Guided Meta-Science Project: Nutrient-phenotype associations

Harvard BMI704: Data Science for Medical Decision Making

Arjun K Manrai, Chirag J Patel

In this project, you will use what you've learned during the course to conduct an investigation on modifiable factors, namely dietary nutrients, and their relevance to health and disease. You might have seen a Time magazine or Newsweek article documenting associations between diet and disease/health outcome (e.g. coffee and mortality: http://time.com/5326420/coffee-longevity-study/). In this guided project, we will explore how such evidence arises. **During Lecture 7, we'll conduct a live *meta-analysis* across all the investigations conducted by the class, and using all available participant data from the CDC NHANES!**

All data/analyses should be submitted to your personal home directory on the RStudio Server VM hosted on Google Cloud Platform (accessible at: bit.ly/bmi704). As always, please reach out to Raj (Arjun_Manrai@hms.harvard.edu) or Chirag (Chirag_Patel@hms.harvard.edu) or the TAs via Canvas should you have any questions.

Please complete the following steps **before 11:59pm, Tuesday, March 12**:

1.) Create a directory in your home directory on the RStudio Server (bit.ly/bmi704) called 'project' (we refer to this directory as 'your project directory' below). [To do this, either click 'New Folder' in the GUI or run the mkdir command in the Terminal.]

2.) Select a single nutrient from the CDC NHANES variable list [available at ../project/nutrients.csv].

3.) Select a single phenotype from the CDC NHANES list of phenotypes [available at ../project/phenotypes.csv].

4.) Create a list of relevant demographic variables that might explain a hypothesized relationship between your nutrient and phenotype (e.g. males/females/both, age group (specific group or all), ethnic group (specific group or all)).

5.) Hypothesize about the relationship between your nutrient and phenotype in two ways: (1) direction [positive or negatively associated with your phenotype for your sample with given inclusion criteria] and (2) effect size [beta coefficient from the linear regression phenotype ~ nutrient + covariates]. Document your hypothesis in a file called hypothesis.csv inside your home folder's 'project' sub-directory with the following format:

nutrient,phenotype,hypothesized_direction(+,-),hypothesized_effect_size [where nutrient and phenotype are the variable names from ../project/nutrients.txt and ../project/phenotypes.txt]

6.) Create a brief 1 min. video of yourself describing your hypothesis and why you came to that hypothesis. Save the video in your project directory.

7.) Find 3 research studies from PubMed (non review articles) to justify your hypothesized effect and direction. Put the associated PubMed IDs (PMIDs) in your project directory in a file called 'hypothesis_support.txt' with one PMID per line.

8.) Find 3 research studies from PubMed that call into question your hypothesized effect. Put their PMIDs (one ID per line) in your project directory in a file called 'hypothesis_against.txt' with one PMID per line.

9.) Load the project CDC NHANES dataset from ../project/proj_data.rds (hint: use readRDS command). Note that there are nine NHANES cohorts in this dataset and the SDDSRVYR variable refers to the cohort.

10.) Subsample the NHANES data to have the median of the sample size for the studies in Step 7.). Conduct a survey-weighted analysis using the svyglm function using a single cohort from the NHANES data frame (Recall Assignment 2 on instructions on how to analyze the data and use the WTDRD1 weight variable). Report the beta coefficient from the survey-weighted linear regression. Repeat the analysis for 2 additional phenotypes from the dataset in 9.) and put one nutrient,phenotype combination per line in a new file called effects.csv in your project directory with format:

exposure,phenotype,SDDSRVYR,beta,standard_error, pvalue,svyglm_call (a string of the formula)

10.) Save all annotated code to project.Rmd in your project directory.

11.) Speculate on the concordance.

In lecture, we will analyze all effects across the class together, as well as integrate data across waves of NHANES in a meta-analysis.