SAS/STAT® 15.1 User's Guide

The STEPDISC Procedure

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.1 User's Guide

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

November 2018

Chapter 114
The STEPDISC Procedure

Contents

Overview: STEPDISC Procedure

Given a classification variable and several quantitative variables, the STEPDISC procedure performs a stepwise discriminant analysis to select a subset of the quantitative variables for use in discriminating among the classes. The set of variables that make up each class is assumed to be multivariate normal with a common covariance matrix. The STEPDISC procedure can use forward selection, backward elimination, or stepwise selection (Klecka 1980). The STEPDISC procedure is a useful prelude to further analyses with the CANDISC procedure or the DISCRIM procedure.

With PROC STEPDISC, variables are chosen to enter or leave the model according to one of two criteria:

- the significance level of an F test from an analysis of covariance, where the variables already chosen act as covariates and the variable under consideration is the dependent variable
- the squared partial correlation for predicting the variable under consideration from the CLASS variable, controlling for the effects of the variables already selected for the model

Forward selection begins with no variables in the model. At each step, PROC STEPDISC enters the variable that contributes most to the discriminatory power of the model as measured by Wilks' lambda, the likelihood ratio criterion. When none of the unselected variables meet the entry criterion, the forward selection process stops.

Backward elimination begins with all variables in the model except those that are linearly dependent on previous variables in the VAR statement. At each step, the variable that contributes least to the discriminatory power of the model as measured by Wilks' lambda is removed. When all remaining variables meet the criterion to stay in the model, the backward elimination process stops.

Stepwise selection begins, like forward selection, with no variables in the model. At each step, the model is examined. If the variable in the model that contributes least to the discriminatory power of the model as measured by Wilks' lambda fails to meet the criterion to stay, then that variable is removed. Otherwise, the variable not in the model that contributes most to the discriminatory power of the model is entered. When all variables in the model meet the criterion to stay and none of the other variables meet the criterion to enter, the stepwise selection process stops. Stepwise selection is the default method of variable selection.

It is important to realize that, in the selection of variables for entry, only one variable can be entered into the model at each step. The selection process does not take into account the relationships between variables that have not yet been selected. Thus, some important variables could be excluded in the process. Also, Wilks' lambda might not be the best measure of discriminatory power for your application. However, if you use PROC STEPDISC carefully, in combination with your knowledge of the data and careful cross validation, it can be a valuable aid in selecting variables for a discrimination model.

As with any stepwise procedure, it is important to remember that when many significance tests are performed, each at a level of, for example, 5% (0.05), the overall probability of rejecting at least one true null hypothesis is much larger than 5%. If you want to prevent including any variables that do not contribute to the discriminatory power of the model in the population, you should specify a very small significance level. In most applications, all variables considered have some discriminatory power, however small. To choose the model that provides the best discrimination by using the sample estimates, you need only to guard against estimating more parameters than can be reliably estimated with the given sample size.

Costanza and Afifi (1979) use Monte Carlo studies to compare alternative stopping rules that can be used with the forward selection method in the two-group multivariate normal classification problem. Five different numbers of variables, ranging from 10 to 30, are considered in the studies. The comparison is based on conditional and estimated unconditional probabilities of correct classification. They conclude that the use of a moderate significance level, in the range of 10 to 25 percent, often performs better than the use of a much larger or a much smaller significance level.

The significance level and the squared partial correlation criteria select variables in the same order, although they might select different numbers of variables. Increasing the sample size tends to increase the number of variables selected when you are using significance levels, but it has little effect on the number selected by using squared partial correlations.

See Chapter 10, "Introduction to Discriminant Procedures," for more information about discriminant analysis.

Getting Started: STEPDISC Procedure

The data in this example are measurements of 159 fish caught in Finland's Lake Laengelmaevesi; this data set is available from the Puranen. For each of the seven species (bream, roach, whitefish, parkki, perch, pike, and smelt) the weight, length, height, and width of each fish are tallied. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail, and from the nose to the end of its tail. The height and width are recorded as percentages of the third length variable. The fish data set is available from the Sashelp library. PROC STEPDISC will select a subset of the six quantitative variables that might be useful for differentiating between the fish species. This subset is used in conjunction with PROC CANDISC and PROC DISCRIM to develop discrimination models.

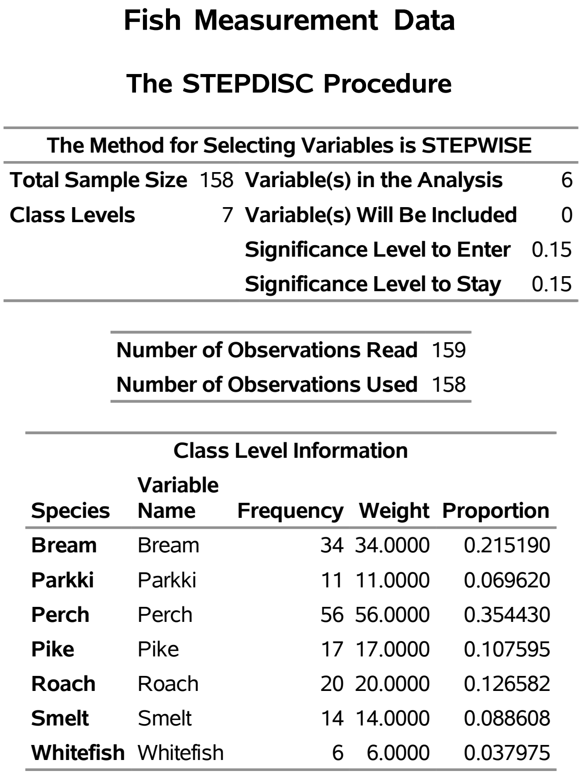
The following steps use PROC STEPDISC to select a subset of potential discriminator variables. By default, PROC STEPDISC uses stepwise selection on all numeric variables that are not listed in other statements, and the significance levels for a variable to enter the subset and to stay in the subset are set to 0.15. The following statements produce Figure 114.1 through Figure 114.5:

```
title 'Fish Measurement Data';

proc stepdisc data=sashelp.fish;
   class Species;
run;
```

PROC STEPDISC begins by displaying summary information about the analysis (see Figure 114.1). This information includes the number of observations with nonmissing values, the number of classes in the classification variable (specified by the CLASS statement), the number of quantitative variables under consideration, the significance criteria for variables to enter and to stay in the model, and the method of variable selection being used. The frequency of each class is also displayed.

Figure 114.1 Summary Information



Fish Measurement Data

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE			
Total Sample Size	158	**Variable(s) in the Analysis**	6
Class Levels	7	**Variable(s) Will Be Included**	0
		Significance Level to Enter	0.15
		Significance Level to Stay	0.15

Number of Observations Read	159
Number of Observations Used	158

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion
Bream	Bream	34	34.0000	0.215190
Parkki	Parkki	11	11.0000	0.069620
Perch	Perch	56	56.0000	0.354430
Pike	Pike	17	17.0000	0.107595
Roach	Roach	20	20.0000	0.126582
Smelt	Smelt	14	14.0000	0.088608
Whitefish	Whitefish	6	6.0000	0.037975

For each entry step, the statistics for entry are displayed for all variables not currently selected (see Figure 114.2). The variable selected to enter at this step (if any) is displayed, as well as all the variables currently selected. Next are multivariate statistics that take into account all previously selected variables and the newly entered variable.

Figure 114.2 Step 1: Variable HEIGHT Selected for Entry

Fish Measurement Data

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 6, 151				
Variable	**R-Square**	**F Value**	**Pr > F**	**Tolerance**
Weight	0.3750	15.10	<.0001	1.0000
Length1	0.6017	38.02	<.0001	1.0000
Length2	0.6098	39.32	<.0001	1.0000
Length3	0.6280	42.49	<.0001	1.0000
Height	0.7553	77.69	<.0001	1.0000
Width	0.4806	23.29	<.0001	1.0000

Variable Height will be entered.

Variable(s) That Have Been Entered
Height

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.244670	77.69	6	151	<.0001
Pillai's Trace	0.755330	77.69	6	151	<.0001
Average Squared Canonical Correlation	0.125888				

For each removal step (Figure 114.3), the statistics for removal are displayed for all variables currently entered. The variable to be removed at this step (if any) is displayed. If no variable meets the criterion to be removed and the maximum number of steps as specified by the MAXSTEP= option has not been attained, then the procedure continues with another entry step.

Figure 114.3 Step 2: No Variable Is Removed; Variable Length2 Added

Fish Measurement Data

The STEPDISC Procedure
Stepwise Selection: Step 2

Statistics for Removal, DF = 6, 151			
Variable	**R-Square**	**F Value**	**Pr > F**
Height	0.7553	77.69	<.0001

No variables can be removed.

Figure 114.3 *continued*

Statistics for Entry, DF = 6, 150				
Variable	**Partial R-Square**	**F Value**	**Pr > F**	**Tolerance**
Weight	0.7388	70.71	<.0001	0.4690
Length1	0.9220	295.35	<.0001	0.6083
Length2	0.9229	299.31	<.0001	0.5892
Length3	0.9173	277.37	<.0001	0.5056
Width	0.8783	180.44	<.0001	0.3699

Variable Length2 will be entered.

Variable(s) That Have Been Entered
Length2 Height

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.018861	157.04	12	300	<.0001
Pillai's Trace	1.554349	87.78	12	302	<.0001
Average Squared Canonical Correlation	0.259058				

The stepwise procedure terminates either when no variable can be removed and no variable can be entered or when the maximum number of steps as specified by the MAXSTEP= option has been attained. In this example at step 7 no variables can be either removed or entered (Figure 114.4). Steps 3 through 6 are not displayed in this document.

Figure 114.4 Step 7: No Variables Entered or Removed

Fish Measurement Data

The STEPDISC Procedure
Stepwise Selection: Step 7

Statistics for Removal, DF = 6, 146			
Variable	**Partial R-Square**	**F Value**	**Pr > F**
Weight	0.4521	20.08	<.0001
Length1	0.2987	10.36	<.0001
Length2	0.5250	26.89	<.0001
Length3	0.7948	94.25	<.0001
Height	0.7257	64.37	<.0001
Width	0.5757	33.02	<.0001

No variables can be removed.

PROC STEPDISC ends by displaying a summary of the steps.

Figure 114.5 Step Summary

No further steps are possible.

Fish Measurement Data

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Height		0.7553	77.69	<.0001	0.24466983	<.0001	0.12588836	<.0001
2	2	Length2		0.9229	299.31	<.0001	0.01886065	<.0001	0.25905822	<.0001
3	3	Length3		0.8826	186.77	<.0001	0.00221342	<.0001	0.38427100	<.0001
4	4	Width		0.5775	33.72	<.0001	0.00093510	<.0001	0.45200732	<.0001
5	5	Weight		0.4461	19.73	<.0001	0.00051794	<.0001	0.49488458	<.0001
6	6	Length1		0.2987	10.36	<.0001	0.00036325	<.0001	0.51744189	<.0001

All the variables in the data set are found to have potential discriminatory power. These variables are used to develop discrimination models in both the CANDISC and DISCRIM procedure chapters.

Syntax: STEPDISC Procedure

The following statements are available in the STEPDISC procedure:

PROC STEPDISC < *options* > ;
  CLASS *variable* ;
  BY *variables* ;
  FREQ *variable* ;
  VAR *variables* ;
  WEIGHT *variable* ;

The BY, CLASS, FREQ, VAR, and WEIGHT statements are described after the PROC STEPDISC statement.

PROC STEPDISC Statement

PROC STEPDISC < *options* > ;

The PROC STEPDISC statement invokes the STEPDISC procedure. Table 114.1 summarizes the options available in the PROC STEPDISC statement.

Table 114.1 STEPDISC Procedure Options

Option	Description
Input Data Set	
DATA=	Specifies input SAS data set

Table 114.1 *continued*

Option	Description
Method Details	
MAXMACRO=	Specifies maximum macro variable lists
METHOD=	Specifies method
SINGULAR=	Specifies singularity
Control Stepwise Selection	
SLENTRY=	Specifies entry significance
SLSTAY=	Specifies staying significance
PR2ENTRY=	Specifies entry partial R square
PR2STAY=	Specifies staying partial R square
INCLUDE=	Forces inclusion of variables
MAXSTEP=	Specifies maximum number of steps
START=	Specifies variables to begin
STOP=	Specifies number of variables in final model
Control Displayed Output	
ALL	Displays all
BCORR	Displays between correlations
BCOV	Displays between covariances
BSSCP	Displays between SSCPs
PCORR	Displays pooled correlations
PCOV	Displays pooled covariances
PSSCP	Displays pooled SSCPs
SHORT	Suppresses output
SIMPLE	Displays descriptive statistics
STDMEAN	Displays standardized class means
TCORR	Displays total correlations
TCOV	Displays total covariances
TSSCP	Displays total SSCPs
WCORR	Displays within correlations
WCOV	Displays within covariances
WSSCP	Displays within SSCPs

ALL

activates all of the display options.

BCORR

displays between-class correlations.

BCOV

displays between-class covariances. The between-class covariance matrix equals the between-class SSCP matrix divided by $n(c-1)/c$, where n is the number of observations and c is the number of classes. The between-class covariances should be interpreted in comparison with the total-sample and within-class covariances, not as formal estimates of population parameters.

BSSCP

displays the between-class SSCP matrix.

DATA=*SAS-data-set*

specifies the data set to be analyzed. The data set can be an ordinary SAS data set or one of several specially structured data sets created by statistical procedures available with SAS/STAT software. These specially structured data sets include TYPE=CORR, COV, CSSCP, and SSCP. If the DATA= option is omitted, the procedure uses the most recently created SAS data set.

INCLUDE=*n*

includes the first *n* variables in the VAR statement in every model. By default, INCLUDE=0.

MAXMACRO=*n*

specifies the maximum number of macro variables with independent variable lists to create. By default, MAXMACRO=100. PROC STEPDISC saves the list of selected variables in a macro variable, &_StdVar. Suppose your input variable list consists of x1-x10; then &_StdVar would be set to x1 x3 x4 x10 if, for example, the first, third, fourth, and tenth variables were selected for the model. This list can be used, for example, in a subsequent procedure's VAR statement as follows:

```
var &_stdvar;
```

With BY processing, one macro variable is created for each BY group, and the macro variables are indexed by the BY-group number. The MAXMACRO= option can be used to either limit or increase the number of these macro variables in processing data sets with many BY groups. The macro variables are created as follows:

With no BY processing, PROC STEPDISC creates the following:

_StdVar	selected variables
_StdVar1	selected variables
_StdNumBys	number of BY groups (1)
_StdNumMacroBys	number of _StdVar*i* macro variables actually made (1)

With BY processing, PROC STEPDISC creates the following:

_StdVar	selected variables for BY group 1
_StdVar1	selected variables for BY group 1
_StdVar2	selected variables for BY group 2
.	
.	
.	
_StdVar*m*	selected variables for BY group *m*, where a number is substituted for *m*
_StdNumBys	*n*, the number of BY groups
_StdNumMacroBys	the number *m* of _StdVar*i* macro variables actually made. This value might be less than _StdNumbys = *n*, and it is less than or equal to the MAXMACRO= value.

MAXSTEP=*n*

specifies the maximum number of steps. By default, MAXSTEP= two times the number of variables in the VAR statement.

METHOD=BACKWARD | BW

METHOD=FORWARD | FW

METHOD=STEPWISE | SW

specifies the method used to select the variables in the model. The BACKWARD method specifies backward elimination, FORWARD specifies forward selection, and STEPWISE specifies stepwise selection. By default, METHOD=STEPWISE.

PCORR

displays pooled within-class correlations (partial correlations based on the pooled within-class covariances).

PCOV

displays pooled within-class covariances.

PR2ENTRY=*p*

PR2E=*p*

specifies the partial R square for adding variables in the forward selection mode, where $p \leq 1$.

PR2STAY=*p*

PR2S=*p*

specifies the partial R square for retaining variables in the backward elimination mode, where $p \leq 1$.

PSSCP

displays the pooled within-class corrected SSCP matrix.

SHORT

suppresses the displayed output from each step.

SIMPLE

displays simple descriptive statistics for the total sample and within each class.

SINGULAR=*p*

specifies the singularity criterion for entering variables, where $0 < p < 1$. PROC STEPDISC precludes the entry of a variable if the squared multiple correlation of the variable with the variables already in the model exceeds $1 - p$. With more than one variable already in the model, PROC STEPDISC also excludes a variable if it would cause any of the variables already in the model to have a squared multiple correlation (with the entering variable and the other variables in the model) exceeding $1 - p$. By default, SINGULAR= 1E–8.

SLENTRY=*p*

SLE=*p*

specifies the significance level for adding variables in the forward selection mode, where $0 \leq p \leq 1$. The default value is 0.15.

SLSTAY=*p*

SLS=*p*

specifies the significance level for retaining variables in the backward elimination mode, where $0 \leq p \leq 1$. The default value is 0.15.

START=*n*

specifies that the first *n* variables in the VAR statement be used to begin the selection process. When you specify METHOD=FORWARD or METHOD=STEPWISE, the default value is 0; when you specify METHOD=BACKWARD, the default value is the number of variables in the VAR statement.

STDMEAN

displays total-sample and pooled within-class standardized class means.

STOP=*n*

specifies the number of variables in the final model. The STEPDISC procedure stops the selection process when a model with *n* variables is found. This option applies only when you specify METHOD=FORWARD or METHOD=BACKWARD. When you specify METHOD=FORWARD, the default value is the number of variables in the VAR statement; when you specify METHOD=BACKWARD, the default value is 0.

TCORR

displays total-sample correlations.

TCOV

displays total-sample covariances.

TSSCP

displays the total-sample corrected SSCP matrix.

WCORR

displays within-class correlations for each class level.

WCOV

displays within-class covariances for each class level.

WSSCP

displays the within-class corrected SSCP matrix for each class level.

BY Statement

BY *variables* **;**

You can specify a BY statement in PROC STEPDISC 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 in the STEPDISC 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* **;**

The values of the CLASS variable define the groups for analysis. Class levels are determined by the formatted values of the CLASS variable. The CLASS variable can be numeric or character. A CLASS statement is required.

FREQ Statement

FREQ *variable* **;**

If a variable in the data set represents the frequency of occurrence for the other values in the observation, include the name of the variable in a FREQ statement. The procedure then treats the data set as if each observation appears *n* times, where *n* is the value of the FREQ variable for the observation. The total number of observations is considered to be equal to the sum of the FREQ variable when the procedure determines degrees of freedom for significance probabilities.

If the value of the FREQ variable is missing or is less than one, the observation is not used in the analysis. If the value is not an integer, the value is truncated to an integer.

VAR Statement

VAR *variables* **;**

The VAR statement specifies the quantitative variables eligible for selection. The default is all numeric variables not listed in other statements.

WEIGHT Statement

WEIGHT *variable* **;**

To use relative weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. This is often done when the variance associated with each observation is different and the values of the WEIGHT variable are proportional to the reciprocals of the variances. If the value of the WEIGHT variable is missing or is less than zero, then a value of zero for the weight is assumed.

The WEIGHT and FREQ statements have a similar effect except that the WEIGHT statement does not alter the degrees of freedom.

Details: STEPDISC Procedure

Missing Values

Observations containing missing values are omitted from the analysis.

Input Data Sets

The input data set can be an ordinary SAS data set or one of several specially structured data sets created by statistical procedures available with SAS/STAT software. For more information about these data sets, see Appendix A, “Special SAS Data Sets.” The BY variable in these data sets becomes the CLASS variable in PROC STEPDISC. These specially structured data sets include the following:

- TYPE=CORR data sets created by PROC CORR by using a BY statement
- TYPE=COV data sets created by PROC PRINCOMP by using both the COV option and a BY statement
- TYPE=CSSCP data sets created by PROC CORR by using the CSSCP option and a BY statement, where the OUT= data set is assigned TYPE=CSSCP with the TYPE= data set option
- TYPE=SSCP data sets created by PROC REG by using both the OUTSSCP= option and a BY statement

When the input data set is TYPE=CORR, TYPE=COV, or TYPE=CSSCP, the STEPDISC procedure reads the number of observations for each class from the observations with _TYPE_='N' and the variable means in each class from the observations with _TYPE_='MEAN'. The procedure then reads the within-class correlations from the observations with _TYPE_='CORR', the standard deviations from the observations with _TYPE_='STD' (data set TYPE=CORR), the within-class covariances from the observations with _TYPE_='COV' (data set TYPE=COV), or the within-class corrected sums of squares and crossproducts from the observations with _TYPE_='CSSCP' (data set TYPE=CSSCP).

When the data set does not include any observations with _TYPE_='CORR' (data set TYPE=CORR), _TYPE_='COV' (data set TYPE=COV), or _TYPE_='CSSCP' (data set TYPE=CSSCP) for each class, PROC STEPDISC reads the pooled within-class information from the data set. In this case, the STEPDISC procedure reads the pooled within-class correlations from the observations with _TYPE_='PCORR', the pooled within-class standard deviations from the observations with _TYPE_='PSTD' (data set TYPE=CORR), the pooled within-class covariances from the observations with _TYPE_='PCOV' (data set TYPE=COV), or the pooled within-class corrected SSCP matrix from the observations with_TYPE_='PSSCP' (data set TYPE=CSSCP).

When the input data set is TYPE=SSCP, the STEPDISC procedure reads the number of observations for each class from the observations with _TYPE_='N', the sum of weights of observations from the variable INTERCEPT in observations with _TYPE_='SSCP' and _NAME_='INTERCEPT', the variable sums from the analysis variables in observations with _TYPE_='SSCP' and _NAME_='INTERCEPT', and the uncorrected sums of squares and crossproducts from the analysis variables in observations with _TYPE_='SSCP' and _NAME_=*variable-names*.

Computational Resources

In the following discussion, let

$$\begin{aligned} n &= \text{number of observations} \\ c &= \text{number of class levels} \\ v &= \text{number of variables in the VAR list} \\ l &= \text{length of the CLASS variable} \\ t &= v + c - 1 \end{aligned}$$

Memory Requirements

The amount of memory in bytes for temporary storage needed to process the data is

$$c(4v^2 + 28v + 3l + 4c + 72) + 16v^2 + 92v + 4t^2 + 20t + 4l$$

Additional temporary storage of 72 bytes at each step is also required to store the results.

Time Requirements

The following factors determine the time requirements of a stepwise discriminant analysis:

- The time needed for reading the data and computing covariance matrices is proportional to nv^2. The STEPDISC procedure must also look up each class level in the list. This is faster if the data are sorted by the CLASS variable. The time for looking up class levels is proportional to a value ranging from n to $n \ln(c)$.

- The time needed for stepwise discriminant analysis is proportional to the number of steps required to select the set of variables in the discrimination model. The number of steps required depends on the data set itself and the selection method and criterion used in the procedure. Each forward or backward step takes time proportional to $(v + c)^2$.

Displayed Output

The displayed output from PROC STEPDISC includes the class level information table. For each level of the classification variable, the following information is provided: the output data set variable name, frequency sum, weight sum, and the proportion of the total sample.

The optional output from PROC STEPDISC includes the following:

The optional output includes the following:

- Within-class SSCP matrices for each group
- Pooled within-class SSCP matrix
- Between-class SSCP matrix
- Total-sample SSCP matrix
- Within-class covariance matrices for each group
- Pooled within-class covariance matrix
- Between-class covariance matrix, equal to the between-class SSCP matrix divided by $n(c-1)/c$, where n is the number of observations and c is the number of classes
- Total-sample covariance matrix
- Within-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the within-class population correlation coefficients are zero
- Pooled within-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the partial population correlation coefficients are zero
- Between-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the between-class population correlation coefficients are zero
- Total-sample correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the total population correlation coefficients are zero
- Simple statistics, including N (the number of observations), sum, mean, variance, and standard deviation for the total sample and within each class
- Total-sample standardized class means, obtained by subtracting the grand mean from each class mean and dividing by the total-sample standard deviation
- Pooled within-class standardized class means, obtained by subtracting the grand mean from each class mean and dividing by the pooled within-class standard deviation

At each step, the following statistics are displayed:

- for each variable considered for entry or removal: partial R-square, the squared (partial) correlation, the F statistic, and $\mathrm{Pr} > F$, the probability level, from a one-way analysis of covariance
- the minimum tolerance for entering each variable. A variable is entered only if its tolerance and the tolerances for all variables already in the model are greater than the value specified in the SINGULAR= option. The tolerance for the entering variable is $1 - R^2$ from regressing the entering variable on the other variables already in the model. The tolerance for a variable already in the model is $1 - R^2$ from regressing that variable on the entering variable and the other variables already in the model. With m variables already in the model, for each entering variable, $m + 1$ multiple regressions are performed by using the entering variable and each of the m variables already in the model as a dependent variable. These $m + 1$ tolerances are computed for each entering variable, and the minimum tolerance is displayed for each.

  The tolerance is computed by using the total-sample correlation matrix. It is customary to compute tolerance by using the pooled within-class correlation matrix (Jennrich 1977), but it is possible for a variable with excellent discriminatory power to have a high total-sample tolerance and a low pooled within-class tolerance. For example, PROC STEPDISC enters a variable that yields perfect discrimination (that is, produces a canonical correlation of one), but a program that uses pooled within-class tolerance does not.
- the variable label, if any
- the name of the variable chosen
- the variables already selected or removed
- Wilks' lambda and the associated F approximation with degrees of freedom and $\mathrm{Pr} < F$, the associated probability level after the selected variable has been entered or removed. Wilks' lambda is the likelihood ratio statistic for testing the hypothesis that the means of the classes on the selected variables are equal in the population (see the section "Multivariate Tests" on page 96 in Chapter 4, "Introduction to Regression Procedures.") Lambda is close to zero if any two groups are well separated.
- Pillai's trace and the associated F approximation with degrees of freedom and $\mathrm{Pr} > F$, the associated probability level after the selected variable has been entered or removed. Pillai's trace is a multivariate statistic for testing the hypothesis that the means of the classes on the selected variables are equal in the population (see the section "Multivariate Tests" on page 96 in Chapter 4, "Introduction to Regression Procedures").
- Average squared canonical correlation (ASCC). The ASCC is Pillai's trace divided by the number of groups minus 1. The ASCC is close to 1 if all groups are well separated and if all or most directions in the discriminant space show good separation for at least two groups.
- Summary to give statistics associated with the variable chosen at each step. The summary includes the following:
  - Step number
  - Variable entered or removed
  - Number in, the number of variables in the model

- Partial R-square
- the F value for entering or removing the variable
- $\mathrm{Pr} > F$, the probability level for the F statistic
- Wilks' lambda
- $\mathrm{Pr} < \mathrm{Lambda}$ based on the F approximation to Wilks' lambda
- Average squared canonical correlation
- $\mathrm{Pr} > \mathrm{ASCC}$ based on the F approximation to Pillai's trace
- the variable label, if any

ODS Table Names

PROC STEPDISC assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 114.2 along with the PROC STEPDISC statement options needed to produce the table. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 114.2 ODS Tables Produced by PROC STEPDISC

ODS Table Name	Description	Option
BCorr	Between-class correlations	BCORR
BCov	Between-class covariances	BCOV
BSSCP	Between-class SSCP matrix	BSSCP
Counts	Number of observations, variables, classes, df	default
CovDF	Nonprinting table of df for covariance matrices	any *COV option
Levels	Class level information	default
Messages	Entry/removal messages	default
Multivariate	Multivariate statistics	default
NObs	Number of observations	default
PCorr	Pooled within-class correlations	PCORR
PCov	Pooled within-class covariances	PCOV
PSSCP	Pooled within-class SSCP matrix	PSSCP
PStdMeans	Pooled standardized class means	STDMEAN
SimpleStatistics	Simple statistics	SIMPLE
Steps	Stepwise selection entry/removal	default
Summary	Stepwise selection summary	default
TCorr	Total-sample correlations	TCORR
TCov	Total-sample covariances	TCOV
TSSCP	Total-sample SSCP matrix	TSSCP
TStdMeans	Total standardized class means	STDMEAN
Variables	Variable lists	default
WCorr	Within-class correlations	WCORR
WCov	Within-class covariances	WCOV
WSSCP	Within-class SSCP matrices	WSSCP

Example: STEPDISC Procedure

Example 114.1: Performing a Stepwise Discriminant Analysis

The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters on 50 iris specimens from each of three species: *Iris setosa*, *I. versicolor*, and *I. virginica*. The iris data set is available from the Sashelp library.

A stepwise discriminant analysis is performed by using stepwise selection.

In the PROC STEPDISC statement, the BSSCP and TSSCP options display the between-class SSCP matrix and the total-sample corrected SSCP matrix. By default, the significance level of an *F* test from an analysis of covariance is used as the selection criterion. The variable under consideration is the dependent variable, and the variables already chosen act as covariates. The following SAS statements produce Output 114.1.1 through Output 114.1.8:

```
title 'Fisher (1936) Iris Data';

%let _stdvar = ;
proc stepdisc data=sashelp.iris bsscp tsscp;
   class Species;
   var SepalLength SepalWidth PetalLength PetalWidth;
run;
```

Output 114.1.1 Iris Data: Summary Information

Fisher (1936) Iris Data

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE			
Total Sample Size	150	**Variable(s) in the Analysis**	4
Class Levels	3	**Variable(s) Will Be Included**	0
		Significance Level to Enter	0.15
		Significance Level to Stay	0.15

Number of Observations Read	150
Number of Observations Used	150

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion
Setosa	Setosa	50	50.0000	0.333333
Versicolor	Versicolor	50	50.0000	0.333333
Virginica	Virginica	50	50.0000	0.333333

Output 114.1.2 Iris Data: Between-Class and Total-Sample SSCP Matrices

Fisher (1936) Iris Data

The STEPDISC Procedure

Between-Class SSCP Matrix

Variable	Label	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	Sepal Length (mm)	6321.21333	-1995.26667	16524.84000	7127.93333
SepalWidth	Sepal Width (mm)	-1995.26667	1134.49333	-5723.96000	-2293.26667
PetalLength	Petal Length (mm)	16524.84000	-5723.96000	43710.28000	18677.40000
PetalWidth	Petal Width (mm)	7127.93333	-2293.26667	18677.40000	8041.33333

Total-Sample SSCP Matrix

Variable	Label	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	Sepal Length (mm)	10216.83333	-632.26667	18987.30000	7692.43333
SepalWidth	Sepal Width (mm)	-632.26667	2830.69333	-4911.88000	-1812.42667
PetalLength	Petal Length (mm)	18987.30000	-4911.88000	46432.54000	19304.58000
PetalWidth	Petal Width (mm)	7692.43333	-1812.42667	19304.58000	8656.99333

In step 1, the tolerance is 1.0 for each variable under consideration because no variables have yet entered the model. The variable PetalLength is selected because its *F* statistic, 1180.161, is the largest among all variables.

Output 114.1.3 Iris Data: Stepwise Selection Step 1

Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 2, 147

Variable	Label	R-Square	F Value	Pr > F	Tolerance
SepalLength	Sepal Length (mm)	0.6187	119.26	<.0001	1.0000
SepalWidth	Sepal Width (mm)	0.4008	49.16	<.0001	1.0000
PetalLength	Petal Length (mm)	0.9414	1180.16	<.0001	1.0000
PetalWidth	Petal Width (mm)	0.9289	960.01	<.0001	1.0000

Variable PetalLength will be entered.

Variable(s) That Have Been Entered
PetalLength

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.058628	1180.16	2	147	<.0001
Pillai's Trace	0.941372	1180.16	2	147	<.0001
Average Squared Canonical Correlation	0.470686				

In step 2, with the variable PetalLength already in the model, PetalLength is tested for removal before a new variable is selected for entry. Since PetalLength meets the criterion to stay, it is used as a covariate in the analysis of covariance for variable selection. The variable SepalWidth is selected because its *F* statistic, 43.035, is the largest among all variables not in the model and because its associated tolerance, 0.8164, meets the criterion to enter. The process is repeated in steps 3 and 4. The variable PetalWidth is entered in step 3, and the variable SepalLength is entered in step 4.

Output 114.1.4 Iris Data: Stepwise Selection Step 2

Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 2

Statistics for Removal, DF = 2, 147

Variable	Label	R-Square	F Value	Pr > F
PetalLength	Petal Length (mm)	0.9414	1180.16	<.0001

No variables can be removed.

Statistics for Entry, DF = 2, 146

Variable	Label	Partial R-Square	F Value	Pr > F	Tolerance
SepalLength	Sepal Length (mm)	0.3198	34.32	<.0001	0.2400
SepalWidth	Sepal Width (mm)	0.3709	43.04	<.0001	0.8164
PetalWidth	Petal Width (mm)	0.2533	24.77	<.0001	0.0729

Variable SepalWidth will be entered.

Variable(s) That Have Been Entered
SepalWidth PetalLength

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.036884	307.10	4	292	<.0001
Pillai's Trace	1.119908	93.53	4	294	<.0001
Average Squared Canonical Correlation	0.559954				

Output 114.1.5 Iris Data: Stepwise Selection Step 3

Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 3

Statistics for Removal, DF = 2, 146

Variable	Label	Partial R-Square	F Value	Pr > F
SepalWidth	Sepal Width (mm)	0.3709	43.04	<.0001
PetalLength	Petal Length (mm)	0.9384	1112.95	<.0001

No variables can be removed.

Statistics for Entry, DF = 2, 145

Variable	Label	Partial R-Square	F Value	Pr > F	Tolerance
SepalLength	Sepal Length (mm)	0.1447	12.27	<.0001	0.1323
PetalWidth	Petal Width (mm)	0.3229	34.57	<.0001	0.0662

Variable PetalWidth will be entered.

Variable(s) That Have Been Entered
SepalWidth PetalLength PetalWidth

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.024976	257.50	6	290	<.0001
Pillai's Trace	1.189914	71.49	6	292	<.0001
Average Squared Canonical Correlation	0.594957				

Output 114.1.6 Iris Data: Stepwise Selection Step 4

Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 4

Statistics for Removal, DF = 2, 145

Variable	Label	Partial R-Square	F Value	Pr > F
SepalWidth	Sepal Width (mm)	0.4295	54.58	<.0001
PetalLength	Petal Length (mm)	0.3482	38.72	<.0001
PetalWidth	Petal Width (mm)	0.3229	34.57	<.0001

No variables can be removed.

Statistics for Entry, DF = 2, 144

Variable	Label	Partial R-Square	F Value	Pr > F	Tolerance
SepalLength	Sepal Length (mm)	0.0615	4.72	0.0103	0.0320

Variable SepalLength will be entered.

All variables have been entered.

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.023439	199.15	8	288	<.0001
Pillai's Trace	1.191899	53.47	8	290	<.0001
Average Squared Canonical Correlation	0.595949				

Since no more variables can be added to or removed from the model, the procedure stops at step 5 and displays a summary of the selection process.

Output 114.1.7 Iris Data: Stepwise Selection Step 5

Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 5

Statistics for Removal, DF = 2, 144

Variable	Label	Partial R-Square	F Value	Pr > F
SepalLength	Sepal Length (mm)	0.0615	4.72	0.0103
SepalWidth	Sepal Width (mm)	0.2335	21.94	<.0001
PetalLength	Petal Length (mm)	0.3308	35.59	<.0001
PetalWidth	Petal Width (mm)	0.2570	24.90	<.0001

No variables can be removed.

Output 114.1.8 Iris Data: Stepwise Selection Summary

No further steps are possible.

Fisher (1936) Iris Data

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	PetalLength		Petal Length (mm)	0.9414	1180.16	<.0001	0.05862828	<.0001	0.47068586	<.0001
2	2	SepalWidth		Sepal Width (mm)	0.3709	43.04	<.0001	0.03688411	<.0001	0.55995394	<.0001
3	3	PetalWidth		Petal Width (mm)	0.3229	34.57	<.0001	0.02497554	<.0001	0.59495691	<.0001
4	4	SepalLength		Sepal Length (mm)	0.0615	4.72	0.0103	0.02343863	<.0001	0.59594941	<.0001

PROC STEPDISC automatically creates a list of the selected variables and stores it in a macro variable. You can submit the following statement to see the list of selected variables:

```
* print the macro variable list;
%put &_stdvar;
```

The macro variable _StdVar contains the following variable list:

```
 SepalLength SepalWidth PetalLength PetalWidth
```

You could use this macro variable if you want to analyze these variables in subsequent steps as follows:

```
proc discrim data=sashelp.iris;
   class Species;
   var &_stdvar;
run;
```

The results of this step are not shown.

References

Costanza, M. C., and Afifi, A. A. (1979). “Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis.” *Journal of the American Statistical Association* 74:777–785.

Fisher, R. A. (1936). “The Use of Multiple Measurements in Taxonomic Problems.” *Annals of Eugenics* 7:179–188.

Jennrich, R. I. (1977). “Stepwise Discriminant Analysis.” In *Statistical Methods for Digital Computers*, edited by K. Enslein, A. Ralston, and H. Wilf, 76–96. New York: John Wiley & Sons.

Klecka, W. R. (1980). *Discriminant Analysis.* Vol. 07-019 of Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage Publications.

Puranen, J. (1917). “Fish Catch data set (1917).” Journal of Statistics Education Data Archive.

Subject Index

Syntax Index