Example for Chapter 8 Stat286

1. In November 2005, Michelin published its first ever guide to hotels and restaurants in New York City. On the other hand, The *Zagat Survey 2006* is purely based on views submitted by customers using mail-in or online surveys. We compare the two restaurant guides to the 164 French restaurants that are included in the *Zagat Survey 2006*. We want to predict θ, the probability that a French restaurant in included in the 2006 *Michelin Guide*, based on customer views from the *Zagat Survey 2006*. The data is saved in the file **MichelinFood.txt**. (from the book website)
   1. Open the data, and look at the arrangement of the data.
   2. Find the logistic regression that uses Food Rating to predict the probability that a French restaurant in included in the 2006 *Michelin Guide*.
   3. Interpret the slope.
   4. Do you think the logistic model is appropriate?
   5. Do the data provide strong evidence that the slope is 0?
   6. Find the Pearson residuals and Deviance residual.
2. Still consider the Michelin and Zagat guide to New York restaurants. This time read in the data: **MichelinNY.csv**. (from the book website)
   1. How many variables are there? Look at the first 10 observations, compare this data to the data in Example1.
   2. Make a graphic comparison for the Food rating, for those in the Michelin guide and those not in Michelin guide. What can you conclude?
   3. Find the logistic regression that uses food rate to predict whether a restaurant is in the Michelin guide or not.
   4. Make the Pearson residual and Deviance residual plots.
   5. Using boxplots to look at the conditional distribution of predictors. What can you conclude?
   6. Get the equation for multiple logistic regression.
   7. Find the Marginal Model Plots. Do you think the model is valid?
3. The personnel director for a firm that manufactures computers has classified the performance of each of the employees in a certain position as either satisfactory or unsatisfactory. The director has two tests that she would like to use to help to determine which future employees will perform in a satisfactory manner and which will not before they are assigned to the position. With this knowledge, she will be better able to suggest jobs to the position within the firm. To help determine whether tests will be useful, she administers them to the current employees and records their scores. Data set **TRAIN10.csv** contains whether the employee is currently classified as satisfactory (1) or unsatisfactory (0). The next two columns are the test score results on each of the two tests. Your job is to help the director to build a model uses the test scores to predict the correct classification of each future employee.
4. The case of **Maloley v. National Revenue Service of Canada** concerned disparate impact on women. The Revenue Service of Canada used General Intelligence Test with a higher than usual passing score as part of screening process for collection clerks (Juriansz, 1986). In the selection process that followed, 68 of 115 men (59%) and 68 of 251 (27%) women passed the exam. (See Table 1 of Gastwirth, Krieger and Rosenbaum, 1994.) The Revenue Service claimed that the difference between the passing rates for females and males is due to the different education levels. The bureau provided the following education data to the court: for those who took the test after the department raised the passing score, 60 of the 115 (52%) males and only 63 out of the 251 (25%) females had some college education (See Table 2 of GKR, 1994). And the Appeals Board agreed even through the Revenue Service did not offer passing rates data by gender and education into evidence. As a statistician, do you think the data is strong enough to justify the significant passing rates between female and male? Analyze the data with your classmates.