Biostatistics 208 -- 03/11/21

Lab #10

Lab Summary
In this lab you will learn to use some simple assessment techniques for logistic regression models, explore the use of fitted logistic regression models as prediction tools, and see how to use Stata to fit regression models to obtain coefficient estimates that have direct relative risk interpretation.

Background
The lab will again use the WCGS data (`lab10.dta` on the course site).

Logistic regression model assessment
Assume that you have arrived at a candidate set of predictors for the binary CHD outcome in the WCGS dataset, including serum cholesterol, age, body mass index (BMI), systolic blood pressure, a binary indicator of type A behavior and a binary indicator distinguishing smokers/non-smokers.

Recall from our previous analyses that 12 individuals are missing measurements of the predictor `chol`. In addition, one individual has an outlying value of 645 for cholesterol. You can see this in the summary for this variable:

```
summ chol, detail
```

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==. | chol==645
```

Note that internally, Stata represents missing values as the largest number possible for the number of fields occupied by a particular variable. Thus, "`drop if chol>644`" or "`drop if chol>=645`" will achieve the same result. All other variables considered today are complete for all individuals.

Recall from the last lecture that one individual had a BMI of 11.2. The age, height and weight values for this person are 47 (years), 70 (inches) and 78 (lbs). Since the weight value seems improbably low give the reported age and height, we will also omit this observation before doing any analysis. (If we had access to the original data collection forms, we could check for the possibility of error before deciding to omit the observation.)

```
drop if bmi<12
```

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 are 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 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 3140 for the last in our case.) 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, coef
```

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.

Goodness of fit

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 a *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 versions ≥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.

Recall from the last lecture that we found evidence for significant interactions between BMI and both cholesterol and SBP. The statements below re-fit the model including these interactions (using centered versions of these predictors to reduce collinearity), and then evaluate goodness of fit again:

```
summ bmi

gen bmic = bmi - r(mean)

summ chol

gen cholc = chol - r(mean)

summ sbp

gen sbpc = sbp - r(mean)

logistic chd69 dibpat age smoke c.bmic##c.cholc c.sbpc##c.bmic

estat gof, table group(10)
```

You should notice that the additions of the interactions results in an apparent improvement in fit. You can check for the presence of collinearity using the same procedure we used for linear models:

```
quietly regress chd69 dibpat age smoke c.bmic##c.cholc c.sbpc##c.bmic

estat vif
```

(Since `vif` is not available as an option after a `logistic` model fit, we use the `regress` command here. This is legitimate since the assessment of variance inflation factors is based only on the relationships between predictors.) Recall that large VIF values (>5) are of potential concern in obtaining reliable estimates. The fact that we centered predictors involved in interactions results in reasonably small values in this case.

Note that although lack of fit diagnosed by a significant test result gives a good indication that a regression model may 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.

Recall from lecture that goodness of fit reflects *calibration* performance of the model. It is also of interest to evaluate the *discrimination* ability of the model to predict individual outcomes. We return to this topic below.

Prediction with logistic regression models

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 and will be used to develop the prediction model; the remaining 10% represent 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 value of the seed is an arbitrary integer. It is set to 123456789 every time Stata starts up.) 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 development dataset to be used to fit the model:

```
set seed 45001

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 dibpat age smoke c.bmic##c.cholc c.sbpc##c.bmic if
group==1
```

Summarize discrimination performance of this model in the development data by calculating the *C*-statistic which measures area under the ROC curve, and storing the value in a scalar for later reference:

```
lroc if group==1

scalar c_dev=r(area)
```

Recall from lecture that the area under the ROC curve is the *C*-statistic, which is a measure of discrimination ability of the model. A value of 0.5 indicates no discrimination ability, with increasing values indicating better performance. We would expect the discrimination performance of the model in the development group to be over-optimistic since the model was fitted using these data. An optimism-corrected measure can be obtained using the validation group:

```
lroc if group==2

scalar c_val=r(area)
```

The optimism correction is calculated as the difference between the two discrimination measures, and represents the reduction in discrimination performance observed in the validation group:

```
dis c_val - c_dev
```

The model fitted above can be used to estimate predicted 10-year risk of CHD in individuals similar to those used in the WCGS cohort. Based on the mediocre optimism-corrected performance, we would not expect this to be very useful in practice.

In some cases we may want to do more than simply estimate a predicted risk for prognostic purposes. For example, treatment decisions or eligibility criteria could be based on predicted outcome status. Given an individual possessing known values of the predictors of interest, 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 candidates for treatment/eligibility. We can use the development dataset to determine a cut-off value that has the desired levels of sensitivity and specificity. The `lsens` command displays sensitivity and specificity for a range of probability cut-off values. The command below provides a graphical display and stores values of the cut-off based on predicted probabilities in the development data, along with the corresponding sensitivity and specificity values.

```
lsens if group==1, genprob(cut) gensens(sens1) genspec(spec1)
```

```
list cut sens1 spec1 if abs(sens1-spec1)<0.001 & group==1
```

The second command lists cut-offs from the development dataset with sensitivity and specificity values that differ by no more than 0.1%. If our goal was to develop a classification rule that weighted sensitivity and specificity equally, we could use this list choose a cut-off of approximately 0.085. (Note that a more formal procedure for choosing cut-off for alternate weighting of sensitivity and specificity can be based on the Youden *J*-statistic (see reference below).) We can see from the list that a classification rule based on this cut-off has approximately 70% sensitivity and specificity in the development data. Again, these are *optimistic* estimates of performance. We can also use the `lstat` command to summarize performance at this cut-off in the development group:

```
lstat if group==1, cutoff(0.085)
```

We can get a sense of optimism-corrected values for sensitivity and specificity of the rule based on the 0.085 cut-off by applying the same command to the validation group:

```
lstat if group==2, cutoff(0.085)
```

Question 4. Comment on the differences between the sensitivity estimates obtained in the last two commands and their implications for expected classification performance.

Fitting relative risk regression models using Stata

Frequently, relative risks are more desirable association measures than odds ratios. Since the WCGS was a prospective study, relative risks are certainly appropriate, and may be preferred to odds ratios.

As discussed in class, it is possible to fit regression models with coefficients that have direct interpretations as relative risks. These models specify that the outcome probability is linear on the log-probability scale (in contrast to the linearity on the log-odds scale assumption used for the logistic model). As discussed in lecture (next week), the log probability model (shown below for a single predictor x) can be fit directly using Stata's `binreg` (i.e. binary regression).

$$\log P(x) = \beta_0 + \beta_1 x$$

This model frequently is difficult to fit because of the constraint that the estimated outcome probabilities must be between 0 and 1.

```
binreg chd69 dibpat age smoke c.bmic##c.cholc c.sbpc##c.bmic, rr
iterate(100)
```

The `iterate(100)` option restricts the number of times Stata iterates the maximum likelihood estimation algorithm to find estimates to 100. You'll note that the model estimation fails to converge after 100 iterations (Note that estimates are typically found in less than 20 iterations for logistic regression, even for models with many predictors.) This is an example where it is not feasible to fit the log relative risk model using the maximum likelihood approach (based on the correct binomial distribution for binary outcomes) used by `binreg`.

A good alternative that avoids this problem is to use a regression approach designed for count outcomes that follow the *Poisson* distribution. (These methods will be discussed more in Biostatistics 209, and are

covered in Chapter 8 of the textbook.) The syntax for fitting Poisson regression models in Stata is familiar, and follows that used for `regress` and `logistic`:

```
poisson chd69 dibpat age smoke c.bmic##c.cholc c.sbpc##c.bmic, irr
robust
```

The `irr` (stands for "incidence rate ratio") option requests that the estimated coefficients be exponentiated, which yields a relative risk interpretation in this context. The `robust` option requests that robust variances be used rather than the maximum likelihood-based standard errors from the Poisson model. This turns out to be more appropriate in this context, since the Poisson distribution for outcomes model is only approximately correct for the binomial distribution.

Question 5. Compare the estimated relative risks with the odds ratios estimated from the logistic model. Are there any notable differences?

Remember that relative risk regression models *don't make sense for data from case-control studies*, where participants are selected based on their outcomes. Because these studies don't have any information about outcome prevalence, estimates like relative risks don't retain their usual interpretation. However, for very rare outcomes, the odds ratios from case-control data estimatyed via logistic regression may approximate the corresponding relative risks. The overall prevalence of the CHD outcome for the WCGS data is around 8%, so the relatively close correspondence between the estimates from the two approaches in this case isn't too surprising.

Reference for Youden index:
https://en.wikipedia.org/wiki/Youden%27s_J_statistic