SAS/STAT® 14.2 User's Guide
The HPREG Procedure

This document is an individual chapter from *SAS/STAT® 14.2 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS/STAT® 14.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.2 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2016

Chapter 61
The HPREG Procedure

Contents

Overview: HPREG Procedure

The HPREG procedure is a high-performance procedure that fits and performs model selection for ordinary linear least squares models. The models supported are standard independently and identically distributed general linear models, which can contain main effects that consist of both continuous and classification variables and interaction effects of these variables. The procedure offers extensive capabilities for customizing the model selection with a wide variety of selection and stopping criteria, from traditional and computationally efficient significance-level-based criteria to more computationally intensive validation-based criteria. PROC HPREG also provides a variety of regression diagnostics that are conditional on the selected model.

PROC HPREG runs in either single-machine mode or distributed mode.

NOTE: Distributed mode requires SAS High-Performance Statistics.

PROC HPREG Features

The main features of the HPREG procedure are as follows:

- **Model specification**
  - supports GLM and reference parameterization for classification effects
  - supports any degree of interaction (crossed effects) and nested effects
  - supports hierarchy among effects
  - supports partitioning of data into training, validation, and testing roles
  - supports a FREQ statement for grouped analysis
  - supports a WEIGHT statement for weighted analysis
- **Selection control**
  - provides multiple effect-selection methods
  - enables selection from a very large number of effects (tens of thousands)
  - offers selection of individual levels of classification effects
  - provides effect selection based on a variety of selection criteria
  - provides stopping rules based on a variety of model evaluation criteria
  - supports stopping and selection rules based on external validation and leave-one-out cross validation
- **Display and output**
  - produces output data sets that contain predicted values, residuals, studentized residuals, confidence limits, and influence statistics

The HPREG procedure supports the following effect selection methods. For a more detailed description of these methods, see the section "Methods" (Chapter 3, *SAS/STAT User's Guide: High-Performance Procedures*).

- Forward selection starts with no effects in the model and adds effects.
- Backward elimination starts with all effects in the model and deletes effects.
- Stepwise regression is similar to forward selection except that effects already in the model do not necessarily stay there.
- Forward-swap selection is a modification of forward selection. Before any addition step, PROC HPREG makes all pairwise swaps of effects in and out of the current model that improve the selection criterion. When the selection criterion is R square, this method coincides with the MAXR method in the REG procedure in SAS/STAT software.
- Least angle regression, like forward selection, starts with no effects in the model and adds effects. The parameter estimates at any step are "shrunk" when compared to the corresponding least squares estimates.
- Lasso adds and deletes parameters based on a version of ordinary least squares in which the sum of the absolute regression coefficients is constrained. PROC HPREG also supports adaptive lasso selection where weights are applied to each of the parameters in forming the lasso constraint.

Hybrid versions of LAR and LASSO methods are also supported. They use LAR or LASSO to select the model, but then estimate the regression coefficients by ordinary weighted least squares.

Because the HPREG procedure is a high-performance analytical procedure, it also does the following:

- enables you to run in distributed mode on a cluster of machines that distribute the data and the computations
- enables you to run in single-machine mode on the server where SAS is installed
- exploits all the available cores and concurrent threads, regardless of execution mode

For more information, see the section "Processing Modes" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*).

PROC HPREG Contrasted with Other SAS Procedures

For general contrasts between SAS high-performance statistical procedures and other SAS procedures, see the section "Common Features of SAS High-Performance Statistical Procedures" (Chapter 3, *SAS/STAT User's Guide: High-Performance Procedures*). The following remarks contrast the HPREG procedure with the GLMSELECT, GLM, and REG procedures in SAS/STAT software.

A major functional difference between the HPREG and REG procedures is that the HPREG procedure enables you to specify general linear models that include classification variables. In this respect it is similar to the GLM and GLMSELECT procedures. In terms of the supported model selection methods, the HPREG procedure most resembles the GLMSELECT procedure. Like the GLMSELECT procedure but different from the REG procedure, the HPREG procedure supports the LAR and LASSO methods, the ability to use external validation data and cross validation as selection criteria, and extensive options to customize the selection process. The HPREG procedure does not support the MAXR and MINR methods that are available in the REG procedure. Nor does the HPREG procedure include any support for the all-subset-based methods that you can find in the REG procedure.

The CLASS statement in the HPREG procedure permits two parameterizations: the GLM-type parameterization and a reference parameterization. In contrast to the GLMSELECT, GENMOD, LOGISTIC, and other procedures that permit multiple parameterizations, the HPREG procedure does not mix parameterizations across the variables in the CLASS statement. In other words, all classification variables are in the same parameterization, and this parameterization is either the GLM or reference parameterization.

Like the REG procedure but different from the GLMSELECT procedure, the HPREG procedure does not perform model selection by default. If you request model selection by using the SELECTION statement then the default selection method is stepwise selection based on the SBC criterion. This default matches the default method used in PROC GLMSELECT.

As with the REG procedure but not supported with the GLMSELECT procedure, you can request observation-wise residual and influence diagnostics in the OUTPUT statement and variance inflation and tolerance statistics for the parameter estimates. If the fitted model has been obtained by performing model selection, then these statistics are conditional on the selected model and do not take the variability introduced by the selection process into account.

Getting Started: HPREG Procedure

The following example is closely modeled on the example in the section "Getting Started: GLMSELECT Procedure" in the *SAS/STAT User's Guide*.

The Sashelp.Baseball data set contains salary and performance information for Major League Baseball players who played at least one game in both the 1986 and 1987 seasons, excluding pitchers. The salaries (*Sports Illustrated,* April 20, 1987) are for the 1987 season and the performance measures are from 1986 (Collier Books, *The 1987 Baseball Encyclopedia Update*). The following step displays in Figure 61.1 the variables in the data set:

```
proc contents varnum data=sashelp.baseball;
   ods select position;
run;
```

Figure 61.1 Sashelp.Baseball Data Set

The CONTENTS Procedure

Variables in Creation Order

#	Variable	Type	Len	Label
1	Name	Char	18	Player's Name
2	Team	Char	14	Team at the End of 1986
3	nAtBat	Num	8	Times at Bat in 1986
4	nHits	Num	8	Hits in 1986
5	nHome	Num	8	Home Runs in 1986
6	nRuns	Num	8	Runs in 1986
7	nRBI	Num	8	RBIs in 1986
8	nBB	Num	8	Walks in 1986
9	YrMajor	Num	8	Years in the Major Leagues
10	CrAtBat	Num	8	Career Times at Bat
11	CrHits	Num	8	Career Hits
12	CrHome	Num	8	Career Home Runs
13	CrRuns	Num	8	Career Runs
14	CrRbi	Num	8	Career RBIs
15	CrBB	Num	8	Career Walks
16	League	Char	8	League at the End of 1986
17	Division	Char	8	Division at the End of 1986
18	Position	Char	8	Position(s) in 1986
19	nOuts	Num	8	Put Outs in 1986
20	nAssts	Num	8	Assists in 1986
21	nError	Num	8	Errors in 1986
22	Salary	Num	8	1987 Salary in $ Thousands
23	Div	Char	16	League and Division
24	logSalary	Num	8	Log Salary

Suppose you want to investigate whether you can model the players' salaries for the 1987 season based on performance measures for the previous season. The aim is to obtain a parsimonious model that does not overfit this particular data, making it useful for prediction. This example shows how you can use PROC HPREG as a starting point for such an analysis. Since the variation of salaries is much greater for the higher salaries, it is appropriate to apply a log transformation to the salaries before doing the model selection.

The following statements select a model with the default settings for stepwise selection:

```
proc hpreg data=sashelp.baseball;
  class league division;
  model logSalary = nAtBat nHits nHome nRuns nRBI nBB
                    yrMajor crAtBat crHits crHome crRuns crRbi
                    crBB league division nOuts nAssts nError;
  selection method=stepwise;
run;
```

The default output from this analysis is presented in Figure 61.2 through Figure 61.6.

Figure 61.2 Performance, Data Access, Model, and Selection Information

The HPREG Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	4

Data Access Information			
Data	**Engine**	**Role**	**Path**
SASHELP.BASEBALL	V9	Input	On Client

Model Information	
Data Source	SASHELP.BASEBALL
Dependent Variable	logSalary
Class Parameterization	GLM

Selection Information	
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None
Stop Horizon	3

Figure 61.2 displays the "Performance Information," "Data Access Information," "Model Information," and "Selection Information" tables. The "Performance Information" table shows that procedure executes in single-machine mode—that is, the model is fit on the machine where the SAS session executes. This run of the HPREG procedure was performed on a multicore machine with four CPUs; one computational thread was spawned per CPU.

The "Data Access Information" table shows that the input data set is accessed with the V9 (base) engine on the client machine.

The "Model Information" table identifies the data source and response and shows that the CLASS variables are parameterized in the GLM parameterization, which is the default.

The "Selection Information" provides details about the method and criteria used to perform the model selection. The requested selection method is a variant of the traditional stepwise selection where the decisions about what effects to add or drop at any step and when to terminate the selection are both based on the Schwarz Bayesian information criterion (SBC). The effect in the current model whose removal yields the maximal decrease in the SBC statistic is dropped provided this lowers the SBC value. When no further decrease in the SBC value can be obtained by dropping an effect in the model, the effect whose addition to the model yields the lowest SBC statistic is added and the whole process is repeated. The method terminates when dropping or adding any effect increases the SBC statistic.

Figure 61.3 displays the "Number of Observations," "Class Levels," and "Dimensions" tables. The "Number of Observations" table shows that of the 322 observations in the input data, only 263 observations are used in the analysis because there are observations with incomplete data. The "Class Level Information" table lists the levels of the classification variables "division" and "league." When you specify effects that contain classification variables, the number of parameters is usually larger than the number of effects. The "Dimensions" table shows the number of effects and the number of parameters considered.

Figure 61.3 Number of Observations, Class Levels, and Dimensions

Number of Observations Read	322
Number of Observations Used	263

Class Level Information

Class	Levels	Values
League	2	American National
Division	2	East West

Dimensions

Number of Effects	19
Number of Parameters	21

The "Stepwise Selection Summary" table in Figure 61.4 shows the effect that was added or dropped at each step of the selection process together with fit statistics for the model at each step. In this case, both selection and stopping are based on the SBC statistic.

Figure 61.4 Selection Summary Table

The HPREG Procedure

Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	**Intercept**		1	-57.2041
1	**CrRuns**		2	-194.3166
2	**nHits**		3	-252.5794
3	**YrMajor**		4	-262.7322
4		**CrRuns**	3	-262.8353
5	**nBB**		4	-269.7804*

*** Optimal Value of Criterion**

Figure 61.5 displays the "Stop Reason," "Selection Reason," and "Selected Effects" tables. Note that these tables are displayed without any titles. The "Stop Reason" table indicates that selection stopped because adding or removing any effect would worsen the SBC value that is used as the selection criterion. In this case, because no CHOOSE= criterion is specified in the SELECTION statement, the final model is the selected model; this is indicated in the "Selection Reason" table. The "Selected Effects" table lists the effects in the selected model.

Figure 61.5 Stopping and Selection Reasons

Stepwise selection stopped because adding or removing an effect does not improve the SBC criterion.

The model at step 5 is selected.

Selected Effects:	Intercept nHits nBB YrMajor

The "Analysis of Variance," "Fit Statistics," and "Parameter Estimates" tables shown in Figure 61.6 give details of the selected model.

Figure 61.6 Details of the Selected Model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	120.52553	40.17518	120.12	<.0001
Error	259	86.62820	0.33447		
Corrected Total	262	207.15373			

Root MSE	0.57834
R-Square	0.58182
Adj R-Sq	0.57697
AIC	-19.06903
AICC	-18.83557
SBC	-269.78041
ASE	0.32938

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.013911	0.111290	36.07	<.0001
nHits	1	0.007929	0.000994	7.98	<.0001
nBB	1	0.007280	0.002049	3.55	0.0005
YrMajor	1	0.100663	0.007551	13.33	<.0001

You might want to examine regression diagnostics for the selected model to investigate whether collinearity among the selected parameters or the presence of outlying or high leverage observations might be impacting the fit produced. The following statements include some options and statements to obtain these diagnostics:

```
proc hpreg data=sashelp.baseball;
  id name;
  class league division;
  model logSalary = nAtBat nHits nHome nRuns nRBI nBB
                    yrMajor crAtBat crHits crHome crRuns crRbi
                    crBB league division nOuts nAssts nError / vif clb;
  selection method=stepwise;
  output out=baseballOut p=predictedLogSalary r h cookd rstudent;
run;
```

The VIF and CLB options in the MODEL statement request variance inflation factors and 95% confidence limits for the parameter estimates. Figure 61.7 shows the "Parameter Estimates" with these requested statistics. The variance inflation factors (VIF) measure the inflation in the variances of the parameter estimates due to collinearities that exist among the regressor (independent) variables. Although there are no formal criteria for deciding whether a VIF is large enough to affect the predicted values, the VIF values for the selected effects in this example are small enough to indicate that there are no collinearity issues among the selected regressors.

Figure 61.7 Parameter Estimates with Additional Statistics

The HPREG Procedure

Selected Model

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	4.013911	0.111290	36.07	<.0001	0	3.79476	4.23306
nHits	1	0.007929	0.000994	7.98	<.0001	1.49642	0.00597	0.00989
nBB	1	0.007280	0.002049	3.55	0.0005	1.52109	0.00325	0.01131
YrMajor	1	0.100663	0.007551	13.33	<.0001	1.02488	0.08579	0.11553

By default, high-performance statistical procedures do not include all variables from the input data set in output data sets. The ID statement specifies that the variable name in the input data set be added as an identification variable in the baseballOut data set that is produced by the OUTPUT statement. In addition to this variable, the OUTPUT statement requests that predicted values, raw residuals, leverage values, Cook's D statistics, and studentized residuals be added in the output data set. Note that default names are used for these statistics except for the predicted values for which a specified name, predictedLogSalary, is supplied. The following statements use PROC PRINT to display the first five observations of this output data set:

```
proc print data=baseballOut(obs=5);
run;
```

Figure 61.8 First 5 Observations of the baseballOut Data Set

Obs	Name	predictedLogSalary	Residual	H	COOKD	RSTUDENT
1	Allanson, Andy	4.73980	.	0.016087	.	.
2	Ashby, Alan	6.34935	-0.18603	0.012645	.000335535	-0.32316
3	Davis, Alan	5.89993	0.27385	0.019909	.001161794	0.47759
4	Dawson, Andre	6.50852	-0.29392	0.011060	.000730178	-0.51031
5	Galarraga, Andres	5.12344	-0.60711	0.009684	.002720358	-1.05510

Syntax: HPREG Procedure

The following statements are available in the HPREG procedure:

```
PROC HPREG < options > ;
  BY variables ;
  CODE < options > ;
  CLASS variable < (options) >. . . < variable < (options) > > < / global-options > ;
  MODEL dependent = < effects > < / model-options > ;
  OUTPUT < OUT=SAS-data-set >
         < keyword < =name > >. . .
         < keyword < =name > > < / options > ;
  PARTITION < partition-options > ;
  PERFORMANCE < performance-options > ;
  SELECTION options ;
  FREQ variable ;
  ID variables ;
  WEIGHT variable ;
```

The PROC HPREG statement and a single MODEL statement are required. All other statements are optional. The CLASS statement can appear multiple times. If a CLASS statement is specified, it must precede the MODEL statement.

PROC HPREG Statement

```
PROC HPREG < options > ;
```

The PROC HPREG statement invokes the procedure. Table 61.1 summarizes the options in the PROC HPREG statement by function.

Table 61.1 PROC HPREG Statement Options

Option	Description
Basic Options	
DATA=	Specifies the input data set
NAMELEN=	Limits the length of effect names
Options Related to Output	
NOPRINT	Suppresses ODS output
NOCLPRINT	Limits or suppresses the display of class levels
User-Defined Formats	
FMTLIBXML=	Specifies a file reference for a format stream
Other Options	
ALPHA=	Sets the significance level used for the construction of confidence intervals

Table 61.1 *continued*

Option	Description
SEED=	Sets the seed used for pseudorandom number generation

Following are explanations of the *options* that you can specify in the PROC HPREG statement (in alphabetical order):

ALPHA=*number*
sets the significance level used for the construction of confidence intervals. The value must be between 0 and 1; the default value of 0.05 results in 95% intervals. This option affects the OUTPUT statement keywords LCL, LCLM, UCL, and UCLM, and the CLB option in the MODEL statement.

DATA=*SAS-data-set*
names the input SAS data set to be used by PROC HPREG. The default is the most recently created data set.

If the procedure executes in distributed mode, the input data are distributed to memory on the appliance nodes and analyzed in parallel, unless the data are already distributed in the appliance database. In that case the procedure reads the data alongside the distributed database. See the section "Processing Modes" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*) about the various execution modes and the section "Alongside-the-Database Execution" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*) about the alongside-the-database model.

FMTLIBXML=*file-ref*
specifies the file reference for the XML stream that contains the user-defined format definitions. User-defined formats are handled differently in a distributed computing environment than they are in other SAS products. See the section "Working with Formats" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*) for details about how to generate a XML stream for your formats.

NAMELEN=*number*
specifies the length to which long effect names are shortened. The default and minimum value is 20.

NOCLPRINT<=*number***>**
suppresses the display of the "Class Level Information" table if you do not specify *number*. If you specify *number*, the values of the classification variables are displayed for only those variables whose number of levels is less than *number*. Specifying a *number* helps to reduce the size of the "Class Level Information" table if some classification variables have a large number of levels.

NOPRINT
suppresses the generation of ODS output.

SEED=*number*
specifies an integer used to start the pseudorandom number generator for random partitioning of data for training, testing, and validation. If you do not specify a seed, or if you specify a value less than or equal to 0, the seed is generated from reading the time of day from the computer's clock.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC HPREG to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the HPREG procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLASS Statement

CLASS *variable* < **(***options***)** >. . . < *variable* < **(***options***)** > > < / *global-options* > **;**

The CLASS statement names the classification variables to be used as explanatory variables in the analysis. The CLASS statement must precede the MODEL statement.

The CLASS statement for SAS high-performance statistical procedures is documented in the section "CLASS Statement" (Chapter 3, *SAS/STAT User's Guide: High-Performance Procedures*). The HPREG procedure also supports the following *global-option* in the CLASS statement:

UPCASE
uppercases the values of character-valued CLASS variables before levelizing them. For example, if the UPCASE option is in effect and a CLASS variable can take the values 'a', 'A', and 'b', then 'a' and 'A' represent the same level and the CLASS variable is treated as having only two values: 'A' and 'B'.

CODE Statement

CODE < *options* > **;**

The CODE statement writes SAS DATA step code for computing predicted values of the fitted model either to a file or to a catalog entry. This code can then be included in a DATA step to score new data.

Table 61.2 summarizes the *options* available in the CODE statement.

Table 61.2 CODE Statement Options

Option	Description
CATALOG=	Names the catalog entry where the generated code is saved
DUMMIES	Retains the dummy variables in the data set
ERROR	Computes the error function
FILE=	Names the file where the generated code is saved
FORMAT=	Specifies the numeric format for the regression coefficients
GROUP=	Specifies the group identifier for array names and statement labels
IMPUTE	Imputes predicted values for observations with missing or invalid covariates
LINESIZE=	Specifies the line size of the generated code
LOOKUP=	Specifies the algorithm for looking up CLASS levels
RESIDUAL	Computes residuals

For details about the syntax of the CODE statement, see the section "CODE Statement" on page 393 in Chapter 19, "Shared Concepts and Topics."

FREQ Statement

FREQ *variable* **;**

The *variable* in the FREQ statement identifies a numeric variable in the data set that contains the frequency of occurrence for each observation. SAS high-performance statistical procedures that support the FREQ statement treat each observation as if it appeared f times, where f is the value of the FREQ variable for the observation. If the frequency value is not an integer, it is truncated to an integer. If the frequency value is less than 1 or missing, the observation is not used in the analysis. When the FREQ statement is not specified, each observation is assigned a frequency of 1.

ID Statement

ID *variables* **;**

The ID statement lists one or more variables from the input data set that are transferred to output data sets created by SAS high-performance statistical procedures, provided that the output data set produces one (or more) records per input observation.

For documentation on the common ID statement in SAS high-performance statistical procedures, see the section "ID Statement" (Chapter 3, *SAS/STAT User's Guide: High-Performance Procedures*).

MODEL Statement

MODEL *dependent*=< *effects* > / < *options* > **;**

The MODEL statement names the dependent variable and the explanatory effects, including covariates, main effects, interactions, and nested effects. If you omit the explanatory effects, the procedure fits an intercept-only model.

After the keyword MODEL, the dependent (response) variable is specified, followed by an equal sign. The explanatory effects follow the equal sign. For information about constructing the model effects, see the section "Specification and Parameterization of Model Effects" (Chapter 3, *SAS/STAT User's Guide: High-Performance Procedures*).

You can specify the following *options* in the MODEL statement after a slash (/):

CLB

requests the $100(1-\alpha)\%$ upper and lower confidence limits for the parameter estimates. By default, the 95% limits are computed; the ALPHA= option in the PROC HPREG statement can be used to change the α level. The CLB option is not supported when you request METHOD=LAR or METHOD=LASSO in the SELECTION statement.

INCLUDE=*n*

INCLUDE=*single-effect*

INCLUDE=(*effects***)**

forces effects to be included in all models. If you specify INCLUDE=*n*, then the first *n* effects listed in the MODEL statement are included in all models. If you specify INCLUDE=*single-effect* or if you specify a list of effects within parentheses, then the specified effects are forced into all models. The effects that you specify in the INCLUDE= option must be explanatory effects defined in the MODEL statement before the slash (/). The INCLUDE= option is not available when you specify METHOD=LAR or METHOD=LASSO in the SELECTION statement.

NOINT

suppresses the intercept term that is otherwise included in the model.

ORDERSELECT

specifies that, for the selected model, effects be displayed in the order in which they first entered the model. If you do not specify the ORDERSELECT option, then effects in the selected model are displayed in the order in which they appear in the MODEL statement.

START=*n*

START=*single-effect*

START=(*effects***)**

is used to begin the selection process in the FORWARD, FORWARDSWAP, and STEPWISE selection methods from the initial model that you designate. If you specify START=*n*, then the starting model consists of the first *n* effects listed in the MODEL statement. If you specify START=*single-effect* or if you specify a list of effects within parentheses, then the starting model consists of these specified effects. The effects that you specify in the START= option must be explanatory effects defined in the MODEL statement before the slash (/). The START= option is not available when you specify METHOD=BACKWARD, METHOD=LAR, or METHOD=LASSO in the SELECTION statement.

STB

produces standardized regression coefficients. A standardized regression coefficient is computed by dividing a parameter estimate by the ratio of the sample standard deviation of the dependent variable to the sample standard deviation of the regressor.

TOL

produces tolerance values for the estimates. Tolerance for a parameter is defined as $1 - R^2$, where R^2 is obtained from the regression of the parameter on all other parameters in the model. The TOL option is not supported when you request METHOD=LAR or METHOD=LASSO in the SELECTION statement.

VIF

produces variance inflation factors with the parameter estimates. Variance inflation is the reciprocal of tolerance. The VIF option is not supported when you request METHOD=LAR or METHOD=LASSO in the SELECTION statement.

OUTPUT Statement

OUTPUT < **OUT=***SAS-data-set* >
< **COPYVARS=***(variables)* >
< *keyword* < *=name* > >. . . < *keyword* < *=name* > > **;**

The OUTPUT statement creates a data set that contains observationwise statistics, which are computed after fitting the model. The variables in the input data set are *not* included in the output data set to avoid data duplication for large data sets; however, variables specified in the ID statement or COPYVARS= option are included.

If the input data are in distributed form, where access of data in a particular order cannot be guaranteed, the HPREG procedure copies the distribution or partition key to the output data set so that its contents can be joined with the input data.

The output statistics are computed based on the parameter estimates for the selected model.

You can specify the following syntax elements in the OUTPUT statement:

OUT=*SAS-data-set*

DATA=*SAS-data-set*

specifies the name of the output data set. If the OUT= (or DATA=) option is omitted, the procedure uses the DATA*n* convention to name the output data set.

COPYVAR=*variable*

COPYVARS=(*variables***)**

transfers one or more *variables* from the input data set to the output data set. Variables named in an ID statement are also copied from the input data set to the output data set.

keyword **< =***name* **>**

specifies the statistics to include in the output data set and optionally names the new variables that contain the statistics. Specify a keyword for each desired statistic (see the following list of keywords), followed optionally by an equal sign and a variable to contain the statistic.

If you specify *keyword=name*, the new variable that contains the requested statistic has the specified name. If you omit the optional *=name* after a *keyword*, then a default name is used.

The following are valid values for *keyword* to request statistics that are available with all selection methods:

PREDICTED

PRED

P

requests predicted values for the response variable. The default name is Pred.

RESIDUAL

RESID

R

requests the residual, calculated as ACTUAL–PREDICTED. The default name is Residual.

ROLE

requests a numeric variable that indicates the role played by each observation in fitting the model. The default name is _ROLE_. For each observation the interpretation of this variable is shown in Table 61.3:

Table 61.3 Role Interpretation

Value	Observation Role
0	Not used
1	Training
2	Validation
3	Testing

If you do not partition the input data by using a PARTITION statement, then the role variable value is 1 for observations used in fitting the model, and 0 for observations that have at least one missing or invalid value for the response, regressors, frequency or weight variables.

In addition to the preceding statistics, you can also use the *keywords* listed in Table 61.4 in the OUTPUT statement to obtain additional statistics. These statistics are not available if you use METHOD=LAR or METHOD=LASSO in the SELECTION statement, unless you also specify the LSCOEFFS option. See the section "Diagnostic Statistics" on page 4624 for computational formulas. All the statistics available in the OUTPUT statement are conditional on the selected model and do not take into account the variability introduced by doing model selection.

Table 61.4 Keywords for OUTPUT Statement

Keyword	Description
COOKD	Cook's D influence statistic
COVRATIO	Standard influence of observation on covariance of betas
DFFIT	Standard influence of observation on predicted value
H	Leverage, $\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_i'$
LCL	Lower bound of a $100(1-\alpha)\%$ confidence interval for an individual prediction. This includes the variance of the error, as well as the variance of the parameter estimates.
LCLM	Lower bound of a $100(1-\alpha)\%$ confidence interval for the expected value (mean) of the dependent variable

Table 61.4 *continued*

Keyword	Description
PRESS	ith residual divided by $(1-h)$, where h is the leverage, and where the model has been refit without the ith observation
RSTUDENT	A studentized residual with the current observation deleted
STDI	Standard error of the individual predicted value
STDP	Standard error of the mean predicted value
STDR	Standard error of the residual
STUDENT	Studentized residuals, which are the residuals divided by their standard errors
UCL	Upper bound of a $100(1-\alpha)\%$ confidence interval for an individual prediction
UCLM	Upper bound of a $100(1-\alpha)\%$ confidence interval for the expected value (mean) of the dependent variable

PARTITION Statement

PARTITION < *partition-options* > **;**

The PARTITION statement specifies how observations in the input data set are logically partitioned into disjoint subsets for model training, validation, and testing. Either you can designate a variable in the input data set and a set of formatted values of that variable to determine the role of each observation, or you can specify proportions to use for random assignment of observations for each role.

The following mutually exclusive *partition-options* are available:

ROLEVAR | ROLE=*variable***(< TEST='***value***' > < TRAIN='***value***' > < VALIDATE='***value***' >)**
names the variable in the input data set whose values are used to assign roles to each observation. The formatted values of this variable that are used to assign observations roles are specified in the TEST=, TRAIN=, and VALIDATE= suboptions. If you do not specify the TRAIN= suboption, then all observations whose role is not determined by the TEST= or VALIDATE= suboptions are assigned to training.

FRACTION(< TEST=*fraction* > < VALIDATE=*fraction* >)**
requests that specified proportions of the observations in the input data set be randomly assigned training and validation roles. You specify the proportions for testing and validation by using the TEST= and VALIDATE= suboptions. If you specify both the TEST= and the VALIDATE= suboptions, then the sum of the specified fractions must be less than 1 and the remaining fraction of the observations are assigned to the training role.

PERFORMANCE Statement

PERFORMANCE < *performance-options* > **;**

The PERFORMANCE statement defines performance parameters for multithreaded and distributed computing, passes variables that describe the distributed computing environment, and requests detailed results about the performance characteristics of the HPREG procedure.

You can also use the PERFORMANCE statement to control whether the HPREG procedure executes in single-machine mode or distributed mode.

The PERFORMANCE statement is documented further in the section "PERFORMANCE Statement" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*).

SELECTION Statement

SELECTION < *options* > **;**

The SELECTION statement performs variable selection. All *options* except the SCREEN option are fully documented in the section "SELECTION Statement" (Chapter 3, *SAS/STAT User's Guide: High-Performance Procedures*). The SCREEN option is described in the following section. The remainder of this section describes specific information about how PROC HPREG implements the METHOD= option and the DETAILS= option.

The HPREG procedure supports the following values of the METHOD= option in the SELECTION statement:

NONE specifies no model selection.

FORWARD specifies the forward selection method, which starts with no effects in the model and adds effects.

BACKWARD specifies the backward elimination method, which starts with all effects in the model and deletes effects.

STEPWISE specifies the stepwise regression method, which is similar to the forward selection method except that effects already in the model do not necessarily stay there.

FORWARDSWAP specifies the forward-swap selection method, which is an extension of the forward selection method. Before any addition step, PROC HPREG makes all pairwise swaps of effects in and out of the current model that improve the selection criterion. When the selection criterion is R square, this method is the same as the MAXR method in the REG procedure in SAS/STAT software.

LAR specifies the least angle regression method. Like forward selection, this method starts with no effects in the model and adds effects. The parameter estimates at any step are "shrunk" when compared to the corresponding least squares estimates. If the model contains classification variables, then these classification variables are split. For more information, see the SPLIT option in the CLASS statement.

LASSO specifies the lasso method, which adds and deletes parameters based on a version of ordinary least squares in which the sum of the absolute regression coefficients is constrained. If the model contains classification variables, then these classification variables are split. For more information, see the SPLIT option in the CLASS statement.

The DETAILS=ALL and DETAILS=STEPS options produce the "ANOVA," "Fit Statistics," and "Parameter Estimates" tables, which provide information about the model that is selected at each step of the selection process.

In addition to other options, which are fully documented in the section "SELECTION Statement" (Chapter 3, *SAS/STAT User's Guide: High-Performance Procedures*), PROC HPREG also supports a SCREEN option, which has the following syntax:

SCREEN < (*global-screen-options***) > < =***screen-options* **>**

You can specify following *global-screen-options*:

DETAILS=NONE | SUMMARY | ALL
specifies the level of detail to be produced about the screening process. You can specify the following values:

NONE suppresses all tables that provide details of the screening process.

ALL produces the following output and shows model selection details at each stage of the screening process:

- a screening table that shows the correlations that are used to obtain the screened effects for the first two stages of the screening process
- a screened effects table that lists the effects that are chosen at each stage of the screening process

SUMMARY produces the following output and shows details about the model selection only for the final stage of the screening process:

- a screening table that shows the correlations that are used to obtain the screened effects for the first two stages of the screening process
- a screened effects table that lists the effects that are chosen at each stage of the screening process

By default, DETAILS=SUMMARY.

SINGLESTAGE
screens effects and selects a model only once.

MULTISTAGE
performs multiple stages, each of which contains a screening and a model selection step.

You can specify the following *screen-options* after an = sign:

SCREEN=*n1* **<** *n2* **>**
specifies the number of effects to be chosen at the first two stages of the screening process. If you specify only *n1*, then *n1* is used for both the first and second stages. If you specify both *n1* and *n2*, then *n1* is used at the first stage and *n2* is used at the second stage. At the first stage, effects are ranked in decreasing order of the magnitude of their pairwise correlations with the response, and the first *n1* effects are used in the selection process at that stage. At the second stage, effects are ranked in decreasing order of the magnitude of their pairwise correlations with the residuals obtained at the first stage, and the first *n2* effects are used in the selection process at that stage.

SCREEN=PERCENT(*p1* < *p2* >)
specifies the percentage of effects in the MODEL statement to be chosen at the first two stages of the screening process. If you specify only *p1*, then *p1* is used for both the first and second stages. If you specify *p1* and *p2*, then *p1* is used at the first stage and *p2* is used at the second stage.

SCREEN=CUTOFF(*c1* < *c2* >)
specifies the minimum value of the screening statistic that effects must have in order to be chosen at the first two stages of the screening process. If you specify only *c1*, then *c1* is used for both the first and second stages. If you specify both *c1* and *c2*, then *c1* is used at the first stage and *c2* is used at the second stage. At the first stage, any effect whose absolute pairwise correlation with the response is less than the first-stage cutoff is not used in the selection process at that stage. At the second stage, any effect whose absolute pairwise correlation with the residuals obtained from the first stage is less than the second-stage cutoff is not used in the selection process at that stage.

If you do not specify any *screen-options*, SCREEN=PERCENT(10) by default.

For a classification effect that has multiple degrees of freedom, pairwise correlations with the response at the first stage and the first stage residuals at the second stage are computed separately for each dummy variable that corresponds to the levels of the classification variables in the effect. The largest magnitude of these correlations is used as a proxy for the correlation statistic for that effect.

WEIGHT Statement

WEIGHT *variable* **;**

The *variable* in the WEIGHT statement is used as a weight to perform a weighted analysis of the data. Observations with nonpositive or missing weights are not included in the analysis. If a WEIGHT statement is not included, all observations used in the analysis are assigned a weight of 1.

Details: HPREG Procedure

Criteria Used in Model Selection

The HPREG procedure supports a variety of fit statistics that you can specify as criteria for the CHOOSE=, SELECT=, and STOP= options in the SELECTION statement. The following statistics are available:

ADJRSQ	Adjusted R-square statistic (Darlington 1968; Judge et al. 1985)
AIC	Akaike's information criterion (Akaike 1969; Judge et al. 1985)
AICC	Corrected Akaike's information criterion (Hurvich and Tsai 1989)
BIC | SBC	Schwarz Bayesian information criterion (Schwarz 1978; Judge et al. 1985)
CP	Mallows C_p statistic (Mallows 1973; Hocking 1976)
PRESS	Predicted residual sum of squares statistic
RSQUARE	R-square statistic (Darlington 1968; Judge et al. 1985)

SL	Significance used to assess an effect's contribution to the fit when it is added to or removed from a model
VALIDATE	Average square error over the validation data

When you use SL as a criterion for effect selection, the definition depends on whether an effect is being considered as a drop or an add candidate. If the current model has p parameters excluding the intercept, and if you denote its residual sum of squares by RSS_p and you add an effect with k degrees of freedom and denote the residual sum of squares of the resulting model by RSS_{p+k}, then the F statistic for entry with k numerator degrees of freedom and $n-(p+k)-1$ denominator degrees of freedom is given by

$$F = \frac{(\mathrm{RSS}_p - \mathrm{RSS}_{p+k})/k}{\mathrm{RSS}_{p+k}/(n-(p+k)-1)}$$

where n is number of observations used in the analysis. The significance level for entry is the p-value of this F statistic, and is deemed significant if it is smaller than the SLENTRY limit. Among several such add candidates, the effect with the smallest p-value (most significant) is deemed best.

If you drop an effect with k degrees of freedom and denote the residual sum of squares of the resulting model by RSS_{p-k}, then the F statistic for removal with k numerator degrees of freedom and $n-p-k$ denominator degrees of freedom is given by

$$F = \frac{(\mathrm{RSS}_{p-k} - \mathrm{RSS}_p)/k}{\mathrm{RSS}_p/(n-p-k)}$$

where n is number of observations used in the analysis. The significance level for removal is the p-value of this F statistic, and the effect is deemed not significant if this p-value is larger than the SLSTAY limit. Among several such removal candidates, the effect with the largest p-value (least significant) is deemed the best removal candidate.

It is known that the "F-to-enter" and "F-to-delete" statistics do not follow an F distribution (Draper, Guttman, and Kanemasu 1971).. Hence the SLENTRY and SLSTAY values cannot reliably be viewed as probabilities. One way to address this difficulty is to replace hypothesis testing as a means of selecting a model with information criteria or out-of-sample prediction criteria. While Harrell (2001) points out that information criteria were developed for comparing only prespecified models, Burnham and Anderson (2002) note that AIC criteria have routinely been used for several decades for performing model selection in time series analysis.

Table 61.5 provides formulas and definitions for these fit statistics.

Table 61.5 Formulas and Definitions for Model Fit Summary Statistics

Statistic	Definition or Formula
n	Number of observations
p	Number of parameters including the intercept
$\hat{\sigma}^2$	Estimate of pure error variance from fitting the full model
SST	Total sum of squares corrected for the mean for the dependent variable
SSE	Error sum of squares
ASE	$\frac{\mathrm{SSE}}{n}$

Table 61.5 *continued*

Statistic	Definition or Formula
MSE	$\frac{\text{SSE}}{n-p}$
R^2	$1-\frac{\text{SSE}}{\text{SST}}$
ADJRSQ	$1-\frac{(n-1)(1-R^2)}{n-p}$
AIC	$n\ln\left(\frac{\text{SSE}}{n}\right)+2p$
AICC	$1+\ln\left(\frac{\text{SSE}}{n}\right)+\frac{2(p+1)}{n-p-2}$
CP (C_p)	$\frac{\text{SSE}}{\hat{\sigma}^2}+2p-n$
PRESS	$\sum_{i=1}^{n}\frac{r_i^2}{(1-h_i)^2}$ where r_i = residual at observation i and h_i = leverage of observation $i=\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_i'$
RMSE	$\sqrt{\text{MSE}}$
SBC	$n\ln\left(\frac{\text{SSE}}{n}\right)+p\ln(n)$

Diagnostic Statistics

This section gathers the formulas for the statistics available in the OUTPUT statement. All the statistics available in the OUTPUT statement are conditional on the selected model and do not take into account the variability introduced by doing model selection.

The model to be fit is $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, and the parameter estimate is denoted by $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$. The subscript i denotes values for the ith observation, and the parenthetical subscript (i) means that the statistic is computed by using all observations except the ith observation.

The ALPHA= option in the PROC HPREG statement is used to set the α value for the confidence limit statistics.

Table 61.6 contains the diagnostic statistics and their formulas. Each statistic is computed for each observation.

Table 61.6 Formulas and Definitions for Diagnostic Statistics

MODEL Option or Statistic	Formula
PRED ($\widehat{\mathbf{Y}}_i$)	$\mathbf{X}_i\mathbf{b}$
RES (r_i)	$\mathbf{Y}_i-\widehat{\mathbf{Y}}_i$
H (h_i)	$\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_i'$
STDP	$\sqrt{h_i\widehat{\sigma}^2}$

Table 61.6 *continued*

MODEL Option or Statistic	**Formula**
STDI	$\sqrt{(1+h_i)\widehat{\sigma}^2}$
STDR	$\sqrt{(1-h_i)\widehat{\sigma}^2}$
LCL	$\widehat{Y}_i - t_{\frac{\alpha}{2}}\text{STDI}$
LCLM	$\widehat{Y}_i - t_{\frac{\alpha}{2}}\text{STDP}$
UCL	$\widehat{Y}_i + t_{\frac{\alpha}{2}}\text{STDI}$
UCLM	$\widehat{Y}_i + t_{\frac{\alpha}{2}}\text{STDP}$
STUDENT	$\frac{r_i}{\text{STDR}_i}$
RSTUDENT	$\frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$
COOKD	$\frac{1}{p}\text{STUDENT}^2\frac{\text{STDP}^2}{\text{STDR}^2}$
COVRATIO	$\frac{\det(\hat{\sigma}^2_{(i)}(\mathbf{x}'_{(i)}\mathbf{x}_{(i)})^{-1}}{\det(\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1})}$
DFFITS	$\frac{(\widehat{\mathbf{Y}}_i - \widehat{\mathbf{Y}}_{(i)})}{(\hat{\sigma}_{(i)}\sqrt{h_i})}$
PRESS(predr$_i$)	$\frac{r_i}{1-h_i}$

Classification Variables and the SPLIT Option

PROC HPREG supports the ability to split classification variables when doing model selection. You use the SPLIT option in the CLASS statement to specify that the columns of the design matrix that correspond to effects that contain a split classification variable can enter or leave a model independently of the other design columns of that effect. The following statements illustrate the use of SPLIT option:

```
data splitExample;
   length c2 $6;
   drop i;
   do i=1 to 1000;
     c1 = 1 + mod(i,6);
     if      i < 250 then c2 = 'low';
     else if i < 500 then c2 = 'medium';
     else                 c2 = 'high';
     x1 = ranuni(1);
     x2 = ranuni(1);
     y = x1+3*(c2 ='low')  + 10*(c1=3) +5*(c1=5) + rannor(1);
     output;
   end;
run;

proc hpreg data=splitExample;
```

```
   class c1(split) c2(order=data);
   model y = c1 c2 x1 x2/orderselect;
   selection method=forward;
run;
```

The "Class Levels" table shown in Figure 61.9 is produced by default whenever you specify a CLASS statement.

Figure 61.9 Class Levels

The HPREG Procedure

Class Level Information

Class	Levels	Values
c1	6	* 1 2 3 4 5 6
c2	3	low medium high

*** Associated Parameters Split**

The SPLIT option has been specified for the classification variable c1. This permits the parameters associated with the effect c1 to enter or leave the model individually. The "Parameter Estimates" table in Figure 61.10 shows that for this example the parameters that correspond to only levels 3 and 5 of c1 are in the selected model. Finally, note that the ORDERSELECT option in the MODEL statement specifies that the parameters be displayed in the order in which they first entered the model.

Figure 61.10 Parameter Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.308111	0.075387	-4.09	<.0001
c1_3	1	10.161702	0.087601	116.00	<.0001
c1_5	1	5.018407	0.087587	57.30	<.0001
c2 low	1	3.139941	0.078495	40.00	<.0001
c2 medium	1	0.221539	0.078364	2.83	0.0048
c2 high	0	0	.	.	.
x1	1	1.317420	0.109510	12.03	<.0001

Using Validation and Test Data

When you have sufficient data, you can subdivide your data into three parts called the training, validation, and test data. During the selection process, models are fit on the training data, and the prediction error for the models so obtained is found by using the validation data. This prediction error on the validation data can be used to decide when to terminate the selection process or to decide what effects to include as the selection process proceeds. Finally, after a selected model has been obtained, the test set can be used to assess how the selected model generalizes on data that played no role in selecting the model.

In some cases you might want to use only training and test data. For example, you might decide to use an information criterion to decide what effects to include and when to terminate the selection process. In this case no validation data are required, but test data can still be useful in assessing the predictive performance of the selected model. In other cases you might decide to use validation data during the selection process

but forgo assessing the selected model on test data. Hastie, Tibshirani, and Friedman (2001) note that it is difficult to give a general rule for how many observations you should assign to each role. They note that a typical split might be 50% for training and 25% each for validation and testing.

You use a PARTITION statement to logically subdivide the DATA= data set into separate roles. You can name the fractions of the data that you want to reserve as test data and validation data. For example, the following statements randomly subdivide the "inData" data set, reserving 50% for training and 25% each for validation and testing:

```
proc hpreg data=inData;
  partition fraction(test=0.25 validate=0.25);
  ...
run;
```

In some cases you might need to exercise more control over the partitioning of the input data set. You can do this by naming both a variable in the input data set and also a formatted value of that variable that correspond to each role. For example, the following statements assign roles to the observations in the "inData" data set based on the value of the variable group in that data set. Observations where the value of group is 'group 1' are assigned for testing, and those with value 'group 2' are assigned to training. All other observations are ignored.

```
proc hpreg data=inData;
  partition roleVar=group(test='group 1' train='group 2')
  ...
run;
```

When you have reserved observations for training, validation, and testing, a model fit on the training data is scored on the validation and test data, and the average squared error (ASE) is computed separately for each of these subsets. The ASE for each data role is the error sum of squares for observations in that role divided by the number of observations in that role.

Using the Validation ASE as the STOP= Criterion

If you have provided observations for validation, then you can specify STOP=VALIDATE as a suboption of the METHOD= option in the SELECTION statement. At step k of the selection process, the best candidate effect to enter or leave the current model is determined. Here "best candidate" means the effect that gives the best value of the SELECT= criterion; this criterion need not be based on the validation data. The validation ASE for the model with this candidate effect added or removed is computed. If this validation ASE is greater than the validation ASE for the model at step k, then the selection process terminates at step k.

Using the Validation ASE as the CHOOSE= Criterion

When you specify the CHOOSE=VALIDATE suboption of the METHOD= option in the SELECTION statement, the validation ASE is computed for the models at each step of the selection process. The smallest model at any step that yields the smallest validation ASE is selected.

Using the Validation ASE as the SELECT= Criterion

You request the validation ASE as the selection criterion by specifying the SELECT=VALIDATE suboption of the METHOD= option in the SELECTION statement. At step k of the selection process, the validation ASE is computed for each model in which a candidate for entry is added or candidate for removal is dropped.

The selected candidate for entry or removal is the one that yields a model with the minimal validation ASE. This method is computationally very expensive because validation statistics need to be computed for every candidate at every step; it should be used only with small data sets or models with a small number of regressors.

Computational Method

Multithreading

Threading refers to the organization of computational work into multiple tasks (processing units that can be scheduled by the operating system). A task is associated with a thread. Multithreading refers to the concurrent execution of threads. When multithreading is possible, substantial performance gains can be realized compared to sequential (single-threaded) execution.

The number of threads spawned by the HPREG procedure is determined by the number of CPUs on a machine and can be controlled in the following ways:

- You can specify the CPU count with the CPUCOUNT= SAS system option. For example, if you specify the following statements, the HPREG procedure schedules threads as if it executes on a system with four CPUs, regardless of the actual CPU count.

```
options cpucount=4;
```

- You can specify the NTHREADS= option in the PERFORMANCE statement to determine the number of threads. This specification overrides the system option. Specify NTHREADS=1 to force single-threaded execution.

The number of threads per machine is displayed in the "Performance Information" table, which is part of the default output. The HPREG procedure allocates one thread per CPU.

The tasks multithreaded by the HPREG procedures are primarily defined by dividing the data processed on a single machine among the threads—that is, the HPREG procedure implements multithreading through a data-parallel model. For example, if the input data set has 1,000 observations and you are running with four threads, then 250 observations are associated with each thread. All operations that require access to the data are then multithreaded. This operations include the following:

- variable levelization
- effect levelization
- formation of the crossproducts matrix
- evaluation of predicted residual sums of squares on validation and test data
- scoring of observations

In addition, operations on matrices such as sweeps might be multithreaded if the matrices are of sufficient size to realize performance benefits from managing multiple threads for the particular matrix operation.

Output Data Set

Many procedures in SAS software add the variables from the input data set when an observationwise output data set is created. The assumption of high-performance statistical procedures is that the input data sets can be large and contain many variables. For performance reasons, the output data set contains the following:

- those variables explicitly created by the statement
- variables listed in the ID statement
- distribution keys or hash keys that are transferred from the input data set

This enables you to add output data set information that is necessary for subsequent SQL joins without copying the entire input data set to the output data set. For more information about output data sets that are produced when PROC HPREG is run in distributed mode, see the section "Output Data Sets" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*).

Screening

Model selection from a very large number of effects is computationally demanding. For example, in analyzing microarray data, where each dot in the array corresponds to a regressor, having 35,000 such regressors is not uncommon. Another source of such large regression problems arises when you want to consider all possible two-way interactions of your main effects as candidates for inclusion in a selected model. See Foster and Stine (2004) for an example that uses this approach to build a predictive model for bankruptcy.

In recent years, there has been a resurgence of interest in combining variable selection methods with an initial screening step that reduces the large number of regressors to a much smaller subset from which the final model is chosen. You can find theoretical underpinnings of this approach in Fan and Lv (2008). See El Ghaoui, Viallon, and Rabbani (2012) and Tibshirani et al. (2012) for examples where screening has also been incorporated in the context of penalized regression methods (such as lasso) for performing model selection.

Screening uses a screening statistic that is inexpensive to compute in order to eliminate from consideration regressors that are unlikely to be selected if you included them in variable selection. For linear regression, you can use the magnitude of the correlation between each individual regressor and the response as such a screening statistic. The square of the correlation between a regressor that has one degree of freedom and the response is the R-square value for the univariate regression for the response with this regressor. Hence, screening by the magnitude of the pairwise correlations is equivalent to fitting univariate models to do the screening.

The first stage of the screening method chooses only the subset of regressors whose screening statistic is larger than a specified cutoff value or by choosing those regressors whose screening statistics are among a specified number or percentage of the largest screening statistic values. Then you perform model selection for the response from this screened subset of the original regressors.

One problem with this approach is that a regressor that is pairwise (marginally) uncorrelated or has very small correlation with the response can nevertheless be an important predictor, but it would be eliminated in the screening. You can address this problem by switching to a multistage approach. The first stage consists

of screening the regressors and selecting the model for the response from the screened subset. The second stage repeats the first stage except that you use the residuals from the first stage as the response variable in this second stage. You can iterate this process by using the residuals from the previous stage as the response for the next stage. The final stage forms the union of all the screened regressors from the first stage with all the selected regressors at the subsequent stages and selects a model for the original response variable from this union.

Experimentation has shown that there is little benefit in practice in using more than one stage where the response is the residual from the previous stage. Hence, PROC HPREG implements a three-stage process by default. However, if you specify the SINGLESTAGE suboption in the SCREEN option in the SELECTION statement, then only the first screening stage is performed.

Displayed Output

The following sections describe the output produced by PROC HPREG. The output is organized into various tables, which are discussed in the order of appearance.

Performance Information

The "Performance Information" table is produced by default. It displays information about the execution mode. For single-machine mode, the table displays the number of threads used. For distributed mode, the table displays the number of compute nodes, and the number of threads per node.

Data Access Information

The "Data Access Information" table is produced by default. For the input and output data sets, it displays the libref and data set name, the engine used to access the data, the role (input or output) of the data set, and path that data followed to the reach the computation.

Model Information

The "Model Information" table displays basic information about the model, such as the response variable, frequency variable, weight variable, and the type of parameterization used for classification variables named in the CLASS statement.

Selection Information

When you specify the SELECTION statement, the HPREG procedure produces by default a series of tables with information about the model selection. The "Selection Information" table informs you about the model selection method; select, stop, and choose criteria; and other parameters that govern the selection. You can suppress this table by specifying DETAILS=NONE in the SELECTION statement.

Screening Information

When you specify the SCREEN option in the SELECTION statement, the "Screening Information" table informs you about the number of screening stages used and informs you about the method and values that are used to determine how many screened effects are chosen at each screening stage.

Screening

When you specify the DETAILS=ALL suboption of the SCREEN option in the SELECTION statement, the "Screening" table displays the model effects and their screening statistic values in descending order of the screening statistic values.

Screened Effects

When you specify the SCREEN option in the SELECTION statement, the "Screened Effects" table displays a list of the screened model effects at each stage of the screening process.

Number of Observations

The "Number of Observations" table displays the number of observations read from the input data set and the number of observations used in the analysis. If you specify a FREQ statement, the table also displays the sum of frequencies read and used. If you use a PARTITION statement, the table also displays the number of observations used for each data role.

Class Level Information

The "Class Level Information" table lists the levels of every variable specified in the CLASS statement. You should check this information to make sure that the data are correct. You can adjust the order of the CLASS variable levels with the ORDER= option in the CLASS statement. You can suppress the "Class Level Information" table completely or partially with the NOCLPRINT= option in the PROC HPREG statement.

If the classification variables are in the reference parameterization, the "Class Level Information" table also displays the reference value for each variable. The "Class Level Information" table also indicates which, if any, of the classification variables are split by using the SPLIT option in the CLASS statement.

Dimensions

The "Dimensions" table displays information about the number of effects and the number of parameters from which the selected model is chosen. If you use split classification variables, then this table also includes the number of effects after splitting is taken into account.

Entry and Removal Candidates

When you specify the DETAILS=ALL or DETAILS=STEPS option in the SELECTION statement, the HPREG procedure produces "Entry Candidates" and "Removal Candidates" tables that display the effect names and values of the criterion used to select entering or departing effects at each step of the selection process. The effects are displayed in sorted order from best to worst of the selection criterion.

Selection Summary

When you specify the SELECTION statement, the HPREG procedure produces the "Selection Summary" table with information about the sequence of steps of the selection process. For each step, the effect that was entered or dropped is displayed along with the statistics used to select the effect, stop the selection, and choose the selected model. For all criteria that you can use for model selection, the steps at which the optimal values of these criteria occur are also indicated.

The display of the "Selection Summary" table can be suppressed by specifying DETAILS=NONE in the SELECTION statement.

Stop Reason

The "Stop Reason" table displays the reason why the selection stopped. To facilitate programmatic use of this table, an integer code is assigned to each reason and is included if you output this table by using an ODS OUTPUT statement. The reasons and their associated codes follow:

Code	Stop Reason
1	All eligible effects are in the model.
2	All eligible effects have been removed.
3	Specified maximum number of steps done.
4	The model contains the specified maximum number of effects.
5	The model contains the specified minimum number of effects (for backward selection).
6	The stopping criterion is at a local optimum.
7	No suitable add or drop candidate could be found.
8	Adding or dropping any effect does not improve the selection criterion.
9	No candidate meets the appropriate SLE or SLS significance level.
10	Stepwise selection is cycling.
11	The model is an exact fit.
12	Dropping an effect would result in an empty model.

The display of the "Stop Reason" table can be suppressed by specifying DETAILS=NONE in the SELECTION statement.

Selection Reason

When you specify the SELECTION statement, the HPREG procedure produces a simple table that contains text informing you about the reason why the final model was selected.

The display of the "Selection Reason" table can be suppressed by specifying DETAILS=NONE in the SELECTION statement.

Selected Effects

When you specify the SELECTION statement, the HPREG procedure produces a simple table that contains text informing you about which effects were selected into the final model.

ANOVA

The "ANOVA" table displays an analysis of variance for the selected model. This table includes the following:

- the Source of the variation, Model for the fitted regression, Error for the residual error, and C Total for the total variation after correcting for the mean. The Uncorrected Total Variation is produced when the NOINT option is used.

- the degrees of freedom (DF) associated with the source
- the Sum of Squares for the term
- the Mean Square, the sum of squares divided by the degrees of freedom
- the *F* Value for testing the hypothesis that all parameters are 0 except for the intercept. This is formed by dividing the mean square for Model by the mean square for Error.
- the Prob>*F*, the probability of getting a greater *F* statistic than that observed if the hypothesis is true. When you do model selection, these *p*-values are generally liberal because they are not adjusted for the fact that the terms in the model have been selected.

You can request "ANOVA" tables for the model at each step of the selection process with the DETAILS= option in the SELECTION statement.

Fit Statistics

The "Fit Statistics" table displays fit statistics for the selected model. The statistics displayed include the following:

- Root MSE, an estimate of the standard deviation of the error term. It is calculated as the square root of the mean square error.
- R-square, a measure between 0 and 1 that indicates the portion of the (corrected) total variation attributed to the fit rather than left to residual error. It is calculated as SS(Model) divided by SS(Total). It is also called the *coefficient of determination*. It is the square of the multiple correlation—in other words, the square of the correlation between the dependent variable and the predicted values.
- Adj R-Sq, the adjusted R-square, a version of R-square that has been adjusted for degrees of freedom. It is calculated as

$$\bar{R}^2 = 1 - \frac{(n-i)(1-R^2)}{n-p}$$

  where i is equal to 1 if there is an intercept and 0 otherwise, n is the number of observations used to fit the model, and p is the number of parameters in the model.
- fit criteria AIC, AICC, BIC, CP, and PRESS if they are used in the selection process. See Table 61.5 for the formulas for evaluating these criteria.
- the average square errors (ASE) on the training, validation, and test data.

You can request "Fit Statistics" tables for the model at each step of the selection process with the DETAILS= option in the SELECTION statement.

Parameter Estimates

The "Parameter Estimates" table displays the parameters in the selected model and their estimates. The information displayed for each parameter in the selected model includes the following:

- the parameter label that includes the effect name and level information for effects that contain classification variables
- the degrees of freedom (DF) for the parameter. There is one degree of freedom unless the model is not full rank.
- the parameter estimate
- the standard error, which is the estimate of the standard deviation of the parameter estimate
- t Value, the *t* test that the parameter is 0. This is computed as the parameter estimate divided by the standard error.
- the Pr > |t|, the probability that a *t* statistic would obtain a greater absolute value than that observed given that the true parameter is 0. This is the two-tailed significance probability.

  When you do model selection, these *p*-values are generally liberal because they are not adjusted for the fact that the terms in the model have been selected.

You can request "Parameter Estimates" tables for the model at each step of the selection process with the DETAILS= option in the SELECTION statement.

Timing Information

If you specify the DETAILS option in the PERFORMANCE statement, the procedure also produces a "Timing" table in which elapsed time (absolute and relative) for the main tasks of the procedure are displayed.

ODS Table Names

Each table created by the HPREG procedure has a name associated with it, and you must use this name to refer to the table when you use ODS statements. These names are listed in Table 61.7.

Table 61.7 ODS Tables Produced by PROC HPREG

Table Name	Description	Required Statement / Option
ANOVA	Selected model ANOVA table	Default output
Candidates	Swap candidates at step	SELECTION DETAILS=ALL|STEPS
ClassLevels	Level information from the CLASS statement	CLASS
DataAccessInfo	Information about modes of data access	Default output
Dimensions	Model dimensions	Default output

Table 61.7 *continued*

Table Name	Description	Required Statement / Option
EntryCandidates	Candidates for entry at step	SELECTION DETAILS=ALL|STEPS
FitStatistics	Fit statistics	Default output
ModelInfo	Information about the modeling environment	Default output
NObs	Number of observations read and used	Default output
ParameterEstimates	Solutions for the parameter estimates associated with effects in MODEL statement	Default output
PerformanceInfo	Information about high-performance computing environment	Default output
RemovalCandidates	Candidates for removal at step	SELECTION DETAILS=ALL|STEPS
ScreenedEffects	List of screened effects	SELECTION SCREEN
ScreeningInfo	Information about the screening method	SELECTION SCREEN
Screening	Screening statistic values for model effects	SELECTION SCREEN(DETAILS=ALL)
SelectedEffects	List of selected effects	SELECTION
SelectionInfo	Information about selection settings	Default output
SelectionReason	Reason for selecting the final model	SELECTION
SelectionSummary	Summary information about the model selection steps	SELECTION
StopReason	Reason selection was terminated	SELECTION
Timing	Timing breakdown by task	SELECTION DETAILS

Examples: HPREG Procedure

Example 61.1: Model Selection with Validation

This example is based on the example "Using Validation and Cross Validation" in the documentation for the GLMSELECT procedure in the *SAS/STAT User's Guide*. This example shows how you can use validation data to monitor and control variable selection. It also demonstrates the use of split classification variables.

The following DATA step produces analysis data that contains a variable that you can use to assign observations to the training, validation, and testing roles. In this case, each role has 5,000 observations.

```
data analysisData;
   drop i j c3Num;
   length c3$ 7;

   array x{20} x1-x20;

   do i=1 to 15000;
      do j=1 to 20;
         x{j} = ranuni(1);
      end;

      c1 = 1 + mod(i,8);
      c2 = ranbin(1,3,.6);

      if      i < 50   then do; c3 = 'tiny';     c3Num=1;end;
      else if i < 250  then do; c3 = 'small';    c3Num=1;end;
      else if i < 600  then do; c3 = 'average';  c3Num=2;end;
      else if i < 1200 then do; c3 = 'big';      c3Num=3;end;
      else                  do; c3 = 'huge';     c3Num=5;end;

      yTrue = 10 + x1 + 2*x5 + 3*x10 + 4*x20  + 3*x1*x7 + 8*x6*x7
                 + 5*(c1=3)*c3Num + 8*(c1=7);

      error = 5*rannor(1);

      y = yTrue + error;

           if mod(i,3)=1 then Role = 'TRAIN';
      else if mod(i,3)=2 then Role = 'VAL';
      else                    Role = 'TEST';

      output;
  end;
run;
```

By construction, the true model consists of main effects x1, x5, x10, x20, and c1 and interaction effects x1*x7, x6*x7, and c1*c3. Furthermore, you can see that only levels 3 and 7 of the classification variable c1 are systematically related to the response.

Because the error term for each observation is five times a value drawn from a standard normal distribution, the expected error variance is 25. For the data in each role, you can compute an estimate of this error variance by forming the average square error (ASE) for the observations in the role. Output 61.1.1 shows the ASE for each role that you can compute with the following statements:

```
proc summary data=analysisData;
   class role;
   ways 1;
   var error;
   output out=ASE uss=uss n=n;
data ASE; set ASE;
   OracleASE = uss / n;
   label OracleASE = 'Oracle ASE';
   keep Role OracleASE;
proc print data=ASE label noobs;
run;


proc print data=ASE label noobs;
run;
```

Output 61.1.1 Oracle ASE Values by Role

Role	Oracle ASE
TEST	25.5784
TRAIN	25.4008
VAL	25.8993

The ASE values shown Output 61.1.1 are labeled as "Oracle ASE" because you need to know the true underlying model if you want to compute these values from the response and underlying regressors. In a modeling context, a good predictive model produces values that are close to these oracle values. An overfit model produces a smaller ASE on the training data but higher values on the validation and test data. An underfit model exhibits higher values for all data roles.

Suppose you suspect that the dependent variable depends on both main effects and two-way interactions. You can use the following statements to select a model:

```
proc hpreg data=analysisData;
   partition roleVar=role(train='TRAIN' validate='VAL' test='TEST');
   class c1 c2 c3(order=data);
   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
             |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2 /stb;
   selection method = stepwise(select=sl sle=0.1 sls=0.15 choose=validate)
                      hierarchy=single details=steps;
run;
```

A PARTITION statement assigns observations to training, validation, and testing roles based on the values of the input variable named role. The SELECTION statement requests STEPWISE selection based on significance level where the SLE and SLS values are set to use the defaults of PROC REG. The CHOOSE=VALIDATE option selects the model that yields the smallest ASE value on the validation data.

The "Number Of Observation" table in Output 61.1.2 confirms that there are 5,000 observations for each data role. The "Dimensions" table shows that the selection is from 278 effects with a total of 661 parameters.

Output 61.1.2 Number of Observations, Class Levels, and Dimensions

The HPREG Procedure

Number of Observations Read	15000
Number of Observations Used	15000
Number of Observations Used for Training	5000
Number of Observations Used for Validation	5000
Number of Observations Used for Testing	5000

Class Level Information

Class	Levels	Values
c1	8	1 2 3 4 5 6 7 8
c2	4	0 1 2 3
c3	5	tiny small average big huge

Dimensions	
Number of Effects	278
Number of Parameters	661

Output 61.1.3 shows the "Selection Summary" table. You see that 18 steps are done, at which point all effects in the model are significant at the SLS value of 0.15 and all the remaining effects if added individually would not be significant at the SLE significance level of 0.1. However, because you have specified the CHOOSE=VALIDATE option, the model at step 18 is not used as the selected model. Instead the model at step 10 (where the validation ASE achieves a local minimum value) is selected. The "Stop Reason," "Selection Reason," and "Selected Effects" in Output 61.1.4 provide this information.

Output 61.1.3 Selection Summary

The HPREG Procedure

Selection Summary

Step	Effect Entered	Number Effects In	Validation ASE	p Value
0	**Intercept**	1	98.3895	1.0000
1	**c1**	2	34.8572	<.0001
2	**x7**	3	32.5531	<.0001
3	**x6**	4	31.0646	<.0001
4	**x20**	5	29.7078	<.0001
5	**x6*x7**	6	29.2210	<.0001
6	**x10**	7	28.6683	<.0001
7	**x1**	8	28.3250	<.0001
8	**x5**	9	27.9766	<.0001
9	**c3**	10	27.8288	<.0001
10	**c1*c3**	11	25.9701*	<.0001
11	**x10*c1**	12	26.0696	0.0109
12	**x4**	13	26.1594	0.0128
13	**x4*x10**	14	26.1814	0.0035
14	**x20*c1**	15	26.3294	0.0156
15	**x1*c3**	16	26.3945	0.0244
16	**x1*x7**	17	26.3632	0.0270
17	**x7*x10**	18	26.4120	0.0313
18	**x1*x20**	19	26.4330	0.0871

*** Optimal Value of Criterion**

Output 61.1.4 Stopping and Selection Reasons

Selection stopped because all candidates for removal are significant at the 0.15 level and no candidate for entry is significant at the 0.1 level.

The model at step 10 is selected where Validation ASE is 25.9701.

Selected Effects: Intercept c1 c3 c1*c3 x1 x5 x6 x7 x6*x7 x10 x20

You can see that the selected effects include all the main effects in the true model and two of the three true interaction terms. Furthermore, the selected model does not include any variables that are not in the true model. Note that these statements are not true of the larger model at the final step of the selection process.

Output 61.1.5 shows the fit statistics of the selected model. You can see that the ASE values on the training, validation, and test data are all similar, which is indicative of a reasonable predictive model. In this case where the true model is known, you can see that all three ASE values are close to oracle values for the true model, as shown in Output 61.1.1.

Output 61.1.5 Fit Statistics for the Selected Model

Root MSE	5.03976
R-Square	0.74483
Adj R-Sq	0.74246
AIC	21222
AICC	21223
SBC	16527
ASE (Train)	25.16041
ASE (Validate)	25.97010
ASE (Test)	25.83436

Because you specified the DETAILS=STEPS option in the SELECTION statement, you can see the "Fit Statistics" for the model at each step of the selection process. Output 61.1.6 shows these fit statistics for final model at step 18. You see that for this model, the ASE value on the training data is smaller than the ASE values on the validation and test data. This is indicative an overfit model that might not generalize well to new data. You see the ASE values on the validation and test data are now worse in comparison to the oracle values than the values for the selected model at step 10.

Output 61.1.6 Fit Statistics for the Model at Step 18

Root MSE	5.01386
R-Square	0.74862
Adj R-Sq	0.74510
AIC	21194
AICC	21196
SBC	16648
ASE (Train)	24.78688
ASE (Validate)	26.43304
ASE (Test)	26.07078

Output 61.1.7 shows part of the "Parameter Estimates" table for the selected model at step 10 that includes the estimates for the main effect c1. Because the STB option is specified in the MODEL statement, this table includes standardized estimates.

Output 61.1.7 Part of the Parameter Estimates Table for the Selected Model

Parameter Estimates

Parameter	DF	Estimate	Standardized Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.479114	0	0.422843	22.42	<.0001
c1 1	1	0.279417	0.009306	0.297405	0.94	0.3475
c1 2	1	0.615589	0.020502	0.297332	2.07	0.0385
c1 3	1	25.678601	0.855233	0.297280	86.38	<.0001
c1 4	1	0.420360	0.014000	0.297283	1.41	0.1574
c1 5	1	0.473986	0.015786	0.297265	1.59	0.1109
c1 6	1	0.394044	0.013124	0.297299	1.33	0.1851
c1 7	1	8.469793	0.282089	0.297345	28.48	<.0001
c1 8	0	0	0	.	.	.

The magnitudes of the standardized estimates and the *t* statistics of the parameters of the effect c1 reveal that only levels 3 and 7 of this effect contribute appreciably to the model. This suggests that a more parsimonious model with similar or better predictive power might be obtained if parameters that correspond to the levels of c1 can enter or leave the model independently. You request this with the SPLIT option in the CLASS statement as shown in the following statements:

```
proc hpreg data=analysisData;
   partition roleVar=role(train='TRAIN' validate='VAL' test='TEST');
   class c1(split) c2 c3(order=data);
   model y =  c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
             |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2 /stb;
   selection method = stepwise(select=sl sle=0.1 sls=0.15 choose=validate)
                      hierarchy=single details=steps;
run;
```

Output 61.1.8 shows the "Dimensions" table. You can see that because the columns in the design matrix that correspond to levels of c1 are treated as separate effects, the selection is now from 439 effects, even though the number of parameters is unchanged.

Output 61.1.8 Dimensions with c1 Split

The HPREG Procedure

Dimensions	
Number of Effects	278
Number of Effects after Splits	439
Number of Parameters	661

Output 61.1.9 shows the selected effects. You can see that as anticipated the selected model now depends on only levels 3 and 7 of c1.

Output 61.1.9 Selected Effects with c1 Split

Selected Effects:	Intercept c1_3 c1_7 c3 c1_3*c3 x1 x5 x6 x7 x6*x7 x10 x20

Finally, the fit statistics for the selected model are shown Output 61.1.10.

Output 61.1.10 Fit Statistics for the Selected Model with c1 Split

Root MSE	5.04060
R-Square	0.74325
Adj R-Sq	0.74238
AIC	21195
AICC	21195
SBC	16311
ASE (Train)	25.31622
ASE (Validate)	25.98055
ASE (Test)	25.76059

If you compare the ASE values for this model in Output 61.1.10 with the oracle values in Output 61.1.1 and the values for the model without splitting c1 in Output 61.1.5, you see that this more parsimonious model produces the best predictive performance on the test data of all the models considered in this example.

Example 61.2: Backward Selection in Single-Machine and Distributed Modes

This example shows how you can run PROC HPREG in single-machine and distributed modes. See the section "Processing Modes" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*) for details about the execution modes of SAS High-Performance Statistics procedures. The focus of this example is to simply show how you can switch the modes of execution of PROC HPREG, rather than on any statistical features of the procedure. The following DATA step generates the data for this example. The response y depends on 20 of the 1,000 regressors.

```
data ex2Data;
   array x{1000};

   do i=1 to 10000;
      y=1;
      sign=1;

      do j=1 to 1000;
         x{j} = ranuni(1);
         if j<=20  then do;
           y = y + sign*j*x{j};
           sign=-sign;
         end;
      end;
      y = y + 5*rannor(1);
      output;
  end;
run;
```

The following statements use PROC HPREG to select a model by using BACKWARD selection:

```
proc hpreg data=ex2Data;
    model y = x: ;
    selection method = backward;
    performance details;
run;
```

Output 61.2.1 shows the "Performance Information" table. This shows that the HPREG procedure executes in single-machine mode using four threads because the client machine has four CPUs. You can force a certain number of threads on any machine involved in the computations with the NTHREADS option in the PERFORMANCE statement.

Output 61.2.1 Performance Information

The HPREG Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	4

Output 61.2.2 shows the parameter estimates for the selected model. You can see that the default BACKWARD selection with selection and stopping based on the SBC criterion retains all 20 of the true effects but also keeps two extraneous effects.

Output 61.2.2 Parameter Estimates for the Selected Model

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.506615	0.419811	3.59	0.0003
x1	1	1.054402	0.176930	5.96	<.0001
x2	1	-1.996080	0.176967	-11.28	<.0001
x3	1	3.293331	0.177032	18.60	<.0001
x4	1	-3.741273	0.176349	-21.22	<.0001
x5	1	4.908310	0.176047	27.88	<.0001
x6	1	-5.772356	0.176642	-32.68	<.0001
x7	1	7.398822	0.175792	42.09	<.0001
x8	1	-7.958471	0.176281	-45.15	<.0001
x9	1	8.899407	0.177624	50.10	<.0001
x10	1	-9.687667	0.176431	-54.91	<.0001
x11	1	11.083373	0.175195	63.26	<.0001
x12	1	-12.046504	0.176324	-68.32	<.0001
x13	1	13.009052	0.176967	73.51	<.0001
x14	1	-14.456393	0.175968	-82.15	<.0001
x15	1	14.928731	0.174868	85.37	<.0001
x16	1	-15.762907	0.177651	-88.73	<.0001
x17	1	16.842889	0.177037	95.14	<.0001
x18	1	-18.468844	0.176502	-104.64	<.0001
x19	1	18.810193	0.176616	106.50	<.0001
x20	1	-20.212291	0.176325	-114.63	<.0001
x87	1	-0.542384	0.176293	-3.08	0.0021
x362	1	-0.560999	0.176594	-3.18	0.0015

Output 61.2.3 shows timing information for the PROC HPREG run. This table is produced when you specify the DETAILS option in the PERFORMANCE statement. You can see that, in this case, the majority of time is spent forming the crossproducts matrix for the model that contains all the regressors.

Output 61.2.3 Timing

Procedure Task Timing

Task	Seconds	Percent
Reading and Levelizing Data	0.16	4.72%
Loading Design Matrix	0.03	0.94%
Computing Moments	0.01	0.45%
Computing Cross Products Matrix	2.39	72.18%
Performing Model Selection	0.72	21.71%

You can switch to running PROC HPREG in distributed mode by specifying valid values for the NODES=, INSTALL=, and HOST= options in the PERFORMANCE statement. An alternative to specifying the INSTALL= and HOST= options in the PERFORMANCE statement is to set appropriate values for the GRIDHOST and GRIDINSTALLLOC environment variables by using OPTIONS SET commands. See the section "Processing Modes" (Chapter 2, *SAS/STAT User's Guide: High-Performance Procedures*) for details about setting these options or environment variables.

The following statements provide an example. To run these statements successfully, you need to set the macro variables GRIDHOST and GRIDINSTALLLOC to resolve to appropriate values, or you can replace the references to macro variables with appropriate values.

```
proc hpreg data=ex2Data;
    model y = x: ;
    selection method = backward;
    performance details nodes = 10
                host="&GRIDHOST" install="&GRIDINSTALLLOC";
run;
```

The execution mode in the "Performance Information" table shown in Output 61.2.4 indicates that the calculations were performed in a distributed environment that uses 10 nodes, each of which uses eight threads.

Output 61.2.4 Performance Information in Distributed Mode

Performance Information	
Host Node	<< your grid host >>
Install Location	<< your grid install location >>
Execution Mode	Distributed
Number of Compute Nodes	10
Number of Threads per Node	32

Another indication of distributed execution is the following message issued by all high-performance statistical procedures in the SAS Log:

```
NOTE: The HPREG procedure is executing in the distributed
      computing environment with 10 worker nodes.
```

Output 61.2.5 shows timing information for this distributed run of the HPREG procedure. In contrast to the single-machine mode (where forming the crossproducts matrix dominated the time spent), the majority of time in distributed mode is spent distributing the data and performing the model selection.

Output 61.2.5 Timing

Procedure Task Timing		
Task	**Seconds**	**Percent**
Distributing Data	0.83	35.52%
Reading and Levelizing Data	0.06	2.75%
Loading Design Matrix	0.02	0.92%
Computing Moments	0.02	0.68%
Computing Cross Products Matrix	0.32	13.64%
Performing Model Selection	0.33	14.12%
Waiting on Client	0.76	32.36%

Example 61.3: Forward-Swap Selection

This example highlights the use of the forward-swap selection method, which is a generalization of the maximum R-square improvement (MAXR) method that is available in the REG procedure in SAS/STAT software. This example also demonstrates the use of the INCLUDE and START options.

The following DATA step produces the simulated data in which the response y depends on six main effects and three 2-way interactions from a set of 20 regressors.

```
data ex3Data;
   array x{20};
   do i=1 to 10000;
      do j=1 to 20;
         x{j} = ranuni(1);
      end;
      y = 3*x1 + 7*x2 -5*x3 + 5*x1*x3 +
          4*x2*x13 + x7 + x11 -x13  + x1*x4 + rannor(1);
      output;
   end;
run;
```

Suppose you want to find the best model of each size in a range of sizes for predicting the response y. You can use the forward-swap selection method to produce good models of each size without the computational expense of examining all possible models of each size. In this example, the criterion used to evaluate the models of each size is the model R square. With this criterion, the forward-swap method coincides with the MAXR method that is available in the REG procedure in SAS/STAT software. The model of a given size for which no pairwise swap of an effect in the model with any candidate effect improves the R-square value is deemed to be the best model of that size.

Suppose that you have prior knowledge that the regressors x1, x2, and x3 are needed in modeling the response y. Suppose that you also believe that some of the two-way interactions of these variables are likely to be important in predicting y and that some other two-way interactions might also be needed. You can use this prior information by specifying the selection process shown in the following statements:

```
proc hpreg data=ex3Data;
    model y = x1|x2|x3|x4|x5|x6|x7|x8|x9|x10|X11|
              x12|x13|x14|x5|x16|x7|x18|x19|x20@2 /
                  include=(x1 x2 x3) start=(x1*x2 x1*x3 x2*x3);
    selection method=forwardswap(select=rsquare maxef=15 choose=sbc)
              details=all;
run;
```

The MODEL statement specifies that all main effects and two-way interactions are candidates for selection. The INCLUDE= option specifies that the effects x1, x2, and x3 must appear in all models that are examined. The START= option specifies that all the two-way interactions of these variables should be used in the initial model that is considered but that these interactions are eligible for removal during the forward-swap selection.

The "Selection Summary" table is shown in Output 61.3.1.

Output 61.3.1 Selection Summary

The HPREG Procedure

Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	Model R-Square
0	Intercept		1		
	x1		2		
	x2		3		
	x1*x2		4		
	x3		5		
	x1*x3		6		
	x2*x3		7	3307.6836	0.8837
1	x2*x13		8	1892.8403	0.8992
2	x7*x11	x1*x2	8	618.9298	0.9112
3	x1*x4	x2*x3	8	405.3751	0.9131
4	x13		9	213.6140	0.9148
5	x7		10	180.4457	0.9152
6	x11	x7*x11	10	1.4039*	0.9167
7	x10*x11		11	2.3393	0.9168
8	x3*x7		12	4.5000	0.9168
9	x6*x7		13	10.0589	0.9169
10	x3*x6		14	13.1113	0.9169
11	x5*x20		15	19.4612	0.9169
12	x13*x20	x3*x6	15	18.3678	0.9169
13	x5*x5	x6*x7	15	12.1398	0.9170*

*** Optimal Value of Criterion**

You see that starting from the model with an intercept and the effects specified in the INCLUDE= and START= options at step 0, the forward-swap selection method adds the effect x2*x13 at step one, because this yields the maximum improvement in R square that can be obtained by adding a single effect. The forward-swap selection method now evaluates whether any effect swap yields a better eight-effect model (one with a higher R-square value). Because you specified the DETAILS=ALL option in the SELECTION statement, at each step where a swap is made you obtain a "Candidates" table that shows the R-square values for the evaluated swaps. Output 61.3.2 shows the "Candidates" for step 2. By default, only the best 10 swaps are displayed.

Output 61.3.2 Swap Candidates at Step 2

Best 10 Candidates			
Rank	Effect Dropped	Effect Added	R-Square
1	x1*x2	x7*x11	0.9112
2	x2*x3	x7*x11	0.9112
3	x1*x2	x7	0.9065
4	x2*x3	x7	0.9065
5	x1*x2	x7*x7	0.9060
6	x2*x3	x7*x7	0.9060
7	x1*x2	x4*x7	0.9060
8	x2*x3	x4*x7	0.9060
9	x1*x2	x11	0.9058
10	x2*x3	x11	0.9058

You see that the best swap adds x7*x11 and drops x1*x2. This yields an eight-effect model whose R-square value (0.9112) is larger than the R-square value (0.8992) of the eight-effect model at step 1. Hence this swap is made at step 2. At step 3, an even better eight-effect model than the model at step 2 is obtained by dropping x2*x3 and adding x1*x4. No additional swap improves the R-square value, and so the model at step 3 is deemed to be the best eight-effect model. Although this is the best eight-effect model that can be found by this method given the starting model, it is not guaranteed that this model that has the highest R-square value among all possible models that consist of seven effects and an intercept.

Because the DETAILS=ALL option is specified in the SELECTION statement, details for the model at each step of the selection process are displayed. Output 61.3.3 provides details of the model at step 3.

Output 61.3.3 Model Details at Step 3

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	108630	15519	15000.3	<.0001
Error	9992	10337	1.03455		
Corrected Total	9999	118967			

Root MSE	1.01713
R-Square	0.91311
Adj R-Sq	0.91305
AIC	10350
AICC	10350
SBC	405.37511
ASE	1.03373

Output 61.3.3 *continued*

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.012095	0.045712	0.26	0.7913
x1	1	3.087078	0.076390	40.41	<.0001
x2	1	7.775180	0.046815	166.08	<.0001
x3	1	-4.957140	0.070995	-69.82	<.0001
x1*x3	1	4.910115	0.122503	40.08	<.0001
x1*x4	1	0.890436	0.060523	14.71	<.0001
x7*x11	1	1.708469	0.045939	37.19	<.0001
x2*x13	1	2.584078	0.061506	42.01	<.0001

The forward-swap method continues to find the best nine-effect model, best 10-effect model, and so on until it obtains the best 15-effect model. At this point the selection terminates because you specified the MAXEF=15 option in the SELECTION statement. The R-square value increases at each step of the selection process. However, because you specified the CHOOSE=SBC criterion in the SELECTION statement, the final model selected is the model at step 6.

Example 61.4: Forward Selection with Screening

This example shows how you can use the SCREEN option in the SELECTION statement to greatly speed up model selection from a large number of regressors. In order to demonstrate the efficacy of model selection with screening, this example uses simulated data in which the response y depends systematically on a relatively small subset of a much larger set of regressors, which is described in Table 61.8.

Table 61.8 Complete Set of Regressors

Regressor Name	Type	Number of Levels	In True Model
xIn1–xIn25	Continuous		Yes
xWeakIn1–xWeakIn2	Continuous		Yes
xOut1–xOut500	Continuous		No
cIn1–cIn5	Classification	From two to five	Yes
cOut1–cOut500	Classification	From two to five	No

The labels In and Out, which are part of the variable names, make it easy to identify whether the selected model succeeds or fails in capturing the true underlying model. The regressors that are labeled xWeakIn1 and xWeakIn2 are predictive, but their influence is substantially smaller than the influence of the other regressors in the true model.

The following DATA step generates the data:

```
%let nObs      = 50000;
%let nContIn   = 25;
%let nContOut  = 500;
%let nClassIn  = 5;
```

```
%let nClassOut = 500;
%let maxLevs   = 5;
%let noiseScale= 1;

 data ex4Data;
   array xIn{&nContIn};
   array xOut{&nContOut};
   array cIn{&nClassIn};
   array cOut{&nClassOut};

   drop i j sign nLevs xBeta;

   do i=1 to &nObs;
      sign  = -1;
      xBeta = 0;
      do j=1 to dim(xIn);
         xIn{j} = ranuni(1);
         xBeta  = xBeta + j*sign*xIn{j};
         sign   = -sign;
      end;
      do j=1 to dim(xOut);
         xOut{j} = ranuni(1);
      end;

      xWeakIn1 = ranuni(1);
      xWeakin2 = ranuni(1);

      xBeta  = xBeta + 0.1*xWeakIn1+ 0.1*xWeakIn2;

      do j=1 to dim(cIn);
         nLevs  = 2 + mod(j,&maxlevs-1);
         cIn{j} = 1+int(ranuni(1)*nLevs);
         xBeta  = xBeta + j*sign*(cIn{j}-nLevs/2);
         sign   = -sign;
      end;

      do j=1 to dim(cOut);
         nLevs  = 2 + mod(j,&maxlevs-1);
         cOut{j} = 1+int(ranuni(1)*nLevs);
      end;

      y = xBeta + &noiseScale*rannor(1);

      output;
   end;
run;
```

When you have insufficient prior knowledge of what effects need to be included in a parsimonious predictive model, a reasonable starting point is to use model selection to build a such a model. In such cases, you might want to consider a large number of possible model effects, even though you know that a successful model that generalizes well for predicting unseen data depends on a relatively small number of effects. In such cases, you can dramatically reduce the computational task by including screening in the model selection process. The following statements show how you do this:

```
proc hpreg data=ex4Data;
    class c: ;
    model y = x: c: ;
    selection method=forward screen(details=all)=100 20;
    performance details;
run;
```

The ordered pair of integers that is specified in the SCREEN option in the SELECTION statement requests that screening be used to reduce the set of regressors to 100 regressors at the first screening stage and to 20 regressors at the second screening stage. This information is reflected in the "Screening Information" table shown in Output 61.4.1.

Output 61.4.1 Screening Information

The HPREG Procedure

Screening Information	
Screening Stages	Multiple
Screening Criterion	Maximum Absolute Correlation
Stage 1 Number of Screened Effects	100
Stage 2 Number of Screened Effects	20

The "Number Of Observations" table in Output 61.4.2 confirms that the data contain 50,000 observations and the "Dimensions" table shows that the selection is from 1,033 effects that have a total of 2,295 parameters.

Output 61.4.2 Number of Observations and Dimensions

Number of Observations Read	50000
Number of Observations Used	50000

Dimensions	
Number of Effects	1033
Number of Parameters	2295

Because you specified the DETAILS=ALL suboption of the SCREEN option, you obtain the "Screening" table in Output 61.4.3, which shows how the screened subset of 100 effects is obtained at the first screening stage. For display purposes, some ranks in this table have been suppressed.

Output 61.4.3 First Stage Screening Details

Effect Screening for Response		
Rank	Effect	Maximum Absolute Correlation
1	xIn25	0.31785
2	xIn24	0.30697
3	xIn23	0.29734
.	.	.
.	.	.
98	xOut338	0.00932
99	cOut363	0.00922
100	cOut194	0.00920
101	xOut125	0.00919*
102	xOut220	0.00916*
103	cOut310	0.00916*
104	cOut49	0.00915*
105	cOut11	0.00915*

The "Screened Effects" table shown in Output 61.4.4 lists the effects from which a model is selected at the first screening stage.

Output 61.4.4 First Stage Screened Effects

Screened Effects:	xIn25 xIn24 xIn23 xIn22 xIn21 xIn20 xIn19 xIn18 xIn17 xIn16 xIn15 xIn14 xIn13 cIn5 xIn12 cIn3 xIn11 xIn10 xIn8 xIn9 xIn7 cIn4 cIn2 xIn6 xIn5 xIn4 xIn3 cIn1 xIn2 cOut498 cOut110 cOut450 cOut441 cOut272 xOut82 cOut45 cOut6 cOut281 cOut134 cOut15 xOut310 xOut252 xOut485 xOut365 cOut138 cOut123 cOut337 cOut195 cOut423 cOut283 cOut62 cOut114 xOut489 cOut14 cOut158 cOut437 xOut64 cOut301 cOut311 cOut187 cOut431 cOut464 cOut388 cOut213 cOut46 xOut329 cOut403 cOut305 cOut171 cOut85 cOut99 cOut249 xOut267 cOut455 cOut457 cOut271 cOut78 xOut93 cOut259 cOut417 cOut258 cOut326 cOut291 cOut263 cOut107 cOut402 cOut17 cOut237 cOut129 cOut198 cOut58 cOut428 cOut135 cOut206 cOut139 cOut113 cOut486 xOut338 cOut363 cOut194

You see that the magnitude of the pairwise correlations of effects xIn1, xWeakIn1, and xWeakIn2 with response are too small for those effects to be included as candidates for selection at the first screening stage.

The first stage continues with forward selection from the screened effects that are shown in Output 61.4.4. The effects in the selected model at this stage are shown in Output 61.4.5.

Output 61.4.5 First Stage Selected Effects

Selected Effects:	Intercept xIn2 xIn3 xIn4 xIn5 xIn6 xIn7 xIn8 xIn9 xIn10 xIn11 xIn12 xIn13 xIn14 xIn15 xIn16 xIn17 xIn18 xIn19 xIn20 xIn21 xIn22 xIn23 xIn24 xIn25 cIn1 cIn2 cIn3 cIn4 cIn5

You see that the selected model at this stage includes only effects that are systematically related to the response. If you had requested that only a single-stage screening method be used by specifying the SINGLESTAGE suboption of the SCREEN option, then the selected model at this stage would have been the final selected model. However, multistage screening is used in this example. The second stage repeats the steps of the first stage except that the modeled response is the residuals from the selected model at the first stage.

Output 61.4.6 shows the screening details at the second stage. You see that 20 effects are chosen by screening at this stage as specified. Because the selected effects from the first stage are orthogonal to the residuals at

the first stage, none of these effects are in the screened subset. Furthermore, you see that although the effects xIn1, xWeakIn1, and xWeakIn2 are weakly correlated with y, they are the most strongly correlated effects with the residuals from the first stage.

Output 61.4.6 Second Stage Screening Details

Screening Stage 2: Residual Fit

Effect Screening for Stage 1 Residuals

Rank	Effect	Maximum Absolute Correlation
1	xIn1	0.27373
2	xWeakIn1	0.02352
3	xWeakin2	0.02132
4	cOut295	0.01524
5	cOut35	0.01443
6	cOut323	0.01417
7	cOut202	0.01406
8	xOut6	0.01401
9	cOut154	0.01263
10	cOut54	0.01160
11	cOut181	0.01159
12	cOut115	0.01150
13	cOut403	0.01144
14	xOut332	0.01142
15	xOut409	0.01141
16	cOut267	0.01137
17	cOut374	0.01132
18	cOut254	0.01128
19	xOut204	0.01121
20	cOut147	0.01120
21	xOut113	0.01116*
22	xOut427	0.01115*
23	cOut259	0.01111*
24	cOut170	0.01106*
25	cOut107	0.01102*

*** Screened Out**

Screened Effects:	xIn1 xWeakIn1 xWeakin2 cOut295 cOut35 cOut323 cOut202 xOut6 cOut154 cOut54 cOut181 cOut115 cOut403 xOut332 xOut409 cOut267 cOut374 cOut254 xOut204 cOut147

Output 61.4.7 shows the selected effects at the second screening stage. You see that the selected effects are precisely the remaining effects that are systematically predictive of y but that were not in the screened subset at the first screening stage.

Output 61.4.7 Second Stage Selected Effects

Selected Effects:	Intercept xIn1 xOut6 xWeakIn1 xWeakin2

In the third and final screening stage, model selection is performed from the union of the screened effects from the first stage (which are shown in Output 61.4.4) and the selected effects from the second stage (which are shown in Output 61.4.7). The selected effects from this final stage are shown in Output 61.4.8.

Output 61.4.8 Final Stage Selected Effects

Selected Effects:	Intercept xIn1 xIn2 xIn3 xIn4 xIn5 xIn6 xIn7 xIn8 xIn9 xIn10 xIn11 xIn12 xIn13 xIn14 xIn15 xIn16 xIn17 xIn18 xIn19 xIn20 xIn21 xIn22 xIn23 xIn24 xIn25 xOut6 xWeakIn1 xWeakin2 cIn1 cIn2 cIn3 cIn4 cIn5

You see that the final selected model contains all the true underlying model effects and just one noise effect (xOut6). Because you specified the DETAILS option in the PERFORMANCE statement, the "Timing" table shown in Output 61.4.9 is displayed.

Output 61.4.9 Timing for Model Selection with Screening

Procedure Task Timing		
Task	**Seconds**	**Percent**
Reading and Levelizing Data	2.31	19.89%
Loading Design Matrix	2.08	17.88%
Computing Moments	0.95	8.20%
Computing Cross Products Matrix	1.90	16.40%
Performing Model Selection	4.37	37.63%

You see that even though the selected model was obtained by selecting from thousands of effects, screening enabled the entire modeling task to be completed in about 10 seconds. You can perform the same model selection without screening as shown in the following statements:

```
proc hpreg data=ex4Data;
    class c: ;
    model y = x: c: ;
    selection method=forward;
    performance details;
run;
```

In this case, the model that is selected without screening is identical to model that is obtained with screening. However, there is no guarantee that you will get identical selected models. Output 61.4.10 shows the "Timing" table for the model selection without screening.

Output 61.4.10 Timing for Model Selection without Screening

Procedure Task Timing		
Task	**Seconds**	**Percent**
Reading and Levelizing Data	2.28	2.59%
Loading Design Matrix	1.14	1.30%
Computing Moments	0.27	0.30%
Computing Cross Products Matrix	35.34	40.22%
Performing Model Selection	48.85	55.59%

You see that the model selection without screening took about 83 seconds, which is substantially slower than the approximately 10 seconds it took when screening was included in the selection process.

References

Akaike, H. (1969). "Fitting Autoregressive Models for Prediction." *Annals of the Institute of Statistical Mathematics* 21:243–247.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. 2nd ed. New York: Springer-Verlag.

Collier Books (1987). *The 1987 Baseball Encyclopedia Update*. New York: Macmillan.

Darlington, R. B. (1968). "Multiple Regression in Psychological Research and Practice." *Psychological Bulletin* 69:161–182.

Draper, N. R., Guttman, I., and Kanemasu, H. (1971). "The Distribution of Certain Regression Statistics." *Biometrika* 58:295–298.

El Ghaoui, L., Viallon, V., and Rabbani, T. (2012). "Safe Feature Elimination for the Lasso and Sparse Supervised Learning Problems." *Pacific Journal of Optimization* 8:667–698. Special issue on conic optimization.

Fan, J., and Lv, J. (2008). "Sure Independence Screening for Ultrahigh Dimensional Feature Space." *Journal of the Royal Statistical Society, Series B* 70:849–911.

Foster, D. P., and Stine, R. A. (2004). "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy." *Journal of the American Statistical Association* 99:303–313.

Harrell, F. E. (2001). *Regression Modeling Strategies*. New York: Springer-Verlag.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Hocking, R. R. (1976). "The Analysis and Selection of Variables in a Linear Regression." *Biometrics* 32:1–50.

Hurvich, C. M., and Tsai, C.-L. (1989). "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76:297–307.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*. 2nd ed. New York: John Wiley & Sons.

Mallows, C. L. (1973). "Some Comments on C_p." *Technometrics* 15:661–675.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–464.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. (2012). "Strong Rules for Discarding Predictors in Lasso-Type Problems." *Journal of the Royal Statistical Society, Series B* 74:245–266.

Time Inc. (1987). "What They Make." *Sports Illustrated* (April 20): 54–81.

Subject Index

Syntax Index