THE ANALYST
TUTORIAL REVIEW
www.rsc.org/analyst

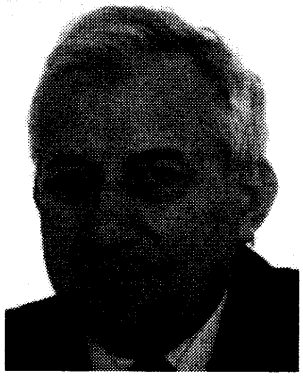
Crucial problems in regression modelling and their solutions

Milan Meloun,*[a] Jiří Militký,[b] Martin Hill[c] and Richard G. Brereton[d]

[a] *Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, CZ532 10 Pardubice, Czech Republic. E-mail: milan.meloun@upce.cz*
[b] *Department of Textile Materials, Technical University, CZ461 17 Liberec, Czech Republic. E-mail: jiri.militky@vslib.cz*
[c] *Institute of Endocrinology, Národní 8, CZ116 94 Prague 1, Czech Republic. E-mail: mhill@endo.cz*
[d] *School of Chemistry, University of Bristol, Cantock's Close, Bristol, UK BS8 1TS. E-mail: r.g.brereton@bristol.ac.uk*

Received 23rd November 2001, Accepted 14th January 2002
First published as an Advance Article on the web 19th March 2002

Professor Milan Meloun, RNDr., PhD., DrSc., graduated from Purkyně University Brno in 1965. In 1973 he obtained his Ph.D., in 1975 an RNDr. and in 1990 a Dr.Sc. degree in Chemometrics and Analytical Chemistry. In 1990 he habilitated Docent and since 1995 has been the Professor of Analytical Chemistry and Chemometrics at the University of Pardubice. He has read Instrumental Methods of Analytical Chemistry and Chemometrics for over 30 years. In 1987–1989 as visiting professor he gave courses on Chemometrics at the Department of Inorganic Chemistry, The Royal Institute of Technology, Stockholm. He has published over 100 papers, co-authored 18 books and 9 textbooks and has given 190 lectures at conferences. His research interests lie in the computer-assisted interpretation of solution equilibria, and regression analysis of potentiometric, spectrophotometric and extraction data, exploratory data analysis, calibration and multivariate spectral resolution. Professor Meloun is secretary of the Chemometrics Section of the Czechoslovak Chemical Society. He is a member of the Editorial Board of the journals Talanta *and* Analitica Chimica Acta. *Further information on Professor Meloun is available on the internet, http://meloun.upce.cz*



DOI: 10.1039/b110779h

Statistical models, particularly regression models, are extremely useful devices for extracting and understanding the essential features of a set of data. These models, however, are nearly always approximate descriptions of more complicated processes, and because of this inexactness the study of the variation in the results of an analysis with minor modifications of the way the problem is formulated becomes important. There are a number of common difficulties associated with real datasets. The first involves the detection and elimination of outliers in the original data. A problem with outliers is that they can strongly influence the model, especially when using least squares criteria, so a multi-step procedure is required, first to identify whether there are any samples that are atypical of the dataset, then to remove them, and finally to reformulate the model. A second problem is that of correlation between parameters in the model. Strong correlation often leads to an unstable model, although the data may be predicted well, there is little physical meaning to the model and predictions on samples left out of the training set can be poor. This article addresses these two problems, surveys several methods, recommends solutions and illustrates these with case studies.

The first part of this paper describes a series of powerful general diagnostics for detecting observations that differ from the bulk of the data. These may be individual observations that do not belong to the general model, *i.e.* influential points or outliers. The identification of influential points and regression diagnostics is a relatively new topic in chemometrics literature, but is rapidly gaining recognition and acceptance by practitioners as a supplement to the traditional analysis of residuals. Outliers in multivariate data can severely affect the results of regression analyses. We think of data as being divided into two classes (1) good observations (the majority of data) reflecting population scatter of data and (2) the outliers (if any), being a part of the so-called influential points. The goal of any outlier detection is to find this true partition and, thus, separate good from outlying observations. The detection, assessment, and understanding of influential points are major problems in regression model building, as is evident from the many measures of influential points that have been proposed in the literature over the last two decades.[3–37] This area was initially studied in regression analysis as a single case approach to the detection of outliers 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] A single case approach to the detection of outliers can, however, fail because of the masking effect, in which outliers go undetected because of the presence of another, usually adjacent, observation. A single masked outlier is easily detected by deletion diagnostics, in which one observation at a time is deleted, followed by the calculation of new residuals and parameter estimates. With two outliers, pairs of observations can be deleted, and the process can be extended to the deletion of several observations at a time. A difficulty both for computation and interpretation is the explosion of the number of combinations to be considered. An alternative is the repeated application of single deletion methods. Regression diagnostics represent procedures for an examination of the *regression triplet* (*data, model, method*) for 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 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 an ideal state are discovered. Our concept of exploratory regression analysis is based on the question: 'does the user know more about the data than the computer?'. Numerous influence measures have been proposed for data analysis assessing the influence of individual cases over the last two decades;[12-35] they also represent a relatively new topic in the chemometrics literature, especially in the last ten years. In this paper influence diagnostics are critically surveyed, commented on and compared using as an illustrative example a model of the cadmium content in wheat for food, in particular the variation of the cadmium content of the ear, the leaf, culm and node, and the root system.

The second part of the paper examines the problem of collinearity in multiple linear regression (MLR), defined as approximate linear dependencies among the independent variables. This problem arises when at least one linear combination of the independent variables is very nearly equal to zero, but the term collinear is often applied to the linear combination of two variables. It is known that given strong multicollinearity the parameter estimates and hypotheses tests are affected more by the linear 'links' between independent variables than by the regression model itself. The classical *t*-test of significance is highly inflated owing to the large variances of regression parameter estimates and the results of statistical analysis are often unacceptable. The problem of multicollinearity has been addressed by means of variable transformation, several biased regression methods, Stein shrinkage,[40] ridge regression,[41,42] and principal component regression and its variations:[43,44] for a brief review, see for example Wold *et al.*[45] Belsey,[46] Bradley and Srivastava,[47] and Seber,[48] among others, have discussed the problems that can be caused by multicollinearity in polynomial regression, and have suggested certain approaches to reduce the undesirable effects of multicollinearity. Although ridge regression has received the greatest acceptance, all have been used with apparent success in various problems. Biased regression methods attack the multicollinearity problem by computationally suppressing the effects of the collinearity, but should be used with caution.[49] While ridge regression does this by reducing the apparent magnitude of the correlations, principal component regression attacks the problem by regressing y on the important component variables to the original variables. In this paper discussion is limited to examples of polynomial regression, but the results can be readily extended to other forms of regression models providing the number of variables is less than the number of samples. Regression estimators based on generalized principal components are adopted. This generalization of classical principal component regression (PCR) avoids problems with 'jump' in regression results due to neglecting small principal components. The results of regression are continuously changed as the precision parameter P varies. Suitable bias selection based on the *mean error of prediction* (MEP) is used. The method of generalized principal component regression is demonstrated on an illustrative example solving a problem in clinical biochemistry: for the age dependence of 17-hydroxypregnenolone, a polynomial regression model was built and the question answered as to whether age-related changes in the concentration of this steroid in men are significant.

1. Indication of single influential points, outliers and high-leverages

1.1 Theoretical

1.1.1 Best linear unbiased estimate of the regression parameters (BLUE). A *linear* regression model is a model which is formed by a linear combination of explanatory variables $\boldsymbol{x}$ or their functions, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. Vector $\boldsymbol{y}$ has dimensions ($n \times 1$) and matrix $\boldsymbol{X}$ dimensions ($n \times m$ and $m < n$). Linear means linear according to model parameters. For linear parameters, the sensitivity:

$$g_j = \frac{\delta f(x,\beta)}{\delta \beta_j} = \text{constant}, j = 1,\ldots,m$$

Individual explanatory variables x_j define geometrically the m-dimensional co-ordinate system or the hyperplane L in n-dimensional Euclidean space E^n. The vector $\boldsymbol{y}$ usually does not have to lie in this hyperplane L. The least-squares method is the most frequently used method in regression analysis and the estimates $\boldsymbol{b}$ of parameters β may be calculated by minimization of the distance between the vector $\boldsymbol{y}$ and the hyperplane L. This is equivalent to finding the minimal length of the *residual vector* $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}}_P$, where $\hat{\boldsymbol{y}}_P = \boldsymbol{Xb}$ is the predictor vector. In Euclidean space the length of vector $\hat{\boldsymbol{e}}$ is expressed by the relation $\sqrt{[\hat{\boldsymbol{e}}, \hat{\boldsymbol{e}}]}$. The square of the length of vector $\hat{\boldsymbol{e}}$ is consistent with the residual sum of squares criterion $U(\boldsymbol{b})$ of the least-squares method, so that the estimates of model parameters $\boldsymbol{b}$ minimize the expression

$$U(\boldsymbol{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_{P,i})^2 = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{m} X_{ij} b_j \right]^2 \approx \text{minimum} \quad (1.1)$$

The residual vector $\hat{\boldsymbol{e}}$ for which the function $U(\boldsymbol{b})$ is minimal lies in an $(n - m)$-dimensional hyperplane $L^{\perp}$ that is perpendicular to the hyperplane L. The perpendicular projection of $\boldsymbol{y}$ into the hyperplane L can be made using projection matrix $\boldsymbol{H}$ and may be expressed[1]

$$\hat{\boldsymbol{y}}_P = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{Hy} \quad (1.2)$$

The conventional least-squares estimator $\boldsymbol{b}$ has the form

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad (1.3)$$

with the corresponding variance

$$D(\boldsymbol{b}) = \sigma^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1} \quad (1.4)$$

Statistical analysis related to least-squares (LS) is based on the normality of estimates $\boldsymbol{b}$. The projection matrix $\boldsymbol{P}$ is orthogonal to the hyperplane L, $\boldsymbol{P} = \boldsymbol{E} - \boldsymbol{H}$, where $\boldsymbol{E}$ is an $(n \times n)$ identity matrix. 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

$$\boldsymbol{y} = \boldsymbol{Hy} + \boldsymbol{Py} = \hat{\boldsymbol{y}}_P + \hat{\boldsymbol{e}} \quad (1.5)$$

The geometric interpretation is that vector $\boldsymbol{y}$ is decomposed into two mutually perpendicular vectors, the prediction vector $\hat{\boldsymbol{y}}_P$ and the vector of residuals $\hat{\boldsymbol{e}}$, Fig. 1.

In the determination of the statistical properties of random vectors $\hat{\boldsymbol{y}}_P$, $\hat{\boldsymbol{e}}$, and $\boldsymbol{b}$, some basic assumptions are necessary for the least-squares method 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 for the measurement of errors is valid, $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$. (3) The matrix of non-random controllable values of the explanatory variable $\boldsymbol{X}$ has a column rank equal to m. This means that the two columns $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ are not collinear vectors. This is the same as saying that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is a symmetric regular invertible matrix with a non-zero determinant. That is, 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 valid for all correlation type models and models having intercept term. (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 and therefore $\text{cov}(\varepsilon_i, \varepsilon_i) = E(\varepsilon_i, \varepsilon_i) = 0$. 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 ε_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 $\sigma^2\boldsymbol{E}$ where $\boldsymbol{E}$ is the identity matrix.

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: (i) The term *best* estimates $\boldsymbol{b}$ means that any linear combination of these estimates has the smallest variance of all the linear unbiased estimates. That is, the variances of the individual estimates $D(b_j)$ are the smallest from all possible linear unbiased estimates (the Gauss-Markov theorem). (ii) 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 $\boldsymbol{x}_j, j = 1, \ldots. m.$ and $\boldsymbol{O} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ for the weight matrix, so we can say $b_j = \sum_{i=1}^{n} Q_{ij} y_i$. 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 $D(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. When condition (7) is valid, all estimates $\boldsymbol{b}$ have a normal distribution, even for a finite sample size n. (iii) The term *unbiased* estimates means that $E(\beta - \boldsymbol{b}) = 0$ and the mean value of an estimate vector $E(\boldsymbol{b})$ is equal to a vector of regression parameters β. It should be noted that there exist *biased estimates*, the variance of which can be smaller than the variance of estimates $D(b_j)$.

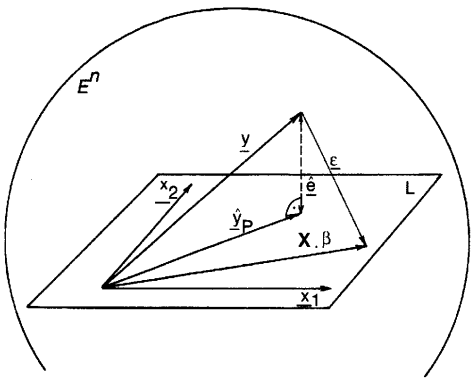


Fig. 1 Geometric illustration of a linear regression model for two independent variables.

1.1.2 Criteria for regression model building. Various test criteria for a search of 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 'cross-validation relationship'

$$\text{MEP} = \frac{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{(i)})^2}{n} \quad (1.6)$$

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

$$\hat{R}_P^2 = 1 - \frac{n - \text{MEP}}{\sum_{i=1}^{n} y_i^2 - n \times \bar{y}^2} \quad (1.7)$$

Another statistical characteristic in quite general use is derived from information theory and entropy,[12] and known as the *Akaike information criterion*,

$$\text{AIC} = n \ln\left(\frac{U(\boldsymbol{b})}{n}\right) + 2m \quad (1.8)$$

The most suitable model is the one which gives the lowest value of the mean quadratic error of prediction MEP and Akaike information criterion (AIC) and the highest value of the predicted determination coefficient, R_p^2.

1.1.3 Residuals analysis. 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] the influential points may instead be classified according to data location into: (i) *Outliers* (denoted in graphs by the letter O), which differ from the other points in value on the y-axis; (ii) *High-leverage points*, also called *extremes* (denoted in graphs by the letter E), which differ from the other points in value on the x-axis or (iii) *Both O and E*, standing for a combination of outlier and high-leverage together. Outlier identification by examination of the residuals is relatively simple, and 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 no account of information contained in $\boldsymbol{y}$ as the high-leverages are found from the diagonal elements H_{ii} of the projection hat matrix $\boldsymbol{H}$.

Analysis of various types of residuals, or some transformation of the residuals, is useful for detecting inadequacies in the model or influential points in the data.

(a) Ordinary residuals. $\hat{e}_i$ are defined by $\hat{e}_i = y_i - \boldsymbol{x}_i\boldsymbol{b}$, where $\boldsymbol{x}_i$ is the ith row of matrix $\boldsymbol{X}$. Classical analysis is based on the wrong assumption that residuals are good estimates of random errors ε_i. Reality is more complex: the residuals $\hat{\boldsymbol{e}}$ are a projection of vector $\boldsymbol{y}$ into a subspace of dimension $(n - m)$, $\hat{\boldsymbol{e}} = \boldsymbol{Py} = \boldsymbol{P}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{P\varepsilon} = (\boldsymbol{E} - \boldsymbol{H})\boldsymbol{\varepsilon}$ and therefore, for the ith residual the following is valid:

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

Each residual $\hat{e}_i$ is a linear combination of all random errors ε_i. The distribution of residuals depends on (i) the random error distribution, (ii) the elements of the projection matrix $\boldsymbol{H}$, (iii) the sample size n. Ordinary residuals have non-constant variance and may not indicate strongly deviant points.

(b) Normalized residuals or scaled residuals. $\hat{e}_{N,i} = \hat{e}_i/\hat{\sigma}$ are often recommended in chemometrics. It is falsely assumed that these residuals are normally distributed quantities with zero mean and variance equal to one, $\hat{e}_{N,i} \approx N(0,1)$. In reality these residuals are correlated and have non-constant variance.

(c) Standardized residuals or internally Studentized residuals. $\hat{e}_{S,i} = \hat{e}_i/(\hat{\sigma}\sqrt{1 - H_{ii}})$ exhibit constant unit variance and their statistical properties are the same as those of ordinary residuals. Here H_{ii} is the ith diagonal element of the $\boldsymbol{H}$ matrix. Standardized residuals behave much like a Student's t random variable except for the fact that the numerator and denominator of $\hat{e}_{S,i}$ are not independent.

(d) Jackknife residuals or externally Studentized residuals. $\hat{e}_{J,i} = \hat{e}_{S,i}\sqrt{\dfrac{n - m - 1}{n - m - \hat{e}_{S,i}^2}}$, are residuals which with an assumption of normality of errors have a Student distribution with $(n - m - 1)$ degrees of freedom. The jackknife residual[25] examines the influence of individual points on the mean quadratic error of prediction, MEP. An approximate rule may be formulated: strongly influential points have squared jackknife residuals $\hat{e}_{J,i}^2$ greater than 10. In the case of high-leverage points, however, these residuals do not give any indication.

(e) Predicted residuals or cross-validated residuals. $\hat{e}_{P,i} = \dfrac{\hat{e}_i}{1 - H_{ii}} = y_i - \boldsymbol{x}_i\boldsymbol{b}_{(i)}$ sensitively monitor the magnitude of shift C in the equation $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Ci} + \boldsymbol{\varepsilon}$, where $\boldsymbol{i}$ is the identity vector with the ith element equal to one and other elements equal to zero. This model expresses the case of an outlier where C is directly equal to the value of deviation, but also the case of a high-leverage point $C = \boldsymbol{d}^T_i\boldsymbol{\beta}$ where $\boldsymbol{d}_i$ is the vector of the deviation of the individual x components of the ith point.

(f) Recursive residuals. have been described by Hedayat and Robson,[26] Brown, Durbin and Ewans.[27] Galpin and Hawkins[28] and Quesenberry.[29] 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 is 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, $\hat{e}_{R,i} = 0$, $i = 1, ..., m$, and then the *recursive residual* is defined as

$$\hat{e}_{R,i} = \frac{y_i - \boldsymbol{x}_i\boldsymbol{b}_{i-1}}{\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}^T_{i-1}\boldsymbol{X}_{i-1})^{-1}\boldsymbol{x}_i^T}}, \quad i = m + 1, \ldots, n \qquad (1.9)$$

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. They allow identification of any instability in a model, for example, instability in time, autocorrelation.

1.1.4 Diagnostics based on the diagonal elements of the hat matrix. Since the introductory paper by Hoaglin and Welsch,[30] the hat matrix $\boldsymbol{H}$ has been studied by many authors from different perspectives. For computational reasons, these measures were originally based on the diagonal elements of H_{ii}. Hoaglin and Welsch[30] suggested declaring observations with $H_{ii} > 2\ m/n$ as high-leverage points. The rationale behind this cut-off point is that m/n is the average of the n diagonal elements of $\boldsymbol{H}$. Therefore, observations with H_{ii} greater than twice the average are declared to be high high-leverage points. Obviously, this cut-off point will fail to nominate any observation when $n \leqslant 2m$, because $0 \leqslant H_{ii} \leqslant 1$.

1.1.5 Diagnostic plots constructed from residuals and hat matrix elements. For analysis of residuals a variety of plots have been widely used in regression diagnostics; Cook and Weisberg,[3] Atkinson,[4] Chatterjee and Hadi,[5] Anscombe,[31] Draper and Smith,[32] Carrol and Ruppert[33] and others. For the identification of influential points, *i.e. outliers* and *high-leverages*, various types of residuals are combined with the diagonal elements H_{ii}, *cf.* page 72 in ref. 1.

(1) The *graph of predicted residuals*[34] has on the x-axis the predicted residuals $\hat{e}_{P,i}$ and on the y-axis the ordinary residuals $\hat{e}_i$. 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.

(2) The *Williams graph*[35] has on the x-axis the diagonal elements H_{ii} and on the y-axis the jackknife residuals $\hat{e}_{J,i}$. Two boundary lines are drawn, the first for outliers, $y = t_{0.95}\ (n - m - 1)$ and the second for high-leverages, $x = 2\ m/n$. Note that $t_{0.95}\ (n - m - 1)$ is the 95% quantile of the Student distribution with $(n - m - 1)$ degrees of freedom.

(3) The *Pregibon graph*[36] has on the x-axis the diagonal elements H_{ii} and on they-axis the square of normalized residuals $\hat{e}_{N,i}^2$. Since the expression $E(H_{ii} + \hat{e}_{N,i}^2) = (m + 1)/n$ 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.

(4) The *McCulloh and Meeter graph*[16] has on the x-axis $\ln[H_{ii}/(m(1 - H_{ii}))]$ and on the y-axis the logarithm of the square of the standardized residuals $\ln(\hat{e}^2_{S,i})$. In this plot the solid line drawn represents the locus of points with identical influence, with slope -1. The 90% confidence 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.

(5) The *Gray's L-R graph*[18] has on the x-axis the diagonal elements H_{ii} and on the y-axis the squared normalized residuals $\hat{e}^2_{N,i} = \hat{e}^2_i/U(\boldsymbol{b})$. All the points will lie under the hypotenuse of a triangle with a 90° angle in the origin of the two axes and the hypotenuse defined by the limiting equality $H_{ii} + \hat{e}^2_{N,i} = 1$. In the Gray's L–R graph, contours of the same critical influence are plotted, and the locations of individual points are compared with them. It may be determined that the contours are hyperbolic as described by $y = \dfrac{2x - x^2 - 1}{x(1 - K) - 1}$, 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.

(6) The *Index graph* has on the x-axis the order index i and on the y-axis the residuals $\hat{e}_{S,i}$, $\hat{e}_{P,i}$, $\hat{e}_{J,i}$, $\hat{e}_{R,i}$, or the diagonal elements H_{ii}, or estimates b_i. It indicates the *suspicious points* only which could be influential, *i.e.* outliers or high-leverages.

(7) The *Rankit graph (Q–Q plot)* has on the x-axis the quantile of the standardized normal distribution u_{Pi} for $P_i = i/(n + 1)$ and on the y-axis the ordered residuals $\hat{e}_{S,i}$, $\hat{e}_{P,i}$, $\hat{e}_{J,i}$, $\hat{e}_{R,i}$ *i.e.* increasingly ordered values of various types of residuals.

1.1.6 Diagnostics based on scalar influence measures. Proper normalization in influence functions[37] leads to scalar measures. These measures express the relative influence of the given point on all parameter estimates.

(1) The *Cook measure* D_i[25] expresses directly the relative influence of the ith point on all parameter estimates and has the form

$$D_i = \frac{(\boldsymbol{b} - \boldsymbol{b}_{(i)})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{b} - \boldsymbol{b}_{(i)})}{m \times \hat{\sigma}^2} = \frac{\hat{e}_{S,i}}{m} \times \frac{H_{ii}}{1 - H_{ii}} \qquad (1.10)$$

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 $\hat{\sigma}^2$. It is generally useful to study cases that have $D_i > 0.5$ and it is always important to study cases with $D_i > 1$. These benchmarks are intended as an aid in finding influential cases, but they do not represent a test. There is no significance test associated with D_i.

(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

$$A_i = |\hat{e}_{J,i}| \times \sqrt{\frac{n - m}{m} \times \frac{H_{ii}}{1 - H_{ii}}} \qquad (1.11)$$

This measure is also convenient for graphical interpretation; Atkinson recommends that signed 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 $\hat{e}_{J,i}$ and A_i could also be large because the ith jackknife residual is large. Large jackknife residuals are due to outliers, points whose response falls far from the fitted function.

(3) The *Belsey DFFITS$_i$ measure*, also called *Welsch–Kuh's distance*,[2] is obtained by 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

$$\text{DFFITS}_i^2 = \hat{e}^2_{J,i} \times \frac{H_{ii}}{1 - H_{ii}} \qquad (1.12)$$

Belsey, Kuh, and Welsch[2] suggest the test that the ith point is considered to be significantly influential on prediction $\hat{y}_P$ when DFFITS$_i$ is larger in absolute value than $2\sqrt{m/n}$.

(4) The *Anders–Pregibon diagnostic* AP_i[36] expresses the influence of the ith point on the volume of the confidence ellipsoid

$$\text{AP}_i = \frac{\det(\boldsymbol{X}^T_{m(i)} \boldsymbol{X}_{m(i)})}{\det(\boldsymbol{X}^T_m \boldsymbol{X}_m)} \qquad (1.13)$$

where $\boldsymbol{X}_m = (\boldsymbol{x}|\boldsymbol{y})$ 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 $\boldsymbol{H}_m$ by the expression $\text{AP}_i = 1 - H_{ii} - \hat{e}^2_{N,i} = 1 - H_{m,ii}$. A point is considered to be influential if $H_{m,ii} = 1 - \text{AP}_i > 2(m + 1)/n$

(5) The *Cook–Weisberg likelihood measures* LD_i[36] represent a general diagnostic defined by

$$\text{LD}_i = 2[L(\hat{\Theta}) - L(\hat{\Theta}_{(i)})] \qquad (1.14)$$

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 the corresponding value when the ith point is omitted. For strongly influential points $\text{LD}_i > \chi^2_{1-\alpha}(m + 1)$ 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 the parameter estimates or on the variance estimate or on both:[36](a) *The likelihood measure* $LD_i(\boldsymbol{b})$ examines the influence of individual points on the parameter estimates $\boldsymbol{b}$ by the relationship

$$\text{LD}_i(\boldsymbol{b}) = n \times \ln\left[\frac{d_i - H_{ii}}{1 - H_{ii}} + 1\right] \qquad (1.15)$$

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

(b) *The likelihood measure* $LD_i(\hat{\sigma}^2)$ examines the influence of individual points on the residual variance estimates by the relationship

$$\text{LD}_i(\hat{\sigma}^2) = n \times \ln\left[\frac{n}{n - 1}\right] + n \ln(1 - d_i) + \frac{d_i(n - 1)}{1 - d_i} - 1 \qquad (1.16)$$

(c) *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

$$\text{LD}_i(\boldsymbol{b}, \hat{\sigma}^2) = n \times \ln\left[\frac{n}{n - 1}\right] + n \ln(1 - d_i) + \frac{d_i(n - 1)}{(1 - d_i)(1 - H_{ii})} - 1 \qquad (1.17)$$

1.2 Methodology

1.2.1 Procedure for regression model building. The procedure for examination of influential points in data and the construction of a linear regression model consists of the following steps:

Step 1. Proposal of a model for original data:. 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. Exploratory data analysis in regression provides a scatter plot of individual variables and all possible pair combinations are examined. Also, in this step the influential points causing multicollinearity are detected.

Step 2. Significance test of parameter estimates:. the parameters of the proposed regression model and the corresponding basic statistical characteristics of this model are determined by the ordinary least-squares method (OLS). Individual parameters are tested for significance by using the Student *t*-test. The following are computed: the correlation coefficient *R* and the determination coefficient or, multiplied by 100%, the regression rabat *100D*. The mean quadratic error of prediction, MEP, and the Akaike information criterion, AIC, are calculated to examine the quality of the model.

Step 3. Detection of influential points:. the statistical analysis of ordinary residuals, different diagnostic graphs and numerical measures are used to examine influential points, namely outliers and leverages. If outliers 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 4. Construction of a more accurate model:. according to the test for fulfillment of the conditions for the least-squares method, and the results of regression diagnostics, a more accurate regression model is constructed.

1.2.2 Software used. For the creation of regression diagnostic graphs and computation of the regression based characteristics, an algorithm in *S-Plus* was written, and the linear regression module of the *ADSTAT* package used, *cf.* ref. 38.

1.3 Case study

1.3.1 Dataset: Cadmium content in wheat for food and the variation of its content in the ear, stem and leaf, and root. The cadmium content was examined in samples of food wheat to determine its variation in the ear, the stem and leaf, and the root system. Cadmium content was determined quantitatively in the grain of wheat *y*; in the ear, *i.e.* in the part that contains the seeds, x_1; in the stem and leaves x_2; and in the root x_3. The main aim is to propose a regression model and to find influential points in the data. Table 1

Table 1 Dataset for cadmium content in the ear of corn x_1, in the stem and leaves of grain x_2, in the root of plant system x_3 and in the corn, grain of wheat *y*

Cadmium content/mg dm^{-3}			
Ear of corn x_1	Stem and leaves of grain x_2	Root of plant system x_3	Corn, grain of wheat *y*
1.50	1.50	1.50	1.60
1.50	1.60	1.30	1.60
2.00	1.90	2.20	2.10
2.00	2.00	2.20	2.10
6.60	7.10	7.60	8.10
7.10	8.20	6.60	7.90
7.80	9.10	7.10	8.40
8.40	10.30	7.80	10.30
8.40	9.60	8.60	9.60
8.60	10.00	9.10	10.80
9.00	12.30	10.50	13.10
10.20	13.10	11.80	15.10
1.30	1.30	1.30	1.30
1.10	1.30	1.20	1.20
1.30	1.60	1.30	1.50
1.50	1.60	1.20	1.50

1.3.2 Proposal of a model for original data. Using the original set of data, the ordinary least-squares method OLS finds the regression model

$$y = -0.073(0.138, \text{R}) - 0.685(0.192, \text{A})\, x_1 + 0.896(0.161, \text{A})\, x_2 + 0.838(0.133, \text{A})\, x_3$$

where the standard deviations of the parameters estimated, are in brackets and the letter R means that β is rejected as a statistically nonsignificant estimate while the letter A means that β is accepted as a statistically significant estimate.

1.3.3 Significance test of parameter estimates. The critical quantile $t_{0.975}(16-4) = 2.179$ of a Student *t*-test at 5% significance level was used to examine the statistical significance of the individual regression parameter estimates: $t_0 = 0.5269$, $t_1 = 3.5746$, $t_2 = 5.5761$, $t_3 = 6.2879$. All values except t_0 are not less than $t_{0.975}(16-4)$ and therefore estimates of parameters β_1, β_2 and β_3 are significant (denoted by the letter A in brackets) while the estimate of parameter β_0 is not significant (denoted by the letter R). The model was described with the correlation coefficient $R = 0.9986$, the determination coefficient $D = 99.72\%$ thus expressing a percentage of points which fulfil the model proposed; the mean error of prediction, MEP = 0.2101, the Akaike information criterion, AIC = −36.18 and the residual standard deviation $s(e) = 0.290$ were also calculated. All these statistics can be used as resolution criteria for the selection of the best model among several plausible ones.

1.3.4 Detection of influential points. (*a*) *Residual analysis*: generally it is valid that outliers are identified by an examination of the residuals while the high-leverage points are found from the diagonal elements H_{ii} of the projection hat matrix, Table 2.

A survey of all the diagnostics for detection of influential points shows that diagnostics plots are the most efficient because they are able to separate influential points into outliers and high-leverages. Table 2 gives numerical values of various types of residuals and diagnostics of influential points. Suspicious points are written in italics. Several types of residuals can be used in statistical tests, and the influential points IP found are written in bold in Table 2. The H_{ii} indicates only leverages and residuals $\hat{e}_J$, $\hat{e}_S$, $\hat{e}_P$ only outliers, while the remaining diagnostics indicate both outliers and leverages, together.

A survey of suspicious points identified by various types of diagnostic measures is given in Table 3. It is clear that there are some local differences arising from the severity of cut-off for individual values but the majority of measures indicate the same points.

Ordinary residuals (Fig. 2a, b) are always associated with a non-constant variance; they may not indicate strongly deviant points. Even though the common practice of chemometrical programs for the statistical analysis of residuals is to examine by use of statistical characteristics such as the mean $\bar{e}$, the variance $\hat{s}^2(e)$, the skewness $\hat{g}_1(e)$ and the kurtosis $\hat{g}_2(e)$, these statistics do not give a correct indication of the influential points. Points (7, 8, 9, 10, 11, 12) may be considered to be suspicious and some testing diagnostics for influential points should be applied.

In the case of *normalized residuals* (Fig. 2c), the rule of 3σ is classically recommended: outliers are quantities with $\hat{e}_{N,i}$ of magnitude greater than $\pm 3\sigma$ of all values and lie outside the interval $\bar{e} \pm 3\hat{\sigma}$. Such assumptions about normalized residuals are misleading. Points (2, 8, 9, 10, 11, 12) may be denoted as suspicious in this graph. However, normalized residuals are not able to indicate high-leverage points.

The statistical properties of *standardized residuals* (Fig. 2d) are the same as those of the ordinary residuals. The maximum values of $\hat{e}_S$ are bounded $\sqrt{n-m} = 3.46$. This influential points criterion also seems to be misleading.

Table 2 A survey of the influential points which were indicated with the use of various tabular diagnostic tools. *Suspicious points* (SP, written in italics) are data points which obviously differ from the others. *Influential points* (IP, written in bold) are points which are detected and separated into outliers and high-leverages with the use of various testing criteria: $n = 16$, $m = 4$, $\hat{e}_i$: no testing limit for IP, it detects SP only; $\hat{e}_{S,i}$: no testing limit for IP, it detects SP only; $\hat{e}_{J,i}$: when $\hat{e}_{J,i}^2 > 10$ then the i-th point is an outlier; $\hat{e}_{P,i}$: no testing limit for IP, it detects SP only; H_{ii}: when $H_{ii} > 2\ m/n = 0.5$ then the i-th point is a high-leverage; $H_{m,\,ii}$: when $H_{m,\,ii} > 2\ m/n = 0.5$ then the i-th point is a high-leverage; D_i: when $D_i > 1$ then the i-th point is an IP; DFFITS$_i$: when $|\text{DFFITS}_i| > 2\sqrt{m/n} = 1$ then the i-th point is an IP; AP$_i$: when AP$_i$ < 1−2 $(m + 1)/n = 0.375$ then the i-th point is an IP; LD$_i$: generally, when LD$_i$ > $\chi^2_{1-\alpha}(m + 1) = 11.07$ then the i-th point is an IP

i	y_i	$\hat{y}_{P,i}$	$s(y_i)$	$\hat{e}_i$	$\hat{e}_{S,i}$	$\hat{e}_{J,i}$	$\hat{e}_{P,i}$	$H_{i,i}$	$Hm_{i,i}$	D_i	A_i	AP$_i$	DFFITS$_i$	LD$_i$(b)	LD$_i$(s^2)	LD$_i$(b,s^2)
1	1.60	1.50	0.10	0.10	0.37	0.35	0.11	0.12	0.13	0.00	0.23	0.87	0.13	0.03	0.02	0.05
2	1.60	1.42	0.11	0.18	0.66	0.64	0.20	0.13	0.16	0.02	0.43	0.84	0.25	0.09	0.01	0.09
3	2.10	2.10	0.11	0.00	−0.01	−0.01	0.00	0.14	0.14	0.00	0.01	0.86	0.00	0.00	0.03	0.03
4	2.10	2.19	0.10	−0.09	−0.34	−0.33	−0.11	0.13	0.14	0.00	0.22	0.86	−0.13	0.02	0.02	0.05
5	8.10	8.14	0.21	−0.04	−0.18	−0.17	−0.07	**0.52**	**0.52**	0.01	0.30	0.48	−0.17	0.04	0.03	0.07
6	7.90	7.94	0.13	−0.04	−0.16	−0.15	−0.05	0.21	0.21	0.00	0.14	0.79	−0.08	0.01	0.03	0.04
7	8.40	8.69	0.16	*−0.29*	*−1.19*	*−1.22*	*−0.42*	0.31	0.39	0.16	*1.42*	0.61	−0.82	0.84	0.03	0.95
8	10.30	9.94	0.18	*0.36*	*1.59*	*1.71*	*0.59*	0.38	**0.51**	*0.39*	*2.32*	0.49	**1.34**	1.94	0.25	2.69
9	9.60	9.98	0.14	*−0.38*	*−1.48*	*−1.57*	*−0.49*	0.22	0.36	0.15	*1.44*	0.64	−0.83	0.80	0.16	1.10
10	10.80	10.62	0.13	0.18	0.69	0.68	0.22	0.20	0.23	0.03	0.59	0.77	0.34	0.16	0.00	0.16
11	13.10	13.58	0.23	*−0.48*	*−2.68*	**−4.06**	*−1.26*	**0.62**	**0.85**	**2.91**	**8.94**	**0.15**	**−5.16**	*10.86*	*7.83*	**44.18**
12	15.10	14.57	0.19	*0.54*	*2.40*	**3.19**	*0.91*	0.41	**0.69**	**1.01**	**4.62**	**0.31**	**2.67**	*4.63*	*3.44*	**13.13**
13	1.30	1.29	0.11	0.01	0.03	0.03	0.01	0.13	0.13	0.00	0.02	0.87	0.01	0.00	0.03	0.03
14	1.20	1.34	0.12	−0.14	−0.54	−0.53	−0.17	0.16	0.18	0.01	0.40	0.82	−0.23	0.07	0.01	0.08
15	1.50	1.56	0.12	−0.06	−0.22	−0.22	−0.07	0.16	0.17	0.00	0.16	0.83	−0.10	0.01	0.03	0.04
16	1.50	1.34	0.11	0.16	0.60	0.58	0.19	0.14	0.17	0.02	0.41	0.83	0.24	0.08	0.01	*10.09*

Table 3 A survey of the influential points which were indicated using various graphical diagnostic tools: *Suspicious points* (SP) are data points in diagnostic graphs which obviously differ from the others; *influential points* (IP) are data points which are detected and are separated into outliers and high-leverages using the following testing criteria: $n = 16$, $m = 4$, 1. *Graph of predicted residuals:* outliers are far from the central pattern on the line $y = x$; 2. *Williams graph*: the first line is for outliers, $y = t_{0.95}(n - m - 1) = 1.796$, the second line is for high-leverages, $x = 2\ m/n = 0.5$; 3. *Pregibon graph*: two constraining lines are drawn, $y = -x + 2(m + 1)/n$, and $y = -x + 3(m + 1)/n$, a strongly influential point is above the upper line; an influential point is between the two lines; 4. *McCulloh and Meeter graph*: the 90% confidence line is for outliers, $y = -x - \ln F_{0.95}(n - m, m)$ while the boundary for high-leverages is $x = \ln[2/(n - m) \times (t^2_{0.95}(n - m)]$; 5. *Gray's L–R graph*: points towards the upper corner are outliers while those towards the right angle of the triangle are high-leverages; 6. D_i: when $D_i > 1$ then the i-th point is an IP; 7. A_i: when $A_i^2 > 10$ then the i-th point is an outlier; 8. DFFITS$_i$: when $|\text{DFFITS}_i| > 2\sqrt{m/n} = 1.0$ then the i-th point is an IP; 9. AP$_i$: when AP$_i$ < $1 - 2(m + 1)/n = 0.375$ then the i-th point is an IP; 10., 11. and 12. LD$_i$: generally, when LD$_i$ > $\chi^2_{1-\alpha}$ $(m + 1) = 11.07$ then the i-th point is an IP. 13. $\hat{e}$: detects SP only; 14. $\hat{e}_N$: when $\hat{e}_{N,i} > |3\sigma|$ then the i-th point is an outlier; 15. $\hat{e}_S$: detects SP only; 16. $\hat{e}_J$: when $\hat{e}_{J,i}^2 > 10$ then the i-th point is an outlier; 17. $\hat{e}_P$: detects SP only; 18. H_{ii}: when $H_{ii} > 2m/n = 0.5$ then the i-th point is a high-leverage; 19. $H_{m,\,ii}$: when $H_{m,ii} > 2m/n = 0.5$ then the i-th point is a high-leverage, 20., 21. and 22. the *rankit graph (Q–Q plot)* examines whether the ordered residuals $\hat{e}_{S,i}$, $\hat{e}_{P,i}$, $\hat{e}_{N,i}$ exhibit a normal distribution

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

For *jackknife residuals* (Fig. 2e) an approximate rule may be applied: strongly influential points (*i.e.* outliers) have $\hat{e}_{J,i}^2 > 10$ but for high-leverages, however, these residuals do not give any indication: according to this criterion the points $\hat{e}_{J,11} = -4.06$ and $\hat{e}_{J,12} = 3.19$ are outliers.

Predictive residuals are able to find suspicious points only (8, 11, 12) as is shown in Fig. 2f.

(*b*) *Diagnostic plots constructed from residuals and hat matrix elements:* a combination of various types of residuals with the diagonal elements of the projection hat matrix $\boldsymbol{H}_{ii}$ leads to five diagnostic graphs of influential points (the data set of size $n = 16$, $m = 4$):

The *graph of predicted residuals* (Fig. 3a), one of the simplest graphs, separates outliers (8, 11, 12) located far from its central pattern on the line $y = x$ from high-leverage points (11), outside and far from the line $y = x$.

The *Williams graph* (Fig. 3b) has two testing boundary lines, the first line for outliers $y = t_{0.95}(n - m - 1) = 1.796$ detecting two outliers (11, 12), and the second for high-leverage points $x = 2m/n = 0.5$ detecting two high-leverages (5, 11).

The *Pregibon graph* (Fig. 3c) is able to distinguish strongly influential points from medium influential points only. The points (11, 12) were found as medium influential.

The *McCulloh-Meeter graph* (Fig. 3d) has two testing boundary lines, the first for outliers $y = \ln[(n - m)t^2_{0.95}(n - m)] = 4.043$ behind which two outliers were indicated and the second for high-leverages $x = \ln[2/(n - 2m)] = -1.386$ behind which two high leverages are found (5, 11).

Gray's L–R graph (Fig. 3e) indicates strongly influential points (8, 11, 12) and separates them into outliers (11, 12) points which lie high in the y-axis, and high-leverages (5, 11) which lie in direction of the x-axis.

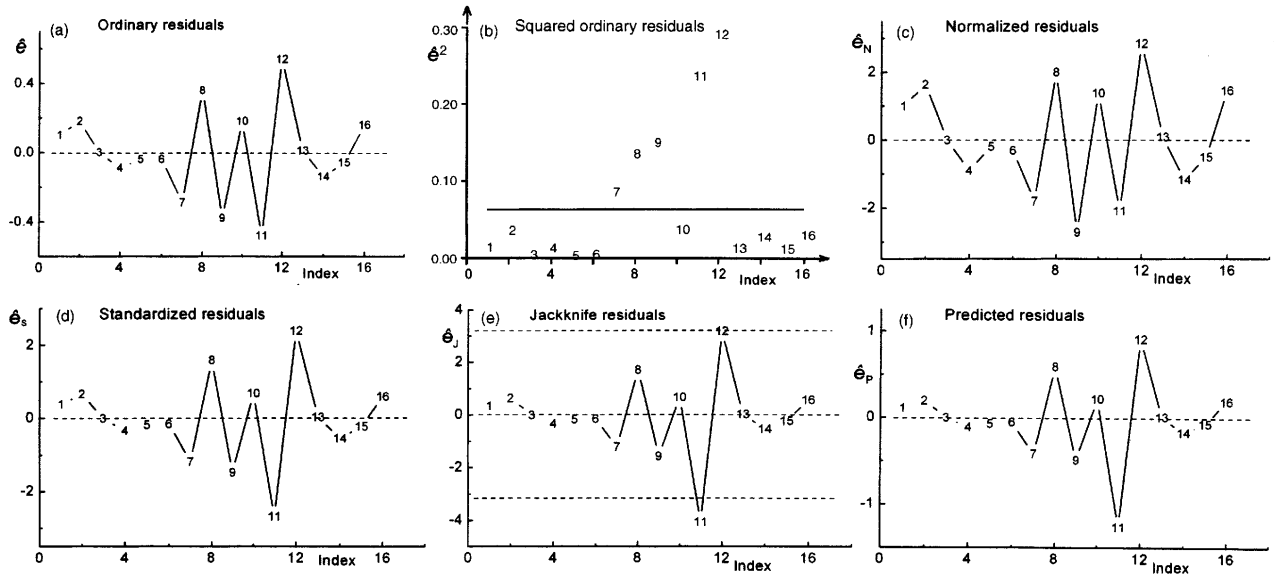


Fig. 2 Index graphs of various residuals for the data set of Example 1: (a) Ordinary residuals; (b) square of ordinary residuals; (c) normalized residuals; (d) standardized residuals; (e) jackknife residuals; (f) predicted residuals.

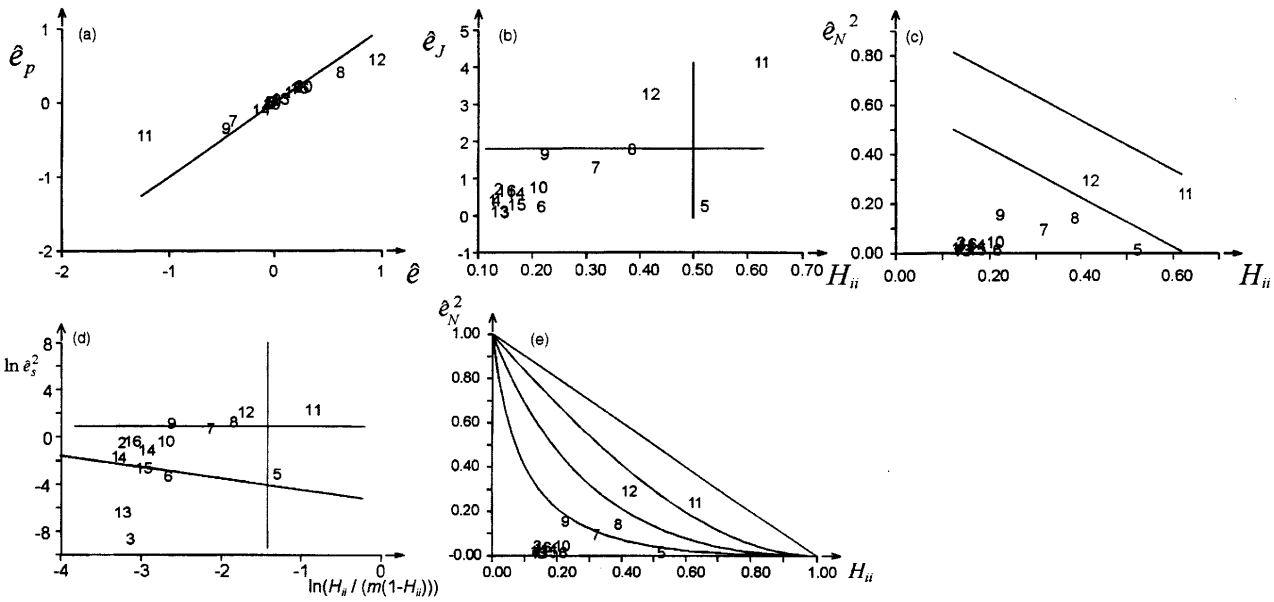


Fig. 3 Diagnostics based on residual plots and hat matrix elements for the data set of Illustrative example 1.1: (a) Graph of predicted residuals, (b) Williams graph, (c) Pregibon graph, (d) McCulloh and Meeter graph, (e) Gray's L–R graph.

(c) *Diagnostics based on scalar influence measures:* in the classification of influential points, it is important to remember that they can affect the various regression characteristics differently. Points affecting the prediction $\hat{y}_{P,i}$, for example, may not affect the parameter variance. The degree of influence of individual points can be classified according to those characteristics that are affected. For the identification of influential points, there are many additional diagnostics which may be divided according to two principal approaches: the first is based on the examination of changes which occur when certain influential points are excluded while the second concerns the validity of the regression model when the variance of errors is abnormal, the so-called the *model of inflated variance*.

For analysis of *the diagonal elements of the projection hat matrix* (Fig. 4a,b) the rule is valid that when $H_{ii} > 2m/n = 0.5$ holds, the actual i-th point is the high-leverage. From that point of view, the points 5 and 11 are high-leverages. For more complex analysis, it is useful to form the extension of matrix $\boldsymbol{X}$ by a vector $\boldsymbol{y}$ to give the matrix $\boldsymbol{X}_m = (\boldsymbol{X}|\boldsymbol{y})$, and the resulting matrix contains the diagonal element $H_{m,ii} = H_{ii} + \hat{e}_i^2/[(n - m)\hat{\sigma}^2]$. According to the same rule, $H_{m,ii} > 2m/n = 0.5$, the diagonal elements of the extended hat matrix $H_{m,ii}$ detect both outliers and high-leverages, (5, 8, 11, 12).

The *Cook measure* D_i (Fig. 4c) is used in connection with an approximative rule: when $D_i > 1$, the shift of parameter estimate $\boldsymbol{b}$ only is greater than the 50% confidence region and the relevant i-th point is rather influential. According to this rule points (11, 12) are influential.

With designed experiments, usually $H_{ii} = m/n$, the *Atkinson measure* (Fig. 4d) is numerically equal to the jackknife residual $\hat{e}_J$. The same rule $\hat{e}_{J,i}^2 > 10$ for the detection of influential points may be used, and points 11 and 12 were found influential.

In the case of *Belsey's DFFITS measure* (Fig. 4e) the i-th point is tested and found to be significantly influential when $\text{DFFITS} > 2\sqrt{m/n} = 1$ is true. Three influential points were indicated (8, 11, 12) with the DFFITS measure.

According to the *Anders–Pregibon measure* (Fig. 4f), the i-th point is tested and considered to be influential if $AP_i < 1 - 2(m + 1)/n = 0.375$, and two influential points were indicated, (11, 12).

There are three *Cook–Weisberg likelihood measures, i. e.* $LD_i(\boldsymbol{b})$ on Fig. 4g, $LD_i(s^2)$ on Fig. 4h and $LD_i(\boldsymbol{b}, s^2)$ on Fig. 4i. All three measures indicate the i-th influential point if it is generally valid that $LD_i > \chi^2(m+1) = 11.07$. According to that criterion $LD_i(\boldsymbol{b})$ detected one influential point (11), $LD_i(s^2)$ two suspicious points (11, 12) and $LD_i(\boldsymbol{b}, s^2)$ two influential points (11, 12).

If the regression model is correct and if there are no influential points then the *rankit Q–Q graph* (Fig. 5) forms a characteristic sigmoidal curve with quite a long linear straight line in the middle part of the graph. The rankit Q–Q graph of jackknife residuals is not among the best diagnostic graphs for influential points. It is based on the phenomenon that the residuals should exhibit a normal distribution. Three suspicious points (8, 11, 12), however, do not fulfil this assumption and therefore they could be tested as they are of an influential nature. The influential points (11, 12) indicated are also located beyond the ends of the straight line on the Q–Q graph of predicted and normalized residuals (Fig. 5b and Fig. 5c).

1.3.5 Construction of a more accurate regression model. The biological meaning of the intercept term β_0 is the cadmium content in wheat corn y when the cadmium content in the ear of corn is zero that in the stem and leaf is zero and that in the root

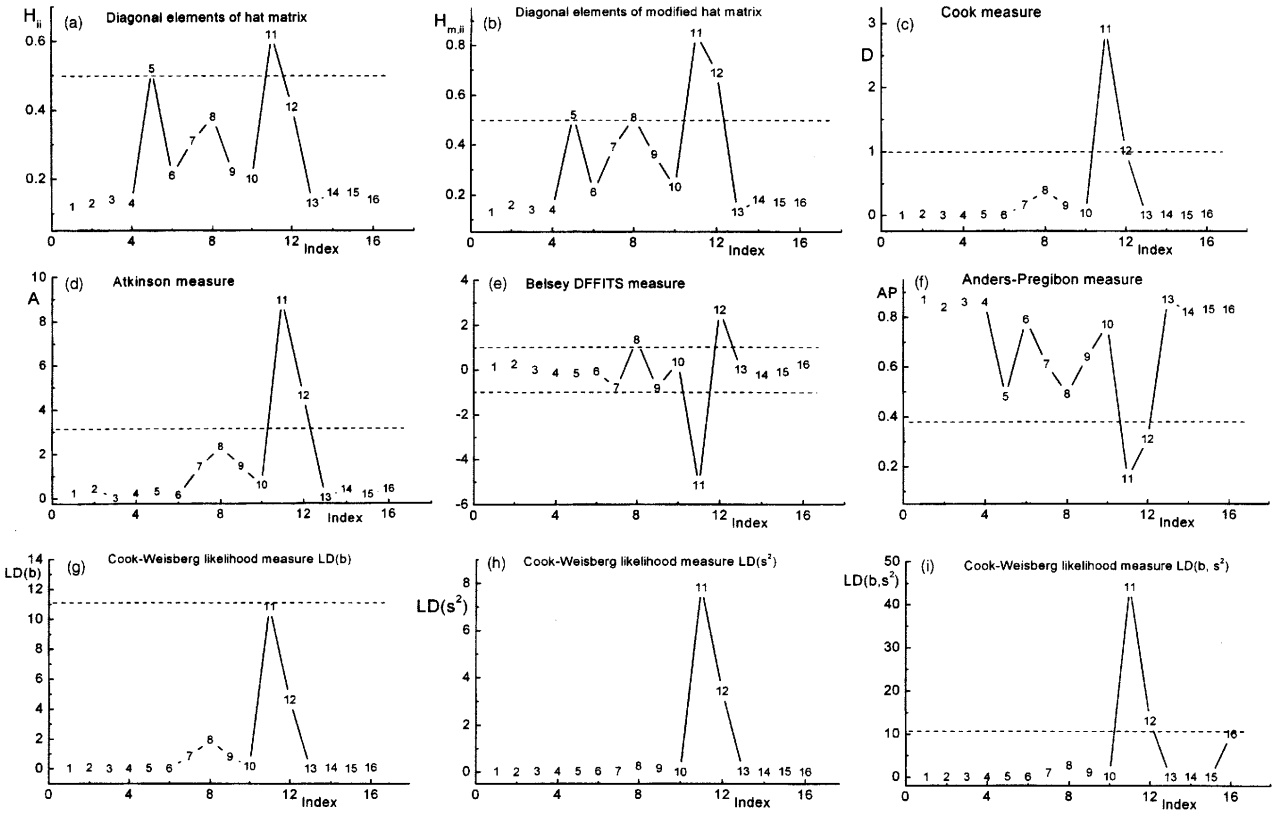


Fig. 4 Index graphs of vector and scalar influence measures: (a) Diagonal elements of the hat matrix; (b) diagonal elements of the modified hat matrix; (c) Cook measure; (d) Atkinson measure; (e) Belsey's DFFITS measure; (f) Anders–Pregibon measure; (g) Cook–Weisberg likelihood measure LD(**b**); (h) Cook–Weisberg likelihood measure LD(s^2); (i) Cook–Weisberg likelihood measure LD(**b**, s^2).

system is also zero. Under such circumstances the cadmium content y must also be equal to zero, and therefore β_0 should be equal to zero. The revised model will then be regarded without this intercept term β_0. Since outliers may influence the regression results they should be treated with care. There are two possible approaches to the data: either to exclude outliers from data or to use robust regression method. One of the greatest disadvantages of the application of robust method is a preference for the regression model proposed, here $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$. If a proposed model is unsuitable (here it is the presence of false parameter β_0), robust methods lead to the suppression of the influence of both individual points and influential points, and therefore also to a suppression of the detection of unsuitable proposed models. Therefore, robust methods should be applied only with careful regard to the peculiarities of the model and data.

On the basis of previous graphical and numerical diagnostics of influential points it may be concluded that the three outliers 8, 11, 12 should be excluded from the original data set, and new parameter estimates should be calculated, Table 4.

1.4 Conclusions

In the interactive PC-aided diagnosis of data, model and estimation method, the examination of data quality involves the detection of *influential points*, outliers and leverages, which cause many problems in regression analysis by shifting the parameter estimates or increasing the variance of the parameters. Regression diagnostics represent the graphical procedures and numerical measures for an examination of the *regression triplet i. e.*, an identification of (i) the data quality for a proposed model, (ii) the model quality for a given data set, (iii) a fulfillment of all least-squares assumptions. Regression diagnostics do not require a knowledge of alternative hypotheses for testing or fulfilling the other assumptions of classical statistical tests. The various types of residuals differ in 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. Data multicollinearity and generalized principal component regression

2.1 Theoretical

2.1.1 Terminology in multiple linear regression. Even if all the assumptions of section 1.1.1 are valid, there may still be significant numerical difficulties with the OLS (ordinary least squares) parameter estimates $\boldsymbol{b}$ found by minimization of the sum of the squared residuals RSS = $U(\boldsymbol{b})$ for various reasons. The reasons for numerical difficulties in the computer evaluation of parameter estimates $\boldsymbol{b}$ are as follows: (1) Neglect of the limited precision of the computer in building the matrix $\boldsymbol{X}^T\boldsymbol{X}$. (2) Inconvenient numerical procedures for matrix inversion or for solving the set of linear equations. (3) Multicollinearity leading to the ill-conditioning of matrix $\boldsymbol{X}^T\boldsymbol{X}$. (4) The linear dependence of some columns of matrix $\boldsymbol{X}^T\boldsymbol{X}$, leading to its non-invertability because of a singularity.

For these reasons, there can be difficulties in producing a stable model. Although a good least squares fit for the data may indicate a certain level of success, if the parameter estimates are to be interpreted physically or used to predict unknown samples, there can be serious problems in the validity of the model. For a test of the simple hypothesis H_0: $\beta_j = \beta_{j,0}$ against the alternative H_A: $\beta_j \neq \beta_{j,0}$, the t-test criterion is defined by

$$t_j = \frac{|b_j - \beta_{j,0}|}{\hat{\sigma}\sqrt{c_{jj}}} \qquad (2.1)$$

where c_{jj} is the jth diagonal element of the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and an unbiased estimate of the variance of errors σ^2 is defined as σ^2 = RSS/$(n - m)$ When H_0 is valid t_j has approximately the Student t-distribution with $(n - m)$ degrees of freedom.

2.1.2 Origins of multicollinearity. The multicollinearity problem in regression refers to the set of problems created when

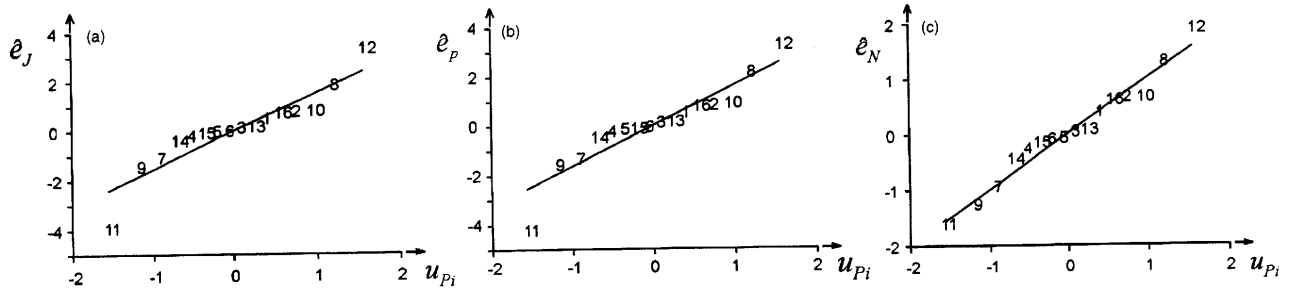


Fig. 5 Rankit Q–Q graph of (a) jackknife residuals., (b) predicted residuals, (c) normalized residuals.

Table 4 Estimates of four unknown parameters of the linear regression model before and after outliers removal in a process of regression model building and testing

Parameter estimate		Standard deviation	Student t-test criterion	Statistical significance of parameter	Estimated significance level
(1) The original data and the model $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$ used: OLS: $t_{1-0.05/2}$ (16−4) = 2.179 R = 0.9986, D = 99.72%, MEP = 0.21014, AIC = −36.18, $s(e)$ = 0.290					
β_0	−0.0727	0.1379	−0.5269	Insignificant	0.608
β_1	−0.6851	0.1917	−3.5746	Significant	0.004
β_2	0.8962	0.1607	5.5761	Significant	0.000
β_3	0.8377	0.1332	6.2879	Significant	0.000
(2) The data without outliers 8, 11, 12 and the model without β_0 used: OLS: $t_{1-0.05/2}$(13−3) = 2.228, R = 0.9992, D = 99.83%, MEP = 0.05101, AIC = −43.55, $s(e)$ = 0.170					
β_0	0.0000	—	—	—	—
β_1	−0.8545	0.3715	−2.3002	Significant	0.044
β_2	0.9542	0.2509	3.8038	Significant	0.003
β_3	0.9155	0.1322	6.9263	Significant	0.000

there are near-singularities among the columns of the X matrix; certain linear combinations of the columns of X are nearly zero. This implies that there are near redundancies among the independent variables; essentially, the same information is being provided in more than one way. Geometrically, collinearity results when at least one dimension of the X-space is very poorly defined in the sense that there is almost no dispersion among the data points in that dimension.

Multicollinearity does not mean a violation of the assumptions for the least-squares methods (LS), .[1] The columns of matrix X are understood as the column vectors which define the hyperplane L in n-dimensional Euclidean space E^n. According to the angle θ_{jk} between two vectors x_j and x_k (or between columns of matrix X) two limiting cases may be distinguished:

(1) *Orthogonality* occurs when the cosine of angle θ_{jk} is zero, $\cos\theta_{jk} = \frac{< x_j, x_k >}{\|x_j\| \times \|x_k\|}$, and also the scalar product $< x_j, x_k > = 0$, where $\|x_j\| = \sqrt{< x_j, x_k >}$ is the length of vector x_j. If all the columns of matrix X are mutually orthogonal, then the matrix X^TX is diagonal and the regression analysis simplifies.

(2) *Collinearity* occurs when the cosine of angle θ_{jk} is equal to 1, $\cos\theta_{jk} = 1$, because the angle between vectors x_j and x_k is zero, $\theta_{jk} = 0$, and the two vectors x_j and x_k are parallel *i.e.* linearly dependent, and the following expression holds for them

$$c_j x_j + c_k x_k = 0 \qquad (2.2)$$

where c_j and c_k are nonzero constants. When eqn. (2.2) holds for q pairs of columns of matrix X, its rank is equal to $m - q$ and the matrix X^TX is singular.

Eqn. (2.2) may be valid for more vectors still, when one of the columns x_j is the result of a linear combination of several columns. This situation is called *perfect collinearity*. The term multicollinearity, however, can include other cases when some columns of matrix X have nearly zero angle and are therefore approximately linearly dependent,

$$\sum_{j=1}^{m} c_j x_j \approx \delta \qquad (2.3)$$

where δ is the vector with components near zero, and the vector c with elements c_j is nonzero, $\|c\| \gg \|\delta\|$. The multicollinearity causes ill-conditioning of the matrix X^TX, and has two consequences: (a) the determinant of matrix X^TX is close to zero; (b) some of the first m eigenvalues of matrix X^TX are close to zero (note that there will never be more than m non-zero eigenvalues).

Multicollinearity causes many difficulties in inversion of matrix X^TX and also numerical errors, depending on the quality of the algorithm for matrix inversion and the machine precision of the computer used. Multicollinearity causes also the following statistical difficulties: (a) *Instability of parameter estimates* is caused by the great sensitivity of parameter estimates to small changes in the data.[50–52] The estimates often have the wrong sign and magnitude, and this damages their physical interpretation. (b) *Large variances* $D(b_j)$ of individual estimates cause t-tests to indicate that parameter β_j in regression model $y = \beta x + \varepsilon$ is statistically insignificant. (c) *Strong correlation* between elements of the estimates vector b means that they cannot be interpreted separately. (d) *Dangerous extrapolation:* prediction is restricted to points within the sample X-space. Extrapolation beyond the data is dangerous in any case, but can quickly lead to serious errors of prediction when the regression equation has been estimated from highly collinear data.

With reference to multicollinearity in data, we can identify three cases of interest: (a) The *over-estimated regression model* contains too many controllable variables expressing the same basic factors. An example is a structure/properties model in which properties of substances are described by various measureable changeable structures. Generating new variables as transformations of other variables can produce a multicollinearity among the set of variables involved. Ratios of variables or powers of variables will frequently be nearly collinear with the original variables.[53](b) *The inappropriate location of experimental points* causes multicollinearity to form 'artificially' because of the choice of location of points. Often the values of significantly important variables oscillate in a small range and seem to be nearly constant, and they are collinear with the vector corresponding to the intercept term.[54](c) *Physical constraints in the model or data* refer to the limits on the values of the controllable variables derived from the chemistry of the system. An example is the investigation of multicomponent compositional mixtures. Orthogonality is impossible in such situations because each variable depends on the others. Similar restrictions may apply to the stoichiometric ratio, *etc.* Another example concerns the various measures of size of an organism, which will show dependencies, as will the amounts of chemicals in the same biological pathway, or measures of rainfall, and elevation in an environmental system. (d) *A bad experimental design* may cause some model effects to be nearly completely confounded with others. This is the result of choosing the levels of experimental factors in such a way that there are near linear dependencies among the columns of X representing the different factors. Factorial designs are usually constructed so as to ensure that linear and interaction effects are orthogonal, or very nearly orthogonal, to each other,[55,56] but this is not possible for designs with squared terms such as the central composite design.

Given knowledge of the controllable variables, and their significances and restrictions, multicollinearity can be removed from the data. In the case of polynomial models, multicollinearity is defined by the model structure. If the experimental strategy cannot be changed, other techniques for decreasing the influence of multicollinearity should be used, despite the fact that the parameter estimates are then biased, as in the case of the generalized principal component regression PCR method described below. One may not always be able to clearly identify the origin of this problem, but it is important to understand its nature as far as is possible.

2.1.3 Multicollinearity diagnostics. It is important to check whether variables show multicollinearity. There are various diagnostics that help demonstrate if there is potential multicollinearity in the data, and so take action to prevent the unstable parameter estimates. A major problem is limited dispersion in an independent variable which results in a very poor (high variance) estimate of the regression parameter for that variable. This can be viewed as a result of the near-collinearity between the variable and the column of ones (for the intercept) in X. This is an example of multicollinearity that is easy to detect by simple inspection of the amount of dispersion in the individual independent variables. The more usual, and less easily detectable, multicollinearity problem arises when the near-singularity involves several independent variables. The dimension of the X-space in which there is very little dispersion is some linear combination of the independent variables, and may not be detectable from inspection of the dispersion of the individual independent variables. The result of multicollinearity involving several variables is high variance in the regression parameters of *all* of the variables involved in the near-singularity.[57]

It is possible to detect multicollinearity from the scatter plots of columns x_j and x_k of matrix X where the approximate linear dependence proves strong multicollinearity. However, multicollinearity may be exposed or masked by the presence of influential points, and especially by high leverage points, and

also there are a large number of possible graphs. Multicollinearity can be removed by, for example, selecting the location of experimental points such that the columns of matrix $\boldsymbol{X}$ will be mutually orthogonal, *i.e.* their scalar product will be zero,

$$< \boldsymbol{x}_j, \boldsymbol{x}_k >= \sum_{i=1}^{n} x_{ij} x_{ik} = 0 \text{ for } j \neq k \tag{2.4}$$

If all the columns of matrix $\boldsymbol{X}$ are mutually orthogonal, then matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is diagonal and a solution of equation $\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ can be expressed in the form

$$b_j = \frac{\sum_{i=1}^{n} x_{ij} y_i}{\sum_{i=1}^{n} x_{ij}^2}, j = 1, \ldots, m \tag{2.5}$$

The significance of the multiple correlation coefficients can be calculated using F-test

$$F_{\mathrm{R}} = \frac{\sum_{j=1}^{m-1} t_j^2}{m-1} = t_{\mathrm{s}} \tag{2.6}$$

where t_{S} is an average value of all the test statistics t_j^2 defined by eqn (2.1) for $\beta = 0$. Note that the β_m is the intercept term. The F_{R} is the F-test criterion for testing multiple correlation coefficient significance.[1]

The presence of multicollinearity can be identified on the basis of numerical and statistical criteria. Instead of the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, its standardized version $\boldsymbol{R}$ is used after $\boldsymbol{X}$ has been scaled so that the length of each vector, the sum of squares of each column, is one. This standardization is necessary to prevent the eigenvalues from being dominated by one or two of the independent variables, and is especially important if the raw variables are of very different magnitudes. The sum of the eigenvalues equals the trace of the matrix being analyzed, which is the sum of the squares of the independent variables including an absolute term. Matrix $\boldsymbol{R}$ is formally identical with the correlation matrix of the controllable variables. For further discussion on this topic, the reader is referred to ref. 58 and the discussions following Belsey's article by Cook,[59] Gunst,[60] Snee and Marquardt[61] and Wood.[62] The following numerical criteria are commonly used to see whether there is multicollinearity in matrix $\boldsymbol{R}$:

(a) If the determinant of matrix $\boldsymbol{R}$, $\det(\boldsymbol{R}) = \prod_{j=1}^{m} \lambda_j$ where λ_j are eigenvalues of the matrix $\boldsymbol{R}$, is less than 10^{-3}, there is good evidence for multicollinearity.

(b) The condition number $K = \lambda_{\max}/\lambda_{\min}$ contains $\lambda_{\max}$ and $\lambda_{\min}$, the largest and the smallest of the m non-zero eigenvalues of a matrix $\boldsymbol{R}$, see ref. 2. The condition number provides a measure of the sensitivity of the solution of the normal equations to small changes in $\boldsymbol{X}$ or $\boldsymbol{y}$. A large condition number indicates that a near-singularity is causing the matrix to be poorly conditioned. Belsey, Kuh, and Welsch[2] suggest that condition numbers K around 10 indicate weak dependencies that may be starting to affect the regression estimates. Condition numbers K of 30 to 100 indicate moderate or strong dependencies and numbers larger than 100 indicate serious collinearity problems. If $K > 1000$, very strong multicollinearity is detected. The number of condition numbers in the critical range indicates the number of near-dependencies contributing to the collinearity problem.

(c) The *variance inflation factor* for the jth regression parameter VIF_j is defined as the ratio of the variance of the jth regression coefficient to the same variance for orthogonal variables when $\boldsymbol{R}$ is the unit matrix. It is given by $\mathrm{VIF}_j = \tilde{R}_{jj}$ where $\tilde{R}_{jj}$ is the jth diagonal element of matrix $\boldsymbol{R}^{-1}$. If $\mathrm{VIF}_j > 10$, strong multicollinearity is detected. The link between VIF_j and collinearity (of the standardized and centred variables) is through the relationship $\mathrm{VIF}_j = 1/(1 - R_j^2)$ where R_j^2 is the coefficient of determination from the regression of X_j on the other independent variables. If there is a near-singularity involving X_j and the other independent variables, R_j^2 will be near 1.0 and VIF_j will be large. If X_j is orthogonal to the other independent variables, R_j^2 will be 0 and VIF_j will be 1.0. Variance inflation factors are simple diagnostics for detecting overall collinearity problems that do not involve the intercept. They will neither detect multiple near-singularities nor identify the source of the singularities. The maximum variance inflation factor has been shown to be a lower bound on the condition number.[63,64] Snee and Marquardt[61] suggest that there is no practical difference between Marquardt's VIF > 10 guideline for serious collinearity, and Belsey, Kuh, and Welsch[2] condition number of 30.

(d) The *Scott multicollinearity criterion*: to examine the suitability of a proposed linear model with regard to possible multicollinearity, a test criterion is usually used,[1]

$$M_{\mathrm{T}} = \frac{\frac{F_{\mathrm{R}}}{t_{\mathrm{S}}} - 1}{\frac{F_{\mathrm{R}}}{t_{\mathrm{S}}} + 1} \tag{2.7}$$

wing rules for identification of multicollinearity:

(i) If $M_{\mathrm{T}} > 0.8$ the model is not suitable because of *strong multicollinearity*, so a model correction is necessary.

(ii) If $0.33 \leqslant M_{\mathrm{T}} \leqslant 0.8$ the model is poor because of *medium multicollinearity*, so some model correction is recommended.

(iii) If $M_{\mathrm{T}} < 0.33$, the model has little trouble from *weak multicollinearity*, so usually no model correction is necessary.

The M_{T} criterion is useful in cases where it is necessary to discover all of the controllable variables which significantly affect the variability of the dependent variable y. When data are approximated by an empirical model, for example by a polynomial, the M_{T} values need not be considered.

2.1.4 Biased regression. As the OLS estimators of the regression parameters are the best, linear and unbiased estimates (of those possible estimators that are both linear functions of the data and unbiased for the parameters being estimated), the LS estimators have the smallest variance. In the presence of collinearity, however, this minimum variance may be unacceptably large. *Biased regression* refers to that class of regression methods in which unbiasedness is no longer required. Generalized principal component regression (GPCR) which will be described below attacks the problem by regressing $\boldsymbol{y}$ on the important principal components and then parcelling out the effect of the principal component variables to the original variables.[41,42] Another biased regression technique, ridge regression, proceeds by adding a small value, k, to the diagonal elements of the correlation matrix of independent variables (from where ridge regression derives its name, since the diagonal of ones in the correlation matrix may be thought of as a ridge). When viewing the ridge trace, the analyst picks a value of k for which the regression parameters β have stabilized. Choosing the smallest value of k possible introduces the smallest bias after which the regression parameters β seem to remain constant. Sometimes increasing k will eventually drive the regression parameters β to zero.

2.1.4.1 Generalized principal component regression. GPCR approaches the collinearity problem from the point of view of eliminating from consideration those dimensions of the X-space that are causing the collinearity problem. This is similar in concept to dropping an independent variable from the model when there is insufficient dispersion in that variable to contribute meaningful information on $\boldsymbol{y}$. However, in GPCR the

dimension dropped from consideration is defined by a linear combination of the variables rather than by a single independent variable. As discussed above it may not always be a single variable that contributes to multicollinearity, but a combination.

GPCR builds on PCA of the matrix of centred and standardized independent variables. As introduced above, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{R}$ where $\boldsymbol{R}$ is the correlation matrix for variables $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{r}$ where $\boldsymbol{r}$ is the correlation vector between $\boldsymbol{y}$ and $\boldsymbol{X}$ variables. To detect ill-conditioning of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, the matrices are decomposed into eigenvalues and eigenvectors. Since the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is symmetrical the eigenvalues are ordered so that $\lambda_1 \leqslant \lambda_2 \leqslant \lambda_3 \leqslant \ldots \lambda_m$, and the corresponding eigenvectors $\boldsymbol{J}_j, j = 1, \ldots, m$, in the form of the sum

$$\boldsymbol{R} = \sum_{j=1}^{m} \lambda_j \boldsymbol{J}_j \boldsymbol{J}_j^{\mathrm{T}} \tag{2.8}$$

The inverse matrix $\boldsymbol{R}^{-1}$ may be expressed in the form

$$\boldsymbol{R}^{-1} = \sum_{j=1}^{m} \lambda_j^{-1} \boldsymbol{J}_j \boldsymbol{J}_j^{\mathrm{T}} \tag{2.9}$$

and therefore the relation for the parameter estimate $\boldsymbol{b}_{\mathrm{N}}$ may be rewritten in the form

$$\boldsymbol{b}_{\mathrm{N}} = \sum_{j=\omega}^{m} [\lambda_j^{-1} \boldsymbol{J}_j \boldsymbol{J}_j^{\mathrm{T}}] \boldsymbol{r} \tag{2.10}$$

and the covariance matrix of normalized estimates $\boldsymbol{b}_{\mathrm{N}}$ may be rewritten in the form

$$D(\boldsymbol{b}_{\mathrm{N}}) = \sigma_{\mathrm{N}}^2 \sum_{j=\omega}^{m} \lambda_j^{-1} \boldsymbol{J}_j \boldsymbol{J}_j^{\mathrm{T}} \tag{2.11}$$

From these equations it follows that when the eigenvalues λ_i are small the estimates $\boldsymbol{b}_{\mathrm{N}}$ and their variances are rather high. Regression problems can be divided into three groups according to the magnitude of the eigenvalues λ_i:

(i) All eigenvalues are significantly higher than zero. The use of the least-squares method (OLS) does not cause any problems. Because matrix $\boldsymbol{R}$ is standardised, the original scale of the variables is not relevant.

(ii) Some eigenvalues are close to zero. This a typical example of multicollinearity when some common methods fail.

(iii) Some eigenvalues are equal to zero: the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ or $\boldsymbol{R}$ is singular and cannot be inverted.

The one way of avoiding difficulties with groups (ii) and (iii) is the use of the *generalized principal component regression GPCR* Here the terms with small eigenvalues λ_i are neglected from the model. One main shortcoming of PCR is neglecting the whole terms which, for the case of higher differences between λ_i, is unacceptable; a better strategy would be to choose a cut-off value that is part way between two PCs. For example the presence of λ_i leads to the unacceptable *high variances of parameters* (small *t*-test) and avoiding of $\boldsymbol{X}_i$ leads to unacceptable high *bias of parameters* and small correlation coefficient, *i.e.* degree of fit. A solution to the dilemma is GPCR. Here the only parts of terms corresponding to λ_i are neglected and therefore the results of regression are continuously changed according to a parameter P which we call *precision.*

1. All eigenvalues ω are retained for which

$$\left[\frac{\sum_{j=1}^{\omega} \lambda_j}{\sum_{j=1}^{m} \lambda_j}\right] > P \tag{2.12}$$

where P can be selected by the user as discussed below but is usually about 10^{-5}. Here m equals the total number of principal components in the datasets: note that the smallest eigenvalue is numbered 1, and the largest m.

2. For the case

$$\left[\frac{\sum_{j=1}^{\omega} \lambda_j}{\sum_{j=1}^{m} \lambda_j}\right] > P \text{ and } \left[\frac{\sum_{j=1}^{\omega-1} \lambda_j}{\sum_{j=1}^{m} \lambda_j}\right] < P \tag{2.13}$$

then part of eigenvalue $\omega-1$ is retained. Define

$$W = \sum_{j=1}^{\omega} \lambda_j \text{ and } E = \sum_{j=1}^{m} \lambda_j \ .$$

When the condition $W/E > P$ is valid, *i.e.* the value ω is not an integer, the summation is made from $\omega-1$ and the eigenvalue $\lambda_{\omega-1}$ is 'weighted' by the factor

$$u = \frac{W - EP}{\lambda_\omega} \tag{2.14}$$

3. Eigenvalues from $\omega-2$ onwards are rejected.

The length of estimates $\|\boldsymbol{b}_{\mathrm{N}}\|$ with their variances may be continuously decreased as a function of increasing precision P. However, it is followed by an increase of the estimate bias and a decrease in the multiple correlation coefficient. The bias of estimates here is caused by neglecting terms in eqn. (2.13) and eqn. (2.14) at $\omega > 1$.

It has been suggested[38] that the squared bias $h_v^2(\boldsymbol{b}_{\mathrm{N}}) = [\beta - E(\boldsymbol{b})]^2$ achieved by the method of GPCR is equal to

$$h_v^2(\boldsymbol{b}_{\mathrm{N}}) = \beta_N^{\mathrm{T}} \left[\sum_{j=1}^{\omega} \boldsymbol{J}_j \boldsymbol{J}_j^{\mathrm{T}}\right] \beta_{\mathrm{N}} \tag{2.15}$$

The optimum magnitude of P may be determined by finding a minimum of the *mean quadratic error of prediction MEP* (in the literature it is also known as the mean squared error of prediction, MSEP).

2.1.4.2 Selection of suitable parameter P. One of the main properties of regression models is a good predictive ability. This predictive ability can also be adopted for the selection of an, in some sense optimum, criterion parameter P. Various criteria for testing prediction ability may be used;[1] one of the most efficient seems to be the mean quadratic error of prediction, MEP. The statistic, MEP uses a prediction $\hat{y}_i$ from an estimate constructed without including the ith point, and is a form of cross-validation, eqn. (1.6). The most suitable model is that which gives the lowest value (minimum) of MEP. Beyond the MEP, the coefficient of determination R^2 (maximum), the predicted coefficient of determination R^2_p (maximum) and the Akaike information criterion, AIC (minimum) can also be used; for the definition of R^2, R^2_p and AIC see pages 41–42 in ref. 1. A suitable P corresponds to some minimum of dependence $\mathrm{MEP}_i = \mathrm{f}(P_i)$. For the selection of this value of P a very simple strategy can be used: for $P \approx 10^{-34}$ the MEP_1 is calculated; for various values P_i, $i = 1, 2, \ldots$, the MEP_i are calculated until $\mathrm{MEP}_i < \mathrm{MEP}_{i-1}$; and in the interval $W_{i-1}/E \leqslant P_i \leqslant W_i/E$ the optimum P is selected by the interval halving method. A trial-and-error procedure can be adopted for selecting a suitable P as reported previously.[39]

2.1.5 Transformation in the case of the non-normality of variable distributions. There are two basic reasons for transforming variables in regression. Transformation of the dependent variable is indicated as a possible remedy for non-normality and for heterogeneous variances of the errors.

Transformations to improve normality have generally been given low er priority than those to simplify relationships or stabilize variance. Fortunately, transformations to stabilize variance often have the effect of also improving normality. Likewise, the power family of transformations, which have been discussed for straightening the one-bend relationship and stabilizing variance, are also useful for increasing symmetry (decreasing skewness) in the distribution. The expectation is that the distribution will also be more nearly normal. An assumption that the residuals are normally distributed is not necessary for estimation of the regression parameters and partitioning of the total variation. Normality is needed only for tests of significance and the construction of confidence interval estimates of the parameters. The *t*-test, *F*-test, and χ^2-test require the underlying random variables to be normally distributed. Likewise, the conventional confidence interval estimates depend on a normal distribution, either directly or through Students's *t*-distribution.

Plots of the observed residuals and skewness and kurtosis coefficients are helpful in detecting non-normality. The skewness coefficient measures the asymmetry of the distribution whereas kurtosis measures the tendency of the distribution to be too flat or too peaked. The skewness coefficient for a normal distribution is 0; the kurtosis coefficient is 3.0. When the sample size is sufficiently large, the frequency distribution of the residuals can be used to judge symmetry and kurtosis.

Transformation of the dependent variable to a form that is more nearly normally distributed is the usual recourse given non-normality. Box and Cox[65] have presented a computational method for determining a power transformation for the dependent variable where the objective is to obtain a simple, normal, linear model that satisfies the usual least squares assumptions. The Box–Cox method uses the parametric family of transformations defined, in standardized form, as

$$y_{trans,i}^{(\gamma)} = \begin{pmatrix} \dfrac{y_i^{\gamma} - 1}{\gamma(\bar{y}_{trans})^{\gamma-1}} & \text{for } \gamma \neq 0 \\ \bar{y}_{trans} \ln(y_i) & \text{for } \gamma = 0 \end{pmatrix} \quad (2.16)$$

where $\bar{y}_g = \exp\sum_{i=1}^{n}[\ln(y_i)]/n$ and $\bar{y}_g$ is the geometric mean of the original observations. The maximum likelihood solution is obtained by performing a least squares analysis on the transformed data for several choices of γ from, say $\gamma = -1$ to +1. Let RSS(γ) be the residual sum of squares from fitting the model to a transformed dependent variable $y^{(\gamma)}$ for the given choice of γ and let $\sigma^2(\gamma)$ = RSS(γ)/n. The likelihood for each choice of γ is then given by

$$L_{max} = -0.5\ln[\sigma^2(\gamma)] \quad (2.17)$$

Maximizing the likelihood is equivalent to minimizing the residual sum of squares. The maximum likelihood solution for $\hat{\gamma}$, then, is obtained by plotting RSS(γ) against γ and reading off the value where the minimum, RSS(γ)$_{min}$, is reached. It is unlikely that the exact power transformation defined by $\hat{\gamma}$ will be used.

2.2 Methodology

2.2.1 Procedure for multiple regression model building. The procedure for the construction of a multiple linear regression model consists of the following steps:

Step 1. *Proposal of the regression model and the statistical significance of parameter estimates.*: The procedure should always start from the simplest regression model and the most convenient one is determined with the use of MEP and AIC.

Step 2. *Exploratory data analysis - examination of multicollinearity, examination of a variable's normality and heteroscedasticity:* The scatter plots of individual variables and all possible pair combinations of the variables are examined. Multicollinearity is examined using the condition number K and the variance inflation factor VIF. Using regression diagnostics a residual's normality and heteroscedasticity are examined. If necessary, the corresponding y variable transformation is applied.

Step 3. *Construction of a more accurate model using GPCR:* On the basis of MEP or AIC a more accurate regression model is constructed (for transformed y variables, if necessary). If some parameters are statistically insignificant the most suitable parameter P is searched for with the use of MEP and AIC.

2.2.2 Software used. For computation of the GPCR an algorithm in S-Plus[66] was written, and the Linear Regression module of the ADSTAT package was used.[38]

2.3 Case study

Many problems in chemometrics concern an approximation of instrumental data of convex (or concave) increasing (or decreasing) values by a polynomial to approximate the shape of the data. For resolution of these types of problems, GPCR with an optimum P minimizing the criteria MEP can be applied.

2.3.1 Dataset: age-related differences in serum levels of 17-hydroxypregnenolone in healthy subjects. 17-hydroxypregnenolone (3β,17α-dihydroxypregn-5-en-20-one), being derived from cholesterol in the metabolic pathway leading to the formation of steroid hormones, represents an important marker in the diagnosis of some gonadal and adrenal defects. Age-related changes in 17-hydroxypregnenolone have been monitored and a detailed study of age- and sex-related changes through childhood, puberty, adulthood and senescence has recently been published.[67] The data concerning serum samples and the monitoring of age-related changes in 17-hydroxypregnenolone [nmol l^{-1}] were obtained from 110 normal males from 2 to 64 years old (Table 5).

2.3.2 Proposal of the regression model. The effects of age on the levels of steroid studied were investigated, and an empirical model describing this dependence was constructed. In step 1 the optimal order of polynomial model m describing the original data on the dependence of 17-hydroxypregnenolone levels for males on age $y = \beta_0 + \beta_1 x + ... + \beta_m x^m$ was established. Using OLS for the highest R^2 (Fig. 6a), highest R^2_P (Fig. 6b), lowest MEP (Fig. 6c) and lowest AIC (Fig. 6d) values and their dependence on the polynomial order m, the global extreme for $m = 9$ was found, while one local extreme was at $m = 6$. In the exploratory data analysis of step 2, the scatter plot of the dependence of ordinary residuals $\hat{e}$ on prediction $\hat{y}$ shows heteroscedasticity, and the Q–Q graph of jackknife residuals exhibits a skewed asymmetric distribution of random errors in variable y.

2.3.3 Examination of multicollinearity, examination of normality of variables and heteroscedasticity. An examination of multicollinearity concerns an estimation of the maximum condition number $K = 8.41 \times 10^{12}$ which is higher than 1000, and the largest value of the variance inflation factor VIF $= 7.34 \times 10^{10}$ which is higher than 10; therefore a strong multicollinearity is suggested. To examine the normality of random error distribution in dependent variable y and to find the most convenient variable transformation y^{γ} (the power transformation) or $(x^{\gamma} - 1)/\gamma$ (the Box–Cox transformation), the RSS(γ) for different values of power γ were calculated. Several resolution criteria were applied to find the optimal power $\hat{\gamma}$. The

most important value was such an estimate of $\hat{\gamma}$ for which the normality of a residual distribution was achieved: for estimate $\hat{\gamma}$= 0.13 the skewness g_1 is nearly zero and the kurtosis g_2 is nearly equal to 3; thus the transformation represents a good approximation of a Gaussian distribution. A search for the optimal polynomial degree m was then repeated for the transformed data, and the same m was determined. From the biochemical point of view, the better fit is for the global minimum, m = 9, when the curve reflects all the fine nuances of the age-dependence.

As the OLS method leads to large variances of regression parameters given a strong multicollinearity in data, the parameter estimates are not statistically significant, and the GPCR method should be used instead. Fig. 7a, b, c, d show a search of the PCR optimum criterion value P with the use of statistical criteria R^2, R^2_P, MEP and AIC and transformed data ($\hat{\gamma}$ = 0.13). The rankit Q–Q plot of jackknife residuals (Fig. 8) then exhibits a Gaussian distribution and the residuals are homoscedastic.

2.3.4 Construction of a more accurate model using GPCR. In step 3 the regression model was constructed: for the identified polynomial degree m = 9 the test criterion F_R = 7.573 was greater than the corresponding quantile of the Fisher-Snedecor F-distribution $F_{0.95}$ (8, 114−9) = 1.975, and therefore the proposed regression model is statistically significant. In contrast, the quantile of the Student t-distribution, $t_{0.975}$ (114−9) = 1.984 is greater than t_0 = −0.812, t_1 = 1.602, t_8 = −1.881, t_9 = 1.762, and therefore the four parameters β_0, β_1, β_8, β_9 are statistically insignificant. Meanwhile, $t_{0.975}$ (114−9) = 1.984 is smaller than t_2 = −2.11, t_3 = 2.234, t_4 = −2.285, t_5 = 2.232, t_6 = −2.129, t_7 = 2.006, and the corresponding six parameters β_2, β_3, β_4, β_5, β_6 are statistically significant. The ordinary least-squares method OLS with P = 10^{-35} has proven the polynomial degree of the 9th order (where brackets of the model equation contain the standard deviation of each parameter and the letter R means that β is rejected while the letter A means β is accepted):

y = −13.18(16.24, R) + 16.43(10.25, R)x − 4.87(2.42, A)x^2 + 0.634(0.284, A)x^3 − 4.29E−02(1.88E02, A)x^4 + 1.67E−03(7.46E−04, A)x^5 − 3.86E−05(1.81E−05, A)x^6 − 3.96E−09(2.10E−09, A)x^7 − 3.96E−09(2.10E−09, R)x^8 + 1.25E−11(7.09E−12, R)x^9

with statistical criteria R^2 = 40.53%, R^2_p = 55.75%, MEP = 26.122, AIC = 362.67. The method of GPCR with P = 1.0 × 10^{-8} found another regression model in which most parameter estimates were statistically significant but biased, in the form

Table 5 The age-related changes in 17-hydroxypregnenolone for 110 normal males from 2 to 64 years old: age x [years], concentration of 17-hydroxypregnenolone y [nmol l^{-1}]

2	5.1000	3	6.0998	5	1.4000	6	1.2946	6	2.5893	6	2.2671
6	2.2573	6	6.5002	6	1.5786	6	1.4000	6	4.8001	7	3.4066
7	0.9149	7	3.8001	8	0.9000	8	4.8001	10	3.8574	10	0.9166
10	3.0473	10	3.1274	10	1.2000	11	3.6561	13	3.3075	13	7.3002
13	2.8999	13	6.1050	13	2.7069	16	10.460	16	28.349	16	12.099
17	20.700	18	19.981	18	11.427	19	38.600	19	10.516	19	12.481
20	4.0029	22	9.9006	22	21.082	22	12.079	23	14.100	24	9.2455
24	4.4665	24	16.263	25	23.656	28	14.278	29	3.0690	29	8.6128
30	14.398	30	6.6234	30	2.2364	31	2.0903	31	5.0297	32	2.1000
32	5.6313	32	7.9572	32	17.651	32	9.6001	33	3.0308	35	15.700
35	4.3706	37	1.3323	37	16.068	38	2.2521	38	3.7740	39	6.3807
39	5.1226	39	5.9714	40	7.3380	41	9.3778	42	5.6996	42	3.7085
43	19.127	44	7.5441	44	9.3468	44	5.2086	45	14.339	46	4.0545
47	6.4874	48	10.732	48	10.327	49	4.9234	49	9.3674	50	5.1617
50	13.210	51	8.7046	51	3.9384	51	3.3616	53	3.5773	53	3.8533
54	4.5176	54	9.6914	55	10.528	55	3.6305	56	7.6003	56	6.2327
57	2.3947	57	6.0150	57	4.4930	58	5.2138	59	1.2867	59	11.529
60	2.5229	60	8.6483	61	2.2264	61	3.6349	62	4.6618	62	5.5733
63	3.8802	64	6.0525								

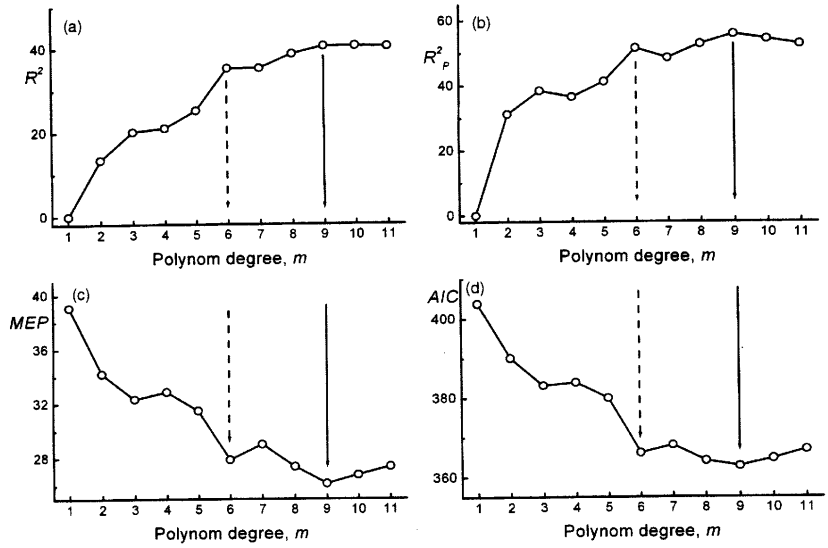


Fig. 6 The search for an optimal polynomial degree m leads to one local (dashed line) and one global extreme (full line) when the following dependences and the ordinary least squares OLS were used: (a) the determination coefficient R^2 on m, (b) the predicted coefficient of determination R^2_P on m, (c) the mean error of prediction MEP on m, (d) the Akaike information criterion AIC on m.

y = 23.77(7.98, A) − 9.05(2.76, A)x + 1.18(0.33, A)x^2 − 5.72E−02(1.72E−02, A)x^3 + 1.00E−03(3.61E−04, A)x^4 + 2.68E−06(6.83E−07, A)x^5 − 2.11E−07(8.41E−08, A)x^6 − 8.85E−10(1.92E−10, A)x^7 + 5.70E−11(2.53E−11, A)x^8 − 3.70E−13(1.97E−13, R)x^9

with more pessimistic values of statistical criteria R^2 = 36.31%, R^2_p = 51.04%, MEP = 28.028, AIC = 370.21 than the previous OLS method, but with polynomial parameters from β_0 to β_8 statistically significant except for β_9. Therefore, data were recalculated for m = 8 and GPCR with the new optimum criterion P = 1.0 × 10^{-7} determined following regression model:

y = 12.9(6.27, A) − 4.39(1.74, A)x + 0.55(0.15, A)x^2 − 2.08E−02(5.18E−03, A)x^3 + 1.69E−04(4.07E−05, A)x^4 + 4.20E−06(1.02E−06, A) x^5 − 2.54E−08(6.58E−09, A)x^6 − 1.32E−09(3.22E−10, A)x^7 + 1.32E−11(3.29E−12, A) x^8

with statistical criteria R^2 = 33.35%, R^2_p = 50.55%, MEP = 28.214, AIC = 373.21 and the curve fitted as presented in Fig. 9a. As the dependent variable y does not exhibit normal distribution, the power transformation of y was used and better goodness-of-fit was achieved with P = 1.0 × 10^{-7} in the form

$y^{0.13}$ = 1.43(0.12, A) − 0.12(0.03, A) x + 1.39E−02(3.05E−03, A)x^2 − 5.20E−04(1.02E−04, A)x^3 + 4.17E−06(8.07E−07, A)x^4 + 1.04E−07(2.03E−08, A)x^5 − 6.31E−10(1.31E−10, A)x^6 − 3.28E−11(6.40E−12, A) x^7 + 3.29E−13(6.54E−14, A)x^8

with statistical criteria R^2 = 40.01%, R^2_p = 57.13%, MEP = 0.011198, AIC = −489.08. Fig. 9b shows the results obtained when the retransformed variable $y^{0.13}$ was used and some diagnostic graphs for a detection of influential points were examined: two scatter graphs of standardized residuals (Fig. 10a and Fig. 10b) indicating outliers and heteroscedasticity but not leverages show that points 29 and 34 are no longer outliers when transformed data are used. The graphs of Cook distance (Fig. 11a and b) prove points 29, 34 and 110 to be influential. After data transformation outliers 29 and 34 may remain in the data set, and Gaussian and homoscedastic distribution results.

It may be concluded that 17-hydroxypregnenolone increases from childhood and reaches a statistically significant maximum at 20 years of age, followed by a fall to a local minimum at 37 years of age. This then increased to a statistically insignificant peak at 49 years of age in men, and was followed by a less pronounced decline as shown in Fig. 9.

2.4 Conclusions

When multicollinearity in data occurs, OLS estimates of regression parameters are unbiased, but their variances are often

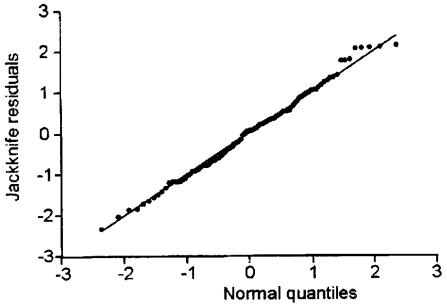


Fig. 8 The rankit Q–Q graph of jackknife residuals proves the normality of the random errors in the transformed dependent variable $y^{0.13}$.

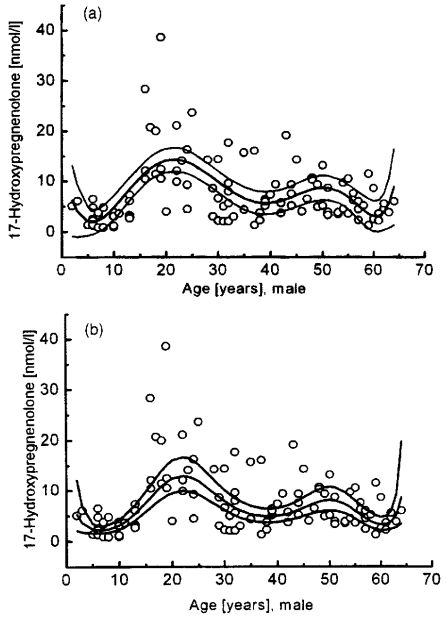


Fig. 9 Linear regression of the 8th degree polynomial of the age-dependence of 17-hydroxypregnenolone when (a) the original variable y was used, (b) the transformed variable $y^{0.13}$ was used.

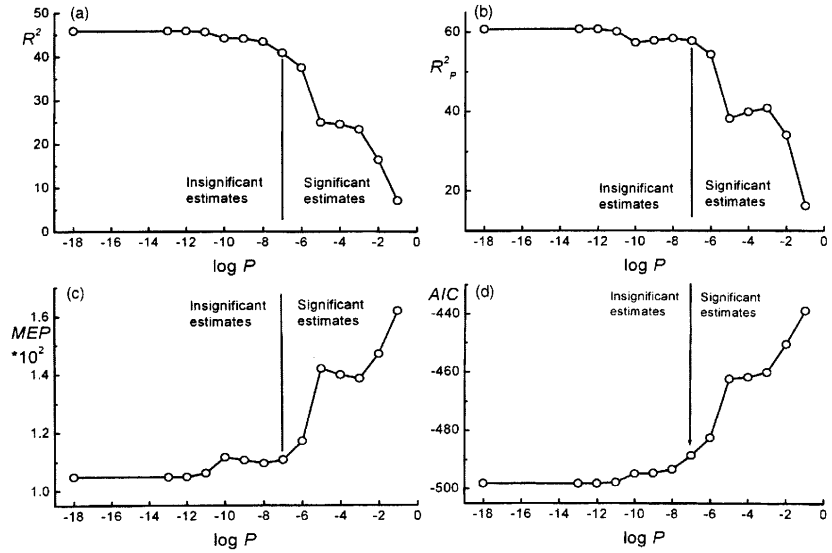


Fig. 7 The search for the GPCR optimal criterion value P separates statistically significant and insignificant parameter estimates when the method of generalized principal component was used: (a) the determination coefficient R^2 on P, (b) the predicted coefficient of determination R^2_P on P, (c) the mean error of prediction MEP on P, (d) the Akaike information criterion AIC on P.

so large that they may be far from the true value. By adding a degree of bias to the regression estimates, GPCR reduces variances. Biased regression methods are generally based on the fact that estimators with smaller mean squared errors can be found if the unbiasedness of the estimators is relaxed. The GPCR method, in combination with the MEP criterion is very useful for constructing biased models. It can also be used for achieving estimates that keep the model course corresponding to the data trend, especially in polynomial-type regression models. In the search for the best degree of polynomial, several statistical characteristics of regression quality should be considered as well.

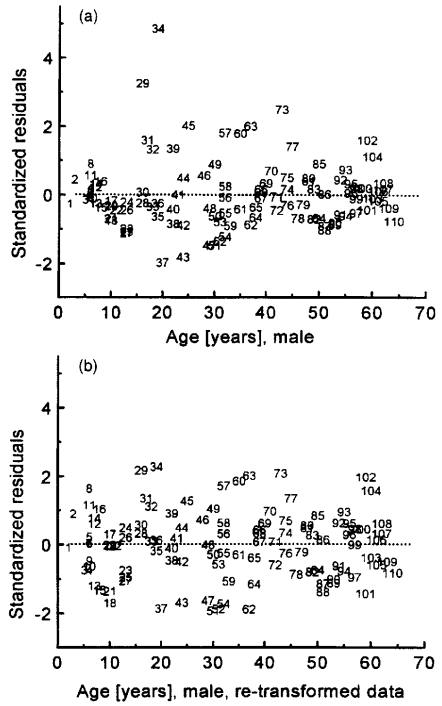


Fig. 10 The scatter plot of standardized residuals $\hat{e}_S$ on age x when (a) the original variable y was used, (b) the transformed variable $y^{0.13}$ was used.

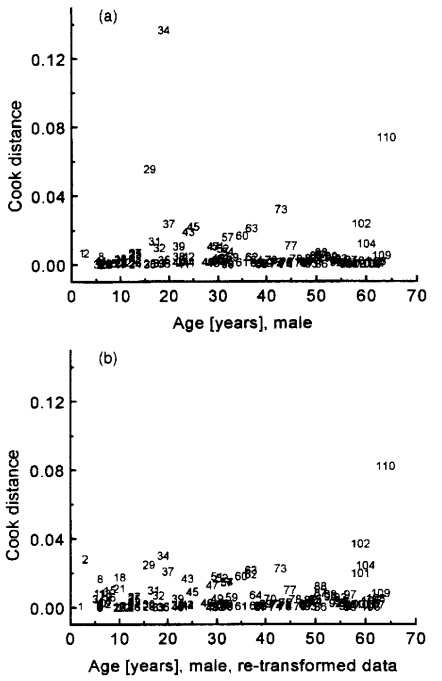


Fig. 11 The scatter plot of Cook distance D on age x when (a) the original variable y was used, (b) the transformed variable $y^{0.13}$ was used.

3. Acknowledgements

The financial support of the Ministry of Education (Grant No MSM253100002) and of the Grant Agency of the Czech Republic (Grant No 303/00/1559) is gratefully acknowledged.

4. References

1 M. Meloun, J. Militký and M. Forina, *Chemometrics for Analytical Chemistry, Vol. 2. PC-Aided Regression and Related Methods*, Horwood, Chichester, 1994.
2 D. A. Belsey, E. Kuh and R. E. Welsch, *Regression Diagnostics: Identifying Influential data and Sources of Collinearity*, Wiley, New York, 1980.
3 R. D. Cook and 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*, Clarendon Press, Oxford, 1985.
5 S. Chatterjee and A. S. Hadi, *Sensitivity Analysis in Linear Regression*, Wiley, New York, 1988.
6 V. Barnett and T. Lewis, *Outliers in Statistical data*, Wiley, New York, 2nd edn., 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 and S. C. Peters, *Proceedings of the Eleventh Interface Symposium on Computer Science and Statistics*, ed. A. R. Gallant and T. M. Gerig, Institute of Statistics, North Carolina State University, Raleigh, 1978.
9 S. Weisberg, *Applied Linear Regression*, Wiley, New York, 1985.
10 P. J. Rousseeuw and 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 and M. B. Wilk, *Biometrics*, 1975, **31**, 387–410.
13 R. D. Cook, *Technometrics*, 1977, **19**, 15–18.
14 R. D. Cook and S. Weisberg, *Technometrics*, 1980, **22**, 495–508.
15 D. C. Hawkins, D. Bradu and G. V. Kass, *Technometrics*, 1984, **26**, 197–208.
16 C. E. McCulloch and D. Meeter, *Technometrics*, 1983, **25**, 152–155.
17 J. B. Gray, *Proceedings of the Statistical Computing Section, American Statistical Association*, 1983, pp. 159–164.
18 J. B. Gray, *Proceedings of the Statistical Computing Section, American Statistical Association*, 1985, pp. 102–107.
19 J. B. Gray and R. F. Ling, *Technometrics*, 1984, **26**, 305–330.
20 S. Chaterjee and A. S. Hadi, *Stat. Sci.*, 1986, **1**, 379–416.
21 B. Walczak and D. L. Massart, *Chemom. Intell. Lab. Syst.*, 1998, **41**, 1–15.
22 R. D. Cook and P. Prescot, *Technometrics*, 1981, **23**, 59.
23 D. A. Pierce and R. J. Gray, *Biometrika*, 1982, **69**, 233.
24 *SAS/STAT User's Guide, Version 6, Volume II*, SAS Institute Inc., Cary, North Carolina, 4th edn., 1989.
25 R. D. Cook, *J. Am. Stat. Assoc.*, 1979, **74**, 169–174.
26 A. Hedayat and D. S. Robson, *J. Am. Stat. Assoc.*, 1970, **65**, 1573.
27 R. L. Brown, J. Durbin and J. M. Evans, *J. Royal Stat. Soc., Ser. B*, 1975, **37**, 149.
28 J. S. Galpin and D. M. Hawkins, *Am. Stat.*, 1973, **68**, 144.
29 C. P. Quesenberry, in *Goodness of Fit Techniques*, ed. R. B. D'Agostino and M. A. Stephens, Marcel Dekker, New York, 1986, ch. 6.
30 D. C. Hoaglin and R. E. Welsch, *Am. Stat.*, 1978, **32**, 17–22.
31 F. J. Anscombe, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistical Problems*, 1961, **I**, pp. 1–36.
32 N. R. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, 1st edn., 1966.
33 R. J. Carrol and D. Ruppert, *Transformation and Weighting in Regression*, Chapman and Hall, New York, 1988.
34 A. S. Hadi, in *Handbook of Statistics*, ed. C. R. Rao, 1993, **9**, 775–802.
35 D. X. Williams, *Appl. Stat.*, 1973, **22**, 407–408.
36 D. Pregibon, *Ann. Stat.*, 1981, **9**, 45–52.
37 F. R. Hampel, *J. Am. Stat. Assoc.*, 1974, **69**, 383–393.
38 ADSTAT (English version), TriloByte Statistical Software, Pardubice 1999.

39 J. O. Rawlings, S. G. Pantula and D. A. Dickey, *Applied Regression Analysis, A Research Tool*, Springer Verlag, New York, 2nd edn., 1998.
40 C. M. Stein, in *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling*, Stanford University Press, Stanford, California, 1960.
41 A. E. Hoerl and R. W. Kennard, *Technometrics*, 1970, **12**, 69.
42 A. E. Hoerl and R. W. Kennard, *Technometrics*, 1970, **12**, 55.
43 D. W. Marquardt, *Technometrics*, 1970, **12**, 591.
44 J. T. Webster, R. F. Gunst and R. L. Mason, *Technometrics*, 1974, **16**, 513.
45 S. Wold, A. Ruhe, H. Wold and W. J. Dunn, *SIAM J. Stat. Comput.*, 1984, **5**, 735.
46 D. A. Belsey, *Condition Diagnostics: Collinearity and Weak Data in Regression*, Wiley, New York, 1991.
47 R. A. Bradley and S. S. Srivastava, *Am. Stat.*, 1979, **33**, 11.
48 G. A. F. Seber, *Linear Regression Analysis*, Wiley, New York, 1977.
49 J. R. Simpson and D. C. Montgomery, *J. Stat. Comp. Simul.*, 1996, **56**, 1.
50 T. Foucart, *Oper. Res.*, 2000, **34**, 199.
51 S. P. Ellis, *Stat. Sci.*, 1998, **13**, 337.
52 P. L. Bonate, *Pharm. Res.*, 1999, **16**, 709.
53 D. Sengupta and P. Bhimasankaram, *J. Am. Stat. Assoc.*, 1997, **92**, 1024.
54 N. Brauner and M. Shacham, *Math. Comput. Simul.*, 1998, **48**, 75.
55 M. Shacham and N. Brauner, *Chem. Eng. Process.*, 1999, **38**, 477.
56 N. Brauner and M. Shacham, *Ind. Eng. Chem. Res.*, 1999, **38**, 4477.
57 S. P. DeCarvalho and C. D. Cruz, *Brazil. J. Genetics*, 1996, **19**, 479.
58 D. A. Belsey, *Am. Stat.*, 1984, **38**, 73.
59 R. D. Cook, *Am. Stat.*, 1984, **38**, 78.
60 R. F. Gunst, *Am. Stat.*, 1984, **38**, 79.
61 R. D. Snee and D. W. Marquardt, *Am. Stat.*, 1984, **38**, 83.
62 F. S. Wood, *Am. Stat.*, 1984, **38**, 88.
63 R. A. Stine, *Am. Stat.*, 1995, **49**, 53.
64 K. N. Berk, *J. Am. Stat. Assoc.*, 1977, **72**, 863.
65 M. Shacham and N. Brauner, *Ind. Eng. Chem. Res.*, 1997, **36**, 4405.
66 S-Plus, MathSoft, Data Analysis Products Division, 1700 Westlake Ave N, Suite 500, Seattle, WA 98109, USA, 1997.
67 M. Hill, D. Lukáč, O. Lapčík, J. Šulcová, R. Hampl, V. Pouzar and L. Stárka, *Clin. Chem. Lab. Med.*, 1999, **37**, 439.