

## Biostatistics 208, Lab #8 02/25/10

### Lab summary

In this lab you will learn how to fit and interpret simple logistic regression models using Stata. You will use the `logistic` command to fit models to explore the relationship between a binary disease outcome and a single categorical or continuous risk factor.

### Background

The lab will use the WCGS data discussed in lecture. Download the dataset `lab8.dta` from the Lab #8 section of the course Web site (<http://www.biostat.ucsf.edu/biostat208/>).

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

```
des
sum
```

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

### Assessing association between a binary outcome and a single categorical risk factor

You already know that logistic regression isn't the only tool that can be used to investigate the association between a binary outcome and a categorical risk factor. As an example of an alternate procedure, use the `cs` command to investigate the association between the primary outcome (`chd69`) and the binary indicator of having *arcus senilis* at baseline (`arcus`):

```
cs chd69 arcus, or
```

The `cs` (Cohort Study) procedure is a part of the `epitab` suite of commands Stata provides to analyze tabular frequency data from a variety of epidemiological study designs. The `or` option requests that odds ratios be included among the reported measures of association. Note the measures of association provided and the results of the chi-squared test for independence of `arcus` and `chd69`.

**Question 1.** What do you conclude about the association between `chd69` and `arcus` based on the odds ratio and its 95% confidence interval? How do these conclusions correspond to the result of the chi-squared test provided?

### Logistic Regression for a Single Categorical Risk Factor

Now use the `logistic` command to repeat the above analysis using a logistic regression model. The syntax of this command specifies the outcome variable first, followed by risk factor(s). For our chosen outcome and risk factor, type:

```
logistic chd69 arcus
```

Note that `logistic` requires outcomes to be coded as binary numeric variables and assumes that the values for non-outcomes are zero. The outcome `chd69` is already coded as 0/1 for outcomes/non-outcomes. If `chd69` had an alternative coding (e.g. no-CHD=1, CHD=2), the `logistic` command would return an error message about the incorrect coding. Since `arcus` is already coded as 0/1, the `logistic` procedure automatically knows how to treat this variable as a two-level categorical predictor in the regression model. In the case where `arcus` was coded using two different numerical values (e.g. no-

arcus=1, arcus=2), the `xi` command should be used with `logistic`, similar to the way `xi` was used with `regress` in previous labs.

Compare the results from the last command issued to the results obtained above using the `cs` procedure.

**Question 2.** Does fitting the logistic regression model add anything to our understanding of the association between these two variables (considering the results from the prior analysis based on the `cs` command)?

### Logistic regression for a single continuous risk factor

Now we'll examine the association between CHD risk and age (in years) of participants at the beginning of the study. A good place to begin the analysis is an exploration of the distribution of the variable `age`. As in standard linear regression, no distributional assumptions are required for independent variables in logistic regression. (Note that age is always positive and takes on a limited range of values, so we would not expect it to follow a normal distribution.) However, it is still useful to examine the distribution of variables empirically (to identify outliers, look at the range, etc.). Use the following commands (included in the `eda` procedure from Lab #3) to do this:

```
graph box age
```

```
histogram age
```

There are at least two ways to get a preliminary assessment of the relationship between CHD risk and age before actually fitting a logistic model. The first is to group age into categories and look at how risk varies across these. Create a five-level categorical version of `age` called `agec`:

```
gen agec = age
```

```
recode agec 35/40=0 41/45=1 46/50=2 51/55=3 56/60=4
```

The following two optional commands create labels for the categories of the newly created variable `agec`, to make output involving it more interpretable:

```
label define agelab 0 "35-40" 1 "41-45" 2 "46-50" 3 "51-55" 4 "56-60"
```

```
label val agec agelab
```

Use the `tabulate` command to cross-tabulate `agec` and `chd69` (including column percentages), and examine the proportions of participants with CHD across age strata:

```
tabulate chd69 agec, col
```

**Question 3.** What do the proportions of participants with CHD reveal about the association between CHD risk and age?

Stata also provides `tabodds`, an analog of the `cs` command that is appropriate for computing measures of association between binary outcomes and categorical predictors with multiple categories. Use `tabodds` as follows:

```
tabodds chd69 agec, or
```

The test for homogeneity of odds in the output uses a chi-squared statistic to test the null hypothesis that all the odds are the same across age strata. A significant result indicates evidence for variation in disease risk (as measured by odds of disease) across strata. The results reported for the score test for trend in odds is a test for linearity of the log odds. A significant result indicates that a linear trend in the log odds with age is supported by the data (against a null hypothesis of no trend). This is a stronger result than demonstrating heterogeneity of odds across strata and indicates a systematic change in disease risk with increasing strata. The trend test only makes sense when applied to a categorical factor with ordered categories (e.g. `agec`).

**Question 4.** What do the results from the `tabodds` analyses allow you to conclude about the dependence of CHD risk on age?

Now we can use the `logistic` command to fit a regression model and obtain similar results:

```
logistic chd69 i.agec
```

Compare these results with those obtained using `tabodds`. Note that for a single categorical risk factor the logistic procedure does not provide any additional information. Also, note that applying `logistic` directly to `agec` without using the `i.` prefix would fit a model which assumes that the log odds of CHD change linearly in the values of the categorical `agec` variable (0,1,2,3). This is a strong assumption and usually not desirable for categorical risk factors.

Suppose we wanted to get the odds ratio comparing the outcome odds between the 2<sup>nd</sup> and 3<sup>rd</sup> age groups. Stata 11 can do this using the following syntax:

```
logistic chd69 ib2.agec
```

The odds ratios in this model are all estimated using category “2” as the comparison group. The following commands in Stata 10 accomplish the same task (the 3<sup>rd</sup> command resets the “0” level as baseline):

```
char agec[omit] 2
```

```
xi: logistic chd69 i.agec
```

```
char agec[omit]
```

We may still be concerned that our chosen (fairly crude) grouping of age obscures some of the actual relationship with CHD risk. In addition to repeating the above steps with alternate (perhaps finer) groupings of age, we can use the scatterplot smoothing methods for binary outcomes to examine the relationship without imposing any groupings (only some assumed smoothness). The following command may take a while to complete:

```
lowess chd69 age, bwidth(0.5) logit generate(chdsm)
```

The `lowess` command above smooths the log odds of the outcome probability (also called the *logit*) against age using a nonparametric regression approach (lowess) and saves the results in a new variable `chdsm`. The `bw(0.5)` option specifies that the nearest half of the observations be used in computing the value of the smoothed estimate at each age. We will use this saved estimate again later.

**Question 5.** What does the resulting plot reveal about the relationship between (the log odds of) CHD risk and age? If you want to see the effects of specifying different amounts of smoothing on the estimated relationship, try again for `bw(0.7)` and `bw(0.8)`, but omit the `gen()` option.

Now fit a standard linear logistic regression model for the risk factor age:

```
logistic chd69 age
```

The logistic model specifies a strictly linear relationship between the log odds of CHD and age. The commands below help us visualize this, and will allow us to compare the results with those from the smoothed estimate saved above.

Note that the default logistic output only shows us the odds ratio for chd69 associated with a one-unit (i.e. one year) change in age. We can view the actual intercept and slope for the linear relationship between the log odds of CHD (the outcome variable) and age by re-fitting the same model and specifying the `coef` option:

```
logistic chd69 age, coef
```

Note the intercept and slope coefficients for this linear model. In addition to the coefficient estimates, a graphical representation of the fitted linear model is needed to compare to the smoothed estimate produced above. After refitting the model again using logistic, the second command below computes the predicted probabilities of chd69 for each individual given their age from the model and saves the results in a new variable `p`. The third command transforms these probabilities into log odds and saves them in the created variable `lp`. (Note that we want `lp` to be on the log scale because the mode assumes linearity with age on that scale.)

```
logistic chd69 age
predict p
gen lp = log(p/(1-p))
```

Now graph the predicted log-odds of CHD from the smoothed estimate and the logistic model on the same plot:

```
twoway (connected chdsm age, sort msymbol(none)) (connected lp age,
sort msymbol(none))
```

**Question 6.** What can you conclude about the adequacy of the linear logistic model for the relationship between CHD risk and age? Do you have an explanation for the apparently increased CHD risk for the youngest (aged 39) participants (relative to the other "younger" age groups)? The risk in this group (as measured by the odds of developing CHD) can be viewed directly using the command

```
tabodds chd69 age
```