# GLM Practical 7 Solutions

1. See Do file for Stata code and output, including some nice bar charts.

a. binary tobacco status and alcohol group

There seems to be an association, with non-smokers drinking less, on average, than smokers. The proportion of non-smokers in the lowest drinking category is 56%, compared to 37% among smokers. Conversely, the proportions in the highest drinking category are 2.3% and 8.6% in non-smokers and smokers respectively.

b. binary tobacco status and age

There appears to be a non-linear association between smoking and age. The highest proportion of smokers is in the 55-64 year group, with the levels of smoking decreasing in younger and older age groups, with the lowest proportion of smokers in the youngest and oldest groups.

c. alcohol and age

There also appears to be a non-linear association between alcohol intake and age. Over half of the youngest age group (25-34 years) are also in the lowest alcohol intake group (0-39 g per day). This proportion declines in each successive ten-year age group, to a minimum of 37% among both 45-54 and 55-64 year-olds. It then increases to 44% and 61% in the two oldest groups.

Discuss: Are there any associations which may affect your analysis?

There appears to be an association between tobacco intake and alcohol intake, with people with high tobacco intake tending to have higher alcohol intake than people using no or small amounts of tobacco. Those using no tobacco or in the highest tobacco intake group tend to be younger than those in the middle categories of tobacco intake. We may expect alcohol intake and age as well as tobacco intake to be associated with the outcome, oesophageal cancer. These preliminary investigations suggest that tobacco is associated with both alcohol and age and it is therefore sensible to consider these variables as potential confounders in our analysis.

2. a.

|  |  |  |
| --- | --- | --- |
|  | Non-smoker | Smoker |
| Control | 255 | 521 |
| Case | 9 | 191 |

The estimated odds ratio is (255x191)=(9x521) = 10.39 and the estimated variance of the log OR estimate, using Woolf’s method, is 1/255 + 1/191 + 1/9 + 1/521 = 0.122.

We can the 95% for the log OR estimate using . The 95% CI for the log OR is (1.66, 3.03). Taking the exponentials gives a 95% for the odds ratio is therefore (5.24, 20.60).

b. The odds of being a case is 10.39 times higher in smokers compared to non-smokers. The 95% CI for this estimate excludes the null value of 1 (by a long way). There is strong evidence for an association between use of tobacco and oesophageal cancer.

c. The mhodds command gives exactly the same point estimate as above, but a different confidence interval (ie it does not use Woolf's method). The 95% CI given is (5.11, 21.12).

3.

a.

Let Y = 1 denote a case and Y= 0 denote a control, and let X= 1 denote a smoker and X= 0 a non-smoker. Using the notation from the lecture notes, the logistic regression model for which considers cancer to be the outcome is given by:

Note we could also write this as where S=1 denotes being in the case-control sample.

b. Using the same notation as above, when smoking is the outcome the logistic regression model is:

Fitting these models using ‘glm’, gives an OR estimate of 10.39. The odds ratio estimates from models (a) and (b) are identical, as we expect. Also as we expect, the \_cons terms are different - these are the exponentiated estimates of the model intercepts. The intercept in model (a) does not have a sensible interpretation. The form of intercept α\* is given in equation 6.23. In model (b) the intercept λ\* is the log odds of being a smoker among the controls. So 2.04 is the estimated odds of being a smoker among the controls.

Using the logit (or logistic) command again gives the same point estimate. The standard error, and hence 95% CI is the same as by Woolf’s method. Note that the SE given by the logistic command, and glm with the eform option, is an attempt by Stata to report an SE on the odds ratio scale. It should not be used.

4. Note that when using the mhodds command it doesn't matter how we order the variables; this is because of the symmetry of odds ratios (see equation 6.10 in the

lecture notes). The adjusted OR is 8.59, as compared with the unadjusted OR of

10.39. That is, after adjusting for alcohol intake, the OR for being a case in smokers

compared with non-smokers is slightly attenuated (closer to 1).

5.

Discuss: Compare and contrast the results from the two models, and the two mhodds commands. What do the intercept parameters represent in each logistic model?

Note that we get an adjusted OR of 8.39. This is slightly different to that obtained from mhodds, due to a slightly different method used for the adjustment. We conclude that people who use tobacco have an odds of oesophageal cancer which is over 8 times that in non-users of tobacco, after conditioning on alcohol intake. Both the intercepts and the coefficients for the alcohol categories are different in models (a) and (b). In model (a) the intercept does not have a sensible interpretation. In model (b) the intercept is the log odds of being a smoker in controls who are in the lowest alcohol intake group.

From Model (a) we can see that increasing alcohol intake is independently associated with higher odds of oesophageal cancer after conditioning on smoking status. People consuming 120+ grams of alcohol per day had an odds of oesophageal cancer which is 22 times greater than that of those in the lowest alcohol intake group (0-39 grams per day), after conditioning on tobacco intake.

The coefficients for alcohol intake in model (b) relate to their association with smoking status, after conditioning on case/control status. People consuming 120+ grams of alcohol per day had an odds of being a smoker which is 2.3 times greater than that of those in the lowest alcohol intake group, after conditioning on case/control status. This is not likely to be an interpretation that is of interest to us.

6.

Discuss: What impact does adding age to the model have on the odds ratio for tobacco status? What are two possible reasons for any impact that you see?

Adding age group into the model has reduced the OR for tobacco intake to about 7. The change in the OR compare to the model with adjustment for alcohol intake only could be due to a combination of confounding by age and non-collapsibility. The investigations in question 1, and subject-matter knowledge suggest that age is expected to be a confounder. The inclusion of age has not had much of an effect on the odds ratios for alcohol intake. Increasing age is strongly associated with oesophageal cancer risk given tobacco use and alcohol intake.

7.

Discuss: Compare the results from this model to your results from Q6. What are the advantages and disadvantages of the two approaches?

There is still a strong association between smoking and oesophageal cancer given alcohol intake and age, though the odds ratio estimate has changed a little from that found above. When including age and alcohol intake as continuous variables, we need to choose a suitable form in which they will be entered into the model. Here we assumed that age and alcohol act linearly on the log odds of being a case. However, a non-linear association may exist in reality. In general, when using continuous variables it is be a good idea to investigate whether non-linear terms are needed, e.g. by including quadratic and cubic terms. Alternatively, age and alcohol intake could be included using flexible non-linear functions, e.g. using splines (to be mentioned briefly in session 9). The most compelling reason to include age and alcohol intake as continuous variables is that by doing so we use the full information on these variables. By categorising variables we throw away information.

8.

|  |  |  |
| --- | --- | --- |
| **Model** | **OR** | **95% CI** |
| Unadjusted | 10.39 | (5.24, 20.61) |
| Adjusted for age and alcohol intake (as categorical variables) | 7.14 | (3.43, 14.87) |
| Adjusted for age and alcohol intake (as continuous variables, assuming a linear association on the log odds scale) | 6.87 | (3.35, 14.09) |