Mirko Savić
Dejan Brcanov
Stojanka Dakić

Discriminant Analysis – Applications and Software Support

Article Info:

Management Information Systems,
Vol. 3 (2008), No. 1,
pp. 029-033

Received 12 Januar 2008
Accepted 24 April 2008

UDC 311.42:004

Summary

Discriminant analysis is a tool for classifying new observational units into defined segments on the basis of the discriminant model. Also, this method is used to provide a visual representation of the structure when large numbers of variables, groups and relations exist. On the basis of this way of interpretation, complex structure is simplified and better for understanding. This method is useful in the process of grading and measuring of product and service quality and can be used for solving different problems in this field

Key words

Multivariate statistics, Statistical software, Statistical computing

Introduction

Discriminant analysis is statistical, multivariate method which can be used in the research areas where dependent variable (Y) consists of categories rather than a continuous metric scale (interval or ratio). Discriminant analysis is a tool for classification of new observational units, especially new respondents, into groups or categories in which most probably it belongs. Number of groups or categories must be more than two. Also, discriminant analysis as a result gives the probability of group membership. Classification is conducted on the basis of measured values for group of characteristics for each observational unit separately. That means that each observational unit possesses certain number of characteristics than could be measured and those values vary from one observational unit to the other.

The goal of discriminant analysis is to construct the model on the basis of observational unit's variation. On the basis of the discriminant model the classification of new observational units into the groups or categories will be conducted. Some authors (Timm, 2002) indicate that goals of a discriminant analysis are to construct a set of discriminants that may be used to describe or characterize group separation based upon a reduced set of variables, to analyze the contribution of the original variables to the separation, and to evaluate the degree of separation.

There are other models which can solve the same problem, but discriminant analysis is specific because of the following differences:

- models themselves,
- number of groups in the dependent variable,
- assumptions and requirements for input data (X),
- ease of application.

Discriminant analysis gets its name from the way the model is constructed. Each of the groups in the dependent variable must have a set of measurements. For example, if there are five different brands of mobile phones, each brand must have its own set of ratings on performance (X) by a sample of respondents. Each respondent could evaluate all five mobile phone brands, each respondent could evaluate only the brands he or she has used or separate samples of respondents could rate each brand. On the basis of collected data discriminant model calculates set of coefficients for each brand separately. The set of coefficients for each brand distinguishes or discriminates the brand among the others.

With five brands of mobile phones, computer develops five sets of coefficients. Multiple regression method will develop single set of coefficients for the same problem.

1. Comparison with Other Techniques

For most applications for the prediction of group membership, discriminant analysis is much easier to use and to understand, especially if there are more than two groups. Logit regression, for example, requires separate regression models for each of the groups except one, which is determined by subtraction. With discriminant analysis

everything is in one model which simultaneously classifies respondents into groups and calculates membership probability for each group separately.

With this situation, why would analyst ever want to use any other method when he could just use discriminant analysis? When there are only two groups in the dependent variable (Y), the following three methods produce essentially the same regression coefficients for the same data sample: ordinary least square regression, logit regression, and discriminant analysis (Myers & Mullet, 2003). The choice among these methods is simply a matter of what kind of prediction output or format researcher wants. Problem occurs when the dependent variable has more than two groups or categories and when required assumptions about the characteristics of the independent variables cannot be met.

2. Classification Coefficients

Classification coefficients in discriminant analysis are used to score each new case for membership prediction purposes. The coefficients numerically show the nature of the differences among groups or categories in the dependent variable. The larger the coefficient for a variable for a particular group, the more that variable characterizes that group. If there are five groups in the dependent variable, each new case is evaluated five times by means of the coefficients for each of the groups in turn. After that, the observational unit is classified in the group whose coefficients produce the highest score.

Coefficients for each group defined classification functions. In general, the functions have the following form:

$$R_i = b_{i0} + b_{i1}X_1 + b_{i2}X_2 + b_{i3}X_3 + .. . + b_{ij}X_j,$$

where:

R_i – score for one respondent for group i,

b – classification coefficient for each independent variable for group i,

X – independent variable,

b_{i0} – a constant for group i,

i – ordinal number for groups or categories, $i = 1, 2 \dots , n,$

j – ordinal number for independent variables, $j = 1, 2, ..., m$.

It is obviously simple linear model, similar to that for ordinary regression. However, whereas ordinary regression requires only one set of coefficients, discriminant analysis requires a set for each group in the dependent variable (i).

In addition to calculating the scores for classification purposes, the computer calculates the probability that a given case belongs in each of the groups. On the basis of calculated probability we can conclude into which group case might belong in other than the one into which is classified. Also, calculated probabilities indicate groups that have low probabilities of membership and that the case clearly does not belong in.

3. Assumptions and Requirements for Application of Discriminant Model

In comparison with other multivariate method, discriminant analysis has more restrictive assumptions and requirements. The major assumptions are:

1. There must be two or more groups or categories.
2. There must be at least two respondents (observational units) per group.
3. The number of discriminating variables in the model must be less than the total number of respondents minus 2.
4. Discriminating variables are measured at the interval or ratio scale level. Dummy variables also work well.
5. No discriminating variable may be a linear combination of the other discriminating variables.
6. The covariance matrices must be approximately equal for each group, unless special formulas are used.
7. Each group has been drawn from a population with normal distribution on the discriminating variables.
8. Group sizes should not be too different. If they are, the units will tend to have overpredicted membership in the largest group.

In praxis, some researches have shown that discriminant analysis is quite robust, which means that it can tolerate some failure to meet assumptions 6 and 7.

4. Software Support

Discriminant analysis could be applied in great number of research areas. Very often it is used for two purposes:

- to predict group membership for new cases, especially when there are more than two groups,
- to provide a visual representation of structure underlying the relationships among large numbers of variables and groups.

Some computer software packages have separate programs for each of these two application, for example – SAS. In order to evaluate and measure the quality of products and services it is possible to efficiently use discriminant analysis, especially for classification of costumers (respondents, observational units) from the aspect of their preferences toward some product or service.

Manly (2005, p. 122) indicates that major statistical packages generally have a discriminant function option that applies for both applications based on the assumption of normally distributed data. Manly wa rns that manuals may have to be studied carefully to determined precisely what is done by any individual program because the details of the order of calculations, the way the output is given, and the terminology vary considerably. There is also resourceful literature about these problems because many authors gave their contribution to software support of multivariate statistical methods in recent years (Marques de Sa, 2003), (Myers & Mullet, 2003), (Manly, 2005).

Visual displays are important aid in discrimination and classification. Sophisticated computer graphics allow us to visually examine multivariate data in two or three dimensions. Johnson and Wichern (2007, p. 649) state that software support is very useful in graphical presentation of discriminant models. Use of visual displays is likely to increase as the hardware and associated software become readily available.

5. Prediction of Group Membership and Classification of New Observational Units into the Groups or Categories

Big supermarket chain has started the research for market segmentation. Customers were surveyed and study asked respondent to indicate how much they wanted each of 36 benefits that might be delivered by a supermarket chain. The respondent should evaluate each benefit on a 10-point scale.

After the data collection, cluster analysis was conducted and four segments were formed on the basis of customer's preferences. Each segment has show statistically significant differences to the others.

Although the knowledge about market segments is useful by itself for strategic planning, supermarket chain is interested in the model for classification of new customers into four defined segments. Through identifying the group membership for each customer in time, the chain will be in position to adjust services and promotional activities. Discriminant analysis offers the solution for this problem. This is an example of an application at the operating level.

In the survey, each customer should evaluate 36 supermarket benefits with grades from 1 to 10. First, these grades were used to form market segments on the basis of cluster analysis. After that, the discriminant analysis is used to develop membership classification model for membership prediction of new customer into existing segments. Ideally, all new respondents would be asked to rate all 36 benefits, just as the original sample of respondents did. However, management believed that this would be far too much of a burden for most businesses and asked if a much smaller number of attributes could provide reasonable classification accuracy.

To reduce the number of question in survey, stepwise discriminant analysis was conducted. It proceeds in the same general manner as stepwise regression. The computer first selects the single benefit that best distinguishes or discriminates among the segments. After that, computer selects the next benefit that adds the most discrimination to the first one selected. This procedure is repeated until no other variable improves classification accuracy.

At each step, computer calculates the membership prediction accuracy for selected group of benefits. The procedure will be conducted until classification probability changes significantly at a specified level of statistical confidence. After each step, the computer prints out a classification matrix that shows predicted versus empirical group membership for the insight of model accuracy.

6. Classification Accuracy of Discriminant Analysis

Table 1 shows what happens when we score each respondent in the original sample by applying the coefficients of this prediction model to all 36 benefits evaluated in the research. Because there are four segments, the computer scores each respondent four times using the set of coefficients for each group (segment) in turn. After that, the computer program classifies a particular respondent into the segment who's set of coefficients produce the highest score.

In practice, it is not usual to score a sample of respondents using sets of coefficients that were developed on this same sample, but on the new observational units, because it overestimates

prediction accuracy. However, the program we used has control procedure which corrects estimate of accuracy, so it became unbiased. The other way of getting unbiased estimates is to randomly select 50% to 70% of the original sample to develop the prediction model and then to apply this model to the remaining cases.

Table 1 Classification of respondent into 4 segments on the basis of evaluation of 36 supermarket benefits

Actual Segment	Percent of Accuracy	Predicted segment on the basis of discriminant model				Total
		1	2	3	4	
1	97,9	420	8	1	0	429
2	96,4	0	352	11	2	365
3	86,6	2	3	258	35	298
4	93,1	3	11	0	190	204
Total	94,1	425	374	270	227	1296

Using all 36 ratings for supermarket benefits from each respondent, Table 1 shows that the model correctly classifies 94,1% of customers through discrimination model. Even though we cannot obtain all 36 ratings for each new customer, this analysis can serve as a benchmark against which to compare prediction accuracy using many fewer attribute ratings. Also, accuracy this high shows us that the segments are indeed different from one another in terms of the services they want from the supermarket and that we can predict segment membership with reasonable accuracy.

Table 2 Classification of respondent into 4 segments on the basis of evaluation of 3 supermarket benefits

Actual Segment	Percent of Accuracy	Predicted segment on the basis of discriminant model				Total
		1	2	3	4	
1	90,4	312	18	11	4	345
2	84,5	0	262	18	30	310
3	34,2	41	73	88	55	257
4	64,5	36	41	0	140	217
Total	71,0	389	394	117	229	1129

In the next step, model was constructed like it was asked for only 3 discriminating benefits. The results of this step are shown in the Table 2.

The number of questions could be reduced to only 3, which will result in easier and faster data collection and membership classification. On the other hand, it is also logical that accuracy of the model will be reduced because the criterion for membership classification is also reduced.

For classification into first two segments we have rather high percent of accuracy, 90,4% for the first segment and 84,5% for the second. If we are interested for membership classification of new customers into only these two segments, discriminant analysis could be stopped here because percent of accuracy is high enough to tell that discriminant model is reliable enough for membership classification into first two segments.

However, if we are interested in classification of new customers also into third segment, it is necessary to correct classification criteria, because percent of accuracy for third segment is rather low (34,2%). It actually means that three questions about three benefits to new customers are not enough to reliably evaluate group membership. It is necessary to enlarge the number of questions (benefits) in the survey.

Table 3 shows the results of discriminant model based on the rates of 5 most important supermarket benefits. If membership prediction is conducted on the basis of five benefits, the model will have membership prediction accuracy of 81%. It is obvious that percent of accuracy for third segment is significantly larger if we have 5 benefits in the model than only three.

Table 3 Classification of respondent into 4 segments on the basis of evaluation of 5 supermarket benefits

Actual Segment	Percent of Accuracy	Predicted segment on the basis of discriminant model				Total
		1	2	3	4	
1	92,5	321	18	4	4	347
2	84,7	0	265	18	30	313
3	66,9	21	19	192	55	287
4	75,8	16	37	0	166	219
Total	81,0	358	339	214	255	1166

If membership classification into 4 segments is conducted on the basis of 6 benefits, in other words, if we add one more benefit in the survey, total percent of accuracy will be 82,4%. By grading 7 benefits the percent of accuracy will have very small increase. At this point, it is up to management to decide about the number of benefits included in the model. It is necessary to make a balance between number of questions in the survey and achieved percent of accuracy. This is the trade-off between slightly increased prediction accuracy and the increased time, effort and patience of respondents.

In this case, for management of supermarket chain it would be good enough to choose the survey with 5 questions with grading only 5 benefits because the gained percent of accuracy (81%) is high enough and not much smaller

comparing to the survey with 6 questions. On the other hand, the 5 question survey will not be to long for the new respondents because they have to grade only 5 supermarket benefits.

Through this example, it is obvious that step by step discriminant analysis is not only good method for membership classification, but is also the simplified and reliable procedure for evaluation of new observational units. Because of all this, discriminant analysis is very useful tool for evaluation of product and service quality especially if it is the part of integral approach implemented through user oriented statistical software.

Understanding of statistical software is very important in order to create an integral approach to the knowledge. Trninić and Tumbas (2007, p. 533) argue that modern trends imposed by competitive environment need knowledge as the main product and quality management systems develop integrative components in order to monitor and evaluate the key activities of use, dissemination, creation, and preservation of knowledge.

References

Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. New Jersey: Pearson Education Inc.

Manly, B. F. (2005). Multivariate Statistical Methods - A primer (3rd Edition ed.). New York: Chapman & Hall/CRC.

Marques de Sa, J. P. (2003). Applied Statistics Using SPSS, Statistica and Matlab. Chicago: Springer.

Myers, J. H., & Mullet, G. M. (2003). Managerial Applications of Multivariate Analysis in Marketing. Chicago: American Marketing Association.

Timm, N. H. (2002). Applied Multivariate Analysis. Chicago: Springer.

Trninić, J., & Tumbas, P. (2007). Knowledge Management Systems and Quality Processes Management. Dependability and Quality Management, 10th International Conference (pp. 529-534). Prijevor: Research Center of Dependability and Quality Management.

Mirko Savić

University of Novi Sad, Faculty of Economics Subotica
Segedinski put 9-11
24 000 Subotica
Serbia

Email: savicmirko@ef.uns.ac.rs

Dejan Brcanov

University of Novi Sad, Faculty of Economics Subotica
Segedinski put 9-11
24 000 Subotica
Serbia

Email: brcanovd@ef.uns.ac.rs

Stojanka Dakić

University of Novi Sad, Faculty of Economics Subotica
Segedinski put 9-11
24 000 Subotica
Serbia

Email: stojankad@ef.uns.ac.rs