Biostatistics 208, Lab 9 03/08/18

Lab Summary
In this lab you will use logistic regression to explore relationships between binary outcomes and multiple risk factors, focusing on interpretation of interaction, and on likelihood ratio tests.

Background
The lab will again use the WCGS data. Download the dataset `lab9.dta` from the lab section for lab #9 from the course Web site.

The primary outcome 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 usually required to investigate the relationship between a binary outcome and one or two categorical predictors with small numbers of categories. In such cases, frequency table methods may suffice. 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 `by(smoke)` option of the `cs` command requests 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`", and 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 (multiplicative) 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 `ncig` in the dataset.

Investigating two-way interactions

Two risk factors are said to interact in their effect on a binary outcome when their estimated 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 be preferred if additional predictors and/or continuous predictors were being considered.) Fitting a logistic model including interaction follows the same approach we used for linear regression models.

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

The estimated odds ratio and 95% confidence interval for the Stata-generated product variable `arcus#dage` 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 OR is different than 1, 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 i.arcus##i.dage, coef
```

The regression equation (*not a Stata command*) for the log odds of CHD is:

log(odds) = -2.882853 + 0.8932677 `dage` + 0.6479628 `arcus` - 0.5920552 `arcus*dage`

Now we can use the display command and the coefficients from the above equation stored by Stata 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 `arc*dage` =1 into the above equation then yields the desired log odds. In Stata's notation from the fitted model:

```
display _b[_cons] + _b[1.arcus]*1 + _b[1.dage]*1 + _b[1.arcus#1.dage]*1
```

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

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

Now, compute the log odds ratio comparing odds of CHD for participants with arcus to those without arcus, restricted to 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 `arc*dage`:

```
display _b[1.arcus] + _b[1.arcus#1.dage]
```

Finally, exponentiate to get the desired odds ratio:

```
display exp(_b[arcus] + _b[1.arcus#1.dage])
```

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:

arcus	dage	_cons	arcus	dage	arcus#dage
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 1.arcus + 1.arcus#1.dage
```

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 1.dage + 1.arcus#1.dage
lincom 1.arcus
lincom 1.dage
lincom 1.darcus + 1.dage + 1.arcus#1.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. What is the interpretation of the last OR estimate requested?

As discussed in lecture, the logistic model evaluates multiplicative interaction because the outcome is linked to the predictors on the log odds scale. For a simple model including categorical predictors, a quick way to evaluate the presence of interaction on an additive scale is to fit a regression model that is linear on the probability scale (and the coefficients have a direct risk difference interpretation). We will discuss this model in more detail in lecture 11. For now, let's use it to look at the effect of the arcus-age interaction again:

```
binreg chd69 i.arcus##i.dage, rd
```

Question 6. Do you see evidence for an additive interaction?

We'll revisit this interaction on an additive scale again below, when we consider age as a continuous predictor.

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. In this case, the `c.` abbreviation is used for the continuous predictor age:

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

Question 7. What do the estimated coefficient for the product variable and its 95% confidence interval imply about the presence of a possible arcus-age interaction?

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(odds of CHD) = -6.788086 + 0.089647 `age` + 2.754185 `arcus` - 0.0498298 `arcus*age`

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 + c.age*55 + 1.arcus + 1.arcus#c.age*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 + c.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 1.arcus + 1.arcus#c.age*55
```

Question 8. Compute the above odds ratio for a 40 year old. Does it make any sense to use the above model to estimate the corresponding 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(odds of CHD) = -6.788086 + 0.089647 `age`

Similarly, for `arcus`=1:
log(odds of CHD) = (-6.788086 + 2.754185) + (0.089647 - 0.0498298) `age`
= -4.033901 + 0.0398172 `age`

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 lor, xb
```

The `xb` option requests predictions of the log odds of outcome occurrence for each individual in the dataset. (Using `predict` with no options returns estimated outcome probabilities.)

Plot the results for the two groups separately as follows:

```
twoway (line lor age if arcus==0, sort lcol(red) lw(medthick)
msymbol(none)) (line lor age if arcus==1, sort lcol(blue)
lw(medthick) msymbol(none)), ytitle("log odds(CHD)")
legend(order(1 "no arcus" 2 "arcus") pos(6) ring(0))
```

Note: The command above should be entered without any line breaks.

Question 9. 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 plot specified below requests separate lowess estimates for the relationship between the log odds of CHD and age in each of the two arcus groups, along with the lines plotted in the previous graph.

```
twoway (line lor age if arcus==0, sort lcol(red) lw(medthick)
msymbol(none)) (line lor age if arcus==1, sort lcol(blue)
lw(medthick) msymbol(none)) (lowess chd69 age if arcus==0, logit
sort bw(.8) lcol(red) lw(medthick) msymbol(none)) (lowess chd69
age if arcus==1, logit sort bw(.8) lcol(blue) lw(medthick)
msymbol(none)), ytitle("log odds(CHD)")  legend(order(1 "no
arcus" 2 "arcus") pos(6) ring(0))
```

Note: The command above should be entered without any line breaks.

Question 10. What do you conclude from the graph about the adequacy of the linear representation of age in the two groups?

Finally, let's revisit the presence of additive interaction for the relationship between age and arcus, again assuming that a linear representation of age is adequate.

```
binreg chd69 i.arcus##c.age, rd
```

Question 11. Is there evidence for additive interaction in this model?

Note that fitting the risk difference model frequently doesn't work numerically because of the need to constrain estimated probabilities to be between zero and one. In this case, the approach based on estimating the relative excess risk for interaction ($RERI_{RR}$) (e.g. by fitting a relative risk regression model, or using logistic regression and the $RERI_{OR}$ analog in the case of rare outcomes) for detecting the presence of additive interaction discussed in lecture 9 can be used. (See the discussion for this lab for a demonstration of the approach for this example.)

Likelihood ratio tests for logistic regression models

As discussed in lecture, the LR test can be used to assess the contribution of any predictor or group of predictors to the overall fit of the model, similar to the use of the *F* test in standard linear regression. This is accomplished using the `lrtest` command in Stata. The command is issued first to save the likelihood for a larger model, and again to compare the likelihood for smaller model(s) based on a subset of the predictors in the larger model.

As an example, let's assess the contribution of behavior pattern as measured by the four-level behavior pattern indicator `behpat` to a model already containing `chol, sbp, age, smoke and bmi`. Note that `behpat` is coded as 1, 2, 3 or 4, corresponding to behavior classifications A1, A2, B3, and B4, respectively, so we need to include it as a categorical predictor using the "`i.`" syntax:

```
logistic chd69 chol sbp age smoke bmi i.behpat
```

Type A behaviors have the lowest numbered categories in `behpat`. (You can see the numerical values associated with the labels using the "`codebook behpat`" command.) Recall that Stata chooses the lowest (type A1 in this case) as the default reference level. If we wish to switch the reference level to type B4, (so that the risk associated with type A behavior will appear elevated) we can create a new categorical variable in which the type A and B categories are reversed:

```
gen behpat2 = 5 - behpat
```

Note: an equivalent (but slower) way to create a variable with reversed category labels is via the `recode` command.) The following command shows how the two versions of the behavior pattern variable match up:

```
tab behpat behpat2
```

Now re-fit the model with the new behavior variable:

```
logistic chd69 chol sbp age smoke bmi i.behpat2
```

In this model, the odds ratios compare CHD risk among individuals with types B3, A2 and A1 behavior types to those with type B4 behavior. Now let's proceed with the LR test using this new representation of behavior pattern. To begin, use the `estimates` (`est`) command to save the likelihood of this (larger) model, and reference the result with the label `A1`

```
estimates store A1
```

Next fit the reduced model without `behpat2`, and run the LR test comparing this smaller model to the larger model that includes `behpat2`. Recall that for the LR test to make sense, all

concerned models need to be fitted on the same set of observations. Since the only change in the reduced model is the exclusion of `behpat2`, we will be fine as long as this variable is free of missing values (a check using "`codebook behpat`", confirms that this is the case).

```
logistic chd69 chol sbp age smoke bmi
lrtest A1
```

Question 12. What do you conclude about the contribution of behavior pattern in our model for CHD risk?

Now try to use `testparm` to look at the same hypothesis:

```
quietly logistic chd69 chol sbp age i.smoke bmi i.behpat2
testparm i.behpat2
```

The results are quite similar in this case. Note that `testparm` uses a different approach to evaluate this hypothesis (a Wald test), and, as mentioned in lecture, the LR test would generally be preferred in smaller samples.

Examination of the estimated odds ratios for `behpat2` in the above model indicates that differentiating between the two subtypes (1 and 2) of patterns A and B does not seem to contribute much to predicting CHD risk. (Why?) We can evaluate whether a dichotomous indicator of type A vs. B behavior would do as well via an additional LR test as follows: First, note that the dataset already contains a dichotomous indicator of type A (either A1 or A2) behavior in the variable `dibpat` used above. Re-fit the last model including `behpat2`, replacing the latter with `dibpat`:

```
logistic chd69 chol sbp age 1.smoke bmi 1.dibpat
```

Notice that if the coefficients (and corresponding odds ratios) in the model including the four-level factor `behpat2` were equal within levels of type A and B behavior, that model would be identical to the model including the two-level behavior indicator just fitted. (i.e. the odds ratio for type B3 would be 1, and the odds ratios for types A1 and A2 would be equal.) Thus, a test of the hypothesis that the true coefficients are equal within levels of behavior types would accomplish our goal of evaluating whether it's necessary to model behavior with the four subtypes versus the two-level version. Because the reduced model based on `dibpat` is a special case of the model including `behpat2`, we can view it as *nested* within the latter and use the LR test to compare them. This test evaluates the hypothesis of interest. Since we've already saved the likelihood from the more complex model, the test is simple to perform:

```
lrtest A1
```

Question 13. What does the LR test just performed allow you to conclude about the relative contributions of the two different representations of behavior pattern in our model for CHD risk?