Multiple Discriminant Analysis

DR. HEMAL PANDYA

Introduction

- Discriminant Analysis is a dependence technique.
- Discriminant Analysis is used to predict group membership.
- This technique is used to classify individuals/objects into one of alternative groups on the basis of a set of predictor variables (Independent variables) .
- The dependent variable in discriminant analysis is categorical and on a nominal scale, whereas the independent variables are either interval or ratio scale in nature.
- When there are two groups (categories) of dependent variable, it is a case of two group discriminant analysis.
- When there are more than two groups (categories) of dependent variable, it is a case of multiple discriminant analysis.

Introduction

- Discriminant Analysis is applicable in situations in which the total sample can be divided into groups based on a non-metric dependent variable.
- Example:- male-female
  - high-medium-low
- The primary objective of multiple discriminant analysis are to understand group differences and to predict the likelihood that an entity (individual or object) will belong to a particular class or group based on several independent variables.

Example

- Heavy product users from light users
- Males from females
- National brand buyers from private label buyers
- Good credit risks from poor credit risks

Objectives

- To find the linear combinations of variables that discriminate between categories of dependent variable in the best possible manner.
- To find out which independent variables are relatively better in discriminating between groups.
- To determine statistical significance of the discriminant function and whether any statistical difference exists among groups in terms of predictor variable.
- To evaluate the accuracy of classification, i.e., the percentage of customers that is able to classify correctly.

Discriminant Analysis Model

- The mathematical form of the discriminant analysis model is:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + \varepsilon$$

where, Y = Dependent variable

b_s = Coefficients of independent variable

X_s = Predictor or independent variable

- Y should be categorized variable and it should be coded as 0, 1 or 1,2,3 similar to dummy variables.
- X should be continuous.
- The coefficients should maximize the separation between the groups of the dependent variable.

Accuracy of classification

- The classification of the existing data points is done using the equation, and the accuracy of the model is determined.
- This output is given by the classification matrix (also called confusion matrix), which tells what percentage of the existing data points is correctly classified by the model.

Relative importance of independent variable

- Suppose we have two independent variables X_1 and X_2.
- How do we know which one is more important in discriminating between groups?
- Coefficients of both the variables will provide the answer.

Predicting the group membership for a new data point

- For any new data point that we want to classify into one of the groups, the coefficients of the equation are used to calculate Y discriminant score.
- A decision rule is formulated – to determine the cutoff score, which is usually the midpoint of the mean discriminant score of two groups.

Coefficients

- There are two types of coefficients.

1. Standardized coefficients
2. Unstandardized coefficients

- Main difference➔ standardized coefficient will not have constant ‘a’.

Apriori probability of classification into groups

- The discriminant analysis algorithm requires to assign an apriori (before analysis) probability of a given case belonging to one of the groups.

1. We can assign equal probabilities of assignments to all groups.
2. We can assign proportional to the group size in the sample data.

TERMS

- Classification matrix:
  Means of assessing the predictive ability of the discriminant functions.
- Hit ratio:
  Percentage of objects (individual, respondent, firms etc.) correctly classified by the discriminant function.

i.e.

$$\frac{Number\ correctly\ classified}{Total\ number\ of\ observations}\ x\ 100$$

Terms

- Cutting score:
  Criterion against which individual's discriminant Z score is compared to determine Predicted group membership.

$$C = \frac{n_2\bar{y}_1 + n_1\bar{y}_2}{n_1 + n_2}$$

(Basic formula for computing the optimal cutting score between any two groups)

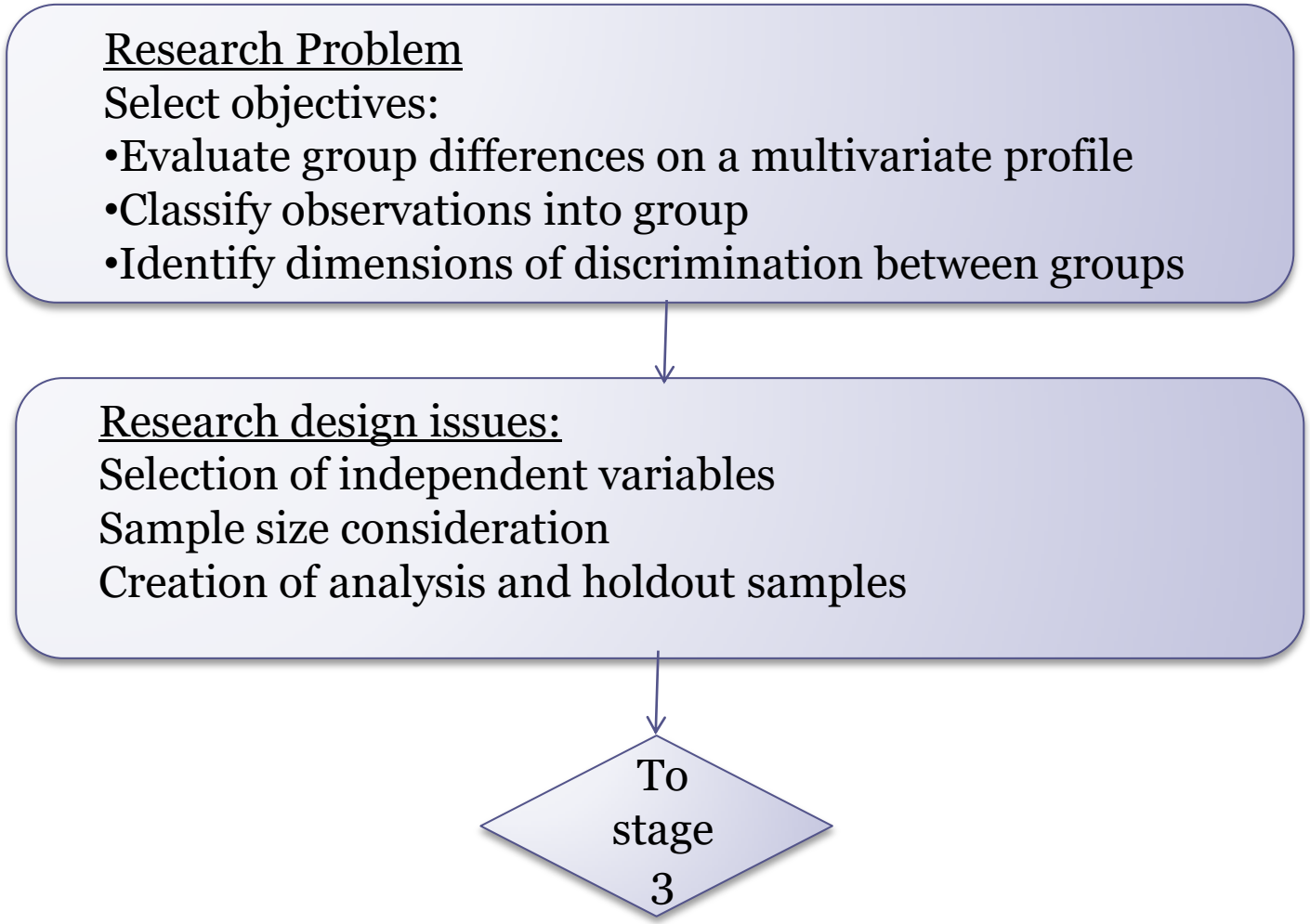
Terms

- Discriminant loadings:
  Discriminant loadings are calculated whether or not an independent variable is included in the discriminant function(s).
- Discriminant weights: (Discriminant coefficients)
  Independent variables with large discriminatory power usually have large weights, and those with little discriminatory power usually have small weights.
- Centroid:
  Mean value for the discriminant Z scores of all objects within a particular category or group.

Flow chart

STAGE 1

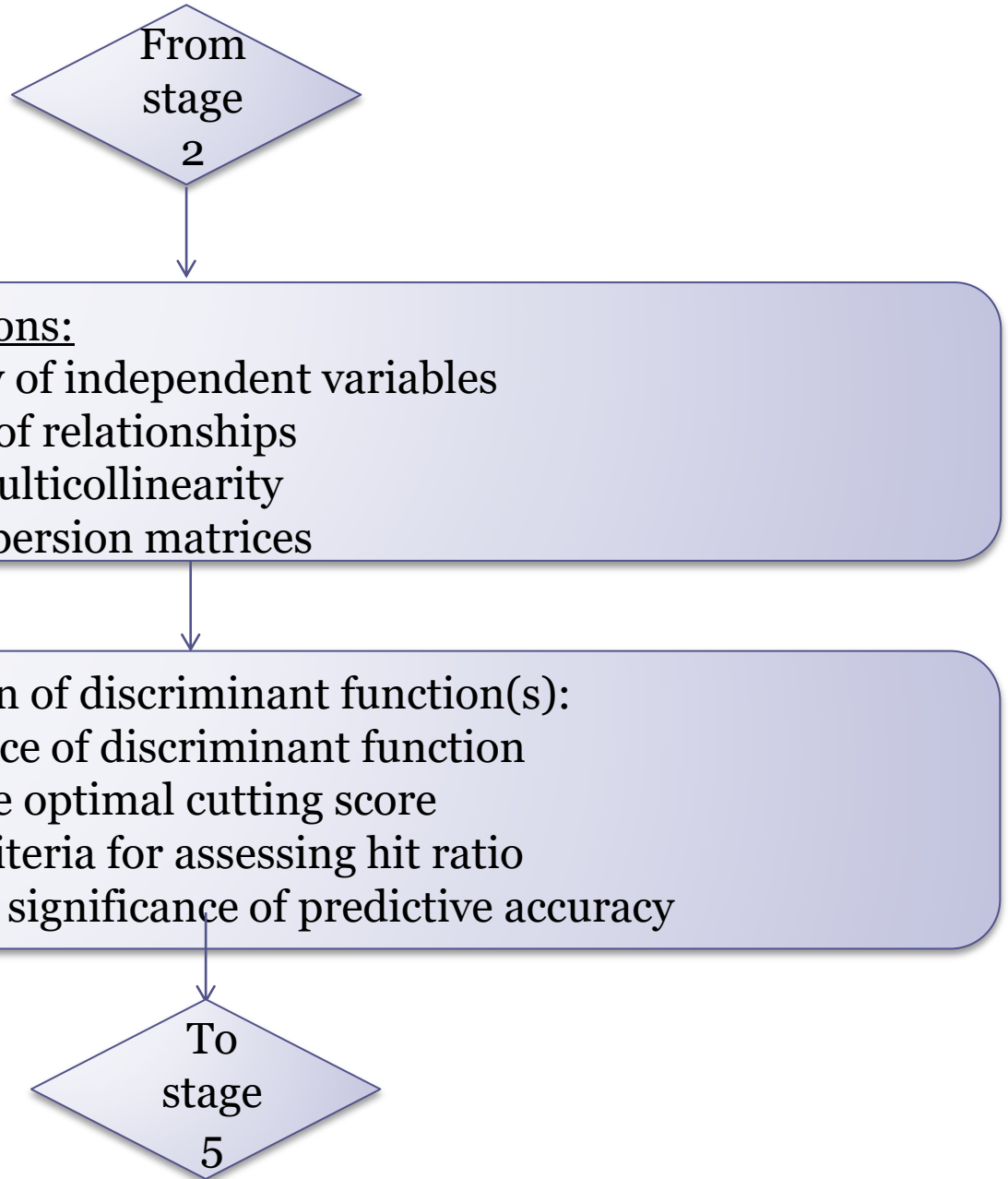
STAGE 2



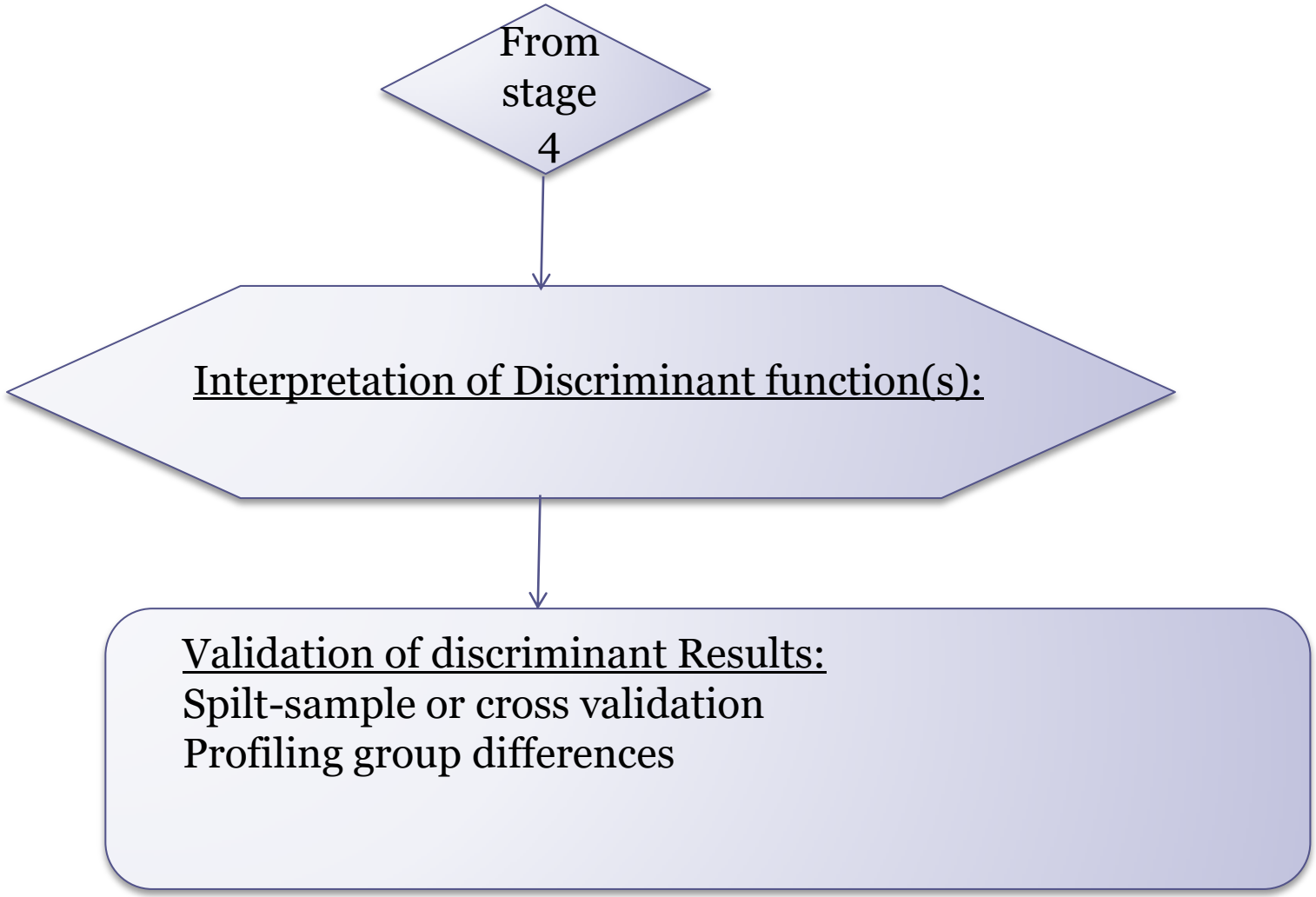
Flow chart

STAGE 3

STAGE 4



Flow chart


From stage 4
STAGE 5
Interpretation of Discriminant function(s):
STAGE 6
Validation of discriminant Results:
Spilt-sample or cross validation
Profiling group differences

Question:

- A retail outlet wants to know the consumer behavior pattern of the purchase of products in two categories- national brands and local brands respectively, which would help it to place orders depending on demand and requirements of customer. This retail outlet uses data from a retail outlet in another location to arrive at a decision about customers visiting at their end.
- This retail outlet wants to use discriminant analysis to screen the responsiveness of customers towards national brand and local brand categories and find out the following:

1. The percentage of customers that it is able to classify correctly.
2. Statistical significance of discriminant function.
3. Which variable (annual income and household size) are relatively better in discriminating between consumers for national and local brand.
4. Classification of new customers into one of the two groups namely- national (group-1) and local brand (group-2)acceptors.

Data

Sr. No.	Brand	Annual income	Household size
1	1	16.8	3
2	1	21.4	2
3	2	17.3	4
4	2	15.4	3
5	1	17.3	4
6	1	18.4	1
7	2	14.3	4
8	2	14.5	5
9	1	23.2	2
10	1	21.1	5
11	2	17.4	2
12	2	16.7	6
13	1	14.5	4
14	1	18.9	1
15	2	13.9	7
16	2	12.4	7
17	1	17.8	2
18	1	19.3	1
19	2	15.3	6
20	2	13.3	4

Results

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.473	12.721	2	.002

- Wilks' Lambda is the ratio of within-groups sums of squares to the total sums of squares.
- This is the proportion of the total variance in the discriminant scores not explained by differences among groups.
- A lambda of 1.00 occurs when observed group means are equal.
- A small lambda indicates that group means appear to differ.
- The associated significance value indicate whether the difference is significant.
- Here, the Lambda of 0.473 has a significant value (Sig. = 0.002); thus, the group means appear to differ.

Results

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.113[a]	100.0	100.0	.726

- An Eigen value indicates the proportion of variance explained
- A large Eigen value is associated with a strong function.

Results

Canonical Discriminant Function Coefficients

	Function
	1
INCOME	.335
HOUSEHOLD SIZE	-.313
(Constant)	-4.545

Unstandardized Coefficients

- The Canonical Discriminant Function Coefficients indicate the unstandardized scores concerning the independent variables.
- It is the list of coefficients of the unstandardized discriminant equation.
- Each subject's discriminant score would be computed by entering his or her variable values (raw data) for each of the variables in the equation.

Discriminant function:

$$Y = -4.545 + 0.335X_1 - 4.545X_2$$

Results

Standardized Canonical Discriminant Function Coefficients

	Function
	1
INCOME	.722
SIZE	-.490

Standardized Coefficients

- The coefficients of standardized discriminant function are independent of units of measurement.
- The absolute value of coefficient in standardized discriminant function indicates the relative contribution of variables in discriminating between the two groups.

Results

Functions at Group Centroids

VAR00001 BRAND	Function
	1
1.00	1.001
2.00	-1.001

Means of Canonical variables

- 'Functions at Group Centroid' indicates the average discriminant score for subjects in the two groups.
- More specifically, the discriminant score for each group when the variable means (rather than individual values for each subject) are entered into the discriminant equation.
- Note that the two scores are equal in absolute value but have opposite signs.

Results

Classification Results[a,c]

			Predicted Group Membership		
		Brand	1.00	2.00	Total
Original	Count	1.00	9	1	10
		2.00	2	8	10
	%	1.00	90.0	10.0	100.0
		2.00	20.0	80.0	100.0
Cross-validated[b]	Count	1.00	8	2	10
		2.00	2	8	10
	%	1.00	80.0	20.0	100.0
		2.00	20.0	80.0	100.0

a. 85.0% of original grouped cases correctly classified.

b. 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.

c. 80.0% of cross-validated grouped cases correctly classified.

- 'Classification Results' is a simple summary of number and percent of subjects classified correctly and incorrectly.
- The 'leave-one out classification' is a cross-validation method, of which the results are also presented.

Results

	Case Number	Actual Group	Predicted Group	Cross Validated	
				Actual Group	Predicted Group
Original	1	1	1	1	1
	2	1	1	1	1
	3	2	1**	2	1**
	4	2	2	2	2
	5	1	1	1	2**
	6	1	1	1	1
	7	2	2	2	2
	8	2	2	2	2
	9	1	1	1	1
	10	1	1	1	1
	11	2	1**	2	1**
	12	2	2	2	2
	13	1	2**	1	2**
	14	1	1	1	1
	15	2	2	2	2
	16	2	2	2	2
	17	1	1	1	1
	18	1	1	1	1
	19	2	2	2	2
	20	2	2	2	2

Results

- In last table, the stared value indicates that respondent 3 and 11 were wrongly classified in group 1.
- Respondent 3 and 11 were actually belongs to group 2.
- Respondent 13 was wrongly classified in group 2.
- It originally belongs to group 1.
- Hit ratio= (no. of correct predictions/ total no. of cases)*100

  =(17/20)*100

  =85%

Using discriminant function

- $Y = -4.545 + 0.335X_1 - 4.545X_2$
- If $X_1 = 15$ and $X_2 = 2$
- Than $Y = -4.545 + 0.335(15) - 4.545(2)$

  $= -4.545 + 5.025 - 9.05$

  $Y = -8.57$
- Cut-off score $= \frac{n_2\bar{y}_1 + n_1\bar{y}_2}{n_1 + n_2} = \frac{(10)(1.001) + (10)(-1.001)}{10+10} = 0$
- Therefore this respondent will belong to group-2 i.e. local brand buyer.

THANK YOU