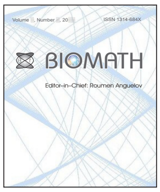
Original article

Biomath 2 (2013), 1309089, 1–11

Biomath Forum

BIOMATH

http://www.biomathforum.org/biomath/index.php/biomath/



Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example

Sorana D. Bolboacă
Department of Medical Informatics and Biostatistics
”Iuliu Haţieganu” UMF Cluj-Napoca
Cluj-Napoca, Romania
Email: sbolboaca@gmail.com

Lorentz Jäntschi
Department of Physics and Chemistry
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Email: lorentz.jantschi@gmail.com

Received: 4 May 2013, accepted: 8 September 2013, published: 18 September 2013

***Abstract*—Quantitative structure-activity relationships are mathematical models constructed based on the hypothesis that structure of chemical compounds is related to their biological activity. A linear regression model is often used to estimate and/or to predict the nature of the relationship between a measured activity and some measure or calculated descriptors. Linear regression helps to answer main three questions: does the biological activity depend on structure information; if so, the nature of the relationship is linear; and if yes, how good is the model in prediction of the biological activity of new compound(s). This manuscript presents the steps on linear regression analysis moving from theoretical knowledge to an example conducted on sets of endocrine disrupting chemicals.**

***Keywords*-robust regression; validation; diagnostic; predictive power; quantitative structure-activity relationships (QSARs);**

I. Linear Regression on QSAR Analysis

Quantitative structure-activity relationships (QSARs) are mathematical models linking chemical structure and pharmacological activity/property in a quantitative manner for a series of compounds [1]. The approaches are based on the assumption that the structure of chemical compounds (such as geometric, topologic, steric, electronic properties, etc.) contains features responsible for its physical, chemical and/or biological properties [2]. This assumption could be summarized as ”*similar compounds have similar properties*” [3].

The two main fields where linear regression analysis found its applicability are drug discovery [4], [5] and toxicology prediction [6], [7]. In both of these fields, the linear regression is used mainly to predict not to estimate (the model is used to quickly determine the activity/property of new/un-investigated compounds) [8].

The linear regression is used in QSAR analysis to linearly link the activity/property of chemical compounds (measured or observed value - outcome variable abbreviated as Y) and some values translated from the structure of the compounds and generally called descriptors (assumed error non-affected independent variables abbreviated as X(s)). The multiple linear regression (MLR) expression is presented in Eq(1):

$$\hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i \qquad (1)$$

where $\hat{Y}$ = estimated activity/property; b_0 = intercept; b_i = coefficient of the i^{th} independent variable / descriptor variable ($1 \leq i \leq k, 5 \times k \leq n$ [9]), k = number of descriptors (independent/descriptor variables) in the model, n = number of observations in the sample. The regression coefficients b_i could be interpreted as the change in Y when X_i increased or decreased by 1 unit

Citation: Sorana D. Bolboacă, Lorentz Jäntschi, Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example, Biomath 2 (2013), 1309089, http://dx.doi.org/10.11145/j.biomath.2013.09.089

when all other independent variables are held constant (b_0 and b_1 estimate the population parameters β_0 and β_i, [10]). The identified values of b_0 and b_i are calculated to minimize the squared error for all n observations.

A. Linear Regression Assumptions

The main assumptions of linear regression (Table I) could be summarized as:

1) Linearity. The relation between Y and each of descriptors X_i are linear.
2) Independence of the errors. Both the experimental values (Y) and experimental/calculated descriptors (X_i) are measured without errors.
3) Homoscedasticity. The variance of the errors is constant.
4) Normality. The dependent variable (Y) is normal distributed.
5) Absence of multicolinearity. The independent variables (X_i) are linearly independent of each other. Please note that this constrain did not exclude a certain degree of collinearity.

Since it has been recognized that "normal law ... is not valid in a great many cases which are both common and important" [11] a series of transformation could be used to reach normal distribution [29] (see Table II).

1) Model Selection and Diagnostic: Selection of the regression model is an important task that researchers must to accomplish. The main criteria useful in this step are:

- Determination coefficient (R^2) and its adjustment form (R^2_{adj} - R^2 adjusted with the number of coefficients in the model → the value will not necessary increase with the addition of X's). Generally, the R^2 increase with the number of parameters in the model but R^2_{adj} penalizes according to the number of parameters (the model with higher number of descriptors does not necessary has the higher value of R^2_{adj}).
- Standard error of the estimate: the average error predicting the activity/property of interest by the identified model.
- Statistics of overall model performances (F-value and associated p-value): assess the overall ability of a model to explain as much as possible from the observed variability in Y.
- Models performances in cross-validation by the leave-one-out analysis. It is say that a model with Q^2 (determination coefficient in cross-validation by the leave-one-out analysis) >0.6 and $|R^2 - Q^2| < 0.1$ is a desired model in QSAR analysis [30]. However, the value of F-statistics and its associated probability are as important as Q^2 in assessment of internal validation of a QSAR model.
- Mallows C_p-statistic ($C_p = SS_{res}/MS_{res} - n + 2 \cdot (k+1)$, k = number of descriptor variables in the model) [31], [32], [33]: measures the overall bias or mean square error in the estimated model parameters. This is a useful parameter when models with different X(s) are compared on the same sample of compounds. A low C_p value indicates good model prediction or a model with a small positive/negative discrepancy between C_p and $(k+1)$ - could be used in evaluating candidate regression models.
- Akaikes information criterion and derivative formulas: assess the degree of fit by involving the goodness-of-fit of the model (R^2): Akaike information criterion ($AIC = n \cdot ln(RSS/n) + 2 \cdot (k+1)$ for the model with intercept and $AIC = n \cdot ln(RSS/n) + 2 \cdot k$ for the model without intercept, where n = sample size, RSS = residual sum of squares; k = number of X_i) [34]; AIC based on the determination coefficient ($AIC_{R2} = ln[(1 - R^2)/n] + 2 \cdot (k+1)$); McQuarrie and Tsai corrected AIC ($AIC_u = ln[RSS/(n-k+1)] + (n+k+1)/(n-k-1)$) [35]; Bayesian Information Criterion ($BIC = n \cdot ln[RSS/(n-k+1)] + (k+1) \cdot ln(n)$) [36]; Amemiya Prediction Criterion ($APC = RSS/n \cdot (n-k+1)/(n+k+1)$) [37]; Hannan-Quinn Criterion ($HQC = n \cdot ln(RSS/n) + 2 \cdot (k+1) \cdot ln[ln(n)]$ [38]. The smallest the AIC, BIC, APC and HQC values are the better the model is considered. In addition to AIC values, the Akaike weights are also used in models assessment: $w_i = [exp(-0.5 \cdot \Delta_i)/[\Sigma_{j=1}^{J} exp(-0.5 \cdot \Delta_j)]]$ [39] where $\Delta_i = AIC_i min(AIC)$, Δ_i = difference between the AIC of the best fitting model and that of the model i^{th}, $min(AIC)$ = minimum AIC value out of all models, j = the number of the models.
- Kubinyi function (FIT) [40], [41]: $FIT = [R^2 \cdot (n-k)]/[(n+(k+1)^2) \cdot (1-R^2)]$. The highest the FIT value the better the model is considered.

The diagnosis of a regression model when the dependent variable is continuous could be conducted by analyzing of residuals or rescaled residuals:

- Look to the largest and/or smallest experimental values ← detect if the values are in the plausible range. Also look to descriptive statistics value: mean, standard deviation, histogram.

TABLE I
ASSUMPTIONS OF LINEAR REGRESSION: EFFECT - IDENTIFICATION - METHODS

Assumption	**What is the effect?**	**How to detect it?**	**How to fix it?**
Normality	Unreliable coefficients and confidence intervals	Plot: normal probability plot Statistics: skewness & kurtosis [12] Test[c]: Kolmogorov-Smirnov [13], [14], Anderson-Darling [15], Chi-Square [16]; Shapiro-Wilks test [^{1}7] (n < 50)	Identify and withdrawn influential outliers (if any) - Grubs test [18]
Linearity	Estimations and predictions are in error	Plot ▪ observed vs estimated values ▪ residuals versus estimated values	Transformation (see Table II)
Independence	Important in models where time is important	Plot: autocorelation plot of residuals Test: Durbin-Watson [a] [19], [20]. If no autocorrelation exists in the sample under independence DW ~ 2	D-W < 1.00 → structural problem → reconsider the transformation (if any). Add more independent variables.
Homoscedasticity	Too wide or too narrow confidence intervals	Plot (pattern of errors): residuals vs predicted value Test: Breusch-Pagan[b] [21], Bartlett [22], modified Levene [23]	Use variance stabilizing transformation. Use Generalized Least Square.
Collinearity (independent variables)	The estimated coefficients are unstable [24]. Standard error of the estimated regression slope is inflated[d] [25]	▪ Correlation matrix: r ≥ 0.80 or 0.90 indicates collinearity [26] ▪ VIF ≥ 10 and/or T(tolerance) < 0.01 indicates the existence of collinearity [26]	Remove the variable(s) that is(are) correlated with others [25]. Principal component analysis of the descriptors [27]. Apply a ridge regression by adding a constant to the normal equation [28]. Be aware that collinearity is not bad all time.

[a] the errors are serially uncorrelated; WD ∈ [0, 4], DW = 2 → no autocorrelation; [b] the variance of the residuals is the same for all values of *Y*; [c] EasyFit program uses it to test the normality of *Y*; [d] The overall regression equation could be significant but none of the individual regression slope are significantly different from zero.

- Plot the independent variable(s) vs dependent variable.
- Plot the values associated to studentized residuals (s_i), leverage (h_i), Cook's (D_i) vs individual X_i values. The hat values ($0 \leq hi \leq 1$) are used to evaluate the leverage of observations in the dimensional space of independent variables (covariates). If the h_i value of a compound exceeds the threshold value ($2\cdot(k+1)/n$ for a regression model with intercept and $2\cdot k/n$ for a model without intercept, where k = number of X_i [42]) it is considered influential whenever if by its removal determine a significant improvement of the model. Cook's distance consider in its formula both residuals and hat matrix to identify influential compound(s) (threshold $D_i > 4/n$, where $D_i = 1/(k+1)\cdot s_i^2 \cdot [h_i/(1-h_i)]$ for the model with intercept and $D_i = 1/k \cdot s_i^2 \cdot [h_i/(1-h_i)]$ for the model without intercept, s_i = studentized residuals [43]).

Several parameters that can found their usefulness in diagnosis of a MLR are presented in Table III. Several parameters presented in Table III are also used by some authors as measures of model predictivity power (see for example MAE [44]).

B. Model Predictive Power

The ability to predict the activity/property of new compounds is of major importance in QSAR/QSPR analysis. Several parameters were proposed and are used to assess model predictivity power and are presented in Table IV.

The diagnosis of a linear regression model could be conducted using a series of statistical parameters calculated on contingency table [58] after transforma-

TABLE II
METHODS FOR DATA TRANSFORMATION

Transformation	**Applied to:**	**Appropriate when:**
'log' *Y' = logY*	▪ Stabilize the variance of *Y* ▪ Normalized the dependent variable ← positive skewed distribution of the residuals for *Y* ▪ Linearize the regression model	*Y* have positive values
'square root' *Y'* = √*Y*	▪ Stabilize the variance (the variance is proportional with the mean of *Y*)	*Y* has the Poisson distribution
'reciprocal' *Y' = 1/Y*	▪ Stabilize the variance	the variance is proportional to the fourth power of the mean of *Y*
'square' $Y' = Y^2$	▪ Stabilize the variance (the variance decrease with the mean of *Y*) ▪ Normalized the dependent variable ← negative skewed distribution of the residuals for *Y* ▪ Linearize the regression model ← the original relation with some independent variable is curvilinear downward (such as decrease of slope with the increase of independent variable)	
'arcsine' *Y' = asin*√*Y*	▪ Stabilize the variance	*Y* is a proportion or a percentage

tion of the observed and estimated/predicted logRBA as dichotomial variables using criteria for classification of compounds as active or inactive. The total fraction of compounds correctly classified (parameter called concordance / accuracy / non-error rate) is one parameter that could bring useful information in choosing which model to be applied.

II. Practical Considerations

Three data sets of endocrine disrupting chemicals with experimental values of relative binding affinity expressed in logarithmic scale (logRBA) [59] were used for exemplification. The investigated compounds could be classified according to their logRBA values as weak binders ($logRBA < -2.0$), moderate binders ($-2.0 = logRBA = 0$) and strong binders ($logRBA > 0$) [60].

The following descriptors were previously calculated on the investigated structures [59] and were used here to illustrate how linear regression analysis works: TIE = E-state topological parameter; TIC1 = Total information content index (neighbourhood symmetry of 1-order); ATS4m = Broto-Moreau autocorrelation of a topological structure - lag 4 / weighted by atomic masses; EEig02d = Eigenvalue 02 from edge adj. matrix weighted by dipole moments; E1s = 1st component accessibility directional WHIM index / weighted by atomic electrotopological states; and Dv = total accessibility index / weighted by atomic van der Waals volumes.

The first set was used to identify the model and comprised 132 compounds (training set; 1 withdrawn, 60 weak binders, 41 moderate binders and 30 strong binders). The second dataset was used to test the performances of the model (test set) and comprised 23 compounds (3 weak binders, 16 moderate binders and 4 strong binders). The third dataset was used as external validation set and consists of 9 compounds (4 weak binders and 5 moderate binders).

A. MLR in Training Sets

The first step in the linear regression analysis was to investigate the distribution of logRBA in training set. One out of three tests rejected the null hypothesis of normality (Chi-Square statistics = 14.862, p-value = 0.03781). No outlier had been identified when the Grubbs test was applied but there was one compound with studentized residuals higher than 3 standard deviations. The experimental data in training test proved not normal distributed according just with the Chi-Square test (see Table V), the normality test that is known to be affected by the presence of outlier(s) [12], even if in this example no outlier has been identified. The normality was not achieved even by withdrawing that compounds but the correlation coefficient increased from 0.810 to 0.837. The studentized residuals, hat matrix and Cook's distance values were plotted against logRBA to identify how data were distributed (Figure 1). Three models obtained on the same datasets were investigated:

TABLE III
STATISTICAL PARAMETERS FOR DIAGNOSIS OF MLR

Parameter (Abbreviation)	Formula [ref]	Remarks
Residual Mean Square (RMS) - Error variance	$RMS = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}$	RMS: the smaller the better 0 < RMS < ∞
Average Prediction Variance (APV)	$APV = \frac{RMS}{n}\cdot(n+k)$ [45]	The smaller the better
Total Squared Error (TSE)	$TSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\hat{\sigma}^2} + 2\cdot k - n$ [46] $TSE = \frac{SSE}{MSE} - (n - 2\cdot k) + 2$ [33]	The smaller the better TSE > (k+1) → bias due to incompletely specified model TSE< (k+1) → the model is over specified (contains too many variables)
Average Prediction Mean Squared Error (APMSE)	$APMSE = \frac{RMS}{n-k-1}$ [47]	The smaller the better
Mean Absolute Error (MAE) - Measures the average magnitude of the errors; could be also used to compare two models	$MAE = \frac{\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert}{n}$	MAE = 0 → perfect accuracy 0 < MAE < ∞
Root Mean Square Error (RMSE): - Measures the average magnitude of the error	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	RMSE > MAE → variation in the errors exists 0 < RMSE < ∞
Mean Absolute Percentage Error (MAPE) - Measure of accuracy expressed as percentage	$MAPE = \frac{\sum_{i=1}^{n} \lvert (y_i - \hat{y}_i)/y_i \rvert}{n}$ [48], [49]	MAPE ~ 0 → perfect fit
Standard Error of Prediction (SEP)	$SEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-1}}$	The smaller the better
Relative Error of Prediction (REP%)	$REP(\%) = \frac{100}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$	The smaller the better

n = sample size; k = number of independent variables in the model; $\bar{y}$ = the mean of estimated/predicted activity/property; $\hat{y}_i$ = predicted value of the ith compound in the sample; y_i = observed/measured activity/property of ith compound; SSE = sum of squared errors; MSE = mean of squared errors

full-model (the model comprised all compounds assigned to training test), Di-model (the model comprised just the compounds that did not exceeded the imposed Cooks distance threshold), and hi-model (the model comprised just the compounds that did not exceeded the imposed hat matrix threshold).

The Cook's distance and hat matrix approaches were applied to withdrawn compounds of the training sample until two criteria were accomplished: logRBA proved normal distributed and withdrawing the compound(s) did not led to an improvement in determination coefficient. Both models proved smaller RMSE and RMSEP values. The characteristics of all investigated models are presented in Table V.

The analysis of the models (Table V) revealed that none model proved collinearity (the highest correlation coefficient did not exceeded 0.8 and VIF values are less than 10). The Di-model is twice better in terms of internal validity when the $|R^2 - Q^2|$ difference is evaluated compared to h_i-model and three times better compared to the full-model. The Mallows C_p-statistic did not found its applicability in our example because the same descriptors are used in all models. The smallest values of information criteria parameter were systemat-

TABLE IV
STATISTICS FOR ASSESSMENT THE PREDICTIVE POWER OF MLR

Parameter (abbr.)	**Formula [ref]**	**Remarks**
Predictive Squared Correlation Coefficient in Training Set (Q_{F1}^2)	$Q_{F_1}^2 = 1 - \frac{\sum_{i=1}^{n_{TS}}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TS}}(y_i - \bar{y}_{TR})^2}$ [50]	Prediction is considered accurate if the predictive power of the model is > 0.6 [51]
Predictive Squared Correlation Coefficient in Test Set (Q_{F2}^2)	$Q_{F_2}^2 = 1 - \frac{\sum_{i=1}^{n_{TS}}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TS}}(y_i - \bar{y}_{TS})^2}$ [52]	
External Predictive Ability (Q_{F3}^2)	$Q_{F_3}^2 = 1 - \frac{\sum_{i=1}^{n_{TS}}(\hat{y}_i - y_i)^2 / n_{TS}}{\sum_{i=1}^{n_{TS}}(y_i - \bar{y}_{TR})^2 / n_{TR}}$ [53]	
r_m^2 metrics	$r_m^2 = r^2 \cdot \left\lfloor 1 - \sqrt{r^2 - r_0^2} \right\rfloor$ [44], [54] $\Delta r_m^2 = \mid r_m^2 - r_m^{'2} \mid$	Values higher than 0.5 indicate an acceptable model [44] , [54] Δr_m^2 indicate an acceptable model
Concordance Correlation Coefficient (CCC)	$CCC = \frac{2 \cdot \sum_{i=1}^{n}(y_i - \bar{y}) \cdot (\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + n \cdot (\bar{y} - \bar{\hat{y}})^2}$ [55]	Strength of agreement between observed and predicted values [56]: > 0.99 almost perfect; [0.95; 0.99) substantial; [0.90; 0.95) moderate; < 0.90 poor
Predictive Power (PP): Fisher's approach	$t = \frac{\overline{res_{TS}} - 0}{stdev(res_{TS}) / \sqrt{n_{TS}}}$ [57] *p = TDIST(abs(t),n_{TS}-1,1)*	Evaluate if the mean of residual is statistically different by the expected value (0)

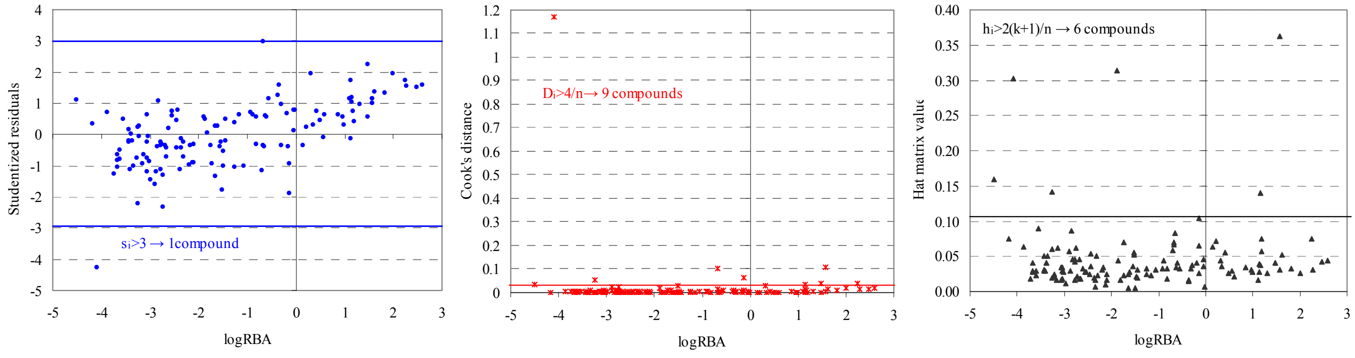
n = sample size; v = number of independent variables in the model; $\overline{y}$ = the mean of observed/measured activity/property; $\overline{\hat{y}}$ = the mean of estimated/predicted activity/property; $\hat{y}_i$ = predicted value of the ith compound in the sample; y_i = observed/measured activity/property of ith compound; $\overline{res}$ = mean of residuals; stdev = standard deviation; TR = training set; TS = test set; r^2_m = a metric calculated using observed (y-axis) and estimated/predicted (x-axis)values; r'^2_m = a metric calculated using observed (x-axis) and estimated/predicted (y-axis)values; r^2_0 = determination coefficient calculating by forcing the origin of axis; Δr^2_m = absolute difference between r^2_m and r'^2_m; EXT = external set; abs = absolute value

ically obtained by D_i-model which was follow by h_i-model while the full-model systematically obtained the highest values (see Table V).

The concordance correlation coefficient for training sets had values closed to the correlation coefficients and for all models were higher than 0.80 (see Table 5).

Looking to the weights of Akaike's information criteria, which can be interpreted as probability that a certain model is the best model, it could not be identify any model with robust inference (none of the model had the values of weights higher than 0.9 [61]). The D_i-model had the weights around 0.37 that is far away from 0.90 but are a little higher than those obtained by the full model where the weights are around 0.30 or by those obtained by the h_i-model which are around 0.32. Recall that the D_i-model could be considered the preferred model and from the inspection of the Akaike weights in Table V, this model is 1.2 ($w_i - AIC_{R2}$) to 1.4 ($w_i - AIC_c$) times more likely in terms of Kullback-Leible discrepancy, a measure of distance between the probability generated by the model and reality [62], compared with h_i-model.

Significant differences between models could also been observed if the BIC and HQC parameters are analyzed; the smallest value of BIC was obtained by D_i-model while the smallest value of HQC was obtained by h_i-model. The plots of residuals versus predicted values for the investigated models are presented in Figure 2. The analyses of residuals allow to identify if the assumptions of the regression appear to have been met or not (specifically linearity and homoscedascity) - the residual plot look like a horizontal band. Thus, according



a) b) c)

Fig. 1. Studentized residuals (a), Cook's distance (b) and hat matrix values (c) versus logRBA in model with all compounds in training set (n=132)

TABLE V

MLR in Training Sets: Models Characteristics

Statistical parameter	**Full-model (n=132)**	**D_i-model (n=115)[a]**	**h_i-model (n=123)[b]**
Normality tests: KS-AD-CS	0.116* - 2.409* - 14.862**	0.124* - 2.432* - 12.613*	0.120* - 2.428* - 12.083*
Durbin-Watson	1.275	1.292	1.263
Collinearity: highest R higher VIF & lower T	0.7700 TIE: 3.367& 0.297	0.7889 ATS4m: 4.082&0.245	0.7752 ATS4m: 4.516&0.221
R^2	0.6559	0.7797	0.6928
R^2_{adj}	0.6394	0.7675	0.6769
RMSE	1.0701	0.8293	0.9977
F-value (p-value)	39.711 ($9.89\cdot10^{-27}$)	63.721 ($3.12\cdot10^{-33}$)	43.59 ($1.62\cdot10^{-27}$)
Q^2	0.5832	0.7543	0.6497
RMSEP	1.1827	0.8764	1.0668
F_{loo}-value (p-value)	28.74 ($9.49\cdot10^{-22}$)	55.17 ($1.85\cdot10^{-31}$)	($1.62\cdot10^{-27}$)
$\|R^2-Q^2\|$	0.0727	0.0254	0.0431
Concordance Correlation Coefficient (CCC)	0.8108 [0.7476 to 0.8595]	0.8762 [0.8278 to 0.9117]	0.8185 [0.7545 to 0.8671]
r^2_m (Δr^2_m)	0.6071 (0.1324)	0.7797 (0.1278)	0.6921 (0.1586)
C_p-statistic	7.00	7.00	7.00
AIC (w_i-AIC)	18.9639 (0.2856)	18.3078 (0.3965)	18.7490 (0.3180)
AIC_{R2} (w_i- AIC_{R2})	8.0504 (0.3137)	7.7421 (0.3659)	8.0077 (0.3204)
AIC_c (w_i- AIC_c)	1.2657 (0.2990)	0.7766 (0.3819)	1.1358 (0.3191)
BIC	52.0750	9.8317†	33.1255
HQC	26.2887	34.7113†	7.8043
FIT	1.3058	2.3097	1.5076

* p ≥0.05; ** p = 0.0378; † = absolute values; KS = Kolmogorow-Smirnov; AD = Anderson Darling; CS = Chi-Square; R = correlation coefficient; VIF = Variance Inflation Factor; T = tolerance; R^2 = determination coefficient; R^2_{adj} = adjusted determination coefficient; RMSE = root mean square error; F-value = Fisher's statistics; Q^2 = determination coefficient in cross-validation by the leave-one-out analysis; RMSEP = root mean square error in prediction; CCC = concordance correlation coefficient [95% confidence interval]; Cp-statistic = Mallows' statistic; AIC = Akaike's information criterion; AIC_{R2} = AIC based on the determination coefficient; AIC_c = AIC corrected by McQuarrie and Tsai; BIC = Bayesian Information Criterion; HQC = Hannan-Quinn Criterion; FIT = Kubinyi's function;

[a] 56 weak binders, 35 moderate binders, and 24 strong binders; withdrawn (16 compounds): 4 weak binders, 6 moderate binders and 6 strong binders;

[b] 57 weak binders, 38 moderate binders, and 28 strong binders; withdrawn (8 compounds): 3 weak binders, 3 moderate binders and 2 strong binders;

to the pattern of the residuals [63], the most appropriate model is the D_i-model since the distribution indicates a homoscedastic model. Furthermore, both full-model and h_i-model showed evidence of heteroscedascity, the error in estimating logRBA increasing as the value of logRBA increase. However, both these models could be accepted because none of them showed the presence of systematic errors or inadequacy [63]. If assumption of linearity and/or of homoscedascity is violated, the residual plots show an increasing and narrow pattern if systematic error exists or depict a Gaussian trend when the model is inadequate [64]. Other proposed plot methods, such as linear residual plots, show to be useful in identification of non-linearity while squared residual plots proved utility in detection of non-constant variances [65].

The normal probability plots (right graphical representations in Figure 2) can be used to verify normality assumption of the residuals. Figure 2 showed that the hi-model fit better a straight line compared to both full-model and D_i-model.

The results obtained on our data associated to the statistical parameters useful in model diagnosis introduced in Table III are presented in Table VI. The total square error is the single parameter that has the same value for all models and in all cases is equal to 7 (obtained by adding 1 to the number of descriptors in the model 6 in our example), indicating that none of the models were not over-specified or did not contain bias due to incompletely specified model. The classification of our models based on parameters presented in Table VI led to the classification obtained according to the parameters presented in Table V: D_i-model, h_i-model, and full model.

Several parameters were used to assess the predictive power of the models and their results are presented in Table VII. The analysis of results presented in Table VII revealed the followings:

- External predictive ability parameter (Q^2_{F3}) [53] systematically took negative values for both external and withdrawn sets. At least for the external set, this result could be explained by the distribution of logRBA values (min=-3.3, max=-0.6) compared to training (min=-4.5, max=2.6) and test (min=-2.51, max=1.41) sets. It could be also of interest to analyze how different are the compounds containing in external and withdrawn data sets compared to the compounds from training set (in terms of similarity of their structure for example).
- D_i-model achieve the criterion of exceeding 0.6 [52] in just one of 6 possible case while the h_i-model reach this criterion in four out of 6 cases. The h_i-model accomplished more frequently the criteria of having values higher than 0.6 while the full-model did not accomplished at all this criterion. Thus, it seems that the compounds in test and external sets are uniformly distributed over the range of training set at least in h_i-model, in view of the fact that otherwise the Q^2_{F1} and the Q^2_{F2} suffer from drawbacks [66].
- The concordance correlation coefficients obtained values higher than 0.70 in test sets. The abilities of prediction the external sets proved smaller than 0.5 for all investigated models but had values higher than 0.50 (D_i-model and h_i-model) when the withdrawn set is investigated.
- The residual of the models proved significantly different by zero in test set for full-model and D_i-model and in external set for all models. Both D_i- and h_i-models proved to have residual not significantly different by zero in samples that contain the withdrawn compounds. According to this criterion, just hi-model proved prediction power.

The r^2_m metric and associated Δr^2_m obtained in test sets were as follows: 0.3726 (0.1743) for full model, 0.3134 (0.1796) for D_i-model, and 0.5248 (0.1494) for h_i-model. These metrics showed that the h_i-model is acceptable model. The r^2_m is a parameter computed by forcing the regression through origin [54] with certain applicability and limitations (fails to detect the differences between experimental and predicted values when the slopes of the regression line are not near to 1) [67]. The values of these metrics were smaller than the determination coefficient in all investigated models and the highest value was observed in D_i-model when training (see Table V) set was investigated but acceptable values were obtained just by the h_i-model when the test set was investigated ($r^2_m > 0.5$ and $\Delta r^2_m < 0.2$).

The classification of the models according to results presented Table VII is as follows: h_i-model, D_i-model, and full-model.

One remark about the parameters used to assess the predictive power, namely Q^2_{F1}, Q^2_{F2} and Q^2_{F3}, can be made. Even the symbols contain ”square”, these parameters could take both positive and negative values according to their formula (see Table IV). A simulation study of these parameters needs to be done to identify their possible values as well as their proper interpretation.

The best way to see the abilities of a MLR model is to plot the measured values against the estimated / predicted values to visualize how well each model works (see Figure 3). With one exception, represented by h_i-model in external set (p-value = 0.0632), all other correlation coefficients proved statistically significant ($p < 0.04$).

The analysis of models presented in Figure 3 revealed the followings:

- The distribution of compounds in training set is narrower in D_i-model compared to both full-model and h_i-model.

- D_i-model obtained higher determination coefficients in training and external sets while the h_i-model obtained the higher determination coefficients in training and withdrawn sets.
- The h_i-model in more stable compared to D_i-model if the difference in determination between training and test set is concerned.
- Both D_i-model and h_i-model performed better in training and test sets compared to full-model.

Whenever applicable, the accuracy of a model will show its ability in correct classification of compounds. The overall accuracy as well as the accuracy on each class (weak binder, moderate binder and strong binder) were computed and the obtained results are presented in Figure 4.

The analysis of Figure 4 revealed the followings:

- The accuracy of all three models was identical for strong binders in test set (75%) and weak binders in external set (25%). Overall, out of 16 possibilities, all models (full-model, D_i-model, and h_i-model) proved highest accuracy in almost 38% of cases.
- Full-model proved highest overall accuracy in both test and external sets, and highest accuracy for moderate binders in test and external sets.
- D_i-model proved highest overall accuracy in training set, highest accuracy for strong binders in training set, highest accuracy for weak binders in training set, and highest accuracy of moderate binders in training set.
- h_i-model proved highest overall accuracy, as well as higher accuracy for weak binders, moderate binders and strong binders for withdrawn compounds.
- No model proved abilities in correct classification of weak binders in test set or of strong binders in external set.

Regarding the accuracy of investigated models it is impossible to classify them since their performances are generally the same (38%). It could be observed that models had abilities to accurately identify the compounds on average of two sets out of three or four. The absence of accurate classification of weak binders in test set and strong binders in externals set could be explained by differences in the chemical structure or measured logRBA of compounds included in these sets.

III. SUMMARY AND FURHER WORK

Choosing a proper linear model is crucial in QSAR analysis because a model able to predict accurately the activity of interest of new chemical compounds is desired under the hypothesis that changes in molecular structure directly reflect in the compound activity/property. Input data and data preparation for regression analysis are of great importance but these subjects were beyond the aim of the present manuscript.

Linear regression analyses identify in QSAR analysis the linearity between compound's activity and calculated descriptors based on chemical structure. Regression analysis answer to the following questions: ***Does the biological activity depend on structural information?*** If so, ***the nature of the relationship is linear?*** If yes, ***how good is the model in prediction of the biological activity of new compounds?***

In this manuscript, some rules had been presented: ① test the assumption of linear regression (normality, linearity, independence, homoscedascity, and/or collinearity); ② construct the model(s) if assumptions are accomplished - analyze the data (choose the best performing model); ③ assess and diagnose the alternative models - analyze the MLR; ④ decide which model fit best to your objectives.

Following these steps in linear regression analysis certainly led to a performing estimation model but the prediction power of the model will always depend on the structure of compounds and their biological activity on which the model is used to predict; in other words, will be dependent by similarity in terms of structure and activity.

Researches on linear regression analysis are of general interest since MLR found its applicability in many research fields. The classical approach implemented in available dedicated software deal with maximization of correlation coefficient. Maximization of the observed probability under assumption of random error affecting all variables in the model is an ongoing research and will be reported somewhere else. It is known that the classical method is exposed to type I errors (to accept a regression model obtained by maximization of determination correlation even if it does not exist) while this new approach does not because it maximize just the observation chance having as hypothesis that the errors between observed value and value obtained by the model is random and depend just by the observed/measured value (therefore being symmetric relative to its arithmetic mean).

ACKNOWLEDGMENT

The authors are grateful to the organizers of the BIOMATH 2013 for the opportunity to present our results.

References

[1] L. P. Hammett, ”The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives,” J. Am. Chem. Soc., vol. 59, no. 1, 1937, pp. 96-103. http://dx.doi.org/10.1063/1.1749914

[2] P. Gramatica, ”A short history of QSAR evolution,” [online] [Accessed January 26, 2012]. Available from: http://qsarworld.com/Temp_Fileupload/Shorthistoryofqsar.pdf.

[3] A. M. Johnson and G.M. Maggiora, ”Concepts and Applications of Molecular Similarity”, New York: John Willey & Sons, 1990.

[4] T. Arodź and A.Z. Dudek, ”Multivariate modeling and analysis in drug discovery,” Curr. Comput. Aided Drug Des., vol. 3, no. 4, 2007, pp. 240-247. http://dx.doi.org/10.2174/157340907782799381

[5] J. Galvez, M. Galvez-Llompart, R. Zanni, and R. Garcia-Domenech, ”Advances in the molecular modeling and quantitative structure-activity relationship-based design for antihistamines,” Expert Opin. Drug Discov., vol. 8, no. 3, 2013, pp. 305-317. http://dx.doi.org/10.1517/17460441.2013.748745

[6] M. P. Gleeson, S. Modi, A. Bender, R. L. Marchese Robinson, J. Kirchmair, M. Promkatkaew, S. Hannongbua, and R. C. Glen, ”The challenges involved in modeling toxicity data in silico: A review,” Curr. Pharm. Des., vol. 8, no. 9, 2012, pp. 1266-1291. http://dx.doi.org/10.2174/138161212799436359

[7] S. Kar, O. Deeb, and K. Roy, ”Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor,” Ecotoxicol. Environ. Saf., vol. 82, 2012, pp. 85-95. http://dx.doi.org/10.1016/j.ecoenv.2012.05.013

[8] M. Goodarzi, B. Dejaegher, and Y. V. Heyden, ”Feature selection methods in QSAR studies,” J. AOAC Int., vol. 95, no. 3, pp. 636-651, 2012. http://dx.doi.org/10.5740/jaoacint.SGE_Goodarzi

[9] D. M. Hawkins, ”The problem of overfitting,” J. Chem. Inf. Comput. Sci., vol. 44, no. 1, 2004, pp. 1-12. http://dx.doi.org/10.1021/ci0342472

[10] S. Chatterjee and A. S. Hadi, ”Regression Analysis by Example,” New Jersey: John Wiley & Sons, 2006.

[11] G. U. Yule, ”On the significance of Bravais formulae for regression in the case of skew correlation,” Proc. R. Soc. Lond., vol. 60, 1897, pp. 477-489.

[12] L. Jäntschi and S. D. Bolboacă, ”Distribution Fitting 2. Pearson -Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics,” Bulletin UASVM Horticulture, vol. 66, no. 2, 2009, pp. 691-697.

[13] A. Kolmogorov, ”Confidence Limits for an Unknown Distribution Function,” Ann. Math. Stat., vol. 12, no. 4, 1941, pp. 461-463. http://dx.doi.org/10.1214/aoms/1177731684

[14] N. V. Smirnov, ”Tables for estimating the goodness of fit of empirical distributions,” Ann. Math. Stat., vol. 19, no. 2, 1948, pp. 279-281. http://dx.doi.org/10.1214/aoms/1177730256

[15] T. W. Anderson and D. A. Darling, ”Asymptotic theory of certain ”goodness-of-fit” criteria based on stochastic processes,” Ann. Math. Stat., vol. 23, no. 2, 1952, pp. 193-212. http://dx.doi.org/10.1214/aoms/1177729437

[16] K. Pearson, ”On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” Philos Mag, vol. 50, 1900, pp. 157-175.

[17] A. A. Shapiro and M. B. Wilks, ”An analysis of variance test for normality (complete sample),” Biometrika, vol. 52, no. 3/4, 1965, pp. 591-611. http://dx.doi.org/10.2307/2333709

[18] F. Grubbs, ”Procedures for Detecting Outlying Observations in Samples,” Technometrics, vol. 11, no. 1, 1969, pp. 1-21. http://dx.doi.org/10.1080/00401706.1969.10490657

[19] J. Durbin and G. S. Watson, ”Testing for Serial Correlation in Least Squares Regression. I,” Biometrika, vol. 37, no. 3/4, 1950, pp. 409-428. http://dx.doi.org/10.2307/2332391

[20] J. Durbin and G. S. Watson, ”Testing for Serial Correlation in Least Squares Regression. II,” Biometrika, vol. 38, no. 1/2, 1951, pp. 159-177. http://dx.doi.org/10.2307/2332325

[21] T. S. Breusch and A. R. Pagan,. ”Simple test for heteroscedasticity and random coefficient variation,” Econometrica, vol. 47, no. 5, 1979, pp. 1287-1294. http://dx.doi.org/10.2307/1911963

[22] M. S. Bartlett,. ”Properties of sufficiency and statistical tests,” Proc. Roy. Stat. Soc. A, vol. 160, 1937, pp. 268-282. http://dx.doi.org/10.1098/rspa.1937.0109

[23] W. G. S. Hines and R. J. O. Hines, ”Increased power with modified forms of the Levene (med) test for heterogeneity of variance,” Biometrics, vol. 56, no. 2, 2000, pp. 451-454. http://dx.doi.org/10.1111/j.0006-341X.2000.00451.x

[24] T. E. Philippi, ”Design and Analysis of Ecological Experiments. Multiple regression: Herbivory,” New York: Chapman & Hall, 1993.

[25] G. P. Quinn and M. J. Keough, ”Experimental Design and Data Analysis for Biologists, 6. Multiple Regression and Correlation,” UK: Cambridge University Press, 2002, pp. 124-174. http://dx.doi.org/10.1017/CBO9780511806384

[26] R. H. Myers, ”Classical and Modern Regression With Applications,” 2nd edition, PWS-Kent, 1990.

[27] J. O. Rawlings, S. G. Pantula, and D. A. Dickey, ”Applied Regression Analysis; A Research Tool,”, 2nd edition, New York: Springer-Verlag, 1998. http://dx.doi.org/10.1007/b98890

[28] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, ”Applied Linear Statistical Models,” 4th edition, Illinois: Irwin, 1996.

[29] D. G. Kleinboum, L .L. Kupper, A. Nizam, and K. E. Muller, ”Applied Regression Analysis and Other Multivariate Methods. Chapter 14. Regression Diagnostics,” Forth edition, Canada: Duxbury, 2008, pp. 287-348.

[30] A. Tropsha, ”Best practices for QSAR model development, validation, and exploitation,” Mol. Inf., vol. 29, no. 6-7, 2010, pp. 476-488. http://dx.doi.org/10.1002/minf.201000061

[31] C. L. Mallows, ”Some comments on Cp,” Technometrics, vol. 15, no. 4, 1973, pp. 661-675.

[32] C. L. Mallows, ”More comments on Cp,” Technometrics, vol. 37, no. 4, 1995, pp. 362-372.

[33] C. L. Mallows, ”Cp and prediction with many regressors: comments on Mallows,” Technometrics, vol. 39, no. 1, 1997, pp. 115-116.

[34] H. Akaike, ”Fitting Autoregressive Models for Prediction,” Ann. I. Stat. Math., vol. 21, no. 1, 1969, pp. 243-247. http://dx.doi.org/10.1007/BF02532251

[35] A. D. R. McQuarrie and C.-L. Tsai, ”Regression and time series model selection in small samples,” Singapore: World Scientific Pub Co Inc, 1998.

[36] G. Schwarz, ”Estimating the dimension of a Model,” Ann. Stat., vol. 6, no. 2, 1978, pp. 461-464. http://dx.doi.org/10.1214/aos/1176344136

[37] T. Amemiya, ”Qualitative response models: A survey,” J. Econ. Lit., vol. 19, no. 4, 1981, pp. 1483-1536.

[38] E. J. Hannan and B. G. Quinn, ”The determination of the Order of an Autoregression,” J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 41, no. 2, 1979, pp. 190-195.

[39] S. T. Buckland, K. P. Burnham, and N. H. Augustin, ”Model selection: An integral part of inference,” Biometrics, vol. 53, no. 2, 1997, pp. 603-618.

[40] H. Kubinyi, ”Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution,” QSAR Comb. Sci., vol. 13, no. 4, 1994, pp. 393-401. http://dx.doi.org/10.1002/qsar.19940130403

[41] H. Kubinyi, ”Variable Selection in QSAR Studies. I. An Evolutionary Algorithm,” QSAR Comb. Sci, vol. 13, no. 13, 1994, pp. 285-294. http://dx.doi.org/10.1002/qsar.19940130306

[42] D. C. Hoaglin and R. E. Welsch, ”The hat matrix in regression and ANOVA,” Am. Stat., vol. 32, no. 1, 1978, pp. 17-22. http://dx.doi.org/10.1080/00031305.1978.10479237

[43] K. A. Bollen and R. Jackman, ”Regression diagnostics: An expository treatment of outliers and influential cases,” In: Modern Methods of Data Analysis, Fox, J.; Scott, and J. Long (Eds.), Sage: Newbury Park, 1990, pp. 257-291.

[44] N. Chirico and P. Gramatica, ”Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection,” J. Chem. Inf. Model., vol. 52, no. 8, 2012, pp. 2044-2058. http://dx.doi.org/10.1021/ci300084j

[45] C. L. Mallows, ”Choosing a subset regression,” Unpublished report, Bell Telephone Laboratories.

[46] J. W. Gorman and R. J. Toman, ”Selection of variables for fitting equations to data,” Technometrics, vol. 8, no. 1, 1966, pp. 27-51. http://dx.doi.org/10.1080/00401706.1966.10490322

[47] J. W. Tukey, ”Discussion,” J. R. Statisti. Soc., vol. 29, 1967, pp. 47-48.

[48] J. S. Armstrong, ”Long-range Forecasting: From Crystal Ball to Computer,” United States of America: John Wiley & Sons, 1978.

[49] B. E. Flores, ”A pragmatic view of accuracy measurement in forecasting,” Omega (Oxford), vol. 14, no. 2, 1986, pp. 93-98. http://dx.doi.org/10.1016/0305-0483(86)90013-7

[50] L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland, and D. M. Sheehan, ”QSAR Models Using a Large Diverse Set of Estrogens,” J. Chem. Inf. Comput. Sci., vol. 41, no. 1, 2001, pp. 186-195. http://dx.doi.org/10.1021/ci000066d

[51] A. Golbraikh and A. Tropsha, ”Beware of q2!”, J. Mol. Graphics Mod., vol. 20, no. 4, 2002, pp. 269-276. http://dx.doi.org/10.1016/S1093-3263(01)00123-1

[52] G. Schüürmann, R. U. Ebert, J. Chen, B. Wang, and R. Kühne, ”External Validation and Prediction Employing the Predictive Squared Correlation Coefficient Test Set Activity Mean vs Training Set Activity Mean,” J. Chem. Inf. Model., vol. 48, no. 11, 2008, pp. 2140-2145. http://dx.doi.org/10.1021/ci800253u

[53] V. Consonni, D. Ballabio, and R. Todeschini, ”Comments on the Definition of the Q2 Parameter for QSAR Validation,” J. Chem. Inf. Model., vol. 49, no. 7, 2009, pp. 1669-1678. http://dx.doi.org/10.1021/ci900115y

[54] P. K. Ojha, I. Mitra, R. N. Das, and K. Roy, ”Further exploring r2m metrics for validation of QSPR models,” Chemom. Intell. Lab. Syst., vol. 107, no. 1, 2011, pp. 194-205. http://dx.doi.org/10.1016/j.chemolab.2011.03.011

[55] L. I. Lin, ”A concordance correlation coefficient to evaluate reproducibility,” Biometrics, vol. 45, 1989, pp. 255-268.

[56] G. B. McBride, ”A proposal for strength-of-agreement criteria for Lin’s Concordance Correlation Coefficient, ” NIWA Client Report: HAM2005-062, 2005, [online] [accs. March 14, 2013]. http://medcalc.org/download/pdf/McBride2005.pdf

[57] R. A. Fisher, ”The goodness of fit of regression formulae, and the distribution of regression coefficients,” J. Royal Statist. Soc., vol. 85, no. 4, 1922, pp. 597-612.

[58] S. D. Bolboacă and L. Jäntschi, ”Predictivity Approach for Quantitative Structure-Property Models. Application for Blood-Brain Barrier Permeation of Diverse Drug-Like Compounds,” Int. J. Mol. Sci., vol. 12, no. 7, 2011, pp. 4348-4364. http://dx.doi.org/10.3390/ijms12074348

[59] J. Li and P. Gramatica, ”The importance of molecular structures, endpoints’ values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders,” Mol. Divers., vol. 14, no. 4, 2010, pp. 687-696. http://dx.doi.org/10.1007/s11030-009-9212-2

[60] R. M. Blair, H. Fang, W. S. Branham, B. S. Hass, S. L. Dial, C. L. Moland, W. Tong, L. Shi, R. Perkins, and D. M. Sheehan, ”The Estrogen Receptor Relative Binding Affinities of 188 Natural and Xenochemicals: Structural Diversity of Ligands,” Toxicol Sci., vol. 54, no. 1, 2000, pp. 138-153. http://dx.doi.org/10.1093/toxsci/54.1.138

[61] K. P. Burnham and D. R. Anderson, ”Model selection and multimodel inference: A practical information-theoretic approach,” New York: Springer-Verlag, 2002.

[62] K. P. Burnham and D. R. Anderson, ”Kullback-Leibler information as a basis for strong inference in ecological studies,” Wildlife Res., vol. 28, no. 2, 2001, pp. 111-119. http://dx.doi.org/10.1071/WR99107

[63] J. W. Osborne and E. Waters, ”Four Assumptions Of Multiple Regression That Researchers Should Always Test,” Practical Assessment, Research, and Evaluation, vol. 8, 2002, [online] [Accessed February 26, 2013]. Available from: http://PAREonline.net/getvn.asp?v=8&n=2

[64] N. R. Draper and H. Smith, ”Applied Regression Analysis,” (2nd ed.). New York: Wiley, 1981.

[65] C.-L. Tsai, Z. Cai, and X. Wu, ”The Examination of Residual Plots,” Stat. Sin., vol. 8, 1998, pp. 445-465.

[66] V. Consonni, D. Ballabio, and R. Todeschini, ”Evaluation of model predictive ability by external validation techniques,” J. Chemom., vol. 24, no. 3-4, 2010, pp. 194-201. http://dx.doi.org/10.1002/cem.1290

[67] N. Chirico and P. Gramatica, ”Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient,” J. Chem. Inf. Model., vol. 51, no. 9, 2011, pp. 2320-2335. http://dx.doi.org/10.1021/ci200211n