Assignment 1 – BMI in NHANES
Analysis of Large-Scale Data
2025

Assignment: The goal of this assignment is to write a do-file that computes some basic summary statistics and regressions of BMI using the NHANES. **Be sure to use lots of comments throughout the do-file**. See the "hw_instructions" document for submission details.

If you use generative AI programs to help you complete your homework, you must disclose which program you used for each question (and, briefly, how you used it). For people working in pairs, only submit one assignment per group. List your name(s) and UNI(s) on the first page.

For this assignment, you will use the 2017-2018 NHANES data. You will need the body measures file and the demographics file. Please use the copies available on the NHANES website that we accessed in class.

1. Report summary statistics on body mass index (BMI) for people who are 20 years old or older versus people who are under 20 years old. Include the mean, standard deviation, median, and range.

2. Using BMI, construct an "obesity class" variable that classifies people as underweight, healthy weight, overweight, class 1 obesity, class 2 obesity, and severely obese, using the CDC guidelines to the right. Focus only on people ages 20 and older who are not pregnant (you can assume that people whose pregnancy status "could not be ascertained" are not pregnant). Keep the entire sample in your data set, but set the obesity class variable equal to missing for those who do not meet the inclusion criteria. Report the share of people in each obesity class, including those in the missing category.

BMI Category	BMI Range (kg/m^2)
Underweight	Less than 18.5
Healthy Weight	18.5 to less than 25
Overweight	25 to less than 30
Obesity	30 or greater
Class 1 Obesity	30 to less than 35
Class 2 Obesity	35 to less than 40
Class 3 Obesity (Severe Obesity)	40 or greater

3. Using the "ratio of family income to poverty" variable, create a dummy variable for whether someone is at or below 200% of the federal poverty line (FPL) (this is a common threshold for social programs and subsidies). Report the share of people at or below 200% FPL versus those above, using all subjects.

4. Focusing on people 20 years old or older and non-pregnant, report the mean BMI for those at or below 200% FPL versus those above 200% FPL. Perform a t-test of whether BMI is different for these two groups of people. Report the difference in means and your conclusion about the significance, using alpha=0.05, in 2 sentences or less. [Note: you will need to find the command for implementing a t-test in Stata yourself.]

5. Perform a linear regression of BMI on the poverty dummy variable you created, focusing on the same people as in question 4. Provide the coefficient, p-value, and 95% confidence interval. Compare the result to the t-test from question 4. Briefly (<4 sentences) discuss why the results are or are not the same.

6. Extend the above regression to include a dummy variable for female (vs male), a dummy variable for whether the individual was born in the US (vs another country), education level (categorical), and the age (in years) of the individual as additional independent variables. Report all of the regression coefficients and their 95% confidence intervals.

7. Instead of using BMI as a continuous outcome, use a dummy variable indicating whether an individual is obese (versus not obese), based on the categories you created in question 2. Perform a logistic regression using the same independent variables as in question 6 and the same inclusion/exclusion criteria. [Note: we did not code a logistic regression example in Stata in class. Use the "help" function in stata to figure out how to use the command]. In 1 sentence, interpret the association for people who have the highest level of education versus people who have an education less than 9^{th} grade.

8. One or more of the variables you used in this assignment used "top-coding." Write a brief (1 sentence each) response telling us

(1) *which* variable(s) included in this assignment used top coding and at what value(s),
(2) *why* NHANES used top coding,
(3) *how* top coding can affect analyses and study results, and
(4) *at least one alternative* to top coding that can be used in surveys like NHANES

Hint: reviewing the NHANES codebooks may help.

9. BMI as a measure of health has been criticized. Waist circumference has been promoted as a potential alternative (especially with respect to cardiovascular outcomes). Using all subjects in the 2017-2018 NHANES data, report the number of subjects with valid (i.e., non-missing) waist circumference data and the number of subjects with valid (i.e., non-missing) BMI data. What are some questions you might ask yourself or descriptive analyses you might run when deciding whether the waist circumference variable is a good variable to use in your study? (<= 4 sentences)