Biostatistics 6640
Python and R in Data Science
R Data Analysis Project: due October 26 by 5:00 PM

In this project, you will be analyzing weather and (simulated) malaria incidence from Mozambique. The objective of this project is to describe the temporal and spatial variation in these data and to draw preliminary conclusions about the relationships between the variables (i.e., no statistical models are necessary). Information about malaria can be found on the Centers for Disease Control and Prevention's website (`https://www.cdc.gov/malaria/`).

The data for this project can be found in the data folder on Canvas. The file is called MozSyntheticMalaria.csv. There are also shape files for Mozambique districts.

The variables in the data are:

`rain` is the weekly ave rainfall in mm
`rainTot` is the weekly total rainfall in mm
`tavg` is the weekly average temperature in Celcius
`rh(%)` is the relative humidity in %
`sd(mmHg)` is the saturation vapor pressure deficit in mm of mercury (another measure of humidity)
`psfc(hPa)` is the surface barometric pressure (a general indicator of large-scale weather activity and exhibits a strong seasonal cycle)
`Population_UN` is the total population of the district
`u5weight` is the proportion of the total population under 5 years of age
`malaria` is the number of cases under 5 reported that week (this is simulated)
`DISTCODE` is the unique identifier for a district, which can be linked to the shape file
`tabove[]` all of the tabove are number of days that week that temperature was above a threshold (the number next to tabove)
`pabove[]` all of the pabove are number of days that week that rainfall was above a threshold (the number next to pabove)

Malaria incidence in cases per 1,000 population in children under 5 needs to be created. You're given the district population and the proportion of kids under 5 and cases under 5.

Malaria tends to be related to weather (increased rainfall and warmer temperatures, etc) in a lagged fashion. This is because there is a 7-14 day incubation period between exposure 

to an infective bite by a mosquito and the onset of symptoms. The incidence data today are likely related to exposure up to 14 days prior and the effects of weather and temperature, etc, are likely related to exposure at an uncertain time before that. This time is typically thought to be 2, 4 or 8 weeks from the day the person showed up in the health center. You are expected to create the lagged variables and explore their relationships with malaria incidence.

In your report, you should address the following questions, but do not number the questions and write your answer explicitly. We expect you to address other questions as well, but this list should help you get started. Please make the report flow in a professional way, like you're writing a manuscript or a report for the country officials.

1. Which lags are most associated with malaria incidence for temperature and total rainfall?

2. Which region has the most malaria?

3. In which regions is rainfall high? Temperature? How much do these variables vary across the country?

4. Are cases clustered in a particular area of Mozambique? How does this overlap with rainfall and temperature? (think about maps here)

5. Is malaria incidence going down or up over time? Does this depend on which region we're looking at?

Other instructions:

1. Please generate the report using RMarkdown. Submit your RMarkdown document and a rendered file (either pdf or .doc).

2. As part of the project, we would like to you to set up a github repository (`https://github.com/`) to store the code you used for your project.

3. We expect you to produce exploratory analysis figures (histograms, splines, etc), maps of incidence (by district, across years, etc) and explore, at the very least, basic relationships between the independent variables and the outcome (malaria incidence).

Write-up: This project should be completed by all individuals in the class. You may work in groups, but if you do, each person needs to submit their own write-up and please include the names of the individuals with whom you worked. The write-up should include a

background section with a literature review and citations (at least one half page), a description of the problem and data (approximately one half page), results (no page expectation, but there should be 4-5 figures and at least one should be a map), conclusions, references and any supplemental material you choose to include.