ELSEVIER

Analytica Chimica Acta 439 (2001) 169–191

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Tutorial

Detection of single influential points in OLS regression model building

Milan Meloun [a,*], Jiří Militký [b]

[a] *Department of Analytical Chemistry, Faculty of Chemical Technology, University of Pardubice, CZ-532 10 Pardubice, Czech Republic*
[b] *Department of Textile Materials, Technical University, CZ-461 17 Liberec, Czech Republic*

Received 11 December 2000; received in revised form 15 March 2001; accepted 6 April 2001

Abstract

Identifying outliers and high-leverage points is a fundamental step in the least-squares regression model building process. Various influence measures based on different motivational arguments, and designed to measure the influence of observations on different aspects of various regression results, are elucidated and critiqued here. On the basis of a statistical analysis of the residuals (classical, normalized, standardized, jackknife, predicted and recursive) and diagonal elements of a projection matrix, diagnostic plots for influential points indication are formed. Regression diagnostics do not require a knowledge of an alternative hypothesis for testing, or the fulfillment of the other assumptions of classical statistical tests. In the interactive, PC-assisted diagnosis of data, models and estimation methods, the examination of data quality involves the detection of *influential points*, outliers and high-leverages, which cause many problems in regression analysis. This paper provides a basic survey of the influence statistics of single cases combining exploratory analysis of all variables. The graphical aids to the identification of outliers and high-leverage points are combined with graphs for the identification of influence type based on the likelihood distance. All these graphically oriented techniques are suitable for the rapid estimation of influential points, but are generally incapable of solving problems with masking and swamping. The powerful procedure for the computation of influential points characteristics has been written in Matlab 5.3 and is available from authors.

Keywords: Influential observations; Regression diagnostics; Outliers; High-leverages; Diagnostic plot; Influence measures

Contents

Abstract 169
1. Introduction 170
2. Theoretical 172
  2.1. Terminology 172
    2.1.1. Linear regression model 172
    2.1.2. Conditions of the least-squares . 173
  2.2. Diagnostics for influential points detection 174
    2.2.1. Diagnostics based on residuals analysis 174
    2.2.2. Diagnostics based on the diagonal elements of the hat matrix 176
    2.2.3. Diagnostics based on residuals plots 176
    2.2.4. Diagnostics based on influence measures 178
    2.2.5. Diagnostics based on scalar influence measures 179

* Corresponding author. Tel.: +42-40-603-7026; fax: +42-40-603-7068.
E-mail address: milan.meloun@upce.cz (M. Meloun).

0003-2670/01/$ – see front matter
PII: S0003-2670(01)01040-6

3. Procedure 181
3.1. Procedure for regression model building 181
3.2. Software used 181
4. Illustrative examples 181
5. Conclusions 189
Acknowledgements 189
References 189

1. Introduction

Multiple linear regression model building is one of the standard problems solved in chemometrics [1]. The method of least-squares is generally used; this method, however, does not ensure that the regression model proposed is fully acceptable from the statistical and physical points of view. One of the main problems is the quality of data used for parameter estimation and model building. The term regression diagnostics has been introduced for a collection of methods for the identification of influential points and multicollinearity [2]; by regression diagnostics are understood methods of exploratory data analysis, for the analysis of influential points, and for the identification of violations of the assumptions of least-squares. In other words, regression diagnostics represent procedures for an examination of the *regression triplet* (data, model, method), i.e. procedures for the identification of (a) the data quality for a proposed model; (b) the model quality for a given set of data; (c) a fulfillment of all least-squares assumptions.

The detection, assessment, and understanding of influential points are the major areas of interest in regression model building. They are rapidly gaining recognition and acceptance by practitioners as supplements to the traditional analysis of residuals. Numerous influence measures have been proposed, and several books on the subject written is available, including these by Belsey et al. [2], Cook and Weisberg [3], Atkinson [4], Chatterjee and Hadi [5], Barnett and Lewis [6], Welsch [7], Welsch and Peters [8], Weisberg [9], Rousseeuw and Leroy [10], and Brownlee [11], published in the 1980s or earlier. A recent approach is given in a book by Atkinson [4].

The most commonly used graphical approaches in regression diagnostics, seen for example in Chatterjee and Hadi [5], are useful for distinguishing between "normal" and "extreme", "outlying" and "non-outlying" observations. An often used approach is the single-deletion method. One of the earliest methods for detecting influential observations was proposed by Gentleman and Wilk [12]. Cook [13], and Cook and Weisberg [14], made some algorithmic suggestions and used an upper bound for the Cook's index to identify influential subsets. Hawkins et al. [15] recommended the use of elemental sets to locate several outliers in multiple regression.

As single-case diagnostic measures can be written in terms of two fundamental statistics, the residual e_i and the diagonal elements of the hat matrix (leverage measure) H_{ii}, being expressed as the product of two general functions $f(n, m) \times g(H_{ii}, e_i)$. Thus, a reasonable strategy for diagnosing single-case influence is to jointly analyze the leverage and residual values to identify cases that are unusual in one or both, and then follow up by computing one or more influence measures for those cases. McCulloh and Meeter [16], Gray [17–21] and Hadi [22] proposed contour plots for summarizing influence information for single influential observations. A review of the standard diagnostics for one case and subset diagnostics has been written by Gray [17]. For evaluating the influence of cases there are two approaches: one is based on case deletion and the other on differentiation. In surveying the methodologies of influence analysis two major tools are often met with both, based on differentiation, in influence analysis in statistical modeling. One is Hampel's [23] influence function; the other is Cook's local influence [24,25].

Numerous influence measures have been proposed for assessing the influence of individual cases over the last two decades [12–74], and also represent a relatively new topic in the chemometrical literature of the last 10 years. Standardized residuals are also called studentized residuals in many references (e.g. [3,25,28,47]). Cook and Weisberg [3] refer to studentized residuals with internal studentization, in contrast to the external studentization of the cross-validatory or jackknife residuals according to Atkinson [4], and RSTUDENT by Belsey et al. [2] and SAS Institute, Inc. [47]. In any event, jackknife residuals are often used for the identification of outliers. Recursive residuals were introduced by Hedayat and Robson [29], Brown et al. [30], Galpin and Hawkins [31] and Quesenberry [32]. These residuals are constructed so that they are independent and identically distributed

when the model is correct. Rousseeuw and van Zomeren [42] used the minimum volume ellipsoid (MVE) for multivariate analysis to identify leverage points and the least median of squares (LMS) to distinguish between good and bad leverages. Seaver and Triantis [44] proposed a fuzzy clustering strategy to identify influential observations in regression; once the observations have been identified, the analyst can then compute regression diagnostics to confirm their degree of influence in regression. The hat matrix $\boldsymbol{H}$ has been studied by many authors from different perspectives. Dodge and Hadi [45] proposed a simple procedure for identifying high-leverages based on the upper and lower bounds for the diagonal and off-diagonal elements of $\boldsymbol{H}$. Cook and Critchley [46] used the theory of regression graphics based on central subspaces to construct a graphical solution to the long-standing problem of estimating the central subspace to identify outliers without specifying a model, and this method is also used in STAT [47]. Tanaka and Zhang [48] discussed the relationship between the influence analyses based on influence function (R-mode analysis) and on Cook's local influence (Q-mode analysis); generally, many such measures have certain disadvantages:

1. They highlight observations which influence a particular regression result but may fail to point out observations influential on other least-squares results.
2. There is no consensus among specialists as to which measures are optimal, and therefore several of the existing measures should be used.
3. Many existing measures are not invariant to location and scale in the response variable or nonsingular transformation of explanatory variables, and are unable to identify all of the observations that are influential on several regression characteristics.

Kosinski [49] has developed a new method for the detection of multiple multivariate outliers which is very resistent to high contamination (35–45%) of the data with outliers, and improved performance was also noted for data with smaller contamination fractions (15–20%) when outliers were situated closer to the "good" data. An adding-back model and two graphical methods with contours of constant measure values have been proposed by Fung [50] for studying multiple outliers and influential observations, using a logarithmic functional form for some influence measures having a better justification for plotting purposes. Barrett and Ling [51] have suggested using general high-leverage and outlier matrices for summarizing multiple-case influence measures. The concept of local influence was introduced by Cook [24] and motivated by other authors [52–70], for example, Gupta and Huang [53], Kim [54], Barrett and Gray [61], Muller and Mok [62] and Rancel and Sierra [70] have derived new criteria for detecting influential data as an alternative to Cook's measure. Barrett and Gray [61] have proposed a set of three simple, yet general and comprehensive, subset diagnostics referred to as leverage, residual, and interaction that have the desirable characteristics of single-case leverage and residual diagnostics. The proposed measures are the basis of several existing subset influence measures, including Cook's measure.

In chemometrics, several papers have addressed the robust analysis of rank deficient data. Liang and Kvalheim [56] have provided a tutorial on robust methods. The application of robust regression and the detection of influential observations is discussed by Singh [55]. Egan and Morgan [68] have suggested a method for finding outliers in multivariate chemical data. Regression diagnostics have been constructed from the residuals and diagonal elements of a hat matrix for detecting heteroscedasticity, influential observations and high-leverages in some papers [57,58,64,66–69,71–74,78]. Walczak and Massart [77,83] adapted the robust technique of ellipsoidal multivariate trimming (MTV) and LMS to make principal component regression (PCR) robust. Regression diagnostics as influence measures have been applied to the epidemiological study of oesophageal cancer in a high incidence area in China to identify the influence estimates of structures [50], to identify the influential observations in the calibration, to refine cell parameters from powder diffraction data [59], in medical chemistry [60], the groundwater flow model of a fractured-rock aquifer though regression [65], etc. Cook and Weisberg [87] detected influential points and argued that useful graphs must have a context induced by associated theory, and that a graph without a well-understood statistical context is hardly worth drawing. Singh [88] published the control-chart-type quantile–quantile (Q–Q) plot of Mahalanobis distance providing a formal graphical outlier identification. The procedure works effectively in correctly identify-

ing multiple univariate and multivariate outliers, and also in obtaining reliable estimates in multiple linear regression and principal component analyses; it identifies all regression outliers, and distinguishes between bad and good leverage points. Pell [89] has examined the use of robust principal component regression and iteratively reweighted partial least-squares (PLS) for multiple outlier detection in an infrared spectroscopic application, as in the case of multiple outliers the standard methods for outlier detection can fail to detect true outliers, and even mistakenly identify good samples as outliers. Walczak [90,91] published the program aiming to derive a clean subset from the contaminated calibration data in the presence of the multivariate outliers.

Statistical tests are needed for real data to decide how to use such data, in order to satisfy approximately the assumptions of hypothesis testing. Unfortunately, a bewilderingly large number of statistical tests, diagnostic graphs and residual plots have been proposed for diagnosing influential points, and it is the time to select those approaches that are suitable in chemometrics. This paper provides a survey of single point influence diagnostics, illustrated with data examples to show how they successfully characterize the influence of a group of cases even in the well-studied stackloss data [11].

2. Theoretical

2.1. Terminology

2.1.1. Linear regression model

Consider the standard linear regression model $y_i = \beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_m x_{m,i}+\varepsilon_i$, $i = 1,\ldots,n$, written in matrix notation $\boldsymbol{y} = \boldsymbol{X\beta}+\boldsymbol{\epsilon}$, where $\boldsymbol{y}$, the response (dependent) variable or regresand, is an $n \times 1$ vector of observations, $\boldsymbol{X}$ is a fixed $n \times m$ design regressors matrix of explanatory (independent) variables or regressors, predictors, factors ($n > m$), β is the $m \times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors, which are assumed to be independent and identically distributed with mean zero and an unknown variance σ^2. The regressor matrix $\boldsymbol{X}$ may contain a column vector of ones if there is a constant term β_0 to be estimated. Columns $\boldsymbol{x}_j$ geometrically define the m-dimensional co-ordinate system or the hyperplane L in n-dimensional Euclidean space. The vector $\boldsymbol{y}$ generally does not lie in this hyperplane L except where $\boldsymbol{\epsilon} = 0$. Using the least-squares estimation method provides the vector of fitted values $\hat{\boldsymbol{y}}_P = \boldsymbol{Hy}$, and the vector of residuals $\hat{\boldsymbol{e}} = (\boldsymbol{E}-\boldsymbol{H})\boldsymbol{Y}$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the so-called hat matrix, and $\boldsymbol{E}$ is the identity matrix. The quantity $s^2 = e^Te/(n-m)$ is an unbiased estimator of σ^2.

For geometric interpretation the residual vector $\hat{\boldsymbol{e}}$, for which the residual-square sum function $U(\boldsymbol{b})$ is minimal, lies in the $(n-m)$-dimensional hyperplane $L^\perp$, being perpendicular to the hyperplane L,

$$U(\boldsymbol{b}) = \sum_{i=1}^{n}(y_i - \hat{y}_{P,i})^2 = \sum_{i=1}^{n}\left[y_i - \sum_{j=0}^{m} x_{ij}b_j\right]^2 \approx \text{minimum} \quad (1)$$

The perpendicular projection of $\boldsymbol{y}$ into hyperplane L can be made using the projection matrix $\boldsymbol{H}$, and may be expressed [1] by $\hat{\boldsymbol{y}}_P = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{Hy}$, or the projection matrix $\boldsymbol{P}$ for perpendicular projection into a hyperplane $L^\perp$ that is orthogonal to hyperplane L and is $\boldsymbol{P} = \boldsymbol{E}-\boldsymbol{H}$. With the use of these two projection matrices, $\boldsymbol{H}$ and $\boldsymbol{P}$, the total decomposition of vector $\boldsymbol{y}$ into two orthogonal components may be written as $\boldsymbol{y} = \boldsymbol{Hy}+\boldsymbol{Py} = \hat{\boldsymbol{y}}_P + \hat{\boldsymbol{e}}$. The vector $\boldsymbol{y}$ is decomposed into two mutually perpendicular vectors, the prediction vector $\hat{\boldsymbol{y}}_p$ and the vector of residuals $\hat{\boldsymbol{e}}$.

It is well known that $(D(\hat{\boldsymbol{y}}_P) = \sigma^2)\,\boldsymbol{H}$ and $D(\hat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{E}-\boldsymbol{H})$. The Tukey's hat matrix $\boldsymbol{H}$ is commonly referred to as the "hat" matrix (because it puts the "hat" on $\boldsymbol{y}$), but is also known as the "projection" matrix (because it gives the orthogonal projection of $\boldsymbol{y}$ onto the space spanned by the columns of $\boldsymbol{X}$) or the "prediction" matrix (because $\boldsymbol{Hy}$ gives the vector of predicted values). The matrix $\boldsymbol{H}$ is also known as the "high-leverage" matrix, because the ith fitted value of the predictor vector $\hat{\boldsymbol{y}}_P$ can be written as $\hat{\boldsymbol{y}}_P = \sum H_{ij}y_i$, where H_{ij} is the ijth element of $\boldsymbol{H}$. Thus, H_{ij} is the high-leverage (weight) of the jth observation y_j, in determining the ith fitted prediction value $\hat{y}_{P,i}$. It can also be seen that H_{ii} represents the high-leverage of the ith observation y_i in determining its own prediction value $\hat{y}_{P,i}$. Thus, observations with large H_{ij} can excerpt an undue influence on least-squares results, and it is thus important for the data analysis to be able to identify such observations. The square of vector $\hat{\boldsymbol{e}}$ length is consistent with criterion $U(\boldsymbol{b})$ of the least-squares

method, so that the estimates of model parameters $\boldsymbol{b}$ minimizes a function $U(\boldsymbol{b})$.

2.1.2. Conditions of the least-squares

There are some basic conditions necessary for the least-squares method LS to be valid [1]:

1. The regression parameters β are not bounded. In chemometric practice, however, there are some restrictions on the parameters, based on their physical meaning.
2. The regression model is linear in the parameters, and an additive model of the measurement errors is valid, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$.
3. The matrix of non-random controllable values of the regressors $\boldsymbol{X}$ has a column rank equal to m. This means that the all pairs $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ are not collinear vectors. This is the same as saying that the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is a symmetric regular invertible matrix with a non-zero determinant, i.e. plane L is m-dimensional, and vector $\boldsymbol{Xb}$ and the parameter estimates $\boldsymbol{b}$ are unambiguously determined.
4. The mean value of the random errors ε_i is zero; $E(\varepsilon_i) = 0$. This is automatically valid for all regression type models containing intercept. For models without intercept the zero mean of errors has to be tested.
5. The random errors ε_i have constant and finite variance, $E(\varepsilon_i^2) = \sigma^2$. The conditional variance σ^2 is also constant and therefore the data are said to be homoscedastic.
6. The random errors ε_i are uncorrelated, i.e. $\mathrm{cov}(\varepsilon_i, \varepsilon_i) = E(\varepsilon_i, \varepsilon_i) = 0$. When the errors follow the normal distribution they are also independent. This corresponds to independence of the measured quantities $\boldsymbol{y}$.
7. The random errors $\boldsymbol{\epsilon}_i$ have a normal distribution $N(0, \sigma^2)$. The vector $\boldsymbol{y}$ then has a multivariate normal distribution with mean $\boldsymbol{X}\beta$ and covariance matrix $\sigma^2\boldsymbol{E}$.

When the first six conditions are met, the parameter estimates $\boldsymbol{b}$ found by minimization of a least-squares are the best linear unbiased estimate (BLUE) of the regression parameters β:

1. The term best estimates $\boldsymbol{b}$ means that any linear combination of these estimates has the smallest variance of all linear unbiased estimates. That is, the variance of the individual estimates $D(b_j)$ are the smallest of all the possible linear unbiased estimates (the Gauss–Markov theorem).
2. The term linear estimates means that they can be written as a linear combination of measurements $\boldsymbol{y}$ with weights Q_{ij} which depend only on the location of variables x_j, $j = 1, \ldots, m$, and $\boldsymbol{Q} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ for the weight matrix; thus $b_j = \sum_{i=1}^{n} Q_{ij} y_i$. Each estimate b_j is the weighted sum of all measurements. Also, the estimates $\boldsymbol{b}$ have an asymptotic multivariate normal distribution with covariance matrix $D(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$. When condition (7) is valid, all estimates $\boldsymbol{b}$ have a normal distribution, even for finite sample size n.
3. The term unbiased estimates means that $E(\boldsymbol{\beta}-\boldsymbol{b}) = 0$ and the mean value of an estimate vector $E(\boldsymbol{b})$ is equal to a vector of regression parameters $\boldsymbol{\beta}$. It should be noted that there exist biased estimates, the variance of which can be smaller than the variance of estimates $D(b_j)$.

Various test criteria for testing regression model quality may be used [1]. One of the most efficient seems to be the mean quadratic error of prediction (MEP), being defined by the relationship

$$\mathrm{MEP} = \frac{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_{(i)})^2}{n}$$

where $\boldsymbol{b}_{(i)}$ is the estimate of regression parameters when all points except the ith one were used and x_i is the ith row of matrix $\boldsymbol{X}$. The statistic MEP uses a prediction $\hat{y}_{\mathrm{P},i}$ from an estimate constructed without including the ith point. Another mathematical expression is $\mathrm{MEP} = \sum_{i=1}^{n} \hat{e}_i^2/(1 - H_{ii}^2)n$. For large sample sizes n the element H_{ii} tends to zero ($H_{ii} \approx 0$) and then $\mathrm{MEP} = U(\boldsymbol{b})/n$. The MEP can be used to express the predicted determination coefficient (known in chemometrics as the CVR square or as the Q square),

$$\hat{R}_{\mathrm{P}}^2 = 1 - \frac{n \times \mathrm{MEP}}{\sum_{i=1}^{n} y_i^2 - n \times \bar{y}^2}$$

Another statistical characteristic in quite general use is derived from information and entropy theory [12], and is known as the Akaike information criterion, $\mathrm{AIC} = n \ln(U(\boldsymbol{b})/n) + 2m$. The most suitable model is the one which gives the lowest value of the mean quadratic error of prediction MEP, the Akaike information

criterion AIC and the highest value of the predicted determination coefficient, $\hat{R}^2_{\mathrm{P}}$.

2.2. Diagnostics for influential points detection

Data quality has a strong influence on any proposed regression model. Examination of data quality involves detection of the influential points, which cause many problems in regression analysis by shifting the parameter estimates or increasing the variance of the parameters. According to the terminology proposed by Rousseeuw [10], influential points may alternatively be classified according to data location into either: (i) outliers (denoted on graphs by the letter O) which differ from the other points in y-axis value; (ii) high-leverage points, also called extremes (denoted on graphs by the letter E), which differ in the structure of $\boldsymbol{X}$ variables; or (iii) both O and E (denoted by letters O and E), standing for a combination of outliers and high-leverages together. Outliers can be identified by examination of the residuals relatively simply and this can be done once the regression model is constructed. Identification of all high-leverage points is based on the $\boldsymbol{X}$ space only, and takes into account no information contained in y, as high-leverages are found from the diagonal elements H_{ii} of the projection hat matrix $\boldsymbol{H}$.

If the data contains a single outlier or high-leverage point, the problem of identifying such a point is relatively simple. If the data contains more than one outlier or high-leverage (which is likely to be the case in most data), the problem of identifying such points becomes more difficult, due to masking and swamping effects. Masking occurs, when an outlying subset goes undetected because of the presence of another, usually adjacent, subset. Swamping occurs when "good" points are incorrectly identified as outliers because of the presence of another, usually remote, subset of points.

Influence statistics are to be used as diagnostic tools for identifying the observations having the greatest impact on regression results. Although some of the influence measures resemble test statistics, they are not to be interpreted as tests of significance for influential observations.

The large number of influence statistics that can be generated can cause confusion; one should thus concentrate on that diagnostic tool that measures the impact on the quantity of primary interest.

There are various diagnostic measures designed to detect individual cases that differ from the bulk of data and which may be classified according to [34,77] into four groups: (i) diagnostics based on the prediction matrix, (ii) diagnostics based on residuals, (iii) diagnostics based on the volume of confidence elipsoids, and (iv) diagnostics based on influence function. Each diagnostic measure is designed to detect a specific phenomenon in the data. They are closely related, as they are functions of the same basic building blocks in model construction, i.e. of various types of residuals $\hat{\boldsymbol{e}}$ and the elements of the hat matrix $\boldsymbol{H}$. Since there is a great deal of redundancy in them, the diagnostics within the same class can vary little, and the analyst does not have to consider all of them. The authors wish to show some efficient shortcuts in statistical diagnostic tools which, according to their experience, can pinpoint the influential points.

2.2.1. Diagnostics based on residuals analysis

Analysis of various types of regression residuals, or of some transformation of such residuals, is very useful for detecting inadequacies in the model or influential points in data. The true errors in the regression model are assumed to be normally and independently distributed random variables with zero mean and variance $\varepsilon \approx N(0, \boldsymbol{I}\sigma^2)$.

1. *Ordinary residuals* $\hat{e}_i$ are defined by $\hat{e}_i = y_i - x_i^{\mathrm{T}}\boldsymbol{b}$, where x_i is the ith row of matrix $\boldsymbol{X}$. Classical analysis is based on the incorrect assumption that residuals are good estimates of errors ε_i; the reality is more complex, residuals $\hat{\boldsymbol{e}}$ being a projection of vector $\boldsymbol{y}$ into a subspace of dimension $(n - m)$,

$$\hat{\boldsymbol{e}} = \boldsymbol{P}\boldsymbol{y} = \boldsymbol{P}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{P}\boldsymbol{\epsilon} = (\boldsymbol{E} - \boldsymbol{H})\boldsymbol{\epsilon}$$

and therefore for the ith residual is valid:

$$\hat{e}_i = (1 - H_{ii})y_i - \sum_{j \neq i}^{n} H_{ii}y_i$$

$$= (1 - H_{ii})\varepsilon_i - \sum_{j \neq i}^{n} H_{ij}\varepsilon_j$$

Each residual $\hat{e}_i$ is a linear combination of all errors ε_i. The distribution of residuals depends on (i) the error distribution, (ii) the elements of the projection matrix $\boldsymbol{H}$, (iii) the sample size n. Because

the residual $\hat{e}_i$ represents a sum of random quantities with bounded variance, the supernormality effect appears for small sample sizes: "Even when the errors ε do not have a normal distribution, the distribution of residuals is close to normal". In small samples the elements of the projection matrix $\boldsymbol{H}$ are larger and the contribution of the sum is prevalent; the distribution of this sum is closer to a normal one than the distribution of errors ε. For large sample sizes, where $1/n \approx 0$, we find that $\hat{e}_i \approx \varepsilon_i$ and analysis of the residual distribution gives direct information about the distribution of errors. Ordinary residuals have non-constant variance and may not often indicate strongly deviant points. The common practice of chemometrics programs for the statistical analysis of residuals is to use for examination some statistical characteristics of ordinary residuals, such as the mean, the variance, the skewness and the kurtosis. As has been shown above, for small and moderate sample sizes the ordinary residuals are not good for diagnostics or the identification of influential points.

2. *Normalized residuals or scaled residuals* $\hat{e}_{\mathrm{N},i} = \hat{e}_i/\hat{\sigma}$ are often recommended in chemometrics. It is falsely assumed that these residuals are normally distributed quantities with zero mean and variance equal to one, $\hat{e}_{\mathrm{N},i} \approx N(0,1)$, but in reality these residuals have non-constant variance. When normalized residuals are used, the rule of 3σ is classically recommended: quantities with $\hat{e}_{\mathrm{N},i}$ of magnitude greater than $\pm 3\sigma$ are classified as outliers: this approach is quite misleading, and may cause wrong decisions to be taken regarding data.
3. *Standardized residuals or internally studentized residuals* $\hat{e}_{\mathrm{S},i} = \hat{e}_i/(\hat{\sigma}\sqrt{1-H_{ii}})$ exhibit constant unit variance, and their statistical properties are the same as those of ordinary residuals. The standardized residuals behave much like a Student's t random variable except for the fact that the numerator and denominator of $\hat{e}_{\mathrm{S},i}$ are not independent.
4. *Jackknife residuals or externally studentized residuals* $\hat{e}_{\mathrm{J},i} = \hat{e}_{\mathrm{S},i}\sqrt{(n-m-1)/(n-m-\hat{e}^2_{\mathrm{S},i})}$, are residuals which with an assumption of normality of errors have a Student distribution with $(n-m-1)$ degrees of freedom. Belsey et al. [2] have suggested standardizing each residual with an estimate of its standard deviation that is independent of the residual. This is accomplished by using, as the estimate of σ^2 for the ith residual, the residual mean square from an analysis where that observation has been omitted. This variance is labeled $s^2_{(i)}$, where the subscript in parentheses indicates that the ith observation has been omitted for the estimate of σ^2; the result is a jackknife residual, also called a fully studentized residual. It is distributed as Student's t with $(n-m-1)$ degrees of freedom when the normality of errors ε holds. As with $\hat{e}_i$ and $\hat{e}_{\mathrm{S},i}$, the $\hat{e}_{\mathrm{J},i}$ are not independent of each other. Belsey et al. [2] have shown that $s_{(i)}$ and jackknife residuals can be obtained from ordinary residuals without rerunning the regression with the observation omitted. The standardized residuals $\hat{e}_{\mathrm{S},i}$ are called studentized residuals in many references (e.g. [3,25,28,47]). Cook and Weisberg [3] refer to $\hat{e}_{\mathrm{S},i}$ as the studentized residual with internal studentization, in contrast to the external studentization of $\hat{e}_{\mathrm{J},i}$. The $\hat{e}_{\mathrm{J},i}$ are called cross-validatory or jackknife residuals by Atkinson [4] and RSTUDENT by Belsey et al. [2] and SAS Institute, Inc. [47]. In any event jackknife residuals are often used for the identification of outliers. The jackknife residual [24] examines the influence of individual points on the mean quadratic error of prediction. When the condition $\hat{e}^2_{\mathrm{J},i} \leq F_{1-\alpha/n}(1, n-m-1, 0.5)$ is valid, no influential points are present in the data. Here, $F_{1-\alpha/n}(1, n-m-1, 0.5)$ means the $100(1-\alpha/n)\%$ quantile of the non-central F-distribution with non-centrality parameters 0.5 and 1, and $(n-m-1)$ degrees of freedom. An approximate rule may be formulated: strongly influential points have squared jackknife residuals $\hat{e}^2_{\mathrm{J},i}$ greater than 10. In the case of high-leverage points, however, these residuals do not give any indication.
5. *Predicted residuals or cross-validated residuals* $\hat{e}_{\mathrm{P},i} = (\hat{e}_i/(1-H_{ii})) = y_i - x_i\boldsymbol{b}_{(i)}$ are equal to shift C in the equation $\boldsymbol{y} = \boldsymbol{Xb} + C\boldsymbol{i} + \boldsymbol{\epsilon}$, where $\boldsymbol{i}$ is the identity vector with the ith element equal to one, and other elements equal zero. This model expresses not only the case of an outlier where C is equal to the value of deviation, but also the case of a high-leverage point $C = \boldsymbol{d}_i^{\mathrm{T}}\boldsymbol{\beta}$ where $\boldsymbol{d}_i$ is the vector of the deviation of the individual x components of the ith point.
6. *Recursive residuals* have been described by Hedayat and Robson [29], Brown et al. [30].

Galpin and Hawkins [31] and Quesenberry [32]. These residuals are constructed so that they are independent and identically distributed when the model is correct. They are computed from a sequence of regression starting with a base of m observations (m being the number of parameters to be estimated), and adding one observation at each step. The regression equation computed at each step is used to compute the residual for the next observation to be added. This sequence continues until the last residual has been computed. There will be $(n - m)$ recursive residuals; the residual from the first m observations will be zero, $\hat{e}_{R,i} = 0, i = 1, \ldots, m$, and then the recursive residual is defined as

$$\hat{e}_{R,i} = \frac{y_i - x_i \boldsymbol{b}_{i-1}}{\sqrt{1 + x_i (\boldsymbol{X}_{i-1}^T \boldsymbol{X}_{i-1})^{-1} x_i^T}}, \quad i = m+1, \ldots, n \tag{5}$$

where $\boldsymbol{b}_{i-1}$ are estimates obtained from the first $(i-1)$ points. The recursive residuals are mutually independent and have constant variance σ^2. Each is explicitly associated with a particular observation and, consequently, recursive residuals seem to avoid some of the "spreading" of model defects that occurs with ordinary residuals. They allow the identification of any instability in a model, such as in time or autocorrelation, and are often used in normality tests or in tests of the stability of regression parameters β.

The various types of residuals differ in their suitability for diagnostic purposes: (i) standardized residuals $\hat{e}_{S,i}$ serve for the identification of heteroscedasticity only; (ii) jackknife residuals $\hat{e}_{J,i}$ or predicted residuals $\hat{e}_{P,i}$ are suitable for the identification of outliers; (iii) recursive residuals $\hat{e}_{R,i}$ are used for the identification of autocorrelation and normality testing.

2.2.2. Diagnostics based on the diagonal elements of the hat matrix

Since the introductory paper by Hoaglin and Welsch [79], the hat matrix $\boldsymbol{H}$ has been studied by many authors from different perspectives. Hoaglin and Welsch [79] suggested declaring observations with $H_{ii} > 2m/n$ as high-leverage points. For regression models containing an intercept and a full rank of $\boldsymbol{X}$, $\sum_{i=1}^{n} H_{ii} = m$ is valid; the mean value of the diagonal elements is therefore n/m. Obviously, this cut-off point will fail to nominate any observation when $n \leq 2m$, because $0 \leq H_{ii} \leq 1$. In some cases it is useful to compute the extension of matrix $\boldsymbol{X}$ by a vector $\boldsymbol{y}$ to give matrix $\boldsymbol{X}_m = (\boldsymbol{X}|\boldsymbol{y})$. It can be shown that $H_m = H + \hat{e}\hat{e}^T/\hat{e}^T\hat{e}$. The diagonal elements of $\boldsymbol{H}_m$ contain information about leverage and outliers because $H_{m,ii} = H_{ii} + \hat{e}_i^2/[(n-m)\hat{\sigma}^2]$.

2.2.3. Diagnostics based on residuals plots

A variety of plots have been widely used in regression diagnostics for the analysis of residuals ([3–5,80–82] among others). Three types of plots can indicate inaccuracy in a proposed model, and some trends, heteroscedasticity or influential points in data: Plot type I is a plot of some kind of residuals against the index i; Plot type II is a plot of residuals against the independent variable x_{ji}; Plot type III is a plot of residuals against the predicted value $\hat{y}_{P,i}$. Fig. 1 shows possible graph shapes which can occur in plots of residuals. If the graph shape is a random pattern (Fig. 1a), the least-squares assumption is correct. A systematic pattern indicates that the approach is incorrect in some way: the sector pattern in graph types I–III indicates heteroscedasticity in the data (Fig. 1b) while a band pattern in graph types I and II indicates some error in calculation or the absence of x_j in the model, (Fig. 1c). A band pattern may be also caused by outlying points or, in type III, by a missing intercept term in the regression model. In all three graph types (I–III) a non-linear pattern indicates that the model proposed is incorrect (Fig. 1d).

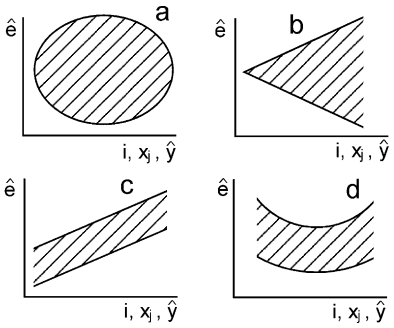

Fig. 1. Possible shapes of residual plots: (a) random pattern shape, (b) sector pattern shape, (c) band shape, (d) non-linear curved band shape.

It should be noted that the plot $\hat{e}_i$ against dependent variable y_i is not useful, because the two quantities are strongly correlated. The smaller the correlation coefficient, the more linear is this plot.

For the identification of influential points, i.e. outliers (denoted in plots by the letter O) and high-leverages (denoted in plots by the letter E), various types of residuals are combined with the diagonal elements H_{ii} of the projection hat matrix $\boldsymbol{H}$, cf. p. 72 in [1]. Most of the characteristics of influential points may be expressed in the form $K(m, n) \times f(H_{ii}, \hat{e}^2_{\mathrm{N},i})$, where $K(m, n)$ is a constant dependent only on m and n.

1. The graph of predicted residuals [76] has the predicted residuals $\hat{e}_{\mathrm{P},i}$ on the x-axis and the ordinary residuals $\hat{e}_i$ on the y-axis. The high-leverage points are easily detected by their location, as they lie outside the line $y = x$, and are located quite far from this line. The outliers are located on the line $y = x$, but far from its central pattern, (Fig. 2).
2. The Williams graph [76] has the diagonal elements H_{ii} on the x-axis and the jackknife residuals $\hat{e}_{\mathrm{J},i}$ on the y-axis. Two boundary lines are drawn, the first for outliers, $y = t_{0.95}(n - m - 1)$ and the second for high-leverages, $x = 2m/n$. Note that $t_{0.95}(n - m - 1)$ is the 95% quantile of the Student distribution with $(n - m - 1)$ degrees of freedom, (Fig. 3).
3. The Pregibon graph [75] has the diagonal elements H_{ii} on the x-axis and the square of normalized residuals $\hat{e}^2_{\mathrm{N},i}$ on the y-axis. Since the expression $E(H_{ii} + \hat{e}^2_{\mathrm{N},i}) = (m + 1)/n$ is valid for this graph, two different constraining lines can be drawn: $y = -x + 2(m + 1)/n$, and $y = -x + 3(m + 1)/n$. To distinguish among influential points the following rules are used: (a) a point is strongly influential if it is located above the upper line; (b) a point is influential if it is located between the two lines. The influential point can be either an outlier or a high-leverage point, (Fig. 4).
4. The McCulloh and Meeter graph [16] has $\ln[H_{ii}/(m(1 - H_{ii}))]$ on the x-axis and the logarithm of square of the standardized residuals $\ln(\hat{e}^2_{\mathrm{S},i})$ on the y-axis. In this plot the solid line drawn represents the locus of points with identical influence, with slope -1. The 90% confidence

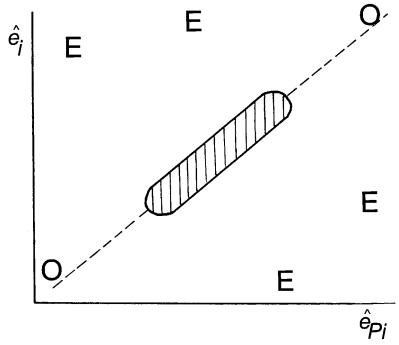

Fig. 2. Graph of predicted residuals: E means a high-leverage point and O an outlier; outliers are far from the central pattern on the line $y = x$.

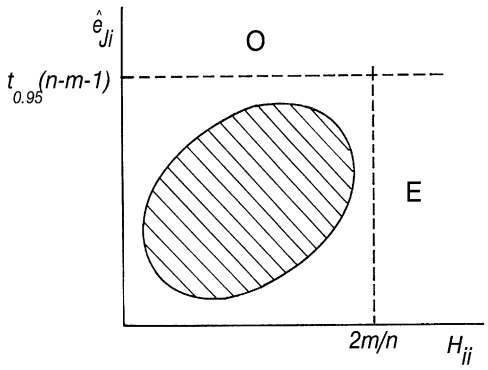

Fig. 3. Williams graph: E means the leverage point and O the outlier; the first line is for outliers, $y = t_{0.95}(n - m - 1)$, the second line is for high-leverages, $x = 2m/n$.

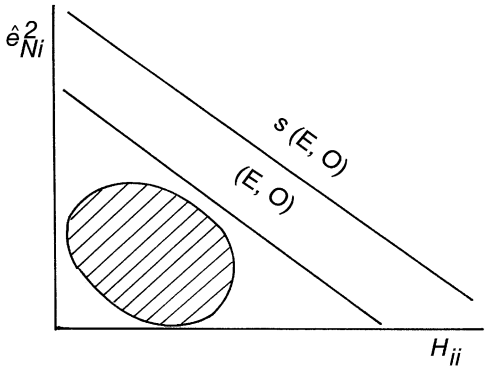

Fig. 4. Pregibon graph: (E, O) are influential points, and s(E, O) are strongly influential points; two constraining lines are drawn, $y = -x + 2(m + 1)/n$, and $y = -x + 3(m + 1)/n$, the strongly influential point is above the upper line; the influential point is between the two lines.

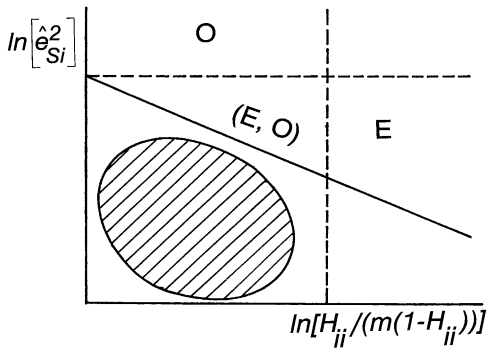

Fig. 5. McCulloh and Meeter graph: E means a high-leverage point and O an outlier, (E, O) an influential point; the 90% confidence line is for outliers, $y = -x - \ln F_{0.95}(n-m, m)$ while the boundary for high-leverages is $x = \ln[2/(n-m) \times (t^2_{0.95}(n-m)]$.

line is defined by $y = -x - \ln F_{0.9}(n-m, m)$. The boundary line for high-leverage points is defined as $x = \ln[2/(n-m) \times (t^2_{0.95}(n-m)]$ where $t^2_{0.95}(n-m)$ is the 95% quantile of the Student distribution with $(n-m-1)$ degrees of freedom (Fig. 5).

5. The Gray's L–R graph [18] has the diagonal elements H_{ii} on the x-axis and the squared normalized residuals $\hat{e}^2_{\mathrm{N},i} = \hat{e}^2_i / U(b)$ on the y-axis. All the points will lie under the hypotenuse of a triangle with the right angle in the origin of the two axes and the hypotenuse defined by the limiting equality $H_{ii} + \hat{e}^2_{\mathrm{N},i} = 1$. Contours of the same critical influence are plotted in the Gray's L–R graph, and the locations of individual points are compared with them. It may be determined that the contours are hyperbolic as described by the equation $y = (2x - x^2 - 1)/(x(1-K) - 1)$ where $K = n(n-m-1)/(c^2 m)$ and c is a constant. For $c = 2$, the constant K corresponds to the limit $2/\sqrt{m/n}$. The constant c is usually equal to 2, 4 or 8, (Fig. 6).
6. The index graph has the order index i on the x-axis and the residuals $\hat{e}_{\mathrm{S},i}, \hat{e}_{\mathrm{P},i}, \hat{e}_{\mathrm{J},i}, \hat{e}_{\mathrm{R},i}$, or the diagonal elements H_{ii}, or estimates b_i on the y-axis. It indicates the suspicious points which could be influential, i.e. outliers or high-leverages.
7. The rankit graph (Q–Q plot) has the quantile of the standardized normal distribution u_{P_i} for $P_i = i/(n+1)$ on the x-axis and the ordered residuals $\hat{e}_{\mathrm{S},i}, \hat{e}_{\mathrm{P},i}, \hat{e}_{\mathrm{J},i}, \hat{e}_{\mathrm{R},i}$, i.e. increasingly ordered values of various types of residuals on the y-axis.

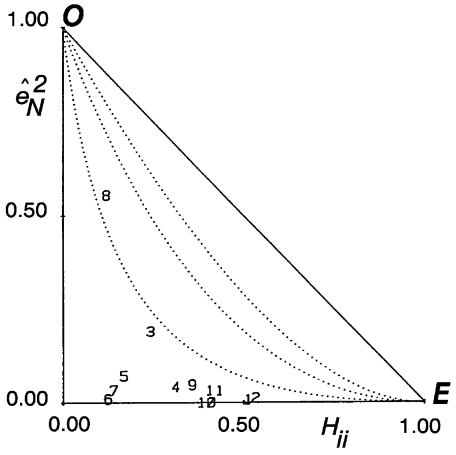

Fig. 6. L–R graph: E means a high-leverage point and O an outlier, and digits in the triangle stand for the order index i of the response (dependent) variable y_i; points towards to the upper part are outliers while towards the right angle of triangle are high-leverages.

2.2.4. Diagnostics based on influence measures

Influential points can affect regression characteristics differently. Points affecting the prediction, $\hat{y}_{\mathrm{P},i}$, for example, may not affect the variance parameter. The degree of influence of individual points can be classified according to the characteristics that are affected. Numerical diagnostics of influential points may be divided by to two principal approaches:

1. Examination of changes which occur when certain points are omitted.
2. Examination of changes which occur when the variances of influential points are inflated. Let the model of inflated variance be assumed; in this model random errors exhibit normal distribution with a constant variance σ^2, i.e. $N(0, \sigma^2)$ except that the ith influential point has a normal distribution $N(0, \sigma^2/w_i)$. The weight, also called the perturbance parameter, lies in the interval $0 < w_i < 1$.

If $\boldsymbol{b}(w_i)$ denotes the parameter estimate where the variance of the ith influential error is equal to σ^2/w_i, then the following expression is valid,

$$\boldsymbol{b}(1) - \boldsymbol{b}(w_i) = \frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} x_i (1-w_i)\hat{e}_i}{1 - (1-w_i)H_{ii}} \qquad (10)$$

where x_i is the ith row of matrix $\boldsymbol{X}$ which contains x components of the ith point.

There are two limiting cases to the perturbance parameter:

1. For $w_i = 1$ the $\boldsymbol{b}(1)$ are equal to parameter estimates obtained by the LS method.
2. For $w_i = 0$: Eq. (10) leads to the relationship $\boldsymbol{b}(1) - \boldsymbol{b}(0) = \boldsymbol{b} - \boldsymbol{b}_{(i)}$, where $\boldsymbol{b}_{(i)}$ is the estimate reached by the least-squares method with the use of all points except the ith one.

Leaving out the ith point is therefore the same as the case when this point has unbounded infinite variance. To express the sensitivity of parameter estimates to the perturbance parameter w_i, the sensitivity function $\delta\boldsymbol{b}(w_i)/\delta w_i$ can be used,

$$\frac{\delta\boldsymbol{b}(w_i)}{\delta w_i} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}x_i\hat{e}_i\frac{s_A + (1-w_i)H_{ii}}{s_A^2} \tag{11}$$

where $s_A = 1 - (1 - w_i)H_{ii}$. The following types of sensitivity function of parameter estimates can be defined:

1. *The Jackknife influence function JC_i:* the sensitivity function of parameter estimates at the value $w_i = 0$ is given by

$$\left|\frac{\delta\boldsymbol{b}(w_i)}{\delta w_i}\right|_{w_i=0} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}x_i\frac{\hat{e}_i}{(1-H_{ii})^2} = \frac{\mathrm{JC}_i}{n-1} \tag{12}$$

   The term JC_i is the jackknife influence function. It is related to the sensitivity function of parameter estimates, i.e. lies in the vicinity of $\boldsymbol{b}(0)$ in cases where the ith point is omitted, because $\boldsymbol{b}(0) = \boldsymbol{b}_{(i)}$.
2. *The empirical influence function EC_i:* the sensitivity function of parameter estimates at the value $w_i = 1$ is given by

$$\left|\frac{\delta\boldsymbol{b}(w_i)}{\delta w_i}\right|_{w_i=1} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}x_i\hat{e}_i = \frac{\mathrm{EC}_i}{n-1} \tag{13}$$

   The term EC_i is the empirical influence function. It is related to the sensitivity function of parameter estimates, i.e. lies in the vicinity of $\boldsymbol{b}(1)$.
3. *The sample influence function SC_i*: the sample influence function is directly proportional to the change in the vector of parameter estimates when the ith point is left out. With the use of Eq. (11) we can write

$$\mathrm{SC}_i = n(\boldsymbol{b} - \boldsymbol{b}_{(i)}) = n(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}x_i\frac{\hat{e}_i}{1-H_{ii}} \tag{14}$$

   All three influence functions differ only in the single term $(1 - H_{ii})$, so they are not identically sensitive to the presence of high-leverage points, for which $H_i \to 1$. The disadvantage of all these influence functions is that they are m-dimensional vectors. Their components define the influence of the ith point on the estimate of the jth parameter.

2.2.5. Diagnostics based on scalar influence measures

Proper normalization of influence functions [23] leads to scalar measures. These measures express the relative influence of the given point on all parameter estimates.

1. The Cook measure D_i directly expresses the relative influence of the ith point on all parameter estimates and has the form

$$D_i = \frac{(\boldsymbol{b} - \boldsymbol{b}_{(i)})^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{b}_{(i)})}{m \times \hat{\sigma}^2} = \frac{\hat{e}_{\mathrm{S},i}^2}{m} \times \frac{H_{ii}}{1-H_{ii}} \tag{15}$$

   It is related to the confidence ellipsoid of the estimates but is really a shift of estimates when the ith point is left out. It is approximately true that when $D_i > 1$, the shift is greater than the 50% confidence region, so the relevant point is rather influential. Another interpretation of D_i is based on the Euclidean distance between the prediction vector $\hat{\boldsymbol{y}}_{\mathrm{P}}$ and the prediction vector $\hat{y}_{\mathrm{P},(i)}$ estimated when the ith point is left out. The Cook measure D_i expresses the influence of the ith point on the parameter estimate $\boldsymbol{b}$ only; when the ith point does not affect $\boldsymbol{b}$ significantly, the value of D_i is low. Such a point, however, can strongly affect the residual variance $\hat{\sigma}^2$.
2. The Atkinson measure A_i enhances the sensitivity of distance measures to high-leverage points. This modified version of Cook's measure D_i suggested by Atkinson [4] is even more closely related to Belsey's DFFITS_i and has the form

$$A_i = |\hat{e}_{\mathrm{J},i}| \times \sqrt{\frac{n-m}{m} \times \frac{H_{ii}}{1-H_{ii}}} \tag{16}$$

This measure is also convenient for graphical interpretation; Atkinson recommends that absolute values of A_i be plotted in any of the ways customary for residuals. With designed experiments, usually $H_{ii} = m/n$, and the Atkinson measure A_i is numerically equal to the jackknife residual $\hat{e}_{\mathrm{J},i}$.

3. The Belsey DFFITS$_i$ measure, also called Welsch-Kuh's distance [2], is obtained by the normalization of the sample influence function, and using the variance estimate $\hat{\sigma}^2_{(i)}$ obtained from estimates $\boldsymbol{b}_{(i)}$. This measure has the form

$$\mathrm{DFFITS}_i^2 = \hat{e}^2_{\mathrm{J},i} \times \frac{H_{ii}}{1 - H_{ii}} \tag{17}$$

Belsey et al. [2] suggest that the ith point is considered to be significantly influential on prediction $\hat{\boldsymbol{y}}_{\mathrm{P}}$ when DFFITS$_i$ is larger in absolute value than $2\sqrt{m/n}$.

4. The Anders–Pregibon diagnostic AP$_i$ [23] expresses the influence of the ith point on the volume of the confidence ellipsoid

$$\mathrm{AP}_i = \frac{\det(\boldsymbol{X}^{*\mathrm{T}}_{(i)}\boldsymbol{X}^*_{(i)})}{\det(\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{X}^*)} \tag{18}$$

where $\boldsymbol{X}^* = (x/y)$ is the matrix having as least column the vector $\boldsymbol{y}$. The diagnostic AP$_i$ is related to the elements of the extended projection matrix $\boldsymbol{H}^*$ by the expression $\mathrm{AP}_i = 1 - H_{ii} - \hat{e}^2_{\mathrm{N},i} = 1 - H_{\mathrm{m},ii}$. A point is considered to be influential if $H_{\mathrm{m},ii} = 1 - \mathrm{AP}_i > 2(m+1)/n$.

5. The Cook–Weisberg likelihood measure LD$_i$ [23] represents a general diagnostic defined by

$$\mathrm{LD}_i = 2[L\hat{\Theta} - L(\hat{\Theta}_{(i)})] \tag{19}$$

where $L(\hat{\Theta})$ is the maximum of the logarithm of the likelihood function when all points are used and $L(\hat{\Theta})_{(i)}$ is corresponding value when the ith point is omitted. The parametric vector Θ contains either the parameter $\boldsymbol{b}$ or the variance estimate $\hat{\sigma}^2$. For strongly influential points $\mathrm{LD}_i > \chi^2_{1-\alpha}(m+1)$ where $\chi^2_{1-\alpha}(m+1)$ is the quantile of the χ^2 distribution.

With the use of different variants of LD$_i$ it is possible to examine the influence of the ith point on parameter estimates, or on the variance estimate, or on both [23]:

5.1. The likelihood measure LD$_i(\boldsymbol{b})$ examines the influence of individual points on the parameter estimates $\boldsymbol{b}$ by the relationship

$$\mathrm{LD}_i(\boldsymbol{b}) = n \times \ln\left[\frac{d_i \times H_{ii}}{1 - H_{ii}} + 1\right] \tag{20}$$

where $d_i = \hat{e}^2_{\mathrm{S},i}/(n-m)$.

5.2. The likelihood measure LD$_i(\hat{\sigma}^2)$ examines the influence of individual points on the residual variance estimates by the relationship

$$\mathrm{LD}_i(\hat{\sigma}^2) = n \times \ln\left[\frac{n}{n-1}\right] + n\ln(1-d_i) + \frac{d_i(n-1)}{1-d_i} - 1 \tag{21}$$

5.3. The likelihood measure LD$_i(\boldsymbol{b}, \hat{\sigma}^2)$ examines the influence of individual points on the parameters $\boldsymbol{b}$ and variance estimates $\hat{\sigma}^2$ together by the relationship

$$\mathrm{LD}_i(\boldsymbol{b}, \hat{\sigma}^2) = n \times \ln\left[\frac{n}{n-1}\right] + n\ln(1-d_i) + \frac{d_i(n-1)}{(1-d_i)(1-H_{ii})} - 1 \tag{22}$$

6. Hadi's influence measure [22] is a new measure and graphical display for the characterization of overall potential influence in linear regression models. This influence measure is based on the simple fact that potentially influential observations are outliers in the X- or y-space. This measure is expressed in the form

$$K_i^2 = \frac{m}{1 - H_{ii}} \frac{\hat{e}^2_{\mathrm{N},i}}{1 - \hat{e}^2_{\mathrm{N},i}} + \frac{H_{ii}}{1 - H_{ii}} \tag{23}$$

where $\hat{e}_{\mathrm{N},i}$ is ith normalized residual. Large K_i^2 indicates influential points; the potential residual plot is then the scatter plot of the first versus second components of K_i^2. This plot is related but not equivalent to the L–R plot. The extensions of K_i^2 for the identification of influential subsets is very simple [22].

3. Procedure

3.1. Procedure for regression model building

The procedure for regression model building including the examination of influential points comprises the following steps:

Step 1. *Proposal of a model*: the procedure usually starts from the simplest model, with individual explanatory controllable variables not raised to powers other than the first, and with no interaction terms of the type $x_j x_k$ included.

Step 2. *Exploratory data analysis in regression:* the scatter plot of individual variables and all the possible pair combinations are examined. Also in this step, the influential points causing multicollinearity are detected.

Step 3. *Parameter estimation:* the parameters of the proposed regression model and the corresponding basic statistical characteristics of this model are determined by the classic least-squares method (LS). Individual parameters are tested for significance by way of Student's *t*-test. The mean quadratic error of prediction MEP and the Akaike information criterion AIC are calculated to examine the quality of model.

Step 4. *Analysis of regression diagnostics:* different diagnostic graphs and numerical measures are used to examine influential points, namely outliers and high-leverages. If influential points are found, it has to be decided whether these points should be eliminated from the data. If points are eliminated, the whole data treatment must be repeated.

Step 5. *Construction of a more accurate model:* according to the test for fulfillment of the conditions for the least-squares method, and the result of regression diagnostics, a more accurate regression model is constructed.

3.2. Software used

For the creation of regression diagnostic graphs and the computation of all regression characteristics an algorithm was written in Matlab 5.3 and also a unique module of the ADSTAT package was used, cf. [86]. The matrix oriented programming leads here to the very effective computations.

4. Illustrative examples

Regression model building and the discovery of influential points in three datasets from literature have been investigated extensively. These data are suitable for a demonstration of the efficiency of diagnostic tools for influential points indication. The majority of multiple outliers and high-leverages has been better detected by diagnostic plots than by values in tables.

Example 1. *Operation of a plant for the oxidation of ammonia to nitric acid*: the stackloss dataset, originally studied by Brownlee [11], is by far the most often cited sample data in the regression diagnostic literature: these are observations from 21 days in the operation of a plant for the oxidation of ammonia as a stage in the production of nitric acid. The independent variables are: x_1 the rate of operation of the plant measured by air flow, x_2 the inlet temperature of the cooling water circulating through coils in the countercurrent absorption tower for nitric acid, x_3 proportional to the concentration of nitric acid in the absorbing liquid in the tower expressed as $10 \times$ (acid concentration $-$ 50), and dependent variable y representing the stack loss, i.e. 10 times the percentage of ingoing ammonia escaping unconverted as unabsorbed nitric oxides. This is an inverse measure of the yield of nitric acid for the plant Table 1.

These data have been much analyzed as a tested for methods of outlier and leverage detection. A summary of many analyses is given by Atkinson, p. 266 [4]. While Gray and Ling [19] identified the subset (1, 2, 3) as the most influential triple in the stackloss data, other authors described the joint influential points (1, 2, 3, 4, 21). Investigation of plant operations indicates that the following sets of runs can be considered as replicates: (1, 2), (4, 5, 6), (7, 8), (11, 12) and (18, 19). While the runs in each set are not exact replicates, the points are sufficiently close to each other in x-space for them to be used as such.

As the conditioning number $K = \lambda_{\max}/\lambda_{\min}$ (the ratio of the maximal and minimal eigenvalue of the regressor matrix $\boldsymbol{X}$) was 9.99 which was less than 1000, no multicollinearity was proven. Also, all three values of the variance inflation factor VIF = 2.87, 2.40 and 1.41 of matrix $\boldsymbol{X}$ were less than 10, so no multicollinearity was indicated.

Table 1
The stackloss dataset written in the order x_1, x_2, x_3, y

80.0	27.0	89.0	42.0	80.0	27.0	88.0	37.0	75.0	25.0	90.0	37.0
62.0	24.0	87.0	28.0	62.0	22.0	87.0	18.0	62.0	23.0	87.0	18.0
62.0	24.0	93.0	19.0	62.0	24.0	93.0	20.0	58.0	23.0	87.0	15.0
58.0	18.0	80.0	14.0	58.0	18.0	89.0	14.0	58.0	17.0	88.0	13.0
58.0	18.0	82.0	11.0	58.0	19.0	93.0	12.0	58.0	18.0	89.0	8.0
50.0	18.0	86.0	7.0	50.0	19.0	72.0	8.0	50.0	19.0	79.0	8.0
50.0	20.0	80.0	9.0	56.0	20.0	82.0	15.0	70.0	20.0	91.0	15.0

The stackloss data have been re-examined to show the detection power of various regression diagnostics for influential points detection. This analysis begins by fitting a model in the three explanatory variables to the untransformed data. The majority of the partial correlation coefficients between the percentage of the input ammonia and chemical descriptors were significant; therefore, the total OLS regression model is determined

$$y = -37.68(12.01, \mathrm{S}) + 0.7336(0.1388, \mathrm{S})x_1 + 1.3883(0.3565, \mathrm{S})x_2 - 0.2167(0.1613, \mathrm{N})x_3$$

where brackets contain the standard deviation of the parameter estimated. The quantile $t_{1-\alpha/2}(21–4) = 2.110$ of the Student's t-test (5% significance level) is used to examine the test statistics (t's) of the individual regression parameters: $t_0 = -3.137$, $t_1 = 5.285$, $t_2 = 3.894$, $t_3 = -1.343$. All values except t_3 are equal to or greater than $t_{1-\alpha/2}(21–4) = 2.110$ and are significant, denoted with the letter S, while for b_3 the letter N stands for non-significant. The model is described with the mean error of prediction MEP = 13.61, the predicted coefficient of determination $\hat{R}^2_{\mathrm{P}} = 0.9283$ and the Akaike information criterion AIC = 53.09 (Figs. 7 and 8).

Detecting influential points, three blocks of diagnostics have been applied: (i) diagnostic plots based on residuals and hat matrix elements, (ii) index graphs of diagnostics based on vector and scalar influence

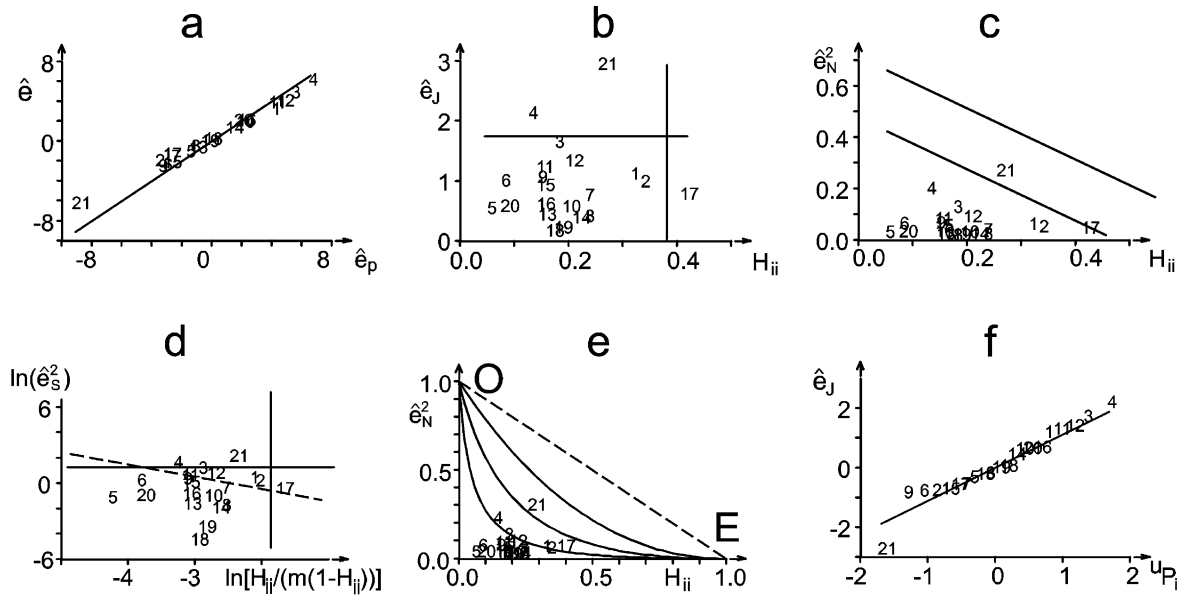

Fig. 7. Diagnostics based on residual plots and hat matrix elements for stackloss data: (a) graph of predicted residuals, (b) Williams graph, (c) Pregibon graph, (d) McCulloh and Meeter graph, (e) Gray L–R graph, (f) rankit Q–Q graph of jackknife residuals.

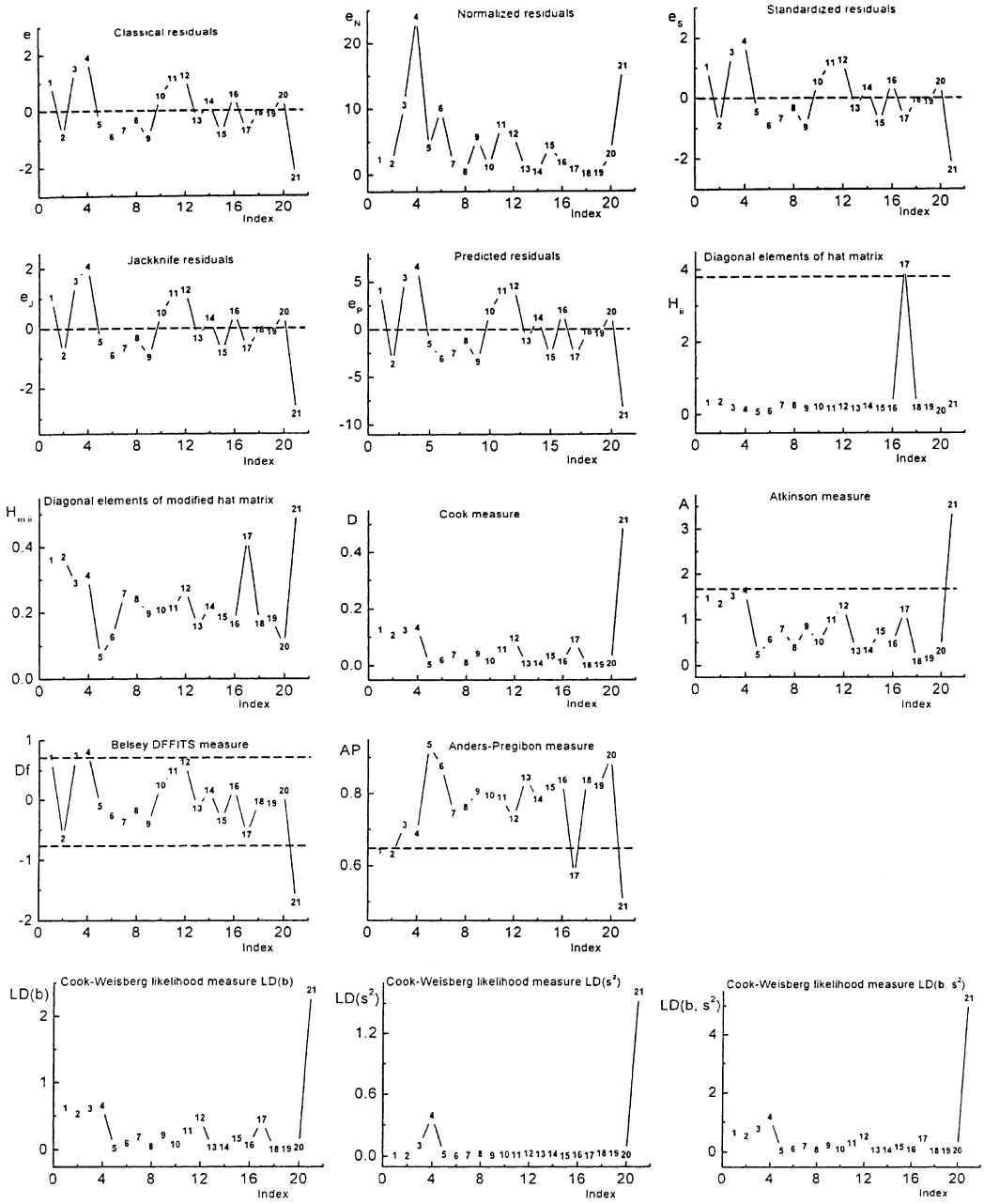

Fig. 8. Index graphs of various residuals and vector and scalar influence measures for stackloss data: suspicious points (SP) are points which obviously differ from the others; influential points (IP) are points which are detected and are separated into outliers and high-leverages using following testing criteria: $n = 16$, $m = 4$, $F_{0.95}(n-m, m) = 5.915$, $t_{0.95}(n-m) = 1.782$, $\chi^2_{1-\alpha}(m+1) = 11.07$, $t_{0.95}(n-m-1) = 1.796$, where $\hat{e}$: detects SP only; $\hat{e}_{\mathrm{N}}$: when $\hat{e}_{\mathrm{N},i} > |3\sigma|$ then the ith point is an outlier; $\hat{e}_{\mathrm{S}}$: detects SP only; $\hat{e}_{\mathrm{J}}$: when $\hat{e}^2_{\mathrm{J},i} > 10$ then the ith point is an outlier; $\hat{e}_{\mathrm{P}}$: detects SP only; H_{ii}: when $H_{ii} > 2m/n$ then the ith point is a high-leverage; $H_{\mathrm{m},ii}$: when $H_{\mathrm{m},ii} > 2m/n$ then the ith point is a high-leverage; D_i: when $D_i > 1$ then the ith point is an IP; A_i: when $A_i^2 > 10$ then the ith point is an outlier; DFFITS_i: when $|\mathrm{DFFITS}_i| > 2\sqrt{m/n}$ then the ith point is an IP; AP_i: when $\mathrm{AP}_i < 1 - 2(m+1)/n = 0.375$ then the ith point is an IP; LD_i: when $\mathrm{LD}_i > \chi^2_{1-\alpha}(m+1)$ then the ith point is an IP.

Table 2
A survey of the influential points which were tested using various diagnostic tools[a]

Diagnostic indicating SP and IP	Suspicious points (SP)	Influential points (IP)	Outliers (O)	High-leverages (E)
(A) Diagnostical plots based on residuals and hat matrix elements				
1. Graph of predicted residuals	1, 3, 4, 11, 12, 21	2, 3, 4, 17, 21	3, 4, 21	1, 2, 17
2. Williams graph	4, 17, 21	4, 17, 21	4, 21	17
3. Pregibon graph	17, 21	17, 21	–	–
4. McCulloh–Meeter graph	4, 17, 21	4, 17, 21	4, 21	17
5. Gray's L–R graph	1, 2, 4, 17, 21	1, 2, 4, 17, 21	4, 21	1, 2, 17
6. Q–Q graph of jackknife residuals	1, 3, 4, 11, 12, 21	–	–	–
(B) Diagnostics based on scalar and vector influence measures				
7. Cook measure D	1, 2, 3, 4, 21	21	–	–
8. Atkinson measure A	1, 2, 3, 4, 17, 21	4, 21	–	–
9. Belsey measure DFFITS	1, 2, 3, 4, 12, 17, 21	4, 21	–	–
10. Anders–Pregibon diagnostic AP	1, 2, 17, 21	1, 2, 17, 21	–	–
11. Cook–Weisberg likelihood measure LD(b)	1, 2, 3, 4, 12, 17, 21	21	–	–
12. Cook–Weisberg likelihood measure LD(s^2)	4, 21	21	–	–
13. Cook–Weisberg likelihood measure LD(b, s^2)	4, 21	21	–	–
(C) Index graphs of residuals and hat matrix elements				
14. Ordinary residuals e	1, 3, 4, 21	21	–	–
15. Normalized residuals e_N	4, 21	4, 21	–	–
16. Standardized residuals e_S	4, 12, 21	4, 21	–	–
17. Jackknife residuals e_J	1, 4, 21	21	–	–
18. Predicted residuals e_P	1, 3, 4, 21	4, 21	–	–
19. Diagonal elements of hat matrix H_{ii}	17	17	–	–
20. Diagonal elements of modified hat matrix $H_{m,ii}$	1, 2, 17, 21	2, 17, 21	–	–

[a] Suspicious points (SP) are points in diagnostic graphs which obviously differ from the others; influential points (IP) are points which are detected and are separated into outliers and high-leverages using following testing criteria: $n = 16, m = 4$, $F_{0.95}(n-m, m) = 5.915$, $t_{0.95}(n-m) = 1.782$, $\chi^2_{1-\alpha}(m+1) = 11.07$, $t_{0.95}(n-m-1) = 1.796$.

measures, and (iii) index graphs of residuals and hat matrix elements. Table 2 shows that five diagnostic plots and the Q–Q graph of the jackknife residuals indicate suspicious points which obviously differ from others. The statistical criteria in the diagnostic plots were then used to prove influential points. These plots also separate influential points into outliers and high-leverages. The diagnostic plots found two outliers, 4 and 21, and one high-leverage point, 17. Index graphs of diagnostics based on vector and scalar influence measures, as well as index graphs of residuals and hat matrix elements, tested all the suspicious points and found statistically significant influential points. However, these index graphs are not able to separate these points into outliers and high-leverages. Most index graphs agreed that there were three influential points (4, 17, 21). Removing a non-significant regressor from the equation has a negligible effect on the fitted values and deletion of one outlier (21) formed the regression model

$$y = -53.21(4.55, \mathrm{S}) + 0.824(0.123, \mathrm{S})x_1 + 0.999(0.336, \mathrm{S})x_2$$

with MEP = 9.35, $\hat{R}^2_P = 0.9536$ and AIC = 43.80, while when deleting 2 outliers (4, 21) the final regression model achieved was

$$y = -52.28(3.86, \mathrm{S}) + 0.888(0.106, \mathrm{S})x_1 + 0.756(0.297, \mathrm{S})x_2$$

with MEP = 6.94, $\hat{R}^2_P = 0.9656$ and AIC = 35.30. The lowest values of MEP and AIC and the highest value of $\hat{R}^2_P$ proved, to be the final model. On its own, the leverage 17 is not particularly informative about the behavior of the units, since it does not depend on the observed values of y.

Example 2. *Aerial biomass and five physicochemical properties of the substrate (salinity, pH, K, Na and Zn)* [67]: the purpose is to find a regression model

to detect influential points and to identify important soil characteristics (salinity, pH, K, Na and Zn) influencing the aerial biomass production of the marsh grass Spartina alterniflora. The 45 data values used involve 1 month of sampling are part of a larger study by Linthurst (cf. p. 161 in [84,85]) and five substrate measurements: x_1 salinity (‰), x_2 acidity as measured in water pH, x_3 potassium (ppm), x_4 sodium (ppm), and x_5 zinc (ppm); y the dependent variable is aerial biomass (g/m^2), (Table 3).

The purpose of this study was to identify the important soil characteristics influencing aerial biomass production of the marsh grass and to identify the substrate variables showing the stronger relationship to biomass to relate total variability in biomass production to total variability in the five substrate variables.

The initial model assumes the biomass, y, can be adequately characterized by linear relationship with the five independent variables plus an intercept,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

where y is the vector of biomass measurements and $\boldsymbol{X}$ (45×6) consists of the column vector $\mathbf{1}$, the 45×1 column vector of ones, and the five column vectors of data for substrate variables.

As the conditioning number $K = \lambda_{\max}/\lambda_{\min}$ (the ratio of the maximal and minimal eigenvalue of the regressor matrix $\boldsymbol{X}$) was 17.791 which was less than 1000, no multicollinearity was proven. Also, all five values of the variance inflation factor VIF = 2.217, 3.331, 2.983, 3.334 and 4.310 of matrix $\boldsymbol{X}$ were less than 10, and thus no multicollinearity was indicated. The OLS method found the regression model

$$\begin{aligned} y = {} & 1252.3(1234.8, \mathrm{N}) - 30.3(24.0, \mathrm{N})x_1 \\ & +305.5(87.9, \mathrm{S})x_2 - 0.285(0.348, \mathrm{N})x_3 \\ & -0.00867(0.01593, \mathrm{N})x_4 - 20.68(15.06, \mathrm{N})x_5 \end{aligned}$$

where brackets contain the standard deviations of the parameters estimated. The quantile $t_{1-\alpha/2}(45–6) = 2.023$ of a Student's t-test (5% significance level) was used to examine the test statistics (t's) of the individual regression parameters: $t_0 = 1.014$, $t_1 = -1.260$, $t_2 = 3.477$, $t_3 = -0.818$, $t_4 = -0.544$ and $t_5 = -1.373$. All values except t_2 are less than $t_{1-\alpha/2}(45–6) = 2.023$, are not significant, and are denoted by the letter N. The model was described with a mean error of

Table 3
Aerial biomass and five physicochemical properties of the substrate [67][a]

x_1	x_2	x_3	x_4	x_5	y
33	5.00	1441.67	35185.50	16.4524	676
35	4.75	1299.19	28170.40	13.9852	516
32	4.20	1154.27	26455.00	15.3276	1052
30	4.40	1045.15	25072.90	17.3128	868
33	5.55	521.62	31664.20	22.3312	1008
33	5.05	1273.02	25491.70	12.2778	436
36	4.25	1346.35	20877.30	17.8225	544
30	4.45	1253.88	25621.30	14.3516	680
38	4.75	1242.65	27587.30	13.6826	640
30	4.60	1281.95	26511.70	11.7566	492
30	4.10	553.69	7886.50	9.8820	984
37	3.45	494.74	14596.00	16.6752	1400
33	3.45	525.97	9826.80	12.3730	1276
36	4.10	571.14	11978.40	9.4058	1736
30	3.50	408.64	10368.60	14.9302	1004
30	3.25	646.65	17307.40	31.2865	396
27	3.35	514.03	12822.00	30.1652	352
29	3.20	350.73	8582.60	28.5901	328
34	3.35	496.29	12369.50	19.8795	392
36	3.30	580.92	14731.90	18.5056	236
30	3.25	535.82	15060.60	22.1344	392
28	3.25	490.34	11056.30	28.6101	268
31	3.20	552.39	8118.90	23.1908	252
31	3.20	661.32	13009.50	24.6917	236
35	3.35	672.15	15003.70	22.6758	340
29	7.10	528.65	10225.00	0.3729	2436
35	7.35	563.13	8024.20	0.2703	2216
35	7.45	497.96	10393.00	0.3205	2096
30	7.45	458.38	8711.60	0.2648	1660
30	7.40	498.25	10239.60	0.2105	2272
26	4.85	936.26	20436.00	18.9875	824
29	4.60	894.79	12519.90	20.9687	1196
25	5.20	941.36	18979.00	23.9841	1960
26	4.75	1038.79	22986.10	19.9727	2080
26	5.20	898.05	11704.50	21.3864	1764
25	4.55	989.87	17721.00	23.7063	412
26	3.95	951.28	16485.20	30.5589	416
26	3.70	939.83	17101.30	26.8415	504
27	3.75	925.42	17849.00	27.7292	492
27	4.15	954.11	16949.60	21.5699	636
24	5.60	720.72	11344.60	19.6531	1756
27	5.35	782.09	14752.40	20.3295	1232
26	5.50	773.30	13649.80	19.5880	1400
28	5.50	829.26	14533.00	20.1328	1520
28	5.40	856.96	16892.20	19.2420	1560

[a] The 45 data values used involve 1 month of sampling are part of a larger study by Linthurst (cf. p. 161 in [84,85]) and five substrate measurements: x_1 salinity (‰), x_2 acidity as measured in water pH, x_3 potassium (ppm), x_4 sodium (ppm), and x_5 zinc (ppm); y the dependent variable is aerial biomass (g/m^2).

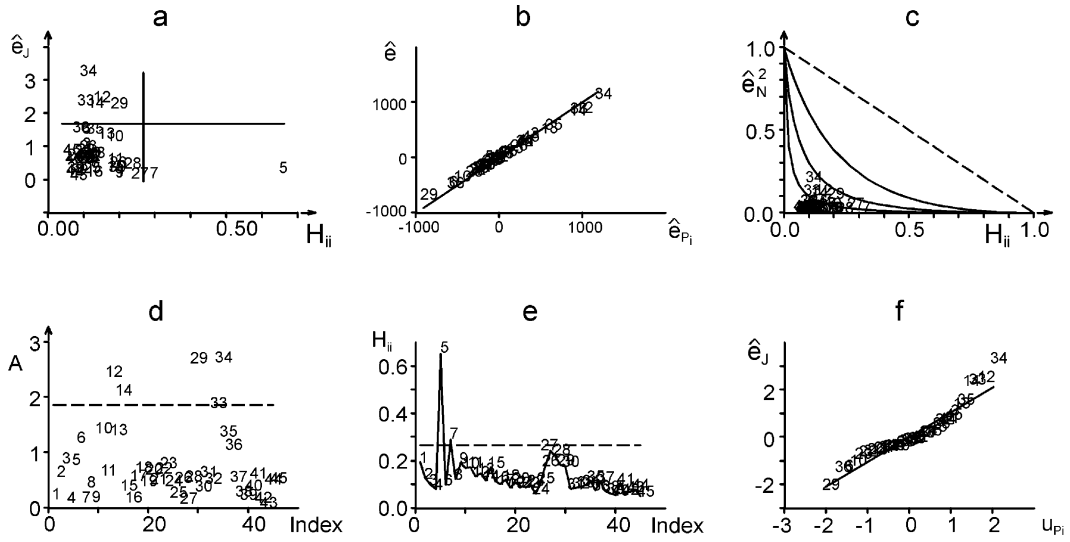

Fig. 9. Selected diagnostics plots for the detection of influential points in aerial biomass data: (a) Williams graph, (b) graph of predicted residuals, (c) Gray L–R graph, (d) index graph of Atkinson measure, (e) index graph of diagonal elements of the hat matrix, (f) rankit Q–Q graph of jackknife residuals.

prediction MEP $= 1.7678 \times 10^5$, the predicted coefficient of determination $\hat{R}^2_{\mathrm{P}} = 0.7649$ and the Akaike information criterion AIC $= 544.40$.

Fig. 9 shows six reliable diagnostic graphs for the detection of influential points. While the first three graphs separate outliers from high-leverages, the other three graphs indicate influential points only without their separation: 7 influential points were detected in 45 data values, and separated into five outliers (12, 14, 29, 33, 34) and two high-leverages (5, 7). The *t*-test of the partial regression parameters H_0: $\beta_j = 0$ would seem to suggest that four of the five independent variables are unimportant and could be dropped from the model. The corrected regression model was then re-calculated using a data set without five outliers 12, 14, 29, 33 and 34 and without statistically insignificant parameters in the form

$$y = -1133.0(198.15, \mathrm{S}) + 447.5(42.1, \mathrm{S})x_2$$

with MEP $= 1.0633 \times 10^5$, $\hat{R}^2_{\mathrm{P}} = 0.8527$ and AIC $=$ 463.67. The lowest values of MEP and AIC and the highest value of $\hat{R}^2_{\mathrm{P}}$ proved, to be the best final model.

Example 3. *The quantitative structure-activity relationship analysis of positively charged sulfonamide inhibitors of the zinc enzyme carbonic anhydride* [60]. The synthetic routes of a series of tri-, tetra-, and penta-substituted 1-2(sulfonamido-1,3,4-thiadiazol-5-yl) pyridinium agents have been published [60]. The 50% inhibitory concentration [μM] of carbonic anhydrase for 28 compounds was determined, transformed by natural logarithm and signified by the dependent variable y. A number of physicochemical descriptors for all 28 compounds were used: x_1 the energy (eV) of the highest occupied molecular orbital (HOMO), x_2 the diagonal component P_{xx} of polarizability along the smallest principal moment of inertia, x_3 the diagonal component P_{yy} of polarizability along the intermediate principal moment of inertia, x_4 the diagonal component P_{zz} of polarizability along the intermediate principal moment of inertia, x_5 the sum of charges of 2, 6-carbon atoms on the pyridinium ring, and x_6 the sum of charges of 3, 5-carbon atoms on the pyridinium ring, Table 4.

As the conditioning number $K = \lambda_{\max}/\lambda_{\min}$ (the ratio of the maximal and minimal eigenvalue of the regressor matrix $\boldsymbol{X}$) was 36.461, which was less than 1000, no multicollinearity was proven. Also, all six values of the variance inflation factor VIF of matrix $\boldsymbol{X}$

Table 4
Original dataset of the quantitative structure–activity relation analysis [60][a]

x_1	x_2	x_3	x_4	x_5	x_6	y
−13.690	236.100	131.100	106.800	−0.207	−0.217	1.498
−13.600	248.700	153.300	117.500	−0.198	−0.249	1.373
−13.660	246.700	149.000	119.200	−0.214	−0.226	1.330
−13.610	250.700	159.800	124.100	−0.197	−0.243	1.290
−13.580	257.600	168.100	130.000	−0.212	−0.235	1.262
−13.590	271.900	189.700	141.100	−0.205	−0.284	1.199
−13.610	297.300	180.300	108.700	−0.223	−0.229	0.778
−13.600	290.700	177.600	142.900	−0.206	−0.265	0.785
−13.600	304.900	195.000	150.400	−0.221	−0.261	0.869
−12.790	247.500	232.500	135.600	−0.170	−0.200	−0.301
−12.630	241.800	219.400	162.600	−0.158	−0.209	−0.097
−12.570	234.600	211.500	175.400	−0.161	−0.196	0.462
−12.680	337.900	133.800	133.800	−0.237	−0.191	−0.523
−12.650	350.100	148.700	148.700	−0.250	−0.206	−0.699
−12.640	343.600	182.000	182.000	−0.238	−0.213	−0.155
−12.620	340.500	180.300	180.300	−0.254	−0.217	−0.155
−12.720	322.300	141.200	141.200	−0.221	−0.119	0.176
−12.660	331.200	163.800	163.800	−0.263	−0.128	0.176
−12.610	345.700	186.700	186.700	−0.193	−0.171	1.980
−12.880	265.300	195.000	195.000	−0.112	−0.192	2.071
−12.580	276.800	214.300	214.300	−0.094	−0.199	2.095
−12.520	305.600	271.200	204.300	−0.090	−0.200	2.171
−12.520	342.500	274.400	247.200	−0.135	−0.161	2.121
−11.180	616.700	217.200	124.200	−0.185	−0.277	0.041
−10.970	776.000	255.600	108.700	−0.183	−0.287	0.176
−13.570	246.700	144.100	112.500	−0.234	−0.160	1.487
−13.520	271.700	170.000	130.300	−0.274	−0.091	0.079
−13.630	281.800	195.400	161.600	−0.275	−0.065	0.531

[a] Physicochemical descriptors for all 28 compounds were used: x_1 the energy (eV) of the highest occupied molecular orbital (HOMO), x_2 the diagonal component P_{xx} of polarizability along the smallest principal moment of inertia, x_3 the diagonal component P_{yy} of polarizability along the intermediate principal moment of inertia, x_4 the diagonal component P_{zz} of polarizability along the intermediate principal moment of inertia, x_5 the sum of charges of 2, 6-carbon atoms on the pyridinium ring, and x_6 the sum of charges of 3, 5-carbon atoms on the pyridinium ring.

were less than 10, and thus no multicollinearity was indicated.

The majority of the partial correlation coefficients between the biological activity y and chemical descriptors $\boldsymbol{X}$ were significant. The total OLS regression model determined was

$$y = -20.86(4.06, \mathrm{S}) - 1.83(0.28, \mathrm{S})x_1 + 0.0090(0.0018, \mathrm{S})x_2 - 0.0085(0.0036, \mathrm{S})x_3 + 0.0102(0.0037, \mathrm{S})x_4 + 20.44(2.90, \mathrm{S})x_5 + 3.92(1.98, \mathrm{N})x_6$$

where brackets contain the standard deviations of the parameter estimated. The critical quantile $t_{1-\alpha/2}(28\text{–}7) = 2.080$ of a Student's t-test (5% significance level) was used to examine the test statistics (t's) of the individual regression parameters: $t_0 = -5.14$, $t_1 = -6.56$, $t_2 = 4.92$, $t_3 = -2.33$, $t_4 = 2.79$, $t_5 = 7.05$ and $t_6 = 1.99$. All values except t_6 were higher than $t_{1-\alpha/2}(28\text{–}7)$, were significant, and are denoted by the letter S. The model was described with a mean error of prediction MEP = 0.2230, the predicted coefficient of determination $\hat{R}^2_{\mathrm{P}} = 0.83640$ and the Akaike information criterion AIC = −41.57.

After testing the significance of multicollinearity, the presence of influential points (outliers and high-leverages) should be examined. The discovery of outliers in data has been investigated extensively (Fig. 10). Of the three indicated influential points there were two outliers (compounds 19 and 26) and

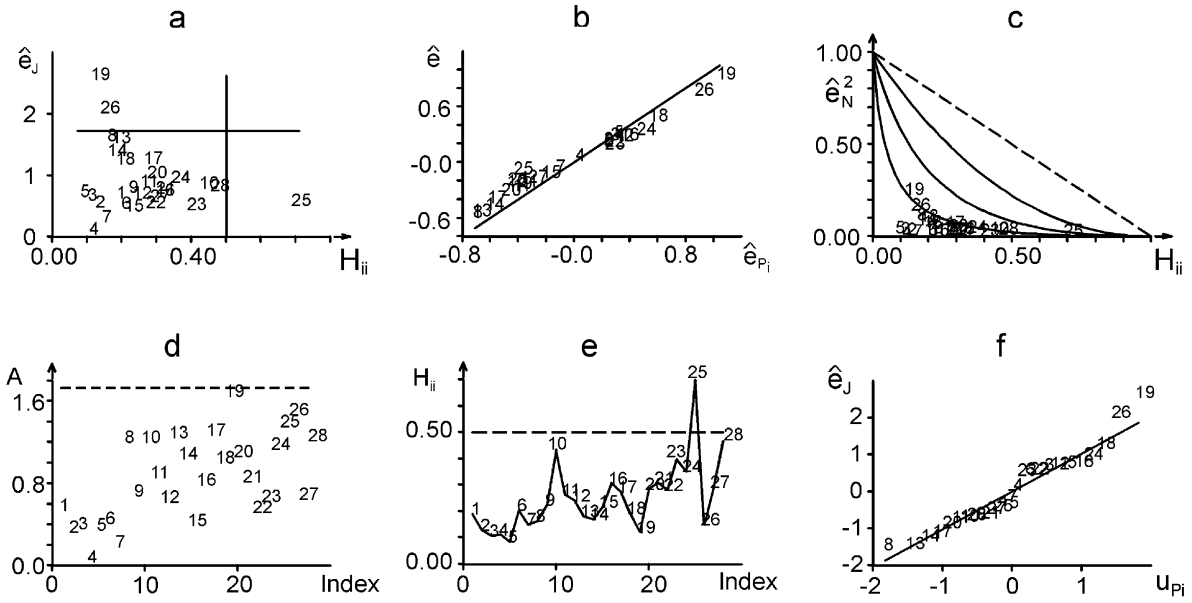

Fig. 10. Selected diagnostics plots for the detection of influential points in structure–activity data: (a) Williams graph, (b) graph of predicted residuals, (c) Gray L–R graph, (d) index graph of Atkinson measure, (e) index graph of diagonal elements of the hat matrix, (f) rankit Q–Q graph of jackknife residuals.

one high-leverage point (compound 25). While Mager [60] found two high-leverage points (compounds 25 and 28), one outlier (compound 19), and two collinearity-creating points (compounds 24 and 25), but he found no influential points; despite Mager's conclusion [60], with the use of graphical diagnostics three influential points (compound 19, 25 and 26) were detected and separated into two outliers (compounds 19 and 26) which should be excluded from the original dataset, and one high-leverage (compound 25) which can remain in the data.

With the omission of one insignificant regressor x_6 and the deletion of two outliers (19, 26) the regression model was

$$y = -20.51(3.39, \mathrm{S}) - 1.67(0.23, \mathrm{S})x_1 + 0.0074(0.0014, \mathrm{S})x_2 - 0.0058(0.0031, \mathrm{N})x_3 + 0.0113(0.0030, \mathrm{S})x_4 + 17.18(2.07, \mathrm{S})x_5$$

where brackets contain the standard deviation of parameter estimated. The quantile $t_{1-\alpha/2}(26\text{–}6) = 2.086$ of a Student's t-test (5% significance level) was used to examine the test statistics (t's) of the individual regression parameters: $t_0 = -6.06$, $t_1 = -7.43$, $t_2 = 5.17$, $t_3 = -1.88$, $t_4 = 3.70$, and $t_5 = 8.29$. All values except t_3 were higher than $t_{1-\alpha/2}(26\text{–}6)$, were significant, and are denoted by the letter S. The model was described with a mean error of prediction MEP = 0.1586, the predicted coefficient of determination $\hat{R}^2_{\mathrm{P}} = 0.8831$ and the Akaike information criterion AIC = −48.37.

With the omission of another insignificant regressor x_3 the final regression model was

$$y = -20.48(3.58, \mathrm{S}) - 1.61(0.24, \mathrm{S})x_1 + 0.0063(0.0014, \mathrm{S})x_2 + 0.0085(0.0028, \mathrm{S})x_4 + 15.00(1.82, \mathrm{S})x_5$$

and the critical quantile $t_{1-\alpha/2}(26\text{–}5) = 2.080$ of a Student's t-test (5% significance level) was used to examine the test statistics (t's) of the individual regression parameters: $t_0 = -5.71$, $t_1 = -6.84$, $t_2 = 4.56$, $t_4 = 3.01$ and $t_5 = 8.24$. All values were significant, and are denoted by the letter S. The model was described with a mean error of prediction MEP = 0.1746, the predicted coefficient of

determination $\hat{R}_P^2 = 0.8703$ and the Akaike information criterion AIC $= -46.142$.

5. Conclusions

Regression diagnostics do not require a knowledge of an alternative hypothesis for the testing or fulfilling of the other assumptions of classical statistical tests. In the interactive PC-assisted diagnosis of data, the examination of data quality involves the detection of influential points, outliers and high-leverages, which cause many problems in regression analysis by shifting the parameter estimates or increasing the variance of the parameters. The main difference between the use of regression diagnostics and classical statistical tests is that there is no necessity for an alternative hypothesis, but all kinds of deviations from the ideal state are discovered. In statistical graphics, information is contained in observable shapes and patterns: regression diagnostics represent the graphical procedures and numerical measures for an examination of the regression triplet, i.e. the identification of (i) the data quality for a proposed model, (ii) the model quality for a given dataset, and (iii) a fulfillment of all least-squares assumptions. The authors' concept of exploratory regression analysis is based on the facts that "the user knows more about the data than the computer does" and graphs are more informative than tables. Selected diagnostic plots were chosen as suitable to give reliable results of influential points detection and to separate influential points into outliers and high-leverages. The graphical aids for the identification of outliers and high-leverage points are combined with graphs for the identification of influence type based on likelihood distance. All these graphically oriented techniques are suitable for the rapid estimation of influential points, but are generally incapable of solving problems with masking and swamping. The Matlab 5.3 procedure for influential points characteristics computation is very useful for practitioners and can be used very simply for routine calculations. Results are comparable with those obtained with the use of expensive statistical packages.

Acknowledgements

The financial support of the Grant Agency of the Czech Republic (Grant no. 303/00/1559) is gratefully acknowledged. Karel Kupka is acknowledged for his help with some figures made in ADSTAT 3.0.

References

[1] M. Meloun, J. Militký, M. Forina, Chemometrics for Analytical Chemistry, Vol. 2, PC-Aided Regression and Related Methods, Ellis Horwood, Chichester, 1994.
[2] D.A. Belsey, E. Kuh, R.E. Welsch, Regression Diagnostics: Identifying Influential data and Sources of Collinearity, Wiley, New York, 1980.
[3] R.D. Cook, S. Weisberg, Residuals and Influence in Regression, Chapman & Hall, London, 1982.
[4] A.C. Atkinson, Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis, Claredon Press, Oxford, 1985.
[5] S. Chatterjee, A.S. Hadi, Sensitivity Analysis in Linear Regression, Wiley, New York, 1988.
[6] V. Barnett, T. Lewis, Outliers in Statistical Data, 2nd Edition, Wiley, New York, 1984.
[7] R.E. Welsch, Linear regression diagnostics, Technical Report 923-77, Sloan School of Management, Massachusetts Institute of Technology, 1977.
[8] R.E. Welsch, S.C. Peters, Finding influential subsets of data in regression models, in: A.R. Gallant, T.M. Gerig (Eds.), Proceedings of the 11th Interface Symposium on Computer Science and Statistics, Institute of Statistics, North carolina State University, Raleigh, 1978.
[9] S. Weisberg, Applied Linear Regression, Wiley, New York, 1985.
[10] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[11] K.A. Brownlee, Statistical Theory and Methodology in Science and Engineering, Wiley, New York, 1965.
[12] J.F. Gentleman, M.B. Wilk, Detecting outliers. II. Supplementing the direct analysis of residuals, Biometrics 31 (1975) 387–410.
[13] R.D. Cook, Detection of influential observations in linear regression, Technometrics 19 (1977) 15–18.
[14] R.D. Cook, S. Weisberg, Characterization of an empirical influence function for detecting influential cases in regression, Technometrics 22 (1980) 495–508.
[15] D.C. Hawkins, D. Bradu, G.V. Kass, Location of several outliers in multiple regression data using elemental sets, Technometrics 26 (1984) 197–208.
[16] C.E. McCulloch, D. Meeter, Discussion of "outliers" by R.J. Beckman and R.D. Cook, Technometrics 25 (1983) 152–155.
[17] J.B. Gray, The L–R plot: a graphical tool for assessing influence, in: Proceedings of the Statistical Computing Section, Vols. 159–164, American Statistical Association, 1983.
[18] J.B. Gray, Graphics for regression diagnostics, Proceedings of the Statistical Computing Section, Vols. 102–107, American Statistical Association, 1985.
[19] J.B. Gray, R.F. Ling, K-clustering as a detection tool for influential subsets in regression, Technometrics 26 (1984) 305–330.

[20] J.B. Gray, A classification of influence measures, J. Statist. Comput. Simul. 30 (1988) 159–171.
[21] J.B. Gray, A simple graphic for assessing influence in regression, J. Statist. Comput. Simul. 24 (1986) 121–134.
[22] A.S. Hadi, A new measure of overall potential influence in linear regression, Comput. Statist. Data Anal. 14 (1992) 1–27.
[23] F.R. Hampel, The influence curve and its role in robust estimation, J. Am. Statist. Assoc. 69 (1974) 383–393.
[24] R.D. Cook, Influential observation in linear regression, J. Am. Statist. Assoc. 74 (1979) 169–174.
[25] R.D. Cook, P. Prescot, On the accuracy of Bonferroni significance levels for detecting outliers in linear models, Technometrics 23 (1981) 59.
[26] D.F. Andrews, D. Pregibon, Finding outliers that matter, J. Roy. Statist. Soc. B 40 (1978) 85–93.
[27] S. Weisberg, Developments in linear-regression methodology: 1959–1982 discussion, Technometrics 25 (1983) 240–244.
[28] D.A. Pierce, R.J. Gray, Testing normality of errors in regression models, Biometrika 69 (1982) 233.
[29] A. Hedayat, D.S. Robson, Independent stepwise residuals for testing homoscedasticity, J. Am. Statist. Assoc. 65 (1970) 1573.
[30] R.L. Brown, J. Durbin, J.M. Evans, Techniques for testing the constancy of regression relationship over time, J. Roy. Statist. Soc., Ser. B 37 (1975) 149.
[31] J.S. Galpin, D.M. Hawkins, The use of recursive residuals in checking model fit in linear regression, Am. Statist. 68 (1973) 144.
[32] C.P. Quesenberry, Some transformation methods in goodness-of-fit, in: R.B. D'Agostino, M.A. Stephens (Eds.), Goodness of Fit Techniques, Marcel Dekker, New York, Chapter 6, 1986.
[33] R.E. Welsch, Influence functions and regression diagnostics, in: R. Launer, A. Siegel (Eds.), Modern Data Analysis, Academic Press, New York, 1982.
[34] S. Chaterjee, A.S. Hadi, Influential observations, high-leverage points, and outliers in linear regression (with discussion), Statist. Sci. 1 (1986) 379–416.
[35] J. Militký, in: Proceedings of the European Simulation Conference 87, Prague, September 1987.
[36] M.A. O'Gorman, R.M. Myers, Commun. Statist. 16 (1987) 771.
[37] A.S. Hadi, Diagnosting collinearity — influential observations, Comput. Statist. Data Anal. 7 (1988) 143–159.
[38] A.S. Hadi, Two graphical displays for the detection of potentially influential subsets in regression, J. Appl. Statist. 17 (1990) 313–327.
[39] S. Ghosh, On two methods of identifying influential sets of observations, Statist. Probability Lett. 7 (1989) 241–245.
[40] R.D. Cook, S. Weisberg, Regression diagnostics with dynamic graphics, Technometrics 31 (1989) 277–311.
[41] S.R. Paul, K.Y. Fung, A generalized extreme studentized residual multiple-outlier detection procedure in linear regression, Technometrics 33 (1991) 339–348.
[42] P.J. Rousseeuw, B.C. van Zomeren, Unmasking multivariate outliers and leverage points, J. Am. Statist. Assoc. 85 (1990) 871–880.
[43] A.S. Hadi, W.D. Jones, R.F. Ling, A unifying representation of some case-deletion influence measures in univariate and multivariate linear regression, J. Statist. Planning Inference 46 (1995) 123–135.
[44] B. Seaver, K. Triantis, The identification of influential subsets in regression using a fuzzy clustering strategy, Technometrics 41 (1999) 340–352.
[45] Y. Dodge, A.S. Hadi, Simple graphs and bounds for the elements of the hat matrix, J. Appl. Statist. 26 (1999) 817–823.
[46] R.D. Cook, F. Critchley, Identifying regression outliers and mixtures graphically, J. Am. Statist. Assoc. 95 (2000) 781–794.
[47] SAS Institute Inc., SAS/STAT User's Guide, Version 6, 4th Edition, Vol. II, SAS Institute Inc., Cary, NC, 1989.
[48] Y. Tanaka, F. Zhang, R-mode and Q-mode influence analyses in statistical modelling: relationship between influence function approach and local influence approach, Comput. Statist. Data Anal. 32 (1999) 197–218.
[49] A.S. Kosinski, A procedure for the detection of multivariate outliers, Comput, Statist. Data Anal. 29 (1999) 145–161.
[50] W.K. Fung, Graphical summaries for influence of multiple observations, Commun. Statist. — Theory Methods 24 (1995) 415–427.
[51] B.E. Barrett, R.F. Ling, General classes of influence measures for multivariate regression, J. Am. Statist. Assoc. 87 (1992) 184–191.
[52] A.S. Hadi, J.S. Simonoff, Procedures for the identification of multiple outliers in linear models, J. Am. Statist. Assoc. 88 (1993) 1264–1272.
[53] S.S. Gupta, D.-Y. Huang, On detecting influential data and selecting regression variables, J. Statist. Planning Inference 53 (1996) 421–435.
[54] Ch. Kim, Cook's distance in spline smoothing, Statist. Probability Lett. 31 (1996) 139–144.
[55] A. Singh, Outliers and robust procedures in some chemometric applications, Chemomet. Intelligent Lab. Syst. 33 (1996) 75–100.
[56] Y.Z. Liang, O.M. Kvalheim, Robust methods for multivariate analysis: a tutorial review, Chemomet. Intelligent Lab. Syst. 32 (1996) 1–10.
[57] W.K. Fung, A stepwise method for the identification of multiple outliers and influential observations, S. Afr. Statist. J. 29 (1995) 51–64.
[58] F. Kianifard, W.H. Swallow, A review of the development and application of recursive residuals in linear models, J. Am. Statist. Assoc. 91 (1996) 391–400.
[59] T.J.B. Holland, S.A.T. Redfern, Unit cell refinement from powder diffraction data: the use of regression diagnostics, Mineralogical Mag. 61 (1997) 65–77.
[60] P.P. Mager, The role of diagnostic statistics in medicinal chemistry, Medicinal Res. Rev. 17 (1997) 505–522 and data taken from C.T. Supuran, B.W. Clare, Eur. J. Med. Chem. 30 (1995) 687.
[61] B.E. Barrett, J.B. Gray, High-leverage, residuals, and interaction diagnostics for subsets of cases in least squares regression, Comput. Statist. Data Anal. 26 (1997) 39–52.

[62] K.E. Muller, M.C. Mok, The distribution of Cook's D statistic, Commun. Statist. — Theory Methods 26 (1997) 525–546.
[63] A. Ziegler, M. Blettner, C. Kastner, J. Chang-Claude, Identifying influential families using regression diagnostics for generalized estimating equations, Genetic Epidemiol. 15 (1998) 341–353.
[64] D.M. Sebert, D.C. Montgomery, D.A. Rollier, A clustering algorithm for identifying multiple outliers in linear regression, Comput. Statist. Data Anal. 27 (1998) 461–484.
[65] R.M. Yager, Detecting influential observation in nonlinear regression modelling of groundwater flow, Water Resour. Res. 34 (1998) 1623–1633.
[66] I.H. Langford, T. Lewis, Outliers in multilevel data, J. Roy. Statist. Soc. Series A: Statist. Soc. 161 (2) (1998) 121–153.
[67] C.L. Tsai, Z.W. Cai, X.Z. Wu, The examination of residual plots, Statistica Sinica 8 (1998) 445–465.
[68] W.J. Egan, S.L. Mogan, Outlier detection in multivariate analytical chemical data, Anal. Chem. 70 (1998) 2372–2379.
[69] D. Jouan-Rimbaud, E. Bouveresse, D.L. Massart, O.E. de Noord, Detection of prediction outliers and inliers in multivariate calibration, Anal. Chim. Acta 388 (1999) 283–301.
[70] M.M.S. Rancel, M.A.G. Sierra, Measures and procedures for the identification of locally influential observations in linear regression, Commun. Statist. — Theory Methods 28 (1999) 343–366.
[71] J.W. McKean, J.D. Naranjo, S.J. Sheather, Diagnostics for comparing robust and least squares fits, J. Nonparametric Statist. 11 (1999) 161–188.
[72] C.G. Troskie, D.O. Chalton, M. Jacobs, Testing for outliers and influential observations in multiple regression using restricted least squares, S. Afr. Statist. J. 33 (1999) 1–40.
[73] N.S. Tang, B.C. Wei, X.R. Wang, Influence diagnostics in nonlinear reproductive dispersion models, Statist. Probability Lett. 46 (2000) 59–68.
[74] X.P. Zhong, B.C. Wei, W.K. Fung, Influence analysis for linear measurement error models, Ann. Institute Statist. Math. 52 (2000) 367–379.
[75] D. Pregibon, Logistic regression, diagnostics, Ann. Statist. 9 (1981) 45–52.
[76] D.X. Williams, Letter to the editor, Appl. Statist. 22 (1973) 407–408.
[77] B. Walczak, D.L. Massart, Multiple outlier detection revisited, Chemomet. Intelligent Lab. Syst. 41 (1998) 1–15.
[78] J.P. Van Buijten, Anatomical factors influencing wood specific gravity of slash pines and the implications for the development of a high/quality pulpwood, Tappi 47 (1964) 401–404.
[79] D.C. Hoaglin, R.E. Welsch, The hat matrix in regression and ANOVA, Am. Statist. 32 (1978) 17–22.
[80] F.J. Anscombe, Examination of residuals, in: Proceedings of the Fourth Berkeley Symposium on Math. Statist. Prob., Vol. 1, 1961, pp. 1–36.
[81] N.R. Draper, H. Smith, Applied Regression Analysis, 1st Edition, Wiley, New York, 1966.
[82] R.J. Carrol, D. Ruppert, Transformation and Weighting in Regression, Chapman & Hall, New York, 1988.
[83] B. Walczak, D.L. Massart, Robust principal component regression as a detection tool for outliers, Chemomet. Intelligent Lab. Syst. 27 (1995) 41–54.
[84] R.A. Linthurst, Aeration, nitrogen, pH and salinity as factors affecting Spartina Alterniflora growth and dieback, PhD thesis, North Carolina State University, 1979.
[85] J.O. Rawlings, S.G. Pantula, D.A. Dickey, Applied Regression Analysis, A Research Tool, 2nd Edition, Springer, New York, 1998.
[86] ADSTAT (English version), TriloByte Statistical Software, Pardubice, 1999.
[87] R.D. Cook, S. Weisberg, Graphs in statistical analysis: is the medium the message, Am. Statist. 53 (1999) 29–37.
[88] A. Singh, Outliers and robust procedures in some chemometric applications, tutorial, Chemomet. Intelligent Lab. Syst. 33 (1996) 75–100.
[89] R.J. Pell, Multiple outlier detection for multivariate calibration using robust statistical techniques, Chemomet. Intelligent Lab. Syst. 52 (2000) 87–104.
[90] B. Walczak, Outlier detection in bilinear calibration, Chemomet. Intelligent Lab. Syst. 29 (1995) 63–73.
[91] B. Walczak, Outlier detection in multivariate calibration, Chemomet. Intelligent Lab. Syst. 29 (1995) 259–272.