STAT 466, Spring 2022

Project 1

Technical Debrief

The Research Question

What are some of the ways that we can characterize environmental conditions that may be explanatory of disease incidence?

Possible ideas are things like degree days above some threshold, consecutive degree days, etc.

The Research Question

What are some of the ways that we can characterize environmental conditions that may be explanatory of disease incidence?

Possible ideas are things like degree days above some threshold, consecutive degree days, etc.

Take 5 minutes to explain to the ~2 people around you what your team did to answer this question for the DE.

The Research Question

What are some of the ways that we can characterize environmental conditions that may be explanatory of disease incidence?

Possible ideas are things like degree days above some threshold, consecutive degree days, etc.

Research Question: Take-Aways

1. Investigating a (possibly explanatory) relationship between (possibly) multiple variables and a single (response) variable.
   - Correlation
   - Regression
2. Response variable sounds quantitative
3. Explanatory variables sound like they could be anything, BUT some specific examples are given.

The Data

- Hourly, daily, and monthly environmental (weather, soil) data
- Final mortality rate data (disease incidence) of common strawberry cultivars
- Cultivar disease resistance numerical category

Screenshots of data files in the slides!

The Data: Take-Aways

- Mortality rate appears to be quantitative
- There are replicates and different varieties in the mortality file, but these variables are NOT present in the hourly, daily, and monthly files
  - This leads me to wonder/guess that the explanatory weather variables do not vary by replicate or variety
  - If the explanatory variables do not vary by replicate, then some sort of aggregation across replicates may be necessary
  - Aggregation across varieties is probably not useful because the slides seem to describe differences in disease incidence across varieties. So, variety might have some explanatory power
- The mortality rate values are at the yearly level
  - This leads me to believe that some sort of aggregation (to the yearly level) will likely be necessary for the explanatory variables
- We have many more weather variables at the monthly level than we do at the hourly and daily levels
  - This may affect how we use the daily and hourly files, especially because Air Temp and Soil Temp appear to be in all three files
- The couple of specific examples of variables to explore are not included in any of the files!
  - The slides describe how to compute these extra variables and so I should be sure to compute these variables myself
- We have data across multiple years and so I should be sure to consider how this temporal aspect should be incorporated, if at all

The Final Approach

- A final, pre-processed dataset (with all of the relevant variables within it) with a row for each variety, season combination
  - Aggregated across replicates (e.g. average)
  - Aggregated to the yearly level (in a variety of ways for the environmental data)
    - Averages
    - Sums
    - Maximums
    - Degree days for multiple thresholds
    - Consecutive degree days for multiple thresholds
  - 40 total rows (5 seasons, 8 varieties)
- Visualization and correlation analysis first
- Follow-up multiple regression

Questions or concerns about this approach?