**EPI 289. Models for Causal Inference – Spring 1 2023**

**Homework 0 Exercises**

**These exercises are due in lecture on Wednesday, January 25th. Use R to complete the following exercises.**

1. Read in the **nhefs.xlsx** file from the EPI 289 course website. Show your log to demonstrate that the file was successfully assigned.

2. Sort the data set by the variable **seqn**. Print out the ID number, age, and sex for the first 10 observations.

3. Find the mean **systolic blood pressure** and standard error for men and for women.

4. What is the mean, 25th percentile, 50th percentile, 75th percentile, and interquartile range of **weight in 1971 (in kilograms)**.

5a. Using *ifelse* statements, create a new categorical variable corresponding to quartiles of **weight in 1971** as based on the cut-points from Question (4). Give a tabulation of your results.

5b. Create quartiles for **weight in 1971** using *cut* in R. Give a tabulation of your results. Do your results match those of Question (5a)? Why or why not?

6. Using *lm* in R, fit a univariate linear regression model for the outcome **weight in 1971** with **number of cigarettes smoked per day in 1971** as the predictor. Report the parameter estimate for cigarettes smoked per day.

7. Create a cross-tabulation between sex and race.

8. Using *lm* in R, fit a multivariate linear regression model for the outcome **weight in 1971** with **age, sex, and race** as the predictors. From this model, print the observed and predicted values of **weight in 1971** for the first 5 observations. What is the predicted value of weight in 1971 for an individual of age 40, female, and of Black or other race/ethnicity?

9. Fit the same model from Question (8) using *glm* in R and compare your results.

10. Using *glm* with family specified as binomial in R, fit a multivariate logistic regression model for the outcome **asthma diagnosis in 1971** with **age, sex, race, and usual physical activity status (var active)** as the predictors. Print the predicted probabilities of asthma diagnosis for the individuals with the first 5 ID numbers.

11. (Optional) Create a graph that plots **systolic blood pressure** on the Y-axis and **usual physical activity status (var active)** on the X-axis.