Chapter 60
The STEPDISC Procedure

Chapter Table of Contents

Chapter 60
The STEPDISC Procedure

Overview

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 using 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 meets 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 may 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 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 percent 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 may select different numbers of variables. Increasing the sample size tends to increase the number of variables selected when using significance levels, but it has little effect on the number selected using squared partial correlations.

See Chapter 7, "Introduction to Discriminant Procedures," for more information on discriminant analysis.

Getting Started

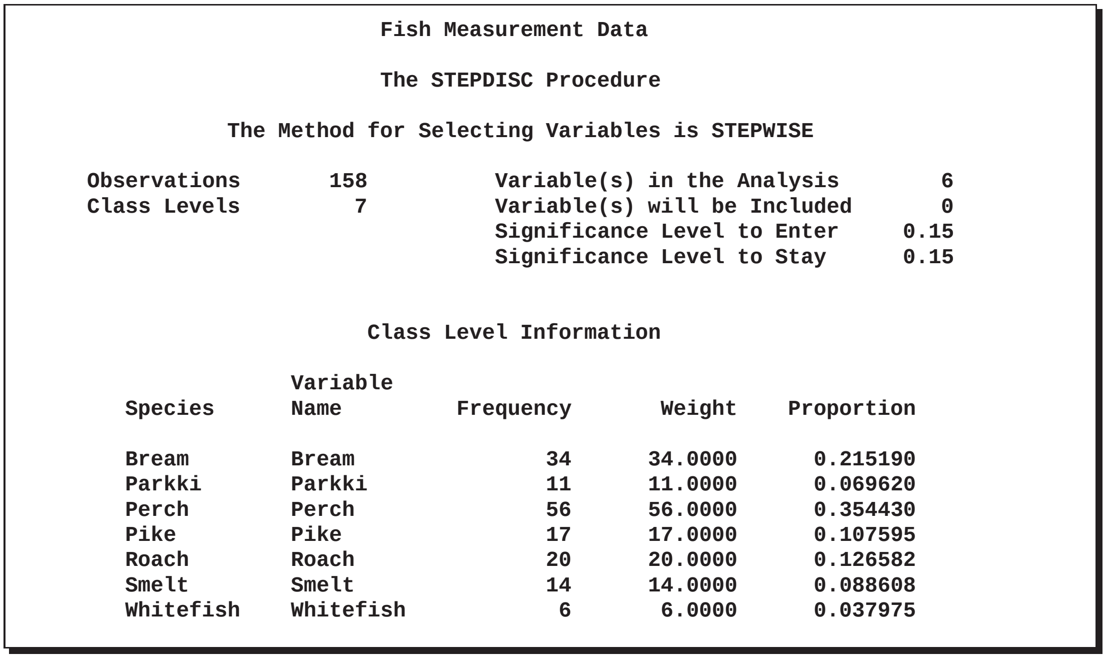
The data in this example are measurements on 159 fish caught off the coast of Finland; this data set is available from the Data Archive of the *Journal of Statistics Education*. For each of the seven species (bream, parkki, pike, perch, roach, smelt, and whitefish), the weight, length, height, and the 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. PROC STEPDISC will select a subset of the six quantitative variables that may 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 program creates the data set fish and uses 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.

```
proc format;
   value specfmt
      1=’Bream’
      2=’Roach’
      3=’Whitefish’
      4=’Parkki’
      5=’Perch’
      6=’Pike’
      7=’Smelt’;
data fish (drop=HtPct WidthPct);
   title ’Fish Measurement Data’;
   input Species Weight Length1 Length2 Length3 HtPct WidthPct @@;
   Height=HtPct*Length3/100;
   Width=WidthPct*Length3/100;
   format Species specfmt.;
   datalines;
1  242.0 23.2 25.4 30.0 38.4 13.4 1  290.0 24.0 26.3 31.2 40.0 13.8
1  340.0 23.9 26.5 31.1 39.8 15.1 1  363.0 26.3 29.0 33.5 38.0 13.3
1  430.0 26.5 29.0 34.0 36.6 15.1 1  450.0 26.8 29.7 34.7 39.2 14.2
1  500.0 26.8 29.7 34.5 41.1 15.3 1  390.0 27.6 30.0 35.0 36.2 13.4
1  450.0 27.6 30.0 35.1 39.9 13.8 1  500.0 28.5 30.7 36.2 39.3 13.7
1  475.0 28.4 31.0 36.2 39.4 14.1 1  500.0 28.7 31.0 36.2 39.7 13.3
1  500.0 29.1 31.5 36.4 37.8 12.0 1     .  29.5 32.0 37.3 37.3 13.6
1  600.0 29.4 32.0 37.2 40.2 13.9 1  600.0 29.4 32.0 37.2 41.5 15.0
1  700.0 30.4 33.0 38.3 38.8 13.8 1  700.0 30.4 33.0 38.5 38.8 13.5
1  610.0 30.9 33.5 38.6 40.5 13.3 1  650.0 31.0 33.5 38.7 37.4 14.8
1  575.0 31.3 34.0 39.5 38.3 14.1 1  685.0 31.4 34.0 39.2 40.8 13.7
1  620.0 31.5 34.5 39.7 39.1 13.3 1  680.0 31.8 35.0 40.6 38.1 15.1
1  700.0 31.9 35.0 40.5 40.1 13.8 1  725.0 31.8 35.0 40.9 40.0 14.8
1  720.0 32.0 35.0 40.6 40.3 15.0 1  714.0 32.7 36.0 41.5 39.8 14.1
1  850.0 32.8 36.0 41.6 40.6 14.9 1 1000.0 33.5 37.0 42.6 44.5 15.5
1  920.0 35.0 38.5 44.1 40.9 14.3 1  955.0 35.0 38.5 44.0 41.1 14.3
1  925.0 36.2 39.5 45.3 41.4 14.9 1  975.0 37.4 41.0 45.9 40.6 14.7
1  950.0 38.0 41.0 46.5 37.9 13.7
2   40.0 12.9 14.1 16.2 25.6 14.0 2   69.0 16.5 18.2 20.3 26.1 13.9
2   78.0 17.5 18.8 21.2 26.3 13.7 2   87.0 18.2 19.8 22.2 25.3 14.3
2  120.0 18.6 20.0 22.2 28.0 16.1 2    0.0 19.0 20.5 22.8 28.4 14.7
2  110.0 19.1 20.8 23.1 26.7 14.7 2  120.0 19.4 21.0 23.7 25.8 13.9
2  150.0 20.4 22.0 24.7 23.5 15.2 2  145.0 20.5 22.0 24.3 27.3 14.6
2  160.0 20.5 22.5 25.3 27.8 15.1 2  140.0 21.0 22.5 25.0 26.2 13.3
2  160.0 21.1 22.5 25.0 25.6 15.2 2  169.0 22.0 24.0 27.2 27.7 14.1
2  161.0 22.0 23.4 26.7 25.9 13.6 2  200.0 22.1 23.5 26.8 27.6 15.4
2  180.0 23.6 25.2 27.9 25.4 14.0 2  290.0 24.0 26.0 29.2 30.4 15.4
2  272.0 25.0 27.0 30.6 28.0 15.6 2  390.0 29.5 31.7 35.0 27.1 15.3
3  270.0 23.6 26.0 28.7 29.2 14.8 3  270.0 24.1 26.5 29.3 27.8 14.5
3  306.0 25.6 28.0 30.8 28.5 15.2 3  540.0 28.5 31.0 34.0 31.6 19.3
3  800.0 33.7 36.4 39.6 29.7 16.6 3 1000.0 37.3 40.0 43.5 28.4 15.0
4   55.0 13.5 14.7 16.5 41.5 14.1 4   60.0 14.3 15.5 17.4 37.8 13.3
4   90.0 16.3 17.7 19.8 37.4 13.5 4  120.0 17.5 19.0 21.3 39.4 13.7
4  150.0 18.4 20.0 22.4 39.7 14.7 4  140.0 19.0 20.7 23.2 36.8 14.2
4  170.0 19.0 20.7 23.2 40.5 14.7 4  145.0 19.8 21.5 24.1 40.4 13.1
4  200.0 21.2 23.0 25.8 40.1 14.2 4  273.0 23.0 25.0 28.0 39.6 14.8
```

```
4  300.0 24.0 26.0 29.0 39.2 14.6
5    5.9  7.5  8.4  8.8 24.0 16.0 5   32.0 12.5 13.7 14.7 24.0 13.6
5   40.0 13.8 15.0 16.0 23.9 15.2 5   51.5 15.0 16.2 17.2 26.7 15.3
5   70.0 15.7 17.4 18.5 24.8 15.9 5  100.0 16.2 18.0 19.2 27.2 17.3
5   78.0 16.8 18.7 19.4 26.8 16.1 5   80.0 17.2 19.0 20.2 27.9 15.1
5   85.0 17.8 19.6 20.8 24.7 14.6 5   85.0 18.2 20.0 21.0 24.2 13.2
5  110.0 19.0 21.0 22.5 25.3 15.8 5  115.0 19.0 21.0 22.5 26.3 14.7
5  125.0 19.0 21.0 22.5 25.3 16.3 5  130.0 19.3 21.3 22.8 28.0 15.5
5  120.0 20.0 22.0 23.5 26.0 14.5 5  120.0 20.0 22.0 23.5 24.0 15.0
5  130.0 20.0 22.0 23.5 26.0 15.0 5  135.0 20.0 22.0 23.5 25.0 15.0
5  110.0 20.0 22.0 23.5 23.5 17.0 5  130.0 20.5 22.5 24.0 24.4 15.1
5  150.0 20.5 22.5 24.0 28.3 15.1 5  145.0 20.7 22.7 24.2 24.6 15.0
5  150.0 21.0 23.0 24.5 21.3 14.8 5  170.0 21.5 23.5 25.0 25.1 14.9
5  225.0 22.0 24.0 25.5 28.6 14.6 5  145.0 22.0 24.0 25.5 25.0 15.0
5  188.0 22.6 24.6 26.2 25.7 15.9 5  180.0 23.0 25.0 26.5 24.3 13.9
5  197.0 23.5 25.6 27.0 24.3 15.7 5  218.0 25.0 26.5 28.0 25.6 14.8
5  300.0 25.2 27.3 28.7 29.0 17.9 5  260.0 25.4 27.5 28.9 24.8 15.0
5  265.0 25.4 27.5 28.9 24.4 15.0 5  250.0 25.4 27.5 28.9 25.2 15.8
5  250.0 25.9 28.0 29.4 26.6 14.3 5  300.0 26.9 28.7 30.1 25.2 15.4
5  320.0 27.8 30.0 31.6 24.1 15.1 5  514.0 30.5 32.8 34.0 29.5 17.7
5  556.0 32.0 34.5 36.5 28.1 17.5 5  840.0 32.5 35.0 37.3 30.8 20.9
5  685.0 34.0 36.5 39.0 27.9 17.6 5  700.0 34.0 36.0 38.3 27.7 17.6
5  700.0 34.5 37.0 39.4 27.5 15.9 5  690.0 34.6 37.0 39.3 26.9 16.2
5  900.0 36.5 39.0 41.4 26.9 18.1 5  650.0 36.5 39.0 41.4 26.9 14.5
5  820.0 36.6 39.0 41.3 30.1 17.8 5  850.0 36.9 40.0 42.3 28.2 16.8
5  900.0 37.0 40.0 42.5 27.6 17.0 5 1015.0 37.0 40.0 42.4 29.2 17.6
5  820.0 37.1 40.0 42.5 26.2 15.6 5 1100.0 39.0 42.0 44.6 28.7 15.4
5 1000.0 39.8 43.0 45.2 26.4 16.1 5 1100.0 40.1 43.0 45.5 27.5 16.3
5 1000.0 40.2 43.5 46.0 27.4 17.7 5 1000.0 41.1 44.0 46.6 26.8 16.3
6  200.0 30.0 32.3 34.8 16.0  9.7 6  300.0 31.7 34.0 37.8 15.1 11.0
6  300.0 32.7 35.0 38.8 15.3 11.3 6  300.0 34.8 37.3 39.8 15.8 10.1
6  430.0 35.5 38.0 40.5 18.0 11.3 6  345.0 36.0 38.5 41.0 15.6  9.7
6  456.0 40.0 42.5 45.5 16.0  9.5 6  510.0 40.0 42.5 45.5 15.0  9.8
6  540.0 40.1 43.0 45.8 17.0 11.2 6  500.0 42.0 45.0 48.0 14.5 10.2
6  567.0 43.2 46.0 48.7 16.0 10.0 6  770.0 44.8 48.0 51.2 15.0 10.5
6  950.0 48.3 51.7 55.1 16.2 11.2 6 1250.0 52.0 56.0 59.7 17.9 11.7
6 1600.0 56.0 60.0 64.0 15.0  9.6 6 1550.0 56.0 60.0 64.0 15.0  9.6
6 1650.0 59.0 63.4 68.0 15.9 11.0
7    6.7  9.3  9.8 10.8 16.1  9.7 7    7.5 10.0 10.5 11.6 17.0 10.0
7    7.0 10.1 10.6 11.6 14.9  9.9 7    9.7 10.4 11.0 12.0 18.3 11.5
7    9.8 10.7 11.2 12.4 16.8 10.3 7    8.7 10.8 11.3 12.6 15.7 10.2
7   10.0 11.3 11.8 13.1 16.9  9.8 7    9.9 11.3 11.8 13.1 16.9  8.9
7    9.8 11.4 12.0 13.2 16.7  8.7 7   12.2 11.5 12.2 13.4 15.6 10.4
7   13.4 11.7 12.4 13.5 18.0  9.4 7   12.2 12.1 13.0 13.8 16.5  9.1
7   19.7 13.2 14.3 15.2 18.9 13.6 7   19.9 13.8 15.0 16.2 18.1 11.6
;
proc stepdisc data=fish;
   class Species;
run;
```

PROC STEPDISC begins by displaying summary information about the analysis; see Figure 60.1. 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.



```
                          Fish Measurement Data

                          The STEPDISC Procedure

             The Method for Selecting Variables is STEPWISE

      Observations       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


                         Class Level Information

                      Variable
         Species      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
```

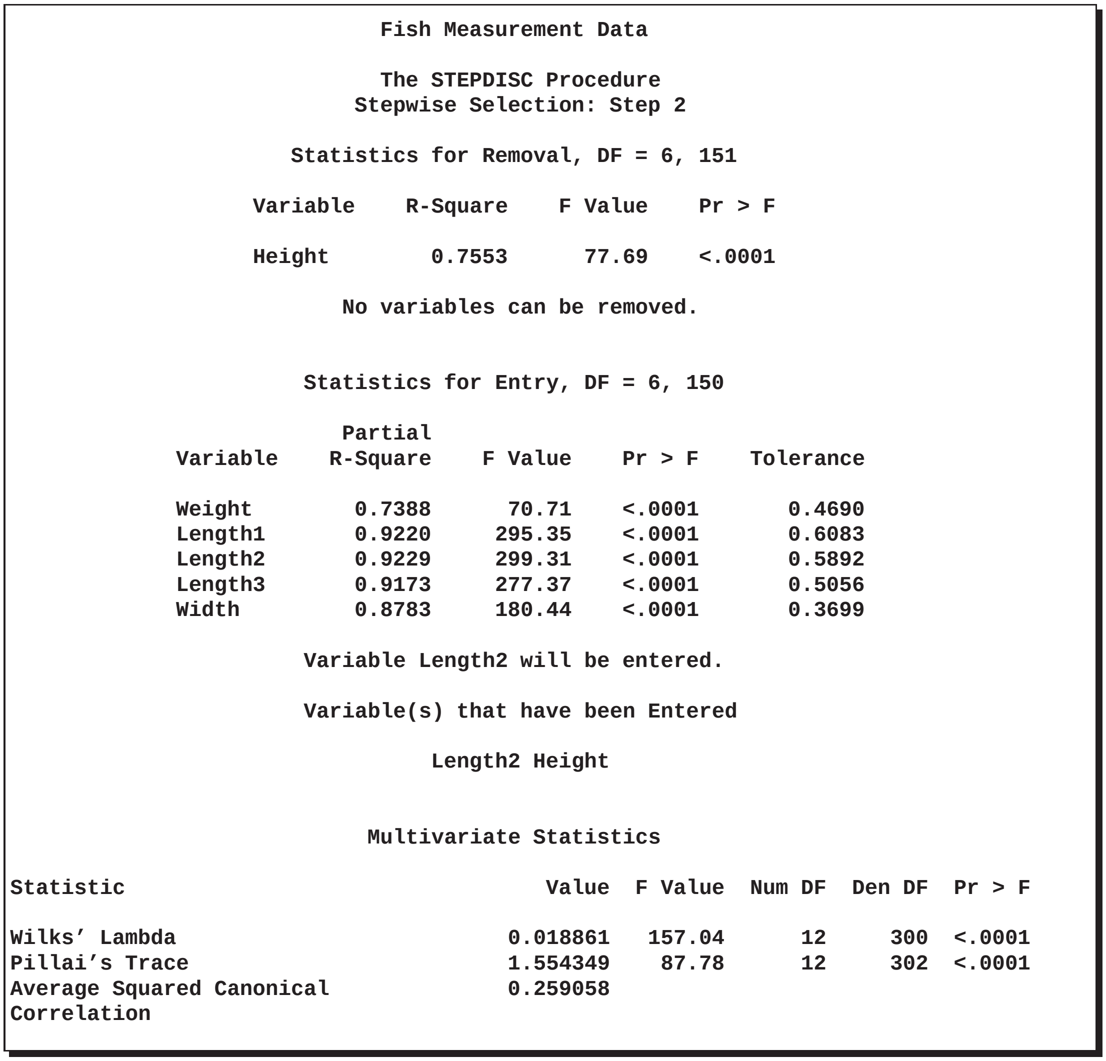
Figure 60.1. Summary Information

For each entry step, the statistics for entry are displayed for all variables not currently selected; see Figure 60.2. 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.

```
                              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              0.125888
Correlation
```

Figure 60.2. Step 1: Variable HEIGHT Selected for Entry

For each removal step (Figure 60.3), 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.



```
                              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.


                       Statistics for Entry, DF = 6, 150

                          Partial
             Variable    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              0.259058
Correlation
```

Figure 60.3. Step 2: No Variable is Removed; Variable Length1 Added

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 60.4). Steps 3 through 6 are not displayed in this document.

```
                          Fish Measurement Data

                         The STEPDISC Procedure
                       Stepwise Selection: Step 7

                  Statistics for Removal, DF = 6, 146

                             Partial
                  Variable    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.

                     No further steps are possible.
```

Figure 60.4. Step 7: No Variables Entered or Removed

PROC STEPDISC ends by displaying a summary of the steps.

```
                                       Fish Measurement Data

                                      The STEPDISC Procedure

                                    Stepwise Selection Summary

                                                                                  Average
                                                                                  Squared
         Number                     Partial                          Wilks'     Pr <   Canonical    Pr >
   Step      In  Entered  Removed  R-Square  F Value  Pr > F        Lambda   Lambda  Correlation    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
```

Figure 60.5. 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

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 PROC STEPDISC statement has the following options.

Table 60.1. STEPDISC Procedure Options

Task	Options
Specify Data Set	DATA=
Select Method	METHOD=
Selection Criterion	SLENTRY= SLSTAY= PR2ENTRY= PR2STAY=
Selection Process	INCLUDE= MAXSTEP= START= STOP=
Determine Singularity	SINGULAR=
Control Displayed Output	
Correlations	BCORR PCORR TCORR WCORR
Covariances	BCOV PCOV TCOV WCOV

Table 60.1. (continued)

Task	Options
SSCP Matrices	BSSCP PSSCP TSSCP WSSCP
Miscellaneous	ALL SIMPLE STDMEAN
Suppress Output	SHORT

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.

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^2 for adding variables in the forward selection mode, where $p \leq 1$.

PR2STAY=*p*
PR2S=*p*

specifies the partial R^2 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 with PROC STEPDISC to obtain separate analyses on observations in groups 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 your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING 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 using the DATASETS procedure (in base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *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

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 on 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

- TYPE=CORR data sets created by PROC CORR using a BY statement
- TYPE=COV data sets created by PROC PRINCOMP using both the COV option and a BY statement
- TYPE=CSSCP data sets created by PROC CORR 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 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

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

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 STEPDISC procedure displays the following output:

- Class Level Information, including the values of the classification variable, the Frequency of each value, the Weight of each value, and the Proportion of each value in the total sample

Optional output includes

- 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
- descriptive 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 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 using the total-sample correlation matrix. It is customary to compute tolerance 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 using 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 "Multivariate Tests" section in Chapter 3, "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 "Multivariate Tests" section in Chapter 3).
- 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 the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 60.2. ODS Tables Produced in PROC STEPDISC

ODS Table Name	Description	PROC STEPWISE 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	DF for covariance matrices, not printed	any *COV option
Levels	Class level information	default
Messages	Entry/removal messages	default
Multivariate	Multivariate statistics	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

Example 60.1. 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 fifty iris specimens from each of three species: *Iris setosa, I. versicolor,* and *I. virginica*.

```
proc format;
   value specname
      1='Setosa    '
      2='Versicolor'
      3='Virginica ';
data iris;
   title 'Fisher (1936) Iris Data';
   input SepalLength SepalWidth PetalLength PetalWidth
         Species @@;
   format Species specname.;
   label SepalLength='Sepal Length in mm.'
         SepalWidth ='Sepal Width in mm.'
         PetalLength='Petal Length in mm.'
         PetalWidth ='Petal Width in mm.';
   datalines;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2 67 31 56 24 3
63 28 51 15 3 46 34 14 03 1 69 31 51 23 3 62 22 45 15 2
59 32 48 18 2 46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3 58 27 51 19 3
68 32 59 23 3 51 33 17 05 1 57 28 45 13 2 62 34 54 23 3
77 38 67 22 3 63 33 47 16 2 67 33 57 25 3 76 30 66 21 3
49 25 45 17 3 55 35 13 02 1 67 30 52 23 3 70 32 47 14 2
64 32 45 15 2 61 28 40 13 2 48 31 16 02 1 59 30 51 18 3
55 24 38 11 2 63 25 50 19 3 64 32 53 23 3 52 34 14 02 1
49 36 14 01 1 54 30 45 15 2 79 38 64 20 3 44 32 13 02 1
67 33 57 21 3 50 35 16 06 1 58 26 40 12 2 44 30 13 02 1
77 28 67 20 3 63 27 49 18 3 47 32 16 02 1 55 26 44 12 2
50 23 33 10 2 72 32 60 18 3 48 30 14 03 1 51 38 16 02 1
61 30 49 18 3 48 34 19 02 1 50 30 16 02 1 50 32 12 02 1
61 26 56 14 3 64 28 56 21 3 43 30 11 01 1 58 40 12 02 1
51 38 19 04 1 67 31 44 14 2 62 28 48 18 3 49 30 14 02 1
51 35 14 02 1 56 30 45 15 2 58 27 41 10 2 50 34 16 04 1
46 32 14 02 1 60 29 45 15 2 57 26 35 10 2 57 44 15 04 1
50 36 14 02 1 77 30 61 23 3 63 34 56 24 3 58 27 51 19 3
57 29 42 13 2 72 30 58 16 3 54 34 15 04 1 52 41 15 01 1
71 30 59 21 3 64 31 55 18 3 60 30 48 18 3 63 29 56 18 3
49 24 33 10 2 56 27 42 13 2 57 30 42 12 2 55 42 14 02 1
49 31 15 02 1 77 26 69 23 3 60 22 50 15 3 54 39 17 04 1
66 29 46 13 2 52 27 39 14 2 60 34 45 16 2 50 34 15 02 1
44 29 14 02 1 50 20 35 10 2 55 24 37 10 2 58 27 39 12 2
47 32 13 02 1 46 31 15 02 1 69 32 57 23 3 62 29 43 13 2
74 28 61 19 3 59 30 42 15 2 51 34 15 02 1 50 35 13 03 1
56 28 49 20 3 60 22 40 10 2 73 29 63 18 3 67 25 58 18 3
```

```
49 31 15 01 1 67 31 47 15 2 63 23 44 13 2 54 37 15 02 1
56 30 41 13 2 63 25 49 15 2 61 28 47 12 2 64 29 43 13 2
51 25 30 11 2 57 28 41 13 2 65 30 58 22 3 69 31 54 21 3
54 39 13 04 1 51 35 14 03 1 72 36 61 25 3 65 32 51 20 3
61 29 47 14 2 56 29 36 13 2 69 31 49 15 2 64 27 53 19 3
68 30 55 21 3 55 25 40 13 2 48 34 16 02 1 48 30 14 01 1
45 23 13 03 1 57 25 50 20 3 57 38 17 03 1 51 38 15 03 1
55 23 40 13 2 66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
51 37 15 04 1 52 35 15 02 1 58 28 51 24 3 67 30 50 17 2
63 33 60 25 3 53 37 15 02 1
;
```

A stepwise discriminant analysis is performed 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 60.1.1 through Output 60.1.8:

```
proc stepdisc data=iris bsscp tsscp;
   class Species;
   var SepalLength SepalWidth PetalLength PetalWidth;
run;
```

Output 60.1.1. Iris Data: Summary Information

```
                                 Fisher (1936) Iris Data

                                 The STEPDISC Procedure

                     The Method for Selecting Variables is STEPWISE

          Observations       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


                                Class Level Information

                           Variable
            Species        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 60.1.2. 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 in mm.      6321.21333    -1995.26667    16524.84000     7127.93333
SepalWidth    Sepal Width in mm.      -1995.26667     1134.49333    -5723.96000    -2293.26667
PetalLength   Petal Length in mm.     16524.84000    -5723.96000    43710.28000    18677.40000
PetalWidth    Petal Width in mm.       7127.93333    -2293.26667    18677.40000     8041.33333


                                  Total-Sample SSCP Matrix

Variable      Label                   SepalLength     SepalWidth    PetalLength     PetalWidth

SepalLength   Sepal Length in mm.     10216.83333     -632.26667    18987.30000     7692.43333
SepalWidth    Sepal Width in mm.       -632.26667     2830.69333    -4911.88000    -1812.42667
PetalLength   Petal Length in mm.     18987.30000    -4911.88000    46432.54000    19304.58000
PetalWidth    Petal Width in 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. Variable PetalLength is selected because its F statistic, 1180.161, is the largest among all variables.

Output 60.1.3. 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 in mm.      0.6187     119.26    <.0001       1.0000
      SepalWidth     Sepal Width in mm.       0.4008      49.16    <.0001       1.0000
      PetalLength    Petal Length in mm.      0.9414    1180.16    <.0001       1.0000
      PetalWidth     Petal Width in 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 variable PetalLength already in the model, PetalLength is tested for removal before selecting a new variable for entry. Since PetalLength meets the criterion to stay, it is used as a covariate in the analysis of covariance for variable selection. Variable SepalWidth is selected because its F statistic, 43.035, is the largest among all variables not in the model and its associated tolerance, 0.8164, meets the criterion to enter. The process is repeated in Steps 3 and 4. Variable PetalWidth is entered in Step 3, and variable SepalLength is entered in Step 4.

Output 60.1.4. 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 in mm.      0.9414    1180.16    <.0001

                                 No variables can be removed.


                              Statistics for Entry, DF = 2, 146

                                                    Partial
       Variable       Label                        R-Square    F Value    Pr > F    Tolerance

       SepalLength    Sepal Length in mm.            0.3198      34.32    <.0001       0.2400
       SepalWidth     Sepal Width in mm.             0.3709      43.04    <.0001       0.8164
       PetalWidth     Petal Width in 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 60.1.5. Iris Data: Stepwise Selection Step 3

```
                                    Fisher (1936) Iris Data

                                    The STEPDISC Procedure
                                  Stepwise Selection: Step 3

                              Statistics for Removal, DF = 2, 146

                                                     Partial
              Variable       Label                  R-Square    F Value    Pr > F

              SepalWidth     Sepal Width in mm.       0.3709      43.04    <.0001
              PetalLength    Petal Length in mm.      0.9384    1112.95    <.0001

                                 No variables can be removed.


                              Statistics for Entry, DF = 2, 145

                                              Partial
       Variable       Label                  R-Square    F Value    Pr > F    Tolerance

       SepalLength    Sepal Length in mm.      0.1447      12.27    <.0001       0.1323
       PetalWidth     Petal Width in 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 60.1.6. Iris Data: Stepwise Selection Step 4

```
                                    Fisher (1936) Iris Data

                                    The STEPDISC Procedure
                                  Stepwise Selection: Step 4

                              Statistics for Removal, DF = 2, 145

                                                     Partial
              Variable       Label                  R-Square    F Value    Pr > F

              SepalWidth     Sepal Width in mm.       0.4295      54.58    <.0001
              PetalLength    Petal Length in mm.      0.3482      38.72    <.0001
              PetalWidth     Petal Width in mm.       0.3229      34.57    <.0001

                                 No variables can be removed.


                              Statistics for Entry, DF = 2, 144

                                              Partial
       Variable       Label                  R-Square    F Value    Pr > F    Tolerance

       SepalLength    Sepal Length in 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 60.1.7. Iris Data: Stepwise Selection Step 5

```
                                    Fisher (1936) Iris Data

                                    The STEPDISC Procedure
                                  Stepwise Selection: Step 5

                              Statistics for Removal, DF = 2, 144

                                                     Partial
              Variable       Label                  R-Square    F Value    Pr > F

              SepalLength    Sepal Length in mm.      0.0615       4.72    0.0103
              SepalWidth     Sepal Width in mm.       0.2335      21.94    <.0001
              PetalLength    Petal Length in mm.      0.3308      35.59    <.0001
              PetalWidth     Petal Width in mm.       0.2570      24.90    <.0001

                                 No variables can be removed.

                                No further steps are possible.
```

Output 60.1.8. Iris Data: Stepwise Selection Summary

```
                                              Fisher (1936) Iris Data

                                               The STEPDISC Procedure

                                             Stepwise Selection Summary

                                                                                                      Average
                                                                                                      Squared
     Number                                              Partial                    Wilks’   Pr <   Canonical   Pr >
Step     In Entered     Removed     Label                R-Square F Value Pr > F     Lambda Lambda Correlation   ASCC

   1      1 PetalLength             Petal Length in mm.   0.9414 1180.16 <.0001 0.05862828 <.0001  0.47068586 <.0001
   2      2 SepalWidth              Sepal Width in mm.    0.3709   43.04 <.0001 0.03688411 <.0001  0.55995394 <.0001
   3      3 PetalWidth              Petal Width in mm.    0.3229   34.57 <.0001 0.02497554 <.0001  0.59495691 <.0001
   4      4 SepalLength             Sepal Length in mm.   0.0615    4.72 0.0103 0.02343863 <.0001  0.59594941 <.0001
```

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*, eds. K. Enslein, A. Ralston, and H. Wilf, New York: John Wiley & Sons, Inc.

Journal of Statistics Education, “Fish Catch Data Set,” [http://www.stat.ncsu.edu/info/jse], accessed 4 December 1997

Klecka, W.R. (1980), *Discriminant Analysis*, Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-019, Beverly Hills: Sage Publications.

Rencher, A.C. and Larson, S.F. (1980), “Bias in Wilks’s Λ in Stepwise Discriminant Analysis,” *Technometrics*, 22, 349–356.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.