HAD 5744: Quantitative Methods I
Dr. Hoagland

Assignment #1

Late assignments will be penalized at the rate of 10% per day

Please submit a knitted R-Markdown file with text and code for this assignment. Don't forget to acknowledge your group members (although each of you must submit your own assignment) and to include your AI statement.

Note: For this problem, we will use publicly available data from the National Health and Nutrition Examination Survey (NHANES), a national survey of US adults. This data includes information on respondent demographics as well as some health measures. Since 1999 approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component of the survey. The health examination is conducted in a mobile examination centre (MEC).

To load the data, run the commands install.packages("NHANES") and data(NHANES). This provides you with a data set of 10,000 respondents to the NHANES(a random sample of the full data). Use ?NHANES to get the data dictionary.

We will work only with the following variables, for ease of interpretation: PhysActiveDays, SleepHrsNight, Gender, Age, Education, HHIncome, Work, BMI, Diabetes, Depressed, and SmokeNow.

Problem 1: Data Cleaning and Univariate Regression. We are interested in understanding the drivers of physical activity. That is, our dependent variable will be the count of days in a week that a respondent is physically active days, PhysActiveDays. Our primary question is: how does a person's sleep impact changes in PhysActiveDays?

a. Why would you care about this question? Write 1-2 paragraphs (total) that motivates understanding these determinants from (a) a physician's perspective, (b) a researcher's perspective, and (c) a policy-maker's perspective.

b. Draw a preliminary DAG (not necessarily in R! Can add in a scan of this later) documenting hypothesized relationships between this outcome and the independent variables listed above. Justify your DAG with text. Where do you think this DAG is imperfect or failing you?

c. Before running any regressions, you will need to clean and summarize your data. Convert all variables to either a numeric or factor form appropriately and deal with missing data. For discrete variables (e.g., Education and Income), decide what groups are relevant and create either dummy or categorical variables. Once this is done, produce a simple summary table reporting (a) the means and standard deviations for continuous variables and (b) the counts and percentages for categorial variables. This table should be well-formatted. Does anything stand out about your sample?

Note: I am indifferent as to whether you use R to construct this table or build it manually in Word.

d. Visualize the relationship between sleep and physical activity in a way that you think is compelling and/or interesting. Report the figure and tell me the story it is showing.

e. Now we will run regressions. First regress PhysActiveDays only on SleepHrsNight. Report the coefficients, standard errors, confidence intervals, p-values, R^2, and sample size in a well-formatted regression table. Interpret the table, noting the economic and statistical significance of the relationship. By how much would sleep need to increase, on average, to produce an additional 0.5 days of physical activity per week?