Course EPIB-621 - Data Analysis for the Health Sciences

Assignment 4

1. There is a data set (adapted from Hosmer and Lemeshow) on the course web site called drugfree.txt. The variables contained in this data set are described in the table below:

Description	Code	Variable Name
Age at Enrollment	Years	age
Beck Depression Score at Admission	0.000-54.000	beck
IV Drug Use History at Admission	1 = Never, 2 = Previous 3 = Recent	ivhx
Number of Prior Drug Treatments	0-40	ndrugtx
Subject's Race	0 = White 1 = Other	race
Treatment Randomization Assignment	0 = Short 1 = Long	treat
Remained Drug Free for 12 Months	1 = Remained Drug Free 0 = Otherwise	drug.free

(a) The main outcome is the drugfree variable. For all continuous variables (age, beck, ndrugtx), present descriptive statistics within the two subgroups defined by drug.free = 1 versus drug.free=0.

(b) Similarly, for all dichotomous (race, treat) or trichotomous (ivhx) variables, present proportions within each category within the two subgroups defined by drug.free = 1 versus drugfree=0.

(c) Run a univariate logistic regression for each of the six independent variables. In each analysis, report the odds ratio with confidence interval. Remember to declare your factor variables.

2. Continuing the same example as above, run a multivariate logistic regression including all six variables.

3. Using your results from the first two questions, Create a table comparing all odds ratios and their confidence intervals between the univariate and multivariate models. Do you see any evidence for confounding? For each set of possibly confounded variables, check the correlation (if two continuous variables are involved) and create a table (if two categorical or a categorical and a continuous variable are involved). In this way, check that the preconditions for confounding are present (i.e., confounded variables are related to each other, and both are related to the outcome of interest).

4. Following the example in the class notes concerning creating a (Hosmer-Lemeshow) graph of predicted versus observed results, divide your fitted values from the full multivariate model into about 20 categories, from lowest predicted probabilities to highest. Create a plot (see plot in class notes for example) of average predicted probabilities versus observed event rates within these 20 subgroups of subjects. Comment on how well the model fits overall, based on this graph.

5. In this question we will repeat number 2, but from a Bayesian viewpoint using WinBUGS. Run a multivariate logistic regression of the full model (i.e., using all six independent variables). In your WinBUGS program, add lines that calculate all odds ratios, and create a probability prediction for the outcome (drug.free) for someone who is aged 30, has a beck depression score of 10, has a recent history of IV drug use, is white, no prior drug treatment, and had a short treatment scheme.

Note: To create the WinBUGS data set, you need to add [ ]'s after each variable name. You also need to split the trichotomous variable ivhx into two separate dummy variables. This is done for you in the data set called drugfree.bugs.txt. The 1's have all been recoded as 0's, while ivhx2 has all 2's coded as 1, and inhx3 has all 3's coded as 1's.