Clin Chem Lab Med 2004;42(3):311–322 © 2004 by Walter de Gruyter • Berlin • New York

New methodology of influential point detection in regression model building for the prediction of metabolic clearance rate of glucose

Milan Meloun[1,*], Martin Hill[2], Jiří Militký[3], Jana Vrbíková[2], Soňa Stanická[2] and Jan Škrha[4]

[1] Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, Pardubice, Czech Republic
[2] Institute of Endocrinology, Prague, Czech Republic
[3] Department of Textile Materials, Technical University, Liberec, Czech Republic
[4] Department of Internal Medicine, Charles University, Prague, Czech Republic

Abstract

Identifying outliers and high-leverage points is a fundamental step in the least-squares regression model building process. The examination of data quality involves the detection of influential points, outliers and high-leverages, which cause many problems in regression analysis. On the basis of a statistical analysis of the residuals (classical, normalized, standardized, jackknife, predicted and recursive) and diagonal elements of a projection matrix, diagnostic plots for influential points indication are formed. The identification of outliers and high leverage points are combined with graphs for the identification of influence type based on the likelihood distance. The powerful procedure for the computation of influential points characteristics written in S-Plus is demonstrated on the model predicting the metabolic clearance rate of glucose (MCRg) that represents the ratio of the amount of glucose supplied to maintain blood glucose levels during the euglycemic clamp and the blood glucose concentration from common laboratory and anthropometric indices. MCRg reflects insulin sensitivity filtering-off the effect of blood glucose. The prediction of clamp parameters should enable us to avoid the demanding clamp examination, which is connected with a higher load and risk for patients.

Keywords: diagnostic plot; high-leverages; influence measures; influential observations; outliers; regression diagnostics.

Introduction

Women with the polycystic ovary syndrome (PCOS) have a high prevalence of insulin resistance. However, controversy exists as to whether insulin resistance results from PCOS or the obesity that is frequently associated with it (1–6). Gennarelli et al. (6) have suggested a prediction statistical model for insulin resistance in PCOS women within a wide range of body mass indices (BMIs) involving waist-to-hip-ratio (WHR) and serum triglycerides or fasting insulin. Cibula et al. (7) recently described such a statistical model in non-obese PCOS women with sex-hormone binding globulin (SHBG) as the best predictor of the insulin sensitivity index. In the present study, a large group of lean and obese women was evaluated (fulfilling the generally accepted diagnostic criteria of PCOS) with the use of a euglycemic clamp, which is considered as the gold standard in the evaluation of insulin sensitivity. Here we attempted to build a regression model for the prediction of one of the clamp parameters, the metabolic clearance rate of glucose (MCRg), which represents the ratio of the amount of glucose supplied to maintain blood glucose levels during the euglycemic clamp and the blood glucose concentration from common laboratory and anthropometric indices. MCRg reflects insulin sensitivity filtering-off the effect of blood glucose. The prediction of clamp parameters should enable us to avoid the demanding clamp examination, which is connected with a higher load and risk for patients. Our primary aim was to demonstrate how to cope with the presence of influential experimental points that deteriorate the quality of prediction.

Statistical models, particularly regression models, are extremely useful tools 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 under modest modifications of the problem formulation becomes important. However, there are a number of common difficulties associated with real data sets. 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 multistep procedure is required, first to identify whether there are any samples that are atypical of the data set, then to remove them, and finally to reformulate the model. This paper describes a proposed new methodology with a series of powerful general diagnostics for detecting observations that differ from the bulk of the data. We think of data as being divided into two classes: (i) good observations (the majority of data) reflecting population scatter of data and (ii) the outliers (if any), which are

*Corresponding author: Prof. RNDr. Milan Meloun, DrSc., Department of Analytical Chemistry, University Pardubice, 532 10 Pardubice, Czech Republic
Phone: +42466037026, Fax: +42466037068,
E-mail: milan.meloun@upce.cz

included in 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 influence measures that have been proposed and the critical survey published (8–13). Regression diagnostics used in this paper 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. In this paper regression diagnostics are critically surveyed and commented upon, and compared on a regression model for prediction of the metabolic clearance rate of glucose as one of the output parameters of the clamp.

Proposed method

Estimate of the regression parameters

A linear regression model is one formed by a linear combination of explanatory variables $\boldsymbol{x}$ or their functions, $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$. Vector y has dimensions $(n\times 1)$ and matrix X $(n\times m)$ and supposes that $m<n$. Linear means ''linear according to model parameters''. The least-squares is the most frequently used method in regression analysis to find the minimal length of the residual vector $\hat{\boldsymbol{e}}=\mathbf{y}-\hat{\boldsymbol{y}}_{\boldsymbol{P}}$, where $\hat{\boldsymbol{y}}_{\boldsymbol{p}}=\boldsymbol{X}\,\boldsymbol{b}$ is the predictor vector. The square of vector $\hat{\boldsymbol{e}}$ length is consistent with the residual sum of squares criterion $U(\boldsymbol{b})$ of the least-squares method, so that the estimates of model parameter $\boldsymbol{b}$ minimizes the expression

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

The conventional least-squares estimator b has the form $\boldsymbol{b}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\,\boldsymbol{y}$ with the corresponding variance $D(\boldsymbol{b})=\sigma^2\,(\boldsymbol{X}^T\,\boldsymbol{X})^{-1}$. However, some basic assumptions are necessary for the least-squares method (LS) to be valid (10):

1. The regression parameters $\boldsymbol{\beta}$ are not bound.
2. The regression model is linear in the parameters, and an additive model for the measurement of errors is valid, $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$.
3. The matrix of non-random controllable values of the explanatory variable $\boldsymbol{X}$ has a column rank equal to m.
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, $E(\varepsilon_i^2)=\sigma^2$ and therefore the data are said to be homoscedastic.
6. The random errors ε_i are uncorrelated and therefore $\mathrm{cov}(\varepsilon_i, \varepsilon_i)=E(\varepsilon_i, \varepsilon_i)=0$. When the errors follow the normal distribution they are also independent. This corresponds to independence of the measured quantities y.
7. The random errors ε_i have a normal distribution $N(0, \sigma^2)$.

In regression analysis the method of least squares is often used. This method, however, does not ensure that the model is fully acceptable from a statistical point of view. A source of problems may be found in components of a regression triplet (data, model and method of estimation). The least squares method provides accurate estimates only when all assumptions about data and about a regression model are fulfilled. When some assumptions are not fulfilled, the least-squares method is inconvenient. Regression diagnostics represent the procedures for identification of (a) the data quality for a proposed model, (b) the model quality for a given data set and (c) fulfillment of all least-squares assumptions.

Examination of data quality

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. Influential points include data classified into: (i) outliers (denoted in the 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 the 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 outliers and high-leverages together. 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. Ordinary residuals $\hat{e}_i$, normalized residuals or also so-called scaled residuals $\hat{e}_{N,i}=\hat{e}_i/\hat{\sigma}$, standardized residuals or also so-called internally Studentized residuals $\hat{e}_{S,i}=\hat{e}_i/(\hat{\sigma}\sqrt{1-H_{ii}})$, Jackknife residuals or also so-called externally Studentized residuals

$$\hat{e}_{J,i}=\hat{e}_{S,i}\sqrt{\frac{n-m-1}{n-m-\hat{e}_{S,i}^2}},$$

predicted residuals or also so-called cross-validated residuals

$$\hat{e}_{P,i}=\frac{\hat{e}_i}{1-H_{ii}}=y_i-x_i\boldsymbol{b}_{(i)},$$

and recursive residuals may be applied and have been described previously (8–10). Some diagnostics are based on the diagonal elements of the hat matrix. For analysis of residuals a variety of plots have been widely used in regression diagnostics:

(a) The graph of predicted residuals (14) 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.

(b) The Williams graph (15) 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, $x = 2m/n$. Note that $t_{0.95}(n-m-1)$ is the 95% quantile of the Student distribution with $(n-m-1)$ degrees of freedom.

(c) The Pregibon graph (16) has on the *x*-axis the diagonal elements H_{ii} and on the *y*-axis the square of normalized residuals $\hat{e}^2_{N,i}$. Since the expression $E\,(H_{ii}+\hat{e}^2_{N,i}) = (m+1)/n$ is valid for this graph, two different constraining lines can be drawn, $y = -x+2(m+1)/n$, and $y = -x+3(m+1)/n$. To distinguish among influential points the following rules are used: (i) a point is strongly influential if it is located above the upper line; (ii) 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.

(d) The McCulloh and Meeter graph (17) 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.

(e) 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

$$y = \frac{2x - x^2 - 1}{x\,(1-K) - 1},$$

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.

(f) The Index graph (10) 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* that could be influential, i.e., outliers or high-leverages.

(g) The Rankit graph (Q-Q plot) (10) 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.

There are diagnostics that are based on scalar influence measures (8–10): Proper normalization in influence functions (19) leads to scalar measures. These measures express the relative influence of the given point on all parameter estimates.

(a) The Cook measure D_i (20) expresses directly the relative influence of the ith point on all parameter estimates and has the form

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

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 is always important to study cases with $D_i > 1$. These benchmarks are intended as an aid to finding influential cases, but they do not represent a test. There is no significance test associated with D_i.

(b) The Atkinson measure A_i (21) enhances the sensitivity of distance measures to high-leverage points and has the form

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

This measure is also convenient for graphical interpretation; Atkinson recommends that signed values of A_i be plotted using any of the customary methods for residuals. 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.

(c) The Belsey $DFFITS_i$ measure, also called Welsch-Kuh's distance (22), 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

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

Belsey, Kuh, and Welsch (22) 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 $2\sqrt{m/n}$.

(d) The Anders-Pregibon diagnostic Ap_i (19) expresses the influence of the ith point on the volume of the confidence ellipsoid

$$AP_i = \frac{det(\boldsymbol{X}^T_{m\,(i)}\,\boldsymbol{X}_{m\,(i)})}{det(\boldsymbol{X}^T_m\,\boldsymbol{X}_m)},$$

where $\boldsymbol{X}_{\mathbf{m}} = (\boldsymbol{x}|\,\boldsymbol{y})$ is the matrix having as the 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 $AP_i = 1 - H_{ii} - \hat{e}^2_{N,i} = 1\text{-}H_{m,ii}$. A point is considered to be influential if $H_{m,ii} = 1\text{-}AP_i > 2(m+1)/n$.

(e) The Cook-Weisberg likelihood measures LD_i (19) represent a general diagnostic defined by

$$LD_i = 2[L\,(\hat{\Theta}) - L\,(\hat{\Theta}_{(i)})],$$

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,

$$LD_i > \chi^2_{1-\alpha}\ (m + 1),$$

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 *i*th point on the parameter estimates or on the variance estimate or on both (37):

(f) The likelihood measure $LD_i(\boldsymbol{b})$ examines the influence of individual points on the parameter estimates $\boldsymbol{b}$ by the relationship

$$LD_i(\boldsymbol{b}) = n \times ln\left[\frac{d_i \times H_{ii}}{1 - H_{ii}} + 1\right],$$

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

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

$$LD_i\ (\hat{\sigma}^2) = n \times ln\left[\frac{n}{n-1}\right] + n\ ln\ (1 - d_i) + \frac{d_i\ (n-1)}{1-d_i} - 1.$$

(h) 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

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

Examination of a proposed regression model

There are many various plots for considering y on x_j, but we limit the choice here to (a) partial regression leverage plots and (b) partial residual plots. Both plots are augmented here by the graph of residual $\hat{e}$ vs. prediction $\hat{y}_P$, which can indicate a false model when the points form a nonlinear pattern (8–10).

(a) Partial regression leverage plots (PRL plot) were introduced by Belsey et al. (22). They permit classification of the quality of a proposed regression model and also indicate the presence of an influential point and lack of fulfillment of the assumptions of the classical least-squares method. They show the dependence between $\boldsymbol{y}$ and a selected controllable variable $\boldsymbol{x}_j$ when the other controllable variables forming columns in the matrix $\boldsymbol{X}$ are kept constant. This linear dependence is valid only when the proposed model is correct. The symbol $\boldsymbol{X}_{(j)}$ denotes a matrix formed by leaving out the *j*th column $\boldsymbol{x}_j$.

(b) Partial residual plots (PR plots) are also termed ''component + residual'' plots. These plots are recommended for indication of different types of nonlinearity in the case of a poorly proposed regression model. The linear dependence shows the suitability of proposed variable $\boldsymbol{x}_j$ in the model.

(c) Sign test for model specification. To check a proposed regression model with reference to the data, all tests of linearity may be applied. The sign test is a single test based on the residuals. Incorrectness of a proposed model causes non-randomness of residuals, and this non-randomness may be tested.

Various test criteria for a search of regression model quality may be used (8–10). One of the most efficient seems to be the mean quadratic error of prediction, *MEP*, being defined by the cross-validation relationship

$$MEP = \left[\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{(i)})^2\right]/n,$$

where $b_{(i)}$ is the estimate of regression parameters when all points except the *i*th one were used and $\boldsymbol{x}_i$ is the *i*th row of matrix $\boldsymbol{X}$. The statistic *MEP* uses a prediction $\hat{y}_{P,i}$ from an estimate constructed without including the *i*th point. The *MEP* also can be used to express the predicted determination coefficient,

$$\hat{R}^2_P = 1 - \frac{n \times MEP}{\sum_{i=1}^{n} y_i^2 - n \times \overline{y}^2}.$$

Another statistical characteristic in quite general use is derived from information theory and entropy (23) and is known as the Akaike information criterion,

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

The most suitable model is the one that 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.

Examination of conditions for the least-squares method

From seven basic conditions for the least-squares method that must be met to give unbiased linear estimates of parameters, the heteroscedasticity, autocorrelation and non-normality of errors ε are the most important.

(a) Identification of heteroscedasticity in data is based on the idea that the variance of a measured quantity at the *i*th point is an exponential function of the variable $\boldsymbol{x}_i\beta$ of the type $\sigma_i^2 = \sigma^2\ exp\ (\lambda\ \boldsymbol{x}_i\ \boldsymbol{\beta})$, where $\boldsymbol{x}_i$ is the *i*th row of matrix $\boldsymbol{X}$. The test for homoscedasticity is carried out by checking the null hypothesis $H_0 > \lambda = 0$. Cook and Weisberg introduced the test criterion

$$S_f = \frac{\left[\sum_{i=1}^{n} (\hat{y}_i - \overline{y}_P)\ \hat{e}_i^2\right]^2}{2\ \hat{\sigma}^4 \left[\sum_{i=1}^{n} (\hat{y}_i - \overline{y}_P)\right]^2}, \text{ where } \overline{y}_P = \left[\sum_{i=1}^{n} \hat{y}_i\right]^2/n.$$

When null hypothesis is valid, the test statistic S_f has approximately the $\chi^2(1)$ distribution with one degree of freedom.

(b) Identification of autocorrelation in data. When data are a time series, the errors ε are not independ-

ent but are correlated with one another. The autoregressive process of the first order is described by the expression $\varepsilon_i = \rho_1\ \varepsilon_{i-1} + u_i$, where $u_i \approx N(0, s^2)$ is an independent random variable with constant variance and $\rho_1 \leq 1$ is the autocorrelation coefficient of the first order. For $\rho_1 = 1$ a case of cumulative errors is defined that appears often in biochemical data. When the model $\boldsymbol{X\beta}$ does not contain all the significant variables and is falsely proposed, the mean values of residuals correspond to a process of the first order, with a positive autocorrelation coefficient of the first order, ρ_1. Tests of autocorrelation can be understood as tests of accuracy of a proposed model, with reference to the number of controllable variables. To test for autocorrelation, the graph of $\hat{e}_i$ against $\hat{e}_{i-1}$is plotted and an approximately linear trend proves significant autocorrelation.

(c) Test of normality of errors in data. To test the normality of residuals, the most convenient test seems to be the Jarque-Bera test (24), which is based on the criterion

$$L(\hat{e}) = n\left[\frac{\hat{g}_1}{6} + \frac{(\hat{g}_2 - 3)^2}{24}\right],$$

where the symbol $\hat{g}_1$ denotes the sample skewness and $\hat{g}_2$ the sample kurtosis of residuals set. When$L(\hat{e}) \rangle \chi^2_{0.95}(2) = 5.99$, the null hypothesis H_0 about the error normality is rejected. In this test, the supernormality effect of small samples may disturb statistical testing.

Experimental

Materials

The study group consisted of 73 oligo/amenorrheic women with PCOS matching NIH criteria (25), all with the clinical manifestations of hyperandrogenemia such as hirsutism and/or acne and with the elevation of free testosterone index (FTI) and/or androstenedione above the upper limit of the normal range or with a decrease below the lower limit for SHBG (i.e., 2.65 nmol/l for testosterone, 5.4 nmol/l for androstenedione and 43 nmol/l for SHBG). None of the patients had taken oral contraceptives or any other steroid medication during the previous 3 months. The patients were evaluated at the clinical departments of both institutions as outpatients, and after signing written informed consent they underwent blood sampling between days 3 and 6 of the menstrual cycle or, in the case of secondary amenorrhea, at any time. After collection of basal blood samples, a 2-hour euglycemic hyperinsulinemic clamp was performed as reported elsewhere (26). Briefly, an indwelling cannula was inserted in the antecubital vein for the simultaneous infusion of 15% glucose solution with addition of 7.5% KCl and for the insulin infusion (1 mIU/kg/min; 25 IU of the regular insulin, Actrapid HM, Novo Nordisk, in 50 ml of sodium saline solution (0.9%)), and another cannula was inserted into the wrist vein on the same hand. The hand was heated in a heating pad at 65°C and blood samples for the determination of blood glucose were taken every 5 minutes. The rate of glucose infusion was manually adjusted to maintain blood glucose within the range of 5.0 mmol/l 5%. The MCRg (ml/kg/min) represents the ratio of the amount of glucose supplied to maintain blood glucose levels during the last 20 minutes of the clamp and the average blood glucose concentration in the same period. This parameter reflects insulin sensitivity filtering-off the effect of blood glucose. Triglycerides were determined enzymatically (reagents from Boehringer Mannheim; Mannheim, Germany, using a Cobas Mira S autoanalyzer; Hoffman-La Roche, Basel, Switzerland); testosterone, androstenedione, and SHBG were determined as described previously (27). Blood glucose was determined in the whole blood using the electrochemical method (Super GL, Freital, Germany).

Proposed procedure

The procedure for examination of influential points in data and the construction of a linear regression model consists of the following steps of the regression triplet examination. 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 1. Data – detection of influential points The statistical analysis of special 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 2. Model – 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 correlation coefficient *R*, the determination coefficient, or multiplied by 100% as the regression rabat 100*D,* are computed. The mean quadratic error of prediction, *MEP*, and the Akaike information criterion, *AIC*, are calculated to examine the quality of the model. A partial regression graph and partial residual graphs show statistical significance of individual parameters.

Step 3. Method – construction of a more accurate model According to the test for fulfillment of the conditions for the least-squares method and the results

Table 1 A survey of the influential points that were indicated using various graphical diagnostic tools.

Diagnostics 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, Figure 2A	7, 23, 29, 51, 33,	7, 23, 29, 51, 33,	7, 23, 29, 51	7
2. Williams graph, Figure 2B	7, 23, 29, 30, 33, 51, 73	7, 23, 29, 33, 51, 73	7, 23, 29, 33 51, 73	4, 7, 21, 36, 38 40, 41, 45
3. Pregibon graph, Figure 2C	7, 21, 23, 36, 38, 40, 45	7, 21, 23, 36, 38, 40, 45	---	---
4. McCulloh-Meeter graph, Figure 2D	7, 23, 29, 36, 38, 40	7, 23, 29, 36, 38, 40	7, 23, 29, 36 38, 40	7, 36, 38, 40
5. Gray's L-R graph, Figure 2E	7, 23, 29, 36	7, 23, 29, 36	7, 29	7, 21, 36, 38, 40, 41, 45, 47
B. Diagnostics based on scalar and vector influence measures				
6. Cook measure *D, Figure 3*C	7, 21, 23, 29, 36, 38, 40, 45	7, 21, 23, 29, 36, 38, 40, 45	--	--
7. Atkinson measure *A*, Figure 3D	7, 23, 29, 36, 38, 40	7, 23, 29, 36, 38, 40	--	--
8. Belsey measure *DFFITS*, Figure 3E	7, 15, 23, 29, 36, 38, 40	7, 15, 23, 29, 36, 38, 40	--	--
9. Anders-Pregibon diagnostic *AP*, Figure 3F	7, 21, 23, 36, 38, 40, 45	7, 21, 23, 36, 38, 40, 45	--	--
10. Cook-Weisberg likelihood measure $LD(b)$, Figure 3G	7, 23, 36, 38, 40	7, 23, 36, 38, 40	--	--
11. Cook-Weisberg likelihood measure $LT(s^2)$, Figure 3H	7, 23, 29, 51	7, 23, 29, 51	--	--
12. Cook-Weisberg likelihood measure $LD(b,s^2)$, Figure 3I	7, 23, 29, 36, 38, 40	7, 23, 29, 36, 38, 40	--	--
C. Index graphs of various residuals and hat matrix elements				
13. Ordinary residuals $\hat{e}$, Figure 1A and B	7, 23, 29, 30, 32, 33, 51, 56, 73	7, 23, 29, 30, 32, 33, 51, 56, 73	--	--
14. Normalized residuals $\hat{e}_N$, Figure 1C	10, 14, 23, 26, 27, 29, 30, 32, 33, 39, 42, 51, 56, 73	10, 14, 23, 26, 27, 29, 30, 32, 33, 39, 42, 51, 56, 73	--	--
15. Standardized residuals $\hat{e}_S$, Figure 1D	7, 23, 29, 30, 32, 51, 56, 73	7, 23, 29, 30, 32, 51, 56, 73	--	--
16. Jackknife residuals $\hat{e}_J$, Figure 1E	7, 23, 29, 30, 32, 51, 56, 73	7, 23, 29, 30, 32, 51, 56, 73	--	--
17. Predicted residuals $\hat{e}_P$, Figure 1F	7, 23, 29, 32, 33, 38, 51, 56, 73	7, 23, 29, 32, 33, 38, 51, 56, 73	--	--
18. Diagonal elements of hat matrix $\hat{H}_{ii}$, Figure 3A	4, 7, 21, 36, 38, 45	4, 7, 21, 36, 38, 45	--	4, 7, 21, 36, 38, 45
19. Diagonal elements of modified hat matrix $\hat{H}_{m,ii}$, Figure 3B	4, 5, 7, 15, 21, 23, 36, 38, 40, 45	4, 5, 7, 15, 21, 23, 36, 38, 40, 45	--	4, 5, 7, 15 21, 23, 36, 38, 40, 45

Suspicious points (SP) are data points in diagnostic graphs that obviously differ from the others; *influential points* (IP) are data points that are detected and are separated into outliers and high-leverages using the following testing criteria: $n=73$, $m=4$, 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, $y=t_{0.95}\ (n-m-1)=1.995$, the second line is for high-leverages, $x=2m/n=0.11$; 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 toward the part are outliers while toward the right angle of the triangle are high-leverages; 6. D_i: when $D_i>1$ then the ith point is an IP; 7. A_i: when $A_i^2>3.5$ then the ith point is an outlier; 8. $DFFITS_i$: when $|DFFITS_i|>2\sqrt{m/n}=0.468$ then the ith point is an IP; 9. AP_i: when $AP_i<1-2(m+1)/n=0.863$ then the ith point is an IP; 10., 11. and 12. LD_i: generally, when $LD_i>\chi^2_{1-a}(m+1)=11.07$ then the ith 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 $\hat{e}^2_{J,i}>3.5$ then the ith point is an outlier; 17. $\hat{e}_P$: detects SP only; 18. H_{ii}: when $H_{ii}>2m/n=0.11$ then the i-th point is a high-leverage; 19. $H_{m,ii}$: when $H_{m,ii}>2m/n=0.11$ then the ith 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. **Conclusion:** Eliminated outliers 7, 23, 29, 33, 51, 73 and eliminated suspicious points 7, 16, 23, 29, 30, 33, 34, 36, 40, 43, 51, 61, 73.

of regression diagnostics, a more accurate regression model is constructed.

Software

For the creation of regression diagnostic graphs and computation of the regression based characteristics, an algorithm Linear regression in S-Plus was written and is available on request.

Results and discussion

The simple screening method based on common laboratory and anthropometric indices, enabling us to avoid expensive and demanding clamp examinations in the majority of the subjects suspected of insulin resistance, is the most desirable. Recently, some authors reported such models in samples of women with PCOS (6, 7). We attempted to suggest a general method for detection and elimination of experimental points deteriorating the informative value of the prediction model. Hence, a model for the prediction of the metabolic clearance rate of glucose reflecting insulin sensitivity and filtering-off the effect of blood glucose was built and evaluated. A number of techniques for detection and elimination of experimental points deteriorating its informative value were demonstrated step-by-step to promote their wider use in biochemistry and medicine.

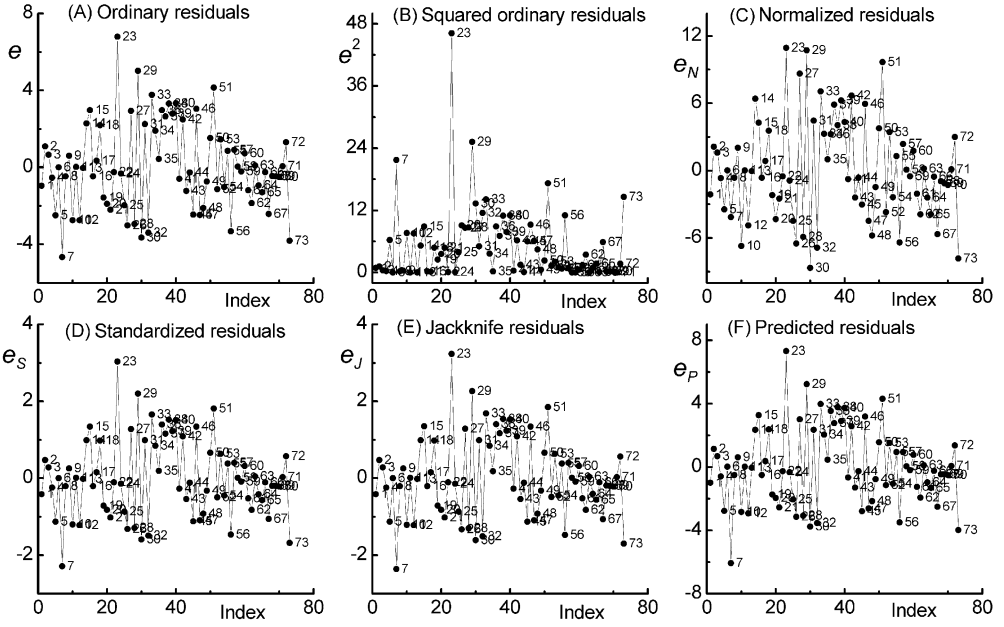

Figure 1 Index graphs of various residuals: (A) Ordinary residuals e; (B) Squared ordinary residuals e^2; (C) Normalized residuals e_N; (D) Standardized residuals e_S; (E) Jackknife residuals e_J; (F) Predicted residuals e_P.

Using backward regression, we previously found in lean PCOS women SHBG as a single reliable predictor of insulin sensitivity. In the present study, we extended this observation for the further common laboratory and anthropometric indices in both lean and obese women. Nevertheless, the primary aim was to propose a regression model and to find influential points in the data and to eliminate them.

1. Data – detection of influential points

The data quality analysis concerns an analysis detection of the influential points, which cause many problems in regression analysis by shifting the parameter estimates or increasing the variance of the parameters:

(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 1.

A survey of all the diagnostics for influential point detection shows that diagnostics plots are the most efficient tool because they are able to separate influential points into outliers and high-leverages. A survey of identified suspicious points by various types of diagnostic measures is given in Table 1. It is clear that there are some local differences given by severity of cut-off values but the majority of measures indicate the same points.

Ordinary residuals (Figure 1A and B) are always associated with a non-constant variance; they may not indicate strongly deviant points. Even though the common practice of 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, 23, 29, 30, 32, 33, 51, 56 and 73 may be considered to be suspicious and some testing diagnostics for influential points should be applied.

In the case of normalized residuals (Figure 1C), 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 10, 14, 23, 26, 27, 29, 30, 33, 39, 42, 51, 56 and 73 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 (Figure 1D) are the same as those of the ordinary residuals and indicate suspicious points 7, 23, 29, 30, 32, 51, 56 and 73 only. The maximum values of $\hat{e}_S$ are bounded $\sqrt{n-m} = 8.31$. This influential points criterion also seems to be misleading.

For jackknife residuals (Figure 1E) an approximate rule may be applied: strongly influential points (i.e., outliers) have $\hat{e}^2_{J,i} > 3.5$, but for high-leverages, however, these residuals do not give any indication: according to this criterion the points 7, 23, 29, 30, 32, 51, 56 and 73 are outliers.

Predictive residuals are able to find suspicious points only – 7, 23, 29, 32, 33, 38, 51, 56 and 73, as shown in Figure 1F.

(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 H_{ii} leads to five diagnostic graphs of influential points (the data set of size $n = 73$, $m = 4$):

The graph of predicted residuals (Figure 2A), one of the simplest graphs, separates outliers 7, 23, 29 and 51 located far from its central pattern on the line $y = x$ from high-leverage point 7 outside and far from the line $y = x$.

The Williams graph (Figure 2B) has two testing boundary lines, the first line for outliers

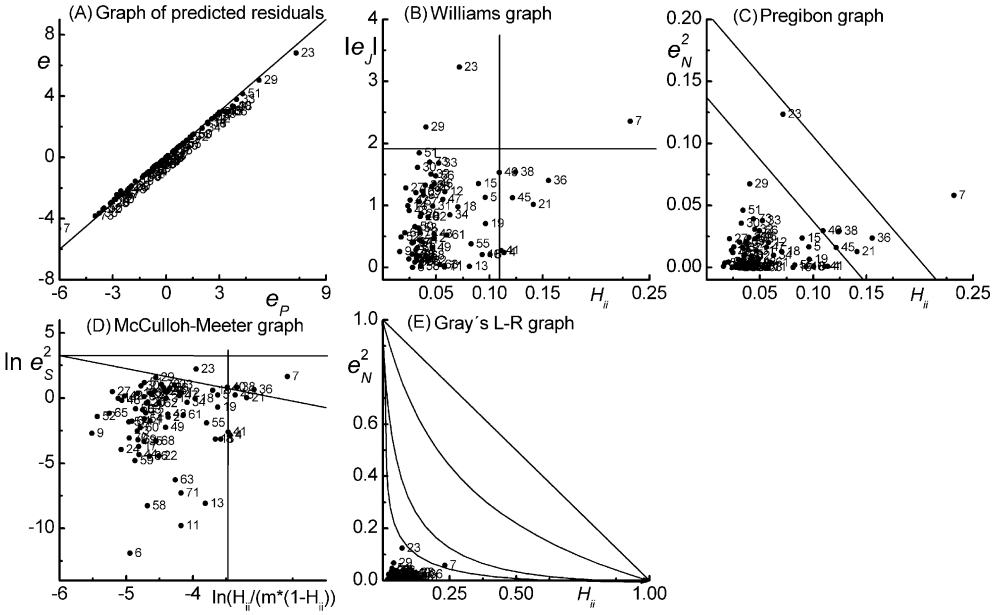

Figure 2 Diagnostics based on residual plots and hat matrix elements: (A) Graph of predicted residuals, (B) Williams graph, (C) Pregibon graph, (D) McCulloh-Meeter graph, (E) Gray's L-R graph.

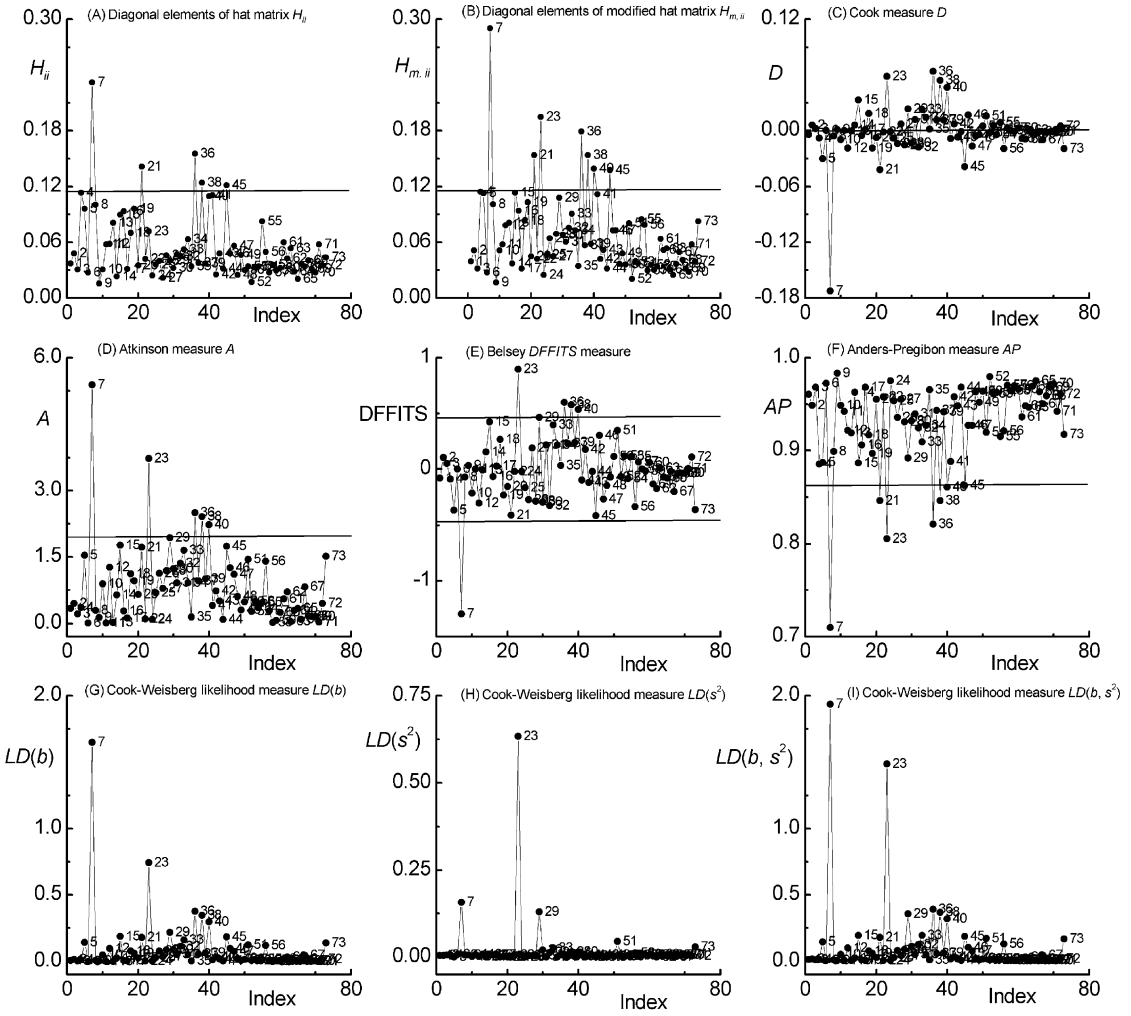

Figure 3 Index graphs of vector and scalar influence measures: (A) Diagonal elements of the hat matrix H_{ii}; (B) Diagonal elements of the modified hat matrix $H_{m,ii}$; (C) Cook measure D; (D) Atkinson measure A; (E) Belsey's $DFFITS$ measure; (F) Anders-Pregibon measure AP; (G) Cook-Weisberg likelihood measure $LD(\boldsymbol{b})$; (H) Cook-Weisberg likelihood measure $LD(s^2)$; (I) Cook-Weisberg likelihood measure $LD(\boldsymbol{b}, s^2)$.

Table 2 Estimates of four unknown parameters of the linear regression model before and after removal of outliers and suspicious points in a process of regression model building and testing.

1st step of regression triplet analysis: original data were used.
Data: 73 original data points used.
Model: $y=\beta_0+\beta_1\ x_1+\beta_2\ x_2+\beta_3\ x_3$ used: $R=0.7073$, $D=50.03\%$, $MEP=5.96$, $AIC=127.22$, $s(e)=2.3$,
$\beta_0=13.55(1.71)$ Parameter is significant. Estimated significance level $p=2.68E\text{-}11$.
$\beta_1=-0.220(0.062)$ Parameter is significant. Estimated significance level $p=0.0007$.
$\beta_2=0.053(0.011)$ Parameter is significant. Estimated significance level $p=1.43E\text{-}05$.
$\beta_3=-2.061(0.721)$ Parameter is significant. Estimated significance level $p=0.0056$.
Method: tests of assumptions about the least squares.
(a) Fisher-Snedecor test for model significance, $F=23.028$, $F(0.95, m-1, n-m)=2.737$, Model is significant,
(b) Scott test of multicollinearity: $SC=-0.067$, Model is correct without multicollinearity.
(c) Cook-Weisberg test of heteroscedasticity: $S_f=2.956$, $\chi^2(0.95, 1)=3.841$, Residuals exhibit homoscedasticity.
(d) Jarque-Bera test for normality: $C=2.237$, $\chi^2(0.95, 2)=5.991$, Normality of residuals is accepted.
(e) Wald test for autocorrelation: $W_a=0.9603$, $\chi^2(0.95, 1)=3.841$, Autocorrelation is not significant.
(f) Sign test for dependence and residuals trend: $S_g=1.032$, $N(0.975)=1.960$, There is no trend in residuals.

2nd step of regression triplet analysis: data without outliers were used.
Data: 73 data points without 6 outliers were used.
Model: $y=\beta_0+\beta_1\ x_1+\beta_2\ x_2+\beta_3\ x_3$ used: $R=0.7136$, $D=50.92\%$, $MEP=3.82$, $AIC=88.4$, $s(e)=1.9$,
$\beta_0=11.75(1.55)$ Parameter is significant. Estimated significance level $p=1.98E\text{-}10$.
$\beta_1=-0.150(0.053)$ Parameter is significant. Estimated significance level $p=0.0071$.
$\beta_2=0.048(0.010)$ Parameter is significant. Estimated significance level $p=2.05E\text{-}05$.
$\beta_3=-1.923(0.600)$ Parameter is significant. Estimated significance level $p=0.0021$.
Method: tests of assumptions about the least squares.
(a) Fisher-Snedecor test for model significance, $F=21.790$, $F(0.95, m-1, n-m)=2.750$, Model is significant,
(b) Scott test of multicollinearity: $SC=-0.052$, Model is correct without multicollinearity.
(c) Cook-Weisberg test of heteroscedasticity: $S_f=0.246$, $\chi^2(0.95, 1)=3.841$, Residuals exhibit homoscedasticity.
(d) Jarque-Bera test for normality: $C=2.180$, $\chi^2(0.95, 2)=5.991$, Normality of residuals is accepted.
(e) Wald test for autocorrelation: $W_a=0.075$, $\chi^2(0.95, 1)=3.841$, Autocorrelation is not significant.
(f) Sign test for dependence and residuals trend: $S_g=0.510$, $N(0.975)=1.960$, There is no trend in residuals.

3rd step of regression triplet analysis: data without suspicious points were used.
Data: 73 data points without 13 suspicious points were used.
Model: $y=\beta_0+\beta_1\ x_1+\beta_2\ x_2+\beta_3\ x_3$ used: $R=0.7505$, $D=56.33\%$, $MEP=3.51$, $AIC=74.6$, $s(e)=1.8$,
$\beta_0=13.44(1.75)$ Parameter is significant. Estimated significance level $p=2.75E\text{-}10$.
$\beta_1=-0.176(0.057)$ Parameter is significant. Estimated significance level $p=0.0032$.
$\beta_2=0.040(0.011)$ Parameter is significant. Estimated significance level $p=0.0010$.
$\beta_3=-2.723(0.637)$ Parameter is significant. Estimated significance level $p=7.63E\text{-}05$.
Method: tests of assumptions about the least squares.
(a) Fisher-Snedecor test for model significance, $F=24.075$, $F(0.95, m-1, n-m)=2.769$, Model is significant,
(b) Scott test of multicollinearity: $SC=-0.011$, Model is correct without multicollinearity.
(c) Cook-Weisberg test of heteroscedasticity: $S_f=0.080$, $\chi^2(0.95, 1)=3.841$, Residuals exhibit homoscedasticity.
(d) Jarque-Bera test for normality: $C=1.278$, $\chi^2(0.95, 2)=5.991$, Normality of residuals is accepted.
(e) Wald test for autocorrelation: $W_a=0.736$, $\chi^2(0.95, 1)=3.841$, Autocorrelation is not significant.
(f) Sign test for dependence and residuals trend: $S_g=1.473$, $N(0.975)=1.960$, There is no trend in residuals.

Conclusion: In the 3rd step the best regression model was found.

$y=t_{0.95}(n-m-1)=1.995$ detecting outliers 7, 23, 29, 33, 51 and 73, and the second for high-leverage points $x=2m/n=0.11$, detecting high-leverages 4, 7, 21, 36, 38, 40, 41 and 45.

The Pregibon graph (Figure 2C) is able to distinguish strongly influential points from medium influential points. Point 7 is strongly influential, while points 21, 23, 36, 38, 40 and 45 are found as medium influential.

The McCulloh-Meeter graph (Figure 2D) has two testing boundary lines, the first for outliers, $y=\ln[(n-m)\ t^2_{0.95}(n-m)]$, behind which two outliers were indicated and the second for high-leverages $x=\ln[2/(n-2m)]$, behind which high leverages are found: 7, 36, 38 and 40.

Gray's L-R graph (Figure 2E) indicates strongly influential points 7, 23, 29 and 36 and separates them into outliers 7 and 29, i.e., points that lie high in the y-axis, and high-leverages 7, 21, 36, 38, 40, 41, 45 and 47, which lie in the direction of the x-axis.

(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 that may be divided according to two principal approaches: the first is based on the examination of changes that 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 model of inflated variance. For analysis of the diagonal elements of the projection hat matrix (Figure 3A and B) the rule is valid that when $H_{ii}>2m/n=0.11$ holds, the actual ith point is the high-leverage. From that point of view, points 4, 7, 21, 36, 38 and 45 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

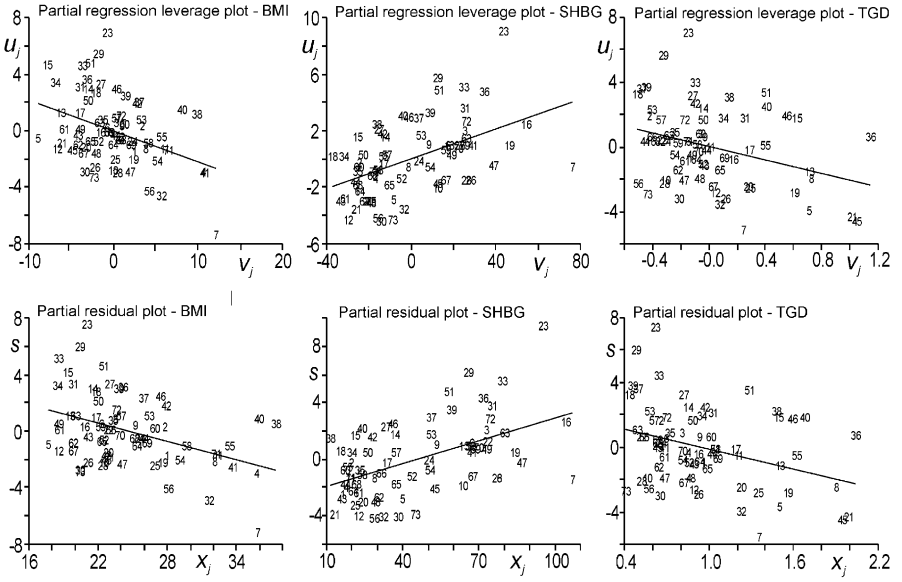

Figure 4 Partial regression leverage plots (upper line) and partial residual plots (lower line) for three regressors, *BMI* (on the left), *SHBG* (in the middle) and *TGD* (on the right).

rule, $H_{m,ii} > 2m/n = 0.11$, the diagonal elements of the extended hat matrix $H_{m,ii}$ detect both outliers and high-leverages 4, 5, 7, 15, 21, 23, 36, 38, 40 and 45.

The Cook measure D_i (Figure 3C) is used in connection with an approximate rule: when $D_i > 1$, the shift of parameter estimate $\boldsymbol{b}$ is greater than the 50% confidence region and the relevant ith point is rather influential. According to this rule, points 7, 21, 23, 29, 36, 38, 40 and 45 are influential.

With designed experiments, usually $H_{ii} = m/n$, the Atkinson measure (Figure 3D) is numerically equal to the jackknife residual $\hat{e}_J$. The same empirical rule $\hat{e}^2_{J,i} > 3.5$ for the detection of influential points may be

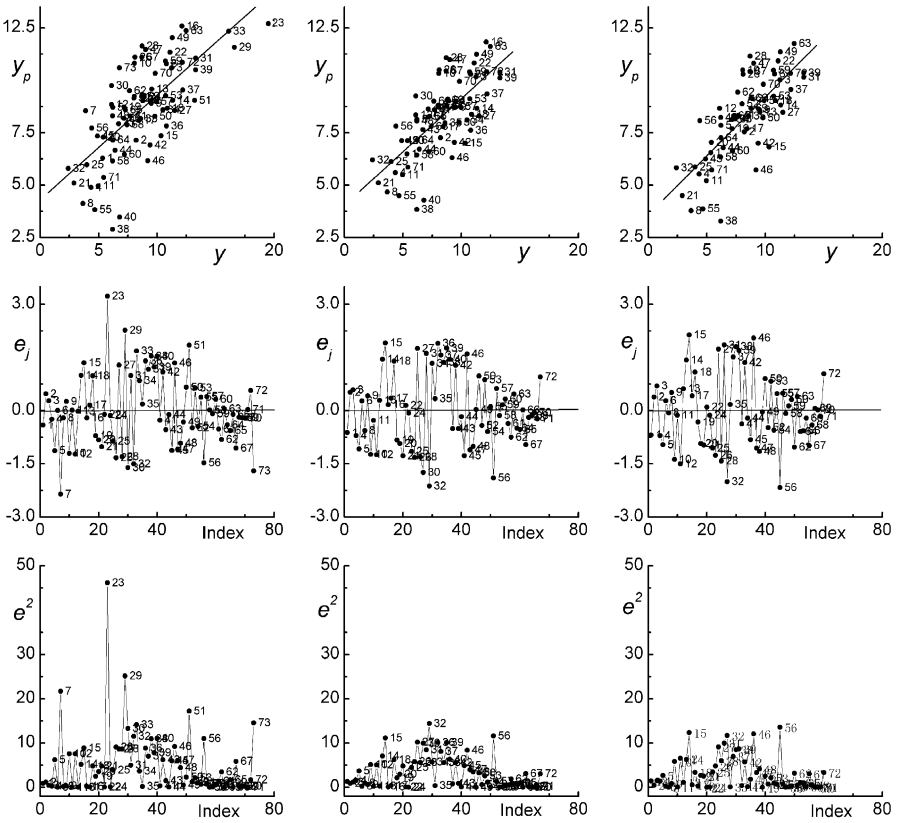

Figure 5 The effect of influential and suspicious points on the prediction ability (upper line), indicated with the use of jackknife residuals e_J (middle line) and indicated with the use of squared residuals (lower line): for 73 points of the original data (on the left), for data without removed outliers 7, 23, 29, 33, 51, 73 (in the middle) and for data without removed suspicious points 7, 16, 23, 29, 30, 33, 34, 36, 40, 43, 51, 61, 73 (on the right).

used, and points 7, 23, 29, 36, 36, 38 and 40 were found influential.

In the case of the Belsey's *DFFITS* measure (Figure 3E) the *i*th point is tested and found to be significantly influential when it is true that $DFFITS > 2\sqrt{(m/n)} = 0.468$. Influential points were indicated (7, 15, 23, 29, 36, 38, 40) with the *DFFITS* measure.

According to the Anders-Pregibon measure (Figure 3F), the *i*th point is tested and considered to be influential if $AP_i < 1-2(m+1)/n = 0.863$, and influential points were indicated to be 7, 21, 23, 36, 38, 40 and 45.

There are three Cook-Weisberg likelihood measures, i.e., $LD_i(\boldsymbol{b})$ on Figure 3G, $LD_i(s^2)$ on Figure 3H and $LD_i(\boldsymbol{b}, s^2)$ on Figure 3I. 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 influential points 7, 23, 36, 38 and 40, $LD_i(s^2)$ suspicious points 7, 23, 29 and 51, and $LD_i(\boldsymbol{b}, s^2)$ influential points 7, 23, 29, 36, 38 and 40.

If the regression model is correct and if there are no influential points, then the rankit *Q–Q* graph forms a nearly 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. The suspicious points, however, do not fulfill this assumption and therefore they could be tested as they are of an influential nature. The influential points indicated are also located beyond the ends of the straight line on the *Q–Q* graph of predicted and normalized residuals.

2. Model – significance test of parameter estimates

The estimates of all parameters β_0, β_1, β_2, and β_3 are significant (denoted by the letter A in brackets), Table 2. The model was described with the correlation coefficient $R = 0.7073$, the determination coefficient $D = 50.03\%$ thus expressing a percentage of variability explained by the regression model; the mean error of prediction $MEP = 5.958$, the Akaike information criterion $AIC = 127.22$ and the residual standard deviation $s(e) = 2.33$ were also calculated. All these statistics excluding *R* can be used as resolution criteria for the selection of the best regression model among several plausible ones. Using the original set of data, the OLS finds the regression model $y = 13.55(1.71, A) - 0.220(0.062, A)\ BMI + 0.053(0.011, A)\ SHBG - 2.06(0.72, A)\ TGD$, where standard deviations of the parameters estimated are in parentheses and the letter A means that is accepted as a statistically significant estimate. Figure 4 shows the partial regression leverage plots and the partial residual plots. The linearity of all partial regression leverage plots and partial residual plots for *BMI*, *SHBG* and *TGD* and non-zero slopes of straight lines proves the correctness of the proposed regression model. The quality of estimates may be classified according to the spread of points around the regression straight line. This spread is connected with a partial regression coefficient between *y* and the corresponding x_j.

The sign test for non-randomness of residuals, caused either by a false model or by the outliers, proves here that there are no more outliers and that the proposed regression model is correct.

3. Method – construction of a more accurate regression model

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 the data or to use a robust regression method. One of the greatest disadvantages of the robust method application is a preference for the regression model proposed, here $y = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_3\ x_3$. If a proposed model is unsuitable, robust methods can 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 all outliers (Figure 5B) or suspicious points (Figure 5C) should be excluded from the original data set and new parameter estimates should be calculated. Figure 5 shows an influence of three steps of data analysis on the prediction ability of the model proposed. The upper part indicates a non-zero straight line that proves a good prediction of the metabolic clearance rate of glucose from individual regressors BMI, SHBG and TBG. The middle part of the Figure indicates influential points with the use of jackknife residuals e_J, while the lower part of the Figure indicates points with the use of squared residuals e^2. It can be seen that the following rule is valid: the less outliers in the data, the more reliable the prediction ability of the proposed regression model.

Conclusions

Statistical conclusion

In the interactive PC-aided diagnosis of the 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, increasing the variance of the parameters or leading to bad prediction ability. 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 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.

Biochemical conclusion

The model for prediction of the metabolic clearance rate of glucose reflecting insulin sensitivity and filtering-off the effect of blood glucose was built and evaluated. A general method for detection and elimination of experimental points deteriorating the informative value of the model was demonstrated to promote its wider use in biochemistry and medicine. From the clinical viewpoint, it is obvious that the clamp parameter MCRg in women with PCOS can be predicted from the regression model including commonly measured indices such as BMI, SHBG and triglycerides and explaining 57% of the total variability found in MCRg. The positive relationship of MCRg with SHBG, as well as the negative correlations with BMI and triglycerides in women including both lean and obese subjects, are in accordance with the report of Cibula et al. (7). The results show that there is no quantitative difference in the prediction of MCRg between lean PCOS women and the sample involving the total population of PCOS women of fertile age.

Acknowledgements

The financial support of the Ministry of Education (Grant No. MSM253100002) and of the Grant Agency of the Ministry of Health of the Czech Republic (Grants NB/7391-3, NB/6696-3 and NH/65583) are gratefully acknowledged.

References

1. Dunaif A, Finegood DT. Beta-cell dysfunction independent of obesity and glucose intolerance in the polycystic ovary syndrome. J Clin Endocrinol Metab 1996;81:942–7.
2. Dunaif A, Segal KR, Futterweit W, Dobrjansky A. Profound peripheral insulin resistance, independent of obesity, in polycystic ovary syndrome. Diabetes 1989;38:1165–74.
3. Holte J, Bergh T, Berne C, Berglund L, Lithell H. Enhanced early insulin response to glucose in relation to insulin resistance in women with polycystic ovary syndrome and normal glucose tolerance. J Clin Endocrinol Metab 1994;78:1052–8.
4. Ovesen P, Moller J, Ingerslev HJ, Jorgensen JO, Mengel A, Schmitz O, et al. Normal basal and insulin-stimulated fuel metabolism in lean women with the polycystic ovary syndrome. J Clin Endocrinol Metab 1993;77:1636–40.
5. Toprak S, Yonem A, Cakir B, Guler S, Azal O, Ozata M, et al. Insulin resistance in nonobese patients with polycystic ovary syndrome. Horm Res 2001;55:65–70.
6. Gennarelli G, Holte J, Berglund L, Berne C, Massobrio M, Lithell H. Prediction models for insulin resistance in the polycystic ovary syndrome. Hum Reprod 2000;15:2098–102.
7. Cibula D, Skrha J, Hill M, Fanta M, Hakova L, Vrbikova J, et al. Prediction of insulin sensitivity in nonobese women with polycystic ovary syndrome. J Clin Endocrinol Metab 2002;87:5821–5.
8. Meloun M, Militký J. Detection of single influential points in OLS regression model building. Anal Chim Acta 2001;439:169–91.
9. Meloun M, Militký J, Hill M, Brereton RG. Crucial problems in regression modelling and their solutions. The Analyst 2002;127:433–50.
10. Meloun M, Militký J, Forina M, editors. Chemometrics for analytical chemistry, vol. 2. PC-aided regression and related methods. Chichester: Ellis Horwood, 1994:pp 64.
11. Belsey DA, Kuh E, Welsch RE, editors. Regression diagnostics: identifying influential data and sources of collinearity. New York: Wiley, 1980.
12. Cook RD, Weisberg S, editors. Residuals and influence in regression. London: Chapman & Hall, 1982:354.
13. Barnett V, Lewis T. Outliers in statistical data. New York: Wiley, 1984.
14. Hadi AS. In: Rao CR, editor. Handbook of statistics. 1993; 9:775–802.
15. Williams DX. Letter to the Editor. Applied Statist 1973;22:407–8.
16. Pregibon D. Logistic regression diagnostics. Ann Statist 1981;9:45–52.
17. McCulloch CE, Meeter D. Technometrics 1983;25:152–5.
18. Gray JB. Graphics for regression diagnostics. Proceedings of the Statistical Computing Section. J Am Statist Assoc 1985;80:102–7.
19. Hampel FR. The influence curve and its role in robust estimation. J Am Statist Assoc 1974;69:383–93.
20. Cook RD. Influential observation in linear regression J Am Statist Assoc 1979;74:169–74.
21. Atkinson AC, editor. Plots, transformations and regression. An introduction to graphical methods of diagnostic regression analysis. Oxford: Clarendon Press, 1985.
22. Belsey DA, Kuh E, Welsch RE, editors. Regression diagnostics. Identifying influential data and sources of collinearity. New York: Wiley, 1980.
23. Gentleman JF, Wilk MB. Detecting outliers. II. Supplementing the direct analysis of residuals. Biometrics 1975;31:387–410.
24. Jarque CM, Bera AK. A test for normality of observations and regression residuals. Int Stat Rev 1987;55:163.
25. Dunaif A. Insulin resistance and the polycystic ovary syndrome: mechanism and implications for pathogenesis. Endocr Rev 1997;18:774–800.
26. DeFronzo RA, Tobin JD, Andres R. Glucose clamp technique: a method for quantifying insulin secretion and resistance. Am J Physiol 1979;237:E214–23.
27. Vrbikova J, Hill M, Starka L, Cibula D, Bendlova B, Vondra K, et al. The effects of long-term metformin treatment on adrenal and ovarian steroidogenesis in women with polycystic ovary syndrome. Eur J Endocrinol 2001;144:619–28.

Received July 24, 2003, accepted January 23, 2004