

Lab #10

Lab Summary

In this lab you will learn to use some simple assessment techniques for logistic regression models and explore the use of fitted logistic regression models as prediction tools.

Background

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

<http://www.biostat.ucsf.edu/biostat208/>

Logistic regression model assessment

Assume that you have arrived at a candidate model for the relationship between CHD (represented by the variable `chd69` in the dataset) and a set of predictors, including serum cholesterol, age, body mass index, systolic blood pressure, a binary indicator of type A behavior and a binary indicator distinguishing smokers/non-smokers. Now, we'll use `logistic` to fit the model.

Start out by defining the smoking indicator:

```
gen smoke = ncig
recode smoke 0 = 0 1/max = 1
```

Note that 12 individuals are missing measurements of the predictor `chol`. In addition, one individual has an outlying value of 645 for cholesterol. You can view this graphically by displaying a histogram for this variable:

```
histogram chol, bin(20)
```

Since we don't want to use the missing values and this outlier in fitting or assessment of the model (which will include `chol`), we can drop these individuals now so that we will be dealing with a consistent set of observations:

```
drop if chol==. (use two "=" signs here)
drop if chol==645
```

The "." character in the first command above symbolizes a missing value. (Note that internally, Stata represents missing values as the largest number possible for the number of fields occupied by a particular variable. From the output of the `describe` command, `chol` occupies nine decimal places. Thus, "`drop if chol>644`" will also drop the missing and/or outlying observations.) All other variables considered today are complete for all individuals.

The following command generates a new variable containing the body mass index for each WCGS participant:

```
gen bmi = ((weight)/(height^2))*703
```

Now, fit the logistic model:

```
logistic chd69 chol sbp dibpat age smoke bmi, coef
```

Question 1. Assess the fitted model based on the confidence intervals for the coefficients and the results of the likelihood ratio test.

Before reporting results of a model, it is wise to perform checks to insure that the results are not due to influential and/or outlying observations, and that the model provides a reasonable description of the data. Similar to conventional linear regression, there are measures of influence of observations on estimated coefficients that can be computed for a fitted model. The following versions of the predict command place the predicted probabilities, and an influence measure for each observation in the new variables `phat` and `dxb`, respectively:

```
predict phat
```

```
predict dxb, dbeta
```

(Specifically, the `dbeta` option requests that Stata compute for each observation a measure of how much the estimated coefficients would change if the observation were deleted.) A plot of `dxb` against predicted probabilities (`phat`) provides a graphical means of identifying influential observations:

```
scatter dxb phat, mlab(id)
```

The `mlab(id)` option requests that Stata label the individual points using the participant ID number. The graph should show one observation with relatively larger influence than the others. List the data for this observation:

```
list id chd69 chol sbp dibpat age smoke bmi if id==10078
```

The following Stata commands sort the observations in reverse order with respect to the magnitude of the `dxb` influence statistics and list the ten observations with the 10 largest values:

```
gsort -dxb  
list id chd69 chol sbp dibpat age smoke bmi if _n<=10
```

(The variable `_n` is a Stata system variable that records the current observation number; 1 for the first, and 3141 for the last in our case. Also, note that the notation "!=" means "not equal to".) To more directly visualize the impact of deleting the observation with the largest `dxb` value on the model results, refit for all observations except this one (which is now first in order):

```
logistic chd69 chol sbp dibpat age smoke bmi if id != 10078
```

Question 2. Examine the differences between the model excluding the most influential observation and the model including all observations. What issues should you consider in deciding whether to retain this observation in further analyses?

In the rest of the lab, we will keep this observation and assume that it is valid. Note that we could conduct further, complementary diagnostic checks using procedures described in the `logistic` entry of the Stata User's manual.

As discussed in lecture, significant test results associated with individual coefficients and the likelihood ratio comparison with the model including no predictors do not necessarily indicate that the model does a good job describing the data. One means of assessing the overall fit of a logistic model is provided by the *goodness of fit test*. The basic idea of such tests is to compare the outcome (CHD) predicted by the fitted logistic model with the actual outcome observed in specified groups of individuals. The *Pearson* test forms these groups based on predictor patterns and the *Hosmer-Lemeshow* test forms groups based on the ordered predicted probabilities from the model. As discussed in the Stata manual on postestimation commands for logistic regression, the Pearson test is preferred when possible combinations of predictor values are small (e.g. small numbers of categorical predictors). The Hosmer-Lemeshow test is better when large numbers of predictor patterns are possible (e.g. when continuous predictors are present). The hypothesis being tested for both goodness-of-fit tests is that the outcomes estimated by the model agree with their empirical counterparts. Thus, *a significant finding indicates lack of fit*. A non-significant finding rules out gross lack of fit. However, failure to find a significant result does not necessarily mean that the model fits the data well. (I.e. the test is not very powerful at picking up lack of fit from a number of possible

sources.) For these reasons, goodness of fit tests are most useful as a (very crude) way to screen for fit problems, and should not be taken as a definitive diagnostic of fit. The Hosmer-Lemeshow test is performed in Stata 11 using the “`estat gof`” command. The `group(10)` option specifies that the test be performed using 10 groups, while `table` requests a table listing observed and expected outcome frequencies for the ten groups:

```
logistic chd69 chol sbp dibpat age smoke bmi  
  
estat gof, table group(10)
```

(We re-fit the model first, including the observation dropped in the last fit.) The groups are defined based on splitting the ordered predicted probabilities of CHD from the model into 10 equal parts (deciles). In each group, the model is used to estimate a predicted frequency of outcomes. These are compared to the observed outcome frequencies using a conventional chi-squared test.

Question 3. Interpret the results of the goodness of fit test.

Note that although lack of fit diagnosed by a significant test result gives a good indication that a regression model will not be useful in making reliable risk predictions, it does not necessarily invalidate the use of the model for exploring issues of confounding and interaction between predictors. In many cases we may simply be unable to achieve a good model fit (as measured by a formal test), due to insufficient sample size and/or lack of key measured predictors. In this case, further analyses and interpretation of results should be made cautiously, paying attention to possible sources of bias.

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. Recall that to include this variable in a logistic model, we need to construct 3 indicator variables for 3 categories of `behpat`, leaving one level as the reference category, or use the “`i.`” syntax as follows:

```
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`

```
est store A1
```

Next fit the model without `behpat2`, and run the LR test comparing this smaller model to the larger model that includes `behpat2`:

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

Question 4. What do you conclude about the contribution of behavior pattern in our model for CHD risk?

Examination of the coefficients 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 smoke bmi 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 5. 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?

Prediction using logistic regression

As discussed in this week's lecture, frequently the goal of fitting a logistic model is to predict risk of the binary outcome given a set of predictors. This is of most interest for individuals different than those in the sample used to fit the model. As an example of a prediction problem, assume that a fraction of the WCGS data are available for developing a prediction tool, and that the remaining observations represent an additional sample which will be used to validate the predictions made with the model. Let's assume that the logistic model fitted above is to be our prediction tool.

To implement this in Stata, first randomly divide the sample into two groups: the first will contain approximately 90% of the observations, with the remaining 10% representing the validation sample. Use the random number generation facility to create a random number for each individual in a new variable `r`. The `seed` statement below is a way to insure that you generate the same string of random numbers. (In this case, it insures that everyone in the lab defined the same two groups of individuals.) The observations are then put in random order by sorting

according to `r`. Next, a new variable `group` is created to assign approximately 90% as the group to be used to fit the model:

```
set seed 386282911

gen r = uniform()

sort r

gen group = 1 if _n <= _N*0.9
```

Recall the definition of `_n` given above. The system variable `_N` is similar but gives the total number of observations in the working dataset. The observations not assigned to the fitting group (`group=1`) are now labeled as the validation set:

```
replace group=2 if group!=1
```

The following command tabulates the proportion of CHD events in the two groups for inspection:

```
tabulate group, summarize(chd69) nostandard
```

Fit the model with the individuals in `group=1`:

```
logistic chd69 chol sbp dibpat age smoke bmi if group==1
```

We can calculate outcome probabilities of CHD for individuals used to fit a model using the `predict` command. However, as discussed above, it is typically of more practical interest to predict outcome risk for a "new" individual with known predictors.

Question 6. Use the logistic model and the `lincom` command to compute the estimated probability of CHD for a hypothetical individual with the following values for the predictors:

```
cholesterol = 350
systolic blood pressure = 210
type A behavior pattern = 1
age = 55
smoking status indicator = 1
bmi = 26
```

Using coefficients from the estimated model and the above values for the predictors:

```
lincom _cons + chol*350 + sbp*210 + dibpat*1 + age*55 + smoke*1 + bmi*26
```

Now use the `display` command to convert the estimated odds for this individual into a probability, noting that the probability and odds are related by the following equation: $\text{probability} = \text{odds}/(1 + \text{odds})$. The following command uses the above model to estimate predicted CHD risk for the 10% of the original sample not used in fitting:

```
predict pht if group==2
```

This command places estimated probabilities of CHD in a new variable `pht`, based on the fitted model and predictor values for the individuals in this group. List the predicted probabilities along with values of the predictors for ten individuals:

```
gsort -group
```

```
list group pht chol sbp dibpat age smoke bmi if _n<=10
```

(The `gsort` command sorts the observations with `group==2` first so these will be listed first.)

Note that it is possible to compute standard errors and 95% confidence limits for predicted probabilities as well. See the following page for more information if you're interested:

<http://www.stata.com/support/faqs/stat/prep.html>

In some cases we may want to do more than simply estimate a predicted risk for prognostic purposes. For example, treatment decisions may be based on outcome status. Given an individual possessing known values of the predictors of interest, but with unknown outcome status, we would like to find a "cut-off" level of predicted outcome risk above which they are likely to be "positive" for the outcome and therefore eligible to receive treatment. In this context, assume that the 10% of the WCGS (selected above as `group=2`) represent a validation set of patients with known outcomes. We can use the model to predict outcomes for this group (as demonstrated above), and given a cut-off, compute sensitivity and specificity of outcome assignments. Stata has two procedures for logistic regression models that allow us to do this for all possible cut-offs. `lsens` computes sensitivity and specificity, while `lroc` estimates and ROC curve:

```
lsens if group==2
```

We can examine predictions and misclassification statistics in more detail for chosen cut-offs using `lstat`:

```
lstat if group==2, cutoff(0.2)
```