Covariate Model Building

Hands-On Session – lasso for covariate model building

In this exercise we will build covariate models using lasso on the moxonidine dataset. The setup is similar to that for SCM.

Data set:

    mx19.csv

Model files:

run1.mod = one compartment model parameterized in CL, V and Ka with inter-individual variability (IIV) and inter-occasion variability (IOV). The model estimates CRCL as a linear covariate on CL and assumes V is proportional to WT.

run2.mod= This is same as above but without CRCL on CL and WT on V.

Tasks:

As for SCM the provided dataset contains 5 continuous and 5 dichotomous covariates. The dichotomous covariates are concomitant medications (digoxin, diuretic and ace inhibitors) NYHA, and sex. Continuous covariates are AGE, WT, CRCL, SCR, and NEHA. Originally, NYHA score had 4 different categories but for the purpose of this exercise NYHA has been dichotomized. In this exercise we will test concomitant medications, NYHA, sex and age on CL and V. Unlike scm, concomitant medication with diuretics (DIU) is excluded from testing as only four patients are off this drug. It is generally not advisable to test a categorical covariate with less than 10-20 subjects in each category. This may be particularly unsuitable for a procedure that requires cross-validation.

(i). Using the below command, run the lasso on a poor selection of potential covariates to investigate (Important covariates already in the starting model).

lasso run1.mod -relations=CL:AGE-2,ACE-1,DIG-1,NYHA-1,SEX-1,,V:AGE-2,ACE-1,DIG-1,SEX-1,NYHA-1 -stratify_on=SEX -seed=1 –threads=2 -retries=0

Note: The above command will test all remaining covariates on both CL and V. AGE-2 indicates that this is a continuous covariate tested for a linear relation. ACE-1 indicates a categorical covariate (nominal scale).
Tip: The pdf format of these exercises only allows copy-pasting line-by-line. You may find the three lasso-commands for these exercises in the file: "Files_provided/command.txt". Use this if you want to copy-paste!

(ii). Read the last part of the screen output. Was the outcome in line with that from the corresponding scm exercise?

(iii) Run the lasso again, but use model run2.mod. Change the command to investigate AGE as a piece-wise linear covariate (AGE-3) on CL and, in addition to the previous covariates, investigate CRCL as a linear covariate on CL and WT as a linear covariate on V.

lasso run2.mod -relations=CL:AGE-3,ACE-1,DIG-1,NYHA-1,SEX-1,CRCL-2,,V:AGE-2,ACE-1,DIG-1,SEX-1,NYHA-1,WT-2 -stratify_on=SEX -seed=1 -threads=2 -retries=0

Tip: This command has longer run time. If you have multiple cores on your work station; you can speed up by running nonmem in parallel. For dual cores you stay with the '-threads=2' in the above command line, whereas quad-core should use at least three threads to speed up execution. For five-fold cross-validation, running five threads in parallel is desirable. Note: With -retries=0 PsN will not make further attempts to run the same model, in case of unsuccessful termination. Termination with rounding error may happen during cross validation, but for this exercise will not harm the outcome. Adding "-significant_digits_accept=2" will avoid warning messages regarding termination, but otherwise will not affect the outcome of the lasso.
Plan for a break while the above lasso is running.
On a <u>slow</u> desktop/laptop it may even take longer than 30 minutes. (with ≥5 CPU:s on multi-core or cluster/grid it is a matter of a few minutes)

(iv)
(a) Read the end of the screen output. What t-value was selected in cross-validation?
(b) Open R with the working directory in the directory from where the model files are (i.e. in the same directory as you executed the lasso commands). Open the file "lassoDiagnostics.R". On the top line of the script; rename the path to the correct LASSO directory, unless it is the currently stated "lasso_dir2". This script is used to visualize the lasso cross-validation. Run the commands in R.
 Note: These graphs illustrate the cross-validation, used to select an appropriate t-value. The final selection of covariates occurs by applying this t-value in a single model fit on the whole (original) dataset. This last step is only performed with one t-value and is not illustrated in any graph.
Tip: The final graph sheet (graphic device) contains two pages. Flip between the two using page-up/page-down.
(c) Investigate the created graphs. Note how OFV for the estimation models is continuously decreasing with increasing model flexibility (increasing model size), represented by the t-value. OFV for (external) prediction however, reaches a minimum at the selected t-value. Which validation groups/sets would have preferred a smaller t-value and which preferred a larger t-value? Which are the two most important covariates in each of the five estimation datasets. *Hint: As covariates have been transformed in the lasso; the magnitude of a coefficient reflects how much parameter variability it accounts for. This is not directly translatable to clinical relevance, but still a good measure of which coefficients/covariate relations are important.*
(d) Investigate the final model produced by the lasso procedure
(found at, for example, "lasso_dir2/m1/optimal_model.mod"). What covariates have been selected?
The two covariates selected by scm in the previous scm exercise have been entered below. Enter the coefficients from the lasso in the table below (next page). For this example with an extensive investigation of a mix of useful and useless covariates; how much have these been shrunk compared to the maximum-likelihood estimates (without lasso restriction)? Try to think about different factors that determine how much lasso will shrink different coefficients.

Covariate relation	Coefficients SCM	Coefficients extensive investigation	Shrinkage extensive investigation	Coefficients Good pre-selection	Shrinkage Good pre-selection
CRCL on CL	0.00727				
WT on V	0.0112				

(v)
Run the lasso again, investigating only a few covariates that are more likely to be useful:
CRCL as linear covariate and WT as piece-wise linear covariate on CL. WT as linear
covariate on V. This represents a good pre-selection of covariates to test for inclusion into the
model.

lasso run2.mod -relations=CL:CRCL-2,WT-3,,V:WT-2 -seed=1 -threads=2 -retries=0

(a) Read the end of the screen output. What t-value was selected in cross-validation?
(b) View the end of the final model produced by the lasso procedure:
(for example "lasso_dir3/m1/optimal_model.mod").
Enter the shrinkage for the two covariates in the table above (as "Good pre-selection"). Try to
come up with two explanations to why shrinkage is less than in (iv) above.

(vi)
Optional assignment: Repeat one of the above lasso executions with a new seed number. Do
you get the same t-value as before?