Biostatistics 208 Lab #7 02/20/2020

The focus of this lab is selecting predictors when the inferential goal is to assess a predictor of primary interest: specifically, whether exercise affects bone mineral density (BMD), a measure of bone strength and resistance to fracture. Several prior studies have suggested such an effect, at least in some populations. However, the unadjusted association between exercise and BMD could be confounded by age, functional decline, medication use, and lifestyle variables.

The dataset is a 50% random subsample from the Study of Osteoporotic Fractures (SOF), and includes women who attended visit 5 and had a BMD measurement. The predictor of central interest, exercise energy use (EEU) in units of 100 Kcals/day, is based on self-report about recreational physical activities in the last week. This version focuses on weight-bearing activities, which may particularly affect BMD. The outcome is hip BMD, which is strongly associated with fracture.
As usual, download the dataset `lab7.dta`. The variables and values are already labeled. The potential confounders and/or mediators of EEU in the dataset include

- age, log weight (`lweight`), and lifestyle variables including alcohol (`drnkspwk`) and caffeine (`caffeine`) consumption, and current smoking (`smoker`)
- physical functioning variables, including performance on the tandem stand (`tandstnd`), use of arms to stand up (`usearms`), gait speed (`gaitspd`), and hip abductor (`has10`) and grip (`gs10`) strength
- poor or fair health by self-report (`poorhlth`), a powerful but difficult-to-interpret predictor of many chronic disease outcomes
- use of medications known to affect BMD, including estrogen (`estrogen`), the bisphosphonate Etidronate (`etid`), calcium supplements (`calsupp`), and diuretics (`diuretic`)
- history of non-spine fracture (`nsfx2`), maternal history of hip fracture (`momhip`)

In addition, this lab will use a new ado-file that does model selection using the change-in-coefficient criterion. Download `cic.ado` and `cic.hlp` from the course website, and put them in the same directory where you put the dataset. Note that you may need to hold the "control" key when clicking the link for the help file to be able to download using macOS (right-click in Windows) it rather than view it in the browser window. You will need to change to that directory for Stata to find these files. You can do this via the "change working directory" menu item, or on the command line using the `cd` command. In macOS, the command I use is (the "~" symbol denotes the home directory folder):

```
cd "~/Documents/teaching/c2019/biostat208/labs/lab7"
```

A priori model

Suppose that on the basis of your knowledge about potential confounders of the association between exercise energy use and BMD, you specify two DAGs, one including an A-list of very well supported confounders, and a second larger DAG, also including a B-list of potential confounders you are less certain about. Suppose also that you have analyzed these two DAGs (possibly using `http://dagitty.net`), and that you come up with two practical MSASs:

1. Based on A-list variables only: age, log weight, use of bone-active medications (estrogen, calcium supplements, diuretics, Etidronate), and maternal history of hip fracture.
2. Based on the A-list in the previous MSAS and the following B-list variables: poor health by self-report, as well as all three lifestyle variables and all five measures of physical functioning.

Question 1. What are some advantages and disadvantages of selecting a model a priori, based on a well-motivated DAG?

Removing less important B-list variables

A possible analytic approach to the situation where the most practical MSASs for two candidate DAGs are inconsistent, is to eliminate B-list variables in the larger MSAS if this does not change the coefficient for the primary predictor by more than 5 or 10%. A primary reason for doing this is to retain precision of the estimated effect of the primary predictor by reducing the number of adjustment variables. The `cic` command automates this procedure, which can involve running a lot of models. In brief, `cic` cycles through the B-list of covariates eligible for removal. On each cycle, a model is fit omitting each of the remaining B-list covariates one at a time, and the absolute percent change in the coefficient for the primary predictor is calculated. Then the covariate that induced the smallest change is removed from the B-list, provided the change is less than specified maximum change criterion. Cycling continues until no remaining B-list variables induce changes smaller than the allowable maximum. To familiarize yourself with some of the options for controlling `cic`, look at the help file (type `help cic` at the command line or in the drop-down Help menu).

Now run the following command. To run this from the command line, you have to remove line separators (///) and enter the entire command without carriage returns; otherwise run it as is from a do-file.

```
cic, model(regress) outcome(bmd) primary(eeu) ///
   alist(age lweight i.estrogen calsupp diuretic etid momhip) ///
   blist(poorhlth usearms i.tandstnd has10 gs10 gaitspd caffeine drnkspwk smoker)
```

Question 2. Interpret the estimated coefficient for EEU from the final model, in the relevant units (g/cm^2 for BMD and 100 Kcal/day for EEU). Suppose that walking a mile uses 150 Kcals. On average, how much increase in BMD would the model lead us to expect if a woman increased her average daily EEU by this amount? Given that average BMD is about 0.73 g/cm^2, would that effect on BMD seem clinically significant? How does the effect of 150 Kcal/day compare in magnitude to the effect of a 10% increase in weight or current use of estrogen?

Here is some code you can use to calculate these effects:

```
lincom eeu*1.5
lincom ((eeu*1.5)/0.73)*100
nlcom _b[lweight]*log(1.1)
lincom 2.estrogen
```

Now try changing some of the `cic` options:

- `criterion(real number)` specifies the maximum allowable change in coefficient, in percentage points. Change it from the default value of 5% to 10%.
- `complete` forces all models to be run using the subset of observations with complete data on all predictors in the initial B list, as well as the outcome, primary predictor, and A-list (which are included in all models). The default is to use observations with complete data on the current B-list. Note that Stata's stepwise selection commands `sw` (`stepwise`) use the `complete` option by default, and do not allow the alternative.

```
cic,  model(regress) outcome(bmd) primary(eeu) ///
   alist(age lweight i.estrogen calsupp diuretic etid momhip) ///
   blist(usearms i.tandstnd has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker)
///
   criterion(10)
```

```
cic,  model(regress) outcome(bmd) primary(eeu) ///
   alist(age lweight i.estrogen calsupp diuretic etid momhip) ///
   blist(usearms i.tandstnd has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker)
///
   complete
```

Question 3. Are the results for `eeu` sensitive to changes in these two options?

Question 4. How much support do these data provide for the hypothesis that exercise increases BMD?

Below are two optional lab exercises. The first covers an alternative to the change in coefficient approach examined above, using a stepwise-based procedure. The second looks at constructing a prediction model using LASSO regression as a modern alternative to the conventional regression-based approach covered in lecture. Do both, or pick one, depending on your interests.

Optional: backwards selection using *p*-values

If we felt unable to draw a persuasive DAG, we might select an A-list of confounders based on face-validity and knowledge of the literature, and then select among the B-list confounders, either using `cic` or using traditional backwards selection based on *p*-values. Stata code to implement the latter procedure follows.

```
xi: sw, lockterm1 pr(.2): ///
regress bmd (eeu age lweight i.estrogen calsupp diuretic etid momhip) ///
usearms (i.tandstnd) has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker
```

Stata Notes:

- The command `prefix sw` (equivalently `stepwise`) is used to run automated model selection procedures; note that `sw` and its options, explained below, come after `xi:`, just before `regress`. If you want to have Stata make indicator variables automatically, you have to use the `xi:` command prefix, even with Version 11-15; `sw` is not compatible with the new machinery for categorical variables.

- Option `lockterm1` forces the inclusion of the first long group of predictors enclosed in parentheses.

- The categorical variable `i.tandstnd` is enclosed in parentheses so that all three indicators are treated as a group in the selection procedure. We don't have to worry about `i.estrogen` because all levels are forced into the model by the `lockterm1` option.

- Option `pr(.2)` specifies backward selection with the inclusion criterion of $p < 0.2$. At each step, the remaining variable not in the `lockterm1` list with the biggest *p*-value > 0.2 is removed. This continues until all variables eligible for removal have *p*-values smaller than this criterion. The model coefficients and *p*-values are re-estimated after each removal.

In the basic backwards selection procedure, the order of the variables in the `lockterm` list as well as the list eligible for removal does not affect the default backwards procedure. However, with the addition of the `hier` option (as used to implement the Allen-Cady procedure described in section 10.3.4 of VGSM), Stata starts with the last variable in the eligible-for-removal list and works back, stopping when it encounters the first variable with a *p*-value smaller than the retention criterion.

Question 5. Consider the effect of modifying the *p*-value required for inclusion. Try out using criteria of $p < 0.10$ and $p < 0.05$. How and why do the results change?

```
foreach pr in 0.1 0.05 {
   dis _newline as result "Backwards deletion with a retention criterion of p<`pr'"
   xi: sw, lockterm1 pr(`pr'): regress bmd (eeu age lweight i.estrogen calsupp
diuretic etid momhip) usearms (i.tandstnd) has10 gs10 gaitspd poorhlth caffeine
drnkspwk smoker }
```

Question 6. We could modify the list of physical function variables considered for inclusion in the model to reflect concerns about the numbers of missing values for tandem stand and hip abductor strength. What happens if we omit them from the list?

```
xi: sw, lockterm1 pr(0.2): regress bmd (eeu age lweight i.estrogen ///
calsupp diuretic etid momhip) usearms gs10 gaitspd poorhlth caffeine drnkspwk smoker
```

Optional: model selection using LASSO regression (Stata 16)

The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression technique that incorporates *shrinkage* of the effect of predictors on the outcome via a penalization technique incorporated into the estimation of coefficients. Coefficients for poorly performing predictors (i.e. those that don't improve prediction performance) are shrunk towards zero based on a user-supplied penalty parameter λ, with higher values implying higher degrees of shrinkage. The effect of this is to produce models that effectively select important predictors and down-weight/eliminate the effects of those that perform poorly. (Predictors with coefficients shrunk to zero are effectively removed from the model.) This approach improves over stepwise selection procedures since all candidate predictors can be included. Further, since model selection for the LASSO involves cross-validation, overfitting is less likely to be an issue. However, note that concerns about nonlinearity and interactions still need to be addressed for LASSO-based models. (We'll ignore these in the example below.)

As of version 16, Stata includes LASSO regression that can be applied to both continuous and binary outcomes. Stata's implementation includes cross-validation in model selection as an option. For each choice of shrinkage penalty (λ), 10-fold cross-validation is used to select the model with the minimum prediction error. By repeating this over a range (grid) of λ values, an overall best-performing model is selected. If a validation set is available (or feasible given overall sample size), the validation performance of this model can also be assessed.

The example below illustrates an application of the LASSO to selecting a regression model predicting BMD based on 17 selected predictors from the SOF data. Following the illustration from the first part of lecture 7, we'll use the following approach:

1. Split the data into randomly selected training and validation sets, with 75(25)% of observations in the training (validation) groups
2. Apply LASSO to develop a best-performing model in the training data
3. Evaluate the prediction performance of the selected model on the validation data

Below is some commented code to implement steps 1 & 2 of this analysis:

```
* split the data into training & validation groups, and create a group indicator
differentiating them
splitsample, generate(group) split(0.75 0.25) rseed(2202020)
* display group and overall mean outcome values
tabulate group, summarize(bmd)
* run the LASSO on outcome and predictors in the training group
lasso linear bmd eeu age lweight i.estrogen calsupp diuretic etid momhip usearms i.tandstnd
has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker if group == 1
* store results for further analysis
estimates store cv
```

In examining the output of the `lasso` command, you'll see that Stata runs through 48 values of λ before reaching one that achieves a minimum prediction error criterion (after which error doesn't decrease appreciably). For each value (starting with the largest), a 10-fold cross-validation is run to select a model that minimizes prediction error. You'll notice from the default output that as λ decreases, more predictors exhibit non-zero coefficients. The `cvplot` command illustrates the results:

```
cvplot
```

The resulting coefficients are displayed below. You'll see that all of the predictors are retained in the final model, but many of the coefficients are shrunk close to zero.

```
lassocoef cv, display(coef) sort(coef, standardized)
```

The following command summarizes prediction performance in the training and validation groups:

```
lassogof cv, over(group)
```

Note that we could have used alternative performance metrics to cross-validated prediction error in the LASSO (e.g. BIC), and compared models across these as well. Once fitted, a LASSO model can be used to aid in variable selection for a more conventional regression, prediction of outcomes, or for inference about predictor effects. Stata's introduction document is a good summary of features:

https://www.stata.com/manuals/lassolassointro.pdf