Comments on Biostatistics 208 Lab #7 – 2/18/2021

A priori model

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

The clearcut advantage is that *p*-values and confidence intervals retain their theoretical meaning; in contrast, after model selection, we know that the *p*-values are usually too small and the CIs too narrow, but not by how much. Of course, you have to be able to motivate the model persuasively, but the determinants of BMD are reasonably well understood. A disadvantage is that we can end up including unimportant variables, potentially inflating the standard error for `eeu` unnecessarily – or exclude something important, opening the door to residual confounding.

Removing less important B-list variables

Question 2. Interpret the estimated coefficient for EEU from the final model, in the relevant units (g/cm2 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/cm2, would that effect on BMD be 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?

Focusing on the estimate of 0.0017 g/cm2 per 100 Kcals from model selected by `cic`, an additional mile of walking per day would ideally be expected to increase BMD by 0.0026 g/cm^2, or about 0.35% of the average level of 0.73 g/cm^2. In contrast, a 10% increase in weight is associated with an increase of 0.074 g/cm^2, and current use of estrogen is associated with an additional 0.079 g/cm^2, both representing increments of more than 10% of the average value.

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

Essentially not: the coefficients for `eeu` differ very little between models, and the *p*-values are all trivially greater than 0.05.

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

My take is that these data do provide weak support for the hypothesis, but suggest that the effect isn't particularly large. Only big increases in exercise energy use would be expected to have substantial effects, and it is not easy to get people to exercise that much, especially older people at high risk for osteoporosis and fractures.

Optional: backwards selection using *p*-values

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?

Note that all three models use just the 2542 observations with complete data on the entire initial list of variables. The reduction in the standard error for `eeu` in the smaller models is relatively minor; the bigger change is the increase in the estimate for `eeu` from about 0.0017 g/cm^2 per 100

Kcal, as in the models selected by `cic`, to about 0.0020, an increase of more than 15%, suggesting that there is some residual confounding of the effect of `eeu` by gait speed and hip abductor strength, the additional variables that get removed.

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?

This retains a larger number of observations, and results in a nominally statistically significant estimate for `eeu`. However, opinion about the effect of EEU on BMD should not be strongly influenced by these minor differences in *p*-values.

Optional: model selection using LASSO regression (Stata 16)

Below is abbreviated output from the LASSO analysis. Note that use of the random seed (**`rseed(2202020)`**) in the **`splitsample`** command should make your results the same as those presented here. As discussed in lecture, seeding a random number generator ensures that the same sequence of random numbers will be used for commands and programs that require them (e.g. sample splitting, bootstrapping, cross-validation & simulations). Any number can be used as a seed.

Note that the displayed shrunken coefficients from the LASSO model (`lassocoef` command) are displayed using the `standardize` option because the LASSO is fit using standardized versions of predictors. (This is done to put all the predictors on a similar scale for estimation purposes.) Also, note that `smoker` is not present because this variable was not selected. (i.e. its coefficient was shrunken to effectively zero.)

```
. * 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)

            |           Summary of BMD
      group |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   .73009549   .13540008       2,084
          2 |   .73184317   .13614306         695
------------+------------------------------------
      Total |   .73053257   .13556385       2,779

.
. * 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

10-fold cross-validation with 100 lambdas ...
Grid value 1:     lambda = .0693681   no. of nonzero coef. =       0
Folds: 1...5....10   CVF = .0179829
Grid value 2:     lambda = .0632056   no. of nonzero coef. =       1
Folds: 1...5....10   CVF = .0172221
Grid value 3:     lambda = .0575906   no. of nonzero coef. =       1
Folds: 1...5....10   CVF = .0165425
Grid value 4:     lambda = .0524744   no. of nonzero coef. =       1
Folds: 1...5....10   CVF = .0159783
Grid value 5:     lambda = .0478127   no. of nonzero coef. =       1
Folds: 1...5....10   CVF = .0155099
```

```
Folds: 1...5....10   CVF = .0120153
```
- Output excluded
```
Grid value 48:    lambda = .0008753   no. of nonzero coef. =      19
Folds: 1...5....10   CVF = .0120152
Grid value 49:    lambda = .0007976   no. of nonzero coef. =      19
Folds: 1...5....10   CVF = .0120154
Grid value 50:    lambda = .0007267   no. of nonzero coef. =      19
Folds: 1...5....10   CVF = .0120156
... change in deviance stopping tolerance reached ... last lambda selected

Lasso linear model                          No. of obs        =      1,900
                                            No. of covariates =         22
Selection: Cross-validation                 No. of CV folds   =         10

------------------------------------------------------------------------------
         |                                No. of      Out-of-      CV mean
         |                               nonzero       sample   prediction
      ID |     Description      lambda     coef.    R-squared        error
---------+--------------------------------------------------------------------
       1 |    first lambda    .0693681         0       0.0012     .0179829
      47 |   lambda before    .0009607        19       0.3327     .0120153
    * 48 | selected lambda    .0008753        19       0.3327     .0120152
      49 |    lambda after    .0007976        19       0.3327     .0120154
      68 |     last lambda    .0001362        20       0.3320      .012028
------------------------------------------------------------------------------
* lambda selected by cross-validation.

. * store results for futher analysis
. estimates store cv

. cvplot

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

------------------------
             |        cv
-------------+----------
     lweight |  .0667905
             |
    estrogen |
    current  |  .0229486
             |
     usearms | -.0113085
         age | -.0087306
    caffeine | -.0083018
             |
    tandstnd |
       good  |   .007472
             |
        etid | -.0063319
     calsupp | -.0045198
      momhip | -.0042051
         eeu |  .0040263
             |
    tandstnd |
   not done  | -.0036527
             |
       has10 |  .0025983
             |
    estrogen |
      never  | -.0019414
             |
    diuretic |  .0012427
     gaitspd |  .0009898
    poorhlth | -.0007748
        gs10 |  .0005548
    drnkspwk |  .0003936
             |
    tandstnd |
       poor  | -.0001638
             |
       _cons |          0
```

```
------------------------
Legend:
  b - base level
  e - empty cell
  o - omitted
```

```
. lassogof cv, over(group)

Penalized coefficients
------------------------------------------------------------------
Name                group |          MSE    R-squared        Obs
------------------------+-----------------------------------------
cv                      |
                      1 |     .011775       0.3460      1,900
                      2 |    .0120212       0.3471        642
------------------------------------------------------------------
```