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.
(Received 1st March 2007 – Accepted 1st June 2007)

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 ($E_i$). 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 ($E_i$) 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 ($E_i$) 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-independente variables 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 is carry on, or the reductions number of poor significant independent variables by means of Principal Component Analysis (PSA) 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. Other important aspect to be considered in multivariate regression analysis is the collinearity 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.

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. In quantitative structure-activity relationships studies (QSAR) principal component analysis followed by sample selection to fit factorial and fractional factorial designs has been reported.

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 ($y = m*x + n$) where $E_i$ is function of three optimal variables chosen of a group of nine variables. $E_i$ is obtained by vector dot product. A similar reduction idea where proposed on V$p$, index by the author applied to saturated hydrocarbons but the calculus for obtained the reduction procedure through vector dot product ($E_i$).

$i$ mean an alcohol, $x_i$ corresponds to $E_{\text{est.}}$ topological index, $y_i$ correspond to partition ration of octanol/water ($\log P$) and $z_i$ corresponds to molecular surface ($A^2$). Other physicochemical parameters considered were: molecular volume, density, refraction index, polarizability, dipolar momentum and hydration energy. None of then gave better results like the three ones mentioned before. All physicochemical values were obtained by Hyperchem 7 program and $E_{\text{est.}}$ index, was obtained by Dragon Software 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 ($y = m*x + n$) resulting by the reduction procedure through vector dot product ($E_i$).

$E_i$ 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 ($E_i$) 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. $E_i$) This was defined like comparative vector [ $X_p, Y_p, Z_p$ ] the $p$ symbol represent average value.

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

The calculus of $E_i$ was obtained by equation 2

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

$i$ 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: $E_{\text{est.}}$, $\log P$, molecular surface area ($S$), ($A^2$) and the boiling point ($Bp^\circ C$) like dependent variable, see Table 1 The particular structure of equation 1 is obtained by Statgraphic program corroborated by Origin 7 program and by the theory based in linear algebra applied to multi-regressions this equation number 3 is.

$$ Bp^\circ C = -70.2249 (\pm 25.1285) - 2.9976 (\pm 0.8105) * E_{\text{est.}} + 15.6801 (\pm 11.5207) * \log P + 0.64830 (\pm 0.6483) * (S) \quad \text{(3)} $$

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

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 (R² > 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 R² values, if its value goes above 90.0% multicollinearity is said to be a problem and is necessary orthogonalized the system or to used PCA method.

Eₑₑₑₑ function of (log P, S) present R-square equal to 85.12%
Log P function of (Eₑₑₑₑ, S) present R-square equal to 95.23%
S function of (Eₑₑₑₑ, 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, Eₐ = [ Xᵢ, Yᵢ, Zᵢ ] * [ Xₚ, Yₚ, Zₚ ] where p indicated the average values from each column (4, 5, 6). The specific equation corresponding to that proposition is: y = m*Eₐ + n named equation 4

Table 2. Factors Standard Error and p-values of Multi-Variable Regression

Table 1 columns 10, 11 are present the calculated boiling points values from linear equation (y = m*Eₐ + 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.
Table 3. Factors (m, n) Standard Error and p-values of \( y = mE^3 + n \)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-51.48</td>
<td>13.91</td>
<td>0.0011</td>
</tr>
<tr>
<td>Slope</td>
<td>0.0018</td>
<td>0.00014</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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 \( E_{\text{est}} \) have not physicochemical significance because the derivative function of boiling point vs. \( E_{\text{est}} \) is positive (derivative of boiling point vs. \( E_{\text{est}} \) is +3.59) in accordance with the following relation: to a greater number of \( E_{\text{est}} \) 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 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 physical 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 CONCLUSIONS**

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, d Bp °C/ d log P, d Bp °C/d(S)

**REFERENCES**

10. Dragon Software.Talete SRL ViaV. Pisani,13-20124 Milano-Italy. E-mail admin@talete.mi.it