Preliminary Analysis Write-Up

Allison Collins & Erika Tyagi

11/5/2019

Our repository can be found here: https://github.com/erika-tyagi/clustering-nhanes.

Overview

For our project, we are utilizing the National Health and Nutrition Examination Survey (NHANES), a large-scale survey conducted annually by the Centers for Disease Control and Prevention (CDC) to assess the health and nutritional status of Americans over time. More specifically, NHANES incorporates interviews (including demographic, socioeconomic, and self-reported health-related information), physical examinations (with medical, dental, and physiological components), and additional clinical laboratory tests.

At the outset, our plan had been to dig into the relationship between clinical and self-reported health outcomes, leveraging the different components of this survey, via running cluster analysis to reveal underlying groupings and patterns in the data.

Given the survey has over 13,000 unique variables, we began by referencing the CDC's guidance on the most common chronic illnesses and associated risk factors to narrow the scope. These include diabetes, stroke, Alzheimer's, heart disease, kidney disease, cancer, lung disease – with lifestyle risk factors such as smoking, alcohol consumption, nutrition and exercise choices.

However, as we dug into the data in the period of time following submission of the proposal, we realized that this would likely not yield a feasible project. Over time, the questions asked in the survey have shifted quite a bit, making it necessary to hone in on a more specific area, where we can find comparable clinical and questionnaire data and a large enough set of complete responses to leverage the techniques we have learned in class.

Given the emphasis of NHANES on nutritional outcomes, and the irrefutable link between nutrition and chronic illness, we have decided to focus our efforts on understanding the patterns in peoples' nutritional profiles and the implication this has for health outcomes. We're particularly interested in this in the context of food security and related policy implications.

Work done to-date

Our README in the `process-raw-data` folder explains our workflow for gathering and transforming the data to use in this analysis. We downloaded the raw NHANES survey data (using the Python package `NHANES-Downloader`). This yielded raw (JSON) and XPT) data, along with CSV data for nine surveys. For each of these years, the data were also subdivided into the survey's five component parts: demographics, dietary, examination, laboratory, and questionnaire.

As mentioned above, the data contained over 13,000 distinct variables. From this, we manually subsetted the set of features to include in our analysis to just those variables that were relevant in the context of food security and related health outcomes. The file `NHANES-varnames_yesflag.csv` includes the manually generated binary flag indicating the variables we included.

We then looped over the full set of survey years, components, and files within each component to create a cleaned file where each row represents a unique individual (identified by the SEQN identifier across tables) and survey year combination, and only the subsetted columns are included. `NHANES-clean.csv` contains the final version of our dataset – though again, as we add additional variables, this will grow.

With this dataset, we then began our preliminary analysis. Specifically, we began exploring our dataset and the variables that we originally chose to include leveraging the EDA techniques learned in class. We also explored the clusterability of our data using hypothesis testing (using Hopkins sparse sampling) and visually (using visual assessment of tendency / ordered dissimilarity images). These led to us believe that our data were likely clusterable (potentially with two natural clusters), although these groups may not be incredibly distinct and of high-quality.

We also began applying clustering techniques. Specifically, we ran a k-means algorithm with two clusters and summarized this fit visually. We also used internal validation techniques to evaluate this fit. Looking at three evaluation metrics (Connectivity, Dunn, and Silhouette), the k-means fit was best with 2-3 clusters (when considering anywhere from 2 to 10 potential clusters).

The visuals summarizing our preliminary analysis are contained in the `preliminary-analysis/images` folder in our repository. This includes a scatterplot matrix of the features that we chose to include related to nutritional outcomes and food security, an ordered dissimilarity image, a cluster plot (leveraging the `fviz_cluster` package's partitioning method visualization functionality) for this k-means fit, and a series of scatterplots of the original datasets with colors separating the two clusters.

The scatterplots show that the clusters seem to delineate the two groups in a way that highly correlates to the level of saturated fat (i.e., the two groups are individuals with low vs. high intake of fat).

Plans for completion of analysis

We intend to take the below steps to complete our project:

- We will add to the list of nutrition-related variables in our scope, both from the dietary data files (clinical nutrient levels) and questionnaire, which includes information regarding peoples' food purchase patterns (including consumption of frozen food, fast food, amount spent).

- Additionally, we will layer on demographic information (e.g., income level, race, age, height, weight, geographic location), as well as create markers for different chronic illnesses (e.g. diabetes, heart disease), to better understand the composition of the clusters. For example, does a cluster with higher prevalence of diabetes emerge that has a certain nutritional profile?

- Our initial data exploration suggested that we need to explore thoughtfully removing certain outliers (e.g. variables coded as 99999 or 77777, which are likely some sort of incomplete rather than an intended response). We drop some of these in our preliminary analysis here, but we intend to explore this in a more systemic and thoughtful way over the broader data set.

- Following the additional data aggregation, we will re-run additional clustering analyses, with the goal of better understanding the relationship between these variables. Running the internal validation on a randomly selected sample suggested that k-means would be a good choice with 2-3 clusters but will explore leveraging other clustering methods.

- Given the richness of the survey data, we are also potentially exploring including a broader set of features, and then applying dimension reduction techniques (e.g., PCA) to see how this affects our clusters.

Finally, as food insecurity and nutritional intake have been known to relate to chronic illnesses according to much of the literature review, our ultimate goal is to take learnings from the unsupervised clustering analysis to feed into a supervised machine learning question of predicting disease outcomes based on individuals' profiles.