Biostatistics 513

Homework 3 - due 4/22/13

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: Information for the study and dataset used these exercises can be found on the class web page: http://courses.washington.edu/b513/ in the Homework directory

StataHelp3: Key Stata commands that are useful for these exercises are described in the text file stataHelp3, also available in the homework directory.

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

1) The datafile **tuyns.dat** was collected by Tuyns et al (1977) in the French department of Ille-et-Villane (Brittany). Cases in this study were 200 males diagnosed with oesophageal cancer in one of the regional hospitals between January 1972 and April 1974. Controls were a sample of 778 adult males drawn from electoral lists in each commune, of whom 775 provided sufficient data for analysis. Both types of subject were administered a detailed dietary interview which contained questions about their consumption of tobacco and of various alcoholic beverages. The goal of the study is to characterize the cancer risk associated with both exposures but we want to control for the effect of age as this is a potential confounder in many cancer studies.

   For this exercise we will look at some dichotomizations of TOB and ALC. Define
   NEWALC = 1 if alcohol consumption is $\geq$ 80 g/day, 0 otherwise;
   NEWTOB = 1 if tobacco consumption is $\geq$ 10 g/day, 0 otherwise.

   a) Provide a table of AGE and case-control status. Summarize the relationship between age and case-control status.

   b) Provide tables of AGE and NEWALC and AGE and NEWTOB in the control group. Summarize the relationship between age and each of the exposures in the control group.

   c) Analyze the relationship between Y (case/control status) and NEWALC, adjusting for AGE using stratification. Provide
      i A crude (unadjusted) odds ratio and 95% CI between Y and NEWALC. Interpret the odds ratio.
      ii Does it seem reasonable to quote a common odds ratio for the different levels of age? Support your answer.
      iii If your answer to (ii) is yes, provide and interpret an adjusted (for AGE) odds ratio along with 95% confidence intervals. If your answer to (ii) is no, provide age-specific odds ratios and 95% CI's.
      iv Does AGE appear to be confounding the relationship between Y and NEWALC?

   d) Analyze the relationship between Y (case/control status) and NEWTOB, adjusting for AGE. Provide

i A crude (unadjusted) odds ratio and 95% CI between Y and NEWTOB. Interpret the odds ratio.
ii Does it seem reasonable to quote a common odds ratio for the different levels of age? Support your answer.
iii If your answer to (ii) is yes, provide and interpret an adjusted (for AGE) odds ratio along with 95% confidence intervals. If your answer to (ii) is no, provide age-specific odds ratios and 95% CI's.
iv Does AGE appear to be confounding the relationship between Y and NEWTOB?

e) Provide a table of NEWALC and NEWTOB in the control group. Considering this table, along with the results you found in (c) and (d), is it plausible that NEWTOB could be confounding the relationship between disease and NEWALC, or visa-versa?

f) Use `mhodds` to estimate the OR between disease and NEWALC after adjusting for age and NEWTOB. Report the adjusted OR, a 95% CI and compare this adjusted OR to the one you obtained in part c.

g) Use `mhodds` to estimate the OR between disease and NEWTOB after adjusting for age and NEWALC. Report the adjusted OR, a 95% CI and compare this adjusted OR to the one you obtained in part d.

2) A study was undertaken to assess predictors of a baby being born with a low birth weight. A total of n = 189 women receiving care at Baystate Medical Center were surveyed and asked about their behavior during pregnancy (including diet, smoking, and prenatal care visits). Babies that are born weighing less than 2500g are considered "low birth weight" (LBW=1 if <2500g, 0 otherwise) and are at increased risk of morbidity and mortality. Documentation for the analysis variables is in the Datasets section of the course web. Briefly, SMOKE=1 if the mother is a smoker, AnyPTL=1 if any history of premature labor (ie. PTL$\geq$1), RACE(2)=1 if mother is Black and 0 otherwise, RACE(3)=1 if mother is neither White nor Black (ie. RACE="other") and 0 otherwise, HYPER=1 if mother has history of hypertension and 0 otherwise, URIRR=1 if presence of uterine irritability and 0 otherwise, AGE20 = maternal age - 20 (years).

In analyzing the data, two logistic models were fit, each involving the dependent variable LBW, but with different sets of independent variables. The variables involved in each model and their estimated coefficients are listed below:

	Model 1	**Model 2**
Variable	Coefficient	Coefficient
Intercept	-2.003	-1.905
SMOKE	0.896	0.713
AnyPTL	1.317	1.295
RACE(2)	0.962	0.870
RACE(3)	0.951	0.906
HYPER	1.364	1.396
URIRR	0.769	0.838
AGE20	-0.051	-0.082
SMOKE×AGE20	---	0.066

(a) For model 1 above, state the form of the logistic model that was used – what is the dependent variable, the interpretation of the probability $\pi(X)$, and the model for $\pi(X)$ in terms of the (unknown) population parameters and the independent variables?

(b) For model 1 in (a) state the form of the estimated log odds functions i.e. $\text{logit}[\pi(X)] = \ldots$ (i.e. put the parameter estimates into the model from a).

(c) Using model 1, compute the estimated risk for LBW (ie. P[LBW=1]) for

i. a smoker (SMOKE=1), without a history of premature labor (AnyPTL=0), black (RACE(2)=1,RACE(3)=0), without hypertension (HYPER=0), without uterine irritability (URIRR=0), who is age 30 (person 1);
ii. a non-smoker (SMOKE=0), without a history of premature labor (AnyPTL=0), black (RACE(2)=1,RACE(3)=0), without hypertension (HYPER=0), without uterine irritability (URIRR=0), who is age 30 (person 2)
iii. what is the estimated relative risk for these two individuals?

(d) Repeat part (c) using model 2. Why is the estimated RR different from what you found in part c? Is the interpretation of the two RR's different?

(e) What is the estimated odds ratio comparing SMOKER=1 to SMOKER=0 for 20 year old women with AnyPTL=0, RACE(2)=0, RACE(3)=0, HYPER=0, and URIRR=0 under model 1 and under model 2 (Note: use the coefficients directly rather than calculate $\pi(X)$).

(f) What is the estimated odds ratio comparing SMOKER=1 to SMOKER=0 for 30 year old women with AnyPTL=0, RACE(2)=0, RACE(3)=0, HYPER=0, and URIRR=0 under model 1 and under model 2 (Note: use the coefficients directly rather than calculate $\pi(X)$).

3) Back to the Ille-et Villaine data! We can use logistic regression to answer many of the same questions we posed in problem 1. We will focus on the analyses in which NEWALC is the exposure of interest. Here is the output of a logistic regression that includes both NEWALC and age (as a series of dummy or indicator variables) in the analysis. The dependent variable is case/control status.

```
Logit estimates                                   Number of obs   =        975
                                                  LR chi2(6)      =     200.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -394.46094                       Pseudo R2       =     0.2027

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      newalc |    1.66989   .1896018     8.81   0.000     1.298277    2.041503
     _Iage_2 |   1.542294   1.065895     1.45   0.148     -.546822     3.63141
     _Iage_3 |   3.198762    1.02314     3.13   0.002     1.193445    5.204079
     _Iage_4 |    3.71349   1.018531     3.65   0.000     1.717207    5.709774
     _Iage_5 |   3.966882   1.023072     3.88   0.000     1.961698    5.972066
     _Iage_6 |    3.96219   1.065024     3.72   0.000      1.87478    6.049599
       _cons |  -5.054348   1.009422    -5.01   0.000    -7.032778   -3.075917
------------------------------------------------------------------------------
```

(a) Write the logistic regression model that produced this result.
(b) Use the output above to estimate the odds ratio comparing NEWALC = 1 to NEWALC = 0, adjusting for age. Compare this estimate to the one you reported in 1b, part iii. Is it of similar magnitude? Is the interpretation similar?

(c) What is the interpretation of the coefficients for the age groups in the output given above?

(d) Suppose I wanted to know the odds ratios for the age groups using the 35-44 year olds (age group 2) as the baseline (comparison) group. How could I compute these odds ratios from the output given above (i.e. without refitting the model)?