Discriminant analysis and applied regression

Constantin Târcolea and Adrian Stere Paris

Abstract. In many applied disciplines it is measure several variables on each subject or experimental unit. Usually the variables are not only correlated with each other, but each variable is influenced by the other variables. With multivariate techniques can be examined the effects of several variables and at the same time the performance. The first part of the paper deals with discriminant analysis. A numerical example illustrates this technique for two groups.The second part of the paper deals with nonlinear regression analysis. In the last part of the paper it is proposed new adaptive cumulative distributions functions of the lifetime data, using this nonlinear technique. Numerical examples are provided for these models and a comparison of them it is discussed.

M.S.C. 2010: 62H30, 62N05.
Key words: : discriminant analysis, nonlinear regression, lifetime laws

1 Introduction

Multivariate analysis (MVA) techniques allow more than two variables to be analysed at once. Multivariate Data Analysis (MDA) consists of a collection of statistical techniques that can be used in many applied disciplines, when several measurements are made on each subject or object of a population. With multivariate techniques can be examined the effects of several variables and at the same time the performance [1], [2], [7]. Discriminant analysis is an analytical technique whereby a multivariate data containing several variables is separated into a number of groups, using discriminant functions. Some of the ideas associated with discriminant analysis go back to the bigining of the XIX century [5]. Multiple regression can be thought of as a multivariate analysis too. Regression analysis is the study of the relationship between one or several predictors (independent variables) and the response (dependent variable). To perform regression analysis on a data set, a regression model is first developed. The best-fit parameters are estimated using something like the least-square technique, the Levenberg-Marquardt method.

BSG Proceedings 18. The Int. Conf. of Diff. Geom. and Dynamical Systems (DGDS-2010), October 8-11, 2010, Bucharest Romania, pp. 95-100.

2 Discriminant analysis for two groups

It is assumed that two populations to be compared have the same covariance matrix Σ but distinct mean vectors μ_1 and μ_2. Let the samples $y_{11}, y_{12}, \ldots, y_{1n1}$ and $y_{21}, y_{22}, \ldots, y_{2n2}$ from the populations. A linear combination $z = a'y$ transforms each observation vector to a scalar [4]. The discriminant function is the linear combination of these variables, which maximizes the distance between the two (transformed) group mean vectors. It is computed the means and the problem is to find a vector that maximizes the squared distance which can be expressed as

$$\frac{(\overline{z_1} - \overline{z_2})^2}{s_z^2} = \frac{[a'(\overline{y_1} - \overline{y_2})]^2}{a' S_{pl} a}. \tag{2.1}$$

The maximum of the function (2.1) occurs when (2.2) or when a is any multiple of this vector [3], [6]:

$$a = S_{pl}^{-1}(\overline{y_1} - \overline{y_2}). \tag{2.2}$$

Example 1 (computed with software StatistiXL)

GROUPS	y_1	y_2
1	48	40
1	55	42
1	65	63
1	70	55
1	60	47
1	68	58
1	72	61
2	30	25
2	30	40
2	35	35
2	45	32
2	42	34
2	41	31

Discriminant Analysis Results for:
Descriptive Statistics

Factor	Variable	Mean	Std Dev.	Std Err	N
1	y_1	62.571	8.715	3.294	7
1	y_2	52.286	9.268	3.503	7
2	y_1	37.167	6.432	2.626	6
2	y_2	32.833	4.956	2.023	6

Pooled Covariance Matrix

	y_1	y_2
y_1	60.232	38.366
y_2	38.366	58.024

Unstandardised Discriminant Function Coefficients

Variable	Function 1
y_1	0.108
y_2	0.029

Standardised Discriminant Function Coefficients

Variable	Function 1
y_1	0.841
y_2	0.223

Group Centroids

Group	Function 1
1	1.533
2	-1.789

3 Regression and curve fitting

Regression analysis is the study of the relationship between one or several predictors (independent variables) and the response (dependent variable). To perform regression analysis on a data set, a regression model is first developed. Then the best-fit parameters are estimated using something like the least-square method. Finally, the quality of the model is assessed using one or more hypothesis tests. From a mathematical point of view, there are two basic types of regression: linear and nonlinear. A model where the fit parameters appear linearly in the Least Squares normal equations is known as a "linear model"; otherwise it is "nonlinear". In many scientific experiments, the regression model has only one or two predictors, and the aim of regression is to fit a curve or a surface to the experimental data. So we may also refer to regression analysis as curve fitting or surface fitting. Fitted curves can be used as an aid for data visualization, to infer values of a function where no data are available, and to summarize the relationships among two or more variables [8], [9], [11].

Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points, possibly subject to constraints. Curve fitting can involve either interpolation, where an exact fit to the data is required, or smoothing, in which a "smooth" function is constructed that approximately fits the data. A related topic is regression analysis, which focuses more on questions of statistical inference such as how much uncertainty is present in a curve that is fit to data observed with random errors. The reliability function is a corner stone of reliability theory. Denoting the lifetime variable of the system by T, the reliability function, $R(t)$, is the probability that the system will survive to time t and the respective failure (cumulative distribution) function is $F(t)$. It results, that is a one-to-one correspondence between $R(t)$ and $F(t)$. Therefore $R(t)$ contains no new information's beyond what is contained in $F(t)$. In other words the failure function of the discrete or continuous random variable is the mapping with definitions domain R_+, whose values are in [0;1]. In this study we analyzed the following continuous failure functions:

$$y(x) = a + bx + cx^2 + dx^2 + ex^4 + fx^5, \tag{3.1}$$

polynomial model failure function

$$y(x) = e^{a+\frac{b}{x}+c\ln x}, \tag{3.2}$$

vapor pressure model failure function

$$y(x) = \frac{a}{(1+e^{b-cx})^{\frac{1}{d}}}, \tag{3.3}$$

Richards model failure function

$$y(x) = ae^{\frac{b}{x}}, \tag{3.4}$$

modified exponential failure function.

The experimental data and the regression curves of the proposed models (eq. (3.1), (3.2), (3.3), (3.4)) were computed with software CurveExpert 3.1 and graphically represented in the figures 1,2,3,4.

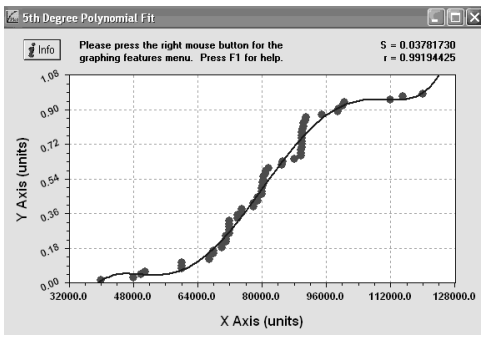

Fig. 1 $y(x) = -13.33 + 0.001x - -3 \star 10^{-8}x^2 + 4.17x^2 \star 10^{-12} - -2.73 \star 10^{-18}x^4 + 6.83 \star 10^{-24}x^5$

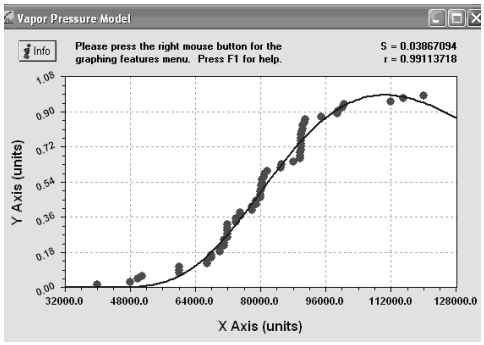

Fig. 2 $y(x) = e^{151.9\frac{1.32-10^6}{x}} - 12.04\ln x$

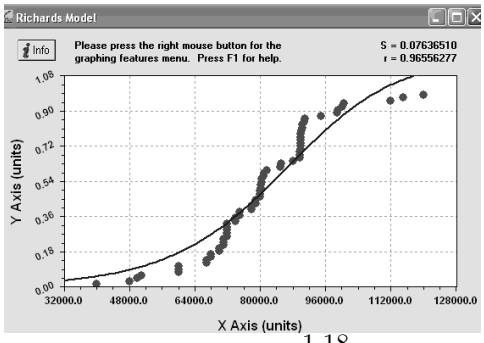

Fig.3 $y(x) = \dfrac{1.18}{(1+e^{7.07-7.8110^{-5}x})^{\frac{1}{1.31}}}$

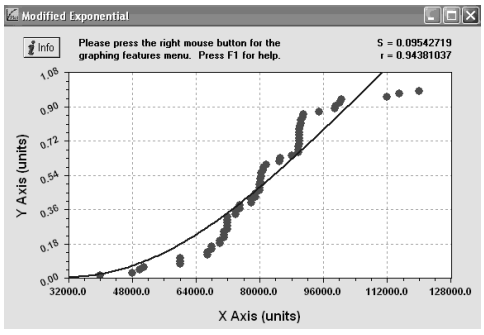

Fig.4 $y(x) = 8.9e^{\frac{-233055}{x}}$

The program uses the Levenberg-Marquardt method to solve nonlinear regressions for the estimation of the parameters of the model. This method combines the steepest-descent method and a Taylor series method to obtain a fast, reliable technique for

nonlinear optimization. This statement will be repeated later, but can bear stating at the beginning of this discussion: neither of the above optimization methods are ideal all of the time; the steepest descent method works best far away from the minimum and the Taylor series method works best close to the minimum.

4 Conclusions

The first part of the paper presents in brief applied multivariate data analysis methods, for modeling relationships among variables, and for exploring data patterns that may exist in more dimensions of the data. The methods presented in the paper usually involve analysis of data consisting of n observations on p variables. In the second part of the paper it was proposed adaptive stochastic distributions for describing the lifetime of the usual manufacturing models. The best results obtained with the adapted models are with the values of the correlation coefficient closed to one and the coefficient of variance very small. In the presented case results values for correlation coefficient greater than 0.95 are obtained for polynomial, vapor pressure and Richards model. Values for the coefficient of variance less than 0.05 are obtained for polynomial and vapor pressure model. The study utility consists in the fact that the obtained results can be useful in the design of the actual strategies and in the practice allowing the prediction of failures for the manufacturing systems. While there are numerous commercial software packages available for descriptive and inferential analysis of multivariate data such as SPSS,S-Plus, Minitab, SYSTAT and OriginPro, among others, we have chosen to use CurveExpert [11] and StatistiXL for Windows [10].

References

[1] A.A. Afifi, V. Clark, S. May, *Computer-Aided Multivariate Analysis*, Chapman and Hall/CRC Press, 2004.

[2] Gh. Amza, A. Paris, C. Târcolea, *Risk and factor analysis*, Proc. of the 3td WSEAS Int. Conf. on Risk Management, Assessment and Mitigation Rima'10, Univ. Politehnica, Bucharest, in vol. Recent Advances in Finite Differences-Finite Elements-Finite Volumes-Boundaries Elements, 146–151.

[3] R. Berndt, *Marketing 1*, Springer-Verlag, Berlin, 1996.

[4] J.H. Friedman, *Regularized Discriminant Analysis*, Journal of the American Statistical Association (American Statistical Association) 84 (405) (1989), 165-175.

[5] K. Pearson, *On Lines and Planes of Closest Fit to Systems of Points in Space Philosophical Magazine*, 2, 6 (1901), 559–572.

[6] A. Rencher, *Methods of multivariate analysis*, John Wiley & Sons, London, 2002.

[7] C. Târcolea, A. Paris, A. Târcolea-Demetrescu, *Statistical methods applied for materials selection*, The Int. Conf. DGDS-2008 & MENP-5 August 29 - September 02, 2008, Mangalia, Romania, In Appl. Sci. (APPS) 11 (2009), 145–150.

[8] C. Târcolea, A. Paris, C. Andreescu, *A comparison of reliability models*, BSG Proc. 16, Geometry Balkan Press, Bucharest 2009, 150–155.

[9] C. Târcolea, A.S. Paris, I. Tănase, *Models for the reliability of the manufacturing Systems*, BSG Proc. 14, Geometry Balkan Press, Bucharest 2007, 175-178.

[10] *** http://www.statistixl.com/
[11] *** http://s91928265.onlinehome.us/curveexpert/

Authors' addresses:

Constantin Târcolea
University Politehnica of Bucharest, Department of Mathematics-Informatics I,
Splaiul Independenţei 313, RO-060042 Bucharest, Romania.
E-mail: constantin_tarcolea@yahoo.com

Adrian Stere Paris
University Politehnica of Bucharest, IMST-Faculty Department of Technology,
Splaiul Independentei 313, RO-060042, Bucharest, Romania.
E-mail: s_paris@clicknet.ro