Categorical Data Analysis

(STAT343)

Assignment 4

NOTE: You can use SAS and/or R to answer the following questions.

1. (GLM for binary data with categorical predictors) The following table displays the coronary artery disease data. The response, presence of coronary artery disease (CA), is dichotomous, as are the explanatory variables, gender and electrocardiogram (ECG) measurement. The primary interest is to investigate the effect of ECG on CA.

		CA		
Gender	ECG	Disease	No Disease	Total
Female	< 0.1	4	11	15
Female	≥ 0.1	8	10	18
Male	< 0.1	9	9	18
Male	≥ 0.1	21	6	27

   (a) Calculate and interpret the marginal OR between CA and ECG, and the conditional ORs between CA and ECG at each level of gender.

   (b) Investigate the independence of CA and ECG by fitting logistic regression models to the data with and without gender.

2. (GLM for binary data including model selection, model checking, inference and prediction) Consider the horseshoe crab data with width (W), color (C), and spline (S) as predictors and whether a female has a satellite or not as a response.

   (a) Select an optimal model by a selection procedure and explain your selection procedure.

   (b) Check residuals and influential points.

   (c) Interpret the model parameters and odds ratios from your selected model.

(d) Predict the responses by your selected model, and then obtain classification tables and calculate the sensitivity, specificity, and prediction accuracy for each of cutoffs 0.1, 0.5, k, and 0.9, where k = the sample proportion of 1's for the response variable. Draw a ROC curve and obtain AUC.