Math 154 Semester Project Revised Proposal

1. Group Members: Oliver Cornelius-Knudsen, David Steffen
    a. Group Dynamic Roles: Project Manager Oliver, Task Manager David
    b. Project Roles: Director of Computation Oliver, Director of Research David
2. Title: Who Dies Where, Of What? Why? The Determinants and Predictors of County-Level Mortality Rates in the United States
3. Purpose

Our key research interest is how rates of different causes of death vary between counties in the US. We will be using a dataset of estimated county-level cause-specific mortality rates that enables analysis of patterns of geographic variation in mortality rates. Our model and visualizations will allow policymakers and concerned parties to better understand what causes the relative prevalence of different causes of death in different areas, and how policy choices that affect an area's characteristics may affect the prevalence of causes of death in that area.

Our models will give policymakers a better understanding of the determinants of the rate of deaths from causes such as opioid overdoses, a very relevant consideration to understanding and predicting the likely effectiveness of policy responses to public problems like the opioid epidemic that are currently being developed and are under discussion.

4. What will people learn?

A. Informative: Increase our understanding of what leads to variation in causes of death

B. Actionable/Predictive: Provide a tool for the analysis of the impact of potential policy decisions on the mortality rates and causes of death of county residents.

5. Data

The key dataset for this analysis is a set of estimates of “annual mortality rates by US county for 21 mutually exclusive causes of death from 1980 through 2014” constructed by Laura Dwyer-Lindgren, Amelia Bertozzi-Villa, Rebecca Stubbs, et al. from the National Center for Health Statistics’ National Vital Statistics System and published in the Journal of the American Medical Association, available at https://jamanetwork.com/journals/jama/fullarticle/2592499. We downloaded the dataset itself from Kaggle, which it was provided to by the Institute for Health Metrics and Evaluation and is at https://www.kaggle.com/IHME/us-countylevel-mortality. It is downloaded as an Excel file with a sheet for each cause of death that we will have to wrangle.

Other county-level data used to model and predict cause-specific mortality rates will come from a variety of government data sources. The American Community Survey (ACS) of the United States Census Bureau provides county demographic data such as race, age, and education. The Economic Research Service of the Department of Agriculture provides county-level data on economic indicators including poverty rates, population, unemployment rates, and education levels at https://www.ers.usda.gov/data-products/county-level-data-sets/. These datasets are all Excel sheets that we can import into RStudio or another program.

Finally, county-level electoral data compiled by the New York Times will be taken from a GitHub repository of a large dataset of county-level data on a variety of topics available at https://github.com/Deleetdk/USA.county.data.

We will have to merge all of these datasets as part of our data wrangling process. One initially apparent data complication (of the many that we will likely encounter) that we know of is that there are some small inconsistencies in the codes used to identify counties between the different datasets that we will have to deal with. We will have to figure out what to do with those inconsistencies as those codes are likely what we will be merging the datasets on.

6. Variables

- Mortality Rate
- Cause of Death
- Employment
- Population
- Income per Capita
- Race
- Vote Share Information
- Education / School Enrollment
- Rural/Urban Continuum
- Health Indicators

7. End Product

We will produce predictive models that, supported by visualizations, will increase our understanding of geographical variation in mortality rates across counties, providing insight for policy consideration and implementation.

One type of model we plan to build is a k-Nearest Neighbors model that will provide predictions of the mortality rate due to a given cause in a given county by taking the k nearest neighbor counties in terms of Euclidean distance in the explanatory variables, calculating the average mortality rate due to that cause in those k counties, and then using that average mortality rate as the predicted mortality rate for that county. We can then compare those estimates to the actual mortality rates and see how accurate our model was.

A second type of model we will build is a regression tree for estimates of mortality rates due to each cause of death. Splits in the trees would be based on county values of explanatory variables, and the counties that end up in each node would be given the average mortality rate

of the counties in that node as their estimated mortality rate. Then, as with the other model, we can compare the estimates to the actual mortality rates and see how accurate our model was.

We hope to develop a shiny application that will effectively visualize the change in mortality rates by cause over time. We would offer a drop down menu for cause of death which would light up counties across the United States with a darker color corresponding to a higher mortality rate by that cause of death. Then a slider or automatic animation will show the change in the mortality rates over time thereby illustrating time trends and shocks in mortality. Examples are visualizing rising mortality as a result of the opioid epidemic or rise of HIV/AIDS in the 80s, showing spikes in mortality as a result of natural disasters, or mapping and visualizing overall trends in deaths from gun violence.

Additional shiny applications would include visualizations of how effective our predictive model are at accurately predicting the the mortality rate by cause in different counties. We would include two drop down menus. The first would consist of a selection of possible models which would cause the map to illustrate by county whether we were able to accurately predict the mortality rate for a given cause of death. The second would allow the viewer to select which cause of death they would like to view. The overall map would then illustrate how accurately each of how selected models predict mortality for selected causes of death.

8. Implications

Our project will hopefully inform policy by demonstrating the impact a number of variables that are related to policy have on mortality rates. Variables that we hope to elucidate the impact on mortality rates of that are related to policy choices include education levels, poverty and unemployment rates, and health indicators of various kinds.

Finally, it will also help inform the public by giving them a way to visualize trends and changes in mortality rates that may be otherwise hard to truly comprehend on a countrywide scale.