On the Regression Discriminant Analysis (RDA), and its Identical Relationship to the Fisher's Discriminant Analysis

Jude C Obi*, Peter Thwaites, John Kent****

* Department of Statistics, COOU[1] Nigeria
**Department of Statistics, The University of Leeds, UK

Abstract- The linear regression has been studied particularly with respect to its identical relationship to the Fisher's Discriminant Analysis (FDA). In this paper, we prove that the coefficient vector of the least squares regression is identical to the vector of coefficients of the Fisher's Discriminant Analysis. We therefore refer to the classification procedure using the least squares regression as Regression Discriminant Analysis (RDA). We further carry out empirical investigations using both real world and simulated datasets, to show that a classification procedure based on both RDA and FDA are identical.

Index Terms- Machine learning, Regression based binary classification, Linear discriminant analysis, Statistical learning

I. Introduction

Regression and classification are two prediction tools, since both tools involve the use of input to generate output[1]. The output, otherwise called predictions are dependent on the input which is either a single explanatory variable or a set of explanatory variables. Both regression and classification differ in their output. For instance, the output of regression consists of continuous response variables, whereas discrete response variables constitute the output of classification.

As tools for prediction, a model or algorithm is involved and in order to carry out prediction, the model is necessarily trained given a set of data [2]. Training here involves making informed choice of a prediction function that best describes a relation between the input and output. To access this function, we often resort to the use of training and test sets. A training set is used to discover potentially predictive relationships, and a test set is useful to access the strength of the relationship [3].

Training of models and the assessment of their performances are often severally repeated before a function that optimally performs is finally obtained. An optimum performing function is the one that upon comparison with all similar functions, that describe a relation between the input and output, gives the smallest prediction error. Prediction error is the error that results from incorrect prediction of the output. As a statistic, it measures how well or badly a model has performed. We usually have preference for a function with zero or smallest prediction error in comparison with other similar functions.

Regression and classification as tools for prediction, have some characteristics that are commonly shared. For instance:

a) There exists a matrix of input data, and vector of output required for training and testing of both regression and classification models.
b) Since the dimension of the input data, in most cases is at least $n \times 2$, there can be concern for numerical stability of data in both cases.
c) In instances where the input data is high dimensional, to obtain a regression or classification function that optimally performs may be a challenging task.
d) Assessment of prediction tools is via prediction error, and the method of calculating such error depends on the prediction tool in question.
e) Finally, we note that a very important characteristic is the fact that the equations describing both regression and classification functions can be similarly expressed. For instance, (1.1) and (1.2) are respectively regression and classification functions.

$$\begin{aligned} y &= f(\mathbf{x}) \\ &= b_1 x_1 + \cdots + b_p x_p + b_0 \end{aligned} \tag{1.1}$$

[1]ChukwuemekaOdumegwuOjukwu University (Former Anambra State University)

$$\begin{aligned} z &= g(\mathbf{x}) \\ &= \mathbf{w}^T\left(\mathbf{x}+\boldsymbol{\mu}\right) \\ &= w_1x_1+\cdots+w_px_p+\mathbf{w}^T\boldsymbol{\mu} \end{aligned} \tag{1.2}$$

Both functions have coefficient vectors respectively denoted by $\mathbf{b}$ and $\mathbf{w}$, and p-dimensional explanatory variables. The constant terms are respectively $\boldsymbol{b}_0$ and $\mathbf{w}^T\boldsymbol{\mu}$. We can infer that (1.1) is a regression function where $\mathbf{b}$ is a regression coefficient vector and similarly, (1.2)is a classification function. Let $\mathbf{w}=\mathbf{d}^T\hat{\boldsymbol{\Sigma}}^{-1}$, and $\boldsymbol{\mu}=-\frac{1}{2}\left(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2\right)$, where $\mathbf{d}=\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2$ and both $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are class mean vectors for a two class classification problem. Also let $\hat{\boldsymbol{\Sigma}}$ be a pooled covariance matrix for both classes, then(1.2) can refer to a Fisher's discriminant function. Hence, $\mathbf{w}$ can be called a coefficient vector of FDA.

Assuming we further examine (1.1) and (1.2) it may be possible to find a situation where the two coefficient vectors are at least proportional, and the explanatory variables exactly the same. In such a situation, one wonders if it could be possible that any informed alteration of say f, can be useful for predicting z, or vice versa.

In the light of these, and particularly considering the commonly shared characteristics, we are of the view that it is possible to use regression as a tool for classification. Thus, we propose a classification function based on the least squares regression, and claim that it is identical to FDA. Hence, we refer to it as Regression Discriminant Analysis (RDA).

II. Literature Review

The logistic regression [4] is a foremost regression based classification procedure and it has been shown that when normality assumption isviolated, it is superior to FDA [5]. If normality is assumed, the logistic regression is an alternative to FDA [6]. Conversely, the logistic regression differs from our main idea because it fits a nonlinear model to a linear combination of explanatory variables. In the case of RDA, our aim is to fit a linear model for classification based on the multiple regression.

One study that involves the use of multiple regression in discriminant analysis is due to [7], but the study considered a multi-class case. We are mainly interested in binary classification problem. Hence, the study carried out by [8] is of interest to us because it concerns a binary classification problem. It considers the relationship of the Minimum Squared Error procedure to the Fisher's Linear Discriminant. The authors showed that with the proper choice of the vector $\mathbf{b}$, say, the MSE discriminant function $\mathbf{b}^T\mathbf{x}$ is directly related to Fisher's Linear Discriminant.

To prove their point, they assumed that we have a set of n p-dimensional samples $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n$, n_1 of which are in subset D_1 labelled ω_1, and n_2 of which are in subset D_2 labelled ω_2. Further, they assumed that a sample $\mathbf{y}_i$ is formed from $\mathbf{x}_i$ by adding a threshold component $x_0=1$ to make an augmented pattern vector. If the sample is labelled by ω_2, then the entire pattern vector is multiplied by -1. Thus without loss of generality, the first n_1 samples are labelled ω_1 and the second n_2 samples are labelled ω_2. The matrix $\boldsymbol{Y}$ can be partitioned as:

$$\boldsymbol{Y}=\begin{pmatrix} 1_1 & \boldsymbol{X}_1 \\ -1_2 & \boldsymbol{X}_2 \end{pmatrix}, \tag{2.1}$$

where 1_i is a column vector of n_i ones, and $\boldsymbol{X}_i$ is an $n_i\times d$ matrix whose rows are samples labelled ω_i. They partitioned $\mathbf{a}$ and $\mathbf{b}$ correspondingly with

$$\mathbf{a}=\begin{pmatrix} w_0 \\ \mathbf{w} \end{pmatrix}, \tag{2.2}$$

and with

$$\mathbf{b} = \begin{pmatrix} \frac{n}{n_1} & \mathbf{1}_1 \\ \frac{n}{n_2} & \mathbf{1}_2 \end{pmatrix}. \tag{2.3}$$

They noted that this special choice of $\mathbf{b}$ links the MSE solution to Fisher's Linear Discriminant.Define sample means $\mathbf{m}_i$ and pooled sample scatter matrix S_W :

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{x} \quad i = 1,2 \tag{2.4}$$

$$S_W = \sum_{i=1}^{2}\sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T, \tag{2.5}$$

and plug into MSE formulation[9]to get

$$\mathbf{w} = anS_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{2.6}$$

Finally, the authors noted that except for an unimportant scale factor, (2.6) is identical to the solution for Fisher's Linear Discriminant. Decision rule: decide ω_1 if $\mathbf{w}^T(\mathbf{x}-\mathbf{m}) > 0$, otherwise decide ω_2 , where $\mathbf{m}$ is the mean of all the samples.

III. Different Procedures

The main ideas expressed by [9]agrees with ours but our procedure for establishing the identical relationship between RDA and FDA differs. Prior to giving the relevant proof, we shall first define some datasets and give some notations.

a. **Data and some notations**

Let $X_1(n_1 \times p)$ and $X_2(n_2 \times p)$ be datasets for two populations Π_1 and Π_2 , and let $n = n_1 + n_2$. Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}(n \times p)$ denote the whole dataset, and $H = I_n - (1/n)1_n 1_n^T$ denote the $n \times n$ centering matrix. In a similar way, let H_1 and H_2 denote the $n_1 \times n_1$, and $n_2 \times n_2$ centering matrices respectively.

Let $\overline{\mathbf{x}}_1$, $\overline{\mathbf{x}}_2$ and $\overline{\mathbf{x}}$ denote the sample means of X_1 , X_2 and X respectively. Note that

$$\overline{\mathbf{x}} = (n_1\overline{\mathbf{x}}_1 + n_2\overline{\mathbf{x}}_2)/n$$

is a weighted average of the two class means. We also need the unweighted average

$$\mathbf{x}_{\text{av}} = (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)/2,$$

and the difference,

$$\boldsymbol{\delta} = \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2. \tag{3.1}$$

b. **Fisher's allocation rule**

Several matrices are of interest in discriminant analysis:

$$T = X^T H X,$$
$$B = (n_1 n_2 / n)\boldsymbol{\delta}\boldsymbol{\delta}^T,$$
$$W = X_1^T H_1 X_1 + X_2^T H_2 X_2.$$

A classic result[10]states that

T = W + B.

The Fisher's allocation rule is based on Fisher's linear discriminant function given by:

$$f(\mathbf{x}) = \boldsymbol{\delta}^T W^{-1}\left(\mathbf{x} - \mathbf{x}_{\text{av}}\right).$$

The allocation rule in respect of a new input $\mathbf{x}$ says: allocate $\mathbf{x}$ to Π_1 if $f(\mathbf{x}) \geq 0$, and to Π_2 otherwise.

It is important to note that sometimes $f(\mathbf{x})$ is constructed using $S_{\text{pooled}} = W/(n-2)$ instead of W, but the allocation rule is the same. Since W is symmetrical, write

$$\boldsymbol{\gamma} = W^{-1}\ \boldsymbol{\delta}; \tag{3.2}$$

then Fisher's discriminant function simplifies to

$$f(\mathbf{x}) = \boldsymbol{\gamma}^T\left(\mathbf{x} - \mathbf{x}_{\text{av}}\right).$$

c. **Multiple Regression**

Let $\mathbf{y} = \begin{pmatrix} +\mathbf{1}_{n_1\times 1} \\ -\mathbf{1}_{n_2\times 1} \end{pmatrix}$ denote a response vector of length n, and consider a regression of $\mathbf{y}$ on X. Then, the ordinary least squares regression function can be written as

$$g(\mathbf{x}) = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\mathbf{x},$$

where $\hat{\alpha} = \bar{y} - \hat{\boldsymbol{\beta}}^T\bar{\mathbf{x}}$, and $\hat{\boldsymbol{\beta}} = \left(X^T HX\right)^{-1} X^T H\mathbf{y} = T^{-1}X^T\left(H\mathbf{y}\right)$.Note that $\hat{\boldsymbol{\beta}}$ is estimated using the centered data matrix HX, we then claim that

$$\hat{\boldsymbol{\beta}} \propto \boldsymbol{\gamma}, \tag{3.3}$$

where $\boldsymbol{\gamma}$ is as defined in(3.2).

d. **Proof**

First note that the centered vector $H\mathbf{y}$ has entries $+1-\bar{y}$ in the first n_1 places and $-1-\bar{y}$ in the final n_2 places. Since $\bar{y} = \left(n_1 - n_2\right)/n$, $H\mathbf{y}$ simplifies to $2n_1n_2/n$ times a vector with $+1/n_1$ in the first n_1 places and $-1/n_2$ in the final n_2 places. Hence,

$$\begin{aligned} X^T(H\mathbf{y}) &= \left(1/n_1\right)X_1^T\mathbf{1}_{n_1} - \left(1/n_2\right)X_2^T\mathbf{1}_{n_2} \\ &= \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \boldsymbol{\delta}, \end{aligned}$$

where $\boldsymbol{\delta}$ is as defined in(3.1).

Showing that $\hat{\boldsymbol{\beta}} \propto \boldsymbol{\gamma}$ is equivalent to showing that $T^{-1}\boldsymbol{\delta} \propto \boldsymbol{\gamma}$, which is true if and only if

$$\begin{aligned} \boldsymbol{\delta} &\propto T\boldsymbol{\gamma} \\ &\propto TW^{-1}\boldsymbol{\delta} \\ &\propto \left(W + B\right)W^{-1}\boldsymbol{\delta} \\ &\propto \left(I + \left(n_1n_2/n\right)\boldsymbol{\delta}\boldsymbol{\delta}^TW^{-1}\right)\boldsymbol{\delta} \\ &\propto \boldsymbol{\delta} + \left(n_1n_2/n\right)\boldsymbol{\delta}\left(\boldsymbol{\delta}^TW^{-1}\boldsymbol{\delta}\right) \\ &= \left\{1 + \left(n_1n_2/n\right)\left(\boldsymbol{\delta}^TW^{-1}\boldsymbol{\delta}\right)\right\}\boldsymbol{\delta} \\ &= u\boldsymbol{\delta}, \end{aligned}$$

where $u = \left\{1 + \left(n_1n_2/n\right)\left(\boldsymbol{\delta}^TW^{-1}\boldsymbol{\delta}\right)\right\}$ is a constant. Hence, the result is proved.

e. **Regression rule**

Set,

$$\begin{aligned} g(\mathbf{x}) &= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x} \\ &= \bar{y} - \hat{\boldsymbol{\beta}}^T \bar{\mathbf{x}} + \hat{\boldsymbol{\beta}}^T \mathbf{x} \\ &= \bar{y} + \hat{\boldsymbol{\beta}}^T \left(\mathbf{x} - \bar{\mathbf{x}}\right), \end{aligned} \tag{3.4}$$

and allocate to Π_1 if $g(\mathbf{x}) \geq 0$, otherwise to Π_2. If on the other hand, we set $\mathbf{x} = \mathbf{x}_{\text{av}}$, then,

$$g\left(\mathbf{x}_{\text{av}}\right) = \bar{y} + \hat{\boldsymbol{\beta}}^T \left(\mathbf{x}_{\text{av}} - \bar{\mathbf{x}}\right) \neq 0,$$

unless $n_1 = n_2$. Hence, the naive regression is different from Fisher's rule. We have used the term naive regression to explain that the function g, specified in(3.4) is identical to FDA if and only if $n_1 = n_2$.

f. **Alternative rule**

Alternatively, we can shift the regression predictor by a constant value to

$$\begin{aligned} g^*\left(\mathbf{x}\right) &= g\left(\mathbf{x}\right) - \left(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_{\text{av}}\right) \\ &= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x} - \hat{\alpha} - \hat{\boldsymbol{\beta}}^T \mathbf{x}_{\text{av}} \\ &= \hat{\boldsymbol{\beta}}^T \left(\mathbf{x} - \mathbf{x}_{\text{av}}\right), \end{aligned} \tag{3.5}$$

and define another rule: allocate $\mathbf{x}$ to Π_1 if $g^*\left(\mathbf{x}\right) \geq 0$ and to Π_2 otherwise.The allocation rule given by f and g^* are identical, hence we call g^* a regression based discriminant function instead of g. In summary, the Fisher's allocation rule based on f is identical to the regression-based allocation rule based on g^*.

IV. EMPIRICAL INVESTIGATION

This investigation will involve the use of some real world and simulated datasets to investigate the identical relationship of FDA and RDA as proved in section III. The majority of the datasets used were sourced from the UCI Machine Learning Repository [11], and KEEL dataset repository [12]. We preprocessed all the datasets to ensure that each class label is identified with the name “class’’, and consists of a vector of $+1$ and -1 discrete variables. This way, we avoid the problem of rewriting the program we used each time a different dataset is involved. The datasets include:

Australia Dataset

The Australia dataset concerns credit card applications, and all attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. It has dimensions $ 690 \times 14 $, with two classes representing approved and not approved. The data source is [12], and website; http://sci2s.ugr.es/keel/dataset.php? cod=53.

Handheight

The Handheight dataset is two dimensional, and consists of heights and stretched hand span of 167 male and female college students. Each student decided which of their hands to measure. Class $+1$ has 89 samples whereas class -1 consists of 78 samples. The source of the data is [13].

Heart

This is a real world binary classification heart disease dataset, and the task is to detect the absence (-1) or presence (1) of heart disease. It contains 270 samples and 13 features, with 120 samples in class +1 and 150 samples in class -1. The data was sourced from the UCI Machine Learning Repository.

Ionosphere

Ionosphere is a radar dataset collected by a system in Goose Bay, Labrador. The system consists of a phased array of 16 high-frequency antennas with a total transmitted power of the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. “Good’’ radar returns are those showing evidence of some type of structure in the ionosphere. “Bad’’ returns are those that do not; their signals pass through the ionosphere.

Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were $ 17 $ pulse numbers for the Goose Bay system. Instances in this database are described by 2 attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal. The dataset is included in the mlbench package [14].

Mammographic
This dataset was used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field, which is the target attribute). The dataset was collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. It has dimensions 830×5, and the source is the KEEL dataset repository.

Twonorm
This dataset is 20 dimensional, and consists of 2 classes. Each class is drawn from a multivariate normal distribution. Class +1 has mean ($a, a, ..., a$) while Class -1 has mean $(-a, -a, ..., -a)$; $a = 2 / sqrt(20)$. The dataset has dimensions 7400×20, and is contained in the KEEL dataset repository.

Wisconsin Diagnostic Breast Cancer (WDBC) Dataset
WDBC is a real-world dataset, and contains 30 features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The number of instances is 569 and the task is to determine if a tumor found is benign or malignant ($-1 = \text{malignant}$, and $1 = \text{benign}$). It was sourced from the UCI Machine Learning Repository.

V. DISCUSSION

Contained in Table 1 are the error rates of FDA and RDA on seven different datasets. Based on the differences in the two error rates, it seems that the two classifiers are identical. In two of the datasets, namely Handheight and Twonorm, we observed small positive differences between the two classifiers. Given the datasets, the error rates of RDA appear to be marginally smaller than the error rates of FDA. Regarding the dataset WDBC, we observed a small negative difference of -0.0175. It shows that on this dataset, the error rate of FDA is marginally smaller than the error rates of RDA. The two classifiers have the same error rates on four different datasets, namely Australia, Heart, Ionosphere and Mammographic datasets.

S/nos	Dataset Name	FDA Error Rate	RDA Error Rate	Diff
1	Australia	0.1353	0.1353	0.0000
2	Handheight	0.1800	0.1200	0.0600
3	Heart	0.1489	0.1489	0.0000
4	Ionosphere	0.1810	0.1810	0.0000
5	Mammographic	0.2088	0.2088	0.0000
6	Twonorm	0.0216	0.0212	0.0004
7	WDBC	0.0234	0.0409	–0.0175

Table 1: *Error rates of FDA and RDA on seven different datasets*

An intuitive appeal from Table 1 is that observed differences in error rates are non-significant. Nevertheless, at a p-value of 0.7893, a non-parametric Wilcoxon signed rank test [15] failed to reject the hypothesis that observed differences in the two error rates are non-significant at 5% level of significance. This development is further confirmation that both classifiers are identical.

VI. SUMMARY/CONCLUSIONS

The main argument presented by [9] is that $\mathbf{w} = anS_w^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right)$ is identical to the Fisher's Linear Discriminant vector of coefficients. We equally observed that the same is true concerning $\hat{\boldsymbol{\beta}}$, because in section III we proved that $\hat{\boldsymbol{\beta}} \propto \boldsymbol{\gamma}$.

The outcome of empirical investigations contained in Table 1, suggests lack of evidence for significant differences in the error rates of the two classifiers. A further test of hypothesis based on the Wilcoxon signed rank test failed to reject the hypothesis that differences in the error rates of the classifiers are non-significant. Thus far, it is our conclusion that the two classifiers are identical. Hence, a choice of either FDA or RDA for a given binary classification problem is appropriate, and therefore recommended.

REFERENCES

[1] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Springer Series in Statistics Springer, Berlin, 2001.

[2] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2011.
[3] Wikipedia, *Test set — Wikipedia, The Free Encyclopedia*. 2016.
[4] D. R. Cox, "The Regression Analysis of Binary Sequences," *J. R. Stat. Soc. Ser. B Methodol.*, pp. 215–242, 1958.
[5] M. Pohar, M. Blas, and S. Turk, "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study," *Metodoloski Zv.*, vol. 1, no. 1, p. 143, 2004.
[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 6. Springer, 2013.
[7] J. Ye, "Least Squares Linear Discriminant Analysis," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 1087–1093.
[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.
[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.
[10] P. J. Hewson, "Multivariate Statistics with R," 2009.
[11] M. Lichman, *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2013.
[12] J. Alcalá *et al.*, "Keel Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework," *J. Mult.-Valued Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2010.
[13] J. Utts and R. F. Heckard, *Mind on Statistics*. Cengage Learning, 2011.
[14] F. Leisch and E. Dimitriadou, "The mlbench Package-A Collection for Artificial and Real-World Machine Learning Benchmarking Problems," *R Package Version*, pp. 1–0, 2004.
[15] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biom. Bull.*, vol. 1, no. 6, pp. 80–83, 1945.