Biostatistics 513

Homework 5 - due 5/6/13

Please submit your assignment in Word or PDF format by 9:30 am (PDT) to the Biost 513 Assignment Dropbox at http://canvas.uw.edu/. You may go to "Assignments" at the left side menus on your Canvas page and then follow the instructions to upload your assignment.

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 Homework directory.

Stata.help5: Key Stata commands that are useful for these exercises are described in the text file stata.help4, also available in the homework directory.

PLEASE START YOUR RESPONSE TO EACH QUESTION ON A SEPARATE PAGE AND PUT YOUR NAME ON EACH 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.dta). 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.

   Note: Be careful to drop any cases with missing values (e.g., 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,

$$\text{logit}\left[\pi(X)\right] = \beta_0 + \sum_{j=2}^{6} \beta_j \text{AGE}(j) + \beta_7 \text{ALC} + \beta_8 \text{TOB}$$

Give interpretations for the estimates β_7 and β_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, one at a time. Should we include either variable as a quadratic function? Support your conclusion with appropriate statistics.

(c) Using the model in (a), estimate the relative risk and 95% CI (using odds ratios) for oesophageal cancer for a 60 year old man who consumes 80 g alcohol/day and uses 15 g tobacco/day compared to a 60 year old man who neither smokes nor drinks (hint: use lincom)

Logistic Regression – Testing for Trend

2) Let's continue the analysis of the data collected by Tuyns et al. (1977). 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.We 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 (Homework3, problem 1b). Here we will characterize the relative risk associated with all 4 levels of ALC (by estimating odds ratios).

   (a) Do a chi-square test for trend for case/control status and ALC. Typically, for a trend test with 4 levels for the exposure, our null hypothesis would be

   Ho: $p_1 = p_2 = p_3 = p_4$

   where p_i = Pr(disease | alcohol level i), but that doesn't seem right with a case-control study (since we have learned that we can't estimate Pr(disease) in such a study). State an equivalent null hypothesis in terms of odds ratios. Use the `tabodds` (described in stata2.help) procedure to test this hypothesis and state your conclusion.

   (b) 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 Ho: $\beta_1 = 0$. Give the estimated value of β_1, a 95% CI and interpret.

   (c) Although the previous result indicates that there is an increasing trend, it may not be exactly linear. Let's consider using dummy variables to characterize the increase in (log)odds. 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. Interpret the coefficient of ALC(3).

   (d) To determine if the model in part (b) is adequate we might consider doing a likelihood ratio test between the model in (b) and the model in (c). However, it doesn't seem as though model (b) is nested in model (c). However, try fitting the following model:

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

   (you'll have to generate the dummy variables ALC(3) and ALC(4)). Clearly, the model in part (b) is nested in this model. Now, complete the following table

		Predicted probabilities			
	likelihood	π(ALC=1)	π(ALC=2)	π(ALC=3)	π(ALC=4)
Model (c)					
Model (d)					

What do you conclude about the models in parts (c) and (d)? Based on this, is the model in part (b) nested in the model in part (c)?

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

(f) Age adjustment: Using ALC as determined by (b)-(e) (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), confidence interval(s) and interpret the adjusted odds ratio(s).

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

(h) Continuous covariates: Consider again the model $\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{ALC}$. Here, ALC takes the values 1,2,3,4 as described in the tuyns documentation. If we had data that recorded the actual g/day of alcohol consumption what would you predict the logistic regression coefficient to be (approximately) for a model that replaces this ALC (groups of alcohol) with the continuous alcohol measurement? Explain. (Hint: each of the categories of alcohol consumption in our current model represents about a 40g/day increase in alcohol consumption)