26 BE 7023 & 26 PH 7023: Advanced Biostatistics

Autumn, 2017

MB Rao

Homework Sheet No. 5 Due Date: October 11, 2017 Maximum points:30

Logistic Regression for Grouped Data

Altman (1991) surveyed a sample of 433 subjects from a certain community with the focus on hypertension. For each subject in the sample, he ascertained responses on the following variables.

X1: Smoking – Yes or No

X2: Obesity – Yes or No

X3: Snoring – Yes or No

Y: Hypertension – Yes or No

His objective is to understand how Hypertension depends on Smoking, Obesity, and Snoring. The response variable and the predictors are all binary. He summarized the results of the survey as follows.

Smoking Obesity Snoring Total Hypertension

No No No 60 5

Yes No No 17 2

No Yes No 8 1

Yes Yes No 2 0

No No Yes 187 35

Yes No Yes 85 13

No Yes Yes 51 15

Yes Yes Yes 23 8

We want to model the probability distribution of Hypertension as a function of Smoking, Obesity, and Snoring.

1. Postulate the logistic regression model. 2 points
2. Fit the model to the data. Exhibit the output. Write the prediction model. 5 points

(Technically, the data should have been reported in 433 rows and 4 columns. The data has been summarized as reported above. This is possible because all variables are categorical (binary). R can be used to fit the logistic regression model. Enter the data into an R console column by column as reported. Try the code: < glm(cbind(Hypertension, Total-Hypertension) ~ Smoking + Snoring + Obesity, data = Yourdata, family = binomial).)

1. Check the adequacy of the model. 3 points
2. Check the significance of the covariates. 6 points
3. Obtain the odds ratios of Hypertension vs Smoking, Hypertension vs Obesity, and Hypertension vs Snoring. 6 points
4. Obtain predicted probabilities as per the fitted model. 3 points
5. Plot the Probability(HypertensionYes) for different scenarios of Smoking, Obesity, and Snoring. Comment on the graph. 5 points