SCM userguide

2013-07-31 PsN 3.6.3

Introduction

The Stepwise Covariate Model (SCM) building tool of PsN implements Forward Selection and Backward Elimination of covariates to a model. In short, one model for each relevant parameter-covariate relationship is prepared and tested in a univariate manner. In the first step the model that gives the best fit of the data according to some criteria is retained and taken forward to the next step. In the following steps all remaining parameter-covariate combinations are tested until no more covariates meet the criteria for being included into the model. The Forward Selection can be followed by Backward Elimination, which proceeds as the Forward Selection but reversely, using stricter criteria for model improvement. The Stepwise Covariate Model building procedure is run by the PsN tool scm.

Data file

In PsN-3.5.3 and later versions the data file must have a header line, and the column headers must match $INPUT of the input model. This is because the procedure for computing covariate statistics has been changed to accommodate functionality in boot_scm and xv_scm.

Valid states for the parameter-covariate relations

There is a fixed set of parameter-covariate relations defined in scm, but the user can redefine one or more of the states using the code option, see below. This gives very high flexibility. The predefined shapes for the parameter-covariate relations for *continuous* covariates are: none/not included, included as a linear relation, included as a piece-wise linear relation with two slopes (“hockey-stick”), included as an exponential relation and included as a power relation. The pre-defined parameterizations are shown in the section Default parameterizations. By default these parameterizations are assigned to numbers or *states* according to 1:none, 2:linear, 3:hockey-stick, 4:exponential and 5:power, but any relation can be assigned to any number. The exception is the number 1, which must always correspond to none/not included. Further information is given in the help text on [valid_states] and [code].

Categorical covariates are by default either not included or linearly included with an extra parameter added for each but the most common category. The default numbers are 1 and 2.

The numbers 1-5, and higher numbers if the user has defined new relations using the [code] section, are used when specifying included parameter-covariate relations. Relationships are tried in the order the corresponding state numbers are listed in [valid_states], see details below. The default is to try a later relationship in the list only if the preceding creates an significant change in the criteria for selection. Per default the relationships are not tried simultaneously, which means they are not compared with each other. The user can choose to have all later relations tried simultaneously by setting the option -parallel_states, see details in the list of options below.

Inclusion and elimination criteria

A combination of a parameter and covariate is included if the change of OFV is significant according to the input p-value. The limit where the change is considered significant depends on the complexity of the relationship and is expressed in degrees of freedom. The linear relation adds one degree of freedom, the hockey-stick relation adds two, exponential only adds one. Categorical covariates adds as many

degrees of freedom as there are categories minus 1. In forward search, for each covariate addition the program computes the lowest p-value for which this addition would still be significant. The covariate addition with the lowest p-value is selected, provided that this p-value is lower than the input forward p-value. In backward search the program computes which p-value would be required to accept going from each candidate reduced model to the current model. The candidate reduced model which would require the highest p-value to be replaced with the current model is selected, provided that this p-value is higher than the input backward p-value.

Input

There are so many options to SCM that it is unpractical to specify them all on the command-line. Therefore the only option required on the command-line is the name of a configuration file, where other options are specified. Additional options are required, but may be specified either on the command-line or in the configuration file. If an option is specified both on the command-line and in the configuration file, the choice on the command-line has precedence. There are also options, of which some are required, that can only be specified in the configuration file. On the command-line options always begin with -, but in the configuration file the '-' must always be omitted.

Configuration file

A set of example and template configuration files are found in the documentation directory, together will all userguides.

The configuration file consists of a list of input options, followed by a set of sections with headers in square brackets. It lists the parameters and covariates, their types and which combinations that should be tested. The header of all sections must be enclosed by square brackets, see examples below.

At the top of the configuration file you can specify any scm command-line options and *some* general PsN options that are normally given on the command line, such as “threads” and “nm_version” (see description of option config_file below for a more complete list). These options can still be given on the command line and will then override any options set in the configuration file. **Important note:** All command-line options that the user chooses to set in the configuration file must come *before* all bracket-header sections in the configuration file, otherwise the command-line options will be ignored.

You can place a comment anywhere in the file by prefixing the line with a semicolon. ***Do not*** add comments at the end of lines, always put comments on separate lines.

Options

Required input which must be given on the command-line

-config_file=<filename> Name of an scm configuration file, including search path if file is not stored in current directory. The configuration file may contain all scm command-line options and some optional general PsN options: -nm_version, -nmfe_options,-abort_on_fail, -compress, -directory, -extra_data_files -extra_files, -picky, -retries, -threads and -tweak_inits. The file common_options_defaults_versions_psn.pdf describes all these options.

As of PsN-3.4.3 it is possible to skip the option name for the configuration file, and call scm with
scm config.scm
instead of the traditional
scm -config_file=config.scm.

Required input which can be given on the command-line or in the configuration file
When set in the configuration file, all these options must be set *before* the first [<section>].

-search_direction=<direction>	No default. Possible choices are 'forward', 'backward' or 'both'. If backward is chosen, included_relations is required input in the configuration file.
-model=<filename>	No default. The name of the basic model file.
-categorical_covariates=<list>	A comma separated list of the categorical covariates to be tested. The listed covariates must be found in the data set. May be omitted if continuous_covariates is specified.
-continuous_covariates=<list>	A comma separated list of the continuous covariates to be tested. The listed covariates must be found in the data set. May be omitted if categorical_covariates is specified.

Optional input which can be given on the command-line or in the configuration file
When set in the configuration file, all these options must be set *before* the first [<section>].

-gof=<>	Default is 'p_value', other alternative is 'ofv'. Option decides which goodness-of-fit criteria should be used for deciding which model is better when comparing.
-global_init=<>	Default is 0.001. With global_init option the initial estimates of parameters in covariate function parameterizations are set to global_init. If using inits section in configuration file individual initial values are used instead of one global.
-logfile=<filename>	Default is scmlog.txt
-p_forward=<>	Using the p_forward option, you can specify the p-value to use for the forward selection. Used when an improvement is good enough for including parameters.
-p_backward=<>	Using the p_backward option, you can specify the p-value to use for the backward deletion.
-p_value=<>	Use this option to set the p_value for both forward and backward steps.
-do_not_drop=<>	Default empty string. Since the number of columns are restricted to 20 for NONMEM (50 for NM7) it is necessary to minimize the number of undropped columns. The scm utility uses the '=DROP' syntax of NONMEM to exclude the covariate columns that are not currently tested in a specific candidate model. If some covariates are used in e.g. PK, PRED or TABLE in the basic model you must list them using the do_not_drop option to prevent them from being dropped. If the -linearize option is used, do_not_drop has a different usage, see the section on the linearization method below.

-only_successful	Default not set. Only consider runs with MINIMIZATION SUCCESSFUL (or equivalent for non-classical estimation methods) when selecting the covariate to add/remove in each step.
-parallel_states	Default not set. Instead of trying valid_states sequentially, try all states after the current one simultaneously. It is currently not possible to stop testing a parameter-covariate relation that has been included in an lower state if there are higher states defined. To achieve this the user has to stop the search and restart with a new configuration file where no higher states are tested for the parameter-covariate pair.
-max_steps	Do not take more that max_steps forward steps, even if there are more covariates left to add and a significant inclusion was made in the last step.

Optional input which can only be given in the configuration file

When set, all these options must be set *before* the first [<section>].

-base_criteria_values=<perl code>	The argument to this option is some perl code (a hash). The base_criteria_values overrides default criteria for the goodness of fit function used. The default goodness of fit function takes its base_criteria_value to be the ofv of the initial model. The ofv of the initial model can be set by base_criteria_values={ofv => 670} where the number 670 should be be replaced by the ofv value of the model evaluated with all included relations. If base_criteria_values is set, then PsN does not have to evaluate the original model with included_relations to get a ofv to use for comparisons.
-logit=<list of parameters>	By default the value of the covariate function for a parameter is a product of individual functions of the included covariates, and the typical value (TV) of the parameter is multiplied with the product. With option -logit the user can specify a list of parameters which should be treated as logit transformations, i.e. the value of the covariate function is a sum of individual functions, the sum should be added to the typical value of the parameter, and the value of each individual covariate function should be 0, not 1, when the corresponding THETA is 0. The thetas in a linear or hockey-stick relation on a logit parameter will be given lower and upper boundaries -20 and 20 by default. Important note: When a parameter is a logit transformation, the power and exponential covariate relations are not appropriate to add on this parameter.
-time_varying=<list of covariates>	Handle time-varying covariates when computing median and mean of covariates. Requires that W is defined in $ERROR/ $PRED. PsN does not check that W is defined. There will be an NMtran error if W is missing.

Assume an ID has the following observations: a,b,c,d
.
Corresponding covariate values: A,B,C,D

Corresponding residual error magnitudes (W): Wa,Wb,Wc,Wd

Corresponding absolute partial derivatives for ETA(1): Ga, Gb, Gc, Gd

Average covariate value for this individual AVCOV(ind) will be

(A*Ga/Wa + B*Gb/Wb + C* Gc/Wc + D* Gd/Wd)/
(Ga/Wa + Gb/Wb + Gc/Wc + Gd/Wd)

The median/mean for this covariate with respect to the parameter which ETA(1) is on is the median/mean of AVCOV over all individuals.

The derivatives needed for computing the new medians and means will be taken from an estimation of the input model without any included relations, even if included_relations is defined in the config file. This is because the medians are needed to add the included relations to the model.

Some optional general PsN options which are extra important for scm

-missing_data_token=<> Default is -99. When some values of the covariates are missing in for some observations in the dataset it is necessary to let PsN know how missing values are coded in the dataset, otherwise the covariate function values will be very wrong. This option cannot be set in the configuration file (it will be ignored if set there). It must be set on the command-line.

-directory=<> The name of the run directory. The default is scm_dirN, which N either 1 or the next unused number if there already are scm_dirN directories present. When set in the configuration file, this option must be set *before* the first [<section>].

Options for the linearization method

Add covariates to a linearized version of the original model. All these options must be set *before* the first [<section>] of the configuration file. The linearization method can only be used with models where the eta on each parameter (CL, V,...) to be tested is unique to that parameter, i.e. if CL is listed in test_relations and CL=TVCL*EXP(ETA(5)) then ETA(5) must not appear anywhere else in the model. Also, the linearization method is restricted to a certain set of models for the inter-individual variation, see below. Different parameters may have different ETA relations as long as each one is

found in the list below.
The recognized forms of inter-individual variation models are (using parameter PAR and ETA(1) as examples):

```
PAR=TVPAR*EXP(ETA(1))          ; exponential
PAR=EXP(MU_1+ETA(1))           ; exponential
PAR=EXP(ETA(1)+MU_1)           ; exponential
PAR = TVPAR+TVPAR*ETA(1)       ; proportional
PAR = TVPAR +ETA(1)*TVPAR      ; proportional
PAR = TVPAR*(1 +ETA(1))        ; proportional
PAR = TVPAR + ETA(1)           ; additive or logit depending on if PAR is flagged as a logit using
                                 option -logit
```

To handle logit transformations, the user can model as follows

```
TVBIO  = THETA(1)
TVPHI = LOG(TVBIO/(1-TVBIO))
PHI    = TVPHI+ETA(1)
BIO    = EXP(PHI)/(1+EXP(PHI))
```

and in the configuration file set PHI as the parameter to test covariates on using [test_relations] and set logit=PHI.

The covariate function will always be multiplied with (or added to if the parameter is a logit) TVPAR, so TVPAR must be used in the model.

-linearize	Turn on the linearization method. By default not set.
-foce	Linearize around empirical Bayes estimates. Default set. If unset linearize around etas equal to 0.
-second_order	This option is not fully implemented yet. Default not set. Use a second order Taylor expansion instead of a first order.
-lst_file=filename	Default not used. Update original model with final estimates from this file before running model to obtain derivatives.
-update_derivatives	Default not set. In each step, after selecting the covariate to add, rerun non-linear model to get updated derivatives for linearization.
-epsilon	Default set. Linearize with respect to epsilons in addition to etas. Disable with epsilon=0 in the configuration file or -no-epsilon on the command-line, then no linearization with respect to epsilon is performed.
-error=<add,prop,propadd,user>	No default. Only allowed to set when -no-epsilon is set. Use an approximate linearization of the error model instead of exact linearization with respect to epsilons. Alternatives are add for additive, prop for proportional, propadd for proportional plus additive and user for user defined. For additive and proportional error (add and prop) it is required that the original model defines a variable W for weighting of EPS(1), for add W = THETA(x)

	Y = IPRED + W*EPS(1) and for prop W = THETA(x)*IPRED Y = IPRED + W*EPS(1) For proportional plus additive error (propadd) it is required that two variables WA and WP are defined so that the following holds WA=THETA(x) WP=THETA(y) W = SQRT(WA**2+(WP*IPRED)**2) Y = IPRED + W*EPS(1)
	If -error=user then the option -error_code must be defined in the configuration file, and probably also do_not_drop (see below).The code can only use IPRED, EPS(x) and parameters listed with do_not_drop. IPRED must be used. F cannot be used. The code must have a \ (backslash) at the end of each line, except the last which cannot have a backslash. The code must not contain blank lines or comments.
-error_code=<nonmem code>	Only if error=user. Can only be set in the configuration file. Define Y, possibly on several lines, with NONMEM code using IPRED, EPS(x), and any parameters listed in do_not_drop. IPRED must be used for Y and F must not be used. When the NONMEM code spans several lines, each line must end with \ (a backslash) except the last one. It is important not to have \ at the end of the last line. There must be no empty lines and no comments in the error_code. THETA(x) cannot be used. It will not work to have etas on the epsilons.
-do_not_drop=list of parameters	When option -linearize is set, option -do_not_drop has a different functionality than otherwise. In combination with -linearize, this option must list all parameters except IPRED and EPS(x) that are used in -error_code. It must also list parameters that are needed for IGNORE or ACCEPT statements. No other parameters should be listed here, for example do not list covariates that are in $PK/ $PRED but not in error_code.
-derivatives_data=filename	Can only be set in the configuration file. It is possible to reuse the derivatives data from a previous run, the file derivatives_covariates.dta, provided that the nonlinear model, the included_relations, the list of covariates and the do_not_drop list is the same. The program does not check that these conditions are fulfilled, so if they are violated NMtran will fail or the program output will be incorrect.
-noabort	Default not set. Only relevant if -linearize is set. With this option set, scm will add NOABORT in $EST of the linearized models.

Configuration file sections

All sections have a header enclosed by square brackets. The sections must come after the single line options. All the following sections except [test_relations] are optional.

test_relations
The section test_relations is required, and in there you list which covariates should be combined with which parameter. Implicitly you also tell the SCM which parameters there are. E.g.:

```
[test_relations]
CL=WGT,AGE
V=SMOK
```

Note the square brackets around the section name. In the example WGT and AGE will be combined with clearance (CL) and SMOK will be combined with volume (V)

valid_states
The section [valid_states] lists which states should be tested for the covariates (default is 1,2 for categorical and 1,2,3 for continuous). Always put 1 first in the list. The scm was implemented to not allow setting different valid states for different covariates, but since any relation can be coupled to any state number except 1, this imposes no limitation. The user can replace the default relations coupled to the states for some parameters and or covariates using the [code] section of the configuration file, see more information below under **code**. The relationships are tried in the order their numbers are listed in [valid_states]. The numbers do not have to be increasing, it is possible to set e.g. 1,5,4 as valid_states for continuous covariates in which case (assuming the default state-relationship definitions) the power relation would be tested before the exponential. Per default, only if a relationship creates an significant change in the criteria for selection, the next relationship in the valid_states list is tried. The relationships are not tried simultaneously so that they can be compared with each other. The user can set option -parallel_states to make scm test all later states in the list simultaneously.

```
[valid_states]
continuous = 1,2,4
categorical = 1,2
```

included_relations
If you want to include some relations from the beginning, you can use the [included_relations] section:

```
[included_relations]
CL=WGT-2,AGE-3
```

Here WGT will be included linearly and AGE with the hockey-stick relation (assuming default relations for states 2 and 3) on CL already from the start. Following states in the list of valid_states will be tried by the SCM, and the combination can be removed as part of the backwards elimination. If option search_direction is set to backward, the section [included_relations] is required (otherwise there will be no relations that PsN can remove during the backward search).

If included_relations is defined, PsN will run the input model with these relations before starting the search, in order to have a reference ofv value for computing the improvements when adding covariates. If base_criteria_values is set with an ofv-value, then the run with included_relations is skipped.

code

The code section allows you to define your own relation code for the different states. Never redefine state 1 (relation not included). The line [code] must only appear once, even if there are several different definitions in the section. For the code section you can write any NONMEM code you like on the right hand side of the first equal sign and it will be inserted into the model when that combination is included. In the code PsN will replace PAR with the parameter, for example CL, and COV with the covariate, for example WGT. In addition, median, mean, maximum and minimum will be replaced with the median, mean, maximum and minimum values of the covariate. The wildcard * can be used to define for many parameters and or covariates at once. E.g.:

```
[code]
V:APGR-7=VAPGR=(THETA(1)*(APGR-7.0))
V:POP-7=VPOP=(THETA(1)*(POP-median))
CL:*-7=IF(COV.EQ.0) CLCOV = 1\
        IF(COV.NE.0) CLCOV = (1 + THETA(1))
```

The THETAs should be numbered from 1, starting over from 1 for each new parameterization. The numbers will be appropriately replaced when the relation is added in the model. You can either specify the code exactly (except for the THETA index) as in the V example, or you can use the wildcards * for the parameter and/or covariate to define multiple relations on one line as in the CL example. If you use * on the left-hand side for the covariate and the name of the covariate is needed in the code, you should use COV. Similarly, if the wildcard * is used for the parameter on the left-hand side, PAR should be used in the code. PAR and COV will be substituted for the parameter and covariate names in the combination being included. If a single definitions spans multiple lines there must be a "\" at the end of all but the last line.

It is possible to use the variables maximum, minimum, mean and median in the code. These words will be replaced by the actual maximum, minimum, mean and median computed for the covariate from the dataset. The words must be in lower case and spelled exactly as listed, otherwise they will not be recognized by PsN but passed on to the modelfile where they will cause errors.

Note that it is possible to use the ANCHOR functionality, see below, to define constants to use in the user-defined code.

You can give the relation any state number you like. You can choose a state number which has a default relation (2-5) or add a new one.

Redefining existing states is useful when wanting to test different valid states for different covariates on different parameters. Starting with PsN-3.2.8 it is possible to use shortcut definitions for the pre-defined states when redefining using the code section. For example, if valid_states for continuous covariates is 1,2,4 but the user wants the last valid state to be the power relation for covariate WGT, the user can redefine state 4 for WGT as follows

```
[code]
```

*:WGT-4=power
Then in state 4 the covariate WGT will always be added as a power relation instead of exponential (the default for state 4). The shortcuts are 'none' for not included, 'linear', 'hockey-stick', 'exponential' and 'power'. The shortcut definitions must be written without quotes, in lower case and be spelled exactly as listed.

inits
The inits and bounds sections lets you set initial estimates for thetas introduced in the code and their corresponding bounds. The default bounds are set so that covariate functions cannot reach negative values, which is appropriate e.g. for parameters V and CL. For parameters that should be allowed both negative and positive values the user may wish to change the boundaries. The inits and bounds sections have very similar syntax, only inits will be shown here:

[inits]
V:SMOK-2=0.01
CL:*-2=0.01
*:AGE-3=0.03,0.03

The left hand side is of the form “PARAMETER:COVARIATE-STATE”. Both PARAMETER and COVARIATE can be an asterisk (*), which is a wildcard and is the same as repeating the line for each parameter and/or covariate. The STATE can be any one of 2-5 or a new state defined by the user in the code section. For the “inits” section the right hand side is a comma separated list of initial values for theta parameters put into the abbreviated code corresponding to the combination. Unless the states are redefined using the code option, for continuous variables the list for state 2 should be one value, for state 3 two values and for states 4 and 5 one value. For categorical covariates where STATE can only be 2 it is the number of categories that decides the number of initial values. The lower_bounds and upper_bounds sections follows the same rules and corresponds to the bounds for the theta parameters put into the abbreviated code. If you specify too few or too many values, those you exclude will be replaced by defaults and the extraneous will be ignored.

It is possible to use the variables maximum, minimum, mean and median in the inits and bounds sections. See more details in the [code] help.

The ANCHOR functionality

SCM by default adds the covariate relationship functions first in $PK/$PRED, before any existing user-written code. Sometimes this is unpractical, for example if there are constants that the user needs in a custom [code] section. In this case the user can add the line
;;;SCM-ANCHOR
at the line where it is okay for SCM to start to add the covariate function code, e.g. the line after the last constant has been defined. The line must look exactly as shown, for example there must be no spaces on the line and SCM-ANCHOR must be in capital letters. If the line is found, SCM will always write the lines up to and including
;;;SCM-ANCHOR
before adding any covariate code. If the linearize option is used, the code preceding the SCM-ANCHOR line will be kept in the linearized models. The ANCHOR functionality has no option, it is turned on and off by adding or removing the line in $PK/$PRED.

Missing coviariate values

The scm program handles missing coviariate values, provided that option missing_data_token is set correctly. If the covariate value is equal to missing_data_token, the missing value will be replaced by the median of the covariate. In the code this is done in a separate IF statement for missing values where the covariate value has been set to the median and the expression simplified, usually leading to the expression being equal to 1. For time-varying covariates the principle is the same: replace missing values with the median. See above help text for option time_varying for a description of how to compute medians for time-varying covariates.

Example

scm -config_file=config_run5.scm -nm_version=7

```
;contents of config_run5.scm
model=pheno_with_cov.mod
search_direction=both
p_forward=0.05
p_backward=0.01
abort_on_fail=0
retries=3
continuous_covariates=WGT,APGR,CV1,CV2,CV3
categorical_covariates=CVD1,CVD2,CVD3

[test_relations]
CL=WGT,APGR,CV1,CV2,CV3,CVD1,CVD2,CVD3
V=WGT,APGR,CV1,CV2,CV3,CVD1,CVD2,CVD3

[valid_states]
continuous = 1,2,4
categorical = 1,2
```

SCM output

The final models are in the final_models subdirectory of the run directory. The logfile is scmlog1.txt, unless a different name was set using option logfile. The logfile contains ofv-values and p-values for the tested models, and tells which covariate relation was added or removed in each step. It also lists the directory where models from each step are found. All parameter estimates are in raw_results1.csv. If option search_direction is 'both' then the first level of the run directory is about the forward search (m1, modelfit_dir1 belong to the forward search). The next levels of the forward search are in forward_scm_dir1 and below. The backward search is found in backward_scm_dir1 inside the main run directory.

Diagnostics of the linearization method

If the linearization method was used, the logfile also lists the ofv values of the nonlinear base model (the derivatives model) and the linearized base model. Those two values should be very similar. If they are not, the results cannot be trusted. Check the files derivatives.lst and base_model_with_included_relations.lst for NONMEM minimization error messages. If the messages

give no obvious explanation for why the values could be different the automatic linearization might have failed. Check if the program detected the correct type of ETA relationship (exponential/additive/proportional) for each parameter.

Known bugs and problems

Two covariates will be mixed up if the complete name of one is the same as the beginning of another. Example: AP and APGR will result in errors, since AP is identical to the beginning of APGR. APX and APGR is okay, since the third letters X and G are different.

Tricks

To build the full model, use p_forward=1. Then any parameter inclusion that gives a drop in ofv will be accepted. Since the initial estimates of the model are updated after each parameter inclusion, the forward search may make it easier to estimate the parameters in the full model, compared to when adding all relations at once.

To only run one step and not accept any relations, use p_forward=0. Then the user can investigate raw_results and the log-file to select the appropriate relation to add, and start a new run with [included_relations] set.

Default parameterizations

Replace PAR with the parameter name, e.g. CL, and COV with the covariate name, e.g. WGT or SEX. If the parameter is listed as a logit the offset 1 in default functions for state 1 and 2 will be changed to 0.

Default state 1:Covariate not included on the parameter
PARCOV=1

Default state 2, continuous covariates: Linear function
PARCOV= (1 + THETA(1)*(COV - median))

Default state 2, categorical covariates: Linear function (will be one extra line for each extra category)
IF(COV.EQ.1) PARCOV = 1 ; Most common
IF(COV.EQ.0) PARCOV = (1 + THETA(1))

Default state 3: Hockey-stick, or piece-wise linear, function. Continuous covariates only.

IF(COV.LE.median) PARCOV = (1 + THETA(1)*(COV - median))
IF(COV.GT.median) PARCOV = (1 + THETA(2)*(COV - median))

Default state 4: Exponential function. Continuous covariates only.
PARCOV= EXP(THETA(1)*(COV – median))

Default state 5: Power function. Continuous covariates only.
PARCOV= ((COV/median)**THETA(1))