XV_SCM userguide

2013-05-28 PsN 3.6.2

Overview

Cross-validated scm, xv_scm, depends heavily on the scm program, and all scm options apply also to xv_scm except that options search_direction, gof, p_value, p_forward, p_backward and update_derivatives are ignored. Please refer to scm_userguide.pdf for help on scm options.

A word of caution: xv_scm produces many files and takes up much disk space. It is wise to delete all the split_X subdirectories once the results are collected.

Example

xv_scm -config_file=config_xv.scm -groups=5 -splits=3 -seed=12345

Input and options

Required input

A configuration file is required (just as for scm). The format of the configuration file follows the format of the scm configuration file exactly.

Optional input

These options are specific to xv_scm, and they can only be given on the command-line, not in the configuration file.

-groups=N	Default 5. The number of cross-validation groups to make an N-fold cross-validation.
-splits=N	Default 1. The number times to perform a complete cross-validation with a new data split.
-stratify_on=variable	Default not used. PsN will try to preserve the relative proportions of individuals with different values of the stratification variable when dividing the data into 'groups' groups during cross-validation. The stratification variable must be found in the original dataset.

Do not set scm option -only_successful in xv_scm. That option would interfere with the xv_scm algorithm.

Algorithm overview

For each split: Divide the dataset into 'groups' equally sized subsets, using stratification if option -stratify_on is set.

For each data subset: Call the selected subset the prediction/test data and the remaining 'groups'-1 subsets the estimation/training data. Run a regular scm on the estimation data, using the scm input option given on the command-line and the configuration file except forcing options search_direction=forward, p_forward=1, gof=p_value, -no-update_derivatives. For the base model and for the model selected by the scm in each iteration a prediction run is performed. The prediction run is

done by copying the model, updating the initial estimates with the final estimates for the same model based on the estimation data, setting MAXEVAL=0 or equivalent for non-classical estimation methods, and running the model with the prediction data. The OFV of the prediction run is then collected and reported in output.

If the linearization method is used (option -linearize to scm), then a prediction step is needed also for the derivatives model. After running the nonlinear derivatives model on the estimation data, a prediction step is run as above for the derivatives model. Then the derivatives output from the derivatives prediction step replaces the original prediction data in the prediction steps for all the linearized models, including the linearized base model.

Output

The files xv_ofv_results.csv, xv_relation_rank_order.csv and xv_percent_inclusion_by_level.csv contain results and summaries of the runs.

Prediction step with automatic data wrapping

If option linearize is used and the number of data items needed for the linearized model exceeds what NONMEM can handle, scm and xv_scm will try to automatically wrap data. NONMEM model files with multiple $PROB are used for the nonlinear model run to obtain derivatives, to output all the items in $TABLE. $MSFI is used in the extra $PROB to transfer parameter values. When also the first $PROB uses $MSFI, as is the case in the prediction step for for the derivatives run, option POPETAS=*nETA*, where *nETA* is the total number of ETAs in the model, is added to $MSFI for all but the first $PROB in the modelfile.