ISSN 2581-3463

ASIAN JOURNAL OF MATHEMATICAL SCIENCES

AJMS

Asian Journal of Mathematical Sciences

RESEARCH ARTICLE

A Discriminant Analysis Procedure for Loan Application using Ranked Data

I. Akeyede, F. G. Ibi, D. T. Ailobhio

Department of Statistics, Federal University of Lafia, PMB 146, Lafia, Nigeria

Received: 01-12-2020; Revised: 01-01-2021; Accepted: 20-01-2021

ABSTRACT

Discriminant analysis is a multivariate techniques concerned with separating distinct sets of objects to previously defined groups. The procedure we intend to develop in this paper uses a less biased statistical technique than the conventional discriminant analysis and parallels to ranking procedure used by loan officers. A variety of univariate and multivariate statistical procedures as well as comprehensive validation methods are used to develop and test a best model. The resulting model obtains using SPSS, a statistical software to run the data, provides more accurate classification than other studies have shown, without violating assumption regarding discrimination.

Key words: Box test, collinearity, correlation, discriminant analysis

INTRODUCTION

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a separate procedure, it is often employed on a time basis to investigate observed differences when causal relationships are not well understood (Conover and Iman, 1980).[1] Discriminant analysis assumes that prior defined groups are usually distributed. If this assumption is not satisfied, certain parts of the analysis may be biased. Typically, logarithm data transformation is made and used in the discriminant procedure. However, such transformations may affect the interrelationships among the variables. An alternative transformation uses ranks (ordinal data), and this has been shown to perform comparably to conventional discriminant analysis which uses interval (cardinal) data while mitigating the multivariate normality problem (Harold *et al*., 2018).[2,3]

A loan application is commonly processed through two stages of evaluation and ranking. The first occurs when the loan officer counsels with each applicant(s) and ranks the applicants from best to worst using factors that the institution considers critical to delineate acceptable borrowers from those that should be rejected. The second stage occurs when the credit committee receives the higher ranked applications from the loan officer and conducts its own ranking from best to worst. The higher ranked applications in the final stage are then approved in order until the supply of loanable funds is exhausted. This is especially true during periods when demand for loan money is high.[4,5] Moore and Smith (2018) presented a discriminant analysis approach that attempted to assess the markets for evidence of discriminatory lending practices on the part of financial institutions. The purpose of their study was to develop a model that allows both lenders and regulators to assess the equity of lending patterns within a given market. Due to the possible bias resulting from the violation of the normality assumption of discriminant analysis, Ingram and Fragler (1982)[5] used discriminant analysis, probit and logit models in a study of mortgage lending discrimination. None of these studies employing discriminant analysis considered ordinal data or the rank transformation approach. However, due to the problems associated with parametric discriminant analysis assumptions, and the ranking procedure used by lending institutions, such a procedure should be considered.

Address for correspondence:
Imam Akeyede
E-mail: akyede.imam@science.fulafia.edu.ng

Rank transformation discriminant analysis also has been used in empirical research.[6,7] Perry, Harry, and Peter (1985)and Perry and Cronan (1982) used this procedure to develop more accurate bond rating models.[8] Cronan *et al*. (2013) also used the procedure as an effective data editing methodology. Much of the lending literature contains research that has been conducted with the objective of detecting discriminatory practices within the models used by lenders to evaluate loan applications. The objective of this paper is to derive a model from ordinal ranking that can delineate accepted from rejected applications. Although detection of discrimination is not a primary objective, the factors contained within the model can be examined closely to determine if it is present. A comparison of ordinal discriminant analysis with the cardinal discriminant analysis found in the literature is useful since it illustrates the difference in results.

The hypothesis tested in this paper is that discriminant analysis using ordinal data will produce a better model for judging future applicants, than the traditional discriminant analysis using cardinal data.[7] The intent of the paper is to provide a less biased procedure for uncovering the critical factors used by a lender to delineate acceptable applications from those that should be rejected and to argues that a better model can be produced since its more closely replicates the loan evaluation process and is statistically more appropriate.

METHODOLOGY

A sample of 350 applications was obtained from the institutions. The data set consists of 200 rejected applications and accepted application. Variables include risk return variables and discriminatory variables. The information for these variables was taken from their applications at the financial institutions. Predictor variables used in the development of the model include the following risk return measures: Applicant's credit rating, applicant's occupation, applicant's tenure in occupation, loan to value ratio, neighborhood crime rate, remaining economic life, applicant's total monthly payment to income ratio, and years to maturity.

Discriminant analysis is the classification of an observation X_0, possibly multivariate, into one of several populations Π_1, Π_2… Π_k which each have density functions. If these densities can be assumed to be normal with equal covariance matrices, then Fisher's linear discriminant function (LDF) method is used. If the matrices are unequal, a quadratic discriminant function (QDF) is appropriate. These methods assume multivariate normality.[6,8] Conover and Iman (1980) and Moore and Smith (1975) suggested a transformation that applies to all distributions equally well, that is, the rank population. Using their terminology, let X_{ij} be the jth observation factor from population i,j =1,2,…n and i=1,2,…k. The p components of X_{ij} are denoted X_{ijm}, m = 1,2,…p. The rank transformation method involves ranking the mth component of all observations X_{ij} from the smallest, with rank 1, to largest, with rank $N=n_1 + n_2 +...+ n_k$. Each component m=1 to m=p is ranked separately. Simply stated, each variable value of the multivariate sample is replaced by its rank from 1 to n of all the groups combined. Sample means and covariance matrices are computed on the ranks and the traditional LDF and QDF are used, hence the rank linear discriminant function. The rank transformation tends to minimize the outlier contamination problem and the non-normality problem caused by outliers. No knowledge of outliers or distribution form is necessary.

RESULTS AND ANALYSIS

Statistical software, SPSS 21 was used to run the data; the various results and the analysis are provided in Table 1:

The classification functions are used to assign cases to groups. There is a separate function for each group. For each case, a classification score is computed for each function. The discriminant model assigns the case to the group whose classification function obtained the highest score. The coefficients for years with current employer

Table 1: Classifying customers as high or low credit risks

Classification function coefficients	Previously defaulted	
	No	Yes
Years with current employer	0.270	0.113
Years at current address	0.165	0.076
Debt to income ratio (×100)	0.261	0.413
Credit card debt in thousands	−0.640	−0.259
(Constant)	−3.624	−4.342

Fisher's linear discriminant functions

and years at current address are smaller for the yes classification function, which means that customers who have lived at the same address and worked at the same company for many years are less likely to default. Similarly, customers with greater debt are more likely to default.

The within-groups correlation matrix in Table 2 shows the correlations between the predictors. The largest correlations occur between credit card debt in thousands and the other variables, but it is difficult to tell if they are large enough to be a concern. Look for differences between the structure matrix and discriminant function coefficients to be sure.

The group statistics Table 3 reveals a potentially more serious problem. For all four predictors, larger group means are associated with larger group standard deviations. In particular, look at debt-to-income ratio (×100) and credit card debt in thousands, for which the means and standard deviations for the yes group are considerably higher. In further analysis, you may want to consider using transformed values of these predictors.

Box's Test of Equality of Covariance Matrices

The ranks and natural logarithms of determinants in Table 4 printed are those of the group covariance matrices tests null hypothesis of equal population covariance matrices. Box's M tests the assumption of equality of covariances across groups. Log determinants are a measure of the variability of the groups. Larger log determinants correspond to more variable groups. Large differences in log determinants indicate groups that have different covariance matrices. Since Box's M is significant, you should request separate matrices to see if it gives radically different classification results. See the section on specifying separate groups covariance matrices for more information.

Assessing Contribution of Individual Predictors

There are several tables that assess the contribution of each variable to the model, including the tests of equality of group means, the discriminant function coefficients, and the structure matrix.

Table 2: Checking collinearity of predictors

Pooled within-groups matrices				
Correlation	**Years with current employer**	**Years at current address**	**Debt-to-income ratio (×100)**	**Credit card debt in thousands**
Years with current employer	1.000	0.189	0.043	0.569
Years at current address	0.189	1.000	0.081	0.212
Debt-to-income ratio (×100)	0.043	0.081	1.000	0.400
Credit card debt in thousands	0.569	0.212	0.400	1.000

Table 3: Checking for correlation of group means and variances

Group statistics				
Previously defaulted	**Mean**	**Std. deviation**	**Valid N (list wise)**	
			Unweighted	**Weighted**
No				
Years with current employer	10.7188	7.62365	96	96.000
Years at current address	9.7708	7.55399	96	96.000
Debt-to-income ratio (×100)	8.7927	5.08223	96	96.000
Credit card debt in thousands	1.4678	1.77583	96	96.000
Yes				
Years with current employer	5.2564	6.02939	39	39.000
Years at current address	5.5897	5.53799	39	39.000
Debt-to-income ratio (×100)	16.6538	9.16730	39	39.000
Credit card debt in thousands	2.3095	2.65118	39	39.000
Total				
Years with current employer	9.1407	7.59533	135	135.000
Years at current address	8.5630	7.26431	135	135.000
Debt-to-income ratio (×100)	11.0637	7.41173	135	135.000
Credit card debt in thousands	1.7110	2.09179	135	135.000

Each test displays the results of a one-way ANOVA for the independent variable using the grouping variable as the factor. If the significance value is greater than 0.10, the variable probably does not contribute to the model. According to the results in Table 5, every variable in the discriminant model is significant. Wilks' lambda is another measure of a variable's potential. Smaller values indicate the variable is better at discriminating between groups. Table 5 suggests that debt-to-income ratio (×100) is best, followed by years with current employer, credit card debt in thousands, and years at current address. The standardized coefficients allow you to compare variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability. Table 6 downgrades the importance of debt-to-income ratio (×100), but the order is otherwise the same.

The structure matrix shows the correlation of each predictor variable with the discriminant function. The ordering in the structure matrix is the same as that suggested by the tests of equality of group means and is different from that in the Table 7. This disagreement is likely due to the collinearity between years with current employer and credit card debt in thousands noted in the correlation matrix. Since the structure matrix is unaffected by collinearity, it's safe to say that this collinearity has inflated the importance of years with current employer and credit card debt in thousands in the standardized coefficients table. Thus, debt-to-income ratio (×100) best discriminates between defaulters and non-defaulters.

Table 4: Checking homogeneity of covariance matrices

Log determinants		
Previously defaulted	**Rank**	**Log determinant**
No	4	11.992
Yes	4	11.863
Pooled within-groups	4	12.370

Table 5: Test results

Parameter	Values
Box's M	55.091
F	
Approx.	5.275
df1	10
df2	25324.568
Sig.	0.000

Table 6: Tests of equality of group means

Tests of equality of group means	Wilks' lambda	F	df1	df2	Sig.
Years with current employer	0.893	15.943	1	133	0.000
Years at current address	0.931	9.790	1	133	0.002
Debt-to-income ratio (x100)	0.767	40.363	1	133	0.000
Credit card debt in thousands	0.966	4.611	1	133	0.034

Table 7: Standardized canonical discriminant function coefficients

Standardized canonical discriminant function coefficients	Function
Years with current employer	−0.682
Years at current address	−0.379
Debt-to-income ratio (x100)	0.598
Credit card debt in thousands	0.475

Assessing Model Fit

In addition to measures for checking the contribution of individual predictors to your discriminant model, the discriminant analysis procedure provides the eigenvalues in Table 8 and Wilks' lambda Table 9 for seeing how well the discriminant model as a whole fits the data.

First canonical discriminant functions were used in the analysis the eigenvalues table provides information about the relative efficiency of each discriminant function.

When there are two groups, the canonical correlation is the most useful measure in the table, and it is equivalent to Pearson's correlation between the discriminant scores and the groups.

Wilks' lambda is a measure of how well each function separates cases into groups. It is equal to the proportion of the total variance in the discriminant scores not explained by differences among the groups. Smaller values of Wilks' lambda indicate greater discriminatory ability of the function. The associated Chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The small significance value indicates that the discriminant function does better than chance at separating the groups.

Table 8: Eigenvalues

Function	Eigenvalue	% of variance	Cumulative %	Canonical correlation
1	0.573[a]	100.0	100.0	0.603

Table 9: Wilks' lambda

Wilks' lambda				
Test of function (s)	Wilks' lambda	Chi-square	df	Sig.
1	0.636	59.324	4	0.000

Table 10: Structure matrix

Structure matrix	Function
Debt-to-income ratio (×100)	0.728
Years with current employer	−0.457
Years at current address	−0.358
Credit card debt in thousands	0.246

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions. Variables ordered by absolute size of correlation within function

Table 11: Model validation

Classification results[b,c,d]	Previously defaulted	Predicted group membership		Total
		No	Yes	
Cases selected				
Original				
Count	No	80	16	96
	Yes	7	32	39
%	No	83.3	16.7	100.0
	Yes	17.9	82.1	100.0
Cross-validated[a]				
Count	No	79	17	96
	Yes	8	31	39
%	No	82.3	17.7	100.0
	Yes	20.5	79.5	100.0
Cases not selected				
Original				
Count	No	39	10	49
	Yes	5	11	16
	Ungrouped cases	102	48	150
%	No	79.6	20.4	100.0
	Yes	31.3	68.8	100.0
	Ungrouped cases	68.0	32.0	100.0

a. Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case. b. 83.0% of selected original grouped cases correctly classified. c. 76.9% of unselected original grouped cases correctly classified. d. 81.5% of selected cross-validated grouped cases correctly classified

The classification Table 11 shows the practical results of using the discriminant model. Of the cases used to create the model, 32 of the 39 people who previously defaulted are classified correctly. Eighty of the 96 non-defaulters are classified correctly. Overall, 83.0% of the cases are classified correctly. About 76.9% of these cases were correctly classified by the model. This suggests that, overall; your model is in fact correct about 3 out of 4 times. The 150 ungrouped cases are the prospective customers, and the results here simply give a frequency table of the model predicted groupings of these customers.

CONCLUSION

In this paper, the hypothesis examined was that a model could be produced that would better explain the factors used by loan officers to delineate acceptable loan applicants from those that should be rejected as shown Table 10. The purpose of the study was to better replicate the ranking conducted by the loan officer and the credit committee before approving or rejecting the loan request. The resulting model that was obtained from the analysis, by the use of discriminant analysis on ordinal data from a set of 350 accepted and rejected loan application, produced a better model. The factors contained in the model developed in this paper did not violate any law or assumption on discrimination, but were discriminatory in the sense that the variables utilized for loan approval or non-approval were financial in nature. The model correctly classified 83% of the loan applications. This implies that if the model is used by lenders, they should correctly classify a loan application as acceptable or unacceptable 83% of the time. This study used a less biased statistical procedure and comprehensive validation procedures. No evidence of discrimination overall was found, however, the classification results suggest that certain discriminatory factors may possibly exist when rejecting a loan application. It should be noted that the variables used in the study were obtained from the applications. Some of these variables represented what they intended, but others were surrogates for risk. The discriminant analysis uses only the variables in the model, was as the lender may use additional values on the applications as additional measures of risk. Furthermore, lenders are conservative. It is their right to ensure an adequate return at a minimum level of risk. The applications that the discriminant analysis model deemed marginally acceptable were considered too risky by the lender, therefore, giving a lower classification rate and the appearance of discrimination. Even if the exact variables were used by the lender and the model, the weighting could by slightly different and again result in a lower rate. The results of this study suggest that lenders are not discriminating; they are doing as expected, by being conservative by rejecting marginal loan applications.

REFERENCES

1. Conover WJ, Iman RL. The rank transformation as a method of discrimination with some examples. Commun Stat Theory Methods 1980;A9:465-87.
2. Cronan P, Epley F. Note on discriminant functions. Biometrika 2013;31:218-20.
3. Harold B, Robert CS, Lewis M. Discrimination in mortgage lending. Am Econ Rev Pap Proc 2018;68:186-91.
4. Harry WR, Peter G. Measuring mortgage delinquency and its determinants. Ann Reg Sci 1985;17:25-34.
5. Ingram FJ, Fragler AE. A test for discrimination in a mortgage market. J Bank Res 1982;12:116-24.
6. Moore K, Smit W. A rank order approach to discriminant analysis. Proc Bus Econ Sec Am Stat Assoc 2018;5:451-5.
7. Thomas PI, Joseph WM. Statistical inference based on ranks. Psychometrika 2012;43:68-79.
8. King AT. Discrimination in Mortgage Lending: A Study of Three Cities. New York: New York University; 1981.