UNIT 12 REGRESSION ANALYSIS, DISCRIMINANT ANALYSIS AND FACTOR ANALYSIS

Objectives

After studying this unit, you should be able to :

- explain the concept of Association that takes place between a dependent variable and a set of independent variables
- describe the various multivariate procedures available to analyse associative data in the context of marketing
- interpret the findings of multivariate analysis in any market research study
- use a particular technique of multivariate analysis suitable for a particular marketing problem.

Structure

12.1 Introduction

12.2 Analysis of Variance

12.3 Regression Analysis

12.4 Discriminant Analysis

12.5 Factor Analysis

12.6 Summary

12.7 Self-Assessment Exercises

12.8 Further Readings

12.1 INTRODUCTION

In the previous block we covered the fundamentals of statistical inference with special emphasis on hypothesis testing as an effective tool for marketing decisions. Univariate analysis forms the foundation for the development of multivariate analysis, which is the topic of discussion in this unit. While the concept of the univariate analysis will continue to draw our attention time and again, our focus in this unit will be on procedures of multivariate analysis which has emerged as the most striking trend in marketing research methodology.

Description and analysis of associative data involves studying the relationship and degree of association among several variables and therefore multivariate procedures become imperative. We shall attempt to highlight the procedures with a marketing orientation. It is important to realise that amongst the many techniques available, some are more intensely used compared to others.

Analysis of variance is suitable for analysing experimental data based on field experiments which is gaining increased attention in marketing research. It is a useful analytical tool for marketing researchers. Whenever we are interested in comparison of means of any number of groups, analysis of variance is an appropriate technique to use. This technique could be applied in new product evaluation, selection of copy theme, effectiveness of an advertising/sales promotional campaigns and the like.

Regression analysis finds out the degree or relationship between a dependent variable and a set of independent variables by fitting a statistical equation through the method of least square. Whenever we are interested in the combined influence of several independent variables upon a dependent variable our study is that of multiple regression. For example demand may be influenced not only by price but also by growth in industrial production, extent of import prices of other goods, consumer's income, taste and preferences etc. Market researchers could use regression for

explaining per cent variation independent variable caused by a number of independent variables and also problems involving prediction or forecasting.

Discriminant analysis is useful in situations where a total sample could be classified into mutually exclusive and exhaustive groups on the basis of a set of predictor variables. Unlike the regression analysis, these predictor variables need not be independent. For example one may wish to predict whether sales potential in a particular marketing territory will be `good' or `bad' based on the territory's personal disposal income, population density and number of retail outlets. You may like to classify a consumer as a user or non-user of one of the five brands of a product based on his age, income and length of time spent in his present job.Here the interest is what variables discriminate well between groups.

Factor analysis provides an approach that reduces-a set of variables into one or more underlying variables. The technique groups together those variables that seem to belong together and simultaneously supplies the weighing scheme. For example one may be interested in the identification of factors that determine the company's image. When the decision maker is overwhelmed by the factor analysis comes to his, help in compressing there many variables into a few meaningful dimensions, like service orientation, quality level and width of assortment in a research project involving 20 retail chains on 35 factors or variables.

12.2 ANALYSIS OF VARIANCE

Analysis of variance abbreviated as ANOVA is the hallmark of statistical methodology pioneered by I.A. Fisher, while analysing design of experiments. This technique is applicable for analysing experimental data. Variation is inherent in nature. The-analysis of variance technique attempts to find out how much variation is caused by random fluctuations called errors and how much by assignable *causes. The* technique breaks down the total variation into meaningful components variation - that produced by treatments and that produced by random errors. Here the word `treatment' implies exposing one set of experimental data to a particular device used in an experiment. The group or set receiving this treatment is called a control group. For example target consumers could be put into two groups, one group is asked to taste the new brand of tea while the other group is asked to taste the existing brand; or each group is asked to taste first the existing brand in the morning and the new brand in-the evening. The objective could be to evaluate whether there is any significant difference in the perception by the two groups.

Whenever we are interested in the comparison of two or more treatment means, the most appropriate tool is the analysis of variance. For example, a test to measure the sales appeal effectiveness of three different package designs or a test to measure the elasticity of four different prices, could be carried out through analysis of variance method. Manly one test variable is used then it is called one way analysis of variance. If two test variables are used then it is called two way analysis of variance. The discussion will begin with one way analysis of variance which lays the foundation for tyro way analysis of variance. We will avoid complicated mathematical derivations and focus only on procedure and hypothesis testing relevant to marketing research. However for-the assumption underlying the model and. computational aspect, certain mathematical rotations and algebra become unavoidable. One thing is sure that you will be very clear when the illustrations are discussed.

One way classification

ANOVA Model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

$$i = 1, 2, ..., a$$

$$j = 1, 2, ...n_i$$

where Y_{ij} is jth observation corresponding to ith treatment drawn independently.

a is the number of treatments

n_i is the number of observations on each treatments

$$\sum_{i=1}^{a} \alpha_i = 0 \varepsilon_{ij} \bigcap IND(0, \sigma^2)$$

μ is the grand mean effect

α_i is the ith treatment effect

IND means independently normally distributed. It can be shown by minimising the sum of squares of errors with respect to.μ, α_i and then squaring

Total sum of squares:

= Treatment sum of squares + Error sum of squares

Computational aspect:

Correction factor $C = \frac{G^2}{n}$ where G is the grand total of all observations, n is the total

number of observations = $\sum_{i=1}^{n} n_i$

Total Sum of Squares (TSS) = TSS = $\sum_{i,j} Yij^2 - C$

Treatment Sum of Squares (TRSS) = TRSS = $\sum_{i,j} Yij^2 - C$

Error Sum of Squares (ESS) = TSS - TRSS

Alternative ESS can also be computed directly as

$$\sum_{i,j} Yij^2 - \sum_{i} \frac{Y_1^2}{n_i}$$

This would help cross checking the result obtained by substraction as before TSS - TRSS)

Y_i = Marginal total of ith treatment

i = 1, 2 ……,a

Form the ANOVA Table as follows

Source of variation	D.F. (Degrees of Freedom)	Sum of Squares	Mean Square	F (ratio)
Due to Treatment	a–1	TRSS	TRSS/a – 1	(TRSS/a – 1) (ESS/n – a)
Due to Error	n – a	ESS	ESS/n – a	
Total	n – 1	TSS		

$$\frac{\text{TRSS}/(a-1)}{\text{ESS}/(n-a)} = \frac{\text{Treatment Mean Square}}{\text{Error Mean Square}}$$

Follows Sredecor's F distribution with (a – 1), (n – a) d.f.

In symbol $\frac{\text{TRSS}/(a-1)}{\text{ESS}/(n-a)} \cap F(a-1, n-a)$

Set up H_o : Treatment Means are equal

H_1 : There is difference amongst treatment means.

If the calculated F in the ANOVA table exceeds the T46le F (a-1, n-a) at 5% level reject H_o and accept H_1.

Easy way to remember computational aspect of ANOVA.

1) If there are a treatments (generally columns) then degrees of freedom for treatment = a-1
2) if the total number of observation is n degrees of freedom associated with Total = n-1
3) Degrees of freedom associated with error = (2) - (1)
   = (n-1) - (a-1)
   = n-a
4) Sum all observations, square it and divide by n, you get correction' factors C.
5) Square each observation; sum them and subtract C, you get total sum of square = TSS

6) Obtain the marginal total for each treatment i (column) square it and divide by the number of observations under column i. Sum for all i and subtract c, you get the Treatment Sum of Squares = TRSS

7): (5) - (6) gives error sum of squares = ESS

Now you should be in a position to form the ANOVA Table.

To make matters clear let us look at an example from marketing.

A consumer product company was interested in evaluating the sales effect of two different colours for the package of one of its products. The firm selected 10 stores with similar monthly sales pattern, and randomly split them into two groups of 5 stores each. One group of stores was stocked with red colour packages while the other group of stores was stocked only with blue packages. All stores were monitored for two weeks to make certain that the packages were properly displayed and that no stockout occurred. The following data shows the number of test packages which were sold in each store for the two-week period. The company wants to know whether there is significant difference in sales effectiveness of the two packages.

Sales by stores

Red package	Blue package
6	16
8	18
10	20
12	22
14	24

Computation:

There are 2 treatments here - Red package and Blue package $n_1 = 5$ (number of observations under treatment 1) and $n_2 = 5$ (number of observations under treatment 2) $n = n_1 + n_2 = 5 + 5 = 10$

	Treatment 1	Treatment 2	Total
	6	16	22
	8	18	26
	10	20	30
	12	22	34
	14	24	38
Total	50	100	150
	G = 150		

$$\text{Correction factor} = \frac{G^2}{n} = \frac{150^2}{10} = \frac{22500}{10} = 2250$$

$$TSS = 6^2 + 8^2 + 10^2 + 12^2 + 14^2 + 16^2 + 18^2 + 20^2 + 22^2 + 24^2 - 2250 = 330$$

$$TRSS = \frac{50^2}{5} + \frac{100^2}{5} - 2250 = 250$$

$$ESS = TSS - TRSS = 330 - 250 = 80$$

Anova Table

Source of variation	D.F.	Sum of square	Mean squares	F Ratio
Due to treatment	1	250	$\frac{250}{1} = 250$	$\frac{250}{10} = 25$
Due to error	8	80	$\frac{80}{8} = 10$	
Total	9	330		

Null Hypothesis H_0: There is no difference between red package and blue package in sales effectiveness.

H_1: There is difference between red and blue package in regard to sales effectiveness.
Table (1, 8) at 5% level = 5.32

Interpretation of results

Since the calculated F = (25) exceeds table F (1, 8) = 5.32, reject the null hypothesis that the treatment means are equal and accept H_i. This means at 5% level or with a confidence level of 95%, we can state that there is difference in sales effectiveness between *red* package and blue package. The analysis of variance does not tell the researchers which of the two package is more effective. By observation, we can conclude that the blue package was more effective than the red package.

Two-way analysis of variance

The basic procedures underlying one way analysis of variance apply also to two way analysis of variance, except for the fact that in two-way analysis of variance, there are two variables being tested rather than only one.

The model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

The assumptions and structure are as before except that we have added one more factor β_j f or which $\sum \beta_j = 0$. If α_i, denotes treatment or column effect β_j denotes block or row effect. The computational aspects, are same as before except that we have to calculate row sum of squares (or block of squares).To recapitulate let us write down the expression.

$$\text{Correction factor C} = \frac{G^2}{n}$$

$$\text{Total sum of Squares (TSS)} = \sum_{i,j} Y_{ij}^2 - C$$

$$\text{Treatment Sum of Squares (TRSS)} = \sum_i \frac{Y_i^2}{n_i} - C$$

$$\text{Block Sum of Squares (BSS)} = \sum_j \frac{Y_i^2}{a} - C$$

ERSS =TSS - TRSS - BSS

BSS involves marginal total for jth block or row, each containing number of observations = a being the number, of treatments in each block.

ANOVA Table for two-way classifications

Source	D.F.	Sum of squares	Mean square	F Ratio
Due to Treatment	a–1	TRSS	$\frac{TRSS}{a-1}$	$\frac{(TRSS)}{(a-1)} \div \frac{(ESS)}{(n-a-b+1)}$
Block	b – 1	BSS	$\frac{BSS}{b-1}$	$\left(\frac{BSS}{b-1}\right) \div \left(\frac{ESS}{n-a-b+1}\right)$
Error	*n–a–b+1	ESS	$\frac{ESS}{n-a-b+1}$	
Total	n–1	TSS		

* obtain by subtraction from total degrees of freedom, degree of freedom for treatment and block you please note that there are two F Ratios in this ANOVA, one to treatment effect and one to block effect.

Ho: Treatment Means are equal
H_1**:** Treatment Means are not equal For treatment

H_o: Block Means are equal
H_1: Block Means are not equal

If the calculated F exceeds Table F at 5%, reject H_o and accept H_I.

Example

The following table gives the quality ratings of 10 service stations by five marketing research professionals. What are your conclusions on service effectiveness difference and difference in professional.

	Service Stations										
Raters	**1**	**2**	**3**	**4**	**5**	**6**	7	**8**	**9**	**10**	**Total**
A	99	70	90	99	65	85	75	70	85	92	830
B	96	65	80	95	70	88	70	51	84	91	790
C	95	60	48	87	48	75	71	93	80	93	750
D	98	65	70.	95	67	82	73	94	86	90	820
E	97	65	62	99	60	80	76	92	90	89	810
Total	485	325	350	475	310	410	365	400	425	455	4000

Computations

$$\text{Correction Factor C} = \frac{G^2}{n} = 320000$$

$$\text{TSS} = 99^2 + 96^2 + 95^2 + 98^2 + 97^2 + \ldots\ldots\ldots + 92^2 + 91^2 + 93^2 + 90^2 + 89^2 - 320000 = 9948$$

$$\text{TRSS} = \frac{485^2}{5} + \frac{355^2}{5} + \ldots\ldots + \frac{455^2}{5} - 320000 = 6810$$

$$\text{BSS} = \frac{830^2}{10} + \frac{790^2}{10} + \frac{750^2}{10} + \frac{820^2}{10} + \frac{810^2}{10} - 320000 = 400$$

$$\text{ESS} = 9948 - 6810 - 400 = 2738$$

ANOVA Table

Source	D.F.	Sum of Squares	Mean Square	F Ratio
Treatment (Column)	9	6810	$\frac{6810}{9} = 756.67$	$\frac{756.67}{76.06} = 9.95$
Block (Row)	4	400	$\frac{400}{4} = 100.00$	$\frac{100.00}{76.06} = 1.31$
Error	36	2738	$\frac{2738}{36} = 76.06$	
Total	49	9948		

Treatment (column) variations :

Null hypothesis H_0 : There is no difference between the effectiveness of service provided by various service stations.

H_1: There is difference amongst service station with regard to effectiveness of service. Since the calculated F 9.95 exceeds table F (9, 36) = 2.15 at 5% level, reject H_o and accept H_1:

Interpretation: With a confidence level of 95%, it can be said that there is difference in the effectiveness of service provided by the various service stations.

Block(Row) variation:

H_0 : There is no difference between rating of service effectiveness by the 5 professionals.

H_1 : There is difference between rating by the 5 professionals.

Since calculated F = 1.31 does not exceed the table F (4, 36) = 2.36 at 5% level, we have no evidence to reject Ha.

Interpretation : With a confidence level of 95% we conclude that there is no difference in rating by the 5 professionals with regard to effectiveness of service:

Activity 1

Mention briefly the usefulness of ANOVA in marketing research.

……

……

……

……

12.3 REGRESSION ANALYSIS

Regression analysis is probably the most widely applied technique amongst the analytical models of association used in marketing research. Regression analysis attempts to study the relationship between a dependent variable and a set of independent variables (one or more). For example, in demand analysis, demand is inversely related to price, for normal commodities. We may write D = A - BP where D is the demand which is the dependent variable, P is the unit price of the commodity, an independent variable. This is an example of a simple linear regression equation. The multiple linear regressions model is the prototype of single criterion/multiple predictor association model where we would like to study the combined influence of several independent variables upon one dependent variable. In the above example if P is the consumer price index, and q is the index of industrial production, we may be able to study demand as a function of two independent variables P and Q and write D = A -- BP + C Q as a multiple linear regression model.

The objectives of the market researchers in using Regression Analysis are

1) To study a general underlying pattern connecting the dependent variable and independent variables by establishing a functional relationship between the two. In this equation the degree of relationship is derived which is a matter of interest to the researcher in his study.
2) To use the well established regression equation for problems involving prediction and forecasting.
3) To study how much of the variation in the dependent variable is explained by the set of independent variables. This would enable him to remove certain unwanted variables from the system. For example if 95% of variation in demand in a study could be explained by price and consumer rating index, the researcher may drop other factors like industrial production, extent of imports, substitution effect etc. which may contribute only 5% of variation in demand provided all the causal variables are linearly independent,

We proceed by first discussing bivariate (simple) regression involving the dependent variables as a function of one independent variable and then on to multiple regression.

Simple linear regression model is given by

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

where Y is the dependent variable,

X_1 is independent variable

ϵ is a random error term

$\beta_0$and β_1 are the regression coefficients to be estimated.

Assumptions of the model

1) The relationship between Y and X_i is linear.
2) Y is a random variable which follows a normal distribution from which sample values are drawn independently.
3) X_1 is fixed and is non-stochastic (non-random).
4) The means of all these normal distribution of Y as conditioned by X_1, lie on a straight line with slope p_i.
5) ϵ is the error term $\sqcap$ IND $(0, \alpha^2)$ and independent of X_1.

Computational aspect

Estimated regression line based on sampling is written $\hat{Y} = a + bX_1$

a and b are estimates of β and β_1 obtained through the method of least square by minimising the error sum of squares.

We state the normal equations without going into any derivations. The normal equations are

$$\sum Y = na + b\sum x_1$$
$$\sum X_1 Y = a\sum X_1 + b\sum X_1^2$$

Solve these two simultaneous equations you get the values of a and b.

Total sum of squares= $\sum (Y - \bar{Y})^2$ (TSS)

Regression sum of squares= $\sum (\hat{Y} - \bar{Y})^2$ (RSS)

Error sum of squares= $\sum (Y - \hat{Y})^2$

From the ANOVA Table for Regression

Source	D.F.	Sum of squares	Mean squares	F ratio
Due to Regression	1	RSS	$\frac{RSS}{1}$	$\frac{RSS}{1} \div \frac{ESS}{(n-2)}$
Due to Error	n–2	ESS	$\frac{ESS}{n-2}$	
Total	n–1	TSS		

H_0 $\beta_1 = 0$ There is no linear relationship between Y and X_1 (Y & X_1 are independent).

$H_1 \beta_1 \neq 0$ There is linear relationship between Y and X_I as stated in our model.

If the calculated F exceeds Table F (1, n - 2) at 5% level, reject H_0 and accept H_1.

Strength of association

It is one thing to find the regression equation after validating the linearity relationship but at this point we still do not know how strong the association is. In other words, how well does X_I predict Y?

This is measured by the co-efficient of determination

$r^2 = \frac{RSS}{TSS}$ = Variation in Y explained by regression compared to total variation.

Higher the r^2, greater is the degree of relationship.

The product moment correlation or simple correlation co-efficient between Y and X_1 is $= \sqrt{r_2} = \sqrt{\frac{RSS}{TSS}}$

r^2 lies between 0 and 1. 0 measuring no correlation and 1 measuring perfect correlation.

r lies between -1 and -;1 and the sign of r is determined by the sign of the sample regression coefficient (b) in the sample regression equation

$\hat{Y} = a + bX_1$

Having given a foundation structure with underlying assumptions and possible. analysis of the model, we now turn our attention to a numerical example to clarify the concepts. It is needless to mention that analysis of data and interpretation of the results are of paramount importance.

Suppose that a marketing researcher is interested in consumers attitude towards nutritional diet of a ready to eat cereal.

X_1: the amount of protein per standard serving

In the nature of a pretest , the researcher obtains consumer's interval-scaled evaluation of the- ten concept descriptions, on a preference rating scale ranging from 1, dislike extremely, upto 9, like extremely well. The data is given below.

Rater	Preference rating (Y)	Protein X_1	
1	3	4	$\Sigma Y = 43\ \bar{y} = 4.3$
2	7	9	
3	2	3	$\Sigma X_1 = 43\ \bar{x}_1 = 4.3$
4	1	1	
5	6	3	$\Sigma YX_1 = 247$
6	2	4	
7	8	7	$\Sigma X_1^2 = 255$
8	3	3	
9	9	8	
10	2	1	

i) Fit a linear regression model of Y on X_1.

ii) Test the validity of the equation statistically.

iii) What do you think of the strength of association?

Answer

i) The normal equations are

$\Sigma Y = na + b\Sigma x_1$. Here n = 10

$\Sigma X_1 Y = a\ \Sigma X_1 + b\ \Sigma X_1^2$

i.e. $10a + 43b = 43$

$43a + 255b = 247$

solving these two simultaneous equations
we have b = 0.886
a = 0.491

Regression Equation is $\hat{Y} = 0.491 + 0.886X_1$
substitute for all x_1^s to get y^s

The regression coefficient b = 0.886 indicates the change in Y per unit change in X_1

ii) Validity of the equation
ANOVA calculation

Total sum of squares (TSS) $= \Sigma (Y - \bar{Y})^2$
$= (3 - 4.3)^2 + (7 - 4.3)^2 + .,.,. + (2 - 4.3)^2$
$= 76.10$

*Regression sum of squares (RSS) $= \Sigma (Y - \bar{Y})^2$
$= (4.034 - 4.3)^2 + (8.464 - 4.3)^2 + ...$
$+ (1.377 - 4.3)^2 = 55.01$

* Error sum of squares $= \Sigma (\hat{Y} - Y)^2 = (-1.034)^2 + (-1.464)^2 + ...$
$+ (0.623)^2 = 21.09$

*Table of Actual Vs predicted

Actual Y	Predicted Y	Error Y – Y
3	4.034	–1.034
7	8.464	–1.464
2	3.148	–1.148
1	1.377	–0.377
6	3.148	2.852
2	4.034	–2.034
8	6.692	1.308
3	3.148	–0.148

	9	7.579	1.422
	2	1.377	0.623

ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F ratio
Due to regression	1	55.01	$\frac{55.01}{1} = 55.01$	$\frac{55.01}{2.64} = 20.84$
Due to error	8	21.09	$\frac{21.09}{8} = 2.64$	
Total	9	76.10		

H_0 : Where is no relationship of linear type between Y and X_1
i.e. $\beta_1 = 0$
H_1 : Y is linearly related to X_1 $\beta \neq 0$

Interpretation of results

Calculated F = 20.84 exceeds Table F (1,8) = 5.32 at 5 % level. Reject H_o and accept H_i. We conclude Y is a linear function of X_I with a confidence level of 95%. In other words preference rating Y is linearly related to amount of (X_1) Protein per standard serving of the cereal with a confidence level of 95%. Thus the equation is a valid one.

iii) Strength of association

$$\text{Coefficient of determination } r^2 = \frac{RSS}{TSS} = \frac{55.01}{76.10} = 0.723$$

This implies that 72.3% of the variation in Y is explained by the regression and only 27.7% of the variation is explained by error. The association is strong to enable X_1 to predict Y.

$r = \sqrt{r2} = 0.850$

Here r will have a positive sign since b is positive. Before starting our discussion on the multiple regression, let us give a brief account of the u efulness of simple linear regression in sales forecasting using time series data in which time t is an independent variable.

1) Linear Trend

$Y = a + bt$

if Y represents the sales data collected for the past many years for example last 10 years from 1979 to 1988, we normalise the year by taking t = 1 corresponding to 1979, t = 2 for 1980 etc. and t = 10 for 1988. Now the simple linear regression model can be directly applied to forecast Y (sales) say for 1989 after fitting the regression equation.

2) Trend in semilog form $Y = ab^t$

Taking log on both sides we have Log

$Y = \text{Log } a + t \log b$

This reduces to $Z = A + Bt$

Where $Z = \text{Log } y$
$A = \text{Log } a$
$B = \text{Log } b$

This can now be solved as a simple linear regression model for forecast where is dependent variable and t is independent variable as before.

3) Double log form

$Y = at^b$

$\text{Log } Y = \log a + b \log t$

i.e. $Z = A + bT$

where $Z = \log Y$ $A = \log a_j$ T Log t.

This can now be solved as normal bivariate regression equation to forecast sales for the next period.

It is now time to introduce the concept of multiple regression model

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots\ldots\ldots + \beta_k X_k + \varepsilon$$

The assumptions are exactly same as simple linear regression except that you add X_I, X_2, ……X_k in the place of X_I because Y is linearly related to X_1……X_k and our aim is to understand the combined influence of the K factors X_I, X_2…….X_k on Y. To understand the concept clearly, let us study a case of 2 independent variables and write the model as

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

so that $Y = a + bX_1 + cX_2$ being the estimated regression equation where we add one more independent variable X_2 in the model. Suppose we extend the previous example of bivariate regression on preference rating Vs protein (X_I) by adding X_2: the percentage of minimum daily requirements of vitamin D per standard serving. Let us see how the multiple regression model emerges to explain the variation in the dependent variable Y caused by X, and X_2. Let us look at the following table giving the data on Y, X_I, and X_2.

Rater	**Preference Rating Y**	**Protein X_1**	**Vitamin D X_2**	
1	3	4	2	
2	7	9	7	$\sum Y = 43$
3	2	3	1	$\sum X_1 = 43$
4	1	1	2	$\sum X_2 = 40$
5	6	3	3	
6	2	4	4	$\sum YX_1 = 247$
7	8		9	$\sum YX_2 = 232$
8	3	3	2	$\sum X^2_1 = 255$
9	9	8	7	$\sum X^2_2 = 226$
10	2	1	3	$\sum X_1X2 = 229$

The normal equations are:

$\sum Y = Na + \sum X_1 + c\sum X_2$

$\sum YX_1 = a\sum X_1 + b\sum X^2_1 + c\sum X_1X_2$

$\sum YX_2 = a\sum X_2 + b\sum X_1X_2 + c\sum X^2_2$

$10a + 43b + 40c = 43$

$43a + 255b + 229c = 247$

$40a + 229b + 226c = 232$

Solving for a, b and c we have

$a=0.247$

$b=0.493$

$c = 0.484$

$Y= 0.247 + 0.493\ X_1 + 0.484\ X_2$

Here b and c are called partial regression coefficients b = 0.493 denotes the change-in Y per unit change in X_I when X_2 is held constant. Similarly c = 0.484 denotes the change in Y per unit change in X_2 when X_I is held constant.

By now you must have notified the cumbersome calculations involved when the number of variables increase and becomes extremely difficult when the number of variables is more than 3. One has to resort to computer based regression models. In fact it may be mentioned here that all multivariate procedures require the help of computer when the variables and observations are large. As before we can calculate the other co-efficients like R^2 co-efficient of multiple determination and R multiple correlation co-efficient and also ANOVA for hypothesis testing. The author has developed his own user friendly program for multiple regression with a conversational

style based on IBM PC MS DOS. We will use the output of the program and interpret the results of our problem which is the most important aspect for us.

Multiple Linear Regression

Number of Variables ? 3
Number of Observations ? 10

WANT TO CHANGE NUMBER OF VARIABLES/NUMBER OF OBSERVATIONS (Y OR N) ? N
NAME OF VARIABLE # 1 ? Y
NAME OF VARIABLE # 2 ? X1
NAME OF VARIABLE # 3 ? X2

DATA GATHERED FOR VARIABLE Y :
— PERIOD # 1 ? 3
— PERIOD # 2 ? 7
— PERIOD # 3 ? 2
— PERIOD # 4 ? 1
— PERIOD # 5 ? 6
— PERIOD # 6 ? 2
— PERIOD # 7 ? 8
— PERIOD # 8 ? 3
— PERIOD # 9 ? 9
— PERIOD # 10 ? 2

DATA GATHERED FOR VARIABLE X1 :
— PERIOD # 1 ? 4
— PERIOD # 2 ? 9
— PERIOD # 3 ? 3
— PERIOD # 4 ? 1
— PERIOD # 5 ? 3
— PERIOD # 6 ? 4
— PERIOD # 7 ? 7
— PERIOD # 8 ? 3
— PERIOD # 9 ? 8
— PERIOD # 10 ? 1

DATA GATHERED FOR VARIABLE X2 :
— PERIOD # 1 ? 2
— PERIOD # 2 ? 7
— PERIOD # 3 ? 1
— PERIOD # 4 ? 2
— PERIOD # 5 ? 3
— PERIOD # 6 ? 4
— PERIOD # 7 ? 9
— PERIOD # 8 ? 2
— PERIOD # 9 ? 7
— PERIOD # 10 ? 3

Correlation Matrix

1	.85	.85
.85	1	.84
.85	.84	1

Variance Covariance Matrix

2.91	6.9	6.67
6.9	2.79	6.33
6.67	6.33	2.71

Variable	Mean	Std Deviation
Y	4.3	2.907844
X1	4.3	2.790858
X2	4	2.708013

CONTINUE ? Y
CONTINUE ? Y

Regression Equation

Dependent Variable: Y

Independent variable	ESTIMATED COEFFICIENT	BETA %	Errors	T-Test
X1	.49	.49	.34	1.47
X2	.48	.45	.35	1.4
CONSTANT	.24711704			

Determination Coefficient	=	.78
Correlation Coefficient	=	.89
F-Test	=	12.65
Degrees of Freedom	=	2,7
Sum of squares of error	=	16.49

CONTINUE?

CONTINUE? **TABLE OF RESIDUAL VALUES**

#	Observation	Estimation	Residual
1	3	3.18	-.18
2	7	8.07	-1.07
3	2	2.21	- .21
4	1	1.71	-.71
5	6	3.18	2.82
6	2	4.15	- 2.15
7	8	8.05	-.05
8	3	2.69	.31
9	9	7.57	1.43
10	2	2.19	-.19

CONTINUE ?Y

Analysis of Variance Table

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Due to Regression	2	59.61	29.81	12.65
Due to Error	7	16.49	2.36	
Total	9	76.1		

Another Analysis (Type Y or N)?

The program output gives many other statistical analysis which we will not touch upon now, and come to our important tests straightway. The residual or error between Y and 1' i.e. between actual and forecast on important measure of reliability of the model is printed out for each observation. If you look at the errors, you get a fairly good idea about the model equation. However for validity of the regression equation, you look first at the coefficient of multiple determination R^2 and multiple correlation coefficient R. In our example $R^2 = 0.78$ and $R = 0.89$ which is a satisfactory one indicating that the preference rating Y is linearly related to protein intake X_I and vitamin D intake X_2. It tells that 78% of variation in Y is explained jointly by the variations in X_I and X2 jointly.

Hypothesis testing for linearity through ANOVA.

H_o : $\beta_1 \& \beta_2 = 0$=> There is no linear relation between Y, X_I and X_2.

H_1 : $\beta_1 \& \beta_2 \neq 0$=> There is linear relationship.

Look at the ANOVA

The calculated F = 12.65

The table F (2, 7) = 4.74 at 5% level. Reject H_o and accept H_1.

That is the linear regression equation between & X_I, X, is statistically valid. We are 9$% confident that preference rating is linearly related to X_1 and X_2, and the equation is $Y= 0.247 + 0.49X_1 + 0.48X_2$

Points to ponder on Multiple regression analysis

1) Equation should be validated statistically.

2) For forecasting the dependent variable, the independent variables should be forecast first.

   For example if demand is a function of price index and production index established by a multiple regression model, then to forecast demand for the next period, it is imperative first to forecast the price index and production index and :then substitute them in the model to get the forecast for demand. This is one of the limitations of regression forecasting.

3) When the variables become too many the analysis is complex and very often the market researcher does not know which variables to retain. This problem could be overcome by doing `stepwise regression' on computer. For example if demand is a function of 20 variables, we first fit demand equation with 3 important variables which we think affect demand. Suppose $R^2 = 0.85$ that is 85% of the variation in demand is explained by these variables, we add another two more variables of importance to make five independent variables. Now if $R^2 = 0.95$ we can as well stop adding further variables as the contribution may not appreciably improve the situation. We can thus visualise demand as a function of just 5 variables. The various permutations of changing and adding variables is possible only with the help of a computer. The important point to remember is that the cat off point for the number of variables to be added should be based on the increase every time you get on R^2. The moment the increase is marginal, stop adding variables.

4) If the independent variables among themselves are highly correlated, then we are facing the problem of `multicollinearity'. Normally we say that the partial regression coefficient-with respect to X_I implies change in Y per unit change in X_I provided X_2, X_3.....are held constant. This poses% serious problem if there is multicollinearity. One way to overcome multicollinearity is to drop certain variables from the model if the corresponding standard error of regression coefficient are unduly large. Another method is to see whether the original set could be transformed into another linear composite so that the new variables are uncorrelated.

12.4 DISCRIMINANT ANALYSIS

It has been pointed out earlier, that the discriminant analysis is a useful tool for situations where the total sample is to be divided into two or more mutually exclusive and collectively exhaustive groups on the basis of a set of predictor variables. For example, a problem involving classifying sales people into successful and unsuccessful; classifying customers into owners or and non-owners of video tape recorder, are examples of discriminant analysis.

Objectives of two group discriminant analysis :

1 Finding linear composites of the predictor variables that enable the analyst to separate the groups by maximising among groups relative to with in-groups variation.

2 Establishing procedures for assigning new individuals, whose profiles hilt not group identity are known, to one of the two groups.

3 Testing whether significant differences exist between the mean predictor variable profiles of the two groups.

4 Determining which variables account most for intergroup differences in mean profiles.

A numerical example

Let us return to the example involving ready-to-eat cereal that was presented in the regression analysis. however in this problem die ten consumer raters are simply asked to classify the cereal into one of two categories like versus dislike. The data is given below: Here again

X_1: The amount of protein (in grams) per standard serving.

X_2 : The percentage of minimum daily requirements of vitamin D per standard serving.

Also shown in the data table are the various sums of squares and cross products; the means on X1 and X2 of each group, and total sample mean,

Consumer evaluations (like versus dislike) of ten cereals varying in nutritional content

Person	Evaluation	**Protein X_1**	**vitamin D X_2**		X^2_1	X^2_2	X_1X_2
1	Dislike	2	4		4	16	8
2	Dislike	3	2		9	4	6
3	Dislike	4	5		16	25	20
4	Dislike	5	4		25	16	20
5	Dislike	6	7		36	49	42
	Mean	4	4.4		Sum 90	110	96
6	Like	7	6		49	36	42
7	Like	8	4		64	16	32
8	Like	9	7		81	49	63
9	Like	10	6		100	36	
10	Like	11	9		121	81	99
	Mean	9	6.4	Sum	415	218	296
	Grand Mean.	6.5	5.4				
	Standard deviation	3.028	2.011				

The grand mean is $X_I = 6.5$, $\bar{X}_2 = 5.4$

We first note from the' table that the two groups are much more widely separated on X_I (protein) than they are on X_2 (Vitamin D). If we were forced to choose just one of the variables, it would appear that X_I is a better bet than X_2. However there is information provided by the group separation on X_2, so we wonder if some linear composite of both X_1 and X_2 could do better than X_1 alone. Accordingly we have the following linear function:

$Z = K_1X_1 + K_2X_2$ where K_1 and K_2 are the weights that we seek.

But how shall we define variability? In discriminant analysis, we are concerned with the ratio of two sums of squares after the set of scores on the linear composite has been computed. One sum of squared deviations represents the variability of the two group means on the composite around their grand mean. The second sum of squared deviations represents the pooled variability of the' individual cases around their respective group means also on the linear composite. One can then find the ratio of the first sum of squares to the second. It is this ratio that is to be maximised through the appropriate selection of K_I and K_2. Solving for K_I and K2 involves a procedure similar to the one encountered in the multiple regression. However in the present case, we shall want to find a set of sums of squares and cross products that relate to the variation within groups. For ease of calculation let us define $x_i = X_1 - X_i$ and $x_2 = X2 - X2$ (i.e. each observation measured from its mean)

Solving for K_1 and K_2

Mean corrected sums of squares and cross products.

	Dislikers	*Likers*	*Total*
$\sum x_1^2 = \sum (x_1 - \bar{x}_1)^2 = \sum X_1^2 - N\bar{X}_1^2 =$	10	10	20
$\sum x_2^2 = \sum (x_1 - \bar{x}_1)^2 = \sum X_1^2 - N\bar{X}_1^2 =$	13.2	13.2	26.4
$\sum x_1 x_2 = \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) = \sum x_1 x_2 - N\bar{X}_1\bar{X}_2$	8	8	16

The normal equations are

$$K_1 \sum x_1^2 + K_2 \sum x_1 x_2 = \overline{X_1}(Li\ker s) - \overline{X_1}(disli\ker s)$$
$$K_1 \sum x_1 x_2 + K_2 \sum x_2^2 = \overline{X_2}(Li\ker s) - \overline{X_2}(disli\ker s)$$
$$20K_1 + 16K_2 = 6.4 - 4.4 = 2$$

Solving these two simultaneous equations, we have $K_1 = 0.368$, $K_2 = -0.147$
Discriminant function $Z = 0.368X_1 - 0.147X_2$

We can also find discriminants scores for the means of the two groups and' the grand mean.

$\bar{Z}$(dislikers) = 0.368 (4) - 0.147 (4.4) = 0.824

$\bar{Z}$(likers) = 0.368 (9) - 0.147 (6.4) = 2.368

$\bar{Z}$(grand means) = 0.368(6.5) - 0.147 (5.4) = 1.596

We note that the discriminant function "favours" X_1 by giving about 2.5 times the (absolute value) weight ($K_1 = 0.368$ versus $^K2 = -0.147$) to X1 as is given to X_2.

The discriminant scores of each person are shown below. Each score is computed by the application of the discriminant function to the persons original'X_1 and X_2 values.

Dislikers		**Likers**	
Person	**Discriminant Score**	**Person**	**Discriminant Score**
1	0.148	6	1.691
2	0.809	7	2.353
3	0.735	8	2.279
4	1.250	9	2.721
5	1.176	10	2.721
Mean	0.824	Mean	2.368

Grand Mean 1.596

Between group variability

$5\ (0.824 - 1.596)^2 + 5\ (2.368 - 1.596)^2 = 5.96$

Within group variability

Dislikers $(0.148 - 0.824)^2 + (0.809 - 0.824)^2 + ...(1.176 - 0.824)^2 = 0.772$

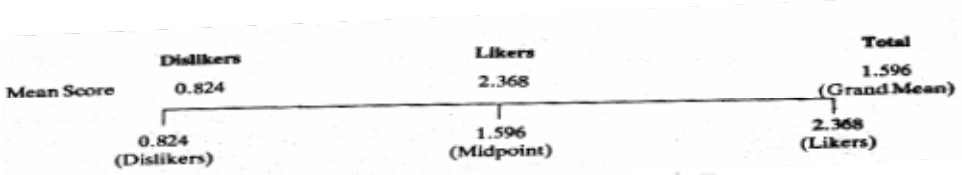
Likers $(1.691-2.368)^2 + (2.353 - 2.368)^2 + (2.721 + 2.368)^2 = \dfrac{0.772}{1.544}$

Discriminant criterion $C = \dfrac{5.96}{1.544} = 3.86$

Since the normal equations for solving k1 and 1c2 are obtained by maximising the ratio between group and within group variance the discriminant criterion as calculated above = 3.86 will be the maximum possible ratio. If We suppress X2 in the discriminant function and calculate another C, it will be less than 3.86: It is rather interesting that the optimal function Z = 0.368 X1 - 0.147 X2 is a difference function in which X2 (Vitamin D) receives a negative weight bringing thereby the importance of XI to the highest order. This means protein is much more important than Vitamin D.

Classifying the persons

It is all well and good to find the discriminant function, but the question is how to assign the persons to the relevant groups.



- Assign all cases with discriminant scores that are on the left of the midpoint (1.596) to the disliker group.

- Assign all cases with discriminant scores that are on the right of the midpoint (1.596) to the liker group.

That is all true dislikers will be correctly classified as such and all true likers will be correctly classified, This can be shown by a 2 x2 table:

	Assigned by Rule		
True State	**Disliker**	**Liker**	**Total**
Disliker	5	**0**	**5**
Liker	0	5	5
Total	5	5	10

Testing Statistical Significance

While the discriminant function does perfectly in classifying the ten cases of the illustration on protein (X_I) and vitamin (X_2) into likers and dislikers, we still have not tested whether the group means differ significantly. This is also based on F ratio which required calculation of Mahalanobis D^2. This calculation of F is a little complicated which is normally an output parameter in the standard package like Biomedical computer program and SPSS of IBM. Biomedical computer program of the University of California press is an outstanding software containing all multivariate procedures. For our illustration let us calculate F

$$F = \frac{n_1 n_2 (n_1 + n_2 - m - 1)}{m(n_1 + n_2)(n_1 + n_2 - 2)} . D^2$$

~F distribution with m_l $n_l + n_2$ - m -1 d.f.

where n_l = number of observations in group 1

n_2 = number of observations in group 2

m = number of independent variables

D^2 = Mahalanobis square distance

In our problem $n_l = 5$

$n_{2.} = 5$

m = 2 (X_1 and X_2)

Simple way of calculating D^2 would be to use the discriminant function

D^2 = $(n_1 + n_2 - 2)$ (0.368 (5.0) - 0.147 (2))

= 8(0.368 X 5 - 0.147 x 2) = 12.353

You please note that the expression within brackets is the discriminant function Z = 0.368 X_1 -0.147) ($_2$ where X_1 and X2 are substituted by the respective group means difference: X_I (likers) - X_1 (dislikers), X_2 (likers) - X2 (dislikers)

$$F = \frac{5 \times 5(5+5-2-1)}{2 \times (5+5)(5+5-2)} \times 12.353$$

$$= \frac{25 \times 7}{2 \times 10 \times 8} \times 12.353 = 13.511$$

Table F(2,7)=4.74at 5% level.

Since the calculated F exceeds table F at 5% level, reject H_o and accept H_1 i.e. the group means are not equal in importance with a probability of 95%. This clearly validates the relative importance of X_I far higher than X_2.

Brief remarks on Multiple Discriminant Analysis :

You would have realised by now the complexity of calculations in the discriminant analysis involving 2 predictors which itself needs computer based solutions when the number of observations increases considerably. Multiple discriminant an analysis is invariably carried out by means of computer programs. One of the most flexible and comprehensive programs in BMD-07M of the biomedical program series of the

University of California press. SPSS also has all multivariate procedures. It may be mentioned that the basic structure of the bivariate analysis remains same in multiple case also. What is important for you is interpretation of the results and findings of the study.

Activity 2

What are the differences between Regression Analysis and Discriminant Analysis?

……
……
……
……
……
……
……..

12.5 FACTOR ANALYSIS

Factor analysis is a generic name given to a class of techniques whose purpose ns data reduction and summarisation. Very often market researchers are overwhelmed by the plethora of data. Factor analysis comes to their rescue in reducing the number of variables. Factor analysis does not entail partitioning the data matrix into criterion and predictor subsets; rather interest is centred on relationships involving the whole set of variables. In factor analysis;

1 The analyst is interested in examining the "strength" of the overall association among variables in the sense that he would like to account for this association in terms of a smaller set of linear composites of the original variables that preserve. most of the information in the full data set. Often his interest will emphasize description of the data rather than statistical inference.

2 No attempt is made to divide the variables into criterion versus prediction sets.

3 The models are primarily based on linear relationships.

Factor analysis is a "search" technique. The researcher-decision maker does not typically have a clear a priori structure of the number of factors to be identified. Cut off points with respect to stopping rules for the analysis are often ad hoc as the output becomes available. Even where the procedures and rules are stipulated in advance, the results are more descriptive than inferential.

The procedure involved in computation of factor analysis is extremely complicated and cannot be carried out effectively without the help of computer. Packages like SPSS, SAS and Biomedical programs (BMD) can be used to analyse various combinations leading to factor reduction. We will make an attempt to conceptualise the scenario of factor analysis with emphasis on the interpretation of figures.

The term "factor analysis" embraces a variety of techniques. Our discussion focuses on one procedure: principal component analysis and the factors derived from the analysis are expressed as linear equations. These linear equations are of the form

$$F_i = a_{1i}x_1 + a_{2i}x_2 + a_{3i}x_3 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + a_{mi}x_m$$

The i factors are derived, and each variable appears in each equation. The a-coefficients indicate the importance of each variable with respect to a particular factor coefficient of zero indicating the variable is of no significance for that factor. In principal component analysis; the factors are derived sequentially, using criteria of maximum reduction in variance and non-correlation among factors.

Let us go to a specific example to explain factor analysis and its output.

Example

A manufacturer of fabricating parts is interested in identifying the determinants of a successful salesperson. The manufacturer has on file the information shown in the following table. He is wondering whether he could reduce these seven variables to two or three factors, for a meaningful appreciation of the problem.

Data Matrix for Factor Analysis of seven variables (14 sales people)

Sales person	Height (x_1)	Weight (x_2)	Education (x_3)	Age (x_4)	No. of Children(x_s)	Size of household	IQ (x_7)
1	67	155	12	27	0	2	102
2	69	175	11	35	3	6	92
3	71	170	14	32	1	3	111
4	70	160	16	25	0	1	115
5	72	180	12	36	2	4	108
6	69	170	11	41	3	5	90
7	74	195	13	30	1	2	114
8	68	160	16	32	1	3	118
9	70	175	12	45	4	6	121
10	71	180	13	24	0	2'	92
11	66	145.	10	39	2	4	100
12	75	210	16	26	0	1	109
13	70	160	12	31	0	3	102
14	71	175	13	43	3	5	112

Can we now collapse the seven variables into three factors? Intuition might suggest the presence of three primary factors : A maturity factor revealed in age/children/size of household, physical size as shown by height *and* weight, and intelligence or training as revealed by education and IQ,

The sales people data have been analysed by the SAS program. This program accepts data in the- original units, automatically transforming them into standard scores. The three factors derived from the sales people data by a principal component analysis (SAS program) are presented below

Three, factor results with seven variables.

Sales people characteristics Factor

Variable		I	II	III	Communality
Height		0.59038	0.72170	- 0.30331	0.96140
Weight		0.45256	0.75932	-0.44273	0.97738
Education		0'.80252	0.18513	0.42631	0.86006
Age		-0.86689	0.41116	0.18733	0.95564
No. of children		-0.84930	0.49247	0.05883	0.96730
Size of household		-0.92582	0.30007	- 0.01953	0.94756
IQ		0.28761	0.46696	0.80524	0.94918
Sum of squares		3.61007	1.85136	1.15709	
Variance summarised		0.51572	0.26448	0.16530	0.94550

Factor Loadings: The coefficients in the factor equations are called "factor loadings" They appear above in each factor column, corresponding to each variable. The equations are

$F_1 = 0.59038x_1 + 0.45256x2 + 0.80252x_1 - 0.86689x_4 - 0.84930x_5\ 0.92582x_6 + 0.28761x,$

$F_2 = 0.72170x_1 + 0.75932x2 + 0.18513x3 + 0.41116x_4 + 0.49247x_5 + 0.30007x_6 + 0.46696x_7,$

$F_3 = -0.30331x_1 - 0.44273x_2 + 0.42631x_3 + 0.18733x_4 + 0.5883x_5 - 0.01953x_6 + 0.80524x_7$

The factor loadings depict the relative importance of each variable with respect to a particular factor. In all the three equations, education (x3) and IQ (x_7) have got positive loading factor indicating that they are variables of importance in determining the success of sales person.

Variance summarised : Factor analysis employs the criterion of maximum reduction of variance -variance found in the initial set of variables. Each factor contributes to reduction. In our example Factor I accounts for 51.6 per cent of the total variance. Factor II for 26.4 per cent and Factor III, for 16.5 per cent. Together the three factors "explain" almost 95 per cent of the variance.

Communality : In the ideal solution the factors derived will explain 100 per cent of the variance in each of the original variables, "Communality" measures the percentage of the variance in the original variables that is captured by the combination of factors in the solution. Thus a communality is computed for each of the original variables, Each variables communality might be thought of as showing the extent to which it is revealed by the system of factors. In our example the communality is over 85 per cent for every variable. Thus the three factors seem to capture the underlying dimensions involved in these variables.

There is yet another analysis called varimax rotation, after We get the initial results. This could be employed if needed by the analyst. We do not intend to dwell on this and those who want to go into this aspect can use SAS program for varimax rotation.

In the concluding remarks, it should be mentioned that there are two important subjective issues which should be properly resolved before employing factor analysis model. They are

1) How many factors should be employed in attempting to reduce the data? What criteria should be used in establishing that number?

2) The labeling of the factors is purely intuitive and subjective.

Activity 3

Mention briefly the purpose and uses of factor analysis

……
……
……
……
……
……
……

12.6 SUMMARY

We have started our discussion by giving an overview of the various multivariate analysis procedures in the context of associative data with a marketing orientation. We have given a brief introduction of the multivariate tools and their applicability in the relevant problem areas.

We have discussed the concept of analysis of variance. We have clearly brought out the assumptions underlying the one way and two-way classification models and the methodology of separation of total variance into meaningful components variations and en-or variations. Hypothesis testing using ANOVA table has been clearly explained using examples from marketing which include testing sales and service effectiveness using experimental data.

The next topic of discussion has been the regression analysis, which is probably .the most widely used technique amongst the analytical models of association. We have started the simple linear regression model first to introduce the concept of regression and then moved on to the multiple linear regression mode All the underlying assumptions of the model have been clearly explained. Both the bivariate and multivariate regression models have been illustrated-using the example of preference rating as a function of protein intakes, and vitamin D intake perception in the case of a ready to eat cereal. The concept of testing the linear equation, contribution made by regression in explaining variation in dependent variables and strength of association have all been explained using ANOVA table. A brief account of the role of regression in sales forecasting involving time series analysis has also been given. The need for resorting to computer solutions for large number of variables and observations has been brought out with an actual print out of the example already discussed. The concept of stepwise regression and the problems encountered in any regression analysis have also been explained.

Then we have gone to the discriminant analysis technique-a technique when the interest is to classify the groups on the basis of a set of predictor variables. We have explained the concept of separation by giving examples of classifying sales people into successful and unsuccessful, customers into owners and non-owners etc. As before, we have begun the discussion with discriminant function involving two predictor variables using the example of `ready to eat cereal problem' but with a difference - classifying the persons into liker group and disliker group. The discriminant function, the discriminant criterion and the assignment rule have all been explained testing the statistical significance using F test based on Mahalanobis D^2 has also been carried out. We have pointed out that the multiple discriminant analysis involving more than two predictor variables require the use of computer although the basic structure of the model does not change.

Factor analysis is the last multivariate tool that we have discussed in this unit. We have first mentioned that the fundamental objective of factor analysis is to reduce the number of variables in the data matrix. Then it has been pointed out that the computation of any factor analysis involves dry complex calculations which will have to be solved using computer packages like SAS. The concepts of "factor loading", "variance summarised" and "communality" have been clearly explained using one practical example that has been solved by SAS program. The subjective issues like "how many factors?" "what criteria to decide this number?" and "labelling of the factors" have been mentioned at the end.

As concluding remarks, it may be mentioned here that 1) all multivariate procedures can be more effectively solved using standard computer packages when the number of variables and number of observations increase significantly, 2) what is more important is the ability to interpret the results of the market research study involving multivariate analysis.

12.7 SELF-ASSESSMENT EXERCISES

1 a) In a demand forecasting study involving a normal commodity, two simple linear regression models are fitted:

$D = 8.5 + 0.22\ p\ (r^2 = 0.75)$

$\text{Log } D = 1.3 + 0.10 \log p\ (r^2 = 0.80)$

i) which model would you prefer and why?

ii) mention the dependent and independent variables.

b) A manufacturer of industrial supplies developed the following model for predicting the number of sales per month

$Y = 41 + .3X_1 + .05X_2 - 7X_3 + 10X_4$

where Y = Sales per month

X_1= Number of manufacturing firms'

X_2 = Number of wholesale and retail firms

X_3 = Number of competing firms

X_4 = Number of full time company sales people.

i) Explain the correct interpretation of all estimated parameters in the equation.

ii) If $R^2 = 0.49$, what` does this figure mean, to you?

iii) Explain how you will go about testing the validity of this multiple linear regression equation:

2 The following discriminant function was developed to classify sales persons into successful and unsuccessful sales person

$\mathbf{Z} = 0.53\ X_I + 2.IX_2 + 1.5X_3$

Where X, = no, of sales call Made by sales persons.

X_2 = no.. of customers developed by sales person.

X_3.= no. of units sold by sales person.

The following decision -rule was developed.

if $\bar{Z} \leq 10$, classify the sales person as successful

if $\bar{Z} < 10$, classify the sales personas unsuccessful.

The sales persons A and B were considered for promotion_ on the basis of being classified as successful or unsuccessful. Only the successful sales person would*be promoted. The relevant data on A and B, is given below :

	A	**B**
X_1	10	11
X_2	2	1.5
X_3	1	0.5

whom will you promote?

3 A large sample of people were asked to rate how much they liked each of 5 beverages -coffee, tea, milk, fruit juice and soft drinks. Through factor analysis

Coffee	-.219	.363	- .338	0.2939
Tea	-.137	.682	.307	0.578t
Milk	.514	-.213	-.277	0.3611
Fruit Juice.	.485	-.117	115	0.2621
Soft drinks	-.358	-.635	.534'	0.8165
Sum of squares	0.6943	1.0592	0.5584	
Variance summarised	0.1389	0.2118	0.1117	0.4624

a) Write the linear equations for all the three factors.

b) Interpret the loading co-efficients, variance summarised and communality ' values of this table.

12.8 FURTHER READINGS

F.E. Brown , Marketing Research Addison – Wesley publishing Company.

Paul E. Green and Donald S. Tull , Research for Marketing Decisions Prentice- Hall of India Pvt Ltd.

Boyd , Westfall and Stasch , Marketing Research Text and Cases D. Irwin Inc Homewood, Illinios.

U.K. Srivastava , G.V., Quantitative Techinques for Managerial .Decision Making wiely Eastern Limited.