STAT 6120 Fall 2019 Data Analysis Project

Due by the end of Dec 12, eastern time. Please email your report to `tianxili@virginia.edu`.

This is a restricted use data set. Do not distribute it or use it for any purpose other than the course project. The dataset is extracted from a survey of people living in a developing country. There are two separate data files. One contains data measured at the household level, the other contains data measured at the individual level. The two files can be linked using the `hhid` variable. This is a longitudinal survey carried out in 9 different years between 1989 and 2011. Each individual and each household has provided data in one or more of the survey waves.

The survey was originally designed to examine the effects of the nutrition and health programs implemented by national and local governments and to analyze how social and economic transformation is affecting the health and nutritional status of a population. The impact on nutrition and health behaviors and outcomes is assessed by changes in community organizations and programs as well as by variation in household and individual economic, demographic, and social factors.

The household file (household.csv) contains the following variables:

- hhid : household id
- commid : community id
- wave : wave (year in which the survey was conducted)
- region : region (categorical codes)
- urban_rural : whether the household was in a rural or urban setting
- hh_size : number of individuals in the household
- dwelling_size : size of the dwelling in square meters
- drinking_water_source
- pay_for_drinking_water : does the household pay for its drinking water
- toilet : does the household have an indoor flush toilet

The following variables are binary variables indicating whether the given item is owned by the household: bicycle, motorcycle, car, bw_tv, color_tv, washing_machine, refrigerator, air_conditioner, sewing_machine, electric_fan, wall_clock, camera, microwave, electric_cooking_pot, pressure_cooker, computer, telephone, dvd, cellphone, satellite_dish.

The individual file (individual.csv) contains the following variables:

- indid : individual id

- hhid : household id
- commid : community id
- wave : survey wave
- carb : average grams of carbohydrate consumed in previous three days
- fat : average grams of fat consumed in previous three days
- kcal : average kilocalories consumed in previous three days
- protein : average grams of protein consumed in previous three days
- region : categorical code for region

Note that some values are missing (blank). Binary variables coded as 8 or 9 should be treated as missing or “other”.

Question: explore the data set to answer **EITHER** of the following questions:

1. What is the relationship between the living condition and other variables?
2. What is the relationship between the nutrition consumption level and other variables?

You can be more specific about the question you want to answer. Then define your own metrics for living condition or nutrition consumption level and use reasonable way to find the answers. You are supposed to use both household level information and individual level information in your analysis. Your report should be at most four pages long, including all graphs and tables. You do not need to submit any code, but your report should include clear descriptions of all the methods that you used. The evaluation will be based on both the quality of analysis and the quality of report.