SCITECH
RESEARCH ORGANISATION

Volume 12, Issue 1

Published online: May 20, 2017|

Journal of Progressive Research in Mathematics
www.scitecresearch.com/journals

Statistical and Operations Research Discrimination Methods for Solving Recovery and Non-Recovery Oil Problems

Taghreed Abdul-Al-Razek Abdul-Motaleb Al-Said, Ph.D.
A lecturer of Statistics, Department of Statistics,
Faculty of Commerce (Women's' Branch),
AL AZHAR University, Cairo, Egypt.

Abstract.

Discriminant analysis is one of the most classical classification procedure used to discriminate and differentiate groups as possible by using one or more attributes. It also assigns observations to one of the pre-defined and separated groups. There are many parametric discriminant statistical methods such Fisher's, quadratic, and logistic methods, and non-parametric operations research methods such as maximize minimum of deviations MMD, and minimize sums of deviation MSD methods to discriminate the difference between groups and classify new observations to the suitable group. Operations research methods are fixable and do not require any assumptions as statistical methods. This study introduces a comparison between three statistical methods and two linear programming models for solving recover and non-recover oil problems.

Keywords: Logistic regression; linear models; leave-one-out cross validation.

1. Introduction

The discrimination analysis is the suitable tool that classify objects into one of two, or more mutually exclusive groups on the bases of measurements.[5] This multivariate analysis techniques focus on the association between multiple independent variables, and a categorical dependent variable by forming a composite of independent variables. It can be used to discriminate between two or more pre-existing separated groups of subjects as possible, and can predict the new group membership to its suitable group. The simplest type of discriminant analysis has only two groups. [1]

The basic statistical methods assumed normality of the measurements, and also equality variances-covariances matrices for groups along many other assumptions. These assumptions are frequently violated and flexible statistical methods such as quadratic and logistic methods can be used for dealing with these problems, along the operation researches methods that play important alternative methods that treating and solving these problems. This study introduces a comparison between three statistical and two operations researches methods in a new field, the oil and gas field using a real data set from the Schlumberger Company, one of greatest company of Gas and Oil in the Middle East. It will reviewed five of frequently used methods in statistics and operation researches for solving the recovery and non-recovery oil problems. Section (2) provides a suitable review of statistical and operations research methods. Section (3) introduces an application of some discrimination methods for solving recovery and non-recovery Oil data problems. Section (4) includes the analysis and recommendation of the study. Finally, at the end of the study, there is the references section in section (5).

2. The Discriminant Methods

This section has the three suggested statistical methods in analyzing the real data used in the application, the Fisher discriminant LDA method, Quadratic discriminant QDA methods, and the logistic discriminant LoDA method, and two of linear programming discriminant models: the maximize minimum of deviations MMD method, and the minimize sums of deviation MSD method.

2.1 The Fisher discriminant method:

The Fisher linear discriminant analysis, LDA is one of the famous and first parametric methods that is used for discriminating groups and allocating observations to its suitable group. It is introduced by Fisher in (1936). The LDA method maximizes correct classification, and minimizes the probability of misclassification under normality, and equal variances-covariances assumptions. The LDA function can be defined as follows:

$$Y = a_0 + a_1 X_{i1} + a_2 X_{i2} + \cdots + a_p X_{ip} = \sum_{j=1}^{p} a_j X_{ij} \qquad (1)$$

Where the j-th coefficient values area_j, j=1,….,p. It is the unknown weight for attributes X_{ij}. It provides a maximum separation between the discriminant scores that can be defined as follows:

$$a = S^{-1}(\bar{x}_1 - \bar{x}_2) \qquad (2)$$

The sample mean for the observations in group I is $\bar{x}_i$, i=1,2, and S is the unbiased estimator of variance-covariance, Σ matrix. The LDA can classify new observations to the correct group membership on the basis of the Mahalanobis distance function that classify x_i to group 1 if $Y = \acute{a}x_i > m$, and classify x_i to group 2 if $Y = \acute{a}x_i \leq m$, $m = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 + \bar{x}_2)$ is the midpoint between the two population means.[10][14]

2.2 The quadratic discriminant method:

The quadratic discriminant, QDA analysis was proposed by Smith in (1974) and assumed one assumption, the equality of variance-covariance matrices of the groups under study. The quadratic discriminant function for two groups can be formed as follows:

$$Q(x) = ln\frac{f_1(x)}{f_2(x)} \qquad (3)$$

$$= \frac{1}{2}\left\{(x - \bar{x}_2)' S_2^{-1}(x - \bar{x}_2) - (x - \bar{x}_1)' S_1^{-1}(x - \bar{x}_1)\right\}$$

Where $f_i(x)$ is the density of the observation x with estimated means $\bar{x}_i$, and estimated variances-covariances are S_i^{-1}, i=1,2. The new observation assigns to group1 if the quadratic function greater than the cut off value c, or assign the observation to group2 otherwise. When the prior probability generated by first distribution $[p_1\, p_2 = 1 - p_1]$, then optimal discriminate rule minimize the expected probability of misclassification, that classifies observation x to group1 if $Q(x) > \ln\left(\frac{p_2}{p_1}\right)$, and to group2 otherwise. [8][16]

2.3 The logistic discriminant method:

The logistic discriminant method is one of the generalized linear models, where the dependent variable is assumed to follow a binomial distribution, and has the log odds; logit when logistic regression is used for discriminant analysis; it is often referred to as logistic discrimination. The normality of the sample, and the homogeneity of the variance-covariance matrices of the groups are relaxed in LoDA. [4] The logistic models have been widely used in many fields such as social, medical, biological, food science, and control researches. The posterior probability of belonging to group1 is π_1, and $\pi_2$for group 2 through the logistic function. The general formulation for linear logistic discriminant function can be defined as follows:

$$\log\left(\frac{\pi_1}{\pi_2}\right) = \acute{\beta} X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{4}$$

Where the logistic regression coefficients are $\beta's$. [9] The predicted posterior probability of an observation to belong to group1 can be defined as follows:

$$\pi_1 = \frac{e^{\beta X}}{1+e^{\beta X}} \tag{5}$$

This classification model can make predictions of new observations, and can classify observation X_i to group 1 if the posterior predicted probability is larger than the cut off value; otherwise, classify the observation to group 2. The costs of misclassification for the two groups are different, and can be assumed to be equal. [15] The LoDA method is less efficient than the LDA method when the normality and variances-covariances assumptions are satisfied. The two methods, LDA and LoDA have similar accuracy when sample size is large relative to the number of predictors.[6][9][13]

2.4 The maximize minimum of distance method

The maximize minimum of deviations, MMD method was suggested by Freed, and Glover in (1981) to solve the classification, and discrimination problems. It has the following form:

$$\text{Maximize d} \tag{6}$$

subject to:

$$\sum_{j=1}^{p} w_j x_{ij} + d_i \le c \quad \text{for all } i = 1,2,\ldots, n_1 \text{in group 1}$$

$$\sum_{j=1}^{p} w_j x_{ij} - d_i \ge c \quad \text{for all } i = 1,2,\ldots, n_2 \text{ in group 2}$$

$$d_i \ge 0\,,$$

$$\sum_{j=1}^{p} w_j = 1$$

w_j; c are unrestricted in sign

Where w_j is the unknown weights for attributes j: j=1, 2,…, p; x_{ij} is the values for observation i of attribute j, n_1 and n_2 are the number of observations in group 1 and group 2 respectively, d_i is the minimum distance between a case score and cut off value c. The cut off value c is an arbitrarily selected positive constant and does not affect classification accuracy except that it rescales the estimated weights. To avoid the trivial solution, the normalization constraint $\sum_{j=1}^{p} w_j = 1$ added by some researcher to the constraints. An observation will be classified into group1 if its classification score is less than c, or into group2 otherwise. [3)] [7][12]

2.5 The minimize sums of deviation method:

The minimize sums of deviation, MSD method is one of the linear mathematical programming technique proposed by Freed and Glover(1986) to solve discrimination problems in two groups, and has not any assumptions. This method focuses on the minimization of total group overlap. The objective has value zero when the paired groups can be separated by a hyperplane. [11]. The MSD method based on minimizing the total absolute distance of all misclassified entities from a cutting hyperplane which separates the two groups between the score $\sum_j w_j x_{ij}$ of entity i and a real-valued cut off score c .[2][8] The MSD model has the following form:

$$\text{Minimize } Z = \sum_{i=1}^{n_1+n_2} d_i \qquad (7)$$

subject to:

$$\sum_{j=1}^{p} w_j x_{ij} - d_i \le c \qquad \text{for all } i = 1,2,\dots,n_1 \text{ in group 1}$$

$$\sum_{j=1}^{p} w_j x_{ij} + d_i \ge c \qquad \text{for all } i = 1,2,\dots,n_2 \text{ in group 2}$$

$$d_i \ge 0$$

$$\sum_{j=1}^{p} w_j + c = \text{constant}$$

w_j; c unrestricted in sign

Where w_j is the unknown weights for attributes j, j=1,2,…,p, x_{ij} is the values for observation i of attribute j, c is the cut off value, n_1 and n_2 are the number of observations in group 1, and group2 respectively, d_i is the distance of entity *i*'s score from the cut off value c, and represents the extent by which entity *i* is misclassified represents, and Z is the value of the objective function. [3] [2] An observation will be classified into group1 if its classification score is greater than, or equal to c, or into group2 otherwise. However, when constraint ($\sum_{j=1}^{p} w_j x_{ij} + d_i \ge c$) is less than, or equal to equality, this creates some potential problems such as improper solutions. The objective function Z has value of zero if the two groups by the hyperplane can be totally separated. The MSD formulation is robust with respect to outlier, whereas MMD formulation is very sensitive to outlier [11]

3. The Recovery and Non-Recovery Oil Application

Oil is a vital source of energy for the world, and will likely remain so for many decades to come, even under the most optimistic assumptions about the growth in alternative energy sources. Most countries are significantly affected by developments in the oil market, either as producers, consumers, or both. In 2008 oil provided about 34% of the world's energy needs, and in the future, oil is expected to continue to provide a leading component of the world's energy combine. [https://www.nrcan.gc.ca.com]. The following table (1) shows the world's top oil reserves ranked from 1 to 7:

TABLE (1): The World's Top Oil Reserves

Rank	Country	Reserves (Billion Barrels)	OPEC Member
1	Saudi Arabia*	262.4	Yes
2	Canada	174.7	No
3	Iran	137.6	Yes
4	Iraq	115.0	Yes
5	Kuwait*	104.0	Yes
6	Venezuela	99.4	Yes
7	U.A.E.	97.8	Yes

* Included half of the Saudi-Kuwaiti "neutral zone" which has 5 billion barrels of proved reserves. Oil and Gas Journal (2009).

The data used in this application was taken from Schlumberger Company. It is the world's leading supplier of technology, integrated project management and information solutions to customers working in the oil, and gas industry worldwide.

It provides the industry's widest range of products, and services from exploration through production. [http://www.slb.com/]

Also, Schlumberger provides two sampling services to obtain sidewall cores, Mechanical Sidewall Coring MSCT, and Perforation Sampling Tool. The first sampling services were used in this study. A hydraulically operated tool, MSCT is a lowered down hole to drill and retrieve core samples (sample of rock) as 0.91 inch, and the diameter is 2.0 inch. It drills into the formation with a diamond-tipped rotary core bit. At the end of its stroke, the bit breaks the core sample off, retracts back into the tool where the core is ejected from the bit into a retrieval tube by a core pusher rod once the core is in the tube, the tool drops a marker on the core sample to separate the samples. This method of recovery reduces damage to the core samples, allowing more accurate assessments of the core data. The core samples are used to verify the collected data that collected from different down whole tool such as Porosity, lithology by gamma ray, grain density, and the presence of hydrocarbons.

In this application three statistical, LDA,QDA, LoDA methods, and also two operations research methods, MMD and MSD that are used to discriminate between two distinct groups, recover and non-recover using four measurements, the density, gamma ray, Porosity/Permeability, and formation mechanical properties by Delta T. These methods may improve the company service quality, and change the client's perception of side wall coring services, and classify the rock sample probability to recover or not. Three different sample sizes n=20, n=25, and n=30 from each group. The error rate, and leave one out cross validation criteria have been applied to select the best model.

The first selected variable Delta T, is the amount of time for a wave to travel a certain distance, proportional to the reciprocal of velocity, typically measured in microseconds per foot by an acoustic log. The range of DT is from 43 (dolostone) to 160 (unconsolidated shales) microseconds per foot. The second variable is the Density that mass per unit of volume. It is typically reported in g/cm3 (for example, rocks) or pounds per barrel (drilling mud). The third selected variable is the Gamma ray which is a nuclear measurement that indicates the radioactive content of the formations naturally. It is the standard value which is used for the correlation in cased and open holes. The result is the fullest possible understanding of lithologies (sand, or shale). The fourth variable is the Porosity, the percentage of pore volume or void space, or that volume within rock that can contain fluids.
[http://www.slb.com/]

The SPSS program is used to test the normality, homogeneity, and difference of group means assumptions. Results showed that the two group means are significantly different from each other for the three sample sizes (n=20, 25, 30) from each group at significant level $\alpha = 0.05$. The tests of normality, homogeneity show that all measurements for all sample sizes satisfy the normality assumption, whereas the equality of variance- covariance matrices assumption is violated

The results of the statistical methods will be done by using SPSS program for the LAD, Minitab program for the QDA, and R-studio software for LoDA for different sample sizes. For LDA, and for sample size 20 from each group, there are 18(90%) of observations from group1 are correctly classified whereas 19(95%) are correctly classified to group 2 although the variance-covariance assumption is violated. There are 17(85%) for each group, apparent hit rate of the leave-one-out cross validation LOOCV which represents the performance measures to be compared between the small sample size methods (proportion of observations classified correctly) will be 85% for all groups. The unstandardized canonical discriminant function will have the following form:

$$D = -19.176 + 0.009\ \mathrm{DT} + 7.184\ \mathrm{density} + 0.027\mathrm{GR} - 5.281\mathrm{Porosity} \qquad (8)$$

The standardized canonical discriminant SCD function which measures the relative importance of the selected variables, the larger absolute value of the coefficient corresponds to greater discriminating ability, and means that the groups differ a lot on that variable. It indicates that the independent variable "Density" is the most powerful discriminating variable, followed by "GR" and "DT". "Porosity" is less successful as predictors:

$$\mathrm{SCD} = 0.566\ \mathrm{deltaT} + 0.929\ \mathrm{density} + 0.601\mathrm{GR} - 0.500\ \mathrm{Porosity} \qquad (9)$$

The sign indicates the direction of the relationship. Density was the strongest predictor while GR was next in importance as a predictor. These two variables with large coefficients stand out as those that strongly predict allocation to the group1 or group2. The DT and porosity were less successful as predictors. The group membership can determine by calculating a cut point halfway between the two centroid, c=0. In many studies the cut point is an arbitrary chosen. The classification score for the first group1 will be as follows:

$$\text{Score}_1 = -195.647 + 0.234\text{ deltaT} + 166.599\text{ density} + 0.293\text{GR} - 171.766\text{ Porosity} \quad (10)$$

and for the second group2 it will be as follows:

$$\text{Score}_2 = -244.594 + 0.258\text{deltaT} + 184.936\text{ density} + 0.361\text{GR} - 185.247\text{ Porosity} \quad (11)$$

The LDA results at sample size n=25 from each group, and the LOOCV validation, showed that 23(92%) of observations from group1 are correctly classified, whereas 19(76%) are correctly classified to group 2. There're 17(88%)

for each group1, and 19(76%) for group 2 apparent hit rate, the LOOCV. The unstandardized canonical discriminant function will be as follows

$$D = -12.923 + 0.012\text{DT} + 3.913\text{ density} + 0.015\text{GR} + 2.105\text{ Porosity} \quad (12)$$

The independent variable "density" was the most powerful for discriminating and differentiating the between group differences. The standardized canonical discriminant function is as follows:

$$\text{SCD} = 0.589\text{deltaT} + 0.744\text{ density} + 0.328 + 0.171\text{ Porosity} \quad (13)$$

The group membership can be equal to 0, and the classification score for the first group1 will be as follows:

$$\text{Score}_1 = -128.708 + 0.309\text{deltaT} + 100.239\text{ density} - 0.080\text{GR} - 68.358\text{ Porosity} \quad (14)$$

and for the group2 it will be as follows:

$$\text{Score}_2 = -128.708 + 0.309\text{ deltaT} + 100.239\text{ density} - 0.80\text{GR} - 68.358\text{ Porosity} \quad (15)$$

The LDA results at sample size n=30 from each group showed that 24(80%) of observations from group1 are correctly classified, whereas 26(86.7%) are correctly classified to group 2 although the variance-covariance assumption is violated. There are 24(80%) for each group1, and 23(76.7%) for group 2 apparent hit rate, the LOOCV. The unstandardized canonical discriminant function will be as follows

$$D = -12.447 + 0.008\ \text{deltaT} + 4.100\ \text{density} + 0.023 - 1.504\ \text{Porosity} \qquad (16)$$

The standardized canonical discriminant function will be as follows:

$$SCD = 0.529\ DT + 0.710\ \text{density} + 0.531GR - 0.149\ \text{Porosity} \qquad (17)$$

The membership is determined by calculating a cut point equal to 0, the SPSS program is used to compute the classification score for the first group1 will be as follows:

$$\text{Score}_1 = -101.774 + 0.120\ \text{deltaT} + 83.172\ \text{density} + 0.055GR - 49.144\ \text{Porosity} \qquad (18)$$

and for the group2 it can be written as follows:

$$\text{Score}_2 = -122.503 + 0.133\ \text{deltaT} + 89.999\ \text{density} + 0.93GR - 51.650\ \text{Porosity} \qquad (19)$$

The program Minitab is used to get the QDA and LOOCV results for all sample sizes. For the sample size n=20 from each group, results show that 19(95%) of observations from group1 are correctly classified, whereas 19(95%) are correctly classified to group 2 although the variance-covariance assumption is violated. There are 18(90%) for group1, and 17(85%) for group 2 apparent hit rate, the LOOCV. The quadratic classification results for sample size n=25 from each group results show that 24(96%) of observations from group1 are correctly classified, whereas 23(92%) are correctly classified to group 2 although the variance-covariance assumption is violated. There are 23(92%) for group1, and 20(80%) for group 2 apparent hit rate, the LOOCV. For the sample size n=30 from each group results showed that 26(86.7%) of observations from group1 are correctly classified, whereas 23(76.7%) are correctly classified to group 2 although the variance-covariance assumption is violated. There are 25(83.3%) for group1, and 22(73.3%) for group 2 apparent hit rate, the LOOCV.

The program SPSS is used to get the results of logistic model, LoDA. It will be shown that at sample size n=20 from each group, 17(85%) of observations from group1 are correctly classified, whereas 18(90%) are correctly classified to group 2. The logistic discriminant function is as follows:

$$\hat{Y}_i = -54.891 + 21.158\ \text{density} \qquad (20)$$

The LoDA results for sample size n=25 from each group showed that, 23(92%) of observations from group1 are correctly classified, whereas 21(84%) are correctly classified to group 2. The logistic discriminant function is as follows:

$$\hat{Y}_i = -42.019 + 12.329\ \text{density} \qquad (21)$$

For the sample size n=30 the results showed that 24(80%) of observations from group1 are correctly classified, whereas 26(86.7%) are correctly classified to group 2. The logistic discriminant function is as follows:

$$\widehat{Y}_i = -19.783 + 6.649\ \text{density} + 0.037\text{GR} \tag{22}$$

The TORA software is used to construct the classification functions of MMD and MSD methods with the three sample sizes (20, 25, and 30) from each group. The Microsoft Excel is used to determine the apparent hit rates (proportion of observations classified correctly). The LOO-cross validation is not practical in linear programming. According to results, for sample size equals n=20, there are only one observation of group1 is misclassified into group2, and seven observations of group2 are misclassified into group1. The hit ratio of MMD method is equal to 80%, and the misclassification is equal to 20%. For sample size equals n=25, there are only three observations from group1 misclassified to group2, and only one observation from group2 is classified to the group1. The hit ratio of MMD method is equal to 92%, and the misclassification is equal to 8%. For sample size equals n=30, there are 10 observations from group1 are misclassified to group2, and five observations from group2 is classified to the group1. The hit ratio of MMD method is equal to 75%, and the misclassification is equal to 25%.

The MSD model of sample size n=20, reveals that there are only three observations of group1 are misclassified into group2, and four observations of group2 are misclassified into group1. Exceed to the cut point give the hit ratio of MSD method equals to 82.5% and the misclassification equals to 17.5%. At sample size 25, there are only three observations from group1 misclassified to group2, and only one observation from group2 is misclassified to the group1. The hit ratio of MSD method is equal to 92%, and the misclassification is equal to 8%. At sample size 30, there are ten observations from group1 are misclassified to group2, and five observations from group2 is misclassified to the group1. The hit ratio of MSD method is equal to 75%, and the misclassification is equal to 25%.

4. Analysis and Recommendations

The Oil and Gas field is a new area for using discriminant analysis either statistical or operations research methods. In statistical methods, the traditional methods such as Anderson's method, Q- method, Fisher method… assumed many assumptions especially the normality, and the equality of variances-coveriances matrices, whereas operations research methods do not have any required assumption; they are flexible methods. Many statistical methods such as quadratic and logistic methods do not have also great assumptions and can be classified as flexible methods. The comparison results show approximately the same rate of correct classification especially the quadratic and the MMD and MSD operations research methods although the equality variance – covariance assumption is violated for these datasets. This study suggests trying many alternative statistical, and operations research methods, along with alternative evaluating criteria in a new application area and also with big data.

.

5. References:

[1] Antonogeorgos, G., Panagiotakos, D. B., Priftis, K. N., & Tzonou, A. (2009). Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10- To 12-Years-old Children: Divergence and Similarity of the Two Statistical Methods. International Journal of Pediatrics, 1-6.

[2] BAL, H., Örkcü, H. H., & Çelebioğlu, S. (2006) a. An Alternative Model to Fisher and Linear Programming Approaches in Two-Group Classification Problem: Minimizing Deviations from the Group Median. *G.U.* Journal of Science, vol.19, 49-55.

[3] Banks, W. J., B., A., C., A., & M., A. B. (1991). New Solution Algorithems for the Classification Problem. MSc. Thesis. McMaster University.

[4] Ben Youssef, S., & Rabai, A. (2007). Comparison Between Statistical Approach and Llinear Programming for Solving Classification Problem. International Mathematical Forum, vol.2, 3125-3141.

[5] Dillon, W. R., & Goldstein, M. (1984). Multivariate Analysis Methods and Applications. John Wiley & Sons.

[6] Fan, X., & Wang, L. (1999). Comparing Linear Discriminant Function with Logistic Regression for the Two-Group Classification Problem. The Journal of Experimental Education, vol.3, 265-286.

[7] Freed, N., & Glover, F. (1979). Simple But Powerful Goal Programming Models for Discriminant Problems. European Journal of Operational Research., vol.7, 44-60.

[8] Joachimsthaler, E. A., & Stam, A. (1990). Mathematical Programming Approaches for the Classification Problem in Two- Group Discriminant Analysis. Multivariate Behavioral Research, vol. 25(4), 427-454.

[9] Johnson , R. A., & Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. Prentice Hall.

[10] Lachbruch, P. A. (1974). Discriminant Analysis. Hafner Press.

[11] Lam, K. F., Choo, E. U., & Moy, J. W. (1996). Minimizing Deviations from the Group Mean: A New Linear Programming Approach for the Two-Group Classification Problem. European Journal of Operational Research, vol.88, 358– 367.

[12] Nath, R., Jacson, W. M., & Jones, T. W. (1992). A Comparison of the Classical and the Linear Programming Approach to the Classification Problem in Disciminant Analysis. Gordon and Breach Science Publishers S.A, vol.41, 73-93.

[13] Rausch, J. R., & Kelley, K. (2009). A Comparison of Linear and Mixture Models for Discriminant Analysis under Nonnormality. Behavior Research Methods, *vol.*1, 85-98.

[14] Rencher, A. C. (2002). Methods of Multivariate Analysis. John Wiley & Sons, Inc.

[15] Smaoui, S., Chabchoub, H., & Aouni, B. (2009). Mathematical Programming Approaches to Classification Problems. Hindawi Publishing Corporation, 1-34.

[16] Timm, N. H.(2002). Applied Multivariat Analysis. Springer-Verlag New York ,Inc.