Assignment #2 for Public Health Surveillance

**Due by 5pm February 26 (Monday)**. Submit your Word document (not .Rmd or .R files) via Canvas. **You should work on this assignment alone**. Feel free to consult the internet and class notes; do **not** discuss with your classmates. Any questions should be e-mailed to me or Zoe ([daniel.weinberger@yale.edu](mailto:daniel.weinberger@yale.edu); Elisabeth.nelson@yale.edu), and I will respond to the whole class (via Canvas) for any issues requiring clarification. **Your answers should be brief** (1-2 sentences per question) and all answers should be typed in this Word document. This is scored out of 100 points.

**All of the materials for this assignment are on posit.cloud under the Assignment\_2 project**

We have been asked to set up a prospective aberration detection algorithm for diarrheal diseases using **daily** data from the NYC syndromic surveillance system.

1. Start with a Poisson regression model, fit to data for the pre-pandemic period (data frame a2)
   1. Is the Poisson regression appropriate with this dataset?
   2. Which variables did you include in the model, and why?
   3. Plot your fitted model with prediction intervals, with the observed cases, include a screenshot
   4. In addition to a time trend and harmonic variables, are there any other variables you might want to adjust for in the analysis?
   5. Name 2 ways you could improve the sensitivity of this approach
2. Next, fit a robust regression to the data using the same set of covariates as in question 1
   1. How do the prediction intervals compare with the intervals from the Poisson regression? How would this influence the sensitivity of the detection?

Now focus on the post-pandemic period (April 2020 onward) (Data frame a3). We are now using weekly data. The goal is to train the model through 3rd quarter of 2021 and then evaluate its performance from 4th quarter of 2021 onwards.

1. Fit a Poisson model to these data (you can use variable ‘cases\_train’ as the outcome variable for model fitting)
   1. What variables did you include in the regression, and why?
2. Set up a cusum-type algorithm using the glrpois\_App(). The goal here is to have a system operational starting 4th quarter of 2022, so we will use the data from April 2020-September 2021 to train the model, and then you can evaluate performance during October 2021-Feb 2022.
   1. What parameters did you set for this algorithm, and why?
   2. What is an advantage or disadvantage of a CUSUM approach vs the Poisson regression method from Question 3?

Nowcasting

1. You have line-list data for 5000 cases containing the date of diagnosis and the date that the case was reported to DPH. This is in the dataset reported\_cases.rds
   1. Plot a time series of the cases based on date of diagnosis and date of case report
   2. Calculate the delay distribution, and plot the delay distribution and the cumulative delay distribution
   3. Use the delay distribution to correct the reported cases based on date of diagnosis. Plot the reported cases based on day of diagnosis and the corrected cases.
   4. What is the interpretation of this adjusted curve compared to the observed curve?