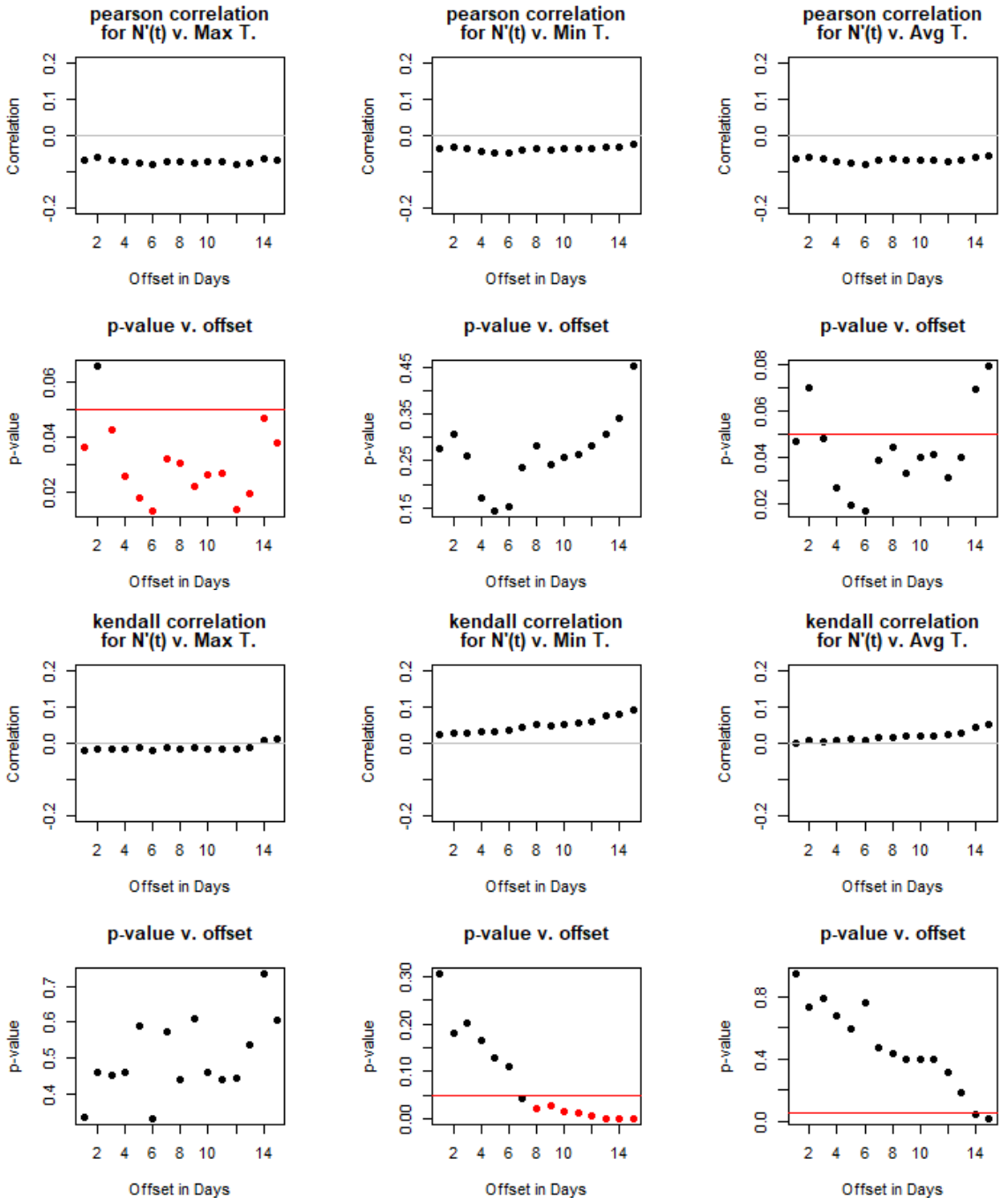
Pooling together data across several U.S. regions should help minimize effects of other variables (e.g. social distancing, demographic data) on COVID spread rate. We pool together pairs of temperature and rate data for the following 27 regions: this list contains 13 of the highest case totals in the U.S.

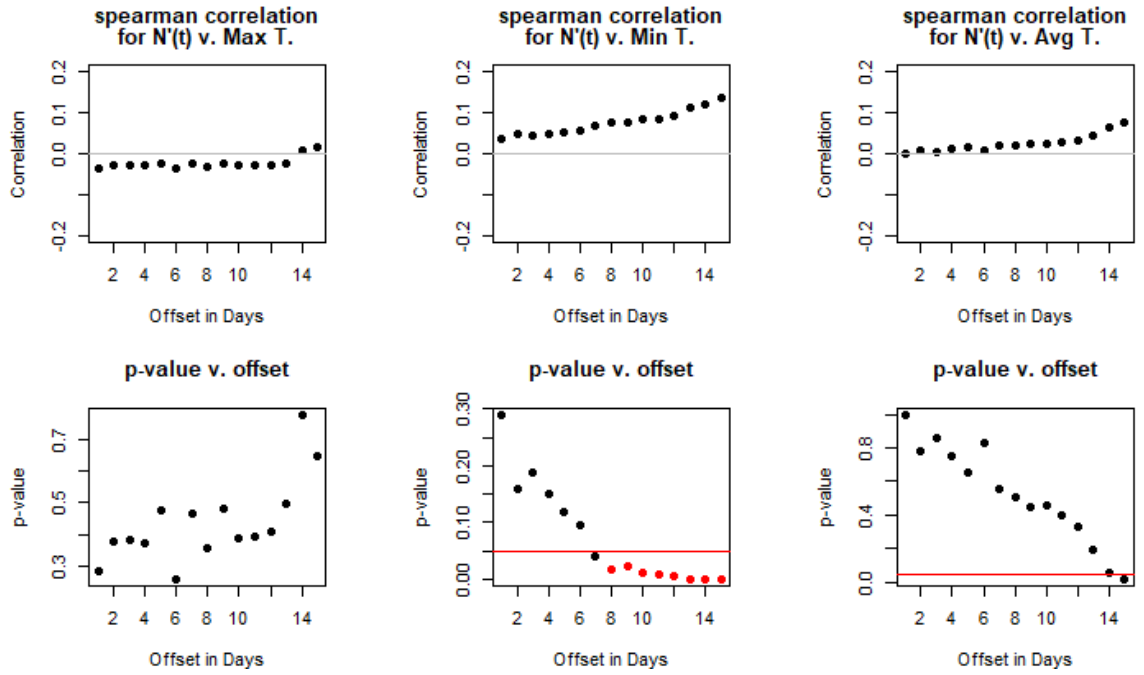
We obtain temperature data for individual counties/cities from the interactive map by the NOAA (URL at the end of this document). This is time series data, recording either the daily high, daily low, or daily average temperature, in the desired county, every day from January 1 to April 16.

Rate Estimate: Linear Difference Rule. We use the linear difference estimate of $N'(t)$, starting at the first time at which $N'(t) > 0$, as our measure of growth rate. This gives $n = 929$ data pairs for each test. We run $15 \times 3 \times 3$ correlation tests, for each of the following combinations:

1. Temperature offset from 0 to 14 days.
2. Max, Min, and Average daily temperature
3. Spearman, Kendall, or Pearson correlation

This is a sequence of **nine** multiple tests (one multiple test for each choice of temperature variable and correlation statistic). For each multiple test, we plot the correlation vs. temperature offset, and also the corresponding p-values. The red line is at $p = 0.05$, and p-values which are highlighted red correspond to null hypotheses rejected by the Benjamini-Hochberg procedure at $q = 0.05$. (i.e. these p-values correspond to **nonzero correlations.)**


pearson correlation for N'(t) v. Max T.
pearson correlation for N'(t) v. Min T.
pearson correlation for N'(t) v. Avg T.
p-value v. offset
p-value v. offset
p-value v. offset
kendall correlation for N'(t) v. Max T.
kendall correlation for N'(t) v. Min T.
kendall correlation for N'(t) v. Avg T.
p-value v. offset
p-value v. offset
p-value v. offset
Correlation
p-value
Offset in Days

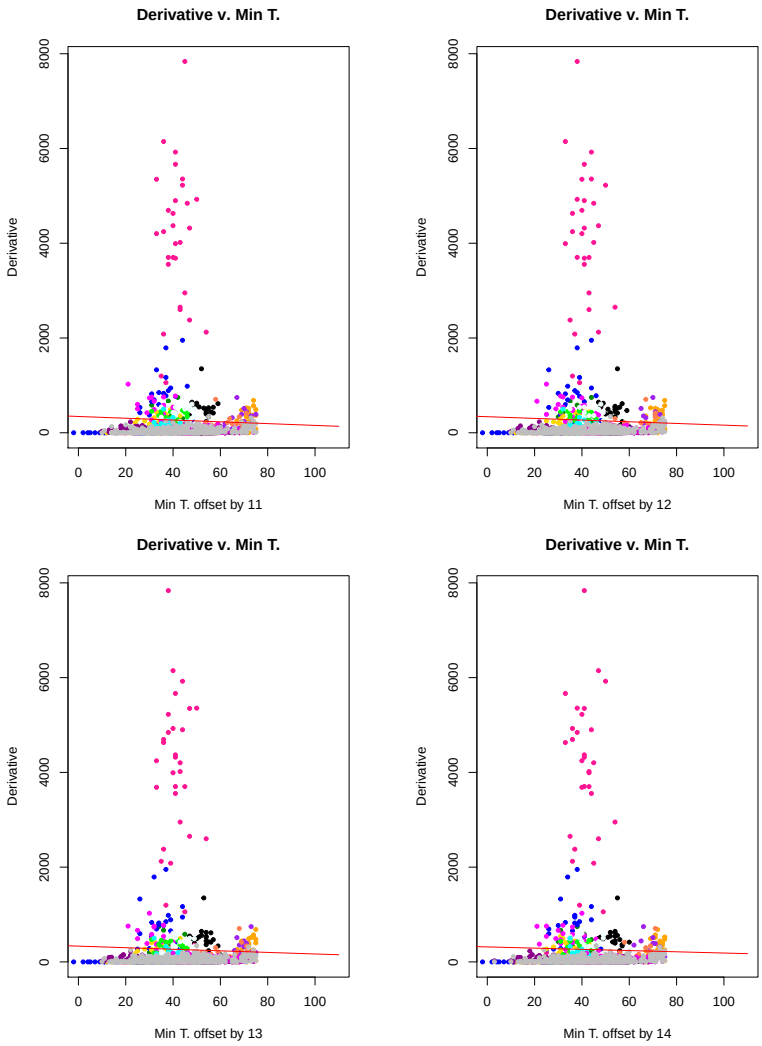


Using Pearson's correlation is questionable since these data are certainly not normally distributed (rate data is nonnegative). The BH procedure rejects almost all null hypotheses (H_0 : correlation $= 0$) for $N'(t)$ v. max temperature, and no others. The standard rule ($p < .05$) implies significant **negative** correlations with between $N'(t)$ and max temperature, and $N'(t)$ and average temperature. Notice, all Pearson correlations are negative. Kendall's τ gives similar results to Spearman's ρ. We see τ is not correlated with max temperature, for any of the time offsets we checked. By the BH procedure at $q = 0.05$, **we fail to reject hypotheses that τ is not correlated with minimum temperature, offset 7-14 days**. The BH rule suggests τ is not correlated with average temperature, for any offset. Most Kendall correlations are **positive**.

In all cases, **the test correlation is *mild*, with absolute value** < 0.15**.** The Kendall and Spearman tests suggest $N'(t)$ is possibly **positively correlated with the minimum temperature from 1-2 weeks before**, while the (less appropriate) Pearson test suggests $N'(t)$ is possibly **negatively correlated with the maximum temperature, from 0 days-2 weeks be-**

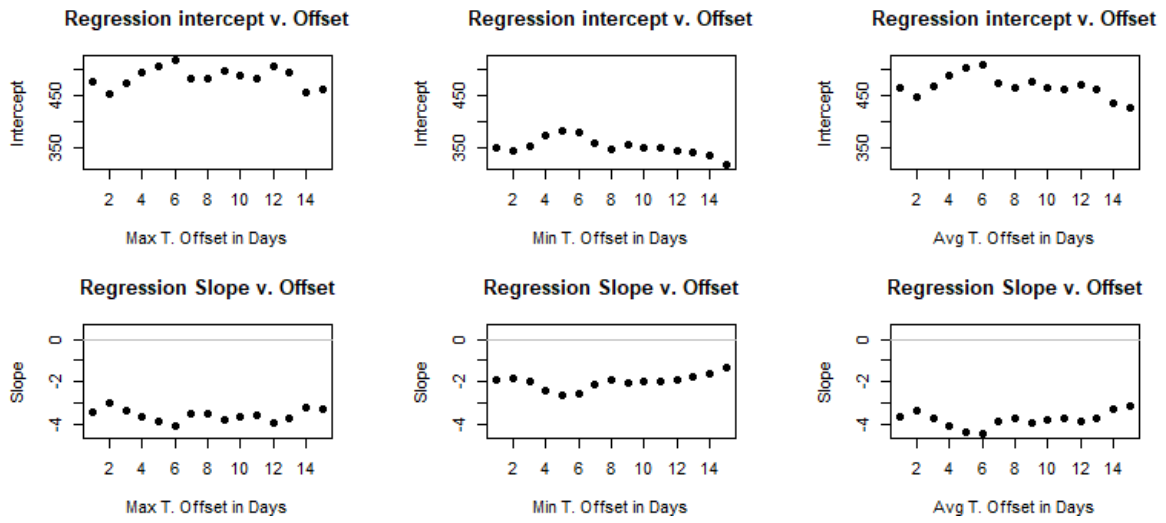
fore.

Below, we plot the rate $N'(t)$ vs. minimum temperature, with offsets between 11 and 14 days (these have significant negative correlation according to the Kendall and Spearman tests), with least-squares lines:



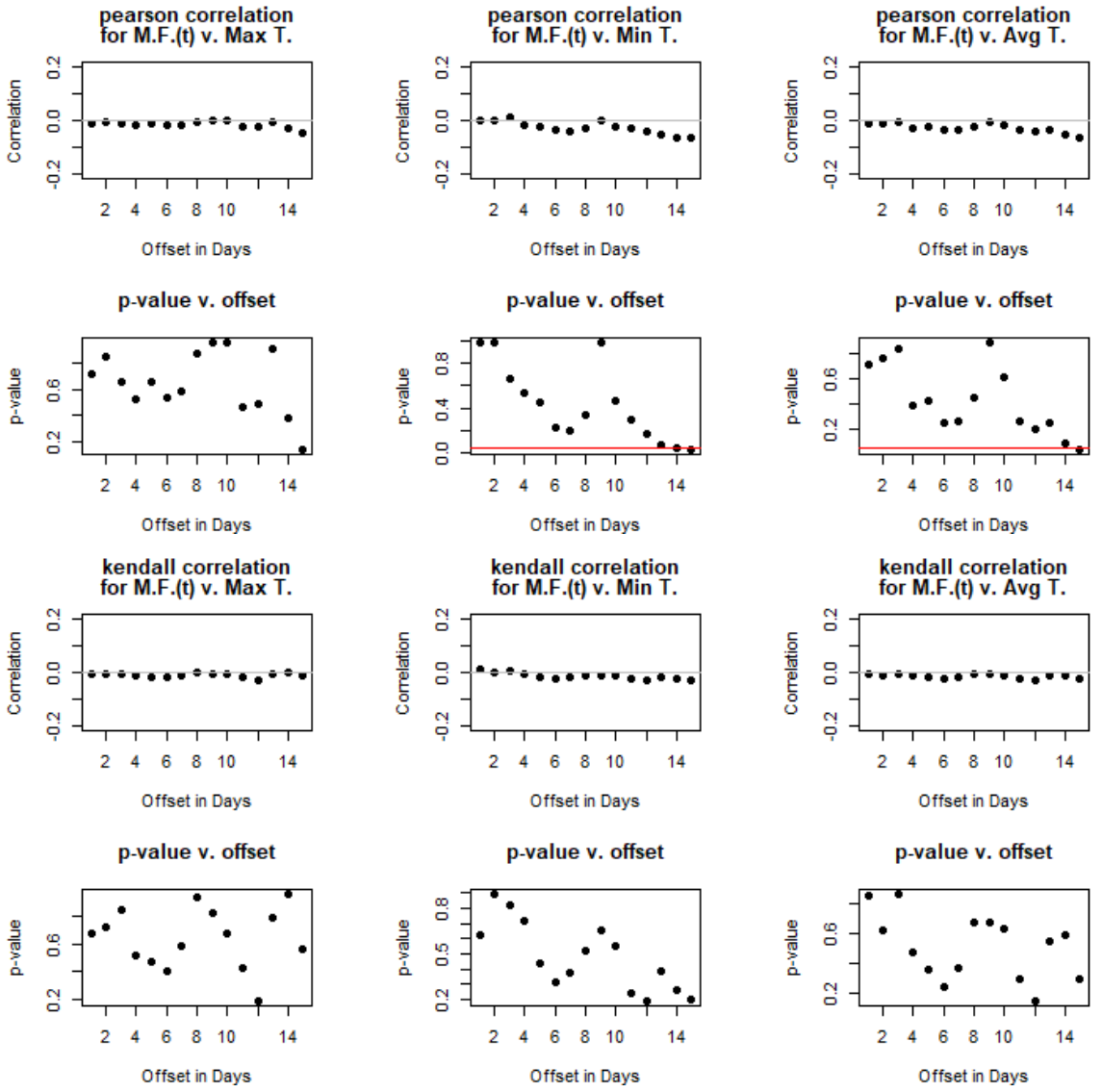
Notice there several points with extremely low temperature and low $N'(t)$. It is possible these points influence the result and influence the significant positive correlation, seen by Kendall's test. Notice the least-squares lines have negative slope.

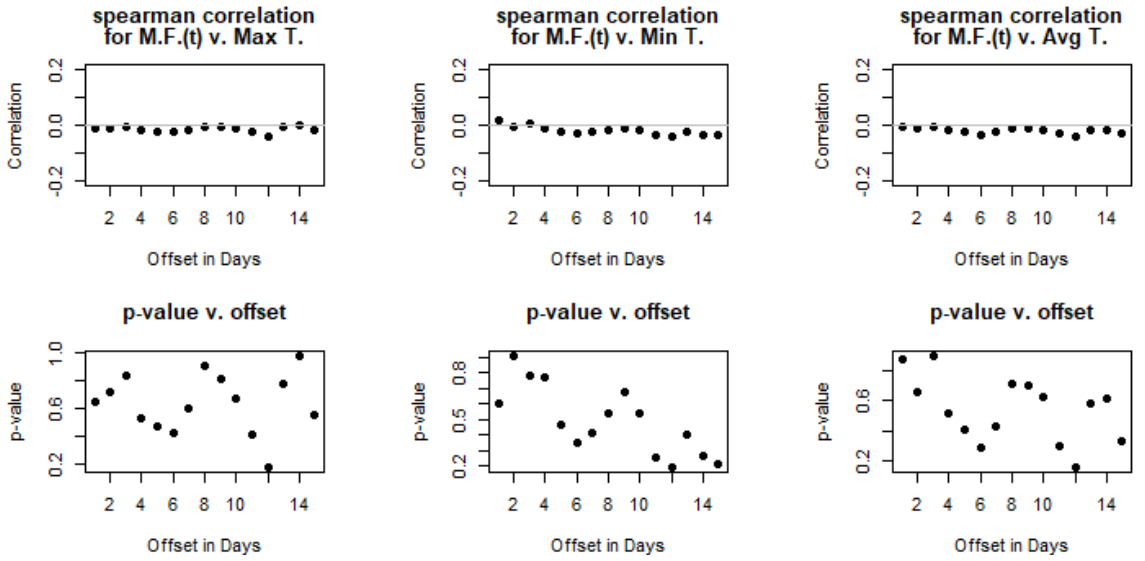
Here, we plot the least squares slope and intercept, for each choice of temperature variable and offset:



Interestingly, (as in the graph on the previous page also) all least-squares slopes are negative, with the strongest negative slopes appearing for $N'(t)$ v. Max temperature and $N'(t)$ v. average temperature. This seems to **agree with the negative correlations from Pearson's test.**

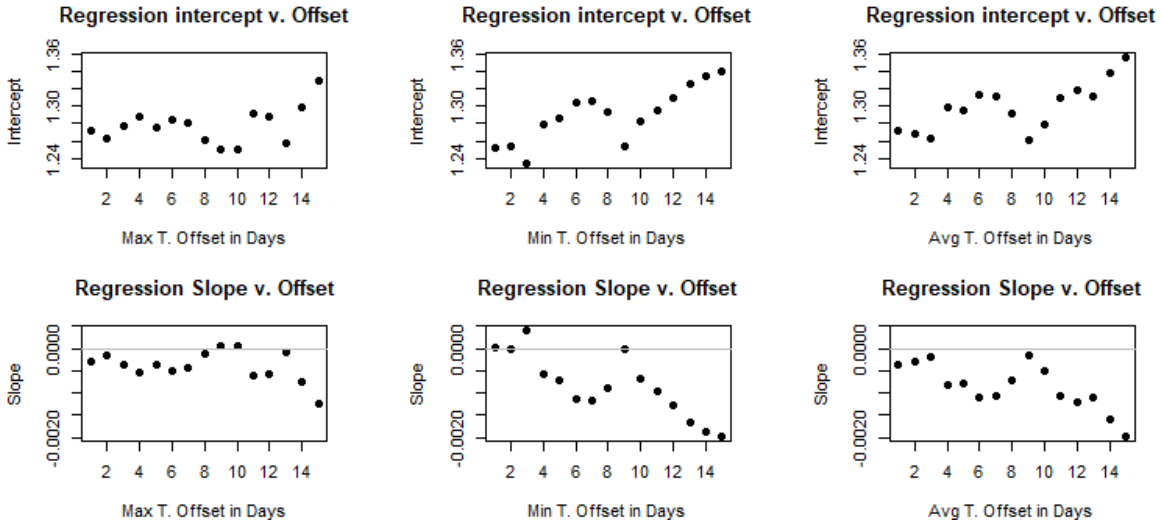
Rate Estimate: Multiplicative Factor Rule. We run the same $15 \times 3 \times 3$ correlation tests as above, but for the daily multiplicative rates $MF(t)$ (after the first day at which the rate was positive). This gives $n = 1074$ data pairs for each test. For each multiple test, we plot the correlation vs. temperature offset, and also the corresponding p-values:


pearson correlation for M.F.(t) v. Max T.
pearson correlation for M.F.(t) v. Min T.
pearson correlation for M.F.(t) v. Avg T.
p-value v. offset
p-value v. offset
p-value v. offset
kendall correlation for M.F.(t) v. Max T.
kendall correlation for M.F.(t) v. Min T.
kendall correlation for M.F.(t) v. Avg T.
p-value v. offset
p-value v. offset
p-value v. offset
Correlation
p-value
Offset in Days



All correlations are bounded in absolute value by 0.08, so if any are true, they are very *mild*. The BH procedure with FDR control at $q = 0.05$ fails to reject any null hypotheses (H_0: correlation $= 0$). The standard rule of $p < .05$ only rejects the zero Pearson correlation hypothesis for Minimum temperature offset 14 days, and average temperature, offset 14 days.

We inlclude the least-squares slopes and intercepts, for each choice of temperature variable and offset:



The slopes are all close to 0, with the most extreme negative slopes happening for $MF(t)$ v. Minimum temperature, offset 14 days, and $MF(t)$ v. average temperature, offset 14 days. As with the linear difference rate, this seems to agree with the Pearson correlation results.

Summary of Results: April 16

In all cases, **the test correlations are *mild*, with absolute value** < 0.15**.**

1. **Linear Difference Rate:** Kendall and Spearman tests suggest $N'(t)$ is possibly **positively correlated with the minimum temperature from 1-2 weeks before**, while the (less appropriate) Pearson test suggests $N'(t)$ is possibly **negatively correlated with the maximum temperature, from 0 days-2 weeks before.** All tests are evaluated with the BH procedure, with FDR control at 0.05. Least-squares lines seem to support the above negative Pearson correlations.
2. **Multiplicative Factor Rate:** All correlations are very mild, with absolute value less than 0.08. Kendall and Spearman tests suggest no correlation between this rate and any temperature variables. Pearson's test suggests $MF(t)$ is **negatively correlated with minimum temperature offset 14 days, and with average temperature, offset 14 days.** The Pearson correlation result is supported by the slope of least-squares lines.

Possible Explanation/ future research

Our results suggest the growth rate of COVID-19 cases may be positively correlated with minimum temp. from 1-2 weeks before, and may be negatively correlated with maximum temperature overall. This suggests the highest growth rates appear when the temperature is not too low and not too high (between 50-80 degrees F). There are many possible explanations for the 2-week delay for minimum temperature effects.

One possible explanation is that the extreme minimum temperatures prolong the COVID-19 incubation period. To test this hypothesis, we need reliable data on the COVID-19 incubation period (i.e. individual patient data. Date of first exposure to infected person, date of first onset of symptoms. This is not easy data to find, since the date of first exposure will often be impossible to determine.)

Data sources

COVID-19 data:

`https://github.com/CSSEGISandData/COVID-19`

Weather from NOAA:

`https://w2.weather.gov/climate/`