Chapter 440

Discriminant Analysis

Introduction

Discriminant Analysis finds a set of prediction equations based on independent variables that are used to classify individuals into groups. There are two possible objectives in a discriminant analysis: finding a predictive equation for classifying new individuals or interpreting the predictive equation to better understand the relationships that may exist among the variables.

In many ways, discriminant analysis parallels multiple regression analysis. The main difference between these two techniques is that regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable. The methodology used to complete a discriminant analysis is similar to regression analysis. You plot each independent variable versus the group variable. You often go through a variable selection phase to determine which independent variables are beneficial. You conduct a residual analysis to determine the accuracy of the discriminant equations.

The mathematics of discriminant analysis are related very closely to the one-way MANOVA. In fact, the roles of the variables are simply reversed. The classification (factor) variable in the MANOVA becomes the dependent variable in discriminant analysis. The dependent variables in the MANOVA become the independent variables in the discriminant analysis.

Technical Details

Suppose you have data for K groups, with N_k observations per group. Let N represent the total number of observations. Each observation consists of the measurements of p variables. The i^{th} observation is represented by X_{ki}. Let M represent the vector of means of these variables across all groups and M_k the vector of means of observations in the k^{th} group.

Define three sums of squares and cross products matrices, S_T, S_W, and S_A, as follows

$$S_T = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(X_{ki} - M\right)\left(X_{ki} - M\right)'$$

$$S_W = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(X_{ki} - M_k\right)\left(X_{ki} - M_k\right)'$$

$$S_A = S_T - S_W$$

Next, define two degrees of freedom values, *df1* and *df2*:

$$df1 = K - 1$$

$$df2 = N - K$$

A discriminant function is a weighted average of the values of the independent variables. The weights are selected so that the resulting weighted average separates the observations into the groups. High values of the average come from one group, low values of the average come from another group. The problem reduces to one of finding the weights which, when applied to the data, best discriminate among groups according to some criterion. The

solution reduces to finding the eigenvectors, *V*, of $S_W^{-1}S_A$. The canonical coefficients are the elements of these eigenvectors.

A goodness-of-fit parameter, Wilks' lambda, is defined as follows:

$$\Lambda = \frac{|S_W|}{|S_T|} = \prod_{j=1}^{m} \frac{1}{1+\lambda_j}$$

where λ_j is the *jth* eigenvalue corresponding to the eigenvector described above and *m* is the minimum of *K-1* and *p*.

The canonical correlation between the j^{th} discriminant function and the independent variables is related to these eigenvalues as follows:

$$r_{c_j} = \sqrt{\frac{\lambda_j}{1+\lambda_j}}$$

Various other matrices are often considered during a discriminant analysis.

The overall covariance matrix, *T*, is given by:

$$T = \left(\frac{1}{N-1}\right) S_T$$

The within-group covariance matrix, *W*, is given by:

$$W = \left(\frac{1}{N-K}\right) S_W$$

The among-group (or between-group) covariance matrix, *A*, is given by:

$$A = \left(\frac{1}{K-1}\right) S_A$$

The linear discriminant functions are defined as:

$$LDF_k = W^{-1} M_k$$

The standardized canonical coefficients are given by:

$$v_{ij}\sqrt{w_{ij}}$$

where v_{ij} are the elements of *V* and w_{ij} are the elements of *W*.

The correlations between the independent variables and the canonical variates are given by:

$$Corr_{jk} = \frac{1}{\sqrt{w_{jj}}} \sum_{i=1}^{p} v_{ik} w_{ji}$$

Discriminant Analysis Checklist

Tabachnick (1989) provides the following checklist for conducting a discriminant analysis. We suggest that you consider these issues and guidelines carefully.

Unequal Group Size and Missing Data

You should begin by screening your data. Pay particular attention to patterns of missing values. When using discriminant analysis, you should have more observations per group than you have independent variables. If you do not, there is a good chance that your results cannot be generalized, and future classifications based on your analysis will be inaccurate.

Unequal group size does not influence the direct solution of the discriminant analysis problem. However, unequal group size can cause subtle changes during the classification phase. Normally, the sampling frequency of each group (the proportion of the total sample that belongs to a particular group) is used during the classification stage. If the relative group sample sizes are not representative of their sizes in the overall population, the classification procedure will be erroneous. (You can make appropriate adjustments to prevent these erroneous classifications by adjusting the prior probabilities.)

NCSS ignores rows with missing values. If it appears that most missing values occur in one or two variables, you might want to leave these out of the analysis in order to obtain more data and hence more accuracy.

Multivariate Normality and Outliers

Discriminant analysis does not make the strong normality assumptions that MANOVA does because the emphasis is on classification. A sample size of at least twenty observations in the smallest group is usually adequate to ensure robustness of any inferential tests that may be made.

Outliers can cause severe problems that even the robustness of discriminant analysis will not overcome. You should screen your data carefully for outliers using the various univariate and multivariate normality tests and plots to determine if the normality assumption is reasonable. *You should perform these tests on one group at a time.*

Homogeneity of Covariance Matrices

Discriminant analysis makes the assumption that the group covariance matrices are equal. This assumption may be tested with Box's M test in the Equality of Covariances procedure or looking for equal slopes in the Probability Plots. If the covariance matrices appear to be grossly different, you should take some corrective action. Although the inferential part of the analysis is robust, the classification of new individuals is not. These will tend to be classified into the groups with larger covariances. Corrective action usually includes the close screening for outliers and the use of variance-stabilizing transformations such as the logarithm.

Linearity

Discriminant analysis assumes linear relations among the independent variables. You should study scatter plots of each pair of independent variables, using a different color for each group. Look carefully for curvilinear patterns and for outliers. The occurrence of a curvilinear relationship will reduce the power and the discriminating ability of the discriminant equation.

Multicollinearity and Singularity

Multicollinearity occurs when one predictor variable is almost a weighted average of the others. This collinearity will only show up when the data are considered one group at a time. Forms of multicollinearity may show up when you have very small group sample sizes (when the number of observations is less than the number of variables). In this case, you must reduce the number of independent variables.

Multicollinearity is easily controlled for during the variable selection phase. You should only include variables that show an R^2 with other X's of less than 0.99.

See the chapter on Multiple Regression for a more complete discussion of multicollinearity.

Data Structure

The data given in the table below are the first eight rows (out of the 150 in the database) of the famous "iris data" published by Fisher (1936). These data are measurements in millimeters of sepal length, sepal width, petal length, and petal width of fifty plants for each of three varieties of iris: (1) Iris setosa, (2) Iris versicolor, and (3) Iris virginica. Note that Iris versicolor is a polyplid hybrid of the two other species. Iris setosa is a diploid species with 38 chromosomes, Iris virginica is a tetraploid, and Iris versicolor is a hexaploid with 108 chromosomes.

Discriminant analysis finds a set of prediction equations, based on sepal and petal measurements, that classify additional irises into one of these three varieties. Here Iris is the dependent variable, while SepalLength, SepalWidth, PetalLength, and PetalWidth are the independent variables.

Fisher dataset (subset)

SepalLength	SepalWidth	PetalLength	PetalWidth	Iris
50	33	14	2	1
64	28	56	22	3
65	28	46	15	2
67	31	56	24	3
63	28	51	15	3
46	34	14	3	1
69	31	51	23	3
62	22	45	15	2

Missing Values

If missing values are found in any of the independent variables being used, the row is omitted. If they occur only in the dependent (categorical) variable, the row is not used during the calculation of the prediction equations, but a predicted group (and scores) is calculated. This allows you to classify new observations.

Example 1 – Discriminant Analysis

This section presents an example of how to run a discriminant analysis. The data used are shown in the table above and found in the Fisher dataset.

Setup

To run this example, complete the following steps:

1 Open the Fisher example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Fisher** and click **OK**.

2 Specify the Discriminant Analysis procedure options

- Find and open the **Discriminant Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Y: Group Variable	**Iris**
X's: Independent Variables	**SepalLength-PetalWidth**
Reports Tab	
All Reports and Plots	**Checked** (Normally you would only view a few of these reports, but we are selecting them all so that we can document them.)
Report Options (*in the Toolbar*)	
Variable Labels	**Column Labels**
Data Labels	**Value Labels**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Group Means Report

Group Means

Variable	Iris Setosa	Versicolor	Virginica	Overall
Sepal Length	50.06	59.36	65.88	58.43333
Sepal Width	34.28	27.7	29.74	30.57333
Petal Length	14.62	42.6	55.52	37.58
Petal Width	2.46	13.26	20.26	11.99333
Count	50	50	50	150

This report shows the means of each of the independent variables across each of the groups. The last row shows the count (number of observations) in the group. Note that the column headings come from the use of value labels for the group variable.

Group Standard Deviations Report

Group Standard Deviations

Variable	Iris Setosa	Versicolor	Virginica	Overall
Sepal Length	3.524897	5.161712	6.358796	8.280662
Sepal Width	3.790644	3.137983	3.224966	4.358663
Petal Length	1.73664	4.69911	5.518947	17.65298
Petal Width	1.053856	1.977527	2.7465	7.622377
Count	50	50	50	150

This report shows the standard deviations of each of the independent variables across each of the groups. The last row shows the count or number of observations in the group.

Discriminant analysis makes the assumption that the covariance matrices are identical for each of the groups. This report lets you glance at the standard deviations to check if they are about equal.

Total Correlation\Covariance Report

Total Correlation\Covariance

Variable	Variable Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	68.56935	-4.243401	127.4315	51.62707
Sepal Width	-0.117570	18.99794	-32.96564	-12.16394
Petal Length	0.871754	-0.428440	311.6278	129.5609
Petal Width	0.817941	-0.366126	0.962865	58.10063

This report shows the correlation and covariance matrices that are formed when the grouping variable is ignored. Note that the correlations are on the lower left and the covariances are on the upper right. The variances are on the diagonal.

Between-Group Correlation\Covariance Report

Between-Group Correlation\Covariance

Variable	Variable Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	3160.607	-997.6334	8262.42	3563.967
Sepal Width	-0.745075	567.2466	-2861.98	-1146.633
Petal Length	0.994135	-0.812838	21855.14	9338.7
Petal Width	0.999768	-0.759258	0.996232	4020.667

This report displays the correlations and covariances formed using the group means as the individual observations. The correlations are shown in the lower-left half of the matrix. The within-group covariances are shown on the diagonal and in the upper-right half of the matrix. Note that if there are only two groups, all correlations will be equal to one since they are formed from only two rows (the two group means).

Within-Group Correlation\Covariance Report

Within-Group Correlation\Covariance

Variable	Variable Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	26.50082	9.272109	16.75143	3.840136
Sepal Width	0.530236	11.53878	5.524354	3.27102
Petal Length	0.756164	0.377916	18.51878	4.266531
Petal Width	0.364506	0.470535	0.484459	4.188163

This report shows the correlations and covariances that would be obtained from data in which the group means had been subtracted. The correlations are shown in the lower-left half of the matrix. The within-group covariances are shown on the diagonal and in the upper-right half of the matrix.

Variable Influence Report

Variable Influence Section

Variable	Removed Lambda	Removed F-Value	Removed F-Prob	Alone Lambda	Alone F-Value	Alone F-Prob	R-Squared Other X's
Sepal Length	0.938463	4.72	0.010329	0.381294	119.26	0.000000	0.858612
Sepal Width	0.766480	21.94	0.000000	0.599217	49.16	0.000000	0.524007
Petal Length	0.669206	35.59	0.000000	0.058628	1180.16	0.000000	0.968012
Petal Width	0.743001	24.90	0.000000	0.071117	960.01	0.000000	0.937850

This report analyzes the influence of each of the independent variables on the discriminant analysis.

Variable

The name of the independent variable.

Removed Lambda

This is the value of a Wilks' lambda computed to test the impact of removing this variable.

Removed F-Value

This is the F-ratio that is used to test the significance of the above Wilks' lambda.

Removed F-Prob

This is the probability (significance level) of the above F-ratio. It is the probability to the right of the F-ratio. The test is significant (the variable is important) if this value is less than the value of alpha that you are using, such as 0.05.

Alone Lambda

This is the value of a Wilks' lambda that would be obtained if this were the only independent variable used.

Alone F-Value

This is an F-ratio that is used to test the significance of the above Wilks' lambda.

Alone F-Prob

This is the probability (significance level) of the above F-ratio. It is the probability to the right of the F-ratio. The test is significant (the variable is important) if this value is less than the value of alpha that you are using, such as 0.05.

R-Squared OtherX's

This is the R-Squared value that would be obtained if this variable were regressed on all other independent variables. When this R-Squared value is larger than 0.99, severe multicollinearity problems exist. You should remove variables (one at a time) with large R-Squared and rerun your analysis.

Linear Discriminant Functions Report

Linear Discriminant Functions

Variable	Iris Setosa	Versicolor	Virginica
Constant	-85.20985	-71.754	-103.2697
Sepal Length	2.354417	1.569821	1.244585
Sepal Width	2.358787	0.707251	0.3685279
Petal Length	-1.643064	0.5211451	1.276654
Petal Width	-1.739841	0.6434229	2.107911

This report presents the linear discriminant function coefficients. These are often called the discriminant coefficients. They are also known as the “plug-in” estimators, since the true variance-covariance matrices are required but their estimates are plugged-in. This technique assumes that the independent variables in each group follow a multivariate-normal distribution with equal variance-covariance matrices across groups. Studies have shown that this technique is fairly robust to departures from either assumption.

The report represents three classification functions, one for each of the three groups. Each function is represented vertically. When a weighted average of the independent variables is formed using these coefficients as the weights (and adding the constant), the discriminant scores result. To determine which group an individual belongs to, select the group with the highest score.

Regression Coefficients Report

Regression Coefficients

Variable	Iris Setosa	Versicolor	Virginica
Constant	0.1182229	1.577059	-0.6952819
Sepal Length	0.006602977	-0.002015369	-0.004587608
Sepal Width	0.02428479	-0.04456162	0.02027684
Petal Length	-0.02246571	0.02206692	0.0003987911
Petal Width	-0.005747273	-0.04943066	0.05517793

This report presents the regression coefficients. These coefficients are determined as follows:

1. Create three indicator variables, one for each of the three varieties of iris. Each indicator variable is set to one when the row belongs to that group and zero otherwise.
2. Fit a multiple regression of the independent variables on each of the three indicator variables.
3. The regression coefficients obtained are those shown in this table.

Hence, predicted values generated by these coefficients will be between zero and one. To determine which group an individual belongs to, select the group with the highest score.

Classification Count Table Report

Classification Count Table for Iris

Actual	Predicted Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
Versicolor	0	48	2	50
Virginica	0	1	49	50
Total	50	49	51	150

Reduction in classification error due to X's = 97.0%

This report presents a matrix that indicates how accurately the current discriminant functions classify the observations. If perfect classification has been achieved, there will be zeros on the off-diagonals. The rows of the table represent the actual groups, while the columns represent the predicted group.

Percent Reduction

The percent reduction is the classification accuracy achieved by the current discriminant functions over what is expected if the observations were randomly classified. The formula for the Reduction in classification error is

[Sum of diagonals – N/k] / [N – N/k].

Misclassified Rows Report

Misclassified Rows Section

Row	Actual	Predicted	Percent Chance of Each Group Pcnt1	Pcnt2	Pcnt3
5	Virginica	Versicolo	0.0	72.9	27.1
9	Versicolo	Virginica	0.0	25.3	74.7
12	Versicolo	Virginica	0.0	14.3	85.7

This report shows the actual group and the predicted group of each observation that was misclassified. It also shows 100 times the estimated probability, P(i), that the row is in each group. For easier viewing, we have multiplied the probabilities by 100 to make this a percent probability (between 0 and 100) rather than a regular probability (between 0 and 1). A value near 100 gives a strong indication that the observation belongs in that group.

P(i)

If the linear discriminant classification technique was used, these are the estimated probabilities that this row belongs to the i^{th} group. See James (1985), page 69, for details of the algorithm used to estimate these probabilities. This algorithm is briefly outlined here.

Let f_i (i = 1, 2, ..., K) be the linear discriminant function value. Let $max(f_k)$ be the maximum score of all groups. Let $P(G_i)$ be the overall probability of classifying an individual into group i. The values of $P(i)$ are generated using the following equation:

$$P(i) = \frac{\exp[f_i - \max(f_k)]P(G_i)}{\sum_{j=1}^{K} \exp[f_j - \max(f_k)]P(G_j)}$$

If the regression classification technique was used, this is the predicted value of the regression equation. The implicit Y value in the regression equation is one or zero, depending on whether this observation is in the i^{th} group or not. Hence, a predicted value near zero indicates that the observation is not in the i^{th} group, while a value near one indicates a strong possibility that this observation is in the i^{th} group. There is nothing to prevent these predicted values from being greater than one or less than zero. They are not estimated probabilities.

You can store these values for further analysis by listing variables in the appropriate *Storage Tab* options.

Predicted Classification Report

Predicted Classification Section

Row	Actual	Predicted	Percent Chance of Each Group Pcnt1	Pcnt2	Pcnt3
1	Setosa	Setosa	100.0	0.0	0.0
2	Virginica	Virginica	0.0	0.0	100.0
3	Versicolo	Versicolo	0.0	99.6	0.4
4	Virginica	Virginica	0.0	0.0	100.0
5	Virginica	Versicolo	0.0	72.9	27.1
6	Setosa	Setosa	100.0	0.0	0.0
7	Virginica	Virginica	0.0	0.0	100.0
8	Versicolo	Versicolo	0.0	96.0	4.0

(report continues for all 150 rows)

This report shows the actual group, the predicted group, and the percentage probabilities of each row. The definitions are given above in the *Misclassified Rows Report*.

Canonical Variate Analysis Report

Canonical Variate Analysis Section

Fn	Inv(W)B Eigenvalue	Ind'l Pcnt	Total Pcnt	Canon Corr	Canon Corr2	F-Value	Numer DF	Denom DF	Prob Level	Wilks' Lambda
1	32.191929	99.1	99.1	0.9848	0.9699	199.1	8.0	288.0	0.0000	0.023439
2	0.285391	0.9	100.0	0.4712	0.2220	13.8	3.0	145.0	0.0000	0.777973

The F-value tests whether this function and those below it are significant.

This report provides a canonical correlation analysis of the discriminant problem. Recall that canonical correlation analysis is used when you want to study the correlation between two sets of variables. In this case, the two sets of variables are defined in the following way. The independent variables comprise the first set. The group variable defines another set, which is generated by creating an indicator variable for each group except the last one.

Inv(W)B Eigenvalue

The eigenvalues of the matrix $W^{-1}B$. These values indicate how much of the total variation explained is accounted for by the various discriminant functions. Hence, the first discriminant function corresponds to the first eigenvalue, and so on. Note that the number of eigenvalues is the minimum of the number of variables and *K-1*, where *K* is the number of groups.

Ind'l Prcnt

The percent that this eigenvalue is of the total.

Total Prcnt

The cumulative percent of this and all previous eigenvalues.

Canon Corr

The canonical correlation coefficient.

Canon Corr2

The square of the canonical correlation. This is similar to R-Squared in multiple regression.

F-Value

The value of the approximate F-ratio for testing the significance of the Wilks' lambda corresponding to this row and those below it. Hence, in this example, the first F-value tests the significance of both the first and second canonical correlations, while the second F-value tests the significance of the second correlation only.

Num DF

The numerator degrees of freedom for this F-test.

Denom DF

The denominator degrees of freedom for this F-test.

Prob Level

The significance level of the F-test. This is the area under the F-distribution to the right of the F-value. Usually, a value less than 0.05 is considered significant.

Wilks' Lambda

The value of Wilks' lambda for this row. This Wilks' lambda is used to test the significance of the discriminant function corresponding to this row and those below it. Recall that Wilks' lambda is a multivariate generalization of R^2. The above F-value is an approximate test of this Wilks' lambda.

Canonical Coefficients Report

Canonical Coefficients

	Canonical Variate	
Variable	**Variate1**	**Variate2**
Constant	-2.105106	6.661473
Sepal Length	-0.082938	-0.002410
Sepal Width	-0.153447	-0.216452
Petal Length	0.220121	0.093192
Petal Width	0.281046	-0.283919

This report gives the coefficients used to create the canonical scores. The canonical scores are weighted averages of the observations, and these coefficients are the weights (with the constant term added).

Canonical Variates at Group Means Report

Canonical Variates at Group Means

	Canonical Function	
Iris	**Function1**	**Function2**
Setosa	-7.6076	-0.215133
Versicolor	1.82505	0.7278996
Virginica	5.78255	-0.5127666

This report gives the results of applying the canonical coefficients to the means of each of the groups.

Std. Canonical Coefficients Report

Std. Canonical Coefficients

	Canonical Variate	
Variable	**Variate1**	**Variate2**
Sepal Length	-0.426955	-0.012408
Sepal Width	-0.521242	-0.735261
Petal Length	0.947257	0.401038
Petal Width	0.575161	-0.581040

This report gives the standardized canonical coefficients.

Variable-Variate Correlations Report

Variable-Variate Correlations

	Canonical Variate	
Variable	**Variate1**	**Variate2**
Sepal Length	0.222596	-0.310812
Sepal Width	-0.119012	-0.863681
Petal Length	0.706065	-0.167701
Petal Width	0.633178	-0.737242

This report gives the loadings (correlations) of the variables on the canonical variates. That is, each entry is the correlation between the canonical variate and the independent variable. This report can help you interpret a particular canonical variate.

Linear Discriminant Scores Report

Linear Discriminant Scores

Row	Iris	Score1	Score2	Score3
1	Setosa	83.86837	38.65921	-6.790054
2	Virginica	1.230765	91.857	104.5692
3	Versicolo	32.19471	83.71141	78.29187
4	Virginica	11.89069	99.97506	113.6244
5	Virginica	19.27056	83.17749	82.18597
6	Setosa	75.06965	33.7306	-9.291955
7	Virginica	26.55469	99.86555	107.6224

(report continues for all 150 rows)

This report gives the individual values of the linear discriminant scores. Note that this information may be stored on the database using the Data Storage options.

Regression Scores Report

Regression Scores

Row	Iris	Score1	Score2	Score3
1	Setosa	0.923755	0.215832	-0.139588
2	Virginica	-0.163732	0.348623	0.815109
3	Versicolo	0.107759	0.471953	0.420288
4	Virginica	-0.082564	0.110031	0.972533
5	Virginica	-0.017776	0.586318	0.431458
6	Setosa	0.915881	0.129902	-0.045782
7	Virginica	0.048718	0.045096	0.906186

(report continues for all 150 rows)

This report gives the individual values of the predicted scores based on the regression coefficients. Even though these values are predicting indicator variables, it is possible for a value to be less than zero or greater than one. Note that this information may be stored on the database using the *Data Storage* options.

Canonical Scores Report

Canonical Scores

Row	Iris	Score1	Score2
1	Setosa	-7.671967	0.134894
2	Virginica	6.800150	-0.580895
3	Versicolo	2.548678	0.472205
4	Virginica	6.653087	-1.805320
5	Virginica	3.815160	0.942986
6	Setosa	-7.212618	-0.355836
7	Virginica	5.105559	-1.992182

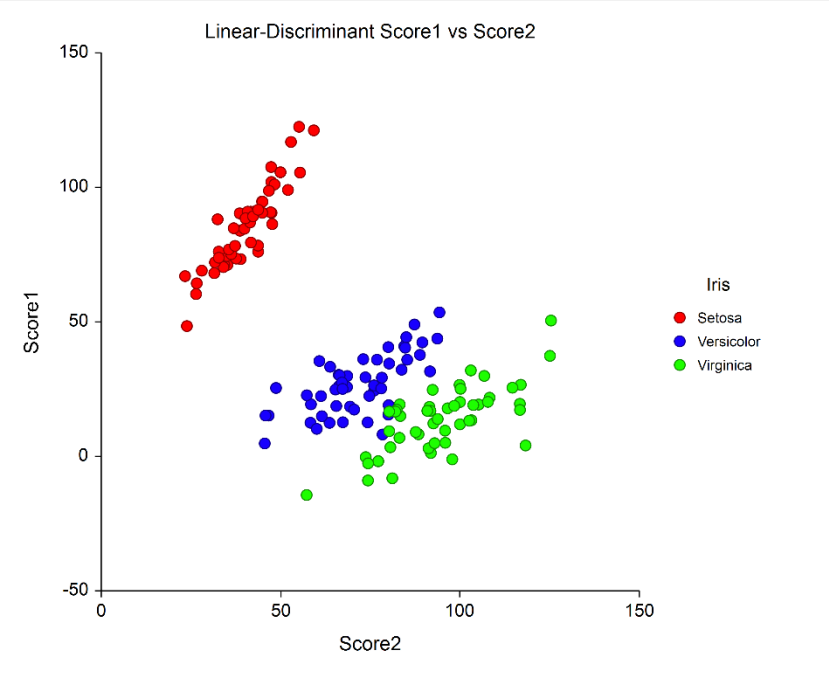
(report continues for all 150 rows)

This report gives the scores of the canonical variates for each row. Note that this information may be stored on the database using the Data Storage options.

Scores Plot(s)

You may select plots of the linear discriminant scores, regression scores, or canonical scores to aid in your interpretation. These plots are usually used to give a visual impression of how well the discriminant functions are classifying the data. (Several charts are displayed on the output. Only one of these is displayed here.)

Linear-Discriminant Scores Plots



This chart plots the values of the first and second linear discriminant scores. By looking at this plot you can see what the classification rule would be. Also, it is obvious from this plot that the first two linear-discriminant functions are necessary in discriminating among the varieties of iris since the groups can be separated along diagonal lines.

Example 2 – Automatic Variable Selection (Brief Report)

The tutorial we have just concluded was based on all four of the independent variables. A common task in discriminant analysis is variable selection. Often you have a large pool of possible independent variables from which you want to select a smaller set (up to about eight variables) which will do almost as well at discriminating as the complete set. **NCSS** provides an automatic procedure for doing this, which will be described next.

The automatic variable selection is run by changing the Variable Selection option to Stepwise. The program will conduct a stepwise variable selection. It will first find the best discriminator and then the second best. After it has found two, it checks whether the discrimination would be almost as good if one were removed. This stepping process of adding the best remaining variable and then checking if one of the active variables could be removed continues until no new variable can be found whose F-value has a probability smaller than the Probability Enter value.

An alternative procedure is to use the Multivariate Variable Selection procedure described elsewhere in this manual. If you have more than two groups, you must create a set of dummy (indicator) variables, one for each group. You ignore the last dummy variable, so if there are *K* groups, you analyze *K-1* dummy variables. The Multivariate Variable Selection program will always find a subset of your independent variables that is at least as good (and usually better) as the stepwise procedure described in this section. Once a subset of independent variables has been found, they can then be analyzed using the Discriminant Analysis program described here.

Once the variable selection has been made, the program provides the reports that were described in the previous tutorial. Note that two report formats may be called for during the variable selection phase: brief and verbose. We will now provide an example of each type of report.

Setup

To run this example, complete the following steps:

1 Open the Fisher example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Fisher** and click **OK**.

2 Specify the Discriminant Analysis procedure options

- Find and open the **Discriminant Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Y: Group Variable	**Iris**
X's: Independent Variables	**SepalLength-PetalWidth**
Variable Selection	**Stepwise**
Reports Tab	
All Reports and Plots	**Unchecked** (We will only view the Variable Selection Report.)
Report Format	**Brief**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Labels**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Variable-Selection Summary Report

Variable-Selection Summary Section

Iteration	Action This Step	Independent Variable	Pct Chg In Lambda	F-Value	Prob Level	Wilks' Lambda
0	None					1.000000
1	Entered	Petal Length	94.14	1180.16	0.000000	0.058628
2	Entered	Sepal Width	37.09	43.04	0.000000	0.036884
3	Entered	Petal Width	32.29	34.57	0.000000	0.024976
4	Entered	Sepal Length	6.15	4.72	0.010329	0.023439

This report shows what action was taken at each step.

Iteration

This gives the number of this step.

Action This Step

This tells what action (if any) was taken during this step. “Entered” means that the variable was entered into the set of active variables. “Removed” means that the variable was removed from the set of active variables.

Pct Chg In Lambda

This is the percentage decrease in lambda that resulted from this step. Note that Wilks’ lambda is analogous to 1 - R-Squared in multiple regression. Hence, we want to *decrease* Wilks’ lambda to improve our model. For example, going from iteration 2 to iteration 3 results in lambda decreasing from 0 .036884 to 0.024976. This is a 32.29% decrease in lambda.

F-Value

This is the F-ratio for testing the significance of this variable. If the variable was “Entered,” this tests the hypothesis that the variable should be added. If the variable was “Removed,” this tests whether the variable should be removed.

Prob Level

The significance level of the above F-Value.

Wilks’ Lambda

The multivariate extension of R-Squared. Wilks’ lambda reduces to 1-(R-Squared) in the two-group case. It is interpreted just backwards from R-Squared. It varies from one to zero. Values near one imply low predictability, while values close to zero imply high predictability. Note that this Wilks’ lambda value corresponds to the currently active variables.

Example 3 – Automatic Variable Selection (Verbose Report)

We will now rerun this example with the "verbose" option.

Setup

To run this example, complete the following steps:

1 Open the Fisher example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Fisher** and click **OK**.

2 Specify the Discriminant Analysis procedure options

- Find and open the **Discriminant Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Y: Group Variable	**Iris**
X's: Independent Variables	**SepalLength-PetalWidth**
Variable Selection	**Stepwise**
Reports Tab	
All Reports and Plots	**Unchecked** (We will only view the Variable Selection Report.)
Report Format	**Verbose**
Variable Names	**Labels**
Value Labels	**Value Labels**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Labels**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Variable-Selection Detail Report

Variable-Selection Detail Section - Step 0

Status	Independent Variable	Pct Chg In Lambda	F-Value	Prob Level	R-Squared Other X's
Out	Sepal Length	61.87	119.26	0.000000	0.000000
Out	Sepal Width	40.08	49.16	0.000000	0.000000
Out	Petal Length	94.14	1180.16	0.000000	0.000000
Out	Petal Width	92.89	960.01	0.000000	0.000000

Overall Wilks' Lambda = 1.000000
Action this step: None

Variable-Selection Detail Section - Step 1

Status	Independent Variable	Pct Chg In Lambda	F-Value	Prob Level	R-Squared Other X's
In	Petal Length	94.14	1180.16	0.000000	0.000000
Out	Sepal Length	31.98	34.32	0.000000	0.759955
Out	Sepal Width	37.09	43.04	0.000000	0.183561
Out	Petal Width	25.33	24.77	0.000000	0.927110

Overall Wilks' Lambda = 0.058628
Action this step: Petal Length Entered

Variable-Selection Detail Section - Step 2

Status	Independent Variable	Pct Chg In Lambda	F-Value	Prob Level	R-Squared Other X's
In	Sepal Width	37.09	43.04	0.000000	0.183561
In	Petal Length	93.84	1112.95	0.000000	0.183561
Out	Sepal Length	14.47	12.27	0.000012	0.840178
Out	Petal Width	32.29	34.57	0.000000	0.929747

Overall Wilks' Lambda = 0.036884
Action this step: Sepal Width Entered

Variable-Selection Detail Section - Step 3

Status	Independent Variable	Pct Chg In Lambda	F-Value	Prob Level	R-Squared Other X's
In	Sepal Width	42.95	54.58	0.000000	0.213103
In	Petal Length	34.82	38.72	0.000000	0.933764
In	Petal Width	32.29	34.57	0.000000	0.929747
Out	Sepal Length	6.15	4.72	0.010329	0.858612

Overall Wilks' Lambda = 0.024976
Action this step: Petal Width Entered

Variable-Selection Detail Section - Step 4

Status	Independent Variable	Pct Chg In Lambda	F-Value	Prob Level	R-Squared Other X's
In	Sepal Length	6.15	4.72	0.010329	0.858612
In	Sepal Width	23.35	21.94	0.000000	0.524007
In	Petal Length	33.08	35.59	0.000000	0.968012
In	Petal Width	25.70	24.90	0.000000	0.937850

Overall Wilks' Lambda = 0.023439
Action this step: Sepal Length Entered

The details are shown for each step.

Step

This gives the number of this step (iteration).

Status

This tells whether the variable is “in” or “out” of the set of active variables.

Pct Chg In Lambda

This is the percentage decrease in lambda that would result if the status of this variable were reversed.

F-Value

This is the F-ratio for testing the significance of changing the status of this variable.

Prob Level

The significance level of the above F-Value.

R-Squared Other X’s

This is the R-Squared that would result if this variable were regressed on the other independent variables that are active (status = “In”). This provides a check for multicollinearity in the active independent variables.

Overall Wilks’ Lambda

This is the value of Wilks’ lambda for all active independent variables. A value near zero indicates an accurate model; a value near one indicates a poor model.