

Biostatistics 208, Lab 9 03/04/10

Lab Summary

In this lab you will use logistic regression to explore relationships between binary outcomes and multiple risk factors.

Background

The lab will use the WCGS data used in the last lab (#8) and discussed in lecture. Download the dataset `lab8.dta` from the lab section for lab #8 from the course Web site:

<http://www.epibiostat.ucsf.edu/courses/schedule/biostatii.html>

Remember to start a log file to record your work. As always begin by listing out the variables and basic summaries for the data using the `describe` and `summarize` commands:

```
des
sum
```

The outcome considered again is `chd69`, a binary indicator of coronary heart disease (CHD) occurrence within the WCGS study period (1960 -1969).

Logistic regression for two binary risk factors

Recall from lecture that logistic regression is not required to investigate the relationship between a binary outcome and one or two categorical predictors. In such cases, frequency table methods are preferred. Let's start by investigating the association between coronary heart disease (CHD) and type A behavior (discussed in lecture), and focus on the question of whether this may be confounded by smoking. Start by creating a binary variable that distinguishes smokers from non-smokers based on reported number of cigarettes smoked per day (`ncig`):

```
gen smoke = ncig
recode smoke 0 = 0 1/max = 1
label define smlab 0 "nonsmoker" 1 "smoker"
label val smoke smlab
```

Now use the `cs` command to investigate the association between the primary outcome (`chd69`) and the binary indicator of type A behavior (`dibpat`, coded as 1 for type A, and 0 for type B behavior) stratified by the binary variable `smoke`:

```
cs chd69 dibpat, by(smoke) or
```

The last command reports separate odds ratios for the relationship between `dibpat` and `chd69` for smokers and non-smokers. In addition, a summary odds ratio (labeled "M-H combined" obtained from the Mantel-Haenszel procedure) and the results of a homogeneity test of the null hypothesis that these odds ratios are equal are provided. Rejection of the hypothesis of homogeneity can be interpreted as evidence that interaction may be present. If this is the case, the summary odds ratio and the associated "Test that combined OR = 1" are not of interest (i.e. the presence of interaction implies that we need to summarize the relationship using separate odds ratios for smokers and non-smokers).

Question 1. Note the estimated odds ratios for nonsmokers and smokers and their 95% confidence intervals, and the estimated summary (combined) odds ratio and its 95% confidence

interval. What do the results of the test for homogeneity and inspection of the odds ratios allow you to conclude about the effect of smoking on the association between type A behavior and CHD? Is there any evidence for interaction? If we assume that no interaction is present, is there any evidence that smoking has a confounding influence on the measured association between `chd69` and `dibpat`?

Now fit a logistic regression model to examine this same issue:

```
logistic chd69 dibpat smoke
```

Question 2. How do the estimated odds ratio and 95% confidence interval for `dibpat` compare to their counterparts from the results reported above?

Notes: We don't expect an exact correspondence between the frequency table and logistic regression results when there are multiple risk factors involved. (This is related to the fact that the logistic model represents the effect of risk factors on the outcome as additive on the log odds scale, and the table-based analysis makes no such assumptions.) Also, a more complete analysis of the relationship between CHD, smoking and behavior type might involve assessing the effect of smoking using the quantitative measure `nc ig` in the dataset.

Investigating two-way interactions

Two risk factors are said to interact in their effect on a binary outcome when their measured effects are interdependent. As discussed in lecture, this phenomenon can be difficult to detect and interpret. Further, interaction is model- dependent: two factors may appear to interact when their effects are measured using a multiplicative risk model (e.g. logistic regression), but not when risk is measured on an alternative scale (e.g. additive or excess risk). Despite these complications, understanding basic approaches to fitting and interpretation of models including interactions is often crucial in data analyses. We begin by investigating the relationship between `arcus` and `age` in the WCGS data. The fundamental research question is whether the observed association between `arcus` and CHD is age-dependent (and the companion issue of whether the CHD-age relationship depends on `arcus`).

For a preliminary assessment of the relationship between these factors and the outcome, create a binary version of `age`, separating the sample into two ten-year age groups:

```
gen dage = age
recode dage 39/49=0 50/59=1
label define dagelab 0 "39-49" 1 "50-59"
label val dage dagelab
```

Now use the `cs` command to examine the relationship between `chd69`, `dage` and `arcus`:

```
cs chd69 arcus, or
cs chd69 dage, or
cs chd69 arcus, by(dage) or
cs chd69 dage, by(arcus) or
```

Question 3. What do the above results allow you to conclude about the effect of age on the association between `arcus` and CHD risk? About the effect of `arcus` on the association between age and CHD risk?

The same analysis can be performed with logistic regression. (Although the previous methods would suffice in this example, the regression approach would be definitely preferred if additional predictors and/or continuous predictors were being considered.) Start out by creating a new variable containing the product of `arcus` (coded 0-1) and the binary age variable `dage` (also coded 0-1):

```
gen arcdage = arcus*dage
```

By definition, this variable takes on the value one only when both `arcus` and `dage` are equal to one, and equals zero for all other combinations of values of `arcus` and `dage`. Now fit a logistic regression model including `arcus`, `age` and `arcdage`:

```
logistic chd69 arcus dage arcdage
```

The estimated coefficient and 95% confidence interval for `arcdage` in this model provide an assessment of evidence of interaction between `arcus` and `age` in their effects on CHD risk. If there is significant evidence that the true coefficient is nonzero, then we can conclude that a multiplicative interaction between these variables is present, and that their influence on CHD risk must be considered jointly. (Note that the logistic model results should be very similar to those obtained from the test of homogeneity referenced above.)

Question 4. Compare these results with those obtained from the last two `cs` commands and comment on any correspondences. What do the results imply about the presence of interaction?

Notice that the estimated odds ratio for `arcus` is the odds ratio comparing CHD risk between subjects with `arcus` to those without `arcus` among the younger age group. Similarly, the estimated odds ratio for `dage` gives the odds ratio for older age among those without `arcus`. The corresponding odds ratios for the older age group and those with `arcus` (respectively) can't be read directly from the logistic output without further calculation. To see why this is true, let's write down the regression equation from the interaction. Re-fit the previous model with the `coef` option to view the regression coefficients:

```
logistic chd69 arcus dage arcdage, coef
```

The regression equation for this model (on the log odds scale) is:

$$\log(\text{odds of CHD}) = -2.882853 + 0.8932677 \text{ dage} + 0.6479628 \text{ arcus} - 0.5920552 \text{ arcdage}$$

Now we can use the `display` command and the coefficients from the above equation to compute the log odds of CHD for individuals with `arcus` and in the older age group. The actual coefficients from the last model fitted are stored by Stata and can be accessed using the `_b[]` commands shown below. Plugging `arcus = 1`, `dage = 1` and `arcdage = 1` into the above equation then yields the desired log odds:

```
display _b[_cons] + _b[arcus]*1 + _b[dage]*1 + _b[arcdage]*1
```

A similar calculation gives the log odds for CHD for individuals without `arcus` in the older age group:

```
display _b[_cons] + _b[arcus]*0 + _b[dage]*1 + _b[arcdage]*0
```

Now, compute the log odds ratio for arcus among older participants, which is the difference between these two numbers. (Recall that the difference between the logarithms of two numbers is equivalent of the logarithm of the ratio of the first to the second number.) Notice that the difference between the above two log odds expressions is just the sum of the regression coefficients for `arcus` and `arcdage`:

```
display _b[arcus] + _b[arcdage]
```

Finally, exponentiate to get the desired odds ratio:

```
display exp(_b[arcus] + _b[arcdage])
```

Compare the result to the corresponding estimate in the output from the last `cs` command.

The following table (introduced in Lab #5 and also discussed in lecture) provides an alternate way to identify the coefficients involved in this particular odds ratio:

<code>arcus</code>	<code>dage</code>	<code>_cons</code>	<code>arcus</code>	<code>dage</code>	<code>arcdage</code>
Yes	Yes	1	1	1	1
No	Yes	1	0	1	0
Difference		0	1	0	1

As we have seen in earlier labs and in lecture, Stata makes such calculations much easier with the `lincom` command. For example, to repeat the calculation of odds ratio just computed:

```
lincom arcus + arcdage
```

This command just adds the requested coefficients and exponentiates the result. In contrast to linear regression, exponentiation of the results of `lincom` is the default after a logistic model is fitted. Notice that the output of the `lincom` command also gives a 95% confidence interval for the true odds ratio as well as a test of the hypothesis that the true log odds ratio is 0 (or, equivalently, that the true odds ratio is 1).

The following commands compute the remaining odds ratios for this example:

```
lincom dage + arcdage
lincom arcus
lincom dage
```

Question 5. Verify the definition of each of these odds ratio estimates by constructing a table like the one shown above, referring back to the equation for the logistic model given above. Check the results by comparing the estimates to the results of the last two `cs` commands issued above.

Note: An alternative (and simpler) way to fit the interaction model that avoids the need to generate the product variable `arcdage` is to use the following `logistic` command (in Stata 11):

```
logistic chd69 i.arcus##i.dage, coef
```

This command will work for fitting models two-way interactions between categorical predictors in Stata 11. Both the main effect terms and the interaction are included automatically.

The `lincom` command can be applied after fitting the model including the “i.” prefix. However, the names for parameters provided to `lincom` must agree with the output from the model fitted. Here is an example `lincom` command for the model specified by the `logistic` command above:

```
lincom 1.arcus + 1.arcus#1.dage
```

Two-way interactions involving continuous variables

The results of the previous analysis depend on an arbitrary dichotomization of age. The logistic model can be used to look at the same issue using age as a continuous variable. Recall from the last lab that a logistic model for a linear relationship between the log-odds of CHD and age (in the absence of other factors) seemed reasonable. This suggests that a linear model may be a good starting point in exploring the arcus-age interaction further. Fitting a logistic model involving an interaction between a continuous and a binary risk factor involves the same procedure use above for two binary risk factors. Begin by generating a variable containing the product of these factors:

```
gen arcusage = arcus*age
```

Fit a logistic regression model including `arcus`, `age` and the new variable:

```
logistic chd69 arcus age arcusage, coef
```

Examine the output, paying special attention to the coefficient and 95% confidence interval for the product variable `arcusage`.

As mentioned above, an alternative way to fit this model which avoids construction of the product variable is to use the `xi` command with `logistic`:

```
logistic chd69 i.arcus##c.age, coef
```

Because the rest of the questions below are based on the results of the fitted model, make sure that you issue the following command before proceeding:

```
logistic chd69 arcus age arcusage, coef
```

Question 6. What do the results imply about the presence of an arcus-age interaction as measured by the coefficient for the product variable `arcusage` and it's 95% confidence interval?

In contrast to the model involving a binary version of age, this model specifies different levels of CHD risk associated with arcus for an arbitrary specified age. The overall regression equation is:

$$\log(\text{odds of CHD}) = -6.788086 + 0.089647 \text{ age} + 2.754185 \text{ arcus} - 0.0498298 \text{ arcusage}$$

The `lincom` command (introduced above) can be used to estimate the odds of CHD associated with arcus for a particular age (say 55) among subjects with arcus =1 as follows:

```
lincom _cons + age*55 + arcus + arcusage*55
```

Note: The regression equation above gives the logs odds of the outcome for any specified values of the predictors. Since `lincom` always exponentiates results when applied after a logistic model

is fit, the `lincom` results give odds in this case. The default Stata output incorrectly labels the result of the above command as an “odds ratio”. (When used following `logistic`, `lincom` labels all results as “odds ratios”.) The same calculation for `arcus=0` gives the odds for being arcus-free and age 55:

```
lincom _cons + age*55
```

Combining these last two calculations, we can see that their difference yields the log odds ratio CHD comparing outcome odds in 55-year old individuals with arcus to the corresponding odds in 55-year olds without arcus.

This odds ratio can be calculated directly as follows:

```
lincom arcus + arcusage*55
```

Question 7. Compute the above odds ratio for a 40 year old. Given what we know about the age range of subjects in the WCGS, does it make any sense to use the above model to estimate this odds ratio for an individual of age 20?

Using the same approach just described, you can write down separate regression equations for the relationship between CHD and age for the two arcus groups as follows:

For those with no arcus, plug `arcus=0` into the above regression equation, yielding the following:

$$\log(\text{odds of CHD}) = -6.788086 + 0.089647 \text{ age}$$

Similarly, for `arcus=1`:

$$\begin{aligned} \log(\text{odds of CHD}) &= (-6.788086 + 2.754185) + (0.089647 - 0.0498298) \text{ age} \\ &= -4.033901 + 0.0398172 \text{ age} \end{aligned}$$

These equations define two separate linear relationships between the log odds of CHD and age, one for each group defined by arcus status. Stata will compute the predicted outcome probabilities for both equations via the `predict` command, and you can store the results in a new variable ("`prb`") for graphing:

```
predict prb
```

Stata defaults to providing predictions from logistic models expressed as outcome probabilities (rather than log odds), and estimates these for each individual in the current dataset. We can convert these to the log odds scale and plot against age to examine the relationship for the two groups:

```
gen lor = log(prb/(1-prb))  
scatter lor age
```

(Note that we can also ask Stata to return predictions on the log odds scale directly by specifying the `xb` option of `predict`.) We prefer to look at this relationship on the log-odds scale because the model is linear on that scale. Now place the predicted values for the two arcus groups in separate variables so separate lines can be plotted for each group (and so they can be used later):

```
gen lor0 = lor if arcus==0  
gen lor1 = lor if arcus==1
```

Plot the results as follows:

```
twoway (connected lor0 age, sort msymbol(none)) (connected lor1  
age, sort msymbol(none))
```

Question 8. What do the plotted lines reveal about the relationship between CHD risk and age in the two arcus groups?

A potential problem with the previous model is that it assumes that the relationship between CHD risk and age is linear for individuals with and without arcus. Although a linear relationship was found in the last lab to provide a reasonable description for the entire sample, this does not necessarily imply that linearity applies in the two groups defined by arcus. To look at this further, we apply the smoothing approach taken in the last lab to the two groups separately.

The following commands fit separate smoothed estimates for the relationship between the log odds of CHD and age in each of the two arcus groups, and store the results in new variables (`sm0` & `sm1`). The `nograph` option suppresses the graphs because we want to plot these later.

```
lowess chd69 age if arcus==0, logit nograph gen(sm0)  
lowess chd69 age if arcus==1, logit nograph gen(sm1)
```

Now graph the smoothed estimates along with the linear estimates saved above. (Sort the observations by age first to facilitate the plot.)

```
twoway (connected lor0 age, sort msymbol(none)) (connected lor1  
age, sort msymbol(none)) (connected sm0 age, sort msymbol(none))  
(connected sm1 age, sort msymbol(none))
```

Question 9. What do you conclude from the graph about the adequacy of the linear representation of age in the two groups?