Metodološki zvezki, Vol. 1, No. 1, 2004, 143-161

Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study

Maja Pohar[1], Mateja Blas[2], and Sandra Turk[3]

Abstract

Two of the most widely used statistical methods for analyzing categorical outcome variables are linear discriminant analysis and logistic regression. While both are appropriate for the development of linear classification models, linear discriminant analysis makes more assumptions about the underlying data. Hence, it is assumed that logistic regression is the more flexible and more robust method in case of violations of these assumptions. In this paper we consider the problem of choosing between the two methods, and set some guidelines for proper choice. The comparison between the methods is based on several measures of predictive accuracy. The performance of the methods is studied by simulations. We start with an example where all the assumptions of the linear discriminant analysis are satisfied and observe the impact of changes regarding the sample size, covariance matrix, Mahalanobis distance and direction of distance between group means. Next, we compare the robustness of the methods towards categorisation and non-normality of explanatory variables in a closely controlled way. We show that the results of LDA and LR are close whenever the normality assumptions are not too badly violated, and set some guidelines for recognizing these situations. We discuss the inappropriateness of LDA in all other cases.

1 Introduction

Linear discriminant analysis (LDA) and logistic regression (LR) are widely used multivariate statistical methods for analysis of data with categorical outcome

[1] Department of Medical Informatics, University of Ljubljana; maja.pohar@mf.uni-lj.si
[2] Postgraduate student of Statistics, University of Ljubljana; mateja.blas@guest.arnes.si
[3] Sandra Turk, Krka d.d., Novo mesto; sandra.turk@krka.biz

variables. Both of them are appropriate for the development of linear classification models, i.e. models associated with linear boundaries between the groups.

Nevertheless, the two methods differ in their basic idea. While LR makes no assumptions on the distribution of the explanatory data, LDA has been developed for normally distributed explanatory variables. It is therefore reasonable to expect LDA to give better results in the case when the normality assumptions are fulfilled, but in all other situations LR should be more appropriate. The theoretical properties of LR and LDA are thoroughly dealt with in the literature, however the choice of the method is often more related to the field of statistics than to the actual condition of fulfilled assumptions.

The goal of this paper is not to discourage the current practice but rather to set some guidelines as to when the choice of either one of the methods is still appropriate. While LR is much more general and has a number of theoretical properties, LDA must be the better choice if we know the population is normally distributed. However, in practice, the assumptions are nearly always violated, and we have therefore tried to check the performance of both methods with simulations. This kind of research demands a careful control, so we have decided to study just a few chosen situations, trying to find a logic in the behaviour and then to think about the expansion onto more general cases. We have confined ourselves to compare only the predictive power of the methods.

The article is organized as follows. Section 2 briefly reviews LR and LDA and explains their graphical representation. Section 3 details the criteria chosen to compare both methods. Section 4 describes the process of the simulations. The results obtained are presented and discussed in Section 5, starting with the case where all the assumptions of LDA are fulfilled and continuing with cases where normality is violated in sense of categorization and skewness. It is shown how violation of the assumptions of LDA affects both methods and how robust the methods are. The paper concludes with some guidelines for the choice between the models and a discussion.

2 Logistic regression and linear discriminant analysis

The goal of LR is to find the best fitting and most parsimonious model to describe the relationship between the outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. The method is relatively robust, flexible and easily used, and it lends itself to a meaningful interpretation. In LR, unlike in the case of LDA, no assumptions are made regarding the distribution of the explanatory variables.

Contrary to the popular beliefs, both methods can be applied to more than two categories (Hosmer and Lemeshow, 1989, p. 216). To simplify, we only focus on

the case of a dichotomous outcome variable (Y). The LR model can be expressed as

$$P(Y_i = 1 | X_i) = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} \tag{2.1}$$

where the Yi are independent Bernoulli random variables. The coefficients of this model are estimated using the maximum likelihood method. LR is discussed further by Hosmer and Lemeshow (1989).

Linear discriminant analysis can be used to determine which variable discriminates between two or more classes, and to derive a classification model for predicting the group membership of new observations (Worth and Cronin, 2003). For each of the groups, LDA assumes the explanatory variables to be normally distributed with equal covariance matrices. The simplest LDA has two groups. To discriminate between them, a linear discriminant function that passes through the centroids of the two groups can be used. LDA is discussed further by Kachigan (1991). The standard LDA model assumes that the conditional distribution of X|y is multivariate normal with mean vector μy and common covariance matrix Σ. With some algebra we can show that we assign x to group 1 as

$$P(1 | x) = \frac{1}{1 + \left(e^{\alpha + \beta x}\right)^{-1}} \tag{2.2}$$

where α and β coefficients are

$$\begin{aligned} \beta &= (\mu_1 - \mu_0)^T \Sigma^{-1} \\ \alpha &= -\log \frac{\pi_1}{\pi_0} + \frac{1}{2}(\mu_1 + \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0) \end{aligned} \tag{2.3}$$

π1 and π0 are prior probabilities of belonging to group 1 and group 0. In practice the parameters π1, π0, μ1, μ0 and Σ will be unknown, so we replace them by their sample estimates, i. e.:

$$\begin{aligned} &\hat{\pi}_1 = \frac{n_1}{n}, \ \hat{\pi}_0 = \frac{n_0}{n}, \\ &\hat{\mu}_1 = \bar{x}_1 = \frac{1}{n_1} \sum_{y_i=1} x_i, \ \hat{\mu}_0 = \bar{x}_0 = \frac{1}{n_0} \sum_{y_i=0} x_i, \\ &\hat{\Sigma} = \left[\sum_{y_i=1} (x_i - \bar{x}_1)(x_i - \bar{x}_1)^T + \sum_{y_i=0} (x_i - \bar{x}_0)(x_i - \bar{x}_0)^T \right] / n \end{aligned} \tag{2.4}$$

(2.2) is equal in form to LR. Hence, the two methods do not differ in functional form, they only differ in the estimation of coefficients.

2.1 Graphical representation: An explanation

When the values of α and β are known, the expression for a set of points with equal probability of allocation can be derived as

$$0.5 = \frac{e^{\alpha+\beta^T x}}{1+e^{\alpha+\beta^T x}} \Rightarrow 0 = \alpha + \beta^T x \qquad (2.5)$$

In two-dimensional perspective this set of points is a line, while in three dimensions it is a plane.

Figure 1 shows the scatterplot for two explanatory variables. Each of the two groups is plotted with a different character. The linear borders presented are calculated on the basis of the estimates of each method. The ellipses indicate the distributions assumed by the LDA.

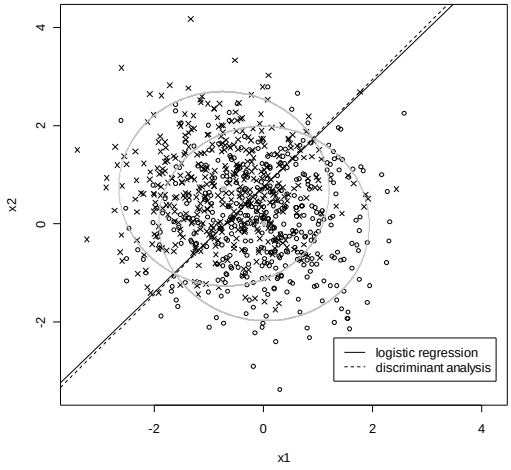

Figure 1: The linear borders between the groups for LR (solid) and LDA (dotted line).

3 Comparison criteria

The simplest and the most frequently used criterion for comparison between the two methods is classification error (percent of incorrectly classified objects; CE). However, classification error is a very insensitive and statistically inefficient measure (Harrell, 1997). The fact is that the classification error is usually nearly the same in both methods, but, when differences exist, they are often overestimated (for example, if the threshold for "yes" is 0.50, a prediction of 0.99 rates the same as one of 0.51). The minimum information gained with the classification error is in the case of categorical explanatory variables. The boundary lines in figures below differ approximately equally in coefficients, but the classification errors provide different information. In Figure 2a, one of the

possible outcomes lies in the area where the lines are different, and therefore the predictions will differ in all objects with this outcome. On the contrary, the area between the lines in Figure 2b covers none of the possible outcomes. The classification error therefore does not reveal any difference.

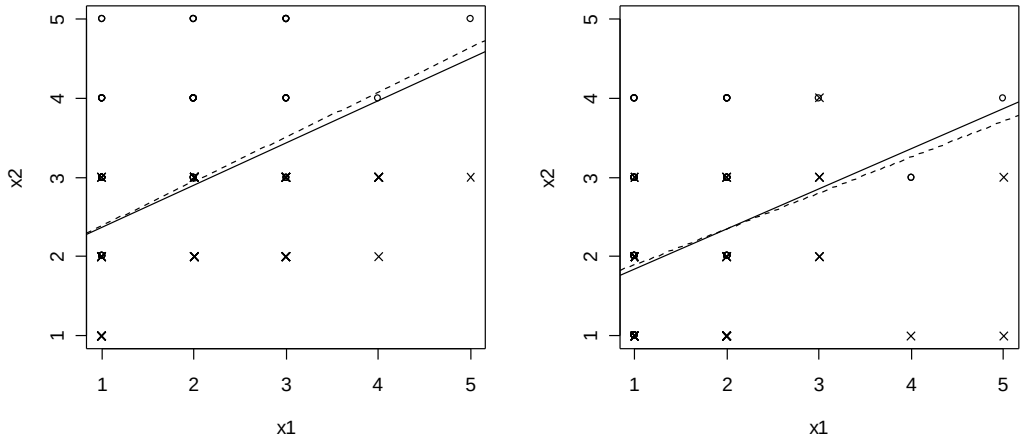

Figure 2a and 2b: Examples of categorised explanatory variables.

Since more information is needed regarding the predictive accuracy of the methods than just a binary classification rule, Harrell and Lee (1985) proposed four different measures of comparing predictive accuracy of the two methods. These measures are indexes A, B, C and Q. They are better and more efficient criteria for comparisons and they tell us how well the models discriminate between the groups and/or how good the prediction is. Theoretical insight and experiences with simulations revealed that some indexes are more and some less appropriate at different assumptions. In this work, we focus on three measures of predictive accuracy, the B, C and Q indexes. Because of its intuitive clearness we sometimes add the classification error (CE) as well.

The C index is purely a measure of discrimination (discrimination refers to the ability of a model to discriminate or separate values of Y). It is written as follows

$$C = \sum_{\substack{i=1 \\ Y_i=0}}^{n} \sum_{\substack{j=1 \\ Y_j=1}}^{n} [I(P_j > P_i) + \frac{1}{2} I(P_j = P_i)] / n_0 n_1 \qquad (3.1)$$

where Pk denotes an estimate of P(Yk=1|Xk) from (2.1) and I is an indicator function.

We can see that the value of the C index is independent of the actual group membership (Y), and as such it is only a measure of discrimination between the groups, and not a measure of accuracy of prediction. A C index of 1 indicates perfect discrimination; a C index of 0.5 indicates random prediction.

The B and Q indexes can be used to assess the accuracy of the outcome prediction. The B index measures an average of squared difference between an estimated and actual value:

$$B = 1 - \sum_{i=1}^{n} (P_i - Y_i)^2 / n \tag{3.2}$$

where Pi is a probability of classification into group i, Yi is the actual group membership (1 or 0), and n is the sample size of both populations. The values of the B index are on the interval [0,1], where 1 indicates perfect prediction. In the case of random prediction in two equally sized groups, the value of the B index is 0.75.

The Q index is similar to the B index and is also a measure of predictive accuracy:

$$Q = \sum_{i=1}^{n} \left[1 + \log_2 (P_i^{Y_i} (1 - P_i)^{1-Y_i}) \right] / n \, . \tag{3.3}$$

A score of 1 of the Q index indicates perfect prediction. A Q index of 0 indicates random predictions, and values less than 0 indicate worse than random predictions. When predicted probabilities of 0 or 1 exist, the Q index is undefined. The B, C and Q indexes are discussed further by Harrell and Lee (1985).

While the C index is purely a measure of discrimination, the B and Q indexes (besides discrimination) also consider accuracy of prediction. Hence, we can expect these two indexes to be the most sensitive measures in our simulations. Instead of comparing the indexes directly, we will often focus only on the proportion of simulations in which LR predicts better than LDA. As we always perform 50 simulations, this proportion will be statistically significant whenever it lies outside the interval [0.36, 0.64].

4 Description of the Simulations

4.1 Basic function

The basic function enables us to draw random samples of size n and m from two multivariate normal populations with different mean vectors, but equal covariance matrix Σ. The mean vector of one group is always set at (0,0). The distance to the other one is measured using Mahalanobis distance, while the direction is set as the angle (denoted by υ) to the direction of the eigenvector of the covariance matrix.

Each sample is then randomly divided into two parts, a training and a test sample. The coefficents of LDA and LR are computed using the first sample and then predictions are made in the second one. The sampling experiment is replicated 50 times. Each time the indexes for both methods are computed. Finally, the average value of indexes and the proportion of simulations in which LR performs better are recorded.

4.2 Categorization

After sampling, the normally distributed variables can be categorised, either only one or both of them. The minimum and maximum value are computed, then the whole interval is divided into a certain number of categories of equal size.

4.3 Skewness

As in the case of categorization, we can also decide here to transform only one of two explanatory variables or both of them. The Box-Cox type of transformation (Box and Cox, 1964) is used to make normal distribution skewed.

4.4 Remarks

To ensure clarity of the graphical representation, we have confined ourselves to a two- dimensional perspective, i.e. two explanatory variables. We have nevertheless made some simulations in more dimensions, but the trends of the results seemed to follow the same pattern.

In most of the simulations we have also set an upper limit for the Mahalanobis distance, in order to prevent LR from failing to converge and LDA from giving unreliable results.

To simplify, we have fixed the two group sizes as the same. As unequally sized groups (or unequal a prori probabilities in LR) only shift the border line closer to the smaller group (the one with the less probable outcome), this only impacts the constant, while the coefficient estimates remain the same.

All the simulations and computations were performed by using the statistical software package R.

5 Results

5.1 Comparison of methods when LDA assumptions are satisfied

We start from the situation where both explanatory variables are normally distributed. We observe the impact of changes connected with the parameters: sample size, covariance matrix, Mahalanobis distance and direction of distance between the group means.

The sample size has the most obvious impact on the difference between methods. LDA assumes normality and the errors it makes in prediction are only due to the errors in estimation of the mean and variance on the sample. On the contrary, LR adapts itself to distribution and assumes nothing about it. Therefore,

in the case of small samples, the difference between the distribution of the training sample and that of the test sample can be substantial. But, as the sample size increases, the sampling distributions become more stable which leads to better results for the LR. Consequently, the results of the two methods are getting closer because the populations are normally distributed.

Table 1: Simulation results for the effect of sample size (n).

	B		**C**		**Q**		**CE**	
n	LR	LDA	LR	LDA	LR	LDA	LR	LDA
40	0.7747	**0.7861**	0.7190	**0.7199**	0.0489	**0.1089**	0.1785	**0.1700**
60	0.7846	**0.7925**	0.7405	**0.7405**	0.1029	**0.1334**	0.1693	**0.1647**
100	0.7939	**0.7993**	0.7593	**0.7590**	0.1313	**0.1541**	0.1591	**0.1527**
200	0.7967	**0.7982**	0.7536	**0.7537**	0.1456	**0.1514**	0.1593	**0.1585**
1000	0.8008	**0.8011**	0.7609	0.7609	0.1595	**0.1608**	0.1550	**0.1543**

The proportion of simulations in which LR performs better

	B		**C**		**Q**		**CE**	
N	LR better	same	LR better	same	LR better	same	LR better	same
40	0.18	0.00	0.36	0.18	0.14	0.00	0.24	0.32
60	0.20	0.00	0.36	0.28	0.20	0.00	0.36	0.28
100	0.20	0.00	0.48	0.16	0.22	0.00	0.26	0.18
200	0.24	0.00	0.48	0.08	0.24	0.00	0.36	0.24
1000	0.26	0.00	**0.62**	0.00	0.30	0.00	0.32	0.18

Parameters: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $\upsilon = \pi/4$

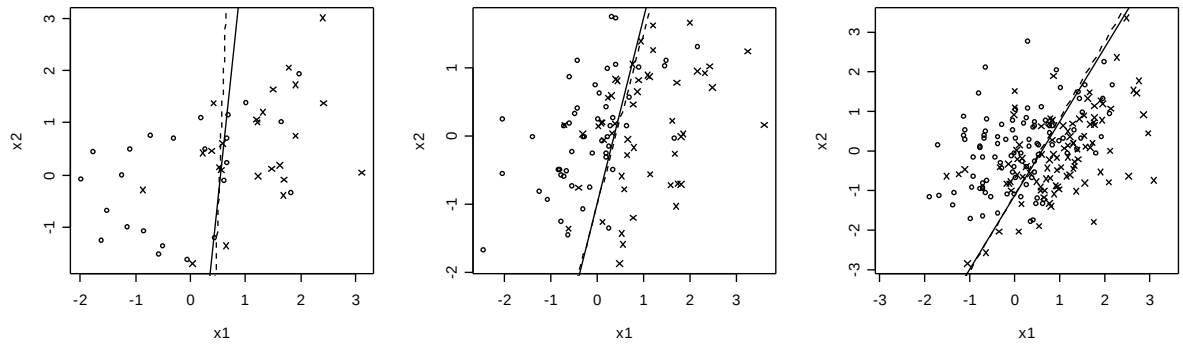

Figure 3: The impact of sample size of n=50 (left), n=100 (middle) and n=200 (right).

The results from Table 1 confirm the consideration above. As the sample size increases, the LDA coefficient estimations become more accurate and therefore all four indexes are improving (bold face is used to highlight the method that performs better). The LR indexes are increasing even faster, thus approaching those of LDA. Decreasing difference between the two methods is best presented with the Q index, which is the most sensitive one. As the differences between index means are negligible, it is also interesting to look at the proportion of simulations where LR performs better. It can be seen that the value of rates to

which we pay special attention, that of B index and of Q index, is constantly increasing.

In the case of other changes (tables below) the results of the two methods remain very close, in fact LDA is only a little bit better than LR. The exception appears in the case of large Mahalanobis distance presented in Table 4. We can see that for low values of Mahalanobis distance LDA yields better results, but as this distance increases and it takes values above 2, LR performs better.

Table 2: Simulation results for the effect of correlation between explanatory variables(σ).

	B		**C**		**Q**		**CE**	
σ	LR	LDA	LR	LDA	LR	LDA	LR	LDA
0	0.7938	**0.7979**	0.7536	0.7533	0.1340	**0.1499**	0.1623	0.1587
0.20	0.7909	**0.7967**	0.7490	0.7495	0.1215	**0.1456**	0.1629	0.1587
0.50	0.7925	**0.7965**	0.7497	0.7498	0.1291	**0.1456**	0.1601	0.1580
0.90	0.7961	**0.7990**	0.7568	0.7567	0.1403	**0.1535**	0.1575	0.1561

The proportion of simulations in which LR performs better

	B		**C**		**Q**		**CE**	
Σ	LR better	same	LR better	same	LR better	same	LR better	same
0	0.20	0.00	0.54	0.12	0.26	0.00	0.30	0.22
0.20	0.12	0.00	0.32	0.12	0.18	0.00	0.20	0.36
0.50	0.20	0.00	0.44	0.12	0.20	0.00	0.34	0.22
0.90	0.20	0.00	0.46	0.18	0.26	0.00	0.32	0.30

Parameters: υ= π/4, m=n=50

Table 3: Simulation results for the effect of direction of distance between group means(υ).

	B		**C**		**Q**		**CE**	
υ	LR	LDA	LR	LDA	LR	LDA	LR	LDA
0	0.7928	**0.7969**	0.7502	0.7501	0.1322	**0.1475**	0.1629	0.1609
π/4	0.7957	**0.7989**	0.7548	0.7547	0.1392	**0.1524**	0.1579	0.1565
π/3	0.7991	**0.8029**	0.7642	0.7645	0.1491	**0.1644**	0.1511	0.1480
π/2	0.7966	**0.8012**	0.7620	0.7619	0.1428	**0.1613**	0.1579	0.1569

The proportion of simulations in which LR performs better

υ	**B**		**C**		**Q**		**CE**	
	LR better	same	LR better	same	LR better	same	LR better	same
0	0.18	0.00	0.44	0.14	0.16	0.00	0.28	0.36
π/4	0.30	0.00	0.44	0.26	0.36	0.00	0.34	0.18
π/3	0.22	0.00	0.40	0.14	0.28	0.00	0.24	0.34
π/2	0.22	0.00	0.36	0.30	0.24	0.00	0.32	0.30

Parameters: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, m=n=50

Table 4: Simulation results for the effect of Mahalanobis distance (M).

	B		C		Q		CE	
M	LR	LDA	LR	LDA	LR	LDA	LR	LDA
0.50	0.7687	**0.7697**	**0.6769**	0.6767	0.0525	**0.0554**	0.1889	**0.1871**
1.00	0.7947	**0.7985**	**0.7552**	0.7551	0.1331	**0.1512**	0.1606	**0.1569**
1.25	0.8014	**0.8067**	0.7741	**0.7747**	0.1568	**0.1799**	0.1486	**0.1458**
2.00	0.8305	**0.8315**	0.8372	**0.8374**	0.2612	**0.2650**	0.1241	**0.1224**
3.00	**0.8570**	0.8557	0.8857	**0.8860**	**0.3575**	0.3492	0.1026	**0.0975**
4.50	**0.8922**	0.8816	**0.9310**	0.9305	**0.4994**	0.4398	0.0756	**0.0747**

The proportion of simulations in which LR performs better

	B		C		Q		CE	
M	LR better	same	LR better	same	LR better	same	LR better	same
0.50	0.46	0.00	0.46	0.22	0.52	0.00	0.38	0.26
1.00	0.24	0.00	0.42	0.24	0.28	0.00	0.30	0.30
1.25	0.20	0.00	0.20	0.22	0.16	0.00	0.22	0.38
2.00	0.56	0.00	0.36	0.28	**0.60**	0.00	0.28	0.30
3.00	**0.60**	0.00	0.38	0.22	**0.70**	0.00	0.26	0.24
4.50	**0.90**	0.00	0.42	0.08	**0.90**	0.00	0.26	0.40

Parameters: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $\upsilon = \pi/4$, m=n=50

To sum up, we can say that in the case of normality LDA yields better results than LR. However, for very large sample sizes the results of the two methods become really close.

5.2 The effect of categorisation

The effect of categorisation is studied under the assumption that the explanatory variables are in fact normally distributed, but measured only discretely. This means they only have a limited number of values or categories. When the number of categories is big enough not to disturb the accuracy of the estimates, the categorisation will not cause any changes in our results. But when the values are forced into just a few categories, we can expect more discrepancies.

All the simulations in this section are performed in the following way: First, the values of the indexes for LR and LDA are calculated for the samples from the normally distributed population. We start from the situation, where the LDA performs better as shown in the previous section (in the tables, these results are denoted with ∞). These samples are then categorised into a certain number of categories and the indexes are again calculated and compared.

As expected, the effect of the categorisation depends somewhat on the data structure (the correlation among the variables), but nevertheless, in all the simulations similar trends can be observed.

Linear discriminant analysis proves to be rather robust. Its prediction power is not much lower when the values are in 5 or more categories, and it usually

performs better than LR. The story changes when the number of categories is low, and LR is the only appropriate choice in the binary case.

The effect of categorisation also depends on the significance of the effect of a certain explanatory variable on the outcome. This is understandable – a nonsignificant variable will not change the model if transformed. On the other hand, if two covariates, equally powerful when predicting the result, are categorised, each of them will have a similar impact on the result.

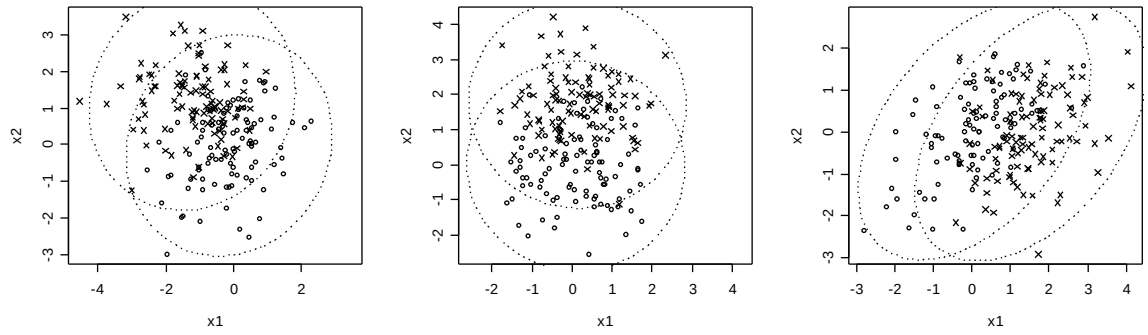

Figure 4a, 4b and 4c: The basic situations used in the study. The ellipses describe the distributions within the groups**.**

We have studied the impact of categorisation in two extreme and one intermediate case. Figures below present the situations that were the basis of our simulations. Figure 4a presents two uncorrelated explanatory variables with a similar impact on the outcome. In Figure 4b only one of the variables is significant, while in Figure 4c the covariates are correlated and both have a significant but different impact on the outcome variable.

Table 5a summarizes the results of the situation shown in Figure 4c. The upper part of this table contains the Q indexes for the case in which both covariates are categorised. It can be seen that the categorisation into only two categories severally lowers the predictive power of the two variables (the Q index falls close to zero) and that this effect is greater with LDA. For better clarity, the lower part of this table concentrates only on the proportion of the simulations in which the LR performs better (with regard to index Q) and compares these results with the categorisation of only one variable at a time. It is obvious that LR always outperforms LDA in the binary case. As discussed above, this effect is greater when we categorise the more significant variable (x2) and even more so when we categorise both explanatory variables.

The results summed up in Table 5b are similar. The effect of both x_1 and x_2 is similar and therefore the trends are even more comparable. However, logistic regression is not truly better even in the two category case. That is probably due to the too big “head start” of LDA. When categorising both covariates the advantages of LR are again more obvious.

Table 5a: Simulation results for different number of categories (Figure 4c).

	Q		
Num. of categ.	LR	LDA	LR better
2	0.0712	0.0579	**0.88**
3	0.0891	0.0839	**0.78**
4	0.1084	0.1076	**0.58**
5	0.1267	0.1281	0.46
10	0.1467	0.1505	0.18
∞	0.1553	0.1595	0.20
	The proportion of simulations in which LR performs better (Q index)		
Num. of categ.	x_1	x_2	Both
2	**0.58**	**0.70**	**0.88**
3	0.50	0.44	**0.78**
4	0.36	0.36	**0.58**
5	0.30	0.26	0.46
∞	0.20	0.20	0.20

Parameters: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, υ=0, m=n=200

Table 5b: Simulation results for different number of categories (Figure 4a).

	The proportion of simulations in which LR performs better (Q index)		
Num. of categ.	x_1	x_2	Both
2	0.48	0.40	**0.74**
3	0.28	0.24	0.46
4	0.26	0.24	0.32
5	0.24	0.26	0.24
∞	0.26	0.26	0.26

Parameters: $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\upsilon=\pi/4$, m=n=200

Table 5c clearly shows the absence of any effect on the result when we categorise an insignificant variable (x1). The results in the second and the third column are practically the same, because categorising only x2 variable is the same as categorising both.

Table 5c: Simulation results for different number of categories (Figure 4b).

	The proportion of simulations in which LR performs better (Q index)		
Num. of categ.	x_1	x_2	Both
2	0.20	**0.78**	**0.76**
3	0.18	0.48	0.48
4	0.22	0.34	0.34
5	0.20	0.30	0.30
∞	0.26	0.20	0.20

Parameters: $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, υ=0, m=n=200

If the study of the categorisation effect is done by taking smaller samples, the advantages of LDA are greater (see the previous section). Therefore they do not tail off even in the case of a small number of categories. Table 5d presents the results of an identical situation as in the lower part of Table 5a, but the sample size is shrunk to 100 units.

Table 5d: Simulation results for different number of categories (Figure 4c).

	The proportion of simulations in which LR performs better (Q index)		
Num. of categ.	x_1	x_2	Both
2	0.42	0.24	0.54
3	0.34	0.32	0.42
4	0.24	0.22	0.26
5	0.24	0.30	0.26
∞	0.22	0.22	0.20

Parameters: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, υ=0, m=n=100

The results in this table tend to vary a bit. Too small a sample size, and at the same time a small number of outcomes, causes the results to be unreliable. This is even more obvious when the Mahalanobis distance is increased, because LR often has problems with convergence.

5.3 The effect of non-normality

In the case of categorical explanatory variables above, the assumption of normality has been preserved and only the consequences of discrete measurement have been studied. Now, we are interested in the robustness of LDA when the normality assumptions are not met and in how much better can LR be in these cases. As non-normality is a very broad term, we have confined ourselves to transforming normal distributions with a Box-Cox transformation and thus making them skewed.

Again we begin with the three situations shown in Figure 4 and transform them into what is shown in Figure 5.

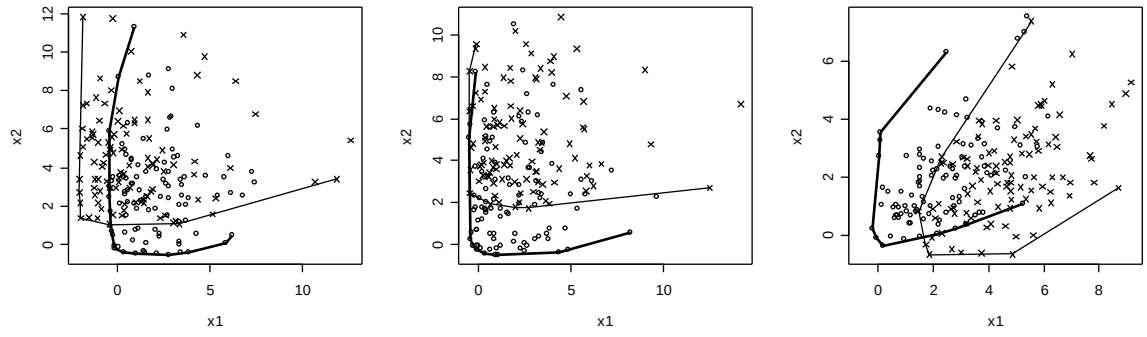

Figure 5a, 5b and 5c: Examples of right skewed distributions (to make groups more discernible, a part of the convex hull has been drawn for each of them).

Table 6a: Simulation results for different degree of skewness (Figures 4c, 5c).

	Q		
CS*	LR	LDA	LR better
-0.5	0.3149	0.2969	**0.88**
-0.4	0.2685	0.2610	**0.78**
-0.2	0.2262	0.2259	**0.60**
-0.1	0.1885	0.1920	0.28
0.1	0.1269	0.1293	0.44
0.2	0.1025	0.1007	**0.60**
0.4	0.0648	0.0494	**0.88**
0.5	0.0505	0.0267	**0.96**

*Coefficient of skewness

Parameters: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, υ=0, m=n=200

The performance of LDA and LR does not depend on the sign of the skewness. Therefore we have used the same transformation function to check the impact of the extent of separation of the groups at the same time. Right skewness thus also mean less separated groups. This is obvious in Table 6a, as index Q is constantly decreasing.

To be able to compare LR and LDA solely in terms of skewness we again focus on the proportion of simulations where LR does better. Tables 6b, 6c and 6d show the results for all the three cases we have described in Figures 4 and 5. The first two columns always show the results when only one of the two explanatory variables is skewed, while in the third column both are transformed.

The trends we can see are rather similar. When the skewness is small and therefore the distribution close to normal, LDA performs better. But when the skewness increases, LR becomes more and more constantly better.

Table 6b: Simulation results for different degree of skewness (Figures 4c, 5c).

	The proportion of simulations in which LR performs better (Q index)		
CS*	x_1	x_2	Both
-0.5	**0.68**	**0.74**	**0.88**
-0.4	0.50	0.44	**0.78**
-0.2	0.38	0.26	**0.60**
-0.1	0.24	0.24	0.28
0.1	0.28	0.28	0.44
0.2	0.38	0.42	**0.60**
0.4	**0.52**	**0.54**	**0.88**
0.5	**0.58**	**0.64**	**0.96**

*Coefficient of skewness

Parameters: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, υ=0, m=n=200

If both explanatory variables are skewed, the highest value of skewness under which LDA is still more appropriate is about ±0,2. We can observe that these boundaries are the same regardless of the separation of the groups.

If only one of the covariates is asymmetric and the other one is left as normal, the LDA is expectedly more robust – the interval widens a bit and the trends again remain similar with positive and negative skewness. The same effect on robustness can be seen by lowering the sample size as discussed in the previous sections.

Table 6d again shows that transforming insignificant variables has no impact on the results. However, it is impossible to control the simulations to the extent where we could say anything exact about the boundaries depending on the significance of the variables.

Table 6c: Simulation results for different degree of skewness (Figures 4a, 5a).

	The proportion of simulations in which LR performs better (Q index)		
CS*	x_1	x_2	Both
-0.5	**0.72**	**0.68**	**0.96**
-0.4	**0.54**	**0.58**	**0.80**
-0.2	0.38	0.32	**0.62**
-0.1	0.16	0.18	0.30
0.1	0.16	0.20	0.32
0.2	0.28	0.28	**0.56**
0.4	0.50	0.40	**0.86**
0.5	**0.58**	**0.50**	**0.94**

*Coefficient of skewness

Parameters: $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\upsilon = \pi/4$, m=n=200

Table 6d: Simulation results for different degree of skewness (Figures 4b, 5b).

	The proportion of simulations in which LR performs better (Q index)		
CS*	x_1	x_2	Both
-0.5	0.26	**0.92**	**0.92**
-0.4	0.26	**0.84**	**0.84**
-0.2	0.26	**0.64**	**0.64**
-0.1	0.26	0.32	0.32
0.1	0.26	0.38	0.38
0.2	0.26	**0.62**	**0.60**
0.4	0.26	**0.92**	**0.92**
0.5	0.26	**0.98**	**0.98**

*Coefficient of skewness

Parameters: $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\upsilon = 0$, m=n=200

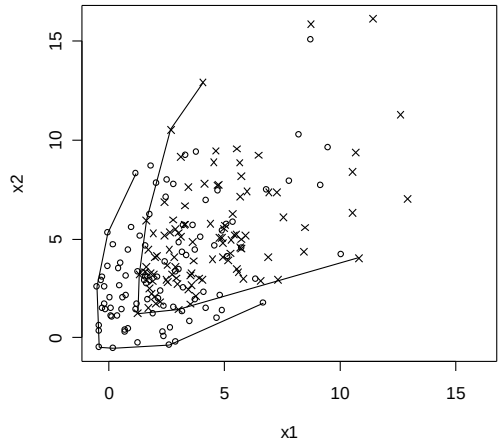

Figure 6: Right skewness with shifted centroids.

In this study we have confined ourselves to situations where the use of either LDA or LR is sensible. Figure 6 presents a situation similar to the one in Figure 4c, but with the centroids of the groups being shifted in a different direction. While we can imagine sensible linear boundaries on the Figures 5a, b and c, the boundary curve in Figure 6 is obviously not linear. Therefore more work should be done before using LDA or LR.

6 Conclusions and discussion

The goal of this paper was to compare logistic regression and linear discriminant analysis in order to set some guidelines to make the choice between the methods easier.

The methods do not differ in their functional forms. The difference rather lies in the estimation of the coefficients and we have focused our study on their predictive power.

The literature offers several criteria for comparison of the two methods. We have discussed some of them and showed that the classification error, although most frequently used, is not appropriate in our case. It is not sensitive enough and can be biased. We preferred the B and Q indexes, both leading to similar results in the sense of comparison of the predictive power of the two methods.

The idea of comparisons was to start with normally distributed covariates and thus satisfy the LDA assumptions, and then to check the robustness of the method by moving away from the assumptions in a closely controlled way.

When the covariates are simulated from the normal distribution, LDA of course seems to be the more appropriate method. However, the results of the two methods are really close when the sample size is large. The main differences can be observed for small samples, as their distributions vary too much for the LR to

be able to give good results. On the other hand, LDA assumes normality. The errors it makes in prediction are only due to the errors in estimation of the mean and the variance on the sample. To conclude, even though LDA is a constantly better method when the normality assumptions are met, the differences between the methods become negligible with a sample size of 50 and more, when the methods differently allocate only about 0.5% of the cases.

When comparing the robustness towards categorisation, we have again assumed the explanatory variables to be normally distributed, but discretely measured. LDA remains the favourite method if the number of categories is big enough to let the estimated mean and variance be close to the population values of the continuous explanatory variables. Usually, 5 categories are enough, but in the case of two or three categories, the advantages of the LR prevail. The impact of categorising the covariates of course depends on its correlation with the outcome variable. A variable with a small predictive power will not considerably affect the final result, whether we categorise it or not.

Whenever the explanatory variables are not normally distributed, the usage of LDA is theoretically wrong, as the assumptions are violated. The goodness-of-fit is therefore only more or less coincidental. On the other hand, the LR fits well to many types of distribution. The rationale for its use has been discussed extensively in the literature (Cox, 1970). It has been shown that many types of underlying assumptions lead to the same logistic formulation (Anderson, 1972).

To illustrate the above conclusions we can take a look at two uniformly distributed explanatory variables shown on the Figure 7a. The boundary between the two groups is obvious and linear. LR therefore finds a straight line that perfectly discriminates the data (7b). The LDA, however, again assumes multivariate normal distribution in both groups – the ellipses in Figure 7c show the assumed normal distribution, calculated on the basis of the means and variances of the two groups. The linear boundary that follows from these calculations, of course, is not optimal.

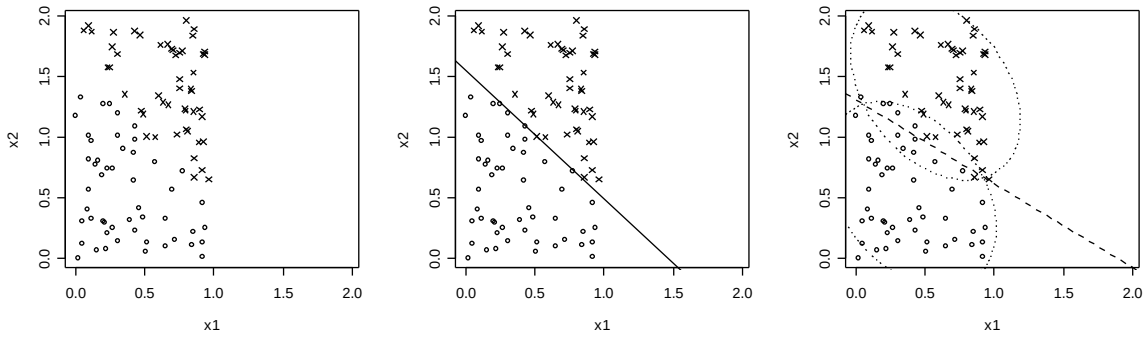

Figure 7a, 7b and 7c: The behaviour of the LR and LDA in the case of uniformly distributed covariates.

Distributions that do not deviate much from normality can be the only exception to the above reasoning. In order to study this, we have skewed the normally distributed explanatory variables. In the case of two covariates, the LDA remains better than LR when the skewness is in the interval [-0.2,0.2]. If only one of the explanatory variables is skewed, this interval is a bit wider. Whenever the distribution is obviously skewed (more than ±0.5), the LR constantly gives better results.

To conclude, we can say that LDA is a more appropriate method when the explanatory variables are normally distributed. In the case of categorised variables, LDA remains preferable and fails only when the number of categories is really small (2 or 3). The results of LR, however, are in all these cases constantly close and a little worse than those of LDA. But whenever the assumptions of LDA are not met, the usage of LDA is not justified, while LR gives good results regardless of the distribution. As the estimates for LR are obtained by the maximum likelihood method, they have a number of nice asymptotic properties as well.

References

[1] Anderson, J.A. (1972): Separate Sample Logistic Discrimination. *Biometrika*, **59**, 19-35.

[2] Box, G.E.P. and Cox, D.R. (1964): An analysis of transformations. *JRSS*, **26**, 211-246.

[3] Cox, D.R. (1970): *The Analysis of Binary Data*. London: Methuen & Co.

[4] Ferligoj, A. (2003): Multivariate analysis. Lecture notes.

[5] Harrell, F.E. and Lee, K.L. (1985): A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. In P. K. Sen (Ed.): *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*. North-Holland: Elsevier Science Publishers, 333-343.

[6] Harrell, F.E. (1997): Translating probability models into clinical decisions. Lecture Notes.

[7] Hosmer, D.W. and Lemeshow, S. (1989): *Applied Logistic Regression*. New York: Wiley.

[8] Johnson, R.A. and Wichern D.W. (2002): *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

[9] Kachigan, S.K. (1991): *Multivariate Statistical Analysis*. New York: Radius Press.

[10] Marcoulides, G.A. and Hershberger S.L. (1997): *Multivariate Statistical Methods*. New Jersey: Publishers.

[11] Portier, K.M.: Discriminant analysis pattern recognition, STA 4702 Multivariate statistical methods, STA 5701 Applied multivariate methods.

[12] Worth, A.P. and Cronin, M.T.D. (2003): The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Theochem*, **622**, 97-111.