Biostatistics 513

Homework 5 Solutions

1)

a) Here is the stata output

```
. xi: logistic case i.agegp tob alc
i.agegp           _Iagegp_1-6        (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        976
                                                  LR chi2(7)      =     294.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -347.74036                       Pseudo R2       =     0.2975

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |   5.851322   6.433949     1.61   0.108     .6781047    50.49069
   _Iagegp_3 |   34.90621    36.8794     3.36   0.001     4.401332    276.8351
   _Iagegp_4 |   59.30331   62.43568     3.88   0.000      7.53212     466.918
   _Iagegp_5 |   101.4181   107.7435     4.35   0.000     12.64226    813.5914
   _Iagegp_6 |   105.0871   116.9321     4.18   0.000     11.86871    930.4551
         tob |   1.041306   .0082032     5.14   0.000     1.025352    1.057509
         alc |   1.026306   .0026633    10.01   0.000     1.021099    1.031539
------------------------------------------------------------------------------
```

The interpretation of the odds ratio for tob is that 1.0413 is the odds of oesophageal cancer associated with a one gram/day increase in tobacco consumption, after adjusting for age and alcohol consumption. The OR for alc indicates that the odds of oesophageal cancer increases by a factor of 1.0263 for each one gram/day increase in alcohol consumption, after adjusting for age and tobacco use.

b) Stata output for adding TOB^2:

```
. xi: logistic case i.agegp tob alc tob2
i.agegp           _Iagegp_1-6        (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        976
                                                  LR chi2(8)      =     296.01
                                                  Prob > chi2     =     0.0000
Log likelihood = -346.96762                       Pseudo R2       =     0.2990

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |   5.318698   5.844352     1.52   0.128     .6172746    45.82814
   _Iagegp_3 |   31.73698    33.4744     3.28   0.001     4.015739     250.822
   _Iagegp_4 |   54.35493   57.10336     3.80   0.000     6.934209    426.0701
   _Iagegp_5 |   91.37431    96.9344     4.26   0.000     11.42424    730.8376
   _Iagegp_6 |    96.9282    107.657     4.12   0.000     10.99083    854.8101
         tob |   1.066909   .0226092     3.06   0.002     1.023503    1.112155
         alc |   1.026023   .0026677     9.88   0.000     1.020808    1.031265
        tob2 |   .9994097   .0004758    -1.24   0.215     .9984777    1.000343
------------------------------------------------------------------------------
```

The Wald test for TOB^2 has p-value equal to .21 suggesting that a quadratic term is not necessary for tobacco.

Stata output for adding ALC^2:

```
. xi: logistic case i.agegp tob alc alc2
i.agegp           _Iagegp_1-6         (naturally coded; _Iagegp_1 omitted)

Logistic regression                                 Number of obs   =        976
                                                    LR chi2(8)      =     294.84
                                                    Prob > chi2     =     0.0000
Log likelihood = -347.55224                         Pseudo R2       =     0.2978

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |   5.924965   6.543814     1.61   0.107     .6801085    51.61708
   _Iagegp_3 |   36.12426   38.36408     3.38   0.001     4.506302    289.5861
   _Iagegp_4 |   61.16769   64.71265     3.89   0.000     7.691116    486.4686
   _Iagegp_5 |   104.5539   111.6036     4.36   0.000     12.90461    847.1014
   _Iagegp_6 |   105.9373   118.2657     4.18   0.000     11.87925    944.7326
         tob |   1.041279   .0082051     5.13   0.000     1.025321    1.057486
         alc |   1.021612   .0081205     2.69   0.007     1.005819    1.037652
        alc2 |    1.00003   .0000504     0.60   0.547     .9999316    1.000129
------------------------------------------------------------------------------
```

The Wald test for ALC^2 has p-value equal to .55 suggesting that a quadratic term is not necessary for alcohol.

c) Since both individuals are 60 years old, I can ignore the age term in the model in calculating the OR (a reasonable estimate of the RR). After refitting the model in (a) (to make it the current model) I can do

```
. lincom 80*alc + 15*tob

( 1)  15 tob + 80 alc = 0

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   14.64964   3.535333    11.12   0.000     9.128729    23.50952
------------------------------------------------------------------------------
```

Given that oesophageal cancer risk is rare, I estimate that the *risk* of oesophageal cancer for the drunken smoker is 14.6 times the risk of the clean-livin' teetotaler.

2)

a) An equivalent null hypothesis in terms of odds ratios is

Ho: 1 = OR_2 = OR_3 = OR_4 (i.e., take Ho: $odds_1$=$odds_2$=$odds_3$=$odds_4$ and divide by $odds_1$ to get Ho as OR's.)

where $OR_i = p_i/(1-p_i)/[p_1/(1-p_1)]$ and p_i = Pr(disease | alcohol level i). Each odds is compared to the lowest alcohol group. You could, of course, choose any of the alcohol groups as the reference category and get an equivalent null hypothesis.

The stata output is

```
.  tabodds y alc [freq=count]

--------------------------------------------------------------------------
       alc  |      cases     controls       odds      [95% Conf. Interval]
------------+-------------------------------------------------------------
   <40g/day |         29          386    0.07513        0.05151   0.10957
  40-79g/~y |         75          280    0.26786        0.20760   0.34560
  80-119g~y |         51           87    0.58621        0.41489   0.82826
  120+g/day |         45           22    2.04545        1.22843   3.40587
--------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(3)  =   158.79
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =   152.97
                                  Pr>chi2  =   0.0000
```

We conclude there is a highly significant ($p < .001$) trend in the odds of disease with level of alcohol exposure.

b) The stata output for a logistic regression is

```
. logit y alc [freq=count]

Logit estimates                                   Number of obs   =        975
                                                  LR chi2(1)      =     144.64
                                                  Prob > chi2     =     0.0000
Log likelihood =  -422.4246                       Pseudo R2       =     0.1462

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.046772   .0935048    11.19   0.000     .8635064    1.230038
       _cons |  -3.530124   .2279715   -15.48   0.000    -3.976939   -3.083308
------------------------------------------------------------------------------
```

The test of Ho: $\beta_1 = 0$ is rejected with $p < .001$. We estimate that the log odds of disease increases by 1.05 (95% CI 0.86 – 1.23) for each categorical increase of alcohol consumption.

c) Here’s the stata output

```
. xi: logit y i.alc [freq=count]
i.alc             _Ialc_1-4          (naturally coded; _Ialc_1 omitted)

Logit estimates                                   Number of obs   =        975
                                                  LR chi2(3)      =     146.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -421.49545                       Pseudo R2       =     0.1481

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Ialc_2 |    1.27124    .232332     5.47   0.000     .8158777    1.726602
     _Ialc_3 |   2.054459   .2611044     7.87   0.000     1.542704    2.566214
     _Ialc_4 |   3.304162   .3236511    10.21   0.000     2.669817    3.938506
       _cons |  -2.588542   .1925445   -13.44   0.000    -2.965922   -2.211161
------------------------------------------------------------------------------
```

Point estimates and confidence intervals for each level of alcohol use are given. For instance, the coefficient for _Ialc_3 is 2.05 (95% CI 1.54 – 2.57) indicating that the logodds for disease associated with 80 – 119 gms/day alcohol consumption is 2.05 greater than the logodds for disease associated with 0 - 39 gms/day alcohol consumption.

d) Here is the relevant stata output

```
. generate alc3=0
. replace alc3=1 if alc==3
. generate alc4=0
. replace alc4=1 if alc==4
. xi: logit y alc alc3 alc4 [freq=count]

Logit estimates                                   Number of obs   =        975
                                                  LR chi2(3)      =     146.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -421.49545                       Pseudo R2       =     0.1481

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |    1.27124    .232332     5.47   0.000     .8158777    1.726602
        alc3 |   -.488021   .3685031    -1.32   0.185    -1.210274    .2342318
        alc4 |  -.5095586   .6067233    -0.84   0.401    -1.698714    .6795972
       _cons |  -3.859782    .406446    -9.50   0.000    -4.656401   -3.063162
------------------------------------------------------------------------------
```

		Predicted probabilities			
	likelihood	π(ALC=1)	π(ALC=2)	π(ALC=3)	π(ALC=4)
Model (c)	-421.495	.0699	.2112	.3695	.6716
Model (d)	-421.495	.0699	.2112	.3695	.6716

I conclude that models (c) and (d) are equivalent.

Since it is clear that the model

$\text{logit}[\pi(X)] = \beta_0 + \beta_1\text{ALC}$

is nested within the model

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

and we have shown that the models

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

are equivalent, that means that

$\text{logit}[\pi(X)] = \beta_0 + \beta_1\text{ALC}$

is nested within the model

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

e) The likelihood ratio chi square is equal to 1.86 (2 df) and the p-value is 0.39. Thus we would conclude that the logistic linear model in part b gives an adequate fit to these data.

f) The stata output is

```
. xi: logit y alc i.age [freq=count]
i.age             _Iage_1-6           (naturally coded; _Iage_1 omitted)

Logit estimates                                   Number of obs   =        975
                                                  LR chi2(6)      =     255.91
                                                  Prob > chi2     =     0.0000
Log likelihood = -366.78769                       Pseudo R2       =     0.2586

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.093726   .1032947    10.59   0.000     .8912717     1.29618
     _Iage_2 |   1.575683    1.07541     1.47   0.143    -.532081     3.683448
     _Iage_3 |   3.304031   1.031874     3.20   0.001     1.281595    5.326467
     _Iage_4 |   3.826275   1.027804     3.72   0.000     1.811816    5.840734
     _Iage_5 |   4.214935   1.033778     4.08   0.000     2.188767    6.241102
     _Iage_6 |   4.308531   1.083401     3.98   0.000     2.185104    6.431957
       _cons |  -6.983926   1.050337    -6.65   0.000    -9.042548   -4.925304
------------------------------------------------------------------------------
```

The odds ratio for alc in this model is exp(1.09) = 2.98 (95% CI = 2.44 – 3.66). We interpret this as the adjusted (for age) relative increase in the odds of disease for each categorical increase in alcohol consumption.

g) The likelihood ratio test comparing a model with age categories versus a model with age as a linear term has a chi-square statistic equal to 17.77 with 4 df. The p value is 0.0014 so we would conclude that the linear model for age is not adequate.

h) I would reason like this: each of the categories of alcohol consumption in our current model represents about a 40g/day increase in alcohol consumption. The coefficient in this model is 1.046. So if we had a model with actual grams of alcohol consumption, I might expect that the value of the coefficient would be about 1.046/40 = 0.026.