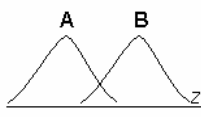


DSC 410/510 Multivariate Statistical Methods Discriminant Analysis


- ### Applications
- Identify the group to which an object or case (e.g. person, firm, product) belongs:
 - ◆ predict the success or failure of a new product
 - ◆ decide whether a student should be admitted to graduate school
 - ◆ classify students as to vocational interests
 - ◆ determine the category of credit risk for a person
 - ◆ predict whether a firm will be successful

- ### What is Discriminant Analysis
- Dependent (response) variable is categorical (nominal or nonmetric), e.g. defining two groups
 - Independent (predictor) variables are metric
 - Empirically derive a variate or discriminant function (linear combination of the predictor variables) that discriminates best between pre-defined groups
 - ◆ calculate discriminant (Z) function scores for each individual (e.g. $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$)
 - ◆ maximize between-group variance of Z relative to within-group variance of Z

Discriminating Two Groups

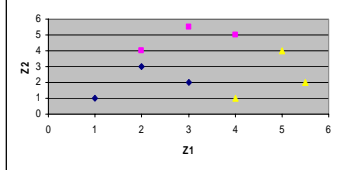


- Little overlap in the discriminant function
 - ◆ High between-group variance relative to within-group variance
- Greater overlap in the discriminant function
 - ◆ Low between-group variance relative to within-group variance

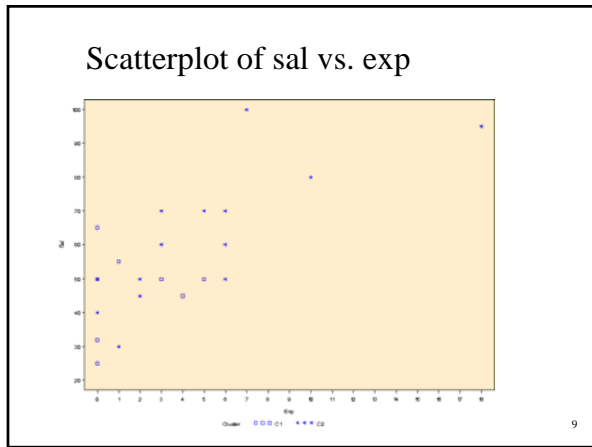
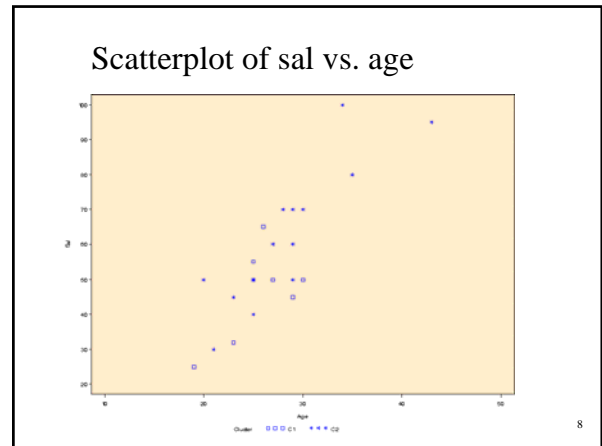
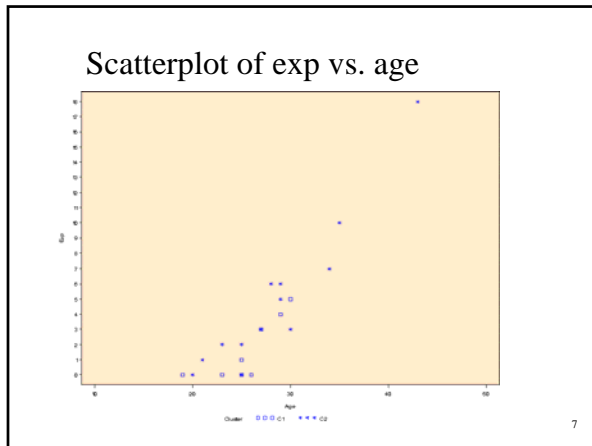


Discriminating More Than Two Groups

- Derive $K - 1$ variates, $K =$ number of groups
- E.g. three groups, two discriminant Z functions
 - ◆ can be represented on a scatterplot:



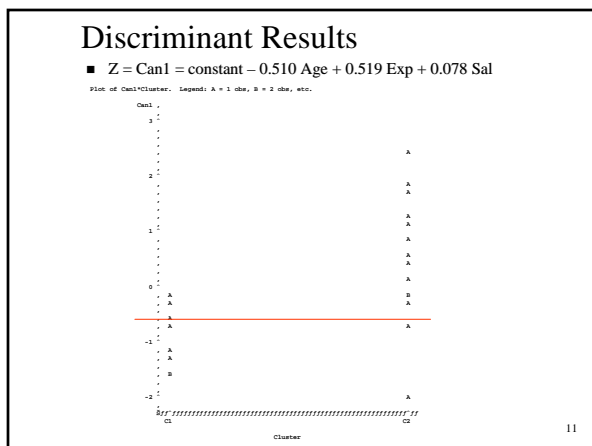
- ### Two Group Discriminant Analysis Example
- Goal: identify cluster of DSC 410/510 students
 - Response: cluster (found in last chapter)
 - Predictors: age, years of work experience (exp), expected salary (sal)
 - Data: 22 students (**discrim.xls**)
 - Question: Can we form a linear combination of age, exp, and sal to discriminate one cluster from the other?



SAS Instructions

- Use SAS procedure “discrim”
- Code: see `discrim1.sas`
- Class variable: cluster
- Predictor variables: age, exp, sal
- Option “canonical” runs a canonical analysis (which is the type we’ll focus on)
- Option “list” displays results for everyone
- Statement “priors proportional” assumes sample sizes reflect population sizes

10



Classification Probabilities

- An alternative to classifying objects using Z is calculating (posterior) probabilities of belonging to each group
 - ◆ classify each object into the group with the highest probability
 - ◆ displayed in SAS output when “list” option used
 - ◆ used to produce classification matrix:

		into	
		C1	C2
from	C1	5	3
	C2	2	12

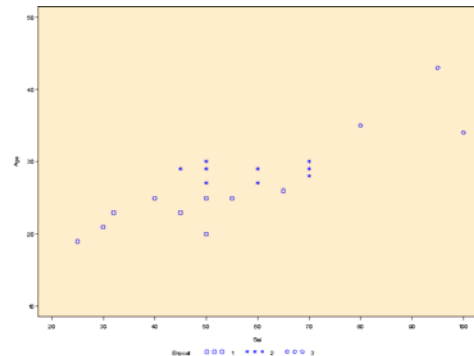
12

Three Group Discriminant Analysis Example

- Goal: identify years of work experience category for DSC 410/510 students
- Response: low (0-2), medium (3-6) or high
- Predictors: age, expected salary (sal)
- Data: 22 students (**discrim.xls**)
- Question: Can we form two linear combinations of age and sal to discriminate students in low (1), medium (2) and high (3) work experience categories?

13

Scatterplot



14

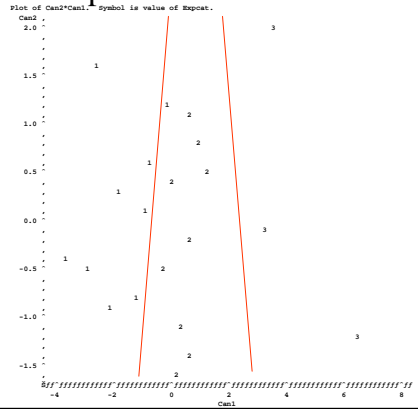
Discriminant Results

- SAS code: see **discrim2.sas**
- $Z_1 = \text{Can1} = \text{constant} + 0.334 \text{ Age} + 0.031 \text{ Salary}$
- $Z_2 = \text{Can2} = \text{constant} - 0.302 \text{ Age} + 0.092 \text{ Salary}$
- Correctly classifies all but one:

		into		
		1	2	3
from	1	9	1	0
	2	0	9	0
	3	0	0	3

15

Scatterplot of Discriminant Functions



16

Discriminant Analysis Objectives

- Determine significant differences between groups on a set of variables
 - ◆ discriminant analysis provides an objective assessment of differences
- Determine which variables account the most for the differences
 - ◆ discriminant analysis lends insight into the role of individual variables
- Classify new objects (individuals, firms, products) into groups based on variable values

17

Selection of Variables

- Dependent variable must be categorical, with two or more mutually exclusive and exhaustive categories
 - ◆ can be created from a metric variable, e.g. work experience category (low, medium, high)
- “Polar Extremes” approach compares only the two most extreme categories
- Independent (predictor) variables can be based on previous research, economic theory, intuition, etc.

18

Sample Considerations

- Aim for 5-20 observations per predictor variable
- Aim for at least 20 observations per group
 - ◆ minimum group size is the number of predictor variables
- Aim for roughly equal group sizes
- Consider dividing the sample in two (50-50, 60-40, or 75-25) if possible (total size > 100):
 - ◆ analysis sample for developing model
 - ◆ holdout sample for validating the model

19

Discriminant Assumptions

- Multivariate normality of predictor variables
 - ◆ logistic regression is a possible alternative if this assumption is violated
- Each group has same variance-covariance matrix
 - ◆ Separate-groups covariance matrices or “quadratic” techniques can be used for classification if this assumption is violated
- All relationships are linear
- No multicollinearity (remove redundant predictors)
- No outliers (remove aberrant cases)

20

Examples

- Two-group Illustrative Example
 - ◆ page 281-296
 - ◆ `hatco.xls` data file
 - ◆ `hatco2group.sas` SAS file
- Three-group Illustrative Example
 - ◆ page 296-314
 - ◆ `hatco.xls` data file
 - ◆ `hatco3group.sas` SAS file

21

Computational Method

- Simultaneous (direct) estimation
 - ◆ all predictor variables are considered concurrently, regardless of discriminating power
- Stepwise estimation
 - ◆ find single best discriminating variable
 - ◆ find variable best able to improve discrimination when paired with first variable, etc....
 - ◆ until remaining predictors do not contribute significantly to further discrimination

22

Statistical Significance

- Various criteria available for assessing discriminant function(s), e.g.
 - ◆ Wilks’ lambda indicates whether group means are significantly different (small: yes, large: no)
 - ◆ Mahalanobis D^2 indicates (generalized) Euclidean distance between groups
- Lead to chi-squared and F hypothesis tests:
 - ◆ used to select predictors in stepwise estimation
 - ◆ used to select number of discriminant functions to retain (up to $K - 1$)
 - ◆ used to assess differences between all groups

23

Discriminant Functions

- Discriminant function coefficients
 - ◆ Unstandardized: used in calculating Z scores
 - ◆ Standardized: used in interpreting relative contribution of predictors to the function(s)
- Fisher’s linear discriminant functions
 - ◆ One for each group
 - ◆ For every case, calculate scores for each group
 - ◆ Used for classifying cases: assign case to group with the highest classification function score
 - ◆ Equivalent to assigning case to group with highest posterior probability

24

Discriminant Function Results

- The discriminant analysis explains a % of the variation in the response variable
- Each discriminant function accounts for a proportion of this
- Proportion of variation explained by 1st function is given by its (canonical correlation)²
- Proportion of *remaining* variation explained by 2nd function given by its (canonical correlation)²
- Etc... until we can finally sum to get the total % of the variation explained by the analysis
 - ◆ $c_1^2 + c_2^2(1 - c_1^2) + c_3^2(1 - c_1^2 - c_2^2(1 - c_1^2)) + \dots$

25

Graphical Representations

- Group membership scatterplot
 - ◆ scatterplot of first two (unstandardized) discriminant functions, with cases identified by symbols indicating group, and centroids marked
- Territorial map
 - ◆ scatterplot of first two (unstandardized) discriminant functions, with classification boundaries and group centroids marked
- Combination! (see p309)

26

Classification

- Cutting scores based on (unstandardized) discriminant function centroids (see p265)
 - ◆ $Z_{cut} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B}$
 - ◆ assign case to A if $Z < Z_{cut}$, to B otherwise
 - ◆ only works with two groups
- Fisher's linear discriminant functions
 - ◆ for every case, calculate scores for each group
 - ◆ assign case to group with the highest classification function score
- Posterior probabilities
 - ◆ assign case to group with the highest posterior probability

27

Classification Matrices

		Predicted Group				
		Actual	1	2	3	Total
Analysis	1					
	2					
	3					
Cross-validated	1					
	2					
	3					
Holdout	1					
	2					
	3					

28

Prediction Accuracy

- Hit ratio = % correctly classified
- Should be at least 25% higher than the larger of:
 - ◆ maximum chance criterion = $\max\left\{\frac{\# \text{ in group}}{\text{sample size}}\right\}$
 - ◆ proportional chance
 - critterion = $p_1^2 + p_2^2 + \dots + p_{NG}^2$
- Press's Q statistic should be "significantly high"
 - ◆ $Q = \frac{[N - nK]^2}{N(K - 1)}$
 - ◆ compare with chi-squared ($df = 1$) critical value

29

Casewise Diagnostics

- Identify misclassified cases
- Profile the correctly classified and misclassified cases on the predictor variables
- Identify significant differences using two-sample t-tests
- Relate to graphical representations if possible
- Patterns in misclassification may indicate the possibility of additional group(s)

30

Interpret Results

1. Assess relative contribution of predictors to the function(s): tabulate and/or plot
 - ◆ standardized discriminant function coefficients
 - ◆ better to use discriminant loadings: correlations between predictor variables and discriminant function(s)
 - ◆ loadings may be “rotated” to aid interpretation
2. Potency index = $\sum_{\text{functions } j} \text{loading}_{ij}^2 \times \text{relative eigenvalue } j$
3. Compare 1 and 2 with univariate F -ratios for each predictor to check consistency

31

Validate Results

- Assess prediction accuracy on holdout sample, or, failing that, using “cross-validation” (“leave-one-out prediction”)
- Profile groups on the predictor variables
 - ◆ assess whether results conform to prior expectations, theory, common sense, etc.
- Profile groups on additional variables expected to reflect differences between groups

32