**BSTT 413 – Introduction to Data Analysis with R**

**Homework 2**

**Due: Monday, June 14th 2021 at 11 PM**

In Homework 1, you were tasked with analyzing NHANES data to explore the relationships between metabolic syndrome (MetS), sleep duration, and a range of demographic variables using the approaches in Smiley et al. (2019) and Moore et al. (2017). In this assignment, we will build on that analysis to further investigate these relationships.

Problem 1: Graphical Analysis (20 pts)

Use the appropriate ggplot2 functions, along with additional supporting methods, to generate the graphs described below. In each case, make sure that the graphs contain appropriate axes labels, titles and captions, legends, and are visually accurate.

1. Univariate Analysis. Generate appropriate graphs to describe the univariate characteristics of the variables MetS (binary), sleep duration, and fasting blood glucose. In the case of the latter, provide a short description of the modality (unimodal or multimodal, ie the number of peaks) and symmetry of the variable. Based on the graph, which measure of centrality is more appropriate: mean or median?
2. Bivariate Analysis. (i) Compare prevalence of MetS across race/ethnicity levels. Does it appear that there is a possible racial disparity in MetS prevalence? Explain briefly. (ii) Compare the distribution of fasting blood glucose in those with MetS and those without by plotting the respective densities on a single figure. What can you glean from this?

Problem 2: Bivariate Inference (20 pts)

In this problem, the primary objectives are to numerically investigate the strength and significance of associations between MetS and predictor variables. In each case, provide the appropriate measure of association and report confidence intervals or p-values where needed. For statistical significance, use alpha=0.05.

1. Create a contingency (2x2) table to summarize the relationship MetS and race/ethnicity. Report the odds ratio (OR) of MetS in each race category versus White (reference level) and the corresponding 95% confidence interval. Which race/ethnicity levels are have significantly disparate odds of MetS versus the reference level?
2. Using t-tests or ANOVA, determine the significance of the association between MetS and (i) sleep duration; (ii) fasting blood glucose. In each case, state your conclusions.
3. Use an appropriate logistic regression model to estimate the OR and 95% confidence interval for a 1 unit increase in fasting glucose. Briefly describe the effect of fasting blood glucose on the odds of MetS (is it protective or a risk factor? Is the association significant?).

Problem 3: Multivariate Analysis (10 pts)

In this problem, your task is to investigate the relationship between MetS (outcome) and sleep duration (primary exposure) **while adjusting for possible confounding variables**.

1. In Smiley et al., the authors show a curvilinear (quadratic) relationship between MetS prevalence and sleep duration. Reproduce similar results using the following steps:
2. Regress MetS on sleep duration using logistic regression and include both linear and quadratic terms (e.g., glm(Y ~ X + I(X^2), data=df);
3. use the predict() function to predict the prevalence of MetS for each subject and append it to the data frame;
4. use ggplot to graph the relationship between predicted MetS prevalence and sleep duration. Do the results appear similar to those in Smiley et al., Figure 1? Briefly explain why or why not.  
   Note: logistic regression and predict can give strange results in the presence of missing values. Therefore, it is strongly recommended that you filter out observations that are missing either MetS or sleep duration prior to modeling.
5. **Bonus question** (optional): Repeat the above analysis and additionally include a ribbon around the plotted line to denote the 95% confidence interval. The predict function can be used to obtain confidence intervals and the geom\_ribbon function can be used to graph them.