Biostat 513: **Exercise Set #4** DUE: May 6, 2002

- Reading: Kleinbaum *Logistic Regression* Chapters 3 & 5
  - *optional*: Kleinbaum *Logistic Regression* Chapters 4

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).

Note: We can use logistic regression for case-control data to obtain valid odds ratio estimates (I will justify this in lecture).

Logistic Regression

1. The data `athens.dat` was analyzed in Exercise #2. Define the variables from that analysis as follows:

   Y = CHD case =1 / control =0 (the outcome variable)
   NEWIRON = 1 if >350mg/month iron / 0 if ≤350mg/month iron
   FEMALE = 1 if female / 0 if male
   AGE = identifier for age group, 1 through 4

(a) Write a logistic regression model using Y as the response and NEWIRON as the predictor of interest.

(b) Interpret the coefficient of NEWIRON in this model in terms of an odds ratio comparison.

(c) Use STATA to fit the model and test whether the NEWIRON coefficient is significant.

(d) Write a logistic regression model that uses NEWIRON, FEMALE, and AGE dummy variables as the predictors.

(e) Interpret the coefficient of NEWIRON in this model in terms of odds ratio comparisons.

(f) Interpret the coefficient of AGE(3) (ie. the indicator variable for the third age group) in this model in terms of odds ratio comparisons.

(g) Interpret the coefficient of AGE(4) (ie. the indicator variable for the fourth age group) in this model in terms of odds ratio comparisons.

(h) In terms of your logistic regression model coefficients, what would be the odds ratio comparing (NEWIRON=0, FEMALE=0, AGE=4) to (NEWIRON=0, FEMALE=0, AGE=3)?

(i) Fit logistic regression models with FEMALE, and AGE dummy variables (no NEWIRON) and then the model with NEWIRON, FEMALE and AGE dummy variables. Use these models to perform a likelihood ratio test of the coefficient of NEWIRON. State the null hypothesis, test statistic, and interpret the p-value.

(j) Compare the estimated NEWIRON odds ratio obtained using logistic regression to that obtained via Mantel-Haenszel in Exercise #2. Do these odds ratios have the same interpretation? Justify.

(k) Are the AGE variables important variables to include in the model? Justify.

(l) Is FEMALE an important variable to include in the model? Justify.

(m) Q: Is the exposure odds ratio the same for men and women? Formulate a logistic regression model that allows the NEWIRON odds ratio to depend on FEMALE. In terms of the regression coefficients in this model, represent the null hypothesis that the odds ratios are homogeneous across gender.

(n) Use logistic regression to answer (m). Compare this result to your use of Mantel-Haenszel homogeneity (interaction) tests from exercise #2.

Logistic Regression

2. Let's continue the analysis of the data collected by Tuyns et al. (1977) (`NewTuyns.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. In class we looked at case/control status (Y) and AGE (in the file `NewTuyns.do` for HW#2 – output is on the web page for HW#3) and found an association between AGE and disease.

In HW#3 we also looked at the association between ALC and disease (Y) by considering a dichotomization of ALC and then adjusting for AGE. Here we will use logistic regression to characterize the risk associated with all 4 levels of ALC (by estimating odds ratios).

(a) Trend test: In HW#2 we calculated a trend test for the distribution of Y across the levels of ALC. A similar test is obtained in logistic regression with the following model:

$\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{ALC}$. Fit this model and test $H_o : \beta_1 = 0$.

(b) Dose response: Rejection of the null using a trend test does not imply a linear increase in risk is necessarily valid. Let's consider using dummy variables to characterize the increase in risk. Fit the model

$$\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{ALC}(2) + \beta_2 \text{ALC}(3) + \beta_3 \text{ALC}(4)$$

where ALC(j) is an indicator variable for ALC=j. Give estimates and 95% confidence intervals for the coefficients. Is the linear model in (a) "nested" within this model?

(c) Is the logistic linear model in (a) adequate? Generate the dummy variables ALC(3) and ALC(4) and fit the model

$$\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{ALC} + \beta_2 \text{ALC}(3) + \beta_3 \text{ALC}(4) \quad .$$

Is the linear model nested within this model? How do the coefficient estimates in this model relate to the estimates obtained in (b)?

(d) Use a likelihood ratio test to see if we'd reject the linear model in favor of the 4 parameter model. Report the null hypothesis, test statistic, and interpret the p-value.

(e) Age adjustment: Using ALC as determined by (a)-(d) (ie. linear or dummy variables) consider a model that adjusts the ALC odds ratio(s) for AGE. Use dummy variables for the AGE categories. Report the adjusted ALC odds ratio(s) and interpret the adjusted odds ratio.

(f) Age adjustment: Can we simplify the model that uses AGE dummy variables by using a linear term in AGE? Justify.

(g) Continuous covariates: Consider the model $\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{ALC}$ where ALC takes the values 1,2,3,4. If we had data that recorded the actual g/day of alcohol consumption what would you predict the logistic regression coefficient to be for a model that replaces our ALC (groups of alcohol) with the continuous alcohol measurement? Explain.
NOTE: I'll provide the data with these continuous measurements next week!