PubH 7400–2 Biostatistics Methods and Modeling

Homework #4 - due 03/06/06 **25 points**

1 Logistic regression analysis

Investigate the data in the Prostate Cancer Study with 380 observations and 9 variables as described on page 25 of Hosmer and Lemeshow (2000) Applied Logistic Regression: 2nd Edition. The variables are:

1. Identification Code, 1 - 380, id
2. Tumor Penetration of Prostatic Capsule, 0 = No and 1=Yes, tumorpen
3. Age in years, age
4. Race, 1= Caucasian and 2 = African-American, race
5. Results of the Digital Rectal Exam, 1 = No Nodule, 2 = Unilobar Nodule (Left), 3 = Unilobar Nodule (Right), and 4 = Bilobar Nodule, dpros
6. Detection of Capsular Involvement in Rectal Exam, 1 = No and 2 = Yes, dcaps
7. Prostatic Specific Antigen Value in mg/ml, psa
8. Tumor Volume Obtained from Ultrasound in cm^3, tumorvol
9. Total Gleason Score from 0 - 10, gleason

The dataset is accessible from the course website,
http://www.biostat.umn.edu/∼tracyb/ph7400/pros.dat, which is space-delimited.

1.1 Analysis

For all analysis below, use the tumor penetration variable as the outcome of interest.

- Fit the full additive model (Model 1):

  $$\begin{aligned} \log \frac{\pi}{1-\pi} &= \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{race} + \beta_3(\text{dpros} = 2) + \beta_4(\text{dpros} = 3) \\ &+ \beta_5(\text{dpros} = 4) + \beta_6 \times \text{dcaps} + \beta_7 \times \text{psa} + \beta_8 \times \text{tumorvol} + \beta_9 \times \text{gleason} \end{aligned}$$

  and a model containing only the statistically significant covariates (Model 2):

  $$\begin{aligned} \log \frac{\pi}{1-\pi} &= \beta_0 + \beta_1(\text{dpros} = 2) + \beta_2(\text{dpros} = 3) + \beta_3(\text{dpros} = 4) + \beta_4 \times \text{psa} \\ & +\beta_5 \times \text{tumorvol} + \beta_6 \times \text{gleason} \end{aligned}$$

  Compare the predictive power of these two models by plotting ROC curves for the fitted values of each. Comment on this comparison and suggest which of the two models should be used for prediction.

- Examine the variance assumption for Model 2. Is the model over- or underdispersed?

- Examine model diagnostics for Model 2. Are the residuals approximately normal?

Your complete analysis write-up should be no more than three typed pages, double spaced, 11pt or bigger font. Your writing should be in several, smoothly flowing paragraphs; it should not be a bulleted list. Include tables and plots within this summary.

Attach your SAS code and output as an appendix to the back of your summary. Your summary, however, should NOT refer to any output, tables, or plots attached to the back. If you want a table or plot in your summary, it must appear within the summary.

Here is the data step in SAS code to import the dataset:

```
DATA pros;

INFILE "pros.dat" FIRSTOBS=1;

INPUT id tumorpen age race dpros dcaps psa tumorvol gleason;

RUN;
```