REVIEW ARTICLE

Discriminant analysis – simplified

Sandhya Jain, Merin Kuriakose

Department of Orthodontics and Dentofacial Orthopedics, Government College of Dentistry, Indore, Madhya Pradesh, India

Correspondence
Dr. Merin Kuriakose, Department of Orthodontics and Dentofacial Orthopedics, Government College of Dentistry, Indore, Madhya Pradesh, India.
Phone: +91-6263435301/+91-8939273396.
E-mail: orthopublications09@gmail.com

Received 10 November 2019;
Accepted 31 December 2019

doi: 10.15713/ins.ijcdmr.139

How to cite the article:
Jain S, Kuriakose M. Discriminant analysis – simplified. Int J Contemp Dent Med Rev, vol.2019, Article ID: 031219, 2019. doi: 10.15713/ins.ijcdmr.139

Abstract

Background: Discriminant function analysis is the statistical analysis used to analyze data when the dependent variable or outcome is categorical and independent variable or predictor variable is parametric. Discriminant function analysis is used to find out the accuracy of a given classification system or predictor variable in predicting the sample into a particular group. Discriminant function analysis includes the development of discriminant functions for each sample and deriving a cutoff score. The cutoff score is used for classifying the samples into different groups. **Aim:** The aim of this review article is to simplify and explain the discriminant function analysis so that it can be used by medical and dental researchers whenever it is applicable. **Conclusion:** Discriminant function analysis is a statistical analysis used to find out the accuracy of a given classification system or predictor variables. This paper explains the basics of discriminant analysis and how to interpret the results along with one simple example of mandibular canine index for gender identification. **Clinical significance:** Whenever a new classification system is introduced or any predictor variable is identified, discriminant function analysis can be used to find out the accuracy with which the classification system or predictor variable can differentiate a sample into different groups. Thus, it is a very useful tool in dental and medical research.

Keywords: Assumptions, data analysis, dependent variable, discriminant function analysis, independent variable, prediction, standard coefficient

Introduction

Discriminant analysis (DA) is a statistical technique for analyzing data when the dependent variable or outcome is categorical or non-parametric and the predictor or independent variables are interval in nature.[1] Discriminant function analysis is a technique to determine in which weightings of independent quantitative variables or predictors best discriminate between two or more than two groups of cases and do so better than chance.[2] It is used when the outcome variables are dichotomous or binary, for example, male or female, dead or alive, etc.

An example of DA is when mandibular canine index (MCI) is used for the prediction of gender. In this case, the predictor, i.e., MCI is an interval or parametric value and the prediction outcome, i.e., gender is categorical.

Objectives of DA

Development of discriminant functions is as follows: [1]

- To find out the accuracy with which a predictor variable can categorize samples into different groups
- To evaluate the accuracy of classification systems
- To find out which predictor variables contribute more to the intergroup differences.

There are two types of DA

- Univariate DA
- Multivariate DA.

The example of univariate DA is the use of MCI for predicting gender, in which one independent variable (MCI) is used for predicting the outcome, i.e. gender. If more than one independent variables or predictors are used for prediction, then multivariate DA should be used.

An example of multivariate DA is when along with canine index, molar index is also used for predicting gender. Here, two or more number of predictor variables can be used. In multivariate DA, there are more than one predictors so we have to use "stepwise function" to remove insignificant predictors and use only significant predictors.

Derivation of Mathematical Formula for DA

The analysis develops a discriminant function which is a linear combination of the weightings and scores on the independent variables or predictors. The maximum number of functions can be the number of predictors or the number of groups minus one, whichever is smaller among these two values.[3]

DA determines a linear equation that will predict the group, in which the sample belongs to.

The equation is formulated like this:

$$D= v1\ X1+ v2\ X2+ v3X3+... + vnXn + a$$

Where, D = discriminant function

v = the discriminant coefficient or weightage for predictor variable, i.e., how much the corresponding independent variable contribute to the prediction or intergroup difference.

X = independent variable or predictor variable

a = a constant also known as intercept which indicates what would be the dependent variable if all the independent variables were zero.

n = the number of predictor variables

The discriminant coefficient or weight for independent variable (v) is unstandardized discriminant coefficients, which is similar to the beta coefficients in the regression analysis. Discriminant coefficient is given for each variable in each discriminant function, and if the coefficient is large, the corresponding variable contributes more for the discrimination between the groups. The discriminant coefficient maximizes the distance between the means of the dependent variable and develops an equation that has strong discriminatory power between groups. For the basic two groups of DA, there is only one discriminant function.

Assumptions of DA

- Sample size: Unequal sample sizes are acceptable. The smallest group sample size has to exceed the number of predictor variables. The maximum number of predictor variables is n-2, where n is the sample size[4]
- Distribution of the data: Predictor variables should have a multivariate normal distribution
- Homogeneity of variances/covariances: Within-group variance-covariance matrices should be equal across groups
- Outliers: DA is sensitive to the outliers. If one group in the study contains extreme values that impact the mean, they will also increase variability[4]
- Low multicollinearity: Multicollinearity of the variables should be low. If the independent variables are correlated to each other, then the discriminant function coefficients cannot reliably predict group membership.

Interpretation of Results

In DA, the discriminant function is calculated for every sample. Let us take the example of MCI used for predicting gender. The sample size is 300, including males and females. In this case, the discriminant function is calculated for 300 samples. Three hundred different discriminant function values are obtained one for each sample.

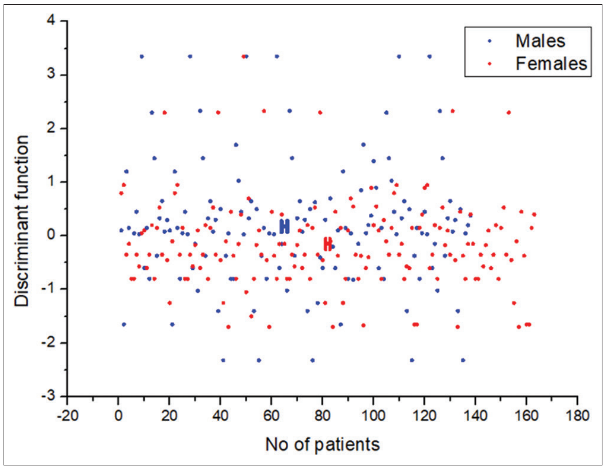
Cutoff score

Cutoff score is the mean of the centroid values. The centroids are given as output by the SPSS software for a set of data. There are two centroid values for two groups, that is, one centroid value for males and one centroid value for females in case of MCI. The mean value of these two centroids is the cutoff score. The discriminant function value of each sample is compared with cutoff score. If the discriminant function value is greater than cutoff score, then the corresponding sample is classified as male and if it is less than cutoff score, then the corresponding sample is classified as female. Likewise, the samples are classified into different groups. A graph is developed [Graph 1] which illustrates the discriminant function values of the individual samples.

Graph 1 shows the discriminant function values for 300 samples of MCI. The centroid for males is 0.177 (blue color H-like structure in the middle) and for females, it is −0.149 (red color H-like structure in the middle). The cutoff score or mean of the centroid is 0.014. The discriminant function values above 0.014 are classified as males and values below 0.014 are classified as females. The accuracy of predicting males using MCI, in this case, is 66% and females is 51.8%. Hence, in this graph, we can see the blue dots which represent males should be above 0.014. However, those blue dots that are below 0.014 are wrongly predicted and likewise the red dots representing females should be below 0.014 and those red dots above 0.014 are wrongly predicted. That is why the accuracy is not 100%.

Difference between DA and Regression Analysis

DA and regression analysis are similar in predicting an outcome. The difference between DA and regression analysis is that DA is used when the prediction outcome is a categorical value and predictors are continuous variables, whereas in regression analysis, both predictor and prediction outcomes are continuous variables. For example, the MCI is used for predicting gender. In this situation, predictor variable, i.e., MCI is continuous or parametric in nature and prediction outcome, i.e., gender is non-parametric or categorical in nature. In such situation, DA has to be used for analyzing data. Another example where



Graph 1: Discriminant function values of 300 samples of MCI

regression analysis has to be used is when MCI is used for predicting mandibular molar width, in which both predictor variable and outcome are continuous or parametric in nature.

Difference between DA and Multivariate Analysis of Variance (MANOVA)

The DA is the reverse of MANOVA. The independent variable of MANOVA becomes the dependent variable in DA and the dependent variable of MANOVA becomes the independent variable in discriminant function analysis. For example, a study in which gender of the sample is used for predicting the MCI. The predictor variable in DA, i.e. MCI has become the outcome in MANOVA and prediction outcome in DA, i.e. gender has become the predictor variable in MANOVA.

Difference between DA and Logistic Regression

Discriminant function analysis and logistic regression are similar. Both can be used for same purpose also. The difference is that logistic regression does not have any assumptions about the distribution of data. DA is more powerful than logistic regression when all the assumptions are met. To make it more clear, DA is used when the data are normally distributed and otherwise, if the data are not normally distributed, logistic regression is used. DA can be used with small sample size also which is not possible with logistic regression.[5]

Advantages of DA

- Discriminant function analysis can be used with small sample sizes
- When sample sizes are equal, and homogeneity of variance/covariance holds, DA is more accurate than any other analysis of the same context
- It can be used to find out which variable best predict the outcome.

Disadvantage/limitation of DA

- DA cannot be used when the data are not normalized.

Conclusion

Discriminant function analysis is a statistical analysis used for predicting the accuracy of a classification system or predictor variables. It has various applications in dental and medical research field as it can be used for validating the newly developed classification system or predictors which can categorize samples into different groups. This paper simplifies DA with giving an example for the better understanding.

References

1. Indian Agricultural Research Institute. Discriminant Analysis: An Overview Venkatesh P. Scientist, Division of Agricultural Economics. New Delhi: Indian Agricultural Research Institute; 2019. Available from: http://www.iari.res.in. [Last accessed on 2019 Oct 2].
2. Cramer D. Advanced Quantitative Data Analysis (Understanding Social Research). 1st ed. United Kingdom: Open University Press; 2003.
3. Ramayah T, Ahmad NH, Halim HA, Zainal SR, Lo MC. Discriminant analysis: An illustrated example. Afr J Bus Manag 2010;4:1654-67.
4. Poulsen J, French A. Discriminant Function Analysis. San Francisco: San Francisco State University; 2008. Available from: http://www.online.sfsu.edu. [Last accessed on 2019 Oct 2].
5. Alayande S, Adekunle B. An overview and application of discriminant analysis in data analysis. IOSR J Math 2015;11:12-5.