User Manual for NSweepPlus

# About NSweepPlus

NSweepPlus is an SPSS macro that reads in a set of variables from an SPSS system file and identifies variables that have very low variance (<CRIT), and variables that are collinear with one or more variables, or a combination of them, on the set of variables (with residual variance <CRIT after other variables have been accounted for). The collinearity is analyzed in the order in which the variables are specified in the parameter VARS. For example, when analyzing variables Var1, Var2 and Var3, where Var1 and Var3 are collinear, it will identify Var3 as collinear with the previous variables. If the variables are entered as Var3, Var2 and Var1, it will identify Var1 as collinear.

With the set of variables that are not colinear and do not have low variance, it conducts a principal components analysis, creates principal component scores and saves them to a file.

When a PROXY is specified, it also carries out a regression analysis using PROXY as the dependent variable and all principal components as independent variables. The selection of the principal components is based on the SPSS forward selection method.

NSweepPlus produces as output a list of all the variables specified for the analysis, with a column indicating if the variable had low variance (LoVariance = 1) or if the variable was collinear with the previous variables in the analysis (IsCollinear = 1). It also produces a file with the principal component scores, and a summary file for the principal components with information about the amount of variance accounted for by each of them.

Due to the nature of the analysis performed by NSweepPlus, all the variables specified in the analysis must be numeric, and these must not have any missing values, either system missing values or user defined system missing values.

Sampling weights, when available, can be used for the analysis, and you can also select and work with a subset of the data.

NSweepPlus identifies low variance and collinear variables following these steps:

1. Creates a variance-covariance matrix with the data
2. Identifies variables with low variance
3. Uses the SWEEP operator on the variance-covariance matrix, column by column, skipping a column when the pivot for the matrix is very small (<CRIT), and identifying the variable for this column as collinear. The variables are swept in the order in which they were specified in the VARS parameter.

# Requirements

NSweepPlus requires a working copy of SPSS installed in the computer where the analysis is conducted. It also requires an SPSS system file with more than one variable stored in numeric format. The data matrix should be complete. Presence of system missing values will yield unpredictable results. User defined missing values are treated as valid values.

The original data is not changed or modified in any way.

# Using NSweepPlus

To use NSweepPlus, you will need to create an SPSS syntax file with an INCLUDE command, followed by a call to the macro. The syntax will look something like this:

include file = "C:\Users\SPSS Macros\NSweepPlus\NSweepPlus.ieasps".

!NSweepPlus

InDir = "C:\Temp\KSA\_IRTCourse201912\Conditioning" /

InFile = asgara2019\_withscores\_BQ\_X\_contrasts /

OutDir = "C:\Temp\Test" /

OutFile = NSweepOutputFile /

Vars = Nationality\_d2 to QS461\_d11 /

ViewCod = Y /

Wgtvar = SampWgt /

Crit = 0.001 /

Mxloops = 10000.

Notice that the first line INCLUDEs the macro, and the following lines are the call to the macro. Some of the parameters are optional and can be left blank. You only need to include the macro once per session.

When calling the macro, you need to use the exclamation point (!) before NSweepPlus, as in !NSweepPlus. Following !NSweepPlus you will have a series of parameters with the corresponding value(s). Each parameter is separated by a slash, much in the same way you would separate subcommands when writing SPSS syntax. The last parameter ends with a period (“.”). Some commands are mandatory, like the name and location of the input and output file, but others are optional. The section on parameters in this User Manual will specify the parameters, their use, and accepted and default values.

The parameters and corresponding values are not case sensitive. The order of the parameters is also not important. Neither is the alignment, other than they should not start on the first column. You can also have multiple parameters in a single line, but always separated by a slash. If you repeat a parameter in the call, the last instance of the parameter will be the one used by the program.

The alignment and sequence of the parameters in the syntax above is done as such for looks and ease of reading.

# Output Files

Running NSweepPlus creates several output files in SAV and CSV format. The files are saved to the directory specified by OUTDIR and are named as [OUTFILE]\_[FileDescriptor].

## [OUTFILE]\_SweepSummary.csv and .sav

This file contains the summary of the sweep performed on the set of variables fed into the program. The columns for this file are the following:

VarName Name of the variable in the [INFILE].

Position Sequential position of the variable in the list of variables provided in the parameter VARS.

Label Variable label, if any was provided. If none was provided, this column displays the text <none>.

Mean Weighted mean for the variable.

SumSqDiff Sum of the squared differences for the variable.

Variance Variance for the variable.

StDev Standard deviation for the variable.

LoVariance Indicator of whether the variable was identified as having low variance (< CRIT).

IsCollinear Indicator of whether the variables is collinear with one or more of the variables that preceded it in the list of variables.

InDir Directory from where the data file was read.

InFile Name of the file from where the data were read.

Selvar Selection variable(s) used to subset the data.

SelCrit Selection criteria used to subset the data.

Weight Name of the variable containing sampling weights, if these were used.

## [OUTFILE]\_PComp.csv and .sav

This file contains one record for each case read from the [INFILE]. If a [SELCRIT] was specified, cases are selected using this selection criteria. If a [WGTVAR] was specified, only cases with [WGTVAR] > 0 are selected. The columns for this file are the following:

CaseSeq The case sequence number in the [INFILE].

IDVARS The identification variable(s), if specified.

WgtVar A weight variable. If none was specified, this takes on the value of 1 for all cases in the file.

PComp[1…] The principal components calculated from the data. There will be as many principal components as there are variables in the [VARS] parameter that are not collinear or have low variance. These are scaled to have mean of 0 and standard deviation of 1.

PCompR[1…] The rescaled principal components calculated from the data. This is the PComp rescaled to have mean of 0 and standard deviation equal to the squared root of the eigenvalue for the principal component.

[PROXY] When a PROXY is specified, this will be included in the file.

## [OUTFILE]\_VarExplained.csv and .sav

This file contains one record for each principal component calculated from the data. The columns for this file are the following:

PComp The sequence number of the principal component.

Eigenvalue The corresponding eigenvalue for the principal component. The eigenvalues over all the principal components add up to the number of variables in the principal components.

Variance The proportion of the overall variance accounted for by the principal component.

VarianceCumulative  
The cumulative proportion of variance accounted for by the current principal component and all previous ones.

NCases Number of cases in the input file

PctOfSample The ratio of the Principal Component to the sample size (NCases). In some applications, principal components are selected to not exceed 5% of the sample size.

## [OUTFILE]\_PCASyntax.ieasps

This is the SPSS syntax used to create the principal components.

## [OUTFILE]\_ComponentMatrix.csv and .sav

This file contains one record for each variable in the analysis. The columns for this file are the following:

Variable The name of the variable in the analysis

@1 to @## The correlation between the variable and the corresponding principal component. The square of this correlation represents the proportion of the variance of the variable accounted by the principal component. Summing across the squared columns will give you the proportion of the variance accounted for, with a maximum of 1 for each variable. Summing down the squared rows for each principal component will give you the eigen value for the corresponding principal component.

## [OUTFILE]\_ModelSummary.csv and .sav

When [PROXY] is specified, a linear regression is calculated using all principal components as independent variables and the [PROXY] variable as a dependent variable. This file has a single record.

R The multiple regression coefficient for the step

RSquared The squared multiple regression coefficient.

AdjustedRSquared  
The adjusted R Squared coefficient.

Std.ErroroftheEstimate  
The standard error of the RSquared.

## [OUTFILE]\_RegCoefficients.csv and .sav

When [PROXY] is specified, a linear regression is calculated using all principal components as independent variables and the [PROXY] variable as a dependent variable. This file has one record for the constant in the equation, and one for each principal component in the equation. The coefficients included are only those of the final solution. The columns in the file are the following:

VarName The name of the variable in the equation, or the constant.

B The regression coefficient.

Std.Error The standard error of the regression coefficient.

Beta The standardized regression coefficient.

t The t-value for the regression coefficient, calculated as (B/Std.Error).

Sig The significance of the t value.

BetaSQ The square standardized regression coefficient. The sum of these across all principal components add up to the RSquared in the model file.

# Parameters

What follows is a list of the parameters that can be used when calling NSweepPlus. They are in the logical order in which you would want to consider them, and not in alphabetical order. But the order of the parameters in the call can be any order. Some parameters are optional, some are mandatory, and some have default values. This is all indicated below.

Depending on the parameter, the values assigned can be one or many. They are not case sensitive. The parameters are assigned a value or values using the equal sign. When more than one value is specified for a parameter, these need to be separated with spaces, and the values from one parameter to the next are always separated by a forward slash (‘/’).

## Data Input and Output

INDIR (not optional; no default)

Directory with the file that has the variables that will be inspected. The directory location must be fully spelled out and enclosed in quotes.

Example:

InDir = "C:\Temp\KSA\_IRTCourse201912\Conditioning" /

INFILE (not Optional; no default)

File with the variables that will be inspected. It does not need the SAV extension for the file name.

Example:

InFile = asgara2019\_withscores\_BQ\_X\_contrasts /

OUTDIR (not optional; no default)

Directory where to write the results. The directory location must exist and be accessible from your computer, be fully spelled out and enclosed in quotes. The program outputs a single file to this directory using the naming convention [INFILE]\_SweepSummary.

Example:

OutDir = "C:\Temp\Test" /

OUTFILE (not optional, no default)

Name used for the output files. Presented as [OUTFILE] in the descriptions above. This value cannot start with a number, cannot have special characters or spaces, and must comply with the SPSS variable naming convention.

Example:

OutFile = PrincipalCompResults /

SELVAR (optional; no default)

A variable or list of variables used to subset the data read from the INFILE. The variable(s) used to subset the data can be one of the VARS. Therefore, you only need to specify the SELVAR when this is not specified as one of the VARS.

Example:

Selvar = var1 var2 /

SELCRIT (optional; no default)

The selection criteria to apply when reading the INFILE. Only records that meet the specific selection criteria are read and accounted for in the processing.

The selection criteria must be specified using SPSS IF/THEN logic, but no IF is needed. Any, and all operations must fit in a single line ending with a forward slash (“/”).

The selection criteria is applied whenever SELCRIT is not blank, regardless of whether the SELVAR parameter if left blank or not.

Example:

Selcrit = (var1 = 1 and var2 > 4) /

In the example above, only cases with VAR1 = 1 and VAR2 > 2 will be read from the INFILE. Parenthesis are used for readability but are not necessary.

IDVARS (optional, no default)

Variable(s) used to identify the cases in the [INFILE] and [OUTFILE]. This is an optional parameter. Regardless of whether this parameter is used, the file with the principal components includes a variable called CaseSeq identifying the record number in the [INFILE].

Examples:

IDVars = IDCNTRY IDSTUD /

VARS (not optional; no default)

The list variable names to inspect for low variance and collinearity. These can be listed one after another, with spaces in between, or using the TO when specifying a set of variables. You can use multiple lines, but none should begin in the first column. The last variable name should be followed by a forward slash. The variables will be read, processed, and assigned a position number (POSITION) based on the order in which they appear on this list, regardless of the order of the items in the INFILE. Variables can only be numeric.

Examples:

Vars = Nationality\_d2 to QS461\_d11 /

Or

Vars = Var1 Var2 Var7 to Var11 /

Or

Vars = Var1 Var2 Var7 Var8 Var9 Var10 Var11 /

The two examples above show two ways in which variables can be specified. Only one VARS is accepted in any call of the macro.

PROXY (optional, no default)

Variable used as a dependent variable in a regression equation calculated with all the principal component treated as independent variables. Only one variable can be specified.

Examples:

PROXY = Reading /

WGTVAR (optional; no default; options = a single variable with the survey or sampling weight)

The name of the variable with the survey or sampling weight. When no WGTVAR is specified, NSweepPlus assigns all cases a weight of 1, therefore all response records contribute equally to the calculations. When a WGTVAR is specified, this is used to weight the response records. Cases with zero, missing and negative sampling weights are excluded from the analysis from the outset.

Example:

wgtvar = TOTWGT /

VIEWCOD (optional; default = N; options = Y or N)

Expands and displays in the output file the code generated by the macro. Useful for troubleshooting the processing and identifying where exactly problems or warnings occur, if any.

Example:

viewcod = N /

CRIT (optional; default = 0.001)

The criteria to use to determine whether a variable has low variance or contributes little additional variance after the previous variables have been accounted for. A variable with variance of < CRIT is flagged as having low variance. A variable with variance of < CRIT not accounted for by previous variables in the matrix is flagged as colinear.

Crit = 0.001 /

MXLOOPS (optional; default = 100000)

By default, SPSS sets the maximum number of loops to 40. However, when inspecting the data, the program needs to loop over the number of cases and the number of variables. The value for MXLOOPS should be set to the greater of these two.

Example:

Mxloops = 26760 /

# Troubleshooting and Bug Report

To report bugs, or troubleshoot the program, please send details in an email to [egonzalez@ets.org](mailto:egonzalez@ets.org) or [eugene.gonzalez@iea-hamburg.de](mailto:eugene.gonzalez@iea-hamburg.de). When troubleshooting, please set VIEWCOD = Y / and submit process the syntax. Accompany your email with the corresponding \*.sps and \*.sps files, as well as any additional information you might consider relevant.

# Examples