Self-assessment test of prerequisite knowledge for Biostatistics III in Stata

Mark Clements, Karolinska Institutet

2017-10-31

- Participants in the course Biostatistics III are expected to have prerequisite knowledge equivalent to the learning outcomes of the courses Biostatistics I and Biostatistics II. In particular, participants should be comfortable interpreting the output from logistic regression models and we expect course participants to understand:
  1. how to interpret regression coefficients after fitting a logistic regression
  2. assessing confounding in a modelling framework
  3. assessing effect modification (interactions) in a modelling framework
  4. how to conduct a formal hypothesis tests (Wald and likelihood ratio tests) in a modelling framework
- This document contains a self-assessment test of the key concepts you are expected to understand prior to the course. Brief answers are provided at the end of this document. If you have difficulty with any questions we recommend you consult previous course notes and/or course texts book or consult a colleague.
- The questions are typical of exam questions from earlier biostatistics courses and the marks (in brackets) reflect the level of difficulty. If you attempt the test under examination conditions (i.e., without referring to the answers) we would recommend:
  1. if you score 70% or more then you possess the required prerequisite knowledge;
  2. if you score 40%-70% you should brush up on the areas where you lost marks;
  3. if you score less than 40% you should, at a minimum, undertake an extensive review of central concepts in statistical modelling and possibly consider studying intermediate-level courses (e.g., Biostatistics II) before taking Biostatistics III.
- Questions about this test should be addressed to Mark Clements (`mailto:mark.clements@ki.se`) via e-mail.

All questions are based on data from a cohort study designed to study risk factors for incidence of coronary heart disease (CHD). We will study three exposures of interest, body mass index (BMI), job type (3 categories) and energy intake (classified as high or low and where high is considered exposed). The Stata output shown on this page is not central to the question but is shown for completeness. The output below shows how a variable for BMI has been created and how job type and energy intake are coded. The data are available on the web (see the use statement below) so it is possible for you to reproduce all analyses shown in this document. There is also a do file available (`http://biostat3.net/download/self-assessment.do`).

We have analysed the data using logistic regression, which is not completely appropriate given that these data are from a cohort study where individuals were at risk for different amounts of time. For the purpose of this exercise you should interpret the results from the models as if logistic regression was appropriate. During Biostatistics III we will reanalyse these data using more appropriate methods (e.g., Cox regression and Poisson regression).

```
. use http://biostat3.net/download/diet, clear
. /** Generate a variable containing BMI **/
. gen bmi=weight/(height/100)^2
--------------------------------------------------------------------------------
bmi                                                                  (unlabeled)
--------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [15.875263,33.292957]        units:  1.000e-06
         unique values:  321                      missing .:  5/337

                  mean:   24.1237
              std. dev:   3.21202

           percentiles:        10%       25%        50%        75%        90%
                           20.0605    21.584   24.1144   26.5157    28.206

. codebook job

--------------------------------------------------------------------------------
job                                                                   Occupation
--------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  job

                 range:  [1,3]                        units:  1
         unique values:  3                        missing .:  0/337

            tabulation:  Freq.   Numeric  Label
                           102         1  driver
                            84         2  conductor
                           151         3  bank

. codebook hieng

--------------------------------------------------------------------------------
hieng                                                  Indicator for high energy
--------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  hieng

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/337

            tabulation:  Freq.   Numeric  Label
                           155         0  low
                           182         1  high
```

```
. codebook bmi
```

We now estimate a logistic regression model where the outcome is CHD (0 = No CHD 1 = CHD) and the exposures are coded as described above.

```
. /*Model 1*/
. logistic chd i.hieng i.job bmi
Logistic regression                             Number of obs     =        332
                                                LR chi2(4)        =       7.77
                                                Prob > chi2       =     0.1003
Log likelihood = -127.84724                     Pseudo R2         =     0.0295

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |
       high  |   .4546316   .1532119    -2.34   0.019     .2348566    .8800685
             |
         job |
   conductor  |   1.793175   .7950121     1.32   0.188     .7520364    4.275695
       bank  |   1.169097   .4660996     0.39   0.695     .5351687    2.553939
             |
         bmi |   1.082693   .0565679     1.52   0.128       .97731     1.19944
       _cons |   .0265452   .0352908    -2.73   0.006     .0019604    .3594388
------------------------------------------------------------------------------
```

1. (1 mark) Interpret the estimated odds ratio for BMI, including a comment on statistical significance.

2. (2 marks) Is it possible to ascertain, using the output on this page, whether the effect of high energy intake is modified by BMI? If so, comment on whether the effect of high energy intake is modified by BMI. If not, describe how you could study this.

3. (1 mark) Both P-values for the parameters representing the effect of occupation are greater than 0.1. Does this mean that there is no evidence of a statistically significant overall association between occupation and CHD risk? If not, how could you test whether there is an association between occupation and CHD risk?

4. (1 marks) What is the estimated odds ratio for individuals working as bankers (job=3) compared to conductors (job=2)?

5. (1 mark) Individuals with a high energy intake ($\geq$ 2750 kcals/day) appear to have a statistically significant lower risk of CHD compared to individuals with a low energy intake ($<$ 2750 kcals/day). Should we recommend individuals with a low energy intake to eat more as a means of reducing CHD risk?

We now fit another model (labelled model 2).

```
. /*Model 2*/
. logistic chd i.hieng bmi
Logistic regression                              Number of obs     =        332
                                                 LR chi2(2)        =       5.91
                                                 Prob > chi2       =     0.0522
Log likelihood =    -128.78                      Pseudo R2         =     0.0224

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |
       high  |    .468139   .1563834    -2.27   0.023     .2432362    .9009932
         bmi |   1.063526   .0535557     1.22   0.221     .9635722    1.173848
       _cons |   .0507886   .0619514    -2.44   0.015     .0046503    .5546935
------------------------------------------------------------------------------
```

6. (1 marks) Based on model 2, among individuals with a BMI of 24, what is the estimated odds ratio for individuals with a high energy compared to those with a low energy intake? You do not have to comment on statistical significane.

7. (2 marks) Based on model 2, what is the estimated odds ratio for individuals with a BMI of 30 compared to individuals with a BMI of 25? Is the difference statistically significant?

8. (2 marks) Is it possible to ascertain, using the output from models 1 and/or 2, whether the effect of high energy intake is modified by job type? If so, comment on whether the effect of high energy intake is modified by job type. If not, describe how you could study this.

9. (2 marks) Is it possible to ascertain, using the output from models 1 and/or 2, whether the effect of high energy intake is confounded by job type? If so, comment on whether the effect of high energy intake is confounded by job type. If not, describe how you could study this.

10. (3 marks) Based on models 1 and/or 2, is there any evidence that job type is associated with CHD incidence? Conduct a formal hypothesis test using output from models 1 and/or 2. You should state the null hypothesis, alternative hypothesis, value of a test statistic, assumed distribution of the test statistic under the null hypothesis, the name of the statistical test you are using, and a comment on statistical significance.

We now refit model 1, but use the `coef` option which causes Stata to report the estimated coefficients rather than the estimated odds ratios. We will label this Model 3 even though it is technically the same model as Model 1 but with estimates presented on a different scale.

```
. /* Model 3 */
. logistic chd i.hieng i.job bmi, coef
Logistic regression                             Number of obs     =        332
                                                LR chi2(4)        =       7.77
                                                Prob > chi2       =     0.1003
Log likelihood = -127.84724                     Pseudo R2         =     0.0295

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |
        high |  -.7882679   .3370023    -2.34   0.019     -1.44878   -.1277555
             |
         job |
   conductor |    .583988   .4433544     1.32   0.188    -.2849705    1.452947
        bank |   .1562318   .3986834     0.39   0.695    -.6251732    .9376368
             |
         bmi |   .0794516   .0522474     1.52   0.128    -.0229513    .1818546
       _cons |  -3.628906    1.32946    -2.73   0.006      -6.2346   -1.023211
------------------------------------------------------------------------------
```

11. (2 marks) What is the standard error and 95 percent confidence interval of the estimate for hieng? That is, what are the numbers indicated by X, Y and Z? You may make use of output from models 12 in your answer.

12. (1 mark) What is the interpretation of the intercept (i.e., the coefficient labelled `_cons`)?

```
. /* Model 4 */
. logistic chd i.hieng##i.job bmi
Logistic regression                             Number of obs     =        332
```

```
                                                LR chi2(6)        =       7.89
                                                Prob > chi2       =     0.2461
Log likelihood = -127.78775                     Pseudo R2         =     0.0300

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hieng |
       high  |   .3792746   .2469698    -1.49   0.137     .1058479    1.359018
             |
         job |
  conductor  |   1.588197   .9160756     0.80   0.423      .512778    4.919028
       bank  |   1.074633   .5513115     0.14   0.888     .3931644    2.937286
             |
   hieng#job |
      high #|
  conductor  |   1.342565   1.189766     0.33   0.740     .2363798    7.625359
  high#bank  |   1.242141   1.018884     0.26   0.792     .2488634    6.199846
             |
         bmi |    1.08078   .0567668     1.48   0.139     .9750546     1.19797
       _cons |   .0297049   .0405208    -2.58   0.010     .0020497    .4304909
------------------------------------------------------------------------------
```

13. (2 marks) What is the OR of high energy intake compared to low for the 3 different job types?

14. (3 marks) Based on models 3 and/or 4, is there any evidence of statistically significant effect modification? Conduct a formal hypothesis test using output from models 3 and/or 4. You should state the null hypothesis, alternative hypothesis, value of a test statistic, assumed distribution of the test statistic under the null hypothesis, the name of the statistical test you are using, and a comment on statistical significance.

Solutions

1. 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 factor of 1.08 (8and every 1 unit increase in BMI.

2. No it is not possible. That would require evaluating if there is an interaction between energy intake and BMI. This can be done by refitting the model with the relevant interaction term and subsequently performing a likelihood ratio test or a Wald test for that effect. Our decision on whether or not effect modification exists should then be based on the size and statistical significance of the interaction effect as well as knowledge of the underlying biology/physiology.

3. The p-values for the parameters representing the effect of occupation represent the pairwise comparison and we should not make a conclusion based on those tests alone. In order to test for a global (overall effect) 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 (see question 10).

4. The OR is given by $\frac{1.169}{1.793} = 0.652$

5. No, we should always be wary of interpreting associations as causal effects. In this specific case we would expect the association to be confounded by, for example, level of physical activity.

6. 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.

7. OR = $(1.064)^5 = 1.364$. The effect is not statistically significant (the scale that is used, i.e. a one unit increase or a five unit increase does not affect the significance).

8. No it is not possible to assess effect modification based on the results from model 1 and/or 2. In order to do so we would need to include an interaction term between high energy intake and attained age.

9. There is no formal test for testing 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.

10. We can perform a likelihood ratio test by testing the null hypotheses that the 2 parameters representing the effect of job type 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 is statistically greater than the likelihood for the reduced model. The test statistic is: $D : 2(lnL_{(submodel)} lnL_{(fullmodel)}) = 2(128.78 + 127.84724) = 1.86552$ 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 less than the critical value we conclude that there is no evidence that job type is statistically significant.

    In Stata

    ```
    . /*Model 1*/
    . logistic chd i.hieng i.job bmi
    [OUTPUT OMITTED]
    ```

```
. est store one
. /*Model 2*/
. logistic chd i.hieng bmi
[OUTPUT OMITTED]
. lrtest one
Likelihood-ratio test                  LR chi2(2)  =      1.87
(Assumption: . nested in one)          Prob > chi2 =    0.3935
```

11. We can retrieve the standard error from model 1 my taking the log of the confidence limits, i.e. Y = -1.44817 and Z = -0.127833. The standard error is thus given by $\frac{0.788(1.44817)}{1.96} =$ 0.336821 by re-organizing the formula for how to calculate .e.g. the lower confidence limit and solving for the standard error.

12. The constant represents the log(odds) for an individual where all covariates are at their reference level (i.e., for a driver with low energy intake and BMI = 0). The constant does not always make any sense in practice (as in this case). We can nevertheless calculate $\exp(3.63) = 0.027$. This is the estimated odds of CHD for a driver with low energy intake and BMI of zero. The estimated odds of CHD for a driver with low energy intake and BMI of 25 is given by $\exp(3.63 + 25 \times 0.079) = 0.19$.

13. For drivers the OR = 0.379, for conductors the OR = 0.379 × 1.343 = 0.509 and for bankers the OR = 0.379 × 1.242 = 0.471

14. Use a likelihood ratio test as in Question 10. The test statistic is 0.12 which follows a χ^2 distribution with 2 df. We conclude that there is no evidence of a statistically significant interaction.