Biostatistics 208, Lab #8 02/25/16

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.

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 "`i.`" prefix should be used with the predictor specification in the `logistic` command, similar to the way "`i.`" 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 not revealed in the parallel 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 predictor variables in logistic regression. However, it is still useful to examine the distribution of variables empirically (to identify outliers, look at the range, etc.). Use the following commands to do this:

```
graph box age, over(chd69) name(age_box)

twoway (kdensity age if chd69==0, bw(3) area(1) lpattern(solid))
(kdensity age if chd69==1, bw(3) area(1) lpattern(longdash)),
ytitle("Density") xtitle("age") legend(order(1 "no CHD" 2 "CHD"))
name(age_den)
```

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`:

```
recode age 36/40=0 41/45=1 46/50=2 51/55=3 56/60=4, gen(agec)
```

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 "36-40" 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`). Note that the test is based on the numerical codes used for the categorical predictor, and does don't assume equal spacing. You can check that equivalent equally spaced categories coded with the midpoints of the age groups given above (i.e. 38, 43, 48, 53, 58) yield the same results as the original codes. However, choosing codes with unequal spacing would be lead to a different test result.

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
```

As in linear regression, applying the `logistic` command directly to `agec` without using the `i.` prefix would fit a model which assumes that the log odds of CHD change linearly in the assigned numerical values for the categorical `agec` variable (0,1,2,3,4) - a very strong assumption. Similarly, we can test for heterogeneity of odds ratios across categories of age as follows:

```
testparm i.agec
```

Here, `testparm` performs a Wald test of the hypothesis that the log-odds ratios for the four highest age levels are all zero (or, equivalently, the corresponding odds ratios are one). Note that this will generally not provide the same results as the (preferred for smaller samples) likelihood ratio (LR) test (to be covered next week). Since this model includes a single categorical predictor, the default LR test provided in the model output evaluates the same hypothesis. In this case, the results for both approaches are almost identical (as might be expected for a fairly large sample).

We can also test for linear trend using the same approaches used for the linear model:

```
contrast q(1).agec, noeffects

test -1.agec + 3.agec + 2*4.agec=0
```

Compare these results with those obtained using `tabodds`. Note that for a single categorical risk factor, using the `logistic` procedure does not provide any additional benefit. However, the test statistics and *p*-values differ somewhat reflecting the different test procedures used. Note that the `contrast` statement in the test command above is exactly the same as what we would use for a linear model (see section 4.3.5 of the text). We can also use contrast to evaluate evidence for departure from linear trend:

```
contrast q(2/4).agec, noeffects
```

We can get the odds ratio comparing the outcome odds between the 2nd and 3rd age groups using the following `lincom` command:

```
lincom 3.agec-2.agec
```

An alternative is to use an alternate choice of the baseline category with `i.`:

```
logistic chd69 ib2.agec
```

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 `lowess` to examine the relationship as a smooth function, without imposing any groupings.

```
lowess chd69 age, bw(0.5) logit generate(chdsm) name(age_lws)
```

The `lowess` command above smooths the log odds of the outcome probability (also called the *logit*) against age 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
```

This 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 shows us the odds ratio for `chd69` associated with a one-unit (i.e. one year) change in age, and the odds of CHD for an individual of age 0 (labeled `_cons`). We can view the coefficients on the log odds scale by specifying the `coef` option (or, equivalently, using the `logit` command in place of `logistic`):

```
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 log odds of `chd69` for each individual given their age from the model and saves the results in a new variable `lp`. (Note that we want `lp` to be on the log odds scale because the mode assumes linearity with age on that scale. By default, `predict` following a `logistic` fit returns predicted probabilities.)

```
predict lp, xb
```

To supplement the above exploratory assessment of linearity, we can also use a restricted cubic spline to account for nonlinearity, and perform a significance test for departure from linearity.

```
* generate restricted cubic spline basis and fit model
mkspline agesp=age, cubic
logistic chd69 agesp*
predict lps, xb
* test for nonlinearity of log outcome odds in age
test agesp2 agesp3 agesp4
```

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

```
twoway (connected chdsm age, sort msymbol(none)) (connected lp age,
sort msymbol(none)) (connected lps age, sort msymbol(none)),
legend(order(1 2 3) label(1 "lowess") label(2 "linear") label(3
"spline") pos(6) ring(0)) name(chdage_pr)
```

Question 6. What can you conclude about the adequacy of the linear logistic model for the relationship between CHD risk and age? Is the apparently increased CHD risk for the youngest (aged 39) participants (relative to the other "younger" age groups) of any concern?

The CHD odds for all ages and associated 95% CIs can be viewed directly using the command:

```
tabodds chd69 age, ciplot name(ciplt)
```

Plotting smoothed outcome estimates with 95% confidence intervals

In some cases, you'll want to present the estimated relationship between an outcome and a continuous predictor that includes 95% confidence limits to illustrate uncertainty. For models including spline terms, this can be challenging using default Stata commands. Here we illustrate the use of the downloadable `xblc` command to generate a cubic spline-based estimate of the dependence of the log odds of CHD on age, along with its 95% confidence limits. The `xblc` command can be located and installed using the following command (look for the package named `st0215_1`):

```
findit xblc
```

The commands for plotting follow:

```
* generate distinct levels of age

levelsof age

* generate log-odds estimates and pointwise 95% CI for unique ages
based on the previously fitted spline model, save results as new
variables for graphing ("x" represents age)

xblc agesp*, covname(age) at(`r(levels)') generate(x losp lolb loub)

* plot, with style options suitable for printing/presentation

twoway (rarea lolb loub x, color(gray)) (connected losp x, sort
msymbol(none) color(black) lpattern(solid)) , xtitle("Age") ytitle("Log
Odds of CHD") legend(order(2 1) label(1 "95% CI") label(2 "Log odds")
pos(5) ring(0) region(lwidth(none))) plotregion(style(none))
scheme(s1mono) name(chdage_lsp)
```

Although it's easier to diagnose nonlinearity on the log-odds scale, presenting graphical summaries of nonlinear relationships from logistic models are often more interpretable for presentation on the probability scale.

```
* Convert to probability scale from log odds & plot again

gen prsp = exp(losp)/(1 + exp(losp))

gen plb = exp(lolb)/(1 + exp(lolb))

gen pub = exp(loub)/(1 + exp(loub))

twoway (rarea plb pub x, color(gray)) (connected prsp x, sort
msymbol(none) color(black) lpattern(solid)) , xtitle("Age")
ytitle("Probability CHD") legend(order(2 1) label(1 "95% CI") label(2
"Probability") pos(6) ring(0) region(lwidth(none)))
plotregion(style(none)) scheme(s1mono) name(chdage_psp)
```