Biostatistics 513

Homework 3 Solutions

1)

a)

. tabodds y age [freq=count]

```
-------------------------------------------------------------------------
       age  |      cases     controls       odds      [95% Conf. Interval]
------------+------------------------------------------------------------
      25-34 |          1          115    0.00870        0.00121   0.06226
      35-44 |          9          190    0.04737        0.02427   0.09244
      45-54 |         46          167    0.27545        0.19875   0.38175
      55-64 |         76          166    0.45783        0.34899   0.60061
      65-74 |         55          106    0.51887        0.37463   0.71864
        75+ |         13           31    0.41935        0.21944   0.80138
-------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(5)  =    96.94
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    83.37
                                  Pr>chi2  =   0.0000
```

It is clear that there is a relationship between age and case control status ($p < .0001$ by either the test for trend or the test for homogeneity). Specifically, cases tend to be older than controls.

b)

i) alcohol use and age in the control group:

Age Group	newalc 0	1	Total
25-34	106	9	115
	92.17	7.83	100.00
35-44	164	26	190
	86.32	13.68	100.00
45-54	138	29	167
	82.63	17.37	100.00
55-64	139	27	166
	83.73	16.27	100.00
65-74	88	18	106
	83.02	16.98	100.00
75+	31	0	31
	100.00	0.00	100.00
Total	666	109	775
	85.94	14.06	100.00

Alcohol use appears to increase with age, then levels off between 45 - 74, then decreases in those over 75.

ii) tobacco use and age in the control group

Age Group	newtob 0	1	Total
25-34	70	45	115
	60.87	39.13	100.00
35-44	107	83	190
	56.32	43.68	100.00
45-54	90	77	167
	53.89	46.11	100.00
55-64	92	74	166
	55.42	44.58	100.00
65-74	68	38	106
	64.15	35.85	100.00
75+	20	11	31
	64.52	35.48	100.00
Total	447	328	775
	57.68	42.32	100.00

Tobacco use is highest in those 45 – 54 and declines with age in either direction

c)

```
. cc y newalc [freq=count], by(age) bd
```

Age Group	OR	[95% Conf.	Interval]	M-H Weight	
25-24	.	0	.	0	(exact)
35-44	5.046154	.9268664	24.86538	.6532663	(exact)
45-54	5.665025	2.632894	12.16536	2.859155	(exact)
55-64	6.359477	3.299319	12.28473	3.793388	(exact)
65-74	2.580247	1.131489	5.857261	4.024845	(exact)
75+	.	4.388738	.	0	(exact)
Crude	5.640085	3.937435	8.061794		(exact)
M-H combined	5.157623	3.562131	7.467743		

```
Test of homogeneity (B-D)      chi2(5) =     9.32  Pr>chi2 = 0.0968

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =     85.01
                                                Pr>chi2 =    0.0000
```

The crude OR is 5.64 (95% CI 3.93 – 8.06) suggesting that high alcohol consumption is associated with increased risk of oesophageal cancer. Since the Breslow-Day test of homogeneity is not significant, it seems reasonable to assume a common OR (further, there is no noticeable pattern in the stratum-specific OR's with age). We estimate this common OR to be 5.16 (95% CI 3.56 – 7.47). This common OR is highly significantly different from 1.0 and suggests that high alcohol consumption is associated with an approximately 5-fold increased risk of oesophageal cancer even after adjusting for age (i.e. within age groups). The adjusted OR differs from the crude OR by about 10% so there is a small confounding effect of age.

d)

```
. cc y newtob [freq=count], by(age) bd

       Age Group |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
           25-24 |          .            0          .             0 (exact)
           35-44 |   4.512048     .8251513    45.2992      .8341709 (exact)
           45-54 |   2.671614     1.270052   5.808776      5.061033 (exact)
           55-64 |   2.536216     1.387893   4.681215      7.644628 (exact)
           65-74 |   1.385399     .6734824   2.833404       7.31677 (exact)
             75+ |   2.121212     .4669619   9.705684           1.5 (exact)
-----------------+-------------------------------------------------
           Crude |   2.131567     1.533217   2.970895               (exact)
    M-H combined |   2.263103     1.609809   3.181516
-----------------------------------------------------------------------
Test of homogeneity (B-D)      chi2(5) =     3.92  Pr>chi2 = 0.5615

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =     22.53
                                                Pr>chi2 =    0.0000
```

The crude OR is 2.13 (95% CI 1.53 – 2.97) suggesting that tobacco use is associated with increased risk of oesophageal cancer. Since the Breslow-Day test of homogeneity is not significant, it seems reasonable to assume a common OR (the stratum-specific OR's do tend to decline with age; note that the B-D test is a general test of homogeneity and is not specifically testing for trend; we can do this with logistic regression). We estimate this common OR to be 2.26 (95% CI 1.61 – 3.18). This common OR is highly significantly different from 1.0 and suggests that tobacco use is associated with a 2.26-fold increased risk of oesophageal cancer even after adjusting for age. The adjusted OR differs from the crude OR by about 6% so there is little evidence of confounding

e)

```
           |        newtob
    newalc |         0          1 |     Total
-----------+----------------------+----------
         0 |       397        269 |       666
         1 |        50         59 |       109
-----------+----------------------+----------
     Total |       447        328 |       775
```

A greater proportion of those that use tobacco also consume higher levels alcohol. Since both are strongly related to disease status, the two exposures could be confounding each other.

f)

```
    Mantel-Haenszel estimate controlling for age and newtob
    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       4.879388      76.85        0.0000         3.294850   7.225951
    ------------------------------------------------------------------
```

After adjusting for both age and NEWTOB, the OR between disease and NEWALC is 4.88 (95% CI 3.29 – 7.22). This is the pooled OR within strata defined by both age and tobacco use (there are 6x2=12 such strata). This is 13% less than the crude OR but only 5% less than the OR adjusted for age that we found in part b. Thus, after controlling for age, there is little additional confounding due to tobacco use. Clearly, the fundamental finding – that risk of oesophageal cancer is associated with increase alcohol consumption is the same in all three cases.

Note that there are four different commands you could use for this analysis:

```
. mhodds y newalc age newtob [freq=count]
. mhodds y newalc [freq=count], by(age newtob)
. mhodds y newalc newtob [freq=count], by(age)
. mhodds y newalc age [freq=count], by(newtob)
```

Each gives the same result for a common OR (whew!) but they give different homogeneity tests. Essentially mhodds does a homogeneity test on the strata defined by the by() variable(s). So if you say by(age) you get a homogeneity test on the OR's within the 6 age strata and if you say by (age newtob) you get a homogeneity test on the OR's in the 6x2=12 agexnewtob strata. But it's a bit tricky – in the last example above you get a homogeneity test on the OR's within the 2 newtob strata. But those are the OR's between Y and newalc, *adjusted for age*. That is, you are doing a homogeneity test on adjusted OR's! In regression terms, we are fitting a model with terms NEWALC + NEWTOB + AGE(k) + NEWALC*NEWTOB + AGE(k)*NEWTOB and testing whether the NEWALC*NEWTOB term is significant.

g)

```
Mantel-Haenszel estimate controlling for age and newalc
------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------
   2.081361      15.96        0.0001         1.440940   3.006416
------------------------------------------------------------------
```

After adjusting for both age and NEWALC, the OR between disease and NEWTOB is 2.08 (95% CI 1.44 – 3.01). This is an estimate of the pooled or common OR within levels of age and tobacco use (there are 12 such strata) This is similar to the crude OR in part c and 8% less than the OR adjusted for age only.

2)

a) The dependent variable is LBW (yes/no) so $\pi(X)$ is the probability that an infant with a given set of covariates, X, is low birthweight. The model for $\pi(X)$ is

$$\pi(X) = \frac{\exp(\beta_0 + \beta_1 \times SMOKE + \beta_2 \times AnyPTL + \beta_3 \times RACE(2) + \beta_4 \times RACE(3) + \beta_5 \times HYPER + \beta_6 \times URIRR + \beta_7 \times AGE20)}{1 + \exp(\beta_0 + \beta_1 \times SMOKE + \beta_2 \times AnyPTL + \beta_3 \times RACE(2) + \beta_4 \times RACE(3) + \beta_5 \times HYPER + \beta_6 \times URIRR + \beta_7 \times AGE20)}$$

b) The estimated logodds function is

$$\text{logit}(\pi(X)) = -2.003 + 0.896SMOKE + 1.317AnyPTL + 0.962RACE(2) + 0.951RACE(3) + 1.364HYPER + 0.769URIRR - 0.051AG$$

c)

i) X = (1, 0, 1, 0, 0, 0, 30-20)
Z = Xβ = -2.003 + 0.896(1) + 1.317(0) + 0.962(1) + 0.951(0) + 1.364(0) + 0.769(0) – 0.051*(30-20) = -.655
π(X)= exp(Z)/(1+exp(Z)) = .342

ii) X = (0, 0, 1, 0, 0 ,0, 30-20)
Z = Xβ = -1.551
π(X)= exp(Z)/(1+exp(Z)) = .175

iii) RR = .342/.175 = 1.95

30 year old black smokers without hypertension and without history of premature labor or uterine irritability are 1.95 times more likely to have a low birthweight baby compared to 30 year old black nonsmokers without hypertension who do not have a history of premature labor or uterine irritability.

d)

i) X = (1, 0, 1, 0, 0, 0, 30-20, 1*(30-20))
Z = Xβ = -1.905 + 0.713(1) + 1.295(0) + 0.870(1) + 0.906(0) + 1.396(0) + 0.838(0) - .082*(30-20) + .066*1*(30-20) = -0.482
π(X)= exp(Z)/(1+exp(Z)) = .382

ii) X = (0, 0, 1, 0, 0 ,0, 30-20,0*(30-20))
Z = Xβ = -1.855
π(X)= exp(Z)/(1+exp(Z)) = .135

iii) RR = .382/.135 = 2.83

This estimate is different from the RR in part c due to the presence of an interaction term in model 2. However, the interpretation is the same as given above. This reminds us that estimates are model dependent!

e) For model 1 the OR due to smoking is exp(0.896) = 2.45. For model 2 the OR due to smoking for a 20 year old woman is exp(0.713) = 2.04.

f) For model 1 the OR is still is exp(0.896) = 2.45. For model 2 the OR due to smoking for a 30 year old woman is exp(0.713+10*.066) = 3.95

3)

a) The logistic regression model is

$$\text{logit}(\pi(X)) = -5.05 + 1.67\,NEWALC + 1.54\,AGE(2) + 3.20\,AGE(3) + 3.71\,AGE(4) + 3.97\,AGE(5) + 3.96\,AGE(6)$$

b) The estimated OR for NEWALC is exp(1.67) = 5.31. This is similar to the adjusted OR found in problem 1b, namely 5.16 (but not exactly the same. The interpretation is similar: it is the estimated ratio of the odds of disease in individuals who consume $\geq$ 80g alcohol per day to the odds of disease in individuals who consume < 80g alcohol per day, adjusted for age group.

c) Each of the age group coefficients gives the estimated log(odds ratio) for disease comparing that age group to the baseline age group (the 25-34 year olds in this example), adjusted for alcohol consumption. For example, we see the risk of oesophageal cancer in men aged 65-74 (age group 5) is estimated to be exp(3.97) = 53 (!) times higher than men in the 25 – 34 year old age group. Note, however, that there is just 1 case of cancer among the 116 men in the 25-34 year old group. Thus, the estimate of risk in that age group is probably not very stable. Further, we see from the output that the coefficient for the 35- 44 year old group is not significantly different from the younger group. In the future, we will combine these two groups.

d) The log odds ratio for any group compared to the second age group is obtained by subtracting the age_2 coefficient from the coefficient of the age group of interest. For example, the log odds ratio comparing group 3 to group 2 is 3.20 – 1.54 = 1.66. The corresponding odds ratio is exp(1.66).