COVID-19 Exercise

In this exercise you will explore using digital traces to detect outbreaks of COVID-19 in the state of Washington. There are many ways to define an outbreak, but for this task you will use a relatively simple definition: exponential growth of COVID-19 cases sustained over a number of weeks. First, download the CSV file for the exercise: `covid_traces_WA.csv`. The first non-date column represents the ground truth we will use to define outbreak occurrences (positive test counts). The remaining columns are various digital traces we will use for outbreak detection: upToDate search intensity, CDC reported ILI, Twitter related tweets, Google searches for 'fever', and Kinsa fever reporting. For ease of demonstration, the data has already undergone some preprocessing, which includes smoothing, which is why you will see fractional case counts.

(a) First, in a 6-panel plot, show the case counts and the five other digital traces over time. First focus on the case count curve. Intuitively, where would you decide are starts of outbreaks? Then compare with the other digital traces. (Due to data limitations, only a small segment of the digital traces are available.) Do any of them show "outbreak" behavior? Are they lagging or leading indicators compared to the COVID-19 case counts?

(b) Now you will work on defining an outbreak mathematically. First, take a sliding window of 11 days over the case counts. In each window, fit a linear regression using days 1-10 to predict days 2-11. The regression should *not* include an intercept. Collect the slope of this regression α_i in a vector the same length as the original data. This will give you a vector of slopes α for each day in the dataset (the first 10 or so will be 0 which is fine).

(c) Make a three-panel plot containing (i) the case count curve over time, (2) α over time, and (3) a binary variable indicating if $\alpha_i > 1$ over time. Interpret the meaning of α.

(d) Next, you will step through an outbreak detection algorithm. Whenever $\alpha_i > 1$ for 10 days in a row (sustained case growth), we detect an outbreak event as being active starting on the 10th day. However, we only want to record the *start* of the outbreak, so we only record the first time such a consecutive event occurs (so even if there are 15 days of sustained growth, only the 10th day is marked). Then once there are 10 consecutive days without growth ($\alpha_i < 1$), we define the outbreak to be over. Once an outbreak is over, the next time there are 10 consecutive days of growth, we record a *new* outbreak as active. Implement such a detection system and apply it to the case count time series.

(e) Regenerate the three-panel case count plot from (c), but now use markers to indicate the starts of outbreaks. Do these locations match your intuition of when outbreaks are starting?

(f) Finally, repeat steps (b)-(e) for each of the other 5 digital traces. Make the six-panel plot from part (a) again but mark the outbreak locations as in (e). The idea is that if a sufficient number of traces show an outbreak ahead of the case counts, we can predict an imminent outbreak.