Lina Stolyar
Intro to Data Science: PPOL 670
Project Proposal: Spring 2020

The goal of my project is to analyze the difference between different U.S. Census Bureau data sources and to see how much they vary in producing the variables of interest related to health policy.

Every year, the United States Census Bureau administers the American Community Survey (ACS) to learn about the state of various demographic information in the U.S. (and Puerto Rico). They generate tables of data from their full sample (a little over 2.4 million responses for the U.S. in 2018) on their website, data.census.gov. These tables provide pre-tabulated and predetermined age categorization, poverty thresholds, etc. For example, they provide information on healthcare coverage for individuals from ages 0 - 6 years old, 6 - 18 years old, 19 - 34 years old, etc.

For more flexibility in analyzing data, such as using different age categorizations or analyzing multi-cross cuts of data (like race, federal poverty level, and age), analysts can use the Integrated Public Use Microdata Sample (IPUMS). The ACS IPUMS utilizes a 1-in-100 national random sample of the population and is weighted (therefore a 1% sample of the ACS total cases).[1] Analysts could also use a final data set known as the Public Use Microdata Sample (PUMS), which includes about two-thirds of the cases used on data.census.gov.[2]

Because it is a smaller subsamples of IPUMS and PUMS, there may be different estimates for the same characteristics that appear on data.census.gov. Given these differences, I would like to explore the difference in calculated rates of uninsurance for children who identify as Hispanic under age 6 in each state from 2008 - 2018. Using this variable allows me to use a few different data slices (age, race, state location, and year), which can affect data reliability with smaller data sets and will allow me to see the differences in calculated rates. In addition, in 2017, Census changed the definition of "child" from those under the age of 18 (0 - 17 years old) to those under the age of 19 (0 - 18 years old). Therefore I would not be able to compare multiple years of data from data.census.gov since their pretabulated groups changed from 0 - 17 years old to 0 - 18 years old. However, they do not change their other predetermined group of children under 6 years old.

Data can be acquired at data.census.gov (specifically Tabel B27001I), IPUMS, and PUMS. I will need to clean the data to categorize insurance categories and race/ethnicity categories. I will also need to make sure that I am weighting the data properly in each data set. I am familiar with how to do that in IPUMS but less so in PUMS so I will consult this document.

[1] https://usa.ipums.org/usa/sampdesc.shtml#us2018a
[2] https://www.census.gov/programs-surveys/acs/technical-documentation/pums/about.html

I would start by comparing the difference in uninsured rates from data.census.gov to IPUMS and data.census.gov to PUMS for each state in 2018. Then, I may pick a big and small state and compare the differences across 2008 - 2018 and visualize this data. My hypothesis is that smaller states will have greater variation in their uninsured rates because of their smaller samples and the difference will be most pronounced in the data.census,gov versus IPUMS comparison.

I would also like to compare the coefficient of variation (CV) for the different data sets. The CV "describes the dispersion of the variable… the lower the CV, the smaller the residuals relative to the predicted value… [suggesting it is] a good fit model."[3] The CV is the ratio of the standard error divided by the sample size (in my case the number of Hispanic children under 6 who are uninsured). There is no industry standard for a CV threshold for when data should be suppressed but I will use two definitions 1) if the total number of Hispanic children under 6 in 2018 in a state plus/minus the margin of error produces a number below zero (i.e. the number of Hispanic children under 6 in the state is below 0), then the sample is not reliable and 2) if the CV is above 25%, then the sample is less reliable.

If there is time left, I would like to run a regression between the total population of a state (independent variable) and the CV (dependent variable) to see if there is a correlation. Perhaps I can use machine learning to see if there is a best fit model for each year of 2008 - 2018. I would like help thinking through what other variables I should include in my regression (and/or how I can determine this).

Success in this project for me will look like:

- Successfully coding all of the data in R
- Successfully keeping track of all variables in each data set
- Successfully weighting the data
- Creating a compelling visualization that illustrates the differences between each data source for the same variable
- Calculating the CV for each data source
- Running one version of machine learning with CV and state size

[3] https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-is-the-coefficient-of-variation/