SAS/STAT® 9.3 User’s Guide
The STEPDISC Procedure
(Chapter)
This document is an individual chapter from SAS/STAT® 9.3 User’s Guide.


Copyright © 2011, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a Web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others’ rights is appreciated.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).


1st electronic book, July 2011

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at support.sas.com/publishing or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.
Chapter 85

The STEPDISC Procedure

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.
The data in this example are measurements of 159 fish caught in Finland’s lake Laengelmavesi; 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 85.1 through Figure 85.5:

```
title 'Fish Measurement Data';
proc stepdisc data=sashelp.fish;
   class Species;
runk;
```

PROC STEPDISC begins by displaying summary information about the analysis (see Figure 85.1). 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 85.1 Summary Information**

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sample Size</td>
<td>158</td>
</tr>
<tr>
<td>Class Levels</td>
<td>7</td>
</tr>
<tr>
<td>Variable(s) in the Analysis</td>
<td>6</td>
</tr>
<tr>
<td>Variable(s) Will Be Included</td>
<td>0</td>
</tr>
<tr>
<td>Significance Level to Enter</td>
<td>0.15</td>
</tr>
<tr>
<td>Significance Level to Stay</td>
<td>0.15</td>
</tr>
<tr>
<td>Number of Observations Read</td>
<td>159</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>158</td>
</tr>
</tbody>
</table>
Chapter 85: The STEPDISC Procedure

Figure 85.1 continued

Class Level Information

<table>
<thead>
<tr>
<th>Species</th>
<th>Name</th>
<th>Frequency</th>
<th>Weight</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bream</td>
<td>Bream</td>
<td>34</td>
<td>34.0000</td>
<td>0.215190</td>
</tr>
<tr>
<td>Parkki</td>
<td>Parkki</td>
<td>11</td>
<td>11.0000</td>
<td>0.069620</td>
</tr>
<tr>
<td>Perch</td>
<td>Perch</td>
<td>56</td>
<td>56.0000</td>
<td>0.354430</td>
</tr>
<tr>
<td>Pike</td>
<td>Pike</td>
<td>17</td>
<td>17.0000</td>
<td>0.107595</td>
</tr>
<tr>
<td>Roach</td>
<td>Roach</td>
<td>20</td>
<td>20.0000</td>
<td>0.126582</td>
</tr>
<tr>
<td>Smelt</td>
<td>Smelt</td>
<td>14</td>
<td>14.0000</td>
<td>0.088608</td>
</tr>
<tr>
<td>Whitefish</td>
<td>Whitefish</td>
<td>6</td>
<td>6.0000</td>
<td>0.037975</td>
</tr>
</tbody>
</table>

For each entry step, the statistics for entry are displayed for all variables not currently selected (see Figure 85.2). 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 85.2 Step 1: Variable HEIGHT Selected for Entry

Fish Measurement Data

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 6, 151

<table>
<thead>
<tr>
<th>Variable</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>0.3750</td>
<td>15.10</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>Length1</td>
<td>0.6017</td>
<td>38.02</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>Length2</td>
<td>0.6098</td>
<td>39.32</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>Length3</td>
<td>0.6280</td>
<td>42.49</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>Height</td>
<td>0.7553</td>
<td>77.69</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>Width</td>
<td>0.4806</td>
<td>23.29</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Variable Height will be entered.

Variable(s) That Have Been Entered

Height

Multivariate Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.244670</td>
<td>77.69</td>
<td>6</td>
<td>151</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>0.755330</td>
<td>77.69</td>
<td>6</td>
<td>151</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Average Squared Canonical</td>
<td>0.125888</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For each removal step (Figure 85.3), 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 85.3** Step 2: No Variable Is Removed; Variable Length2 Added

<table>
<thead>
<tr>
<th>Variable</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>0.7553</td>
<td>77.69</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

No variables can be removed.

<table>
<thead>
<tr>
<th>Partial Variable</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>0.7388</td>
<td>70.71</td>
<td>&lt;.0001</td>
<td>0.4690</td>
</tr>
<tr>
<td>Length1</td>
<td>0.9220</td>
<td>295.35</td>
<td>&lt;.0001</td>
<td>0.6083</td>
</tr>
<tr>
<td>Length2</td>
<td>0.9229</td>
<td>299.31</td>
<td>&lt;.0001</td>
<td>0.5892</td>
</tr>
<tr>
<td>Length3</td>
<td>0.9173</td>
<td>277.37</td>
<td>&lt;.0001</td>
<td>0.5056</td>
</tr>
<tr>
<td>Width</td>
<td>0.8783</td>
<td>180.44</td>
<td>&lt;.0001</td>
<td>0.3699</td>
</tr>
</tbody>
</table>

Variable Length2 will be entered.

Variable(s) That Have Been Entered

Length2 Height

Multivariate Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.018861</td>
<td>157.04</td>
<td>12</td>
<td>300</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>1.554349</td>
<td>87.78</td>
<td>12</td>
<td>302</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Average Squared Canonical Correlation</td>
<td>0.259058</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 85.4). Steps 3 through 6 are not displayed in this document.
PROC STEPDISC ends by displaying a summary of the steps.

**Figure 85.5 Step Summary**

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 PROC STEPDISC:

```
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. The options listed in Table 85.1 are available in the PROC STEPDISC statement.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input Data Set</strong></td>
<td></td>
</tr>
<tr>
<td>DATA=</td>
<td>Specifies input SAS data set</td>
</tr>
<tr>
<td><strong>Method Details</strong></td>
<td></td>
</tr>
<tr>
<td>MAXMACRO=</td>
<td>Specifies maximum macro variable lists</td>
</tr>
<tr>
<td>METHOD=</td>
<td>Specifies method</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Specifies singularity</td>
</tr>
<tr>
<td><strong>Control Stepwise Selection</strong></td>
<td></td>
</tr>
<tr>
<td>SLENTRY=</td>
<td>Specifies entry significance</td>
</tr>
<tr>
<td>SLSTAY=</td>
<td>Specifies staying significance</td>
</tr>
<tr>
<td>PR2ENTRY=</td>
<td>Specifies entry partial R square</td>
</tr>
<tr>
<td>PR2STAY=</td>
<td>Specifies staying partial R square</td>
</tr>
<tr>
<td>INCLUDE=</td>
<td>Forces inclusion of variables</td>
</tr>
<tr>
<td>MAXSTEP=</td>
<td>Specifies maximum number of steps</td>
</tr>
<tr>
<td>START=</td>
<td>Specifies variables to begin</td>
</tr>
<tr>
<td>STOP=</td>
<td>Specifies number of variables in final model</td>
</tr>
<tr>
<td><strong>Control Displayed Output</strong></td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>Displays all</td>
</tr>
<tr>
<td>BCORR</td>
<td>Displays between correlations</td>
</tr>
<tr>
<td>BCOV</td>
<td>Displays between covariances</td>
</tr>
<tr>
<td>BSSCP</td>
<td>Displays between SSCPs</td>
</tr>
</tbody>
</table>
Table 85.1  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCORR</td>
<td>Displays pooled correlations</td>
</tr>
<tr>
<td>PCOV</td>
<td>Displays pooled covariances</td>
</tr>
<tr>
<td>PSSCP</td>
<td>Displays pooled SSCP s</td>
</tr>
<tr>
<td>SHORT</td>
<td>Suppresses output</td>
</tr>
<tr>
<td>SIMPLE</td>
<td>Displays descriptive statistics</td>
</tr>
<tr>
<td>STDMEAN</td>
<td>Displays standardized class means</td>
</tr>
<tr>
<td>TCORR</td>
<td>Displays total correlations</td>
</tr>
<tr>
<td>TCOV</td>
<td>Displays total covariances</td>
</tr>
<tr>
<td>TSSCP</td>
<td>Displays total SSCP s</td>
</tr>
<tr>
<td>WCORR</td>
<td>Displays within correlations</td>
</tr>
<tr>
<td>WCOV</td>
<td>Displays within covariances</td>
</tr>
<tr>
<td>WSSCP</td>
<td>Displays within SSCP s</td>
</tr>
</tbody>
</table>

**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:

```plaintext
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
- _StdVar1
- _StdNumBys
- _StdNumMacroBys

With BY processing, PROC STEPDISC creates the following:

- _StdVar
- _StdVar1
- _StdVar2
- .
- .
- _StdVarm
- _StdNumBys
- _StdNumMacroBys

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 \( p \leq 1 \).
Chapter 85: The STEPDISC Procedure

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.

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

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 with PROC STEPDISC to obtain separate analyses on 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 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 variable=variablenames in observations with _TYPE_='SSCP' and _NAME_='INTERCEPT', and the uncorrected sums of squares and crossproducts from the variable=variablenames in observations with _TYPE_='SSCP' and _NAME_=variablenames.

**Computational Resources**

In the following discussion, let

\[ n = \text{number of observations} \]
\[ c = \text{number of class levels} \]
\[ v = \text{number of variables in the VAR list} \]
\[ l = \text{length of the CLASS variable} \]
\[ t = v + c - 1 \]

**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 \( \text{Pr} > |r| \) to test the hypothesis that the within-class population correlation coefficients are zero
- Pooled within-class correlation coefficients and \( \text{Pr} > |r| \) to test the hypothesis that the partial population correlation coefficients are zero
- Between-class correlation coefficients and \( \text{Pr} > |r| \) to test the hypothesis that the between-class population correlation coefficients are zero
- Total-sample correlation coefficients and \( \text{Pr} > |r| \) 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 \( \text{Pr} > F \), 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 \( \text{Pr} < F \), 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 92 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 \( \text{Pr} > F \), 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 92 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
The document contains information on ODS Table Names in PROC STEPDISC. It explains the purpose and content of specific tables created by the procedure, emphasizing the importance of these names for referencing tables when using the Output Delivery System (ODS) to select tables and create output data sets. A table is provided, listing the ODS table names produced by PROC STEPDISC, along with their descriptions and the options needed to produce them. The table includes columns for ODS Table Name, Description, and Option. Each table entry highlights the type of information each table provides, such as between-class correlations, within-class covariances, and total-sample correlations, among others. The document concludes with a note directing readers to further information about ODS in Chapter 20, “Using the Output Delivery System.”
Example: STEPDISC Procedure

Example 85.1: 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 85.1.1 through Output 85.1.8:

```sas
title 'Fisher (1936) Iris Data';
%let _stdvar = ;
proc stepdisc data=sashelp.iris bsscp tsscp;
   class Species;
   var SepalLength SepalWidth PetalLength PetalWidth;
run;
```

Output 85.1.1 Iris Data: Summary Information

<table>
<thead>
<tr>
<th>Total Sample Size</th>
<th>150</th>
<th>Variable(s) in the Analysis</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Levels</td>
<td>3</td>
<td>Variable(s) Will Be Included</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Significance Level to Enter</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Significance Level to Stay</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Number of Observations Read | 150
Number of Observations Used  | 150

Class Level Information

<table>
<thead>
<tr>
<th>Species</th>
<th>Name</th>
<th>Frequency</th>
<th>Weight</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setosa</td>
<td>Setosa</td>
<td>50</td>
<td>50.0000</td>
<td>0.333333</td>
</tr>
<tr>
<td>Versicolor</td>
<td>Versicolor</td>
<td>50</td>
<td>50.0000</td>
<td>0.333333</td>
</tr>
<tr>
<td>Virginica</td>
<td>Virginica</td>
<td>50</td>
<td>50.0000</td>
<td>0.333333</td>
</tr>
</tbody>
</table>
Example 85.1: Performing a Stepwise Discriminant Analysis

Output 85.1.2 Iris Data: Between-Class and Total-Sample SSCP Matrices

<table>
<thead>
<tr>
<th>Fisher (1936) Iris Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>The STEPDISC Procedure</td>
</tr>
</tbody>
</table>

**Between-Class SSCP Matrix**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>SepalLength</th>
<th>SepalWidth</th>
<th>PetalLength</th>
<th>PetalWidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal</td>
<td>Sepal Length (mm)</td>
<td>6321.21333</td>
<td>-1995.26667</td>
<td>16524.84000</td>
<td>7127.93333</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>Sepal Width (mm)</td>
<td>-1995.26667</td>
<td>1134.49333</td>
<td>-5723.96000</td>
<td>-2293.26667</td>
</tr>
<tr>
<td>Petal</td>
<td>Petal Length (mm)</td>
<td>16524.84000</td>
<td>-5723.96000</td>
<td>43710.28000</td>
<td>18677.40000</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>Petal Width (mm)</td>
<td>7127.93333</td>
<td>-2293.26667</td>
<td>18677.40000</td>
<td>8041.33333</td>
</tr>
</tbody>
</table>

**Total-Sample SSCP Matrix**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>SepalLength</th>
<th>SepalWidth</th>
<th>PetalLength</th>
<th>PetalWidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal</td>
<td>Sepal Length (mm)</td>
<td>10216.83333</td>
<td>-632.26667</td>
<td>18987.30000</td>
<td>7692.43333</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>Sepal Width (mm)</td>
<td>-632.26667</td>
<td>2830.69333</td>
<td>-4911.88000</td>
<td>-1812.42667</td>
</tr>
<tr>
<td>Petal</td>
<td>Petal Length (mm)</td>
<td>18987.30000</td>
<td>-4911.88000</td>
<td>46432.54000</td>
<td>19304.58000</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>Petal Width (mm)</td>
<td>7692.43333</td>
<td>-1812.42667</td>
<td>19304.58000</td>
<td>8656.99333</td>
</tr>
</tbody>
</table>
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 85.1.3** Iris Data: Stepwise Selection Step 1

Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 2, 147

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>R-Square</th>
<th>$F$ Value</th>
<th>$Pr &gt; F$</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalLength</td>
<td>Sepal Length (mm)</td>
<td>0.6187</td>
<td>119.26</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>Sepal Width (mm)</td>
<td>0.4008</td>
<td>49.16</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>PetalLength</td>
<td>Petal Length (mm)</td>
<td>0.9414</td>
<td>1180.16</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>Petal Width (mm)</td>
<td>0.9289</td>
<td>960.01</td>
<td>&lt;.0001</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Variable PetalLength will be entered.

Variable(s) That Have Been Entered

PetalLength

Multivariate Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>$F$ Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>$Pr &gt; F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.058628</td>
<td>1180.16</td>
<td>2</td>
<td>147</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>0.941372</td>
<td>1180.16</td>
<td>2</td>
<td>147</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Average Squared Canonical</td>
<td>0.470686</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
Example 85.1: Performing a Stepwise Discriminant Analysis

### Output 85.1.4 Iris Data: Stepwise Selection Step 2

Fisher (1936) Iris Data

The STEPDIC Procedure
Stepwise Selection: Step 2

Statistics for Removal, DF = 2, 147

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PetalLength</td>
<td>Petal Length (mm)</td>
<td>0.9414</td>
<td>1180.16</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

No variables can be removed.

Statistics for Entry, DF = 2, 146

<table>
<thead>
<tr>
<th>Partial</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalLength</td>
<td>0.3198</td>
<td>34.32</td>
<td>&lt;.0001</td>
<td>0.2400</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>0.3709</td>
<td>43.04</td>
<td>&lt;.0001</td>
<td>0.8164</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>0.2533</td>
<td>24.77</td>
<td>&lt;.0001</td>
<td>0.0729</td>
</tr>
</tbody>
</table>

Variable SepalWidth will be entered.

Variable(s) That Have Been Entered
- SepalWidth
- PetalLength

Multivariate Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.036884</td>
<td>307.10</td>
<td>4</td>
<td>292</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>1.119908</td>
<td>93.53</td>
<td>4</td>
<td>294</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Average Squared Canonical</td>
<td>0.559954</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Output 85.1.5 Iris Data: Stepwise Selection Step 3

<table>
<thead>
<tr>
<th>Fisher (1936) Iris Data</th>
</tr>
</thead>
</table>

### The STEPDISC Procedure

Stepwise Selection: Step 3

#### Statistics for Removal, DF = 2, 146

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalWidth</td>
<td>Sepal Width (mm)</td>
<td>0.3709</td>
<td>43.04</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>PetalLength</td>
<td>Petal Length (mm)</td>
<td>0.9384</td>
<td>1112.95</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

No variables can be removed.

#### Statistics for Entry, DF = 2, 145

<table>
<thead>
<tr>
<th>Partial</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalLength</td>
<td>Sepal Length (mm)</td>
<td>0.1447</td>
<td>12.27</td>
<td>&lt;.0001</td>
<td>0.1323</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>Petal Width (mm)</td>
<td>0.3229</td>
<td>34.57</td>
<td>&lt;.0001</td>
<td>0.0662</td>
</tr>
</tbody>
</table>

Variable PetalWidth will be entered.

Variable(s) That Have Been Entered

SepalWidth PetalLength PetalWidth

### Multivariate Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.024976</td>
<td>257.50</td>
<td>6</td>
<td>290</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>1.189914</td>
<td>71.49</td>
<td>6</td>
<td>292</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Average Squared Canonical Correlation</td>
<td>0.594957</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Output 85.1.6 Iris Data: Stepwise Selection Step 4

```
Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 4

Statistics for Removal, DF = 2, 145

<table>
<thead>
<tr>
<th>Partial Variable</th>
<th>Label</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalWidth</td>
<td>Sepal Width (mm)</td>
<td>0.4295</td>
<td>54.58</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>PetalLength</td>
<td>Petal Length (mm)</td>
<td>0.3482</td>
<td>38.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>Petal Width (mm)</td>
<td>0.3229</td>
<td>34.57</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

No variables can be removed.

Statistics for Entry, DF = 2, 144

<table>
<thead>
<tr>
<th>Partial Variable</th>
<th>Label</th>
<th>R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalLength</td>
<td>Sepal Length (mm)</td>
<td>0.0615</td>
<td>4.72</td>
<td>0.0103</td>
<td>0.0320</td>
</tr>
</tbody>
</table>

Variable SepalLength will be entered.

All variables have been entered.

Multivariate Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.023439</td>
<td>199.15</td>
<td>8</td>
<td>288</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>1.191899</td>
<td>53.47</td>
<td>8</td>
<td>290</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Average Squared Canonical Correlation</td>
<td>0.595949</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

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 85.1.7 Iris Data: Stepwise Selection Step 5

Fisher (1936) Iris Data

The STEPDISC Procedure
Stepwise Selection: Step 5

Statistics for Removal, DF = 2, 144

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>Partial R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SepalLength</td>
<td>Sepal Length (mm)</td>
<td>0.0615</td>
<td>4.72</td>
<td>0.0103</td>
</tr>
<tr>
<td>SepalWidth</td>
<td>Sepal Width (mm)</td>
<td>0.2335</td>
<td>21.94</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>PetalLength</td>
<td>Petal Length (mm)</td>
<td>0.3308</td>
<td>35.59</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>PetalWidth</td>
<td>Petal Width (mm)</td>
<td>0.2570</td>
<td>24.90</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

No variables can be removed.

Output 85.1.8 Iris Data: Stepwise Selection Summary

No further steps are possible.

Fisher (1936) Iris Data

The STEPDISC Procedure

Stepwise Selection Summary

<table>
<thead>
<tr>
<th>Number</th>
<th>Step</th>
<th>In Entered</th>
<th>Removed</th>
<th>Partial Label</th>
<th>Partial R-Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Wilks' Lambda</th>
<th>Pr &lt; Lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PetalLength</td>
<td>Petal Length (mm)</td>
<td>0.9414</td>
<td>1180.16</td>
<td>&lt;.0001</td>
<td>0.05862828</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>SepalWidth</td>
<td>Sepal Width (mm)</td>
<td>0.3709</td>
<td>43.04</td>
<td>&lt;.0001</td>
<td>0.03688411</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>PetalWidth</td>
<td>Petal Width (mm)</td>
<td>0.3229</td>
<td>34.57</td>
<td>&lt;.0001</td>
<td>0.02497554</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>SepalLength</td>
<td>Sepal Length (mm)</td>
<td>0.0615</td>
<td>4.72</td>
<td>0.0103</td>
<td>0.02343863</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Average Squared

<table>
<thead>
<tr>
<th>Number</th>
<th>Step</th>
<th>In Entered</th>
<th>Removed</th>
<th>Canonical Correlation ASCC</th>
<th>Pr &gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>PetalLength</td>
<td>0.47068586</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>SepalWidth</td>
<td>0.55995394</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>PetalWidth</td>
<td>0.59495691</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>SepalLength</td>
<td>0.59594941</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
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:

```sas
* 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:

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

The results of this step are not shown.

---

**References**


**URL** [http://www.amstat.org/publications/jse/datasets/fishcatch.txt](http://www.amstat.org/publications/jse/datasets/fishcatch.txt)
Subject Index

D

discriminant analysis
  stepwise selection, 7143

S

STEPDISC procedure
  average squared canonical correlation, 7158
  computational resources, 7156
  input data sets, 7155
  introductory example, 7145
  memory requirements, 7156
  methods, 7143
  missing values, 7155
  ODS table names, 7159
  Pillai’s trace, 7158
  stepwise selection, 7145
  time requirements, 7156
  tolerance, 7158
  Wilks’ lambda, 7158
stepwise discriminant analysis, 7143
stepwise selection
  STEPDISC procedure, 7145

V

variable selection
  discriminant analysis, 7143
Syntax Index

A

ALL option
   PROC STEPDISC statement, 7150

B

BCORR option
   PROC STEPDISC statement, 7150
BCOV option
   PROC STEPDISC statement, 7150
BSSCP option
   PROC STEPDISC statement, 7150
BY statement
   STEPDISC procedure, 7153

C

CLASS statement
   STEPDISC procedure, 7154

D

DATA= option
   PROC STEPDISC statement, 7150

F

FREQ statement
   STEPDISC procedure, 7154

I

INCLUDE= option
   PROC STEPDISC statement, 7150

M

MAXMACRO= option
   PROC STEPDISC statement, 7150
MAXSTEP= option
   PROC STEPDISC statement, 7151
METHOD= option
   PROC STEPDISC statement, 7151

P

PCORR option
   PROC STEPDISC statement, 7151
PCOV option
   PROC STEPDISC statement, 7151
PR2ENTRY= option
   PROC STEPDISC statement, 7151
PR2STAY= option
   PROC STEPDISC statement, 7151
PROC STEPDISC statement, see STEPDISC procedure
PSSCP option
   PROC STEPDISC statement, 7152

S

SHORT option
   PROC STEPDISC statement, 7152
SIMPLE option
   PROC STEPDISC statement, 7152
SINGULAR= option
   PROC STEPDISC statement, 7152
SLENTRY= option
   PROC STEPDISC statement, 7152
SLSTAY= option
   PROC STEPDISC statement, 7152
START= option
   PROC STEPDISC statement, 7152
STDMEAN option
   PROC STEPDISC statement, 7152
STEPDISC procedure
   syntax, 7149
STEPDISC procedure, BY statement, 7153
STEPDISC procedure, CLASS statement, 7154
STEPDISC procedure, FREQ statement, 7154
STEPDISC procedure, PROC STEPDISC statement, 7149
   ALL option, 7150
   BCOV option, 7150
   BSSCP option, 7150
   DATA= option, 7150
   INCLUDE= option, 7150
   MAXMACRO= option, 7150
   MAXSTEP= option, 7151
   METHOD= option, 7151
   PCORR option, 7151
   PCOV option, 7151
   PR2ENTRY= option, 7151
   PR2STAY= option, 7151
PSSCP option, 7152
SHORT option, 7152
SIMPLE option, 7152
SINGULAR= option, 7152
SLENTRY= option, 7152
SLSTAY= option, 7152
START= option, 7152
STDMEAN option, 7152
STOP= option, 7152
TCORR option, 7152
TCOV option, 7153
TSSCP option, 7153
WCORR option, 7153
WCOV option, 7153
WSSCP option, 7153
STEPDISC procedure, VAR statement, 7154
STEPDISC procedure, WEIGHT statement, 7154
STOP= option
   PROC STEPDISC statement, 7152


TCORR option
   PROC STEPDISC statement, 7153
TCOV option
   PROC STEPDISC statement, 7153
TSSCP option
   PROC STEPDISC statement, 7153


VAR statement
   STEPDISC procedure, 7154


WCORR option
   PROC STEPDISC statement, 7153
WCOV option
   PROC STEPDISC statement, 7153
WEIGHT statement
   STEPDISC procedure, 7154
WSSCP option
   PROC STEPDISC statement, 7153
Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to yourturn@sas.com. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to suggest@sas.com.
SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

SAS® Press

Need to learn the basics? Struggling with a programming problem? You’ll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – free on the Web.
- Hard-copy books.

SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.