SPSS Discriminant Function Analysis

By Hui Bian

Office for Faculty Excellence

Discriminant Function Analysis

- What is Discriminant function analysis
  - It builds a predictive model for group membership
  - The model is composed of a discriminant function based on linear combinations of predictor variables.
  - Those predictor variables provide the best discrimination between groups.

Discriminant Function Analysis

- Purpose of Discriminant analysis
  - to *maximally* separate the groups.
  - to determine the most parsimonious way to separate groups
  - to discard variables which are little related to group distinctions

Discriminant Function Analysis

- Summary: we are interested in the relationship between a group of independent variables and one categorical variable. We would like to know how many dimensions we would need to express this relationship. Using this relationship, we can predict a classification based on the independent variables or assess how well the independent variables separate the categories in the classification.

Discriminant Function Analysis

- It is similar to regression analysis
- A *discriminant score* can be calculated based on the weighted combination of the independent variables
  - $D_i = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$
  - D_i is predicted score (discriminant score)
  - x is predictor and b is discriminant coefficient
- We use maximum likelihood technique to assign a case to a group from a specified cut-off score.
  - If group size is equal, the cut-off is mean score.
  - If group size is not equal, the cut-off is calculated from weighted means.

Discriminant Function Analysis

- Grouping variables
  - Categorical variables
  - Can have more than two values
  - The codes for the grouping variables must be integers
- Independent variables
  - Continuous
  - Nominal variables must be recoded to dummy variables

Discriminant Function Analysis

- Discriminant function
  - A latent variable of a linear combination of independent variables
  - One discriminant function for 2-group discriminant analysis
  - For higher order discriminant analysis, the number of discriminant function is equal to g-1 (g is the number of categories of dependent/grouping variable).
  - The first function maximizes the difference between the values of the dependent variable.
  - The second function maximizes the difference between the values of the dependent variable while controlling the first function.
  - And so on.

Discriminant Function Analysis

- The first function will be the most powerful differentiating dimension.
- The second and later functions may also represent additional significant dimensions of differentiation.

Discriminant Function Analysis

- Assumptions (from SPSS 19.0 help)
  - Cases should be independent.
  - Predictor variables should have a multivariate normal distribution, and within-group variance-covariance matrices should be equal across groups.
  - Group membership is assumed to be mutually exclusive
  - The procedure is most effective when group membership is a truly categorical variable; if group membership is based on values of a continuous variable (for example, high IQ versus low IQ), consider using linear regression to take advantage of the richer information that is offered by the continuous variable itself.

Discriminant Function Analysis

- Assumptions(similar to those for linear regression)
  - Linearity, normality, multilinearity, equal variances
  - Predictor variables should have a multivariate normal distribution.
  - fairly robust to violations of the most of these assumptions. But highly sensitive to outliers.
- Model specification

Discriminant Function Analysis

- Test of significance
  - For two groups, the null hypothesis is that the means of the two groups on the discriminant function-the centroids, are equal.
  - Centroids are the mean discriminant score for each group.
  - Wilk's lambda is used to test for significant differences between groups.
  - Wilk's lambda is between 0 and 1. It tells us the variance of dependent variable that is not explained by the discriminant function.

Discriminant Function Analysis

- Wilk's lambda is also used to test for significant differences between the groups on the individual predictor variables.
- It tells which variables contribute a significant amount of prediction to help separate the groups.

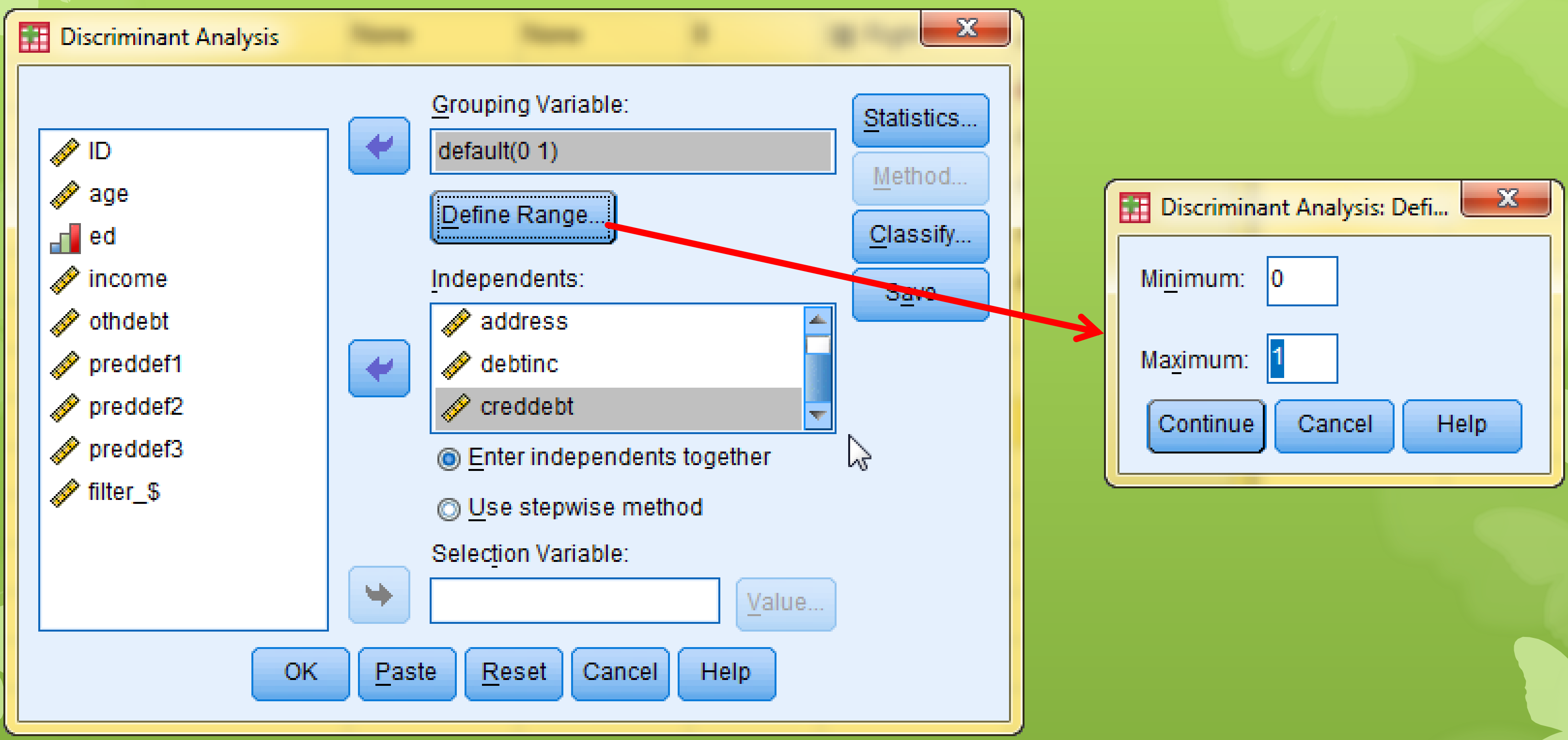
Discriminant Function Analysis

- Two groups using an example from SPSS manual
  - Example: the purpose of this example is to identify characteristics that are indicative of people who are likely to default on loans, and use those characteristics to identify good and bad credit risks.
  - Sample includes a total of 850 cases (old and new/future customers) The first 700 cases are customers who were previously given loans.
  - Use first 700 customers to create a discriminant analysis model, setting the remaining 150 customers aside to validate the analysis.
  - Then use the model to classify the 150 prospective customers as good or bad credit risks.

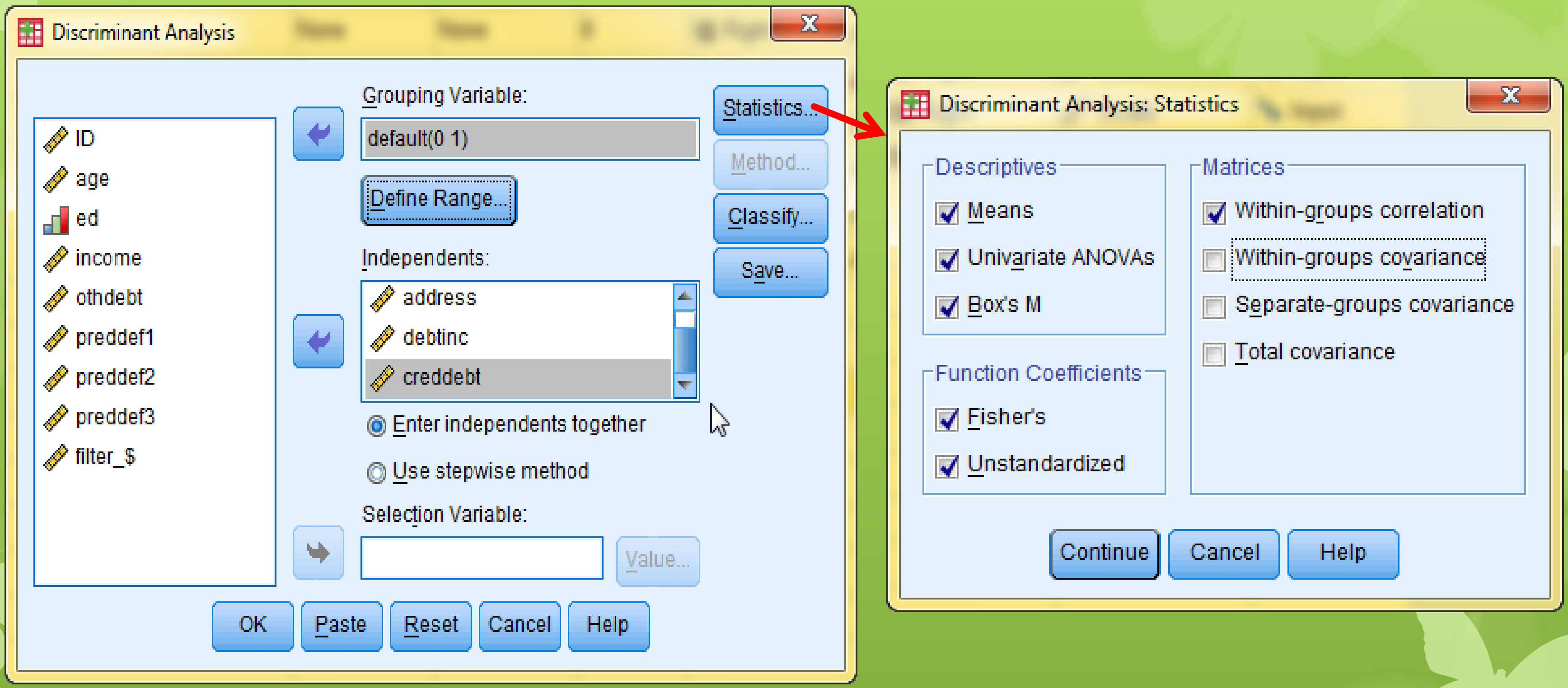
Discriminant Function Analysis

- Grouping variable: *Default*
- Predictors: *employ, address, debtinc, and creaddebt*
- Obtain Discriminant function analysis
  - Analyze > Classify > Discriminant

Discriminant Function Analysis

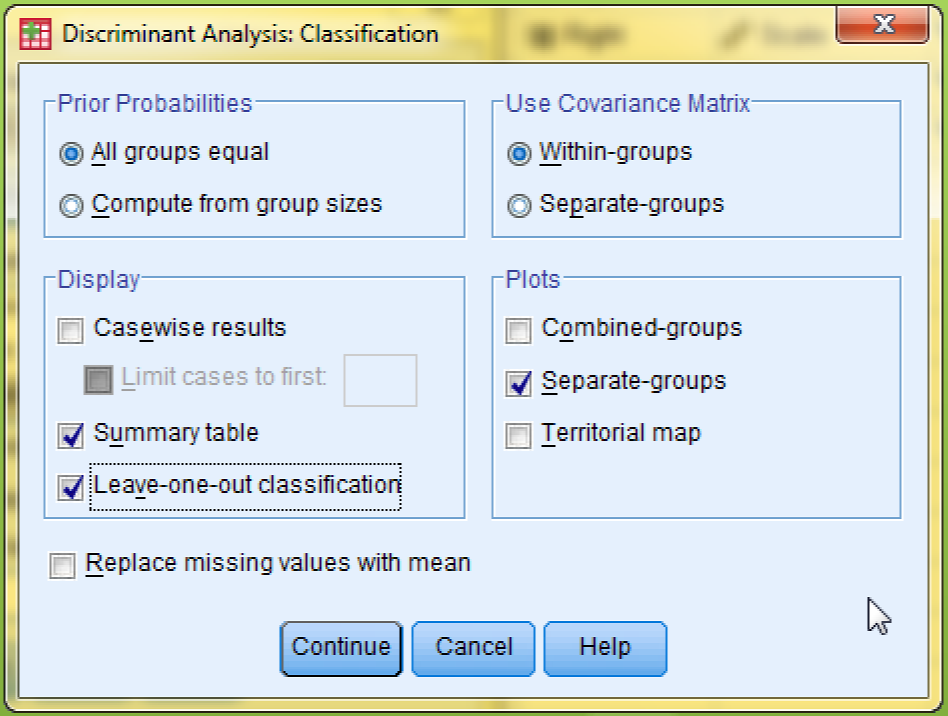


Discriminant Function Analysis



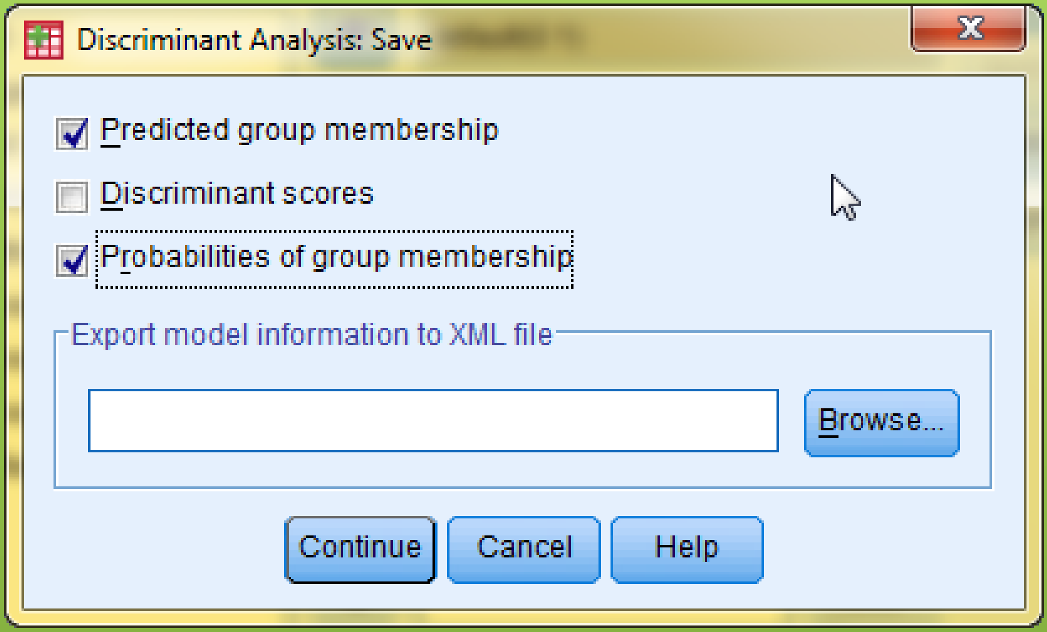
Discriminant Function Analysis

- Click *Classify* to get this window



Discriminant Function Analysis

- Click *Save* to get this window



Discriminant Function Analysis

- SPSS Output: descriptive statistics

Group Statistics

default Previously defaulted		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
0 No	employ Years with current employer	9.5087	6.66374	517	517.000
	address Years at current address	8.9458	7.00062	517	517.000
	debtinc Debt to income ratio (x100)	8.6793	5.61520	517	517.000
	creddebt Credit card debt in thousands	1.2455	1.42231	517	517.000
1 Yes	employ Years with current employer	5.2240	5.54295	183	183.000
	address Years at current address	6.3934	5.92521	183	183.000
	debtinc Debt to income ratio (x100)	14.7279	7.90280	183	183.000
	creddebt Credit card debt in thousands	2.4239	3.23252	183	183.000
Total	employ Years with current employer	8.3886	6.65804	700	700.000
	address Years at current address	8.2786	6.82488	700	700.000
	debtinc Debt to income ratio (x100)	10.2606	6.82723	700	700.000
	creddebt Credit card debt in thousands	1.5536	2.11720	700	700.000

Discriminant Function Analysis

- SPSS output: ANOVA table

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
employ Years with current employer	.920	60.759	1	698	.000
address Years at current address	.973	19.402	1	698	.000
debtinc Debt to income ratio (x100)	.848	124.889	1	698	.000
creddebt Credit card debt in thousands	.940	44.472	1	698	.000

In the ANOVA table, the smaller the Wilks's lambda, the more important the independent variable to the discriminant function. Wilks's lambda is significant by the F test for all independent variables.

Discriminant Function Analysis

- SPSS Output (correlation matrix)

Pooled Within-Groups Matrices

		employ Years with current employer	address Years at current address	debtinc Debt to income ratio (x100)	creddebt Credit card debt in thousands
Correlation	employ Years with current employer	1.000	.292	.089	.509
	address Years at current address	.292	1.000	.083	.260
	debtinc Debt to income ratio (x100)	.089	.083	1.000	.455
	creddebt Credit card debt in thousands	.509	.260	.455	1.000

The within-groups correlation matrix shows the correlations between the predictors.

Discriminant Function Analysis

- SPSS output: test of homogeneity of covariance matrices

Box's Test of Equality of Covariance Matrices

Log Determinants

default Previously defaulted	Rank	Log Determinant
0 No	4	11.156
1 Yes	4	12.270
Pooled within-groups	4	11.970

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		364.962
F	Approx.	36.182
	df1	10
	df2	552413.774
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

The larger the log determinant in the table, the more that group's covariance matrix differs. The "Rank" column indicates the number of independent variables in this case. Since discriminant analysis assumes homogeneity of covariance matrices between groups, we would like to see the determinants be relatively equal.

Discriminant Function Analysis

- SPSS output: test of homogeneity of covariance matrices

Box's Test of Equality of Covariance Matrices

Log Determinants

default Previously defaulted	Rank	Log Determinant
0 No	4	11.156
1 Yes	4	12.270
Pooled within-groups	4	11.970

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		364.962
F	Approx.	36.182
	df1	10
	df2	552413.774
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

1. Box's M test tests the assumption of homogeneity of covariance matrices. This test is very sensitive to meeting the assumption of multivariate normality.
2. Discriminant function analysis is **robust** even when the homogeneity of variances assumption is not met, provided the data do not contain important outliers.
3. For our data, we conclude the groups do differ in their covariance matrices, violating an assumption of DA.
4. when n is large, small deviations from homogeneity will be found significant, which is why Box's M must be interpreted in conjunction with inspection of the log determinants.

Discriminant Function Analysis

- SPSS output: test of homogeneity of covariance matrices

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.395[a]	100.0	100.0	.532

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.717	231.524	4	.000

1. The larger the eigenvalue, the more of the variance in the dependent variable is explained by that function.
2. Dependent has two categories, there is only one discriminant function.
3. The canonical correlation is the measure of association between the discriminant function and the dependent variable.
4. The square of canonical correlation coefficient is the percentage of variance explained in the dependent variable.

Discriminant Function Analysis

- SPSS output: summary of canonical discriminant functions

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.395[a]	100.0	100.0	.532

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.717	231.524	4	.000

When there are two groups, the canonical correlation is the most useful measure in the table, and it is equivalent to Pearson's correlation between the discriminant scores and the groups.
Wilks' lambda is a measure of how well each function separates cases into groups. Smaller values of Wilks' lambda indicate greater discriminatory ability of the function.

The associated chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The small significance value indicates that the discriminant function does better than chance at separating the groups.

Discriminant Function Analysis

- SPSS output: summary of canonical discriminant functions

Standardized Canonical Discriminant Function Coefficients

	Function
	1
employ Years with current employer	-.766
address Years at current address	-.248
debtinc Debt to income ratio (x100)	.470
creddebt Credit card debt in thousands	.642

The standardized discriminant function coefficients in the table serve the same purpose as *beta weights in multiple regression (partial coefficient)* : they indicate the relative importance of the independent variables in predicting the dependent. They allow you to compare variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability.

Discriminant Function Analysis

- SPSS output: summary of canonical discriminant functions

Structure Matrix

	Function
	1
debtinc Debt to income ratio (x100)	.673
employ Years with current employer	-.470
creddebt Credit card debt in thousands	.402
address Years at current address	-.265

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

1.The structure matrix table shows the correlations of each variable with each discriminant function.
2. Only one discriminant function is in this study.
3. The correlations then serve like factor loadings in factor analysis -- that is, by identifying the largest absolute correlations associated with each discriminant function the researcher gains insight into how to name each function.

Discriminant Function Analysis

- SPSS output: summary of canonical discriminant functions

Structure Matrix

	Function
	1
debtinc Debt to income ratio (x100)	.673
employ Years with current employer	-.470
creddebt Credit card debt in thousands	.402
address Years at current address	-.265

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

1. Discriminant function is a latent variable that is created as a linear combination of independent variables.
2. Discriminating variables are independent variables.
3. The table shows the Pearson correlations between predictors and standardized canonical discriminant functions.
4. Loading < .30 may be removed from the model.

Discriminant Function Analysis

- SPSS output: summary of canonical discriminant functions

Canonical Discriminant Function Coefficients

	Function
	1
employ Years with current employer	-.120
address Years at current address	-.037
debtinc Debt to income ratio (x100)	.075
creddebt Credit card debt in thousands	.312
(Constant)	.058

Unstandardized coefficients

This table contains the unstandardized discriminant function coefficients. These would be used like unstandardized b (regression) coefficients in multiple regression -- that is, they are used to construct the actual prediction equation which can be used to classify new cases.

Discriminant Function Analysis

- Discriminant function: Our model should be like this:

  $D_i = 0.058 - 0.12emplo - 0.037addres + 0.075debin + 0.312\ credebt$

Discriminant Function Analysis

- SPSS output: summary of canonical discriminant functions

Functions at Group Centroids

default Previously defaulted	Function
	1
0 No	-.373
1 Yes	1.054

Unstandardized canonical discriminant functions evaluated at group means

Centroids are the mean discriminant scores for each group. This table is used to establish the *cutting point* for classifying cases. If the two groups are of equal size, the best cutting point is half way between the values of the functions at group centroids (that is, the average). If the groups are unequal, the optimal cutting point is the weighted average of the two values. The computer does the classification automatically, so these values are for informational purposes.

Discriminant Function Analysis

- The centroids are calculated based on the function: D_i = 0.058 – 0.12emplo – 0.037addres + 0.075debin + 0.312 credebt
- Centroids are discriminant score for each group when the variable means (rather than individual values for each case) are entered into the function.

Discriminant Function Analysis

- SPSS Output: Classification Statistics

Classification Processing Summary

Processed		700
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		700

Prior Probabilities for Groups

default Previously defaulted	Prior	Cases Used in Analysis	
		Unweighted	Weighted
0 No	.500	517	517.000
1 Yes	.500	183	183.000
Total	1.000	700	700.000

Prior Probabilities are used in classification. The default is using observed group sizes. In your sample to determine the prior probabilities of membership in the groups formed by the dependent, and this is necessary if you have different group sizes. If each group is of the same size, as an alternative you could specify equal prior probabilities for all groups.

Discriminant Function Analysis

- SPSS Output: Classification Statistics

Classification Function Coefficients

	default Previously ...	
	0 No	1 Yes
employ Years with current employer	.267	.096
address Years at current address	.153	.100
debtinc Debt to income ratio (x100)	.277	.384
creddebt Credit card debt in thousands	-.643	-.197
(Constant)	-3.446	-3.850

Fisher's linear discriminant functions

Two sets (one for each dependent group) of unstandardized linear discriminant coefficients are calculated, which can be used to classify cases. This is the classical method of classification, though now little used.

Discriminant Function Analysis

- SPSS Output: Classification Statistics

Classification Results[b,c]

		default Previously defaulted	Predicted Group Membership: 0 No	Predicted Group Membership: 1 Yes	Total
Original	Count	0 No	391	126	517
		1 Yes	42	141	183
	%	0 No	75.6	24.4	100.0
		1 Yes	23.0	77.0	100.0
Cross-validated[a]	Count	0 No	391	126	517
		1 Yes	43	140	183
	%	0 No	75.6	24.4	100.0
		1 Yes	23.5	76.5	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

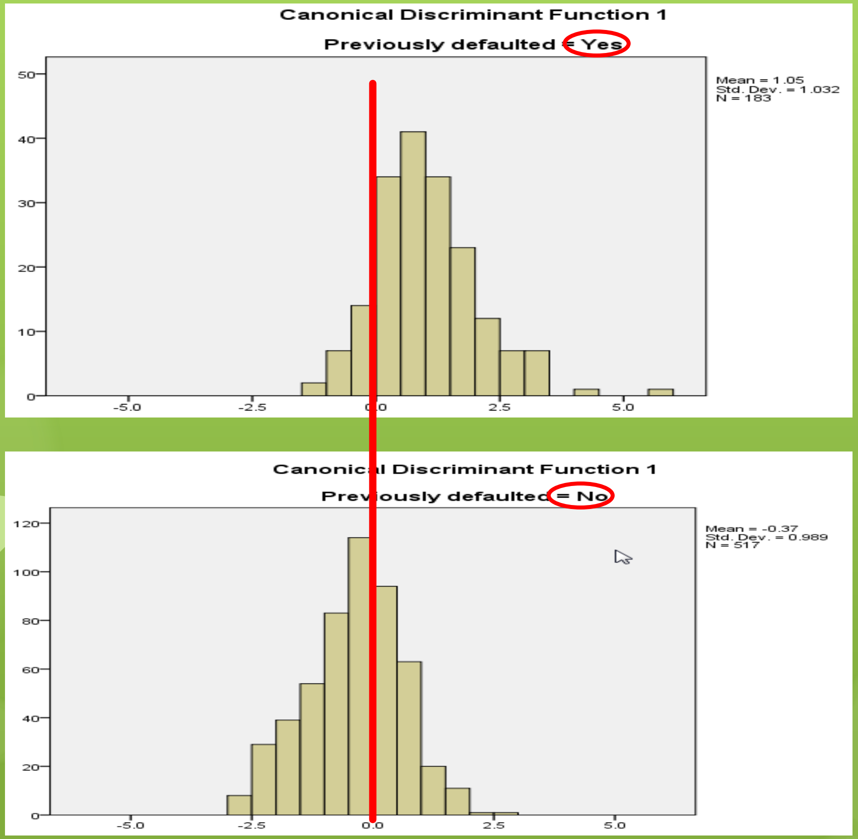
b. 76.0% of original grouped cases correctly classified.

c. 75.9% of cross-validated grouped cases correctly classified.

This table is used to assess how well the discriminant function works, and if it works equally well for each group of the dependent variable. Here it correctly classifies more than 75% of the cases, making about the same proportion of mistakes for both categories. Overall, 76.0% of the cases are correctly classified.

Discriminant Function Analysis

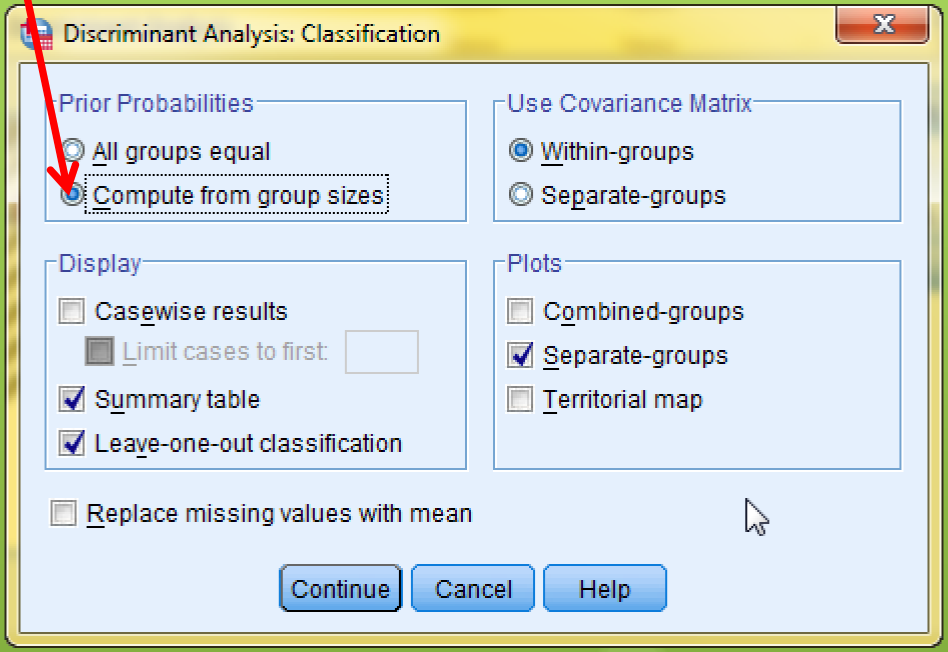
- SPSS Output: separate-group plots



If two distributions overlap too much, it means they do not discriminate too (poor discriminant function).

Discriminant Function Analysis

- Run DA



Discriminant Function Analysis

- We get this classification table

Classification Results[b,c]

		default Previously defaulted	Predicted Group Membership 0 No	1 Yes	Total
Original	Count	0 No	487	30	517
		1 Yes	100	83	183
		Ungrouped cases	129	21	150
	%	0 No	94.2	5.8	100.0
		1 Yes	54.6	45.4	100.0
		Ungrouped cases	86.0	14.0	100.0
Cross-validated[a]	Count	0 No	487	30	517
		1 Yes	101	82	183
	%	0 No	94.2	5.8	100.0
		1 Yes	55.2	44.8	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b. 81.4% of original grouped cases correctly classified.
c. 81.3% of cross-validated grouped cases correctly classified.

Sensitivity = 45.4%; specificity = 94.2%

Discriminant Function Analysis

- The table from previous slide shows how accurately the customers were classified into these groups.
  - Sensitivity: highly sensitive test means that there are few false negative results (Type II error)
  - Specificity: highly specific test means that there few false positive results (Type I error).

Discriminant Function Analysis

- Discriminant methods
  - Enter all independent variables into the equation at once
  - Stepwise: remove independent variables that are not significant.

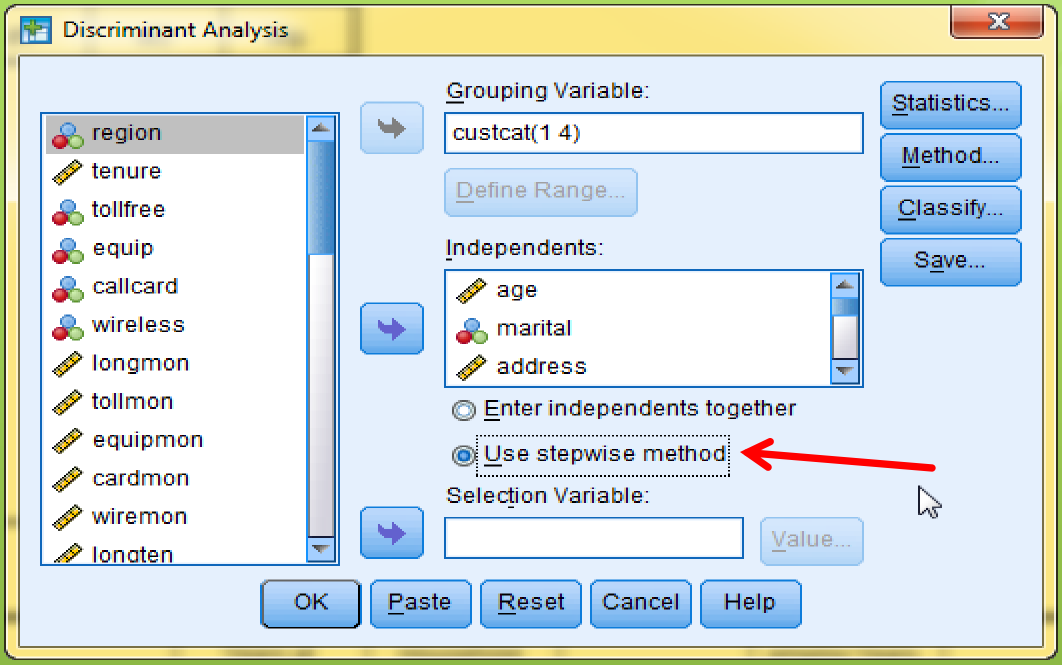
Discriminant Function Analysis

- Discriminant Function Analysis (more than two Groups)
- Example from SPSS mannual.
  - A telecommunications provider has segmented its customer base by service usage patterns, categorizing the customers into four groups. If demographic data can be used to predict group membership, you can customize offers for individual prospective customers.

Discriminant Function Analysis

- Variables in the analysis
  - Dependent variable *custcat* (four categories)
  - Independent variables: demographics
- Obtain discriminant function analysis
  - Analyze > Classify > Discriminant

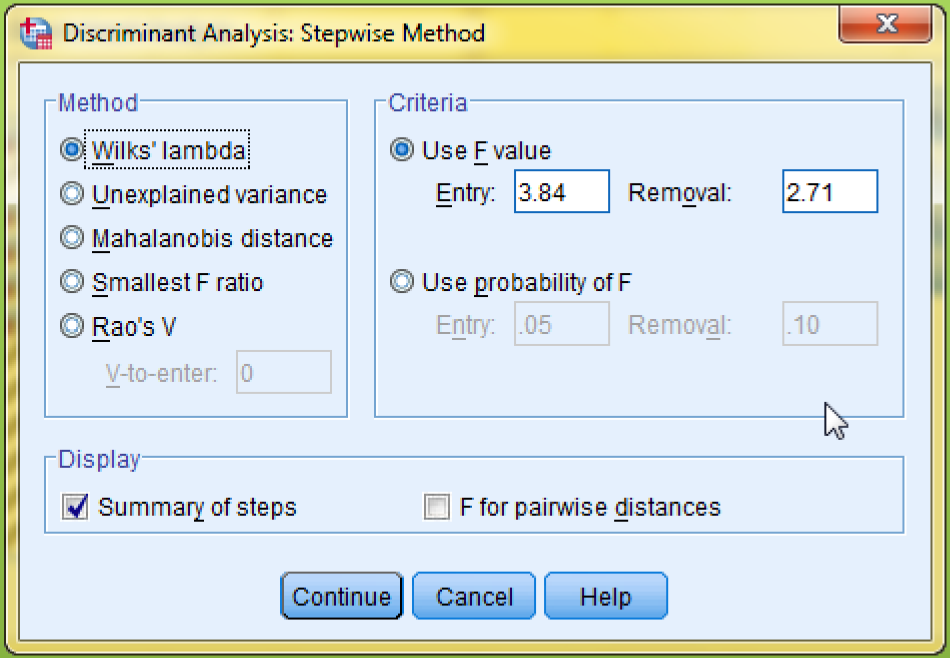
Discriminant Function Analysis



When you have a lot of predictors, the stepwise method can be useful by automatically selecting the "best" variables to use in the model.

Discriminant Function Analysis

- Click Method



Discriminant Function Analysis

- Method
  - Wilks' lambda. A variable selection method for stepwise discriminant analysis that chooses variables for entry into the equation on the basis of how much they lower Wilks' lambda. At each step, the variable that minimizes the overall Wilks' lambda is entered.
  - Unexplained variance. At each step, the variable that minimizes the sum of the unexplained variation between groups is entered.
  - Mahalanobis distance. A measure of how much a case's values on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables.

Discriminant Function Analysis

- Method
  - Smallest F ratio. A method of variable selection in stepwise analysis based on maximizing an F ratio computed from the Mahalanobis distance between groups.
  - Rao's V. A measure of the differences between group means. Also called the Lawley-Hotelling trace. At each step, the variable that maximizes the increase in Rao's V is entered. After selecting this option, enter the minimum value a variable must have to enter the analysis.

Discriminant Function Analysis

- Use F value. A variable is entered into the model if its F value is greater than the *Entry value* and is removed if the F value is less than the *Removal value*. Entry must be greater than Removal, and both values must be positive. To enter more variables into the model, lower the Entry value. To remove more variables from the model, increase the Removal value.
- Use probability of F. A variable is entered into the model if the significance level of its F value is less than the Entry value and is removed if the significance level is greater than the Removal value. Entry must be less than Removal, and both values must be positive. To enter more variables into the model, increase the Entry value. To remove more variables from the model, lower the Removal value.

Discriminant Function Analysis

- SPSS output
  - The stepwise method starts with a model that doesn't include any of the predictors (step 0).
  - At each step, the predictor with the largest F to Enter, value that exceeds the entry criteria (by default, 3.84) is added to the model.

Discriminant Function Analysis

- SPSS output

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
age Age in years	.978	7.521	3	996	.000
marital Marital status	.990	3.500	3	996	.015
address Years at current address	.975	8.433	3	996	.000
income Household income in thousands	.980	6.689	3	996	.000
ed Level of education	.844	61.454	3	996	.000
employ Years with current employer	.951	16.976	3	996	.000
retire Retired	.991	3.005	3	996	.030
gender Gender	.999	.373	3	996	.772
reside Number of people in household	.988	3.976	3	996	.008

Discriminant Function Analysis

- SPSS output: variables in the analysis

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	ed Level of education	1.000	61.454	
2	ed Level of education	.953	59.108	.951
	employ Years with current employer	.953	14.933	.844
3	ed Level of education	.951	60.046	.940
	employ Years with current employer	.934	15.824	.834
	reside Number of people in household	.979	4.841	.807

Tolerance is the proportion of a variable's variance *not accounted for* by *other independent variables* in the equation. A variable with very low tolerance contributes little information to a model and can cause computational problems.
Actually, tolerance is about multicollinearity. <.40 is worthy of concern. <.10 is problematic.

Discriminant Function Analysis

- SPSS output: variables in the analysis

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	ed Level of education	1.000	61.454	
2	ed Level of education	.953	59.108	.951
	employ Years with current employer	.953	14.933	.844
3	ed Level of education	.951	60.046	.940
	employ Years with current employer	.934	15.824	.834
	reside Number of people in household	.979	4.841	.807

F to Remove values are useful for describing what happens if a variable is removed from the current model (given that the other variables remain).

Discriminant Function Analysis

- SPSS output: Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.198[a]	80.2	80.2	.407
2	.048[a]	19.4	99.6	.214
3	.001[a]	.4	100.0	.031

a. First 3 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.796	227.345	9	.000
2 through 3	.953	47.486	4	.000
3	.999	.929	1	.335

Nearly all of the variance explained by the model is due to the first two discriminant functions. We can ignore the third function. For each set of functions, this tests the hypothesis that the means of the functions listed are equal across groups. The test of function 3 has a p value of .34, so this function contributes little to the model.

Discriminant Function Analysis

- SPSS output

The structure matrix table shows the correlations of each variable with each discriminant function. The correlations serve like factor loadings in factor analysis -- that is, by identifying the largest absolute correlations associated with each discriminant function.

Structure Matrix

	Function		
	1	2	3
ed Level of education	.966*	-.090	-.244
employ Years with current employer	-.182	.964*	-.193
age Age in years[a]	-.162	.598*	-.285
income Household income in thousands[a]	.109	.514*	-.190
address Years at current address[a]	-.151	.394*	-.214
retire Retired[a]	-.108	.230*	-.137
gender Gender[a]	.008	.054*	.009
reside Number of people in household	.232	.097	.968*
marital Marital status[a]	.132	.134	.600*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

Discriminant Function Analysis

- Three discriminant functions

Canonical Discriminant Function Coefficients

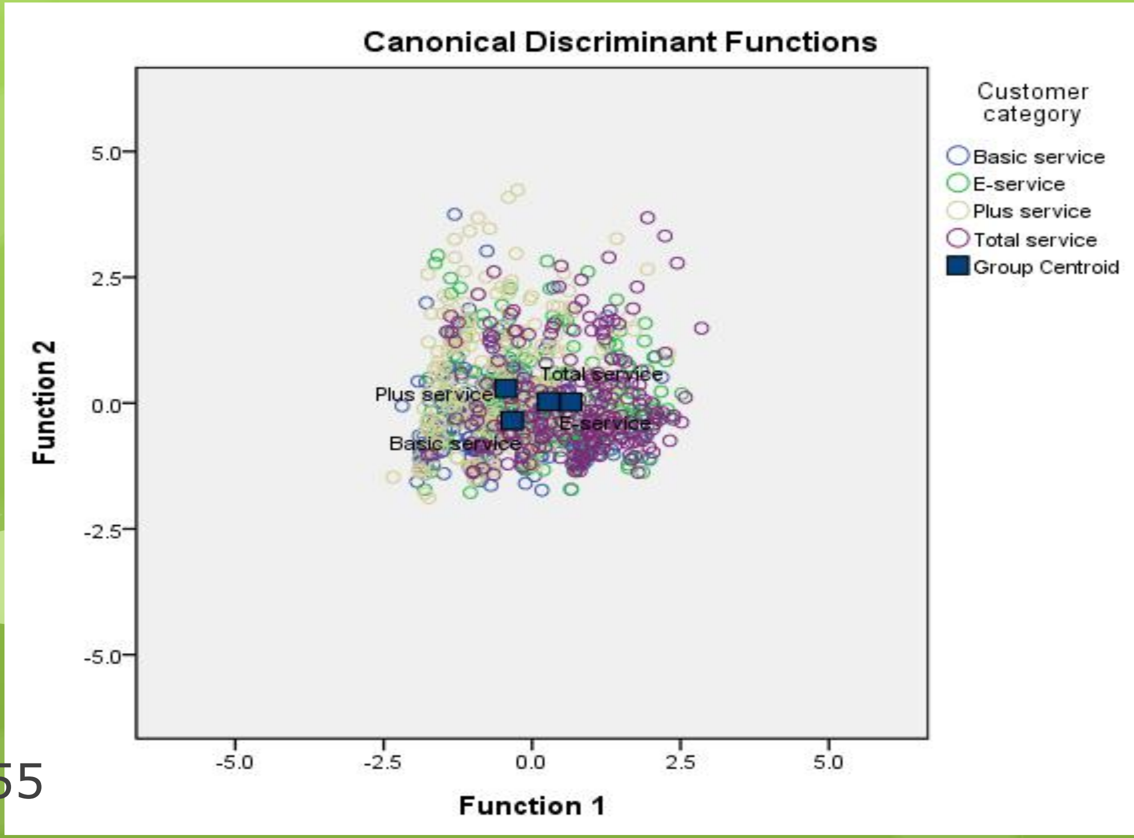
	Function		
	1	2	3
ed Level of education	.876	.121	-.222
employ Years with current employer	.007	.104	-.012
reside Number of people in household	.183	.167	.662
(Constant)	-2.839	-1.858	-.818

Unstandardized coefficients

Function 1: -2.84 + .88ed + .01em + .18re
Function 2: -1.86 + .12ed + .10em + .17re
Function 3: -.82 - .22ed -.01em + .66re

Discriminant Function Analysis

- SPSS Output: combined-group plot



The closer the group centroids, the more errors of classification likely will be.

Discriminant Function Analysis

- SPSS output: classification table

Classification Results[a]

		custcat Customer category	Predicted Group Membership: 1 Basic service	2 E-service	3 Plus service	4 Total service	Total
Original	Count	1 Basic service	125	11	61	69	266
		2 E-service	49	15	58	95	217
		3 Plus service	102	14	112	53	281
		4 Total service	40	16	37	143	236
	%	1 Basic service	47.0	4.1	22.9	25.9	100.0
		2 E-service	22.6	6.9	26.7	43.8	100.0
		3 Plus service	36.3	5.0	39.9	18.9	100.0
		4 Total service	16.9	6.8	15.7	60.6	100.0

a. 39.5% of original grouped cases correctly classified.

The model excels at identifying Total service customers. However, it does an exceptionally poor job of classifying E-service customers. You may need to find another predictor in order to separate these customers.

Discriminant Function Analysis

- We have created a discriminant model that classifies customers into one of four predefined "service usage" groups, based on demographic information from each customer. Using the structure matrix, we identified which variables are most useful for segmenting the customer base. Lastly, the classification results show that the model does poorly at classifying E-service customers. If identifying E-service customers is not the concern, the model may be accurate enough for this purpose.

Discriminant Function Analysis

Thank you