J Comput Aided Mol Des (2013) 27:837–844
DOI 10.1007/s10822-013-9680-4

Estimation of influential points in any data set from coefficient of determination and its leave-one-out cross-validated counterpart

Gergely Tóth · Zsolt Bodai · Károly Héberger

Received: 28 June 2013 / Accepted: 3 October 2013 / Published online: 20 October 2013

Abstract Coefficient of determination (R^2) and its leave-one-out cross-validated analogue (denoted by Q^2 or R^2_{cv}) are the most frequantly published values to characterize the predictive performance of models. In this article we use R^2 and Q^2 in a reversed aspect to determine uncommon points, i.e. influential points in any data sets. The term $(1 - Q^2)/(1 - R^2)$ corresponds to the ratio of predictive residual sum of squares and the residual sum of squares. The ratio correlates to the number of influential points in experimental and random data sets. We propose an (approximate) F test on $(1 - Q^2)/(1 - R^2)$ term to quickly pre-estimate the presence of influential points in training sets of models. The test is founded upon the routinely calculated Q^2 and R^2 values and warns the model builders to verify the training set, to perform influence analysis or even to change to robust modeling.

Keywords Coefficient of determination · Leave-one-out cross-validation · Influence analysis · Quantitative structure activity relationships · Prediction · Training set

G. Tóth · Z. Bodai
Institute of Chemistry, Loránd Eötvös University, Pázmány Sétány 1/a, Budapest 1117, Hungary

K. Héberger (✉)
Institute of Materials and Environmental Chemistry, Research Centre of Natural Sciences, Hungarian Academy of Sciences, Pusztaszeri út 59-67, Budapest 1025, Hungary
e-mail: heberger.karoly@ttk.mta.hu

Introduction

Model validation and evaluation of predictive ability are basic steps in chemometrics, bioinformatics, quantitative structure activity relationship (QSAR) and quantitative structure retention relationship (QSRR). The coefficient of determination (R^2) and the leave-one-out cross-validated R^2 (Q^2 or R^2_{cv}) e.g. in Ref. [1] are performance parameters calculated in most studies. In the last decades there is a plenty of discussion on the qualitative and the quantitative meaning of these parameters in the validation and prediction processes alike. There are other ways of calculations for performance parameters, e.g. they can be calculated on the training set and on the test set of the data [2–7]. The former is called internal validation, the latter is called external one. We can use the mean of the test set or of the training set in external Q^2 calculations [6, 7]. Further functions can be defined, if we take into account the degrees of freedom of the sums of squares in the calculations [1, 7]. In the case of Q^2, most of the calculations are performed with the leave-one-out cross-validation method, but there are many examples for different number of data to leave out [3, 8].

Though the interpretation of R^2 is usually straightforward, in the case of Q^2, the interpretation is not unified or even dubious. Some authors only take into account the value of Q^2 to R^2. If Q^2 is only “slightly” less than the corresponding R^2, the model is considered to be validated [2, 6]. However, the measure for “slight” difference cannot be given, especially not without the degree of freedom. Other users concentrate on the numerical value of Q^2 without the degree of freedom and without any comparison to R^2. If it is larger than e.g. 0.5, the model is thought to be validated [2, 9–11]. It is not necessary that Q^2 calculated on the training set correlates to the external predictive ability as it is stated in the article entitled "Beware of Q^2" written by Golbraikh and Tropsha on QSAR in 2002 [2]. Doweyko repeated this observation in his paper entitled "QSAR: dead or alive?" [12].

The literature on Q^2 is connected mostly on model validation and predictive ability. Leave-one-out Q^2 on the training set is a measure of internal predictive power and it is not the standalone best choice to quantify predictive performance in general, e.g. [4, 13–15].

In this article we focus on a different aspect of R^2 and Q^2. Originally, we tried to develop a statistical test to be used in model validation, where the input data are R^2 and Q^2 calculated on the training set. We tried with different formulas, but none of them indicated reliable correlation to the expected validity of the models. Looking through the calculation details of R^2 and Q^2, we realized that our methods were not connected to the validity or the predictive ability of the models, but they were connected to a different feature of the training set. Here, we suggest using a statistical test to pre-estimate the presence of influential points in the training set. In our study we focus on the average model builders of QSAR or QSRR ones, where ordinary or partial least square regressions are applied, and R^2 and Q^2 are routinely calculated. Influence analysis, identification of x and y outliers, and comparison to robust regression are usually outside of scope in the average QSAR/QSRR publication. Therefore, an introduction of a method that alerts model builders is a valuable aim. There are two general ways to investigate the data set and the model building for uncommon points (e.g. Ref. [16]). The first one is the regression diagnostics pioneered by Cook [17, 18]. Here, the model is fitted to the whole data set first, and thereafter the influential points, x and y outliers are detected via different criteria [1, 18]. The other way is to use robust regression, where the model is built on a subset or on a weighted set of data. Here, the uncommon feature of the points is taken into account in the model building. The outliers are quantified with large robust residuals in the y direction and robust distances in the predictor space. The latter points are usually termed as leverages. Since the aim of our study is to introduce a quick test to alert uncommon points in the data set of QSAR studies, where the model building has been already performed with ordinary or partial least square methods, we have limited ourselves to the first type of regression diagnostics. It does not mean, that we question or neglect the results obtained in the last decades with robust regression, simple the aim and the corresponding preconditions do not allow its use. The suggested test uses *PRESS* (predictive residual sum of squares) and its meaning is at least questionable in the combination of the leave-one-out method and robust regression.

Theory

Calculation of sum of squares

We denote with *TSS*, *RSS* and *MSS* the total, residual and model sum of squares of n data, y_i. The average of the experimental data is denoted by $\bar{y}$ and $\hat{y}_i$-s are the data calculated by a model. The number of the parameters (including intercept) in the model is p.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \quad MSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{1}$$

The coefficient of determination is defined as

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2}$$

because $TSS = MSS + RSS$.

In the case of internal cross-validation with leave-one-out method, we can calculate the predictive residual sum of square as:

$$PRESS = \sum_{i=1}^{n} \left(\hat{y}_{i/i} - y_i\right)^2 \tag{3}$$

where $\hat{y}_{i/i}$ denotes the value calculated for the i-th experiment leaving out the i-th experiment in the parameterization of the model. The cross-validated correlation coefficient is defined in Eq. 4.

$$Q^2 = R^2_{cv} = 1 - \frac{PRESS}{TSS} \tag{4}$$

$R^2 \in [0;1]$, but Q^2 can be negative, if the model performs weakly (worse than modeling with a simple average), therefore $Q^2 \in (-\infty, 1]$.

The basic assumption in our test is that the ratio of two variances sampling from the same normal distribution follows F distribution with the corresponding degree of freedom. Both *RSS* and *PRESS* are sum of squares with df_{RSS} and df_{PRESS} degrees of freedom. If our data set (training set) is correctly chosen, we can reasonably expect that

$$\frac{PRESS/df_{PRESS}}{RSS/df_{RSS}} = \frac{(1-Q^2)/df_{PRESS}}{(1-R^2)/df_{RSS}} \approx F\text{–}distributed \tag{5}$$

Strictly speaking the F distribution in (Eq. 5) is only valid when *PRESS* and *RSS* are independent. The *PRESS* is higher or equal to *RSS*, i.e. they are not fully independent. Hence we emphasize the approximate sign in Eq. (5).

Therefore, a traditional F test (known also as variance ratio test) gives us information that the models on the reduced data sets obtained by the leave-one-out way are significantly different in the aspect of the variance from the one derived on whole data set. Of course it is not easy to identify the direct link between the meaning of "difference in the aspect of the variance" and "model validation".

One of the reviewers suggested that the *PRESS/RSS* test (i.e. the traditional parametric F test) might be substituted with a non-parametric alternative. However, we have not found any reasonable algorithm (i.e. using bootstrap) for our case, where a given *PRESS/RSS* ratio is available from the literature. The bootstrap on the given data set provides very important (but different) information. Namely, it provides the uncertainty (confidence interval, histogram) of the *PRESS/RSS* on the given data set. This issue is detailed in the result and discussion part (3.3).

Identification of influential points

There is no unique mathematical definition of an influential observation in the literature, therefore we used the following "… compared to other observations it has a relatively large impact on the estimated quantities like response, regression coefficient, standard error, etc." [1]. One or more parameters are extremely sensitive to the influential observation. If we omit the observation, there is a reasonable difference in the parameter set causing different models. Outliers and influential points are similar but not identical concepts. Many of the outliers are influential points as well, despite that they are outliers in the y direction (termed often as outliers) or in the x direction (known as leverages). There are mathematical definitions for outliers, there are methods to detect them despite the masking effect, but it is a mismatch to use the outlier definitions for influential observations.

To identify the influential points in data sets, we selected some basic methods. A good survey of the methods was published in 1986 [18]. A comparison of some new methods to robust methods was performed recently, as well [16]. As we mentioned earlier, we did not use all available methods to identify influential points, because we concentrated on the studies, where Q^2 and R^2 are calculated and the model can be obtained with ordinary least squares and partial least square regressions. Robust methods are very efficient to build models with correct treatise of outliers and leverages, but an average QSAR or QSRR developer avoids using robust methods. The aim of our study was to develop a quick pre-estimation tool on uncommon data points for average model builders, who are not interested in model building with special knowledge on robust statistics or influence analysis. Therefore, we deliberately chose several non-robust methods being differently sensitive on the presence of influential points in dissimilar data sets. Here, we outline the selected ones as described in the handbook of Frank and Todeschini [1]. For details see the references therein and the surveys mentioned in refs. [16, 18]. We performed some calculations with robust methods as well (e.g. [16, 19]), but the results are questionable for the aim of the study and because of the combination of ordinary least squares and robust regression.

HD: The i-th observation is called influential point, if the corresponding diagonal element of the hat matrix (h_{ii}) is

$$h_{ii} > 2p/n \tag{6}$$

The hat matrix is calculated as $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ for ordinary least square regression, where $\mathbf{X}$ is the predictor matrix. Since h_{ii} is proportional to the square of the Mahalanobis distance of the observation from the mean of observations, this definition often used to identify leverages, as well.

SR: t_{i} denotes the studentized residuals and calculated as

$$t_i = \frac{r_i}{s_{/i}\sqrt{1-h_{ii}}} \tag{7}$$

where $r_i = \hat{y}_i - y_i$. $s_{/i} = \sqrt{\frac{(n-p)s^2 - r_i^2/(1-h_{ii})}{n-p-1}}$ means the standard error without the i-th observation. The i-th observation is influential, if $t_{\mathrm{i}} > 2$.

COOK: Cook's method is used for regression, where r_{si} denotes the standardized residuals and s_{r} is the residual standard deviation. In this method a d_{i} value is defined and tested in an F test with p and $n-p$ degrees of freedom.

$$COOK_i = \frac{r_{si}^2 h_{ii}}{p(1-h_{ii})} \tag{8}$$

Practically, the F test can be replaced by the comparison of $COOK_{\mathrm{i}}$ to different limit values (constants). We defined influential points as COOK-1, if $COOK_{\mathrm{i}} > 1$ and COOK-2, if $COOK_{\mathrm{i}} > 4/n$ [17]. According to the classification of Chatterjee and Hadi [18], Eq. 8 belongs to the influence function type definitions.

COVRATIO: The covariance ratio method measures the influence of the i-th observation on the variance of the regression coefficients.

$$COVRATIO_i = \left(\frac{s_{/i}}{s}\right)^{2p}\frac{1}{1-h_{ii}} = \frac{1}{(1-h_{ii})[(n-p-1)/(n-p) + t_i^2/(n-p)]^2} \tag{9}$$

We used the definitions of influential points with $|COVRATIO_i - 1| > 3p/n$ [20]. It differs from the (maybe mistyped) definition in the book of Frank and Todeschini [1]. Eq. 9 is related to the volume of confidence ellipsoids according to the classification of Ref. [18].

DFBETAS is calculated using the b_{j} estimated regression coefficient, its $b_{\mathrm{j/i}}$ estimation when the i-th experiment is omitted and c_{jj}, is the diagonal of the $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ matrix.

$$DFBETAS_{ij} = \frac{b_j - b_{j/i}}{s_{/i}\sqrt{c_{jj}}} \tag{10}$$

The i-th data is treated as influential observation, if $DFBETAS_{ij} \geq 2/\sqrt{n}$ [20]. The definition is related to partial influence [18].

DFFITS: The scaled variable is the difference between the predicted and the response of the i-th observation with and without using the observation in the model. It is scaled by the standard error of the observations.

$$DFFITS_i = t_i\left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} \tag{11}$$

A point is influential, if $DFFITS_i > 2\sqrt{p/n}$ (case DFFITS-1) or $DFFITS_i > 2$ (case DFFITS-2). It belongs to the influence function type definitions [18].

Results and discussions

Simulated data sets

Data sets were generated using random numbers to mimic experimental data to be modeled with multivariate ordinary least squares regression. The superscripts denote the dimension of the variables. At first $\mathbf{y}^{\mathrm{p}}$ and $\mathbf{X}^{\mathrm{p}\times(\mathrm{p}-1)}$ variables were filled with uniform random numbers of [0;1). $\mathbf{X}^{\mathrm{p}\times(\mathrm{p}-1)}$ was extended with a p-th column containing ones. The solution of the set of linear equations $\mathbf{X}^{\mathrm{p}\times(\mathrm{p}-1)}\mathbf{p}^{\mathrm{p}} = \mathbf{y}^{\mathrm{p}}$ provided a set of regression parameters $\mathbf{p}^{\mathrm{p}}$, where the last element of the $\mathbf{p}^{\mathrm{p}}$ vector was the intercept in the regression. The number of the rows of $\mathbf{X}$ was extended from p to n and the new rows were filled with random numbers from uniform [0;1) distribution. The last column of $\mathbf{X}$ contained only ones. The dimension of the column vector $\mathbf{y}$ was extended from p to n. The new elements were calculated using the previously obtained $\mathbf{p}^{\mathrm{p}}$ regression parameters and the generated new rows of $\mathbf{X}^{\mathrm{n}\times\mathrm{p}}$ using the equation $\mathbf{y}^{\mathrm{n}} = \mathbf{X}^{\mathrm{n}\times\mathrm{p}}\mathbf{p}^{\mathrm{p}}$. Finally, a white noise of $w*\varepsilon$ was added to each elements of $\mathbf{y}$, where ε was a random number chosen from standard normal distribution and w was a predefined factor. Nine parameter sets were used because of practical reasons: $n = 10$, $p = 5$; $n = 20$, $p = 5$; $n = 20$, $p = 10$; with combinations of $w = 0.05$, $w = 0.10$ and $w = 0.25$ weights of white noise; 10^5 random model calculations were performed for each parameter sets resulted all together 9×10^5 datasets. The limit correlation coefficients for chance correlation for $n = 20$ (or $n = 10$) is 0.444 (or 0.632) at the 5 % level according to the Table C-3 of Bevington [21]; i.e. the medium range for correlation coefficients were used, where the distortions can be effectively observed. Such a way the data sets will contain outliers, influential points randomly.

If we used an F-like test for the ratio defined in Eq. 5, we need to know the degrees of freedom both for RSS and $PRESS$. In the case of ordinary least square regression $df_{\mathrm{RSS}} = n - p$. We found in the literature that $df_{\mathrm{PRESS}} = n - p$ is used without any proof or explanation. To test this we determined df_{PRESS} numerically. $PRESS$ is a sum of squares of residual quantities. If we accept the reasonable assumption of OLS regression that the $PRESS$ residuals are not serially correlated and they are normally distributed, their sum of squares shows χ^2 distribution. The shape of the χ^2 distribution functions can be used to determine the degrees of freedom [22] as it depends strongly on them. We calculated the histograms of our $PRESS$-s (10^5 $PRESS$-s for each parameter set). We scaled the histograms with their standard deviations. Thereafter we calculated the overlap integral of the scaled and normalized histograms and theoretical χ^2-distributions with different degrees of freedom. The maximal overlap integrals (0.96–0.98) were obtained for the theoretical distributions with $n - p - 1$ or $n - p$ degrees of freedom for all of the nine cases. The results encouraged us to use $df_{\mathrm{PRESS}} = n - p$. It has the advantage of simplifying Eq. 5, as well.

The number of influential points for all the 9×10^5 datasets were calculated with the methods detailed above. Different number of influential points was provided according to the different definitions. We found good correlation among the number of the influential points and the $PRESS/RSS$ ratios for the methods SR, COOK-1, COOK-2, DFBETAS, DFFIT-1 and DFFIT-2. We did not detect reliable correlation for the method called HD, and we got negative correlation for the COVRATIO one. This negative correlation is not surprising, because a large SR value causes small COVRATIO, especially, if the HD method did not seem to be decisive for our data sets. The lack of positive correlation of the HD and the COVRATIO methods mean that leverage points are not necessarily influential observations, because these quantities are suitable (only) to identify leverages and not influential observations. HD is related to the Mahalanobis distance of the corresponding point to the centre of the points, and COVRATIO is related to the volume of the confidence ellipsoids [18]. We performed calculations, where PRESS/RSS-s were calculated by ordinary least squares fit and the leverage points were detected by extreme Mahalanobis distances or by extreme robust distances after different robust regression methods. In these cases we did not observe correlation between the number of leverage points and PRESS/RSS values, similarly to non-correlation with the HD and COVRATIO terms. We calculated also the correlation between PRESS/RSS from ordinary least squares regression and the number of the omitted or down-weighted observations in robust regressions, but we did not

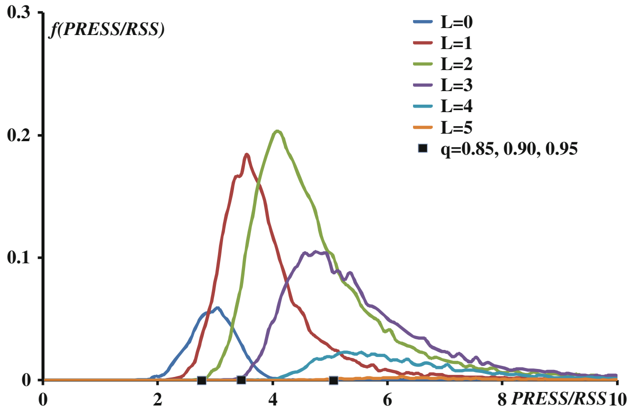

Fig. 1 Relative frequencies of the number of data sets with different number of influential points (L) identified by the COOK-2 method versus the *PRESS/RSS* of the data sets. The *black squares* denote three percentiles of the corresponding F distribution. Parameters: $n = 10$, $p = 5$ and $w = 0.05$

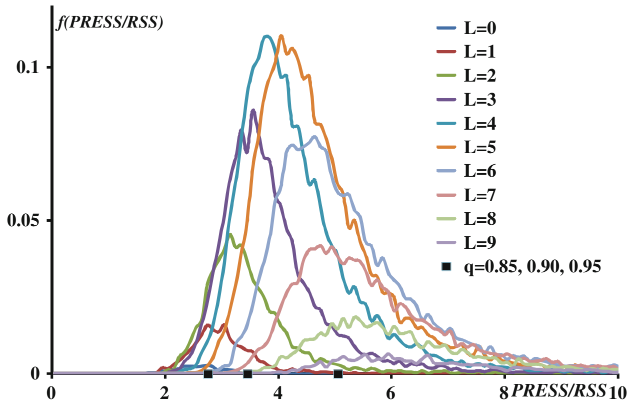

Fig. 2 Relative frequencies of the number of data sets with different number of influential points (L) identified by the DFBETAS method versus the *PRESS/RSS* of the data sets. The *black squares* denote three percentiles of the corresponding F distribution. Parameters: $n = 10$, $p = 5$ and $w = 0.05$

observe any significant correlations. Without going into details and repeating all specific aspects of robust regression, the lack of correlation in these incoherent comparisons can be caused by the differences in the definitions of Euclidean and Mahalanobis distances, by the so-called masking effect and differences in the breakdown points.

We calculated the relative frequencies of the number of data sets with different number of influential points versus the *PRESS/RSS* of the data sets. In Fig. 1 the relative frequencies are shown for the COOK-2 method in the case of $n = 10$, $p = 5$ and $w = 0.05$ parameter set. This method identifies for the most data sets 2–4 influential points. This range is not surprising due to relatively small n/p ratio. It is also known that there is some connection between the expected number of influential observations and the number of the parameters in a model [22]. Data sets with larger number of influential points (defined as COOK-2) had larger *PRESS/RSS* values. We plotted three percentiles of F distributions for 85, 90 and 95 % with $\nu_1 = 5$ and $\nu_2 = 5$ degrees of freedom. Three or four influential points were in the data sets, if the *PRESS/RSS* ratio was higher than the 90 % percentile. Five uncommon points were found, if *PRESS/RSS* was larger than the 95 % percentile.

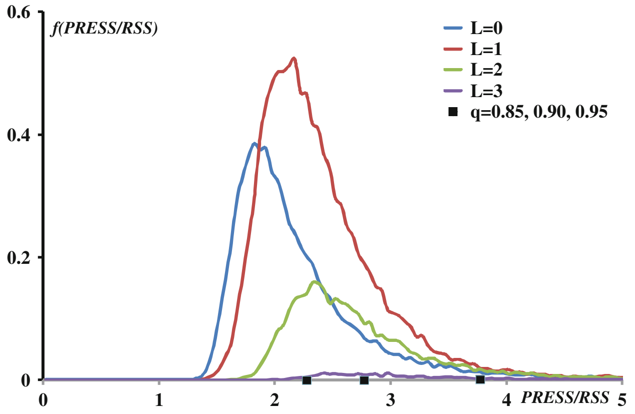

Fig. 3 Relative frequencies of the number of resampled experimental sets with different number of influential points (L) identified by the DFFITS-1 method versus the *PRESS/RSS* of the data sets. The *black squares* denote three percentiles of the corresponding F distribution. Parameters: $n = 10$, $p = 3$

The results of an even more sensitive method can be seen in Fig. 2. The DFBETAS method identified 3–6 influential points for the most of the cases in the same set ($n = 10$, $p = 5$ and $w = 0.05$) 7 and more influential points were found mostly with *PRESS/RSS* larger than the 90 % percentiles. The lack or the small number of influential points depended on the *PRESS/RSS* as well. Zero to two influential points were found mostly for *PRESS/RSS* smaller than the 95 % percentiles.

Experimental data

We tested the method on the results of Zhang et al. [23]. They performed a quantitative structure retention relationship (QSRR) study on the gas chromatographic retention indices using molecular descriptors. They built a multivariate regression model on the experimental retention data of 161 hydrocarbons using a constant and two descriptors: the total number of non-H bonds (constitutional descriptor), R autocorrelation of lag 3 weighted by atomic van der Waals volumes (GETAWAY descriptor) [24].

In order to test the relation between *PRESS/RSS* and the number of influential points we performed resampling on their data. We chose $n = 10$ or $n = 20$ molecules. We performed the regression with $p = 3$ parameters. We calculated *PRESS*, *RSS* and the number of the influential points with the different methods. We repeated the random resampling for 10^5 cases both for $n = 10$ and $n = 20$.

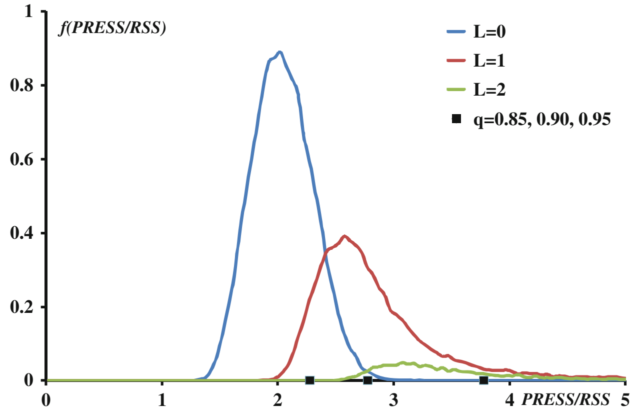

Fig. 4 Relative frequencies of the number of resampled experimental sets with different number of influential points (L) identified by the COOK-1 method versus the *PRESS/RSS* of the data sets. The *black squares* denote three percentiles of the corresponding F distribution. Parameters: $n = 10$, $p = 3$

Table 1 Frequencies a, b, c and d denote the number of occurrences of the sub cases where L is the number of the influential points in the data set

	$0 < L$	$L = 0$
$F_{\mathrm{crit}} \leq$ *PRESS/RSS*	a	b
PRESS/RSS $< F_{\mathrm{crit}}$	c	d

The relative frequencies of the number of data sets versus *PRESS/RSS* are shown for the DFFITS-1 method ($n = 10, p = 3$) in Fig. 3. There was zero or one influential point in the majority of the resampled sets. Three influential points were seldom found and it coincided mostly with *PRESS/RSS* values larger than the 85 % percentage of the corresponding F distribution. The results are shown for the COOK-1 method on the same sets in Fig. 4. This method selected only few influential points, the most of the data sets were without any influential points. Two influential points were found mostly for data sets with larger *PRESS/RSS* than the 90 % percentile of the corresponding F distribution. This percentile served also as an upper limit for the data sets with zero influential point.

We applied the *fi* coefficient e.g. in Ref. [1] to quantify the correlation among *PRESS/RSS* and influential point methods. It is defined as:

$$fi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}, \qquad (12)$$

where the meaning of a, b, c and d is detailed in Table 1.

In this calculation we distinguished according to the absence/presence of influential points, but it was an approximation in the case of our simulated data, because methods identifying the influential points observed many influential points in most of the cases. We did not predefine an F_{crit} value, but we scanned the possible *PRESS/RSS* range (x-axis of Figs. 1, 2, 3, 4) to find an F_{crit} which maximizes the *fi* coefficient. The ranges of the obtained maximal *fi*-s for the different parameter sets in the scanned F_{crit} range were as follows: SR (0.1–0.2), COOK-1 (0.6–0.8), COOK-2 (0.5–0.6), DFBETAS (0.4–0.5), DFFITS-1 (0.2–0.3), DFFITS-2 (0.3–0.4). We note again, that the absence/presence criterion fails due to the high number of the influential points found by the methods for the most data sets.

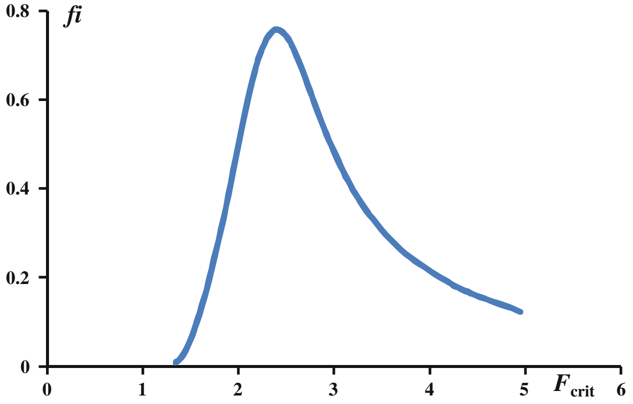

Fig. 5 Dependence of *fi* on the choice of the critical F value for the resampled experimental data of Zhang et al. $n = 10$, $p = 3$

In Figs. 1, 2, 3 and 4 we showed that higher number of influential points (L) causes shift of the relative frequency curves along the *PRESS/RSS* axis and that **the *PRESS/RSS* ratio positively correlates with the number of the influential points.** If the correlation is strong, there is a possibility to use *PRESS/RSS* to detect, or at least to pre-estimate the presence of influential points. An F test on *PRESS* and *RSS* may be used for this purpose, but we have to predefine a significance level and a corresponding F_{crit} value. A reasonable significance level can be identified, if we search the F_{crit} value, where the *PRESS/RSS* and one of the identification methods shows the maximal *fi*. It means an F_{crit} value, where the separation of data sets with and without influential points is maximal. We show the F_{crit} dependence of *fi* for the gas chromatographic retention data of Zhang et al. [23]. The COOK-1 method was chosen for the detection of influential points. It can be seen in Fig. 5, that a clear maximum is obtained at $F_{\mathrm{crit}} = 2.4$ here. $F_{\mathrm{crit}} = 2.4$ corresponds to a percentile of 86 %.

We collected Q^2 and R^2 values of 247 QSAR models from the literature. The sources of the data were mostly collections, for QSAR details see references [6, 10, 25] and references therein. We calculated the *PRESS/RSS* ratios for these data and also the F-percentiles, because n and p were accessible in the data collections. The histogram of the F-percentiles is shown in Fig. 6. Obviously, there are no data less than 0.5, because $1 \leq$ *PRESS/RSS* and $v_1 = v_2$ cause a

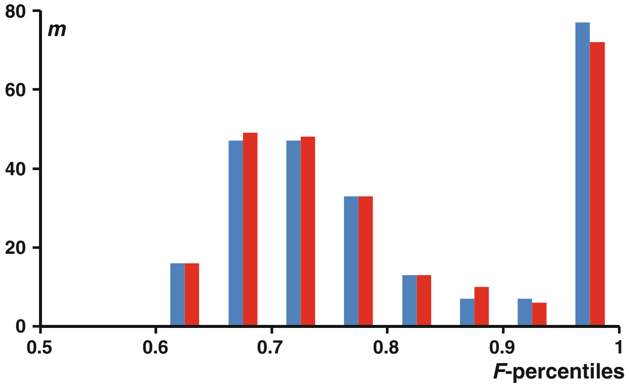

Fig. 6 Frequencies for the percentiles calculated from F distributions at the PRESS/RSS values of 247 QSAR models. *Blue* (*black*): dfPRESS and dfPRESS $= n - p$ *Red* (*gray*): estimation of the pseudo degrees of freedom, dfPRESS and dfRSS $=$ n $-$ 4*p for PLS models

minimal 0.5 percentile of the corresponding F distribution. There were no influential points in the training sets of the two thirds of the models according to our test, but one third of the QSAR models would fail on an F test of the *PRESS/RSS* ratio. It means the training sets of these models probably contained influential points. It can be interpreted that there were problems already with the internal predictive character of the models. We note here that the part of the models were taken in the collections of references [6, 10] to show the existence of better models. We note as well that the most of the models with F-percentiles larger than 0.95 (models with larger probability of influential observations) were 3D QSAR ones collected or calculated by Cramer and Wendt [10].

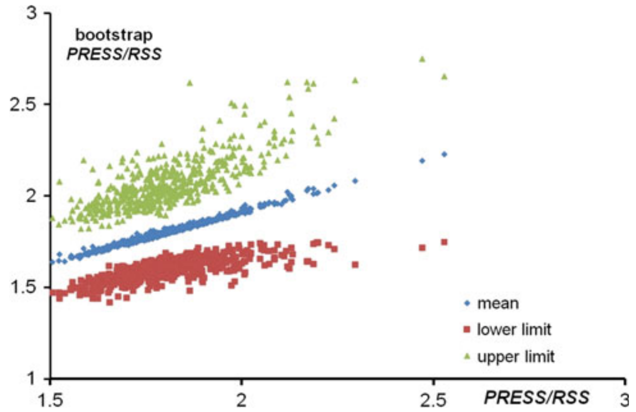

Fig. 7 Scatter plot of bootstrap PRESS/RSS values versus standard ones. $n = 20$, $m = 5$, $w = 0.05$

The determination of the number of the degrees of freedom is not straightforward in the case of partial least square regression (PLS). There are different assumptions and methods to calculate so-called pseudo degrees of freedom for PLS regression [22, 26, 27]. Unfortunately, we were not able to calculate pseudo degrees of freedom for these cases with PLS, because it needs more details on the data sets and the models, than it was accessible in the used literature sources of Q^2 and R^2. Anyway, we plot a second histogram in Fig. 6 (frequencies against percentiles of the F distribution ($p = 0.05$), where the pseudo degrees of freedom of the model was defined as 4*p causing $df_{\mathrm{PRESS}} = n - 4p$ and $df_{\mathrm{RSS}} = n - 4p$. The factor 4 was chosen as an extremely large difference between conventional degrees of freedom and pseudo degrees of freedom. Figure 6 clearly shows that though the majority of the models are acceptable about 80 models are wrong (percentile is above 95 %). The ambiguity problem of degrees of freedom in case of PLS (or principal component regression) cannot cause a serious limitation. The problem disappears asymptotically as '$n - p$' approximates n, whereas there is some uncertainty in the p value only. The test detected the same models as wrong ones even if the degree of freedom value was calculated by other multiplier than 4.

The uncertainty of PRESS/RSS data

We used the bootstrap method of resampling residuals [28] to assess the uncertainty and the confidence intervals of PRESS/RSS calculation. We generated 500 bootstrap samples for each of 500 random datasets corresponding to our test sets with given n, m and w. The bootstrap PRESS/RSS averages, the corresponding 2.5 and 97.5 percentiles are plotted versus traditional PRESS/RSS values of the data sets (Fig. 7).

In the case of low PRESS/RSS values, the bootstrap means are usually larger than the standard one, while at medium and large PRESS/RSS they are smaller. The correlation of the bootstrap averages and the standard ones are strong with less than unit slope. It means, the use of bootstrap average PRESS/RSS enhances the conservative feature of our proposed test. The lower and the upper confidence limits depend strongly on the datasets and they provide rather large uncertainty.

Conclusions

The *PRESS/RSS* ratio calculated from leave-one-out Q^2 and R^2 correlates well with the number of influential points in the training sets. Different identification methods on both simulated and experimental data support the conclusion. The correlation is strong enough, so we suggested a variance ratio test on the *PRESS/RSS* ratio to pre-estimate the presence of influential points in the training set, if degrees of freedoms (df_{PRESS} and df_{RSS}) are known. Some ambiguity in the degrees of freedom does not limit the applicability, because the test is conservative in this sense: i.e. it will detect only the "largely" contaminated models as wrong ones. However, any leave-one-out at a time diagnostic will fail, if influential points are shown up in groups (e.g. in pairs).

There are two possible applications of our results. Q^2, R^2, n and p are usually calculated and published in modeling, especially in QSAR studies. The rapid calculation of *PRESS/RSS* and the F test on it is a fast method to pre-estimate the presence of influential points or with other words the internal predictive character of a model. If a model fails in this test, it is worthwhile to consider changes in the training data. As many fortuitous QSAR models appear in the literature, editors and reviewers can check the submitted models easily: if a model fails the above variance ratio test the model has little generalization ability, if at all.

The other possibility is to apply the method for influential point detection, where not specific data is declared as an influential one, but the whole set is marked as influential point free or infected one. Of course, a hypothesis is necessary for the F test. Our examples suggested using 85–95 % percentiles as critical F values to make decisions between the H_0 hypothesis of influential observation free or H_a alternative hypothesis of presence of influential points.

References

1. Frank IE, Todeschini R (1994) The data analysis handbook, 1st edn. Elsevier, Amsterdam
2. Golbraikh A, Tropsha A (2002) J Mol Graph Model 20:269–276
3. Todeschini R, Consonni V, Mauri A, Pavan M (2004) Anal Chim Acta 515:199–208
4. Kubinyi H (2006) QSAR and molecular modelling in rational design of bioactive molecules. In: Yalcin I, Aki Sener E (eds) Proceedings of the 15th European symposium on QSAR and molecular modelling, Istanbul, Turkey, 2004. CADDD Society, Ankara, pp 30–33
5. Consonni V, Ballabio D, Todeschini R (2009) J Chem Inf Model 49:1669–1678
6. Roy PP, Paul S, Mitra I, Roy K (2009) Molecules 14:1660–1701
7. Consonni V, Ballabio D, Todeschini R (2010) J Chemom 24:194–201
8. Manvar AT, Pissurlenkar RRS, Virsodia VR, Upadhyay KD, Manvar DR, Mishra AK, Acharya HD, Parecha AR, Dholakia CD, Shah AK, Coutinhi EC (2010) Mol Divers 14:285–305
9. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) J Comput Aided Mol Des 17:241–253
10. Cramer RD, Wendt B (2007) J Comput Aided Mol Des 21:23–32
11. Jiménez-Contreras E, Torres-Salinas D, Bailón-Moreno R, Ruiz-Baños R, Delgado-López-Cózar E (2008) Scientometrics 79:201–218
12. Doweyko AM (2008) J Comput Aided Mol Des 22:81–89
13. Chirico N, Gramatica P (2011) J Chem Inf Model 51:2320–2335
14. Chirico N, Gramatica P (2012) J Chem Inf Model 52:2044–2058
15. Roy K, Mitra I, Ojha PK, Kar S, Das RN, Kabir H (2012) Chemom Intell Lab Syst 118:200–210
16. Bagheri A, Midi H, Ganjali M, Eftekhari S (2010) Appl Math Sci 4:1367–1386
17. Cook DR, Weisberg S (1982) Residuals and influence regression. Chapman & Hall, New York
18. Chatterjee S, Hadi AS (1986) Stat Sci 1:379–416
19. Rousseeuw P, Hubert M (1997) Lab statistical procedures and related topics. In: Dodge Y (ed) Papers from the 3rd international conference on lab-norm related methods Neuchatel 1997, Ins. Math Stat. Hayward, pp 201–214
20. Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York
21. Bevington PR (1969) Data reduction and error analysis for the physical sciences. McGraw-Hill Book Co., New York
22. van der Voet H (1999) J Chemom 13:195–208
23. Zhang X, Ding L, Sun Z, Song L, Sun T (2009) Chromatographia 70:511–518
24. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV (2005) J Comput Aid Mol Des 19:453–463
25. Dearden JC, Netzeva TI (2004) QSAR modelling of hERG potassium channel inhibition with low-dimensional descriptors. J Pharm Pharmacol 56(Suppl):S82–S82
26. Seipel HA, Kalivas JH (2004) J Chemom 18:306–311
27. Zhang L, Garcia-Munoz S (2009) Chemometr Intell Lab Syst 97:152–158
28. Fox J (2008) Applied regression analysis and generalized linear models, 2nd edn. SAGE Publications, Thousand Oaks