Solutions, Biostat III exam, 19 Dec 2012

Section 1

1. (a) After adjusting for total energy intake (in two categories) and job type (in three categories) it is estimated that the odds of CHD incidence increases by a multiplicative factor of 1.08 (i.e., by 8%) for each and every 1 unit increase in BMI. The effect is not statistically significant (CI contains 1 and p-value is 0.13).

   (b) The p-values for the parameters representing the effect of occupation indicate the significance of the pairwise comparison between each occupation and the reference and we should not make a conclusion regarding the overall effect based on those tests alone. In order to test for a global (overall effect) of occupation on CHD risk we could conduct a joint test of the two parameters representing occupation, e.g., a likelihood ratio test or a Wald test.

   It's possible, for example, that the outcome (e.g., odds) in the reference category is intermediate to the other two categories such that the two pairwise comparisons are not significant but the overall test is. This, however, is not highly likely. It's rather common, however, to see one or more statistically significant pairwise comparisons while the overall test is not significant. In short, one should not use the pairwise p-values to make inference about the overall effect.

   (c) OR = 0.468. The OR is assumed to be the same within any level of BMI since the model does not account for possible effect modification.

   (d) OR = $(1.064)^5 = 1.364$. The effect is not statistically significant since the p-value is 0.221. Note that changing the scale, i.e., a one unit increase or a five unit increase does not affect the significance.

   (e) There is no formal test for confounding. If the effect of high energy was confounded by job type we would expect to see a substantial difference in the OR representing the effect of energy intake if we include job type in the model compared to when it is left out. The OR for energy intake goes from 0.468 to 0.455 so there is no convincing evidence of confounding by job type.

   (f) For drivers the OR = 0.379, for conductors the OR = $0.379 \times 1.343 = 0.509$ and for bankers the OR = $0.379 \times 1.242 = 0.471$

(g) We can perform a likelihood ratio test by testing the null hypotheses that the 2 interaction parameters are 0 against the alternative hypothesis that at least one of parameters is non-zero. That is, we test whether the likelihood for the more elaborate model (model 3) is statistically greater than the likelihood for the reduced model (model 1). The test statistic is:

$$D : -2(lnL_{\text{(model 1)}} - lnL_{\text{(model 3)}}) = -2(-127.85 + 127.79) = 0.12$$

Under the null hypothesis, the test statistic follows a χ^2 distribution with 2 df (the difference in the number of parameters between the two models). The critical value of a χ^2 with 2 degrees of freedom is 5.99 at the 5 % significance level. Since our test statistic is considerably less than the critical value we conclude that there is no evidence that the effect of high energy is modified by job type.

Note that Stata reports, for each model, a LR test comparing the model to the so-called null model (the model with only one parameter). For model 3 this is labelled `LR chi(6)` and is a joint test of the significance of the additional 6 parameters in model 3 compared to the null model. We note that there is no evidence that this model provides a better fit to the data than the null model (where we assume everyone has the exact same odds of CHD). Mathematically, the test statistic is calculated as the negative of twice the difference between the log likelihood for model 3 and the log likelihood for the null model. If we take the difference (without multiplying by 2) between this test statistic for model 3 and the corresponding test statistic for model 1 then we have an expression that is equivalent to the LR test comparing models 3 and 1. That is, the test statistic can be calculated as $7.89 - 7.77 = 0.12$

Section 2

2. (a) In choosing to fit this model we have effectively reduced the data set to two rows, males with 120 deaths in 3500 person-years and females with 110 deaths in 4000 person-years.

   $\hat{\beta}_0$ is the log rate for males which is equal to $\ln(120/3500) = \ln(0.034) = -3.37$

   $\hat{\beta}_1$ is the log rate ratio (females/males) which is equal to $\ln \frac{110/4000}{120/3500} = \ln(0.802) = -0.22$

   You could also argue as follows:

   When $X = 0$ we have $\ln(\lambda) = \beta_0$ implying $\beta_0 = \ln(120/3500) = -3.37$

   When $X = 1$ we have $\ln(\lambda) = \beta_0 + \beta_1$ implying $\beta_0 + \beta_1 = \ln(110/4000) = -3.59$

   Therefore, $\beta_0 = -3.37$ and $\beta_1 = -3.59 - (-3.37) = -0.22$

   (b) The predicted rate for males is assumed to be the same for both age groups, since we have not adjusted for age, and is equal to 120/3500. The predicted number of deaths is therefore $1500 \times 120/3500 = 51.4$

   (c) The parameter estimates will be unchanged since we are fitting the same model to the same data (although presented in a different manner).

   (d) We would expect the parameter estimates to change since we are now adjusting for the effect of time.

   (e) $\hat{\beta}'_1$ will remain unchanged since it represents the change in the log rate for a one unit change in sex. That is, the effect of sex is the same in both models and is represented by both $\hat{\beta}_1$ and $\hat{\beta}'_1$.

   $\hat{\beta}'_0$ will be different since it represents the log rate when $X'_{\text{sex}} = 0$. In model 1, $X_{\text{sex}} = 0$ corresponded to males whereas in the new coding males have value 1 so $X'_{\text{sex}} = 0$ will correspond to something else (without a simple interpretation).

   In the new coding there are three values of X'_{sex} that are of interest, namely 0 (the reference), 1 (males), and 2 (females). Since the rate for females is lower than the rate for males, and X'_{sex} is modelled as a linear effect, the rate for $X'_{\text{sex}} = 0$ will be higher than the rate for $X'_{\text{sex}} = 1$. Therefore, $\hat{\beta}'_0$ will be higher than $\hat{\beta}_0$.

   You were not required to provide the values of the parameter estimates but they are:

   Model 1: $\ln(\lambda) = -3.37 - 0.22X_{\text{sex}}$

   Model 1A: $\ln(\lambda) = -3.15 - 0.22X'_{\text{sex}}$

   That is, in model 1 the log rate for males is -3.37 (the intercept) whereas in model 1A it is $-3.15 - 0.22 = -3.37$.

   [1 mark for explaining why $\hat{\beta}'_1$ is unchanged; 1 mark for explaining why $\hat{\beta}'_0$ will change; and 1 mark for correctly describing the direction of change]

   (f) β_2 represents the log HR for age. Older individuals have a higher rate for both males and females so the HR will be greater than 1. Therefore $\hat{\beta}_2$ will be greater than zero.

   (g) We are modelling 4 data points with 4 parameters (i.e., a saturated model) so the model will fit the data perfectly. The predicted number of deaths will be the same as the observed number of deaths, namely 90.

   (h) $S(t) = \exp(-\Delta(t))$ where $S(t)$ is the probability of surviving and $\Delta(t)$ the cumulative hazard. If we assume the rate is constant then the cumulative hazard after 2 years will be $2 \times 30/2000 = 0.03$. Therefore, $S(t) = \exp(-0.03) = 0.97$

   The probability that a young male survives 2 years is 0.97.

An analysis of these data in Stata.

```
. list
     +-----------------------------+
     | sex   age   deaths   pyears |
     |-----------------------------|
  1. |   1     0       30     2000 |
  2. |   1     1       90     1500 |
  3. |   2     0       20     2000 |
  4. |   2     1       90     2000 |
     +-----------------------------+

/* part (a) Sex coded as 1 for females and 0 for males (by specifying i.sex) */
. glm deaths i.sex, fam(poiss) lnoff(pyears) nolog

Generalized linear models                          No. of obs      =         4
Optimization     : ML                              Residual df     =         2
                                                   Scale parameter =         1
Deviance         =  99.31185829                    (1/df) Deviance =  49.65593
Pearson          =  95.17045454                    (1/df) Pearson  =  47.58523

Log likelihood   = -61.03875361                    BIC             =  96.53927
------------------------------------------------------------------------------
           |                 OIM
    deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
     2.sex |  -.2205428   .1320009    -1.67   0.095    -.4792598    .0381743
     _cons |  -3.373027   .0912871   -36.95   0.000    -3.551946   -3.194107
ln(pyears) |          1  (exposure)
------------------------------------------------------------------------------

/* part (b) Predictions */
. predict fitted
     +----------------------------------------+
     | sex   age   deaths   pyears     fitted |
     |----------------------------------------|
  1. |   1     0       30     2000   68.57143 |
  2. |   1     1       90     1500   51.42857 |
  3. |   2     0       20     2000         55 |
  4. |   2     1       90     2000         55 |
     +----------------------------------------+

/* part (c) Sex coded as 2 for females and 1 for males */
. glm deaths sex, fam(poiss) lnoff(pyears) nolog

Generalized linear models                         No. of obs      =         4
Optimization     : ML                             Residual df     =         2
                                                  Scale parameter =         1
Deviance         =  99.31185829                   (1/df) Deviance =  49.65593
Pearson          =  95.17045454                   (1/df) Pearson  =  47.58523

Log likelihood   = -61.03875361                   BIC             =  96.53927
------------------------------------------------------------------------------
           |                 OIM
    deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
       sex |  -.2205428   .1320009    -1.67   0.095    -.4792598    .0381743
     _cons |  -3.152484   .2059715   -15.31   0.000     -3.55618   -2.748787
ln(pyears) |          1  (exposure)
------------------------------------------------------------------------------
```

3. (a) Under proportional hazards the residuals will be indpendent of time. It is difficult to assess this from the plot so we typically fit a smooth line to the residuals. Under proportional hazards, the smoothed line will be straight and have zero slope.

   (b) Yes, your colleague has made a very sensible suggestion. In a stratified (by sex) Cox model one has separate baseline hazards for males and females. There is no requirement for the hazard for males to be proportional to the hazard for females. The effect of other covariates are, however, assumed to be proportional to the sex-specific baselines.