

Biostat 208 Lab #7, 2/18/10

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. Variables include predictors of BMD identified in a 1996 SOF paper (Orwoll et al, Axial bone mass in older women, *Ann Intern Med*, 1996; 124: 187-96). The predictor of central interest, exercise energy use (EEU) in units of 100 Kcals/day, is a “Paffenbarger score,” based on self-report about recreational physical activities in the last week. These are assigned caloric expenditure rates from standard tables. 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. In lab 6 and lecture we already verified that BMD residuals have a reasonably normal distribution after adjustment for age and log weight, two of the most powerful BMD predictors. We also saw that the associations of BMD with age and log weight are reasonably linear. 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`)

1 *A priori* model

The primary concern with this inferential goal is to rule out confounding of the effects of the primary predictor of interest. This is a big dataset, with many more than 10 observations for each of the available predictors. Here is a model that might be specified *a priori*, at least in broad outline, without looking at the current data, on the basis of existing knowledge about the determinants of BMD:

- Covariates recognized as causal determinants of BMD (or markers for such causal factors): age, log weight, use of bone-active medications (estrogen, calcium supplements, diuretics, Etidronate), and maternal history of hip fracture.
- Potential confounders less well supported by evidence: poor health by self-report, the three lifestyle variables, and the five measures of physical functioning.

But there are some issues to work out in deciding which variables to include.

- Physical functioning and weight are potential confounders because they may affect the ability to exercise, and could be markers for other age-associated processes that drive changes in BMD. But in addition, exercise may affect both weight and physical functioning, so to some extent these variables may mediate as well as confound the effects of exercise.
- History of non-spine fracture is potentially problematic because fracture risk increases with low BMD, and thus the variable could be seen a common effect of the outcome BMD and many of the factors that potentially confound EEU – in short, a collider. What is the problem with this argument?
- Hip abductor strength and performance on the tandem stand are the most powerful predictors of BMD among the physical function variables, but have the largest numbers of missing values (you can see this using the `sum` command). High levels of missingness can reduce efficiency in a so-called complete case analysis, in which we only analyze observations with complete data on all variables used in the model, the default in STATA. This can also induce bias and call into question the validity of the model.
- In developing an *a priori* model, it would be reasonable to look at the correlation between EEU and potential confounders/mediators, provided we don't assess associations with the outcome. You can use the following code to characterize these associations. As usual, be careful not to type a carriage return after `etid` in the long `foreach` command, which ends with the opening `{`. Since this is a large dataset, I would focus on effect sizes rather than *p*-values, since most of these associations are likely to be statistically significant.

```
pwcorr eeu age weight bmi drnkspwk caffeine gaitspd has10 gs10, sig obs
foreach x in poorhlth tandstnd usearms momhip nsfx2 estrogen calsupp etid
    diuretic smoker {
    tab 'x', sum(eeu)
    }
```

Recognizing that this may not be your area of expertise, nonetheless map out a DAG for plausible causal pathways linking EEU, BMD, and potential confounders and mediators, and run the model you select. In the interests of time, you can skip the model checks we learned about last week.

1. *How did your DAG map on to your model?*
2. *Which of the physical function variables, if any, did you include? How would you justify your decision?*
3. *Did you also control for weight and history of non-spine fracture? Why or why not?*
3. *What other concerns drove your model selection, and what are the results for eeu?*
4. *What are the advantages and disadvantages of selecting a model a priori?*

2 Backwards selection model

Alternatively, we could select among the list of potential confounders using backwards selection. Stata code to implement the procedure is as follows. Remember not to type a carriage return at the end of the first line. If you cut and paste from the handout, you will only be able to copy one line at a time. Details of the command are explained below.

```

* Version 10 or 11, using xi: and i.
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

* Version 11 workaround, making indicators beforehand
tab estrogen, gen(est)
tab tandstand, gen(ts)
sw, lockterm1 pr(.2): regress bmd (eeu age lweight est1 est2 calsupp diuretic etid momhip)
    usearms (ts2 ts3 ts4) has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker

* Allen-Cady procedure
sw, lockterm1 pr(.2) hier: regress bmd (eeu age lweight est1 est2 calsupp diuretic etid momhip)
    usearms (ts2 ts3 ts4) has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker

```

STATA Notes:

- The command prefix `sw` 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. A workaround is also given.
- Option `lockterm1` forces the inclusion of the first long group of predictors enclosed in parentheses.
- The categorical variable `i.tandstnd` (or the indicators `ts2`, `ts3` and `ts4`) 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` (or `est1` and `est2`) 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 t -test 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.
- 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 used to implement the Allen-Cady procedure, 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.

1. 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?

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

3. Interpret the estimated coefficient for *EEU* in the relevant units (g/cm^2 for BMD and 100 Kcal/day for *EEU*). Suppose that walking a mile uses 150 Kcal. 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 $0.73 \text{ g}/\text{cm}^2$, 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?

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