VECTOR DOT PRODUCT USED FOR REDUCED THREE INDEPENDENT VARIABLES OF MULTIVARIATE REGRESSION TO A LINEAR REGRESSION WITH ONE INDEPENDENT VARIABLE. ALCOHOLS USED LIKE A MODEL.

*E. CORNWELL**

*Departamento de Química Inorgánica y Analítica, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Casilla 233, Santiago, Chile.

ABSTRACT

The aim of this work is based in the reduction of independent variables in multivariate regression analysis to one by means a vector dot product (E_3). By this way, it is omit the orthogonalized procedure to obtained valid regression equation without co-linearity variables and valid signs supporting each independent variables factor, also by this procedure (E_3) it is possible to omit variable reduction process by means the Principal Components Analysis (PCA) and the used of others calibrations techniques in order to reach simples valid regressions functions. The reduction of three independent variables to one by (E_3) method, permit to applied linear regression (y = m*x + n) with clear significance on m and n parameters, this not occur in the original three-independent variable parameters regression, if it is not properly treatise.

INTRODUCTION

In the QSPR multivariate regression equations, the real significance of all factors and signs affecting each independent variable are obtained if orthogonal procedure[1] is carry on, or the reductions number of poor significant independent variables by means of Principal Component Analysis (PSA)[2] is applied. By other hand, is very important to considered the number of independent variables used in the mathematical regressions, its must be in accordance with the number of cases treatise, if not, the correlation determination coefficients (R^2) value is false by excess[3]. Other important aspect to be considered in multivariate regression analysis is the collianearity of the independent variables, this occur when the regression of each independent variable is correlated in turn against the other variables and the regressions determinant coefficient (R^2) are superior to 0.900 value[4].

Others multivariate calibrations techniques are frequently applied in conjunction with PSA technique on multivariate functions, these techniques included multiple linear regression (MLR) used in this article, partial least-squares regression (PLS), continuum regression (CR), projection pursuit regression (PPR) locally weighted regression (LWR) and artificial neural network (ANNs) among others. Each of these methods possesses its own strengths and weaknesses, and which works best for a given problem depends on the characteristics of the data and objective of the analysis[5]. In quantitative structure-activity relationships studies (QSAR) principal component analysis followed by sample selection to fit factorial and fractional factorial designs has been reported[6]

More extensive multivariate calibration methodology is not used in this paper because it is an introduction one to propose a new idea, with a few numbers of cases.

METODOLOGY

Reduction method presented in this work, eliminated these troubles by using a linear simple regression ($y = m*E_3 + n$) where E_3 is function of three optimal variables chosen of a group of nine variables. E_3 is obtained by vector dot product. A similar reduction idea where proposed on V_3 index by the author[7] applied to saturated hydrocarbons but the calculus for obtained the variable reduction is different and with statistically results no so good for polar substances (alcohols).

The model used in this work consist in twenty seven alcohols whose boiling points used like dependent variable where extracted from the literature[8] and for each one of then, eight physicochemical parameter where chosen and one well-known topological index named Electrotopological index[9,10] ($E_{-Estate}$) was used. For this reduction procedure is necessary used a maximum three independent variable by each multivariate regression, in accordance with the number of cases treatise[3]. Based on a combination procedure, forty eight regressions were made, each one with three independent variable using alcohols boiling point (Bp °C) like dependent variable, and from this forty eight modeling multi-regression calculated, one of them was chosen like the best in accordance with common regression statistical criteria. The structure of this model correspond to equation 1

$$Bp\ °C = A*x_i + B*y_i + C*z_i \quad (1)$$

i mean an alcohol, x_i corresponds to $E_{-Estate}$ topological index, y_i correspond to partition ration of octanol/water (log P)[11], z_i corresponds to molecular surface ($A°$)^2 (S)[11]. Other physicochemical parameters[11] considered were: molecular volume, density, refraction index, polarizability, dipolar momentum and hydratation energy. None of then gave better results like the three ones mentioned before. All physicochemical values were obtained by Hyperchem 7 program[11] and $E_{-Estate}$ index, was obtained by Dragon Software[10] by this way it was establish the triad elements belonging to nine independent variables set which permit to obtained the best multi-regression and this relation was compared statistically against the linear regression ($y = m*E_3 + n$) resulting by the reduction procedure through vector dot product (E_3)

E_3 parameter was obtained by the following processes:

The Q matrix rows were building by triads of alcohols independent variables corresponding to physicochemical parameters that were used in the optimal multi-regression. To applied mechanism reduction (E_3) was necessary to have defined a vector of three independent variables used like comparative vector. From twenty seven comparative vectors, only one representing the average (p) values of each parameter class produced the best results (an acceptable calculated alcohols boiling point vs. E_3) This was defined like comparative vector [X_p, Y_p, Z_p] the p symbol represent average value.

$$Q = \begin{vmatrix} X_1 & Y_1 & Z_1 \\ X_2 & Y_2 & Z_2 \\ * & * & * \\ X_n & Y_n & Z_n \end{vmatrix}$$

The calculus of E_3 was obtained by equation 2

$$E_3 = [X_i, Y_i, Z_i] * [X_p, Y_p, Z_p] \text{ (scalar number)} \quad (2)$$

i denoted a particular alcohol. The result is a scalar number that is possible to associate with any dependent variable, in this case the alcohols boiling points.

PROCEDURE AND DISCUSSION

Twenty seven alcohols are characterize by a three optimal independent variables: $E_{-Estate}$, log P, molecular surface area (S), ($A°$)^2 and the boiling point (Bp.°C) like dependent variable, see Table 1 The particular structure of equation 1 is obtained by Statgraphic program[12] corroborated by Origin 7 program[13] and by the theory based in linear algebra applied to multi-regressions[14] this equation number 3 is.

$$Bp,\ °C = -70.2249\ (\pm 25.1285) - 2.9976\ (\pm 0.8105)*E_{-Estate} + 15.6801\ (\pm 11.5207)*\log P + 0.64830\ (\pm 0.6483)*(S) \quad (3)$$

$R^2 = 92.524$
s.d. = 6.01
F = 94.88

Corresponding author: e-mail: ecornwell.garcia@gmail.com

Since the P-value in ANOVA analysis is less than 0.01 there is a statistically significant relationship between the variable at the 99% confidence level. The R-Squared statistic indicates that the model as fitted explain 92.52% of the variability in boiling point. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 91.5 %

The mean absolute error (MAE) is 4.51 and indicated the average value of residual

The study of collinearity[4] ($R^2 > 0.90$) present the following relations:

One way to checking for multicollinearity is to regress each independent variable in turn against all other predictors and to examine the statistically R^2 values, if its value goes above 90.0% multicollinearity is said to be a problem and is necessary othogonalized the system or to used PCA method.

E_{Estate} function of (log P, S) present R-square equal to 85.12%
Log P function of (E_{Estate}, S) present R-square equal to 95.23%
S function of (E_{Estate}, log P) present R-square equal to 91.03%

This result indicated collinearity between the independent variables. In part it can be simplified because the P-values of log P on regression is 0.1867, Since the P-value is greater or equal to 0.10, this variables is not statistically significant at the 90% or higher confidence level. Consequently, its possible considers removing log P from the model that is not the case for this study.

Table 1 columns 8, 9 are the calculated boiling points values from multivariate regression and the residuals of experimental and calculated boiling points.

Table 1 column 7 are present vector product values, $E_s = [X_i, Y_i, Z_i] * [X_p, Y_p, Z_p]$ where p indicated the average values from each column (4, 5, 6). The specific equation corresponding to that proposition is: $y = m*E_s + n$ named equation 4

$$Bp, °C = -51.4765(\pm 13.4100) + 0.0018(\pm 1.44 * 10^{-4}) * E \quad (4)$$

R = 0.9322
s.d = 7.64
F = 165.91

Table 1 Topological, Physicochenical and Vector Product Reduction of Aliphatic Alcohols Parameters

	Substances	Boiling Point °C	E_{Estate}	log P	Surface A^2	Vector Product a*b	Calculated Boiling Point °C Mult. Regression	Differences	Calculated Boiling Point °C Linear model	Differences
1	ethanol	78,0	0,752	0,08	214,51	66423,3	63,3	14,7	68,7	9,3
2	propanol	97,1	2,213	0,55	249,51	77273,3	91,8	5,3	89,1	8,0
3	isopropyl alcohol	82,4	2,563	0,49	247,89	76774,7	87,0	-4,6	88,1	-5,7
4	butanol	117,6	4,090	0,94	283,63	87854,5	114,6	3,0	108,9	8,7
5	2-methyl-1-propanol	108,1	4,622	0,95	280,20	86797,1	111,4	-3,3	106,9	1,2
6	2-butanol	99,5	4,885	0,96	273,35	84678,5	108,0	-8,5	102,9	-3,4
7	pentanol	138,0	6,395	1,34	317,78	98448,6	136,1	1,9	128,8	9,2
8	3-methyl-1-butanol	131,0	6,946	1,27	311,34	96459,3	128,0	3,0	125,0	6,0
9	2-methyl-1-butanol	128,0	7,282	1,34	303,71	94100,0	126,6	1,4	120,6	7,4
10	2-pentanol	119,3	7,523	1,36	303,61	94071,2	126,4	-7,1	120,6	-1,3
11	3-pentanol	116,2	7,919	1,43	296,42	91848,6	124,9	-8,7	116,4	-0,2
12	3-methyl-2-butanol	112,9	8,267	1,36	295,55	91582,1	119,9	-7,0	115,9	-3,0
13	hexanol	157,6	9,136	1,73	352,20	109130,2	155,2	2,4	148,8	8,8
14	3-methyl-1-pentanol	153,0	9,970	1,67	336,91	104403,3	142,9	10,1	139,9	13,1
15	4-methyl-1-pentanol	151,9	9,738	1,67	344,87	106865,8	147,1	4,8	144,5	7,4
16	2-methyl-1-pentanol	149,0	10,280	1,74	335,73	104040,7	144,2	4,8	139,2	9,8
17	2-ethyl-1-butanol	147,0	10,494	1,74	321,27	99565,5	137,4	9,6	130,8	16,2
18	2,3-dimethyl-1-butanol	144,5	11,092	1,68	328,16	101703,9	135,0	9,5	134,9	9,6
19	3,3-dimethyl-1-butanol	143,0	11,007	1,71	347,23	107607,7	144,3	-1,3	145,9	-2,9
20	2-hexanol	140,0	10,567	1,75	338,16	104795,6	144,3	-4,3	140,7	-0,7
21	2,2-dimethyl-1-butanol	136,5	11,779	1,85	333,95	103502,8	141,8	-5,3	138,2	-1,7
22	3-hexanol	135,0	11,186	1,82	328,58	101834,9	140,9	-5,9	135,1	-0,1
23	3-methyl-2-butanol	134,3	11,733	1,76	318,38	98681,5	131,8	2,5	129,2	5,1
24	4-methyl-2-pentanol	131,6	11,285	1,69	327,94	101637,5	134,5	-2,9	134,7	-3,1
25	2-methyl-3-pentanol	126,5	12,115	1,83	317,95	98551,8	133,1	-6,6	128,9	-2,4
26	3-methyl-3-pentanol	122,4	13,048	1,51	321,28	99590,5	116,2	6,2	130,9	-8,5
27	3,3-dimethyl-2-butanol	120,4	17,346	1,87	332,93	103235,6	118,2	2,2	137,7	-17,3

Table 1 columns 10, 11 are present the calculated boiling points values from linear equation ($y = m*E_s + n$) and the residuals of experimental and calculated boiling points.

The mean absolute error (MAE) is 6.30 and it indicated the average residual value.

The factors standard error of multivariable model and the linear equation proposed are present in Table 2, and Table 3.

Table2. Factors Standard Error and p-values of Multi-Variable Regression

Parameters	Estimate	Standard Error	P-value
Constants	-70.23	25.13	0.0103
E_{Estate}	-2.99	0.81	0.0012
log P	15.68	11.52	0.1867
S $(A^o)^2$	0.64	0.12	0.0000

Table3. Factors (m, n) Standard Error and p-values of $y = m^*E_s + n$

Parameters	Estimate	Standard Error	P-value
Intercept	-51.48	13.91	0.0011
Slope	0.0018	0.00014	0.000

The factor standard errors of multivariable regression are more significative than n, m factors standard errors of proposed model, see P-values, Table 2 and Table 3 The negative signs of the E_{Estate} have not physicochemical significance because the derivative function of boiling point vs. E_{Estat} is positive (derivative of boiling point vs. E_{Estate} is +3.59) in accordance to the following relation: to a greater number of E_{Estat} correspond a greater boiling point and consequently a greater molecular weight. Standardized skewness and standardized kurtosis are for both differences (Table 1 column 9, 11) within the range of –2 to +2 validating the following statistically parameters. An analysis of the statistically differences between experimental boiling points and calculated boiling point for both regression models (column 9, 11) using Statgraphic[11] software indicated that: there are not statistically significance differences between the means, standard deviation, median and distribution (Kolmogorov-Smirnov test) at 95.0% confidence level. Really, the factors and signs of the multivariate regression correlation do not have physicals sustenance, only is possible to use as a model to obtained calculated dependent variable, with spurious interpretation on independent variables factor and in many cases the signs of factors are wrong. For this reason is necessary applied an orthogonal method to multivariable regression or to use the method described in this paper to obtained a model consistent with a physicochemical interpretation.

DISCUSSION AND CONCLUSSIONS

Both models present similar differences of experimental boiling points vs. calculated boiling points but multivariate regression analysis model have not clear define the signs and magnitude affecting each independent variable. The model proposed in this paper is easy to obtain and its positive slope is on accordance with all positive slopes of the following derivatives: d Bp °C/ d E_{Estate}, d Bp °C/d log P, d Bp °C/d(S)

REFERENCES

1. M. Randic. *J. Chem.Inform.Comput. Sci.* 37, 672 (1997).
2. R. C. Graham "Data Analysis of the Chemical Sciences. A Guide to Statistical Techniques" U.C. Publisher. Inc (1993) page 329-346.
3. J. C. Toplis, R. P. Edwards. *J. Med. Chem.* 22, 1238 (1979).
4. Co lineal http:// 149.170.199.144/multivar/mr.htm
5. P. D. Wentzell, D. T. Andrews. *Anal. Chem.* 69, 2299 (1997).
6. J. Ferré, F. X. Rius. *Anal. Chem.* 68. 1565 (1996).
7. E. Cornwell. *J. Chil. Chem. Soc.* 51(1) 765, (2006).
8. D. R. Lide., H. P. R. Frederikse. "CRC Handbook of Chemistry and Physics" 75th Edition CRC Press, INC, 1995.
9. L. H. Hall. L.B. Kier. *J. Chem. Inf. Comput. Sci.* 40, 784 (2000).
10. Dragon Software.Talete SRL ViaV. Pisani,13-20124 Milano-Italy. E-mail admin@talete.mi.it
11. Hyperchem. Release 7.01 for Windows "Molecular Model System" (Evaluation Copy) Copyright 2002 Hypercube. Inc.
12. Statgraphic Plus 5.1 Copyrighy 1994-2001 Statistically Graphic Corp.
13. Origin 73R1 V7. 0301 (B30019) Copyright © 1991-2002 Origin Lab. Corporation. One Round Plaza Northampton, MA 01060 USA.
14. D. L. Massart., B.G.M. Vadegisnte., S.N.N. Deming., Y Machotte., L.Kaufman "Chemometric a textbook". Elsevier Scientific Publishing Company, Amsterdam, 1998.