Biostat 513: **Exercise Set #5** DUE: May 17, 2002

Reading: • Kleinbaum, *Logistic Regression* – Chapter 6
◦ Kleinbaum, *Logistic Regression* – Chapter 7 (optional)

NOTE: Unless explicitly stated, direct computer output is **not** desired. Typically only part of the computer output is asked for (such as a confidence interval) and then proper interpretation of the statistics is requested.

DATA: The data for these exercises can be found on the class web page:
`http://courses.washington.edu/b513/` in the *Homeworks* directory (ie. click on Homeworks from the main Biostat 513 page).

Logistic Regression Estimation and Interpretation

"The full power of the regression approach to case-control studies is obtained when continuous risk variables are analyzed in the original form in which they were recorded, rather than by grouping into intervals whose endpoints are often arbitrarily chosen. "
Breslow and Day (1980) page 227

1. Let's look at a version of the Tuyns et al. (1977) data that has continuous variable measurements (`TuynsC.dat`). Recall that the goal of the study is to characterize the cancer risk associated with both alcohol and tobacco consumption. In many applications `age` is a potential confounder.

** Be careful to drop any cases with missing values (coded 99 for TOB) ***

Note: use the variables named "Tobacco Amount" = TOB, and "Total Alcohol" = ALC.

(a) Fit the logistic regression model that uses dummy variables for the age categories and ALC and TOB as continuous variables entered as linear terms. That is,

$$\text{logit}[\pi(X)] = \alpha + \sum_{j=2}^{6} \beta_j \text{AGE}(j) \; + \; \beta_7 \text{ ALC} \; + \; \beta_8 \text{ TOB}$$

Give interpretations for the estimates $\widehat{\beta}_7$ and $\widehat{\beta}_8$ (ie. convert to odds ratio summaries for fixed comparisons of each exposure – don't forget to mention what is "controlled").

(b) Check the assumption of a linear increase in the log odds for each variable by considering: a model that adds ALC^2; and a model that adds TOB^2 (don't add both). Should we include either variable as a quadratic function? Support your conclusion with appropriate statistics.

(optional) The rationale for using the total alcohol in the earlier analysis is the belief that the alcohol content of the beverages is responsible for the apparent association with cancer rather than some other characteristic such as impurities. The data set gives a breakdown of the specific types of alcohol consumed: beer; wine; cider; aperitif; and digestive. Explore these variables (what do they drink? and how much?) and the relationship between specific alcohol exposures and disease. Instead of ALC, use a logistic regression model with linear terms for each type of alcohol (Use a model with BEER, WINE, CIDER, APERITIF, and DIGEST – five exposures in place of the total, ALC). Interpret these parameter estimates and compare them to the model in 2(a).

Logistic Regression and Fitted Probabilities

2. The Framingham study (Dawber, Kannel, and Lyell 1963) enrolled 5209 subjects in 1948 from a small Massachusetts town. Biennial exams were conducted for blood pressure, serum cholesterol, and weight. Baseline data and 30 year follow-up data is available on the class web page. The major endpoints include occurrence of coronary heart disease (CHD) and deaths from: coronary heart disease (CHD) including sudden death (MI); cerebrovascular accident (CVA); cancer; and other causes. This study was a major stimulus for the development of logistic regression:

> "It it the function of longitudinal studies, like that of coronary heart disease in Framingham, to investigate the effects of a large variety of variables, both singly and jointly, on the effects of disease. The traditional approach of the epidemiologist, multiple cross-classification, quickly becomes impractical as the number of variables to be investigated increases. Thus if 10 variables are under consideration, and each is to be studied at only 3 levels... there would be 59,049 cells in the multiple cross-classification." Truett, Cornfield and Kannel (1967)

The goal of our analysis is to construct a logistic regression model that quantifies the risk of death due to CHD for male subjects over the age of 40. The "do" file on the web page reads the data, recodes missing values, and subsets to this analysis group. The outcome variable for our analysis is `chd`. We will consider the covariates (predictors) `age`, `smoke`, `dbp`, `sbp`, `chol1`, and `weight` or `bmi`.

(a) Using categorized variables explore the "dose-response" relationship by summarizing

odds ratios, or log odds ratios that compare each covariate level to a reference level. Summarize the patterns suggested for each covariate of interest. (see the "do" file for an example with `bmi`).

(b) Now use the original covariate (continuous scale) to characterize the disease risk associated with different values of the covariate. In particular, consider a linear logistic regression model and compare this to a model that also includes a quadratic term ($\beta_0 + \beta_1 X + \beta_2 X^2$), and a model that includes a square root term ($\beta_0 + \beta_1 X + \beta_2 \sqrt{X}$).

(optional) Consider using linear splines or fractional polynomials to guide determination of an appropriate dose-response relationship. See the "do" file for illustration with `bmi`.

(c) Based on the univariate relationships determined in (a)-(b), construct a single multivariate logistic regression model using all of the covariates. Interpret individual coefficients in this model for predictors that you use in a simple linear fashion. Calculate predicted probabilities and cite the predicted probability of death due to CHD for a few key individuals to illustrate the range of predictions that are obtained and to indicate who appears at high (or low) risk of CHD death.

(d) Use the STATA options `lstat`, `lsens`, and `lroc` to summarize the ability of your logistic regression model to discriminate subjects that are likely to die from CHD from those subjects that do not. Provide interpretation of a few key summaries provided by these commands.