BOOT_SCM userguide

2013-05-28 PsN 3.6.2

Overview

Bootstrap scm, boot_scm, depends heavily on the scm program, and all scm options apply also to boot_scm. Please refer to scm_userguide.pdf for help on scm options.

Example

boot_scm config_run1.scm -samples=10 -seed=12345

Input and options

Required input

A configuration file is required on the command-line. The format of the configuration file follows the format of the scm configuration file exactly. The input model must be set in the configuration file, it cannot be given on the boot_scm commandline. In addition to the configuration file, one command-line option is mandatory:

-samples=N	The number of bootstrapped datasets to run the scm on.

Optional input

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

-methodA	Default not set. If the scm option -linearize is not set, the bootstrap scm non-linear method will be used. If option -linearize is set, by default the bootstrap scm linear method B will be used. If option -linearize is set together with option -methodA then the bootstrap scm linear method A will be used. If -linearize is not set and option -methodA is set this will result in an error message.
-run_final_models	Default not set. If set then boot_scm will run the final models from each scm on the original dataset and collect the ofv values in the output file ofv_final.csv
-dummy_covariates=<comma-separated list of covariates>	Default not used. If used, a new column for each listed covariate will be added to the dataset, containing a randomly permuted copy of the original covariate column and with header X<name of original covariate>. The dummy covariate will be tested for inclusion in the covariate model exactly like the original covariate.
-stratify_on=<column header or column number>	Default not used. 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. Set -stratify_on to the column that defines the two groups, either the column number or the datafile header for

the column. The stratification variable must be found in the original dataset, it cannot be defined in the model.

Some common PsN-options useful with boot_scm

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

-directory=boot_scm_dirN	The run directory can be named. The default name is "boot_scm_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.
-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 boot_scm will print a longer help message.

Algorithm overview

If IGNORE/ACCEPT is found in $DATA (not counting single character IGNORE like e.g. IGNORE=@), the data will be filtered using a dummy model run in the preprocess_data_dir subdirectory of the boot_scm directory. The new dataset is called filtered.dta. A modified input model orig_model_filtered_data.mod is created where the new dataset is used.

If -dummy_covariates is set, a modified input model (either based on orig_model_filtered_data.mod or on the original input model if no filtering was done) called model_with_xcov.mod is created where the dummy covariates are added in $INPUT and $DATA specifies a new dataset called xcov_<old data name> where the dummy covariates are added. The new model and dataset is created in preprocess_data_dir.

When using method A or Non-linear (i.e. if option -linearize is not set, or options -linearize and -methodA are both set): The program creates 'samples' bootstrapped datasets from the possibly pre-processed original dataset. Then a regular scm is run on each of these datasets, using the options set in the configuration file. Filtering on IGNORE/ACCEPT is skipped in these scm runs, since filtering was done during preprocessing if necessary.

When using method B (i.e. if option -linearize is set but not option -methodA): The program runs the possibly pre-processed input model with the possibly pre-processed dataset using the options set in the configuration file and terminates the run directly after the derivatives dataset has been generated. Then 'samples' bootstrapped datasets are created from the derivatives dataset. A regular scm is run on each bootstrapped dataset, using the options set in the configuration file. In addition to the options in the configuration file, the bootstrapped derivatives data is used as input with option -derivatives_data, which makes the scm run faster since the derivatives generation step can be skipped. In these scm runs

the filtering on IGNORE/ACCEPT is skipped, since filtering was done during pre-processing. If there are time-varying covariates (option time_varying is set in the original configuration file) each scm run will include a run with the original, non-linear model on a bootstrapped version of the possibly pre-processed original dataset, using the same individuals in each sample as in the bootstrapping of the derivatives dataset. This extra run is needed to compute medians for the time-varying covariates.

If option -run_final_models is set: Run the final models from each scm on the original, possibly pre-processed, dataset.

Output

The file bs_ids.csv contains one row per bootstrapped dataset and one column per individual in the bootstrapped dataset. The value in each column gives the original data ID of that individual.

The file ofv_final.csv is only created if option run_final_models is set. It contains one row per bootstrapped dataset plus one for the original, possibly pre-processed, model. It lists the ofvs of the final models from the scm, rerun on the original dataset.

The file covariate_inclusion.csv has one row per bootstrapped dataset. There is one column per parameter-covariate-state combination possible given the test_relations and valid_states settings in the configuration file, excluding state 1 (which means 'not included'). For each bootstrapped dataset the value in the column is 1 if the relation is included in the final model, and 0 otherwise.