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

QSARs-LRM Modelling and Validation Strategies by Example

Sorana D. BOLBOACĂ[*]
[*]"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur, 400349 Cluj-Napoca. E-mail: sbolboaca@umfcluj.ro

Lorentz Jäntschi[**]
[**]Technical University of Cluj-Napoca, Department of Physics and Chemistry, 103-105 Muncii Bvd., 400641 Cluj-Napoca, Romania. E-mail: lorentz.jantschi@gmail.com

***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 predict the nature of the relationships 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 compounds*. 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. Brief History of Linear Regression

Linear regression analysis is used in life science researches *to describe* the strength of the association between outcome and factors of interest, *to adjust* data for covariates or co-founders, *to identify predictors* (factors that affect the outcome) and/or *to predict the outcome* [1].

It could be considered that Sir Francis Galton provided the initial inspiration that led to correlation and regression. The fundamentals of correlation were discussed by Bravais [2] who presented the correlation of two and three variables. Galton improved notation as "Galton function" of correlation coefficient (r); this function could be found in Bravais' work but not as a single symbol. Edgeworth indicated in 1892 how to extend the Bravais' method to higher degree of correlation [3] and expressed his results in terms of "Galton's function".

Galton used regression to understand heredity and suggested a slope of 0.33 that showed the relationships between extremely large or small mother peas seed and their less extreme daughter seeds [4,5]. Galton seems to build the regression analysis based on the work of Adolphe Quetelet who is known to be the first scientists that applied in a systematically way a statistical methods to human [6]. Furthermore, Quetelet showed normal distributions in diverse aggregated data [6].

Galton was able to fit all data in a single line and he abbreviated the slope of this line as "r" [7], later this symbol being use to stand for correlation coefficient [8]. Pearson demonstrated in 1896 that optimum values of slope and correlation coefficient could be calculated from the product-moment [8]. On the same time, George Yule refined regression analysis [9], [10], [11], solving his regression problem by minimizing the sum of squares error [9,10], method that was presented for the first time by Legendre in 1805 [12].

II. 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 [13]. 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 [14]. This assumption could be summarized as "*similar compounds have similar properties*" [15].

The two main fields were linear regression analysis found its applicability are drug discovery [16], [17] and toxicology prediction [18], [19]. 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) [20].

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):

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

where $\hat{Y}$ = estimated activity/property; b_0 = intercept; b_i = coefficient of the i^th variable ($1 \leq i \leq k$, $5 \times k \leq n$ [21], where k = number of descriptors (independent variables) in the model, n = number of observations in the sample) and represents the slope of the straight-line relationship between activity/property and descriptor(s), the amount Y changes when X increased or decreased by 1 unit (b_0 and b_1 estimate the population parameters β_0 and β_i), and ε = random error. The identified values of b_0 and b_i are calculated to minimize the squared error for all n observations. However, the model could looks different if the values are obtained under other hypotheses like: maximization of r-value, maximization of F-value, minimization of p-value associated to the F-value, maximization of t-values of bi or minimization of their p-values.

A. *Linear Regression Assumptions*

The main assumptions of linear regression (Table 1) 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*" [10] a series of transformation could be used to reach normal distribution [35] (see Table 2).

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 [22] Test[c]: Kolmogorov-Smirnov [23], [24], Anderson-Darling [25], Chi-Squared [26]; Shapiro-Wilks test [27] ($n < 50$)	Identify and withdrawn the outliers (if any) - Grubs test [28]
Linearity	Estimations and predictions are in error	Plot ▪ observed vs estimated values ▪ residuals versus estimated values	Transformation (see Table 2)
Independence	Important in models where time is important	Plot: autocorelation plot of residuals Test: Durbin-Watson [a] [29], [30]. If no autocorrelation exists in the sample 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] [31], Bartlett [32], Levene [33]	Use different variables. Use Generalized Least Square
Collinearity (independent variables)	Predictors are related to each other	▪ Correlation matrix: $r \geq 0.80$ or 0.90 indicates collinearity [34] ▪ $VIF \geq 10$ and/or T(tolerance) < 0.01 indicates the existence of collinearity [34]	Remove the variable that is correlated with others Be aware that collinearity is not bad all time

[a] the errors are serially uncorrelated; $WD \in [0, 4]$, 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*;

TABLE II. METHODS FOR DATA TRANSFORMATION

Transformation	Formula	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	$Y' = \sqrt{Y}$	▪ 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	$Y' = asin\sqrt{Y}$	▪ stabilize the variance	Y is a proportion or a percentage

B. 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 *Xs*). 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 predictors 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 leave-one-out analysis. It is say that a model with Q^2 (determination coefficient in leave-one-out analysis) > 0.6 and $|R^2-Q^2| < 0.1$ is a desired model in QSAR analysis [36]. 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.

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

a) Look to the five largest and five smallest values ← detect if the values are in the plausible range. Also look to descriptive statistics value: mean, standard deviation ± histogram.

b) Plot the independent variable(s) vs dependent variable.

c) Plot the values associated to studentized residuals (s_i), leverage (h_i), Cook's (D_i) vs individual X_i values. The hat values ($0 \leq h_i \leq 1$) 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 [37]) 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 [38]).

d) Mallows' C_p-statistic ($C_p = SS_{res}/MS_{res} - n + 2*(k+1)$, k = number of dependent variables in the model) [39], [40], [41]: 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.

e) Akaike's information criterion and derivative formulas: assess the degree of fit by involving the goodness-of-fit of the model (R^2): Akaike information criterion ($AIC = n\cdot ln(RSS)/n + 2\cdot k$, where n = sample size, RSS = residual sum of squares; k = number of parameters in the model) [42]; AIC based on the determination coefficient ($AIC_{R2} = ln[(1-R^2)/n] + 2\cdot k$); McQuarrie and Tsai corrected AIC ($AIC_u = ln[RSS/(n-k)] + (n+k)/(n-k-2)$) [43]; Bayesian Information Criterion ($BIC = n\cdot ln[RSS/(n-k)] + k\cdot ln(n)$) [44]; Amemiya Prediction Criterion ($APC = RSS/n\cdot(n-k)/(n+k)$) [45]; Hannan-Quinn Criterion ($HQC = n\cdot ln(RSS/n) + 2\cdot k\cdot ln[ln(n)]$ [46]. 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: $w_i = [exp(-0.5\cdot\Delta_i)/[\sum_{j=1}^{J} exp(-0.5\cdot\Delta_j)]]$ [47] 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.

f) Kubinyi function (FIT) [48], [49]: $FIT = [R^2\cdot(n-k-1)]/[(n+k^2)\cdot(1-R^2)]$. The highest the FIT value the better the model is considered.

Other parameters that can found their usefulness in diagnosis of a MLR are presented in Table 3. Several parameters presented in Table 3 are also used by some authors as measures of model predictivity power (see for example MAE [50]).

C. 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 4.

TABLE III. OTHER 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 $0 < RMS < \infty$
Average Prediction Variance (APV)	$APV = \frac{RMS}{n}\cdot(n+k)$ [51]	The smaller the better
Total Squared Error (TSE)	$TSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\hat{\sigma}^2} + 2\cdot k - n$ [52] $TSE = \frac{SSE}{MSE} - (n - 2\cdot k) + 2$ [39]	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)	$APMSE = \frac{RMS}{n-k-1}$ [53]	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 $0 < MAE < \infty$
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 $0 < RMSE < \infty$
Mean Absolute Percentage Error (MAPE) - Measure of accuracy expressed as percentage	$MAPE = \frac{\sum_{i=1}^{n}\lvert (y_i - \hat{y}_i)/y_i \rvert}{n}$ [54], [55]	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 i^{th} compound in the sample; y_i = observed/measured activity/property of i^{th} compound; SSE = sum of squared errors; MSE = mean of squared errors

TABLE IV. STATISTICS FOR ASSESSMENT THE PREDICTIVE POWER OF MLR.

Parameter (Abbreviation)	Formula [ref]	Remarks
Predictive Squared Correlation Coefficient in Training Set (Q_{F1}^2)	$Q^2_{F_1} = 1 - \frac{\sum_{i=1}^{n_{TS}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TS}} (y_i - \bar{y}_{TR})^2}$ [56]	Prediction is considered accurate if the predictive power of the model is > 0.6 [57]
Predictive Squared Correlation Coefficient in Test Set (Q_{F2}^2)	$Q^2_{F_2} = 1 - \frac{\sum_{i=1}^{n_{TS}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TS}} (y_i - \bar{y}_{TS})^2}$ [58]	
External Predictive Ability (Q_{F3}^2)	$Q^2_{F_3} = 1 - \frac{\sum_{i=1}^{n_{TS}} (\hat{y}_i - y_i)^2 / n_{TS}}{\sum_{i=1}^{n_{TS}} (y_i - \bar{y}_{TR})^2 / n_{TR}}$ [59]	
Predictive Power (PP): Fisher's approach	$t = \frac{\overline{res}_{TS} - 0}{stdev(res_{TS}) / \sqrt{n_{TS}}}$ [60] $p = TDIST(abs(t), n_{TS}-1, 1)$	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; $\bar{y}$ = the mean of estimated/predicted activity/property; $\hat{y}_i$ = predicted value of the i^th compound in the sample; y_i = observed/measured activity/property of i^th compound; $\overline{res}$ = mean of residuals; stdev = standard deviation; TR = training set; TS = test set; EXT = external set; abs = absolute value

The diagnosis of a linear regression model could be conducted using a series of statistical parameters calculated on contingency table [61] whenever classification of compounds activity is useful. 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.

III. PRACTICAL CONSIDERATIONS

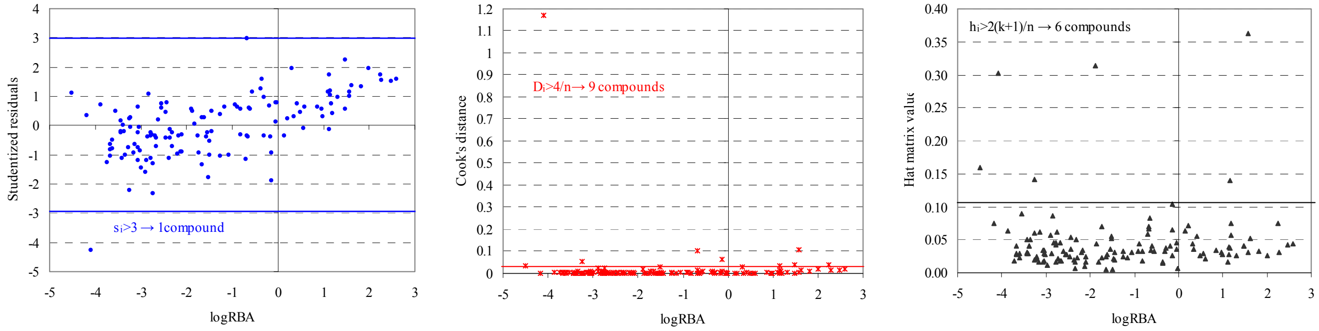
Three data sets of endocrine disrupting chemicals with experimental values of relative binding affinity expressed in logarithmic scale (logRBA) [62] were used for exemplification. The investigated compounds could be classified according to their logRBA values as weak binders (logRBA < -2.0), moderate binders (-2.0 ≤ logRBA ≤ 0) and strong binders (logRBA > 0) [63]. The following descriptors were previously calculated on the investigated structures [62] 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 independent variable (logRBA) in training set. One out of three tests rejected the null hypothesis of normality (Chi-Squared statistics = 14.862, p-value = 0.03781). No outlier had been identified when the Grubb's test was applied but there was one compound with studentized residuals higher than 3 standard deviations, compound which was withdrawn. The experimental data in training test proved not normal distributed according to Chi-Squared test, the normality test that is known to be affected by the presence of outlier(s) [22], 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).

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. The characteristics of the obtained models are presented in Table 5.



a) b) c)

Figure 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).

The analysis of the models (Table 5) revealed that none model proved collinearity (the highest correlation coefficient did not exceeded 0.8). As an overall classification, it could be say that the D_i-model is the first best model and it is followed by the h_i-model. The D_i-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 classification of the models according to information criteria led to the same order as previous: D_i-mode as the first best, h_i-model as the second best and full-model as the last best.

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
s_{est}	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
s_{loo}	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
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	$9.8317^\dagger$	33.1255
HQC	26.2887	$34.7113^\dagger$	7.8043
FIT	1.3058	2.3097	1.5076

* $p \geq 0.05$; ** $p = 0.0378$; † = absolute values; KS = Kolmogorow-Smirnov; AD = Anderson Darling; CS = Chi-Squared; R = correlation coefficient; VIF = Variance Inflation Factor; T = tolerance; R^2 = determination coefficient; R^2_{adj} = adjusted determination coefficient; s_{est} = standard error of the estimate; F-value = Fisher's statistics; Q^2 = determination coefficient in leave-one-out analysis; s_{loo} = standard error of the predict; 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;

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 [64]). The first best 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 is the preferred model and from the inspection of the Akaike weights in Table 5, this model is 1.2 ($w_{i\text{-AICR2}}$) to 1.4 ($w_{i\text{-AICc}}$) times more likely the best model in terms of Kullback-Leible discrepancy, a measure of distance between the probability generated by the model and reality [65], that is the second-best model h_i.

Significant differences between models could also been observed if the BIC and HQC parameters are analyzed; the smallest value of BIG identified the D_i-model as first best while the smallest value of HQC sustain the h_i-model as the first best model.

The plots of residuals versus predicted values for the investigated models are presented in Figure 2.

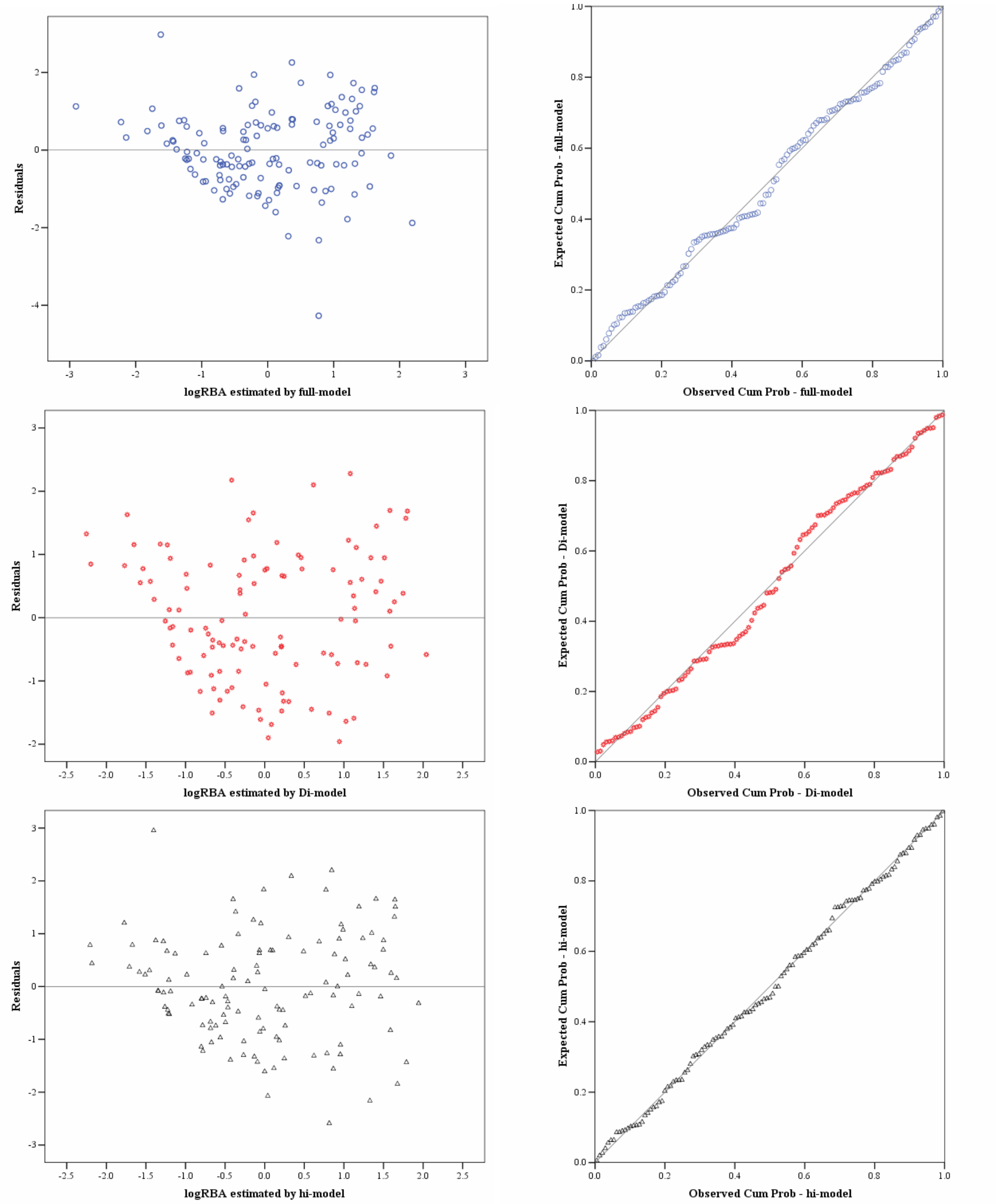


Figure 2. Scatter plots of residuals versus estimated logRBA by full model, Cook's distance (D_i) model and hat matrix leverage (h_i) model and associated normal probability plots.

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 to the pattern of the residuals, the most appropriate model is the D_i-model since the distribution indicates an unbiased and homoscedastic model. Furthermore, both full-model and h_i-model showed clear evidence of heteroscedascity, the error in estimating logRBA increasing as the value of logRBA increase. However, even if both models showed heteroscedascity could be accepted because none of them show the presence of systematic errors or inadequacy. 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 [66]. 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 [67].

As far as the normality is concern, in none of the cases the normal probability plot is far away from a straight line but the h_i-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 3 are presented in Table 6.

The total square error is the single parameter that has the same value for all models and in all cases is equal to the sum between number of independent variables in the model (in our example 6) and 1, 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 6 led to the classification obtained according to the parameters presented in Table 5: D_i-model the first best, hi-model the second best and full model the last.

Four parameters were used to assess the predictive power of the models and their results are presented in Table 7. The analysis of results presented in Table 7 revealed the followings:

- External predictive ability parameter (Q_{F3}^2) [59] 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.

TABLE VI. MLR IN TRAINING SETS: OTHER STATISTICAL PARAMETERS FOR DIAGNOSIS OF LRM

Parameter (Abbreviation)	**Full-model (n=132)**	**D_i-model (n=115)**	**h_i-model (n=123)**
Residual Mean Square (RMS)	1.1361	0.6815	0.9870
Average Prediction Variance (APV)	1.1877	0.7170	1.0351
Total Squared Error (TSE)	7.0000	7.0000	7.0000
Average Prediction Mean Squared Error (APMSE)	0.0091	0.0063	0.0085
Mean Absolute Error (MAE)	0.8356	0.6812	0.7827
Root Mean Square Error (RMSE):	1.0414	0.8037	0.9689
Mean Absolute Percentage Error (MAPE)	1.3033	1.0797	1.1649
Standard Error of Prediction (SEP)	1.0453	0.8072	0.9729
Relative Error of Prediction (REP%)	73.9756	58.0395	70.9144

TABLE VII. RESULTS REGARDING THE PREDICTIVE POWER OF THE MODELS

Criterion	**Full-model (n=132)**		**D_i-model (n=115)**			**h_i-model (n=123)**		
	test[a]	**external[b]**	**test[a]**	**external[b]**	**withdrawn[c]**	**test[a]**	**external[b]**	**withdrawn[d]**
Q_{F1}^2	0.5498	-0.1890	0.4796	-0.4581	0.2009	0.6476	-0.4444	0.7434
Q_{F2}^2	0.4804	0.2010	0.3875	0.1450	0.0443	0.5738	0.1112	0.7431
Q_{F3}^2	0.5527	-16.3066	0.7809	-17.6311	-4.4056	0.7813	-18.5792	-2.9125
PP (p)	-1.7852 (0.0440)	-2.8228 (0.0112)	-2.0961 (0.0239)	-3.0020 (0.0085)	0.1039 (0.4593)	-0.4239 (0.3379)	-2.9139 (0.0097)	0.0489 (0.4812)

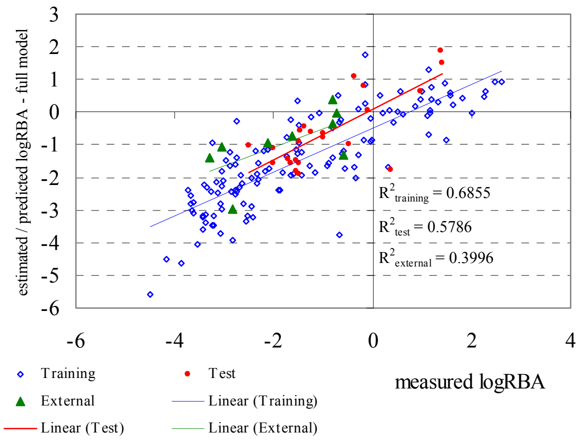
Q^2_{F1} = predicted squared correlation coefficient in training set;
Q^2_{F2} = predicted squared correlation coefficient in test set; Q^2_{F3} = external predictivity ability; PP = predictive power;
PP = Predictive Power: Fisher's approach; [a] n=23; [b] n = 9; [c] n = 16; [d] n = 8

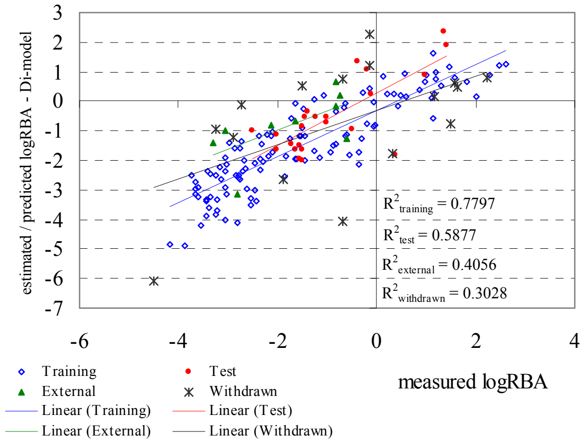
- D_i-model achieve the criterion of exceeding 0.6 [58] in just one of case out 6 possible while the h_i-model reach this criterion in four out of 6 cases. The h_i-model accomplished more frequently the criteria of having values higher than 0.6 while the full-model did not accomplished at all this criterion. 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_{F1}^2 and the Q_{F2}^2 suffer from drawbacks [68].
- 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 h_i-model proved prediction power.

The classification of the models according to results presented in Table 7 is as follows: h_i-model the first best, D_i-model the second best and full-model the last best.

One remark about the parameters used to assess the predictive power, namely Q_{F1}^2, Q_{F2}^2 and Q_{F3}^2, can be made. Even the symbols contain "square", these parameters could take both positive and negative values according to their formula (see Table IV). it is not a definition for quantities with just positive values. Furthermore, a correlation coefficient is expected to take values between -1 and 1 while a determination coefficient is expected to take values from 0 to 1, but for example the Q_{F3}^2 parameter took values that exceeded these ranges. Therefore, these statistical parameters should be considered as biased estimators of the determination.

Other statistics were introduced to test the external predictivity of QSAR. One example is the r_m^2, a parameter computed by forcing the regression through origin [69] with certain applicability like as the line slopes not near to 1 [70].





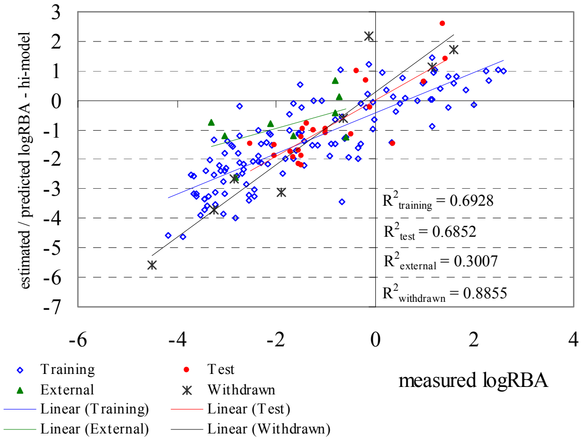


Figure 3. Scatter plots of measured logRBA versus estimated/predicted logRBA by full model (estimated), Cook's distance (D_i) model (predicted), and hat matrix leverage (h_i) model (predicted) on training (estimated), test (predicted), external (predicted), and withdrawn (predicted) sets.

Regarding the r^2_m parameter, the main differences between a model without and with intercept could be summarized as: the degrees of freedom for residuals are not the same, the formula for sum of squares is different, and the coefficient of determination can be absurdly large even for weak correlation between *X(s)* and *Y*. Basically, regression through the origin should not be used in the absence of a strong reason (such as data of *X(s)* in the vicinity of zero).

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.

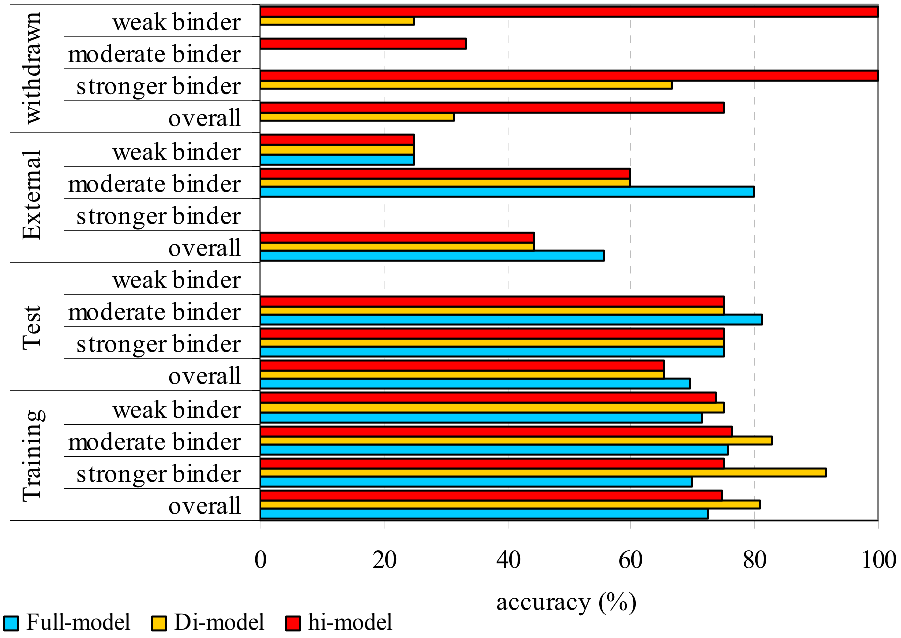


Figure 4. Accuracy of full-, D_i- and h_i- models on training, test, external and withdrawn sets.

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.

IV. 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 paper.

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 under implementation and assessment is our lab. 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).

REFERENCES

[1] Y. M. Chan, "Biostatistics 201: Linear Regression Analysis," Singapore Med. J., vol. 45, no. 2, 2004, pp. 55-61.

[2] A. Bravais, "Analyse mathématique sur les probabilités des erreurs de situation d'un point," Memoires par divers Savans T. IX., Paris, 1844, pp. 255-332.

[3] F. Y. Edgeworth, "Correlated Averages," Phil Mag (5th series), vol. 34, 1892, pp. 194-204.

[4] F. Galton, "Regression towards mediocrity in hereditary stature," J. Anthropol. Inst. Great Brit. Ireland, vol. 15, 1886, pp. 246-263.

[5] F. Galton and J. D. H. Dickson, "Family likeness in stature," Proc. R. Soc. Lond., vol. 40, 1886, pp. 42-73.

[6] A Quetelet, "A Treatise on Man and the Development of his Faculties," (Edinburgh: W and R Chambers, 1842) [online] [accessed March 2, 2013]. Available from: http://archive.org/stream/treatiseonmandev00quet#page/n7/mode/2up

[7] K. Pearson, "The Life, Letters and Labors of Francis Galton," Cambridge University Press, 1930.

[8] K. Pearson, "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia," Philos. Trans. R. Soc. Lond., vol. 187, 1896, pp. 253-318.

[9] G. U. Yule, "An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades (part I)," J. R. Stat. Soc. vol. 62, 1899, pp. 249-295.

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

[11] G. U. Yule, "On the theory of correlation," J. R. Stat. Soc., vol. 60, 1897, pp. 812-854.

[12] A. M. Legendre, "Nouvelles Méthodes pour la Détermination des Orbites des Comètes. Paris: Courcier," 1805. [online] [accessed February 24, 2013]. Available from: http://archive.org/stream/nouvellesmthode00legegoog#page/n9/mode/2up

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] D. M. Hawkins, "The problem of overfitting," J. Chem. Inf. Comput. Sci., vol. 44, no. 1, 2004, pp. 1-12.

[22] 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.

[23] A. Kolmogorov, "Confidence Limits for an Unknown Distribution Function," Ann. Math. Stat., vol. 12, no. 4, 1941, pp. 461-463.

[24] N. V. Smirnov, "Tables for estimating the goodness of fit of empirical distributions," Ann. Math. Stat., vol. 19, 1948, pp. 279.

[25] 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.

[26] 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.

[27] A. A. Shapiro and M. B. Wilks, "An analysis of variance test for normality (complete sample)," Biometrika, vol. 52, 1965, pp. 591-611.

[28] F. Grubbs, "Procedures for Detecting Outlying Observations in Samples," Technometrics, vol. 11, no. 1, 1969, pp. 1-21.

[29] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression, I," Biometrika, vol. 37, 1950, pp. 409-428.

[30] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression, II," Biometrika, vol. 38, 1951, pp. 159-179.

[31] T. S. Breusch and A. R. Pagan,. "Simple test for heteroscedasticity and random coefficient variation," Econometrica, vol. 47, no. 5, 1979, pp. 1287-1294.

[32] M. S. Bartlett,. "Properties of sufficiency and statistical tests," Proc. Roy. Stat. Soc. A, vol. 160, 1937, pp. 268–282.

[33] H. Levene, "Robust tests for equality of variances". In I. Olkin, H. Hotelling, et al. Stanford University Press, 1960, pp. 278-292.

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

[35] 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.

[36] A. Tropsha, "Best practices for QSAR model development, validation, and exploitation," Mol. Inf., vol. 29, 2010, pp. 476-488.

[37] D. C. Hoaglin and R. E. Welsch, "The hat matrix in regression and ANOVA," Am. Stat., vol. 32, 1978, pp. 17-22.

[38] 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.

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

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

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

[42] H. Akaike, "Fitting Autoregressive Models for Prediction," Ann. I. Stat. Math., vol. 21, 1969, pp. 243-247.

[43] A. D. R. McQuarrie and C.-L. Tsai, "Regression and time series model selection in small samples," World Sci., 1988, pp. 32.

[44] G. Schwarz, "Estimating the dimension of a Model," Ann. Stat., vol. 6, 1978, pp. 461-464.

[45] T. Amemiya, "Qualitative response models: A survey," J. Econ. Lit., vol. 19, 1981, pp. 1483-1536.

[46] 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, 1979, pp. 190-195.

[47] 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.

[48] H. Kubinyi, "Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution," QSAR Comb. Sci., vol. 3, 1994, pp. 393-401.

[49] H. Kubinyi, "Variable Selection in QSAR Studies. I. An Evolutionary Algorithm," QSAR Comb. Sci, vol. 13, 1994, pp. 285-294.

[50] 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.

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

[52] J. W. Gorman and R. J. Toman, "Selection of variables for fitting equations to data," Technometrics, vol. 8, 1966, pp. 27-51.

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

[54] J. S. Armstrong, "Long-range Forecasting: From Crystal Ball to Computer," Wiley, 1978.

[55] B. E. Flores, "A pragmatic view of accuracy measurement in forecasting," Omega (Oxford), vol. 14, no. 2, 1986, pp. 93-98.

[56] 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, 2001, pp. 186-195.

[57] A. Golbraikh and A. Tropsha, "Beware of q2!", J. Mol. Graphics Mod., vol. 20, 2002, pp. 269-276.

[58] 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.

[59] V. Consonni, D. Ballabio, and R. Todeschini, "Comments on the Definition of the Q^2 Parameter for QSAR Validation," J. Chem. Inf. Model., vol. 49, no. 7, 2009, pp. 1669-1678.

[60] 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.

[61] 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.

[62] 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.

[63] 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, 2000, pp. 138-153.

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

[65] K. P. Burnham and D. R. Anderson, "Kullback–Leibler infor- mation as a basis for strong inference in ecological studies," Wildlife Res., vol. 28, 2001, pp. 111-119.

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

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

[68] V. Consonni, D. Ballabio, and R. Todeschini, "Evaluation of model predictive ability by external validation techniques," J. Chemom., vol. 24, 2010, pp. 194-201.

[69] P. K. Ojha, I. Mitra, R. N. Das, and K. Roy, "Further exploring r^2_m metrics for validation of QSPR models," Chemom. Intell. Lab. Syst., vol. 107, 2011, pp. 194-205.

[70] 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, 2011, pp. 2320-2335.