STAT 796: Homework 5

Due **Monday**, February 25 at 11:59pm on Canvas.

This homework assignment asks you to investigate the relationship between tobacco use and esophageal cancer. Please provide your R code in an Appendix at the end of your responses.

1. For this question, use the case-control data on esophageal cancer in "long" format provided on Canvas (`esoph_cancer.csv`)
    a. Fit a logistic regression model that includes tobacco consumption as an unordered categorical variable, and also adjusts for age and alcohol use (also as unordered categorical variables).
    b. Is there statistical evidence that tobacco use is associated with esophageal cancer, when accounting for age and alcohol consumption? Answer with a summarizing statement that includes the results of a hypothesis test.
    c. Provide an estimate and 95% confidence interval for each of the odds ratios that correspond to coefficients of the tobacco terms in your model in (a).
    d. Provide the estimated odds ratio for esophageal cancer, comparing individuals who consume 35 g/day tobacco to those who consume 15 g/day and are the same age and consume the same amount of alcohol.
    e. Provide the estimated probability of esophageal cancer for someone who is 30 years old, and does not consume alcohol or tobacco. If this cannot be calculated, explain why not.
2. Is there a linear relationship between tobacco use and esophageal cancer? Use the same data as Question 1 to answer the following.
    a. Fit a logistic regression model that includes tobacco consumption as a grouped linear variable, and also adjusts for age and alcohol use (as unordered categorical variables).
    b. Is there statistical evidence that tobacco use is linearly associated with esophageal cancer, when accounting for age and alcohol consumption? Answer with a summarizing statement that includes the results of a hypothesis test.
    c. Provide an estimate and 95% confidence interval for the odds ratio that corresponds to the coefficient of the tobacco term in your model in (a).
3. Consider the differences between the models in 1 and 2.
    a. In your own words, explain the differences between the models in 1(a) and 2(a). How are the interpretations of the odds ratios different?
    b. Which model fits the data better? Describe what statistical test can answer this question, and then provide the results from the test.
4. Fit the model from Question 1 using the compact version of the dataset, which is in the file `esoph_cancer_short.csv`.
    a. Are the estimates of the coefficients the same as in 1(a)?
    b. Are your conclusions any different fitting the model using this form of the data?