Link between Orthogonal and Standard Multiple Linear Regression Models

Milan Šoškić*

Department of Chemistry, Faculty of Agriculture, University of Zagreb, HR-10000 Zagreb, The Republic of Croatia

Dejan Plavšić and Nenad Trinajstić

The Rugjer Bošković Institute, P.O.B. 1016, HR-10001 Zagreb, The Republic of Croatia

Received December 13, 1995[⊗]

Several topics in connection with a recently proposed method for the orthogonalization of predictor variables (dominant component analysis) are considered. Applying the sequential regression procedure, it is shown that dominant component analysis and the standard multiple linear regression method are directly related to each other. In addition, it is demonstrated that an earlier proposed iterative procedure for the orthogonalization of a correlated variable can be efficiently replaced by one step regression. It is also shown that the coefficient of determination for an orthogonal descriptor coincides with the corresponding squared semipartial correlation coefficient. Finally, the origin of extra information in an orthogonalized predictor variable is discussed.

INTRODUCTION

The strong intercorrelation between predictor variables (multicollinearity) may adversely affect the reliability of derived regression models and their applicability.[1] The regression coefficients of such models are unstable, that is, they are very sensitive to omission or inclusion of other variables in the model. Also, their standard errors can be very high. Various techniques have been proposed to overcome this problem.[1,2] Recently, dominant component analysis, a novel approach to the problem of multicollinearity, was suggested.[3] The method is based on the sequential removing of the superfluous information from the set of correlated variables, which results in the set of new, mutually orthogonal descriptors.[4,5] The order in which predictor variables are orthogonalized is very important, because it strongly affects the information content of an individual orthogonal descriptor.[5] Naturally, a whole set of the orthogonal descriptors, regardless of the sequence of orthogonalization, contains the same total information content as the set of original predictors.

It was found that applying dominant component analysis one may obtain not only more stable regression model[6−10] but also (for a particular orthogonalization ordering) the model with the improved statistics.[5] Another interesting finding that motivated this study is that the regression coefficients of an orthogonal model can be deduced from the gradual development of the corresponding nonorthogonal model.[11]

In the first part of this article we demonstrate that there exists the direct relation between the standard regression and dominant component procedure. In the next section we show that the method of iterative orthogonalization of variables can be replaced by a more effective one-step regression. It is followed by the paragraph in which the relationship between the squared semipartial correlation coefficient and the coefficient of determination for an orthogonalized descriptor is discussed. In the closing section, we give an illustrative example and try to answer the question, why some orthogonalized variables are better predictors than the original variables from which they are derived.

RELATION BETWEEN THE STANDARD REGRESSION AND DOMINANT COMPONENT PROCEDURE

Instead of the standard regression procedure of direct solving a system of normal eqs, a multiple linear regression model can also be derived through a series of simple regressions.[12] Naturally, for computational reasons the latter is not commonly used in practice. We will use it here to demonstrate an algebraic link between regression models based on nonorthogonalized and orthogonalized (dominant component) predictor variables. Suppose we model a physicochemical or a biochemical property (Y) using a set of three correlated variables (X_1, X_2, X_3). The corresponding regression eqs for the sequential developing of the hypothetical model are given in Table 1.

Regression of X_1 on Y gives the eq 1.1, where c is a constant, a is a regression coefficient, and e denotes a residual variation. To reduce the unexplained variance of the eq 1.1, we introduce the variable X_2 in the model. Being correlated with X_1, the part of the information content that X_2 shares with X_1 is previously removed from it (eq 1.3). The proportion of X_2 that is unique concerning X_1 is the residual term e_2. According to the dominant component concept, e_2 is the orthogonalized X_2 ($\Omega(X_2)$). Regressing e_1 against e_2 we obtain the eq 1.5. Since we regress the two sets of residual terms the constant term of the eq 1.5 is zero. Upon inserting expressions for e_1 and e_2 (eqs 1.2 and 1.4) in eq 1.5 and rearranging it we get the eq 1.6. In the same way the third predictor variable X_3 is introduced in the model. The variance in Y that remains unexplained after introduction of the second variable in the model (eq 1.7) is regressed against X_3, previously corrected for the overlapping information with X_1 and X_2 (eq 1.9). By definition, e_4 is the orthogonalized X_3 ($\Omega(X_3)$). It should be pointed out that eq 1.8 can also be derived sequentially, as will be shown in the next section (Table 2). To simplify the derivation, we give here only the final equation. Upon inserting values of e_3 and e_4 in the eq 1.11 and rearranging, the eq 1.12 is obtained.

[⊗] Abstract published in *Advance ACS Abstracts,* April 15, 1996.

S0095-2338(95)00183-1 CCC: $12.00

Table 1. Derivation of the Three Parameter Regression Model through a Series of Simple Regressions[a]

$Y = c_1 + a_1X_1 + e_1$	(1.1)
$e_1 = Y - (c_1 + a_1X_1)$	(1.2)
$X_2 = c_2 + a_2X_1 + e_2$	(1.3)
$e_2 = X_2 - (c_2 + a_2X_1) = \Omega(X_2)$	(1.4)
$e_1 = a_3e_2 + e_3$	(1.5)
$Y = c_1 + a_1X_1 + a_3[X_2 - (c_2 + a_2X_1)] + e_3$	(1.6)
$e_3 = Y - (c_1 + a_1X_1) - a_3[X_2 - (c_2 + a_2X_1)]$	(1.7)
$X_3 = c_3 + a_4X_1 + a_5X_2 + e_4$	(1.8)
$e_4 = X_3 - (c_3 + a_4X_1 + a_5X_2) = \Omega(X_3)$	(1.9)
$e_3 = a_6e_4 + e_5$	(1.10)
$Y = c_1 + a_1X_1 + a_3[X_2 - (c_2 + a_2X_1)] + a_6[X_3 - (c_3 + a_4X_1 + a_5X_2)] + e_5$	(1.11)
$Y = c_1 + a_1\Omega(X_1) + a_3\Omega(X_2) + a_6\Omega(X_3)$	(1.12)
$Y = (c_1 - a_3c_2 - a_6c_3) + (a_1 - a_2a_3 - a_4a_6)\,X_1 + (a_3 - a_5a_6)\,X_2 + a_6X_3$	(1.13)

[a] The derivation was performed in order to demonstrate the direct correspondence between the model based on dominant component descriptors (eq 1.12) and the standard multiple linear regression model (eq 1.13).

This equation is the sought link between the standard regression model and orthogonal (dominant component) model. (The same holds true for eq 1.6, but we will restrict the discussion to the former one.) Namely, it is easy to see that the algebraic expressions in square brackets are the orthogonalized X_2 and X_3. Taking into account that the first orthogonal descriptor remains unchanged in the process of orthogonalization of variables (i.e., $X_1 = \Omega(X_1)$), eq 1.11 can be rewritten in the orthogonal form as it is given in the fitted eq 1.12. Alternatively, we can rewrite eq 1.11 in its standard, nonorthogonal regression form as in fitted eq 1.13. Since both the orthogonal and the standard model are contained in implicit forms in eq 1.11, it is clear that their statistics[11] are the same. From eqs 1.11 and 1.13, it can be seen why the regression coefficient[11] of a variable that is last introduced in a nonorthogonal model is always equal to the regression coefficient of the corresponding orthogonalized variable (i.e., coefficients of X_3 and $\Omega(X_3)$). As is seen from eq 1.13, only the coefficient of X_3 is not influenced by the correction terms that are the result of interrelations between the variables. Naturally, in the case of uncorrelated variables the correction terms are zero.

In conclusion, using the sequential regression procedure one can generate simultaneously both an orthogonal and the corresponding nonorthogonal model. These are statistically equivalent, but the former is more attractive for quantitative structure−activity relationships (QSAR) and the quantitative structure−property relationships (QSPR) studies, since its regression coefficients can be interpreted more directly. In the nonorthogonal models, the direct interpretation of regression coefficients is prevented by the effect of other correlated variables.

A VARIABLE CAN BE ORTHOGONALIZED IN A SINGLE REGRESSION STEP

The method for the orthogonalization of predictor variables has been described in detail elsewhere.[3,11] The reason for the brief repetition of the method in this section is to illustrate that it follows the considered sequential procedure for the derivation of a multiple regression model and consequently that this approach can be replaced by a simple one-step calculation. For the arbitrarily selected order of the orthogonalization, the construction of the orthogonalized descriptors

Table 2. The Computational Procedure for the Orthogonalization of Predictor Variables as Proposed by Randić (Eqs 2.1−2.7)[a]

$X_1 = \Omega(X_1)$	(2.1)
$X_2 = d_1 + b_1X_1 + e_1$	(2.2)
$e_1 = X_2 - (d_1 + b_1X_1) = \Omega(X_2)$	(2.3)
$X_3 = d_2 + b_2X_1 + e_2$	(2.4)
$e_2 = X_3 - (d_2 + b_2X_1)$	(2.5)
$e_2 = b_3e_1 + e_3$	(2.6)
$e_3 = X_3 - (d_2 + b_2X_1) - b_3[X_2 - (d_1 + b_1X_1)] = \Omega(X_3)$	(2.7)
$X_3 = (d_2 - b_3d_1) + (b_2 - b_3b_1)X_1 + b_3X_2 + e_3$	(2.8)

[a] It is clear from eq 2.8 that the orthogonalized X_3 ($\Omega(X_3)$) can also be obtained by a single regression.

goes as follows (Table 2): The first variable remains unchanged. Therefore, $X_1 = \Omega(X_1)$. The second orthogonal descriptor $\Omega(X_2)$ is the residual of X_2, when it is regressed against X_1 (eq 2.3). The X_3 is orthogonalized in two steps. Initially it is regressed against X_1 (eq 2.4), and the residual obtained is than regressed against e_1 (eq 2.6). The residual of the last regression (e_3) is the third orthogonal descriptor $\Omega(X_3)$. For a larger set of descriptors the orthogonalization continues in the analogous way.

It is obvious that the orthogonalization of each subsequent descriptor becomes more and more complex in terms of the regressions to be performed. In general, $n-1$ simple regressions should be carried out to obtain an orthogonalized descriptor n. Turning back to Table 2, it is easy to see that eq 2.7 can also be written as eq 2.8. That means that instead of gradual derivation of the eq 2.7, one can obtain the orthogonalized X_3 ($\Omega(X_3)$) by direct regression of X_3, against the rest of predictor variables (eq 2.8). It is worth nothing that this significantly reduces the computational procedure, that is, for n orthogonalized descriptors the number of the regressions to be computed decreases for the factor $n-1$.

Finally, it should be pointed out that in order to avoid confusion with the algebraic regression coefficients, we use somewhat different notation in Tables 1 and 2. However, in a concrete example, it would be easy to recognize that the eqs 1.8 and 2.8 are identical.

SQUARED SEMIPARTIAL CORRELATION COEFFICIENT

Semipartial (or part) correlation coefficient[13] measures the relationships between the dependent variable Y and the residual of the predictor variable X_i, after regressing X_i on $X_1,...X_p$. Its square value can be calculated as the increase in r^2 caused by the inclusion of the variable X_i in the model:

$$r^2_{Y(i.1...i-1,i+1...p)} = r^2_{Y(1...i-1,i,i+1,...p)} - r^2_{Y(1...i-1,i+1...p)} \quad (3.1)$$

In eq 3.1, $r^2_{Y(i.1...i-1,i+1...p)}$ is the square of semipartial correlation coefficient between Y and X_i, corrected by overlapping effects of variables $X_1,...X_{i-1},X_{i+1},...X_p$, and $r^2_{Y(1...i-1,i,i+1,...p)}$ and $r^2_{Y(1...i-1,i+1...p)}$ are the simple coefficients of determination for the two consecutive models.

It is clear from the definition of the semipartial correlation coefficient that the former is identical with the simple correlation coefficient between Y and orthogonalized X_i. The same is true for their squared values. Hence, for the known simple correlation coefficients, the contributions of an orthogonalized descriptor to the regression can simply be calculated from eq 3.1, without its previous orthogonalization.

Table 3. All Possible Models for the Boiling Points of Nonanes Based on the Original and Orthogonalized Molecular Connectivity Indices ($^0\chi$ -$^2\chi$)[a]

nonorthogonalized descriptors	r^2
$^0\chi$	0.316
$^1\chi$	0.469
$^2\chi$	0.603
$^0\chi$, $^1\chi$	0.582
$^0\chi$, $^2\chi$	0.646
$^1\chi$, $^2\chi$	0.672
χ^0, $^1\chi$, $^2\chi$	0.803

	orthogonalized descriptors	r^2		orthogonalized descriptors	r^2
(1)	$\Omega(^0\chi)$	0.316	(2)	$\Omega(^0\chi)$	0.316
	$\Omega^0(^1\chi)$	0.266		$\Omega^0(^2\chi)$	0.330
	$\Omega^{0,1}(^2\chi)$	0.221		$\Omega^{0,2}(^1\chi)$	0.157
(3)	$\Omega(^1\chi)$	0.469	(4)	$\Omega(^1\chi)$	0.469
	$\Omega^1(^0\chi)$	0.113		$\Omega^1(^2\chi)$	0.203
	$\Omega^{1,0}(^2\chi)$	0.221		$\Omega^{1,2}(^0\chi)$	0.131
(5)	$\Omega(^2\chi)$	0.603	(6)	$\Omega(^2\chi)$	0.603
	$\Omega^2(^0\chi)$	0.043		$\Omega^2(^1\chi)$	0.069
	$\Omega^{2,0}(^1\chi)$	0.157		$\Omega^{2,1}(^0\chi)$	0.131

[a] The squared correlation coefficients (r^2) of the orthogonalized descriptors can be obtained from the coefficients of determination for the nonorthogonalized descriptors, by simple hand computations according to eq 3.1. Superscripts associated with Ω indicate the order in which the descriptors are orthogonalized. Within each of the six orthogonalization orderings ((1)−(6)), the coefficients of determination for any combinations of the orthogonal descriptors can be computed by adding their individual r^2.

AN EXAMPLE

Recently, the QSPR study on the boiling points (BP) of 35 nonanes has been reported, in which the connection between orthogonal and standard regression models was suggested in an indirect way.[11] That is the reason why we take this data set for an illustration of what we said before. For convenience, only the first three molecular connectivity indices[14,15] ($^0\chi$, $^1\chi$, $^2\chi$) from the connectivity basis used in the original paper are considered in the process of the model building. Therefore, the obtained three parameter model is not a representative one, but this is irrelevant for the purpose of the present report. Applying the sequential regression procedure one can derive the eq 4.1, from which it is easy to obtain the eqs for the orthogonal (4.2) and nonorthogonal (4.3) model. The expressions in the square brackets are the residuals of $^1\chi$ and $^2\chi$, after their regression on $^0\chi$ and $^0\chi$ and $^1\chi$, respectively.

$$\mathrm{BP} = 276.9791 - 19.0147^0\chi + 68.9534[^1\chi - (10.4652 - 0.8633^0\chi)] - 72.6696[^2\chi - (49.4058 - 2.7151^0\chi - 6.2788^1\chi)] \quad (4.1)$$

$$\mathrm{BP} = 276.98 - 19.02\Omega(^0\chi) + 68.95\Omega(^1\chi) - 72.67\Omega(^2\chi) \quad (4.2)$$

$$\mathrm{BP} = 3145.67 - 156.79^0\chi - 387.32^1\chi - 72.67^2\chi \quad (4.3)$$

As stated before, eqs 4.2 and 4.3 are statistically equivalent, but the regression coefficients of the former are more meaningful. It is seen from Table 3 that eq 4.2 is only one of the several possible three-parameter orthogonal models. The statistical characteristics of these models are the same, because the corresponding orthogonalized descriptors contain the same total information content. However, the individual contributions of these descriptors are different. That is so because the sequential removing of overlapping information affects the information content of an individual orthogonalized descriptor, while the information content of the whole set of orthogonalized variables remains the same, regardless of the sequence of orthogonalization. (Recall that the information content of the first orthogonalized descriptor is not changed in the process of orthogonalization.) Hence, the choice of the orthogonalization order is very important in the construction of QSAR and QSPR models.[3,5] In eq 4.2 the orthogonalized descriptors are arranged according to the arbitrarily chosen sequence of orthogonalization.

It was reported recently[5] that for a certain orthogonal ordering one can obtain the improved model compared to the model based on the nonorthogonalized variables. In other words, going through all possible orderings (for n predictor variables there are $n!$ of them) one can find a particular ordering with one or several descriptors that do not contribute significantly to the variability in the dependent variable. Consequently, their omission is followed by a negligible reduction in the explained variance of the model. Moreover, the standard deviation of such a model is lower, since the value of standard deviation depends on the number of predictor variables in the model. However, in our example this effect fails to appear in its extreme form: none of the orthogonalized variables displayed in Table 3 possess the above described characteristics. The closest to this is the orthogonalized zero-order molecular connectivity index ($\Omega^2(^0\chi)$). The superscript associated with Ω indicates that $^0\chi$ is regressed against the second-order molecular connectivity index. The removing of $\Omega^2(^0\chi)$ from the three-parameter model is followed by roughly 4% fall in r^2 and 10% increase in standard deviation. Nevertheless, the model ($\Omega(^2\chi)$, $\Omega^{2,0}(^1\chi)$) obtained after omitting $\Omega^2(^0\chi)$ is the best two-parameter model for the given set of variables. Its coefficient of determination is improved for nearly 10%, compared to the corresponding two-parameter nonorthogonalized model ($^2\chi$, $^1\chi$). Since $\Omega(^2\chi)$ and $^2\chi$ represent only different designations for the same descriptor, clearly, improvement of the model must be attributed to $\Omega^{2,0}(^1\chi)$. In other words, the orthogonalized $^1\chi$ is richer in unique information than ordinary $^1\chi$, from which it is generated. The question arises as to where does this extra information in $\Omega^{2,0}(^1\chi)$ come from? It is instructive in that sense to consider the changes in information content of the orthogonalized variables, as a result of the orthogonalization order. Generally, an orthogonalized variable is less informative than the ordinary variable from which it is obtained. Namely, from each orthogonalized descriptor is removed that part of information content that overlaps with the information content of the variables against which the original descriptor is orthogonalized. Hence, one expects that the orthogonal descriptor obtained by orthogonalization against n predictor variables is richer in information content than the one constructed by regression of the same original variable against $n+1$ predictors. Closer inspection of Table 3 reveals that this is not always so. It can be best illustrated by taking the orthogonalization of the variable $^0\chi$ as example (Table 4).

As expected, the orthogonal descriptors ($\Omega^1(^0\chi)$, $\Omega\ ^2(^0\chi)$) obtained by regression of $^0\chi$ against $^1\chi$ and $^2\chi$, respectively, are less informative than the $^0\chi$. Surprisingly, the orthogonal descriptor ($\Omega^{1,2}(^0\chi)$) derived by orthogonalization of $^0\chi$

Table 4. The Squared Correlation Coefficients (r^2) between Boiling Points of Nonanes (BP) and the Original and Orthogonalized (Ω) Zero-Order Molecular Connectivity Indices ($^0\chi$)[a]

	BP	$^0\chi$	$^1\chi$	$^2\chi$
$^0\chi$	0.316	1.000	0.920	0.740
$\Omega^1(^0\chi)$	0.113	0.080	0.000	0.059
$\Omega^2(^0\chi)$	0.043	0.261	0.062	0.000
$\Omega^{1,2}(^0\chi)$	0.131	0.006	0.000	0.000

[a] Molecular connectivity indices against which $^0\chi$ was regressed are indicated by superscripts associated with Ω. The coefficients of determination between the orthogonal descriptors and the molecular connectivity indices used in the orthogonalization procedure are also displayed.

against the two descriptors ($^1\chi$ and $^2\chi$) can explain the greater proportion of the variance in BP than individual $\Omega^1(^0\chi)$ and $\Omega^2(^0\chi)$. Before offering an explanation for this apparent inconsistency, we wish to recall that information content of the correlated variables may be affected by the presence of other variable in the model. Thus, we may have the case[3,16] that neither, for example, of the two independent variables contributes significantly to the explained variance of the model, but taken together they explain the reasonable part of the variations in Y. It appears that this effect is generally present in the models containing correlated descriptors, though not necessarily in such a drastic form. It follows from this that the information content of an orthogonalized descriptor is the result of the two parallel effects: the orthogonalization of an ordinary variable against a certain number of other correlated variables reduces its information content, but at the same time the joint effect of the variables (not necessarily a favorable one[6]) remains saved in the orthogonalized descriptor. From this simple explanation it becomes immediately clear why some orthogonal descriptors are better predictors than corresponding original ones and, hence, why some orthogonal models are superior to the corresponding nonorthogonal models.

CONCLUSIONS

It has been demonstrated that both the orthogonal and the corresponding nonorthogonal models can be simultaneously generated by the sequential regression procedure. Several other topics such as the improving of the orthogonalization algorithm and the relation between the squared semipartial correlation coefficient and the coefficient of determination for a dominant component descriptor as well as why some dominant component descriptors are more informative than ordinary variables have also been discussed.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grants 1-07-159 and 1-07-165.

REFERENCES AND NOTES

(1) Dillon, W. R.; Goldstein, M. *Multivariate Analysis*; Wiley: New York, 1984.
(2) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
(3) Randić, M. Orthogonal Molecular Descriptors, *New J. Chem.* **1991**, 15, 517−525.
(4) Randić, M. Resolution of Ambiguities in Structure−Property Studies by Use of Orthogonal Descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, 31, 311−320.
(5) Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D. The Structure−Property Models Can Be Improved Using the Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 532−538.
(6) Amić, D.; Davidović-Amić, D.; Trinajstić, N. Calculation of Retention Times of Anthocyanins with Orthogonalized Topological Indices. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 136−139.
(7) Randić M. Molecular Profiles: Novel Geometry-Dependent Molecular Descriptors. *New. J. Chem*. **1995**, 19, 781−791.
(8) Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. A Structure−Property Study of the Solubility of Aliphatic Alcohols in Water. *Croat. Chem. Acta*. **1995**, 68, 417−434.
(9) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D.; Jurić, A. A Novel QSPR Approach to Physicochemical Properties of the α-Amino Acids. *Croat. Chem. Acta*. **1995**, 68, 435−450.
(10) Šoškić, M.; Plavšić, D.; Trinajstić, N. 2-Difluoromethylthio-4,6-bis-(monoalkylamino)-1,3,5-triazines as Inhibitors of Hill Reaction: A QSAR Study with Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci*. **1996**, 36, 146−150.
(11) Randić, M.; Trinajstić, N. Isomeric Variations in Alkanes: Boiling Points of Nonanes. *New. J. Chem*. **1994**, 18, 179−189.
(12) Draper N. R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, 1981.
(13) Nie N. H.; Hull C. H.; Jenkins J. G.; Steinbrenner K.; Bent D. H. *SPSS: Statistical Package for Social Sciences*; McGraw-Hill: New York, 1975.
(14) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd: Letchworth, 1986.
(15) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(16) Pike D. J. In *Statistical Procedure in Food Research*; Piggott J. R., Ed.; Elsevier: London, 1986; Chapter 3, p 61.

CI950183M