Stat 6021: Guided Question Set 9

The Western Collaborative Group Study (WCGS) is one of the earliest studies regarding heart disease. Data were collected from 3154 middle-aged males in California. Download the file "wcsg.csv" and load it into R. We will focus 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.

1. Before fitting a model, create some graphical summaries to see if there is a difference in the distributions of the predictors among people who did and did not develop coronoray heart disease. To create a boxplot, type `boxplot(age~chd69)` to see the distribution of the variable *age* differs between people who did and did not develop coronary heart disease. Create similar boxplots for the other variables. Which variables seem to differentiate those who developed and did not develop coronary heart disease?

2. Use R to fit the logistic regression model using all the predictors listed above, and write the estimated logistic regression equation.

3. Interpret the estimated coefficient for *ncigs* in context.

4. What are the estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 mm Hg, diastolic blood pressure of 70 mm Hg, and does not smoke? What is this person's corresponding probability of developing heart disease?

5. Carry out the relevant hypothesis test to check if this logistic regression model with the four predictors is useful in estimating the odds of heart disease. Clearly state the null and alternative hypotheses, test statistic, and conclusion in context.

6. Based on the Wald test, is diastolic blood pressure a significant predictor of heart disease, when the other three predictors are already in the model?

7. Suppose a co-worker of yours suggests fitting a logistic regression model without the two blood pressure variables. Carry out the relevant hypothesis test to check if this model without the blood pressure variables should be chosen over the previous model with all four predictors.

8. Based on all the analysis performed, which of these four predictors would you use in your logistic regression model?