Unit III

Discriminant function analysis.

A variate of the independent variables selected for their discriminantary power used in the prediction of group membership is known as Discriminant function analysis. The predictor value of the discriminant function is the discriminant z score which is calculated for each objects in the analysis. It takes the form of linear equation which is as follows.

$$Z_{jk} = a + w_1 x_{1k} + w_2 x_{2k} + \dots + w_n x_{nk}$$

where Z_{jk} is discriminant z score of discriminant function j for the object k, a is the intercept, w_i is the discriminant weight for independent variable i. x_{ik} is the independent variable for the object k.

For example, consider the students applying for admission to college are given an entrance examination. The students may be a member of one popln consisting of those students who will successfully completed the entrance exam or the students may be a member of other popln those who will not be compiled the entrance successfully. The problem is to discriminate a student applying for admission on the basis of his score on the entrance examination.

Assumptions:

Discriminant function analysis is computationally very similar to MANOVA apply

Sample size: Unequal sample sizes are acceptable. The sample size of the smallest group needs to exceed the no. of predictor variables. As a rule of thumb, the smallest group sample size should be at least 20 for a few (4 or 5) predictor. The max no of ind variable is n-2, where n is the sample size. Ithele the low sample size may work,

Normal distn: It is assumed that the data represent a sample from a multivariate normal distn. You can examine whether or not variables are normally distributed with histograms of frequency distn. However, note that violations of the normality assumption are not "fatal" and the resultant significance test are still reliable as long as non-normality is caused by skewness and not outliers

Homogeneity of variances: DA is very sensitive to heterogeneity of variance covariance matrix. Before accepting final conclusion for an important study, it is a good idea to review the within groups variances and correlation matrices. It is evaluated through scatterplots and correlated by transformation of variables

Outliers :- DA is highly sensitive to the inclusion of outliers. Run a test for for univariate and multivariate outliers for each group and transform or eliminate them. If one group in the study contains extreme outliers that impact the mean, they will also increase variability. Overall significance tests are based on pooled variances, that is, the average variance across all groups.

Non-multicollinearity :- If one of the independent variables is very highly correlated with another, or one is a function of other independents, then the tolerance value for that variable will approach 0 and the matrix will not have a unique discriminant solution. There must also be low multicollinearity of the independents. To the extent that independents are correlated, the standardized discriminant function coefficients will not reliably assess the relative importance of the predictor variables.

Comparison of Discriminant Analysis and Regression Analysis Unit III

Discriminant model is, $d = v_1 x_1 + v_2 x_2 + \ldots + v_n x_n + a$

Regression model is, $Y = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + a$

1) Regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable

2) In Regression analysis to study the relationship b/w the variable we use simple correlation. In case of discriminant analysis we use canonical correlation analysis to know the relationship b/w the variables

3) In Regression analysis the dependent variable is ratio type or scale type, but the case of discriminant analysis the dependent variable is categorical

4) Regression Analysis the independent variables are ratio or scale type. In case of discriminant analysis the independent variables may be of any type (ie) Normal, Ordinal, interval, scale

5) In linear Regression dependent variable relates a set of independent predictor variables and the idea is to find a function of linear parameters that best fits the data in Discriminant analysis it's a procedure for classifying the object into categories.

6) In RA the best fit is desired by its coefficient of determination (R^2). In DA the best model is desired by the prob. value of misclassification

7) DA and RA are exactly proportional to each other that is the computations are difference and the results are based on their coefficient.

Canonical correlation

The canonical correlation coefficient measures the strength of association b/w two canonical variates

A canonical variates is the weighted sum of the variables in the analysis.

characteristic function of wishard distn :-

definition :- A random p×p symmetric matrix $\mathbf{S}$ has a wishart distn with parameters Σ, p, and n

Discriminant

$d = V_1 X_1 + V_2 X_2 + \ldots + V_n X_n + a$

1) Dependent var should be qualitative type with two or more groups

2) Indepnd var should be quantitative type in discriminant analysis

3) $V_1, V_2, V_3 \ldots V_n$ are discriminant coefficients

4) a is a error term

logistic Reg

1) $Y = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + a$

Dep variable should be qualitative with two groups

Y is prob value

2) Indp variable may be quantitative (or) qualitative (or) mixture of both.

3) $b_1, b_2 \ldots b_n$ are regression coefficients

4) Constant — a

Purpose of Discriminant analysis:

The purpose of DA is to classify objects into two (or) more groups based on a set of features that describe the objects In general, we assign an object to one of a no of pre-determined groups based on observation made on the object.

For eg, If we want to know whether a soap product is good (or) bad based on several measurement on the product such as weight, volume, smell, colour contrast etc. Here the object is soap, the class category is good and bad. what we are looking for is dependent variable Each measurement of the product is called feature that describe the object (ie) independent variables

Thus, in DA, dependent variable Y in the group and the independent variable x are the object features that might describe the group

The dependent variables is always categorical variable (nominal scale) while the inde. variable can be any measurement (nominal, ordinal, interval (or) ratio)

While classifying the objects, the classification criterion is to minimize the total error of classification.

We tend to make the proportion of object that the misclassification as small as possible Thees, TEC should be thought as the prob that the rule under consideration will misclassify and the object of the classification rule is to assign the object two the group with highest conditional probability. This is called base rule and this rule also minimise TFC.