Bootstrap user guide

2013-10-16 PsN 3.6.10

Introduction

Bootstrap is a tool for calculating bias, standard errors and confidence intervals of parameter estimates. It does so by generating a set of new datasets by sampling individuals with replacement from the original dataset, and fitting the model to each new dataset, see Efron B, An Introduction to the Bootstrap, Chap. & Hall, London UK, 1993. To compute standard errors for all parameters of a model using the non-parametric bootstrap implemented here, roughly 200 model fits are necessary. To assess 95% confidence intervals approximatly 2000 will suffice.

Examples

bootstrap moxonidine.mod -samples=100

bootstrap pheno.mod -samples=500 -seed=12345 -threads=5

Input and options

bootstrap-specific input

A model file is required on the command-line.

-samples=N	The number of bootstrapped datasets to generate. Default is 200.
-sample_size=M	The number of subjects in each bootstrap data set. The default value is set to the number of individuals in the original data set. When the resampling is stratified, the sample_size option can be used to specify the exact number of samples that should be drawn from each strata. Below follows an example of the syntax that should be used in such a case. Stratification is here done based on the study number, STUD, with the values 1001, 1002 and 1003. -sample_size='1001=>12,1002=>24,1003=>10' This example specifies that the bootstrap should use 12 samples from study 1001, 24 samples from 1002 and 10 from study 1003. If only one sample size is used together with stratified resampling (the default case; sample_size=number of individuals in the data set), the strata are assigned samples in proportion to their size in the data set. Please note that this usage of the sample_size option does not guarantee that the sum of the samples of the strata is equal to the given sample_size since PsN needs to round the figures to the closest integer. For a sample size equal to the number of individuals in the data set, the sum will however always be correct.
-stratify_on=<column name>	It may be necessary to use stratification in the resampling procedure. For example, if the original data consists of two groups of patients - say 10 patients with full pharmacokinetic profiles and 90 patients with sparse steady state concentration measurements - it may be wise to restrict the resampling procedure to resample within the two groups, producing bootstrap data sets that all

	contain 10 rich + 90 sparse data patients but with different compositions. The default is not to use stratification. Set -stratify_on to the column (the name in $INPUT) that defines the two groups. Note that the option sample_size has a different behaviour when stratified resampling is used.
-bca	Default not used. When used, the bootstrap utility will calculate the confidence intervals through the BCa method (see Efron B, An introduction to the Bootstrap, 1993). The BCa is intended for calculation of second-order correct confidence intervals. Warning: Using bca is very time-consuming.
-skip_minimization_terminated	Used by default. When used, the bootstrap will skip all samples where the NONMEM run terminated the minimization step. Disable with -no-skip_minimization_terminated.
-skip_covariance_step_terminated	Default not used. When true/used, the bootstrap will skip all samples where the NONMEM run terminated the covariance step.
-skip_with_covstep_warnings	Default not used. When used the bootstrap will skip all samples where the NONMEM run had warnings from the covariance step.
-skip_estimate_near_boundary	Default not used. When used, the bootstrap will skip all samples where the NONMEM run signal that some estimates are near its boundary.
-allow_ignore_id	Default not used. When not used, i.e. by default, bootstrap will print a message and terminate if an IGNORE or ACCEPT statement based on the ID column is found in the $DATA record. This is done because it would interfere with the internal renumbering of individuals that the script does, producing errors. If -allow_ignore_id is used (not recommended), a warning is printed but the program continues execution. Note: The IGNORE statement can safely be used in conjunction with any other column than ID.
-copy_data	Default not set. By default, the bootstrapped datasets bs_pr1_<sample_no>.dta generated in the m1 subdirectory are not copied to the NM_run subdirectories of modelfit_dir1. Relative data paths, ../../m1/bs_pr1_<sample_no>.dta , are used in $DATA in the modelfiles in NM_run. This saves much disk space. If -copy_data is set, the bootstrapped datasets are instead copied to the NM_run directories, and $DATA has bs_pr1_<sample_no>.dta without any path.
-dofv	Optional, default not used. Evaluate original data with bootstrap parameter estimates and compute delta-ofv. See section Computing delta-ofv.
-mceta=N	Optional, default not used. If option -dofv is set and NM version 7.3 or later is used, setting this option will make PsN set MCETA=N in $ESTIMATION. It is up to the user to check that the estimation method used can in NONMEM be combined with option MCETA, PsN will not do that.

Some common PsN-options useful with bootstrap

For a complete list of common options see common_options_defaults_versions.pdf, or psn_options -h on the commandline.

-directory=bootstrap_dirN	The directory in which the script will run NONMEM can be named. The default name is “bootstrap_dirN” where N is increased by 1 each time you run the script. If the run is aborted or crashes, setting the directory to the one from which the script was running earlier can be done. PsN will then not run the model files that had finished, saving time.
-seed=N	A seed for the random number generator can be specified. This makes the run reproducible. Important note: the results of two runs will be different even if the seed is the same if the lst-file of the base model is present at the start of one run but not the other. Running the base model changes the state of the random number generator, and therefore the bootstrapped datasets will be different depending on if the base model is run or not before generating the new datasets.
-clean=N	The user may choose to remove different sets of intermediate file by setting clean to 3, 2, 1, or 0. The higher the number the more files are removed.
-threads=N	The number of parallel processes to start on a parallel computer.
-help	With -help bootstrap will print a longer help message.

Output

The file bootstrap_results.csv contains statistics and summaries specific for the bootstrap.

The raw_results1.csv file is a standard PsN file containing raw result data for termination status, parameter estimates, uncertainty estimates etc. for all model estimations. The first row is for the original dataset.

included_individuals1.csv: One row per bootstrapped dataset. Each row consists of the ID:s of the individuals in that dataset. Note that different individuals may have the same ID number in a NONMEM dataset. The numbers appear in the order the individuals appear in the datafile.

included_keys1.csv: One row per bootstrapped dataset. Each row consists of the internal and unique order numbers (1-N) of the individuals in that dataset. The numbers appear in the order the individuals appear in the bootstrapped dataset.

sample_keys1.csv: One row per bootstrapped dataset. Each row consists of one number C per individual in the dataset, in the order the individuals appear in the original datafile, where C is the number of times the individual is included in the bootstrapped dataset.

The row order is consistent between the files raw_results, included_individuals, included_keys and sample_keys so that row j (excluding headers if any) in each of the files concerns the same bootstrapped dataset.

Known bugs and problems

If a csv-format dataset has an empty first column, i.e. there is a leading comma on each line in the datafile, bootstrap will crash. Avoid the problem by editing the datafile, removing the leading empty column.

It is recommented to remove all $TABLE from the modelfile, otherwise there will be much extra output produced. For the same reason it is recommended to remove PRINT options from $ESTIMATION.

ACCEPT/IGNORE statements based on ID will cause errors since PsN renumbers the individuals during bootstrapping. There is an input check for this, but it is recommended to not rely on this check.

The results of two runs will be different even if the seed is the same if the lst-file of the base model is present at the start of one run but not the other. Running the base model changes the state of the random number generator, and therefore the bootstrapped datasets will be different depending on if the base model is run or not before generating the new datasets.

Technical overview of algorithm

PsN will rerun the base model if the lst-file of the input model (run1.lst if the input model file is called run1.mod) is not present in the same directory as the model file OR if the model file has a different extension than .mod, for example .ctl. If the model file has the extension .mod and the lst-file is present then PsN will simply read the estimates from that file.

The program creates N (N=number of samples), datasets of size M (M=sample_size) by randomly drawing individuals with replacement from the original dataset. The program creates N new NONMEM modelfiles which are identical to the original modelfile with the exception that each uses a different bootstrapped dataset and that the initial parameter estimates are the final estimates from the original run. The model parameters are estimated with each dataset, including the original, resulting in N+1 estimates for each parameter.

Computing delta-ofv

If option -dofv is set, PsN will perform a MAXEVAL=0 run for each set of bootstrap parameter estimates using the original model and data. This is done after all other runs and computations are completed, including any Bca step. The results will be printed to a table file called delta_ofv.csv where the first column is bootstrap data id and the second delta-ofv which is computed as $ofv_{bs,maxev=0} - ofv_{orig}$. If option -mceta=N is set and NM7.3 or later is used and the estimation method is classical then option MCETA=N will be set in $ESTIMATION. There will be up to 200 $PROBLEMs defined in each model file (control stream) to reduce PsN overhead when calling NONMEM.

It is possible to restart a previously run bootstrap to compute delta-ofv without rerunning all bootstrap estimations, provided that option -clean was set to 2 or less in the original run. Simply run bootstrap again with option -directory set to the existing run directory. PsN will intiate reruns of all bootstrap samples, but reread existing output instead of really starting NONMEM. Then the delta-ofv models will be run. If option -bca was set in the original bootstrap run it is important to set -bca also in the restart, otherwise the jackknife results will be overwritten.

Recovering a crashed bootstrap

All the modelfiles and bootstrapped dataset are in the m1 subdirectory. Models that have finished will also have a lst-file. The lst-files have been copied from the NM_run directories in the modelfit_dir subdirectory.

If there is a computer crash in the middle of a long bootstrap, you can still use the samples that did finish.

1. In NM_run directories of modelfit_dir subdirectory, remove the lst-files (all files ending with .lst) from crashed NONMEM runs.
2. Rerun same command (command.txt) with same -seed (check in version_and_option_info.txt) and add option -directory=<name_of_old_directory>. PsN will see that some models finish and only rerun the ones with missing lst-files.

If some samples terminated due to e.g. rounding errors but most of the samples finish okay:

1. Go to the m1 subdirectory and run execute with -retries option on the model that did not finish.
2. After models in m1 have run, use runrecord script (see runrecord_userguide.pdf) to create a summary of all parameter estimates. Open output file from runrecord in Excel and do statistics manually.

If too many samples were discarded due to e.g. minimization terminated and you wish you had set option -no-skip_minimization_terminated: Open raw_results.csv in Excel and do statistics manually.