Stat 6021: Guided Question Set 10

1. For this question, we will continue using the Western Collaborative Group Study (WCGS) data set, which is from a study regarding heart disease. Data are collected from 3154 middle-aged males in California. Download the file "wcsg.csv" and load it into R. In the previous guided question set, we focused on predicting the likelihood of getting a heart attack based on the following predictors:

   - *age*. Age in years
   - *sbp*. Systolic blood pressure in mm Hg
   - *dbp*. Diastolic blood pressure in mm Hg
   - *ncigs*. Number of cigarettes smoked per day, on average.

   The response variable is *chd69*, with a '1' indicating the person developed coronary heart disease, and a '0' indicating the person did not develop coronary heart disease.

   (a) Refer to your answer from the previous guided question set, which predictors did you use to fit a logistic regression model?

   (b) Validate your logistic regression model using an ROC curve. Randomly split your data set into a testing and training data set, of equal size. For consistency of results among all groups, use `set.seed(199)`. What does your ROC curve tell you?

   (c) Find the AUC associated with your ROC curve. What does your AUC tell you?

   (d) Create a confusion matrix using a cutoff of 0.5. Create another confusion matrix using a cutoff of 0.1. Are these values surprising? What do you think is going on here?

2. For this question, we will use a data set containing information regarding housing in Boston. The data set, Boston, comes from the MASS package in R. We will focus on predicting whether a tract in Boston can be classified as a low-, medium-, or high-crime area, based on two predictors, the weighted distance of the tract from five Boston employment centers, and the student-teacher ratio in the tract.

(a) The variable *crim* is the per capita crime rate of the town that the tract is in. Create a new variable that categorizes *crim* in the following manner. Define a tract to have a

- low crime rate, if its crime rate is less than the median crime rate for this data set
- medium crime rate, if its crime rate is between the median and 75th percentile of the crime rate for this data set
- high crime rate, if its crime rate is higher than the 75th percentile of the crime rate for this data set

(b) Fit a multinomial logistic regression model to predict whether a tract is a low-, medium-, or high-crime area using the variables *dis* and *ptratio*, the weighted distance of the tract from five Boston employment centers and the student-teacher ratio in the tract, respectively.

(c) Compute the Wald statistics and p-values associated with the regression coefficients.

(d) Interpret the results of the Wald statistics associated with the two predictors contextually.