BOOT_SCM user guide

PsN 5.2.6

Revised 2018-03-02

Contents

1 Introduction

The boot_scm, (bootstrap of the scm) tool is an implementation of the method presented in [1]. The program depends heavily on the scm tool, and all scm options except base_ofv apply also to boot_scm. Please refer to scm_userguide.pdf for help on scm options.

Examples

```
boot_scm config_run1_nonlinear.scm -samples=100 -seed=12345
boot_scm config_run1_linearize.scm -samples=100  -seed=12345 -methodA
```

2 Input and options

2.1 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 command line.

In addition to the configuration file, one command line option is mandatory:

-samples = *N*
: Mandatory command line option. The number of bootstrapped datasets to run the scm on.

2.2 Optional input

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

-dummy_covariates = *comma-separated list of covariates*
: Default not set. 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. However, a known bug is that boot_scm will

not correctly create a dummy covariate based on a time-varying covariate.

-methodA

Default not set. If the scm option linearize=1 is not set in the scm config file, the bootstrap scm non-linear method will be used. If option linearize=1 is set in the scm config file, by default the bootstrap scm linear method B (see algorithm description below) will be used. If option linearize=1 is set together with option -methodA on the boot_scm command line (no argument to -methodA) then the bootstrap scm linear method A will be used. If linearize=1 is not set and option -methodA is set this will result in an error message. Setting linearize=1 in the scm config file by default gives linearization using FOCE, for details see the scm userguide.

-missing_data_token = *string*

Default is -99. This option sets the string that PsN accepts as missing data, and needs to be set correctly when PsN computes summary statistics for data set columns.

-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

-stratify_on = *item in $INPUT*

Default not set. 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 (the name in $INPUT in the model) that defines the two groups.

2.3 PsN common options

For a complete list see common_options.pdf or type psn_options -h on the command line.

3 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 called orig_model_filtered_data.mod is created where the new dataset is used.

If -dummy_covariates is set, a modified input model (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=1 is not set in the scm config file, or options linearize=1 and boot_scm option -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=1 is set in the scm config file but not option -methodA on the boot_scm command line): The tool runs the possibly pre-processed input model with the possibly pre-processed dataset using the options set in the scm 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 scm configuration file.

In addition to the options in the scm configuration file, the bootstrapped derivatives data is used as input with option -derivatives_data (this is done automatically, the user should not set this option), 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.

4 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.

References

[1] R. J. Keizer, A. Khandelwal, A. C Hooker, and M. O. Karlsson. "The bootstrap of Stepwise Covariate Modeling using linear approximations". In: *PAGE 20 Abstr 2161* (2011).