Biostat 208, Lab #11

Lab Summary

In this lab you will learn how to analyze data from unmatched and matched case-control studies using Stata.

Background

The lab will be based on data from two studies. The first is the Ille-et-Vilaine study of esophageal cancer, an unmatched design including 200 cases and 775 comparable controls. This study was designed to investigate alcohol and tobacco consumption as risk factors for esophageal cancer in men. In addition to these risk factors, the dataset (`esoph.dta`) contains the variable `age`, an established confounder from previous research.

The second study is the "Leisure World" 1:4 matched case-control study of endometrial cancer, including 63 cases, each matched to 4 controls of similar age and marital status. The dataset is called `endom.dta`. The risk factor of primary interest is a binary indicator of past estrogen use. Additional variables of interest include binary indicators of obesity, hypertension, history of gall bladder disease and use on non-estrogen hormone replacements.

Methods of analysis for unmatched case-control study data

First we'll consider the esophageal cancer study (dataset `esoph.dta`). Begin by describing and summarizing the variables:

```
des
summarize age alc tob
tab case
tab agegp
tab alcgp
tab tobgp
```

Recall from lecture that Stata has commands for analyzing case-control study data in tabular form. Appropriate for small numbers (1-2) of categorical risk factors, these procedures mirror those used for data from prospective and cross-sectional studies, but focus on the odds ratio as a measure of association because it is valid in the case-control setting. First, consider the marginal association between alcohol consumption and esophageal cancer risk. Stata's `epitab` module provides two procedures to do this. The first is the `cc` command. This is appropriate for dichotomous risk factors, but allows for stratification by additional factor(s) similar to the `cs` command used in prior labs for cross-sectional and prospective studies. Alcohol is provided either as a 4-level categorical factor `alcgrp` or the continuous measurement `alc` (alcohol consumption in grams/day),so `cc` can't be used unless we dichotomize alcohol. As we've seen previously, the `tabodds` command is useful for examining odds ratios for risk factors with more than two levels. The following command estimates odds ratios for the association between the outcome variable (`case`) and categorized alcohol consumption (`alcgp`):

```
tabodds case alcgp, or
```

Now use the same procedure to examine the marginal associations with tobacco use (using the 4-level categorical factor `tobgp`) and `age` (the 6-level categorical factor `agegp`):

```
tabodds case tobgp, or
tabodds case agegp, or
```

Question 1. What do the results from `tabodds` indicate about the associations between these three variables and esophageal cancer?

Because age is a known confounding factor from past studies of this outcome, let's examine the effect of adjusting for age in assessing the association between alcohol, tobacco and esophageal cancer. This is accomplished using the `adjust()` option of `tabodds`:

```
tabodds case tobgp, or adjust(agegp)
tabodds case alcgp, or adjust(agegp)
```

Note that the odds ratios presented are adjusted for the presence of age using the Mantel-Haenszel procedure. (We can assume that age doesn't interact with these exposures to influence cancer risk based on previous analyses of these data, including the results presented in lecture.)

Question 2. Do the above results indicate that age has a confounding influence on the measured associations between tobacco and/or alcohol consumption and the outcome?

Since we're only dealing with at most two categorical factors simultaneously in the above analyses, use of logistic regression isn't called for. However, to investigate the independent effects of alcohol and tobacco adjusted for age, or to examine the influence of alcohol and tobacco as continuous variables, the logistic regression model is preferred. As discussed in lecture, the logistic model for case-control data is fit and interpreted in the same way as it was when applied to data from cross-sectional and prospective studies. However, because the proportion of subjects with the outcome is fixed by design in case-control studies, the intercept term in logistic models cannot be interpreted as providing information about the "background" rate of disease in the absence of risk factors. In fact, the intercept term (and consequently the predicted outcome probabilities) is not directly interpretable because it depends on the sampling selection probabilities for cases and controls (which are usually unknown). As a result, the logistic model is not useful for predicting individual disease outcomes for case-control data. However, as demonstrated in lecture, the regression coefficients for risk factors retain their interpretation as log-odds ratios familiar from applications to cross-sectional and prospective studies.

The following commands fit models to examine the age-adjusted associations for alcohol and tobacco, repeating the analyses performed using the `tabodds` command above:

```
logistic case i.alcgp i.agegp
logistic case i.tobgp i.agegp
```

Notice that the age-adjusted odds ratios for tobacco and alcohol are similar but not identical to their counterparts in the table-based analysis. The difference relates to the way in which the logistic model represents the influence of risk factors on the outcome (contrasted to the approach implicit in the corresponding frequency tables methods). Now fit a model to investigate the effects of age and tobacco simultaneously:

```
logistic case i.alcgp i.tobgp i.agegp
```

To assess the separate contribution of alcohol and tobacco to explaining esophageal cancer risk, perform likelihood ratio tests for each factor separately. (If you wish, you can instruct Stata to suppress further regression output by proceeding commands with the word `quietly`):

```
estimates store M0
quietly logistic case i.tobgp i.agegp
lrtest M0
quietly logistic case i.alcgp i.agegp
lrtest M0
```

Question 3. Interpret the results of the likelihood ratio tests. Which risk factor results in the biggest change in log-likelihood?

Question 4. Create a dichotomous indicator of heavy alcohol use with the statements below, and evaluate it as a predictor in the model including `tobgp` and `agegp`. Use the likelihood ratio test to compare this model to the model including the four-level categorical predictor `alcgp` (fitted above). Do we lose anything from modeling alcohol consumption as a binary predictor?

```
. gen alcb = alcgp
. recode alcb 1/2=0 3/4=1
```

Methods of analysis for matched case-control study data

Now let's briefly consider the use of Stata for analysis of data from matched case-control studies. We'll look at data from the matched study of endometrial cancer described above, comprised of 63 cases, each matched according to age and marital status to four controls.

Now load the next dataset `endom.dta` into Stata. Begin by describing the data and summarizing the outcome and selected risk factors:

```
des
summarize age dose
tab case
tab gall
tab est
tab ob
```

The variable `set` takes on the values 1 - 63, and labels each matched set consisting of a case and four controls. Since the women in each set are similar in terms of the factors used for matching, the analyses will need to control for this variable. The following command shows you the structure of the data needed for the analyses we'll perform in Stata:

```
sort set case
list case set age if set<5
```

This structure (listed for the first four "sets") differs from that used for conventional logistic regression only in the presence of the matching variable `set`. You can see that ages of women in each set differ by at most one year, confirming the matching.

First, we'll use frequency table methods to look at the effect of selected risk factors. Stata has two procedures for matched case-control data, `mcc` and `mhodds`. The former is only applicable to studies with a single control matched to each case (i.e. 1:1 matched designs) and requires a different data structure than the present example. The `mhodds` procedure can be applied to studies with multiple controls/case, but is restricted to looking at most one risk factor along with a single adjustment variable. Here's an example of the use of `mhodds` to examine the association between endometrial cancer and the binary indicator `est` of past estrogen use:

```
mhodds case est set
```

As mentioned above, the matching in the study design must be incorporated into analyses to draw valid inferences. The `mhodds` procedure does this using the matching variable `set` in a procedure similar to the stratification used for analyses of data from unmatched studies (and for prospective and cross-sectional studies). When continuous risk factors, factors with multiple levels, and/or several risk factors are involved, a regression analysis approach is needed.

Unlike standard logistic regression for unmatched data, the logistic model for matched studies is *conditional*, and adjusts for the matching without explicitly including indicator variables for the matched sets (of which there are 63 in the current example). The `clogit` procedure is used similarly to `logistic`, but requires an option (`group`) specifying the variable identifying the matched sets. Begin by fitting a model for the marginal association between endometrial cancer occurrence and the binary indicator `est` of past estrogen use:

```
clogit case est, group(set) or
```

(The `or` option instructs `clogit` to return exponentiated regression coefficients similar to the `logistic` procedure.) Now let's extend the analysis to investigate the influence of the additional risk factors `gall`, `hyp`, `non` and `ob` (described in the output to the `describe` command issued above). Recall that `age` was used in matching and therefore should not be included. (It is already accounted for in the estimation through the `group()` option to `clogit`.) Before beginning, notice that the obesity variable `ob` has three levels, and that presence/absence of this condition is unknown for 51 women:

```
tab ob case
```

If we look at the association with the outcome for each level separately (taking non-obese women as the reference group), we see that the women with unknown status are more like the women who are not obese than those that are:

```
clogit case i.ob, group(set) or
```

Based on this observation, and to simplify further analyses, let's combine the women with unknown status with the non-obese women and use this group as the new reference group when examining obesity as a risk factor:

```
gen obes = ob
recode obes 9=0
```

Note that the new variable `obes` is assigned the numerical values of the original indicator `ob`. The "unknown" category has the value "9" which is replaced with "0" by the `recode` command. Now fit an overall conditional logistic model including five binary risk factors:

```
clogit case est hyp non gall obes, group(set) or
```

Question 5. Compare the results of this model to the model fitted above including only the indicator `est` of estrogen use and comment on the apparent effects of including the additional risk factors on the association between the outcome and estrogen use.

A complete analysis would continue with explorations of interactions between risk factors and of quantitative measures of estrogen use (`dose` and `dur`). For example, look at the interaction between `gall` (history of gall bladder disease) and `est`. You may wish try fitting additional models to investigate other potential interactions as well.