1.6.3 The Prostate Cancer Study

A third data set involves a study of patients with cancer of the prostate. These data have been provided to us by Dr. Donn Young at The Ohio State University Comprehensive Cancer Center. The goal of the analysis is to determine whether variables measured at a baseline exam can be used to predict whether the tumor has penetrated the prostatic capsule. The data presented are a subset of variables from the main study. Of

the 380 subjects considered here, 153 had a cancer that penetrated the prostatic capsule. Actual observed variable values have been modified to protect subject confidentiality. These data will be used primarily for exercises. A code sheet for the variables to be considered in this text is shown in Table 1.7.

Table 1.7 Code Sheet for the Prostate Cancer Study

Variable	Description	Codes/Values	Name
1	Identification Code	1 - 380	ID
2	Tumor Penetration of Prostatic Capsule	0 = No Penetration 1 = Penetration	CAPSULE
3	Age	Years	AGE
4	Race	1= White 2 = Black	RACE
5	Results of the Digital Rectal Exam	1 = No Nodule 2 = Unilobar Nodule (Left) 3 = Unilobar Nodule (Right) 4 = Bilobar Nodule	DPROS
6	Detection of Capsular Involvement in Rectal Exam	1 = No 2 = Yes	DCAPS
7	Prostatic Specific Antigen Value	mg/ml	PSA
8	Tumor Volume Obtained from Ultrasound	cm^3	VOL
9	Total Gleason Score	0 – 10	GLEASON

EXERCISES

1. In the ICU data described in Section 1.6.1 the primary outcome variable is vital status at hospital discharge, STA. Clinicians associated with the study felt that a key determinant of survival was the patient's age at admission, AGE.
   (a) Write down the equation for the logistic regression model of STA on AGE. Write down the equation for the logit transformation of this logistic regression model. What characteristic of the outcome variable, STA, leads us to consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between STA and AGE?
   (b) Form a scatterplot of STA versus AGE.
   (c) Using the intervals [15, 24], [25, 34], [35, 44], [45, 54], [55, 64], [65, 74], [75, 84], [85, 94] for AGE, compute the STA mean over subjects within each AGE interval. Plot these values of mean STA versus the midpoint of the AGE interval using the same set of axes as was used in Exercise 1(b).
   (d) Write down an expression for the likelihood and log likelihood for the logistic regression model in Exercise 1(a) using the ungrouped, $n = 200$, data. Obtain expressions for the two likelihood equations.
   (e) Using a logistic regression package of your choice obtain the maximum likelihood estimates of the parameters of the logistic regression model in Exercise 1(a). These estimates should be based on the ungrouped, $n = 200$, data. Using these estimates, write down the equation for the fitted values, that is, the estimated logistic probabilities. Plot the equation for the fitted values on the axes used in the scatterplots in Exercises 1(b) and 1(c).
   (f) Summarize (describe in words) the results presented in the plot obtained from Exercises 1(b), 1(c), and 1(e).
   (g) Using the results of the output from the logistic regression package used for Exercise 1(e), assess the significance of the slope coefficient for AGE using the likelihood ratio test, the Wald test, and, if possible, the Score test. What assumptions are needed for the p-values computed for each of these tests to be valid? Are the results of these tests consistent with one another? What is the value of the deviance for the fitted model?

(h) Using the results from Exercise 1(e) compute 95 percent confidence intervals for the slope and constant term. Write a sentence interpreting the confidence interval for the slope.

(i) Obtain the estimated covariance matrix for the model fit in Exercise 1(e). Compute the logit and estimated logistic probability for a 60-year old subject. Compute a 95 percent confidence intervals for the logit and estimated logistic probability. Write a sentence or two interpreting the estimated probability and its confidence interval.

(j) Use the logistic regression package to obtain the estimated logit and its standard error for each subject in the ICU study. Graph the estimated logit and the pointwise 95 percent confidence limits versus AGE for each subject. Explain (in words) the similarities and differences between the appearance of this graph and a graph of a fitted linear regression model and its pointwise 95 percent confidence bands.

2. Use the ICU Study and repeat Exercises 1(a), 1(b), 1(d), 1(e) and 1(g) using the variable "type of admission," TYP, as the covariate.

3. In the Low Birth Weight Study described in Section 1.6.2, one variable that physicians felt was important to control for was the weight of the mother at the last menstrual period, LWT. Repeat steps (a) – (g) of Exercise 1, but for Exercise 3(c) use intervals [80, 99], [100, 109], [110, 114], [115, 119], [120, 124], [125, 129], [130, 250].

(h) The graph in Exercises 3(c) does not look "S-Shaped". The primary reason is that the range of plotted values is from approximately 0.2 to 0.56. Explain why a model for the probability of low birth weight as a function of LWT could still be the logistic regression model.

4. In the Prostate Cancer Study described in Section 1.6.3, one variable thought to be particularly predictive of capsule penetration is the prostate specific antigen level, PSA. Repeat steps (a) – (g) and (j) of Exercise 1 using CAPSULE as the outcome variable and PSA as the covariate. For Exercises 4(c) use intervals for PSA of [0, 2.4], [2.5, 4.4], [4.5, 6.4], [6.5, 8.4], [8.5, 10.4], [10.5, 12.4], [12.5, 20.4], [20.5, 140].