DATA 512
Part 2 – Extension Plan
Ted Liu

Problem Statement / Motivation:

Smoke inhalation, of varying degrees, can lead to serious respiratory distress and long-term health problems. In the case of smoke from wildfires, it is perhaps unlikely that the amount of smoke individuals inhale will be extremely serious or fatal. However, less serious smoke inhalation can still lead to onset and/or exacerbation of conditions like asthma and heart disease. (https://www.health.ny.gov/environmental/outdoors/air/smoke_from_fire.htm)

Studying the impact of wildfire and smoke impact on the community may help inform new policies for handling wildfires, guidance for the community when wildfires and smoke become an issue, and perhaps create new resources / policies for those who experience negative health outcomes because of the smoke and wildfires.

Impact Focus:

For my extension plan, I hope to investigate the health impacts of wildfires on my specified city of Rochester, NY. Specifically, I aim to investigate how the impact of wildfires on prevalence of reported asthma and high blood pressure in the local population.

Data:

A data set I have sourced is from the Center of Disease Control (CDC). The CDC provides the PLACES dataset in the public domain. The PLACES dataset is which provides model-based estimates of health conditions and behaviors for different locations throughout the United States. Its original purpose is to offer local-level data which can help communities and officials understand the local populations health conditions and risk factors. The PLACES dataset can be found here: https://www.cdc.gov/places/. The data itself can be downloaded as CSVs on the CDC website or other access methods described here: https://www.cdc.gov/places/help/explore-data-portal/index.html#access-data.

The only issue with this dataset is that it does not have the exact city we analyzed in Part 1 – Common Analysis. Rochester, NY is not listed in the data set. But Monroe County, which Rochester is a part of, is listed in the data set. This means for the

extension plan we will need to use the data on Monroe County rather than Rochester for our analysis.

Another dataset we can possibly leverage is the covariate data from the Institute for Health Metrics and Evaluation (IHME) which can be found here: https://ghdx.healthdata.org/record/global-burden-disease-study-2021-gbd-2021-covariates-1980-2021

From this data set, we can use their asthma and blood pressure metric. We can also examine the IHME dataset more thoroughly to determine if there are other relevant metrics we can include in addition to asthma. The data set contains data from 1990 – 2019. The issue with this dataset is that its regional granularity does not go down to the city level, it only goes down to state-level. This means for the extension plan we will need to use the data on all of Ney York State rather than Rochester for our analysis.

Unknowns and Dependencies:

Given that we have not done an in-depth look into all the available data offered in the PLACES dataset, an issue that may arise is a lack of longitudinal data. In other words, the PLACES dataset may not go as far back as the wildfire data does. This will perhaps limit the amount of smoke estimate data we can use from Part 1.

In our quick examination of the PLACES dataset, we verified that there are at least seven annual data points for Monroe / Rochester in the PLACES dataset. This few amount of data points may be problematic when performing any form of statistical analysis. If this does become a problem, perhaps including other counties that neighbor Monroe will resolve the issue – at the cost of being less localized to Rochester.

In the event that the PLACES dataset turns out to be inadequate for our extension plan for answering if smoke impact (through out smoke estimate) affects health outcomes, we will utilize the IHME data set.

Time To Completion:

The extension plan itself is due on November 6th with Part 3 – Project Presentation due on November 27th and the Part 4 – Project Repository due on December 4th. We will

prepare a timeline that begins on the due date of Part 2 – Extension Plan and plan in weekly chunks. This creates the following (business day) week-buckets with the associated assignments:

Week 1: November 4th – November 8th

- Part 2 Extension Plan due on 6th

Week 2: November 11th – November 15th

Week 3: November 18th – November 22nd

Week 4: November 25th – November 29th

- Part 3 Project Presentation due on 27th

Week 5: December 2nd – December 6th

- Part 4 Project Repository due on 4th

Given these week buckets and the associated project due dates, we propose the following timeline and milestones:

Week 1:

Week 1 will primarily consist of acquiring and exploring the PLACES dataset and the completion of Part 2 – Extension Plan. The datasets will likely be acquired by downloading them off of the CDC website. If there is time remaining in the week, then we will also attempt to retrieve the data from the R / Python packages. If there is time remaining, we will begin creating the presentation.

Tasks:

- Turn-in Part 2 Extension Plan
- Download PLACES dataset
- Examine the PLACES dataset
    - How many Years can we get from the dataset?
    - What other relevant features besides asthma are present?
- Download the IHME dataset
- Examine the IHME dataset
    - What does the asthma variable actually represent?
    - Are there more relevant outcomes that we can use?

DATA 512
Part 2 – Extension Plan
Ted Liu

Week 2:

Week 2 will primarily consist of performing exploratory data analysis on the PLACES and IHME dataset and creating a model using our smoke estimate and our health outcomes. We will most likely implement a relative risk regression – also known as a modified Poisson regression – to understand the risk associated with the smoke estimates.

Tasks:

- Perform exploratory data analysis on the two datasets
    - Plot the prevalence of asthma over years
- Create relative risk regression
- Start presentation (if time available)

Week 3:

Week 3 will serve as an overflow week, if the creation of the model is not completed by week 3 then we will use week 3 to complete the model. By week 3, we will, at the very least, begin creating the presentation.

Tasks:

- Start, at the very least, the project presentation
- If the model creation is not implemented yet, complete implementation

Week 4:

Week 4 will primarily focus on completing the project presentation. Week 4 will also mark the beginning of vetting the project repository to ensure it follows all aspects of the best practices for reproducibility.

Tasks:

- Complete project presentation
- Turn in project presentation
- Begin vetting project repository

Week 5:

Week 5 will primarily focus on ensuring the project repository is in an acceptable format that aligns with the best practices of reproducibility.

DATA 512
Part 2 – Extension Plan
Ted Liu

Tasks:

- Complete organizing project repository