Matematické Metody v Ekonometrii 6.

Multiple Regression Analysis - regression diagnostic based on residuals

Blanka Šedivá

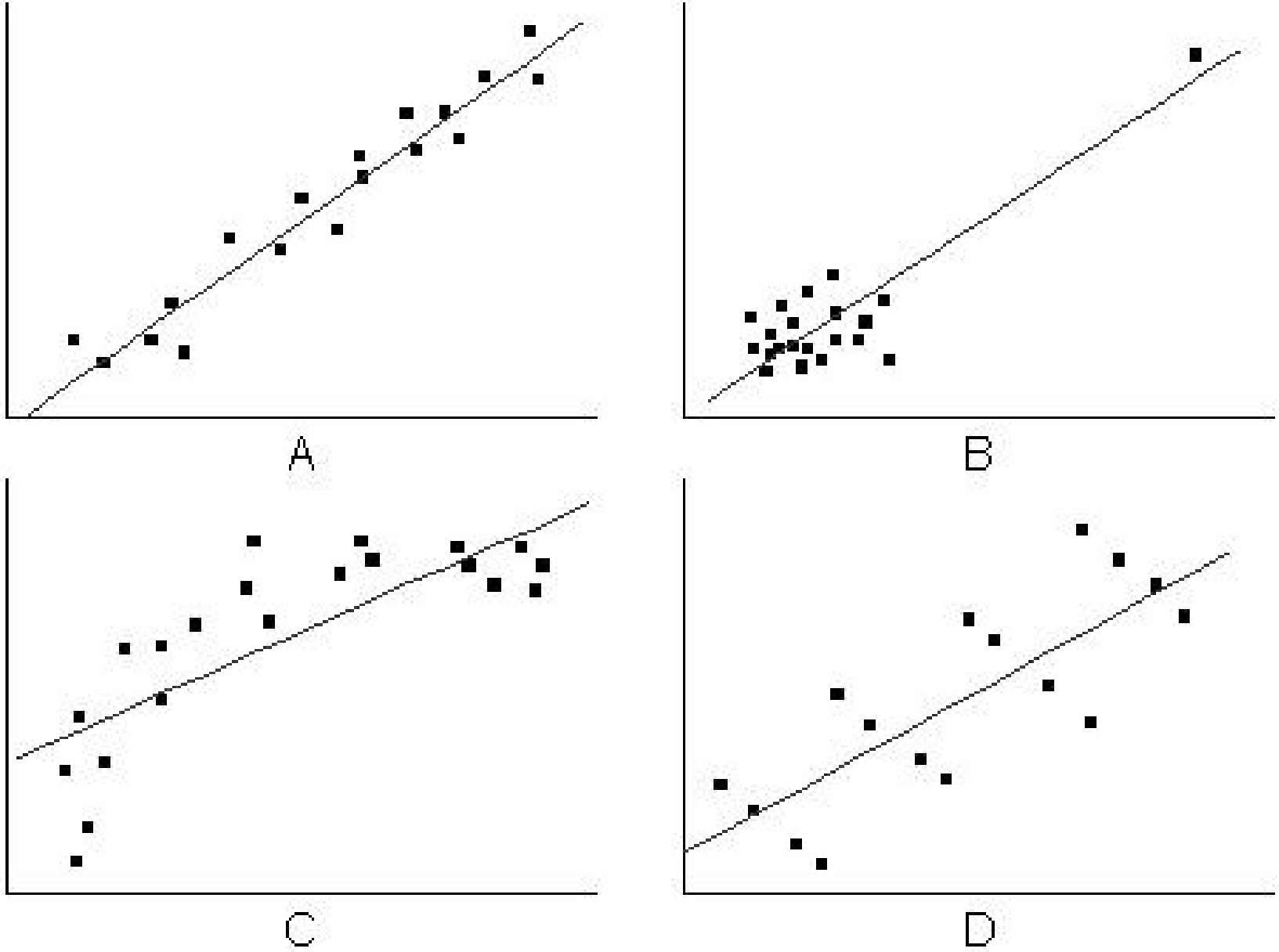
KMA

zimní semestr 2016/2017

Classical Linear Normal Regression Model

- normal linear regression model $\boldsymbol{Y} \sim N\left(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}\right)$
- assumptions of the linear regression model
  - linearity of the regression model
  - full rank of matrix $\boldsymbol{X}$
  - exogeneity of the independent variables
  - homoscedasticity anan non autocorrelation
  - data generation
  - normal distribution
- estimation of coefficients by least square method
- statistical properties of estimators
- confidence intervals for a linear combination of coefficients
- hypothesis tests

Examples of the least square method - residuals diagnostic



Regression diagnostic

. . . represent procedures for an examination of the regression triplet (data, model, method), i.e. procedures for the identification of

(a) the data quality for a proposed model;

(b) the model quality for a given set of data;

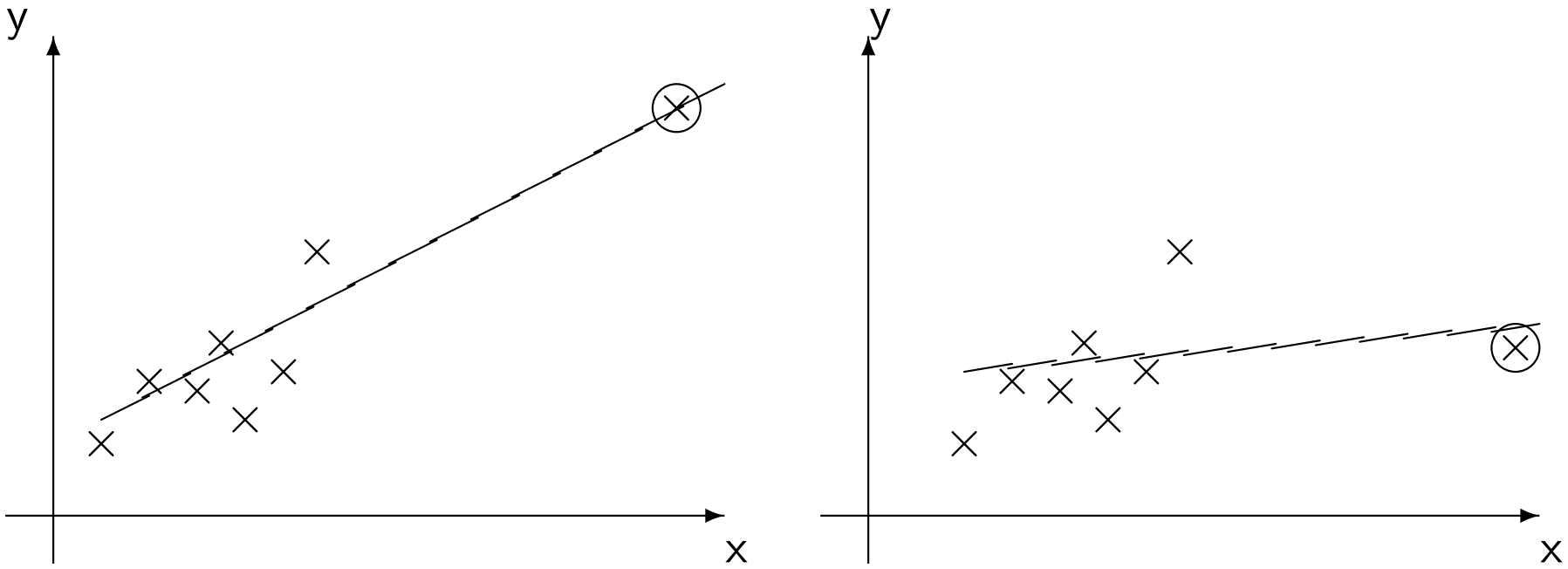
(c) a fulfillment of all least-squares assumptions.

Unusual and Influential Data

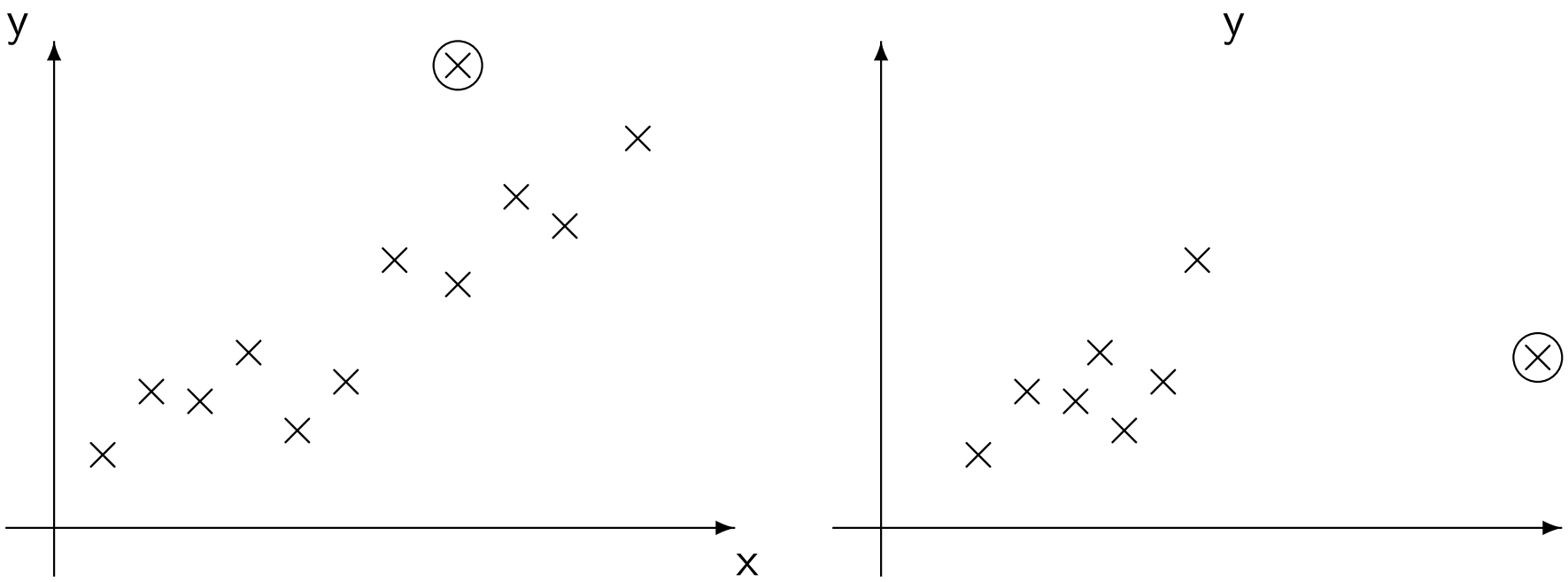
- **Outliers**
  - An observation with large residual.
  - An observation whose dependent-variable value is unusual given its values on the predictor variables.
  - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
- **Leverage**
  - An observation with an extreme value on a predictor variable
  - Leverage is a measure of how far an independent variable deviates from its mean.
  - These leverage points can have an effect on the estimate of regression coefficients.
- **Influence**
  - Influence can be thought of as the product of leverage and outlierness.
  - Removing the observation substantially changes the estimate of coefficients.

Outliers and influential observations

- Outliers – cases with extreme values that differ greatly from the rest of your sample
- More formally: "influential cases"
  Outliers can result from:
  - Errors in coding or data entry
  - Highly unusual cases
  - Or, sometimes they reflect important "real" variation
- Even a few outliers can dramatically change estimates of the slope, especially if n is small.



Obrázek: Uninfluential (left) and influential (right) observations



Obrázek: outliers (left) and leverage points (right)

Diagnostics based on residuals analysis

- $\varepsilon \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$
- Ordinary residuals are defined by $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}$
- Classical analysis is based on the incorrect assumption that residuals are good estimates of errors ε_i, the reality is more complex, residuals e_i being a projection of vector $\boldsymbol{y}$ into a subspace of dimension $n - p$
- $\boldsymbol{e} = \boldsymbol{M} \cdot \varepsilon$, where $\boldsymbol{M}$ is residual matrix
- each residual e_i is a linear combination of all errors ε_i
- $\boldsymbol{e} \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{M}\right)$
- $RSE = \boldsymbol{e}^T \boldsymbol{M}(\boldsymbol{e})$
- $\boldsymbol{X}^T \boldsymbol{e} = \mathbf{0}$

Other type of residuals I

- **Normalized residuals or scaled residuals**

$$e_i^{(N)} = \frac{e_i}{s},$$

where $s^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p}$ is estimator of σ^2.
This residua have $\mathrm{var}\left(e_i^{(N)}\right) = (1 - h_{ii})$ and are dependent.

- **Standardized residuals or internally studentized**

$$e_i^{(S)} = \frac{e_i}{s\sqrt{1 - h_{ii}}},$$

where h_{ii} diagonal elements of the projection matrix $\boldsymbol{H}$
This residua have $\mathrm{var}\left(e_i^{(N)}\right) = 1$.
The standardized residuals behave much like a Student's t random variable except for the fact that the numerator and denominator of $e_i^{(S)}$ are not independent

Other type of residuals II

- **Jackknife residuals or externally studentized residuals**

$$e_i^{(T)} = \frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}},$$

where $s^2_{(-i)}$ is estimator of σ^2 without i sample,

$$s^2_{(-i)} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}_{(-i)})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}_{(-i)})}{n-p-1},$$

where $\boldsymbol{b}_{(-i)}$ is estimator of β based on all samples n without sample $i-$

$$\boldsymbol{b}_{(-i)} = \left(\boldsymbol{X}^T_{(-i)}\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}^T_{(-i)}\boldsymbol{Y}_{(-i)},$$

kde

$$\boldsymbol{Y}_{(-i)} = [y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n]^T$$

a

Other type of residuals III

$$\boldsymbol{X}_{(-i)} = \begin{bmatrix} x_{1\,1} & x_{1\,2} & \cdots & x_{1\,p} \\ x_{2\,1} & x_{2\,2} & \cdots & x_{2\,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i-1\,1} & x_{i-1\,2} & \cdots & x_{i-1\,p} \\ x_{i+1\,1} & x_{i+1\,2} & \cdots & x_{i+1\,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n\,1} & x_{n\,2} & \cdots & x_{n\,p} \end{bmatrix}$$

The residuals which with an assumption of normality of errors have a Student distribution with $(n-p-1)$ degrees of freedom.
The jackknife residual examines the influence of individual points on the mean quadratic error of prediction.

$$e_i^{(T)} = e_i^{(S)} \sqrt{\frac{n-p-1}{n-p-\left(e_i^{(S)}\right)^2}} \qquad s_{(-i)}^2 = \frac{\boldsymbol{e}^T\boldsymbol{e} - \frac{e_i^2}{1-h_{ii}}}{n-p-1}$$

Other type of residuals IV

- **Predicted residuals or cross-validated residuals**

$$e_i^{(P)} = Y_i - \boldsymbol{X}_i \boldsymbol{b}_{(-i)} = \frac{e_i}{1 - h_{ii}}$$

where $\boldsymbol{X}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ is $i-$row of matrix $\boldsymbol{X}$.
are equal to shift C in the equation $\boldsymbol{y} = \boldsymbol{X}\beta + C\boldsymbol{i} + \varepsilon$, where $\boldsymbol{i}$ is the identity vector with the ith element equal to one, and other elements equal zero. This model expresses not only the case of an outlier where C is equal to the value of deviation, but also the case of a high-leverage point $C = d_i\beta$ where d_i is the vector of the deviation of the individual x components of the ith point.

$$e_i^{(S)} = e_i^{(P)} \frac{\sqrt{1 - h_{ii}}}{s}$$

$$e_i^{(T)} = e_i^{(P)} \frac{\sqrt{1 - h_{ii}}}{s_{(-i)}}$$

Other type of residuals V

- **Recursive residuals**
  They are computed from a sequence of regression starting with a base of m observations (m being the number of parameters to be estimated), and adding one observation at each step. The regression equation computed at each step is used to compute the residual for the next observation to be added. This sequence continues until the last residual has been computed. residuals
  - to start with $i-1$ rows of matrix $\boldsymbol{X}$ and $i-1$ rows of vector $\boldsymbol{Y}$, ${}^{(i-1)}\boldsymbol{X}$ a ${}^{(i-1)}\boldsymbol{Y}$

$$
{}^{(i-1)}\boldsymbol{Y} = [y_1, y_2, \ldots, y_{i-1}]^T
$$

$$
{}^{(i-1)}\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i-1\,1} & x_{i-1\,2} & \cdots & x_{i-1\,p} \end{bmatrix}
$$

Other type of residuals VI

- to estimate coefficients ${}^{(i-1)}\boldsymbol{b}$ based on $1, 2, \ldots, i-1$

$$ {}^{(i-1)}\boldsymbol{b} = \left({}^{(i-1)}\boldsymbol{X}^T \; {}^{(i-1)}\boldsymbol{X} \right)^{-1} \; {}^{(i-1)}\boldsymbol{X}^T \; {}^{(i-1)}\boldsymbol{Y} $$

- to compute predictive residuum r_i for$i$ based on $1, 2, \ldots, i-1$,

$$ r_i = Y_i - \boldsymbol{X}_i \; {}^{(i-1)}\boldsymbol{b} $$

- to calculate variability ${}^{(i-1)}\boldsymbol{b}$ based on $1, 2, \ldots, i$,

$$ \mathrm{var}\left({}^{(i-1)}\boldsymbol{b} \right) = {}^{(1-i)}s^2 \left({}^{(i-1)}\boldsymbol{X}^T \; {}^{(i-1)}\boldsymbol{X} \right)^{-1} $$

- to norm residuum r_i

$$ e_i^{(R)} = \frac{r_i}{\sqrt{1 + \boldsymbol{X}_i \left({}^{(i-1)}\boldsymbol{X}^T \; {}^{(i-1)}\boldsymbol{X} \right)^{-1} \boldsymbol{X}_i^T}} $$

Other type of residuals VII

The recursive residuals are mutually independent and have constant variance σ^2.
They allow the identification of any instability in a model, such as in time or autocorrelation, and are often used in normality tests or in tests of the stability of regression parameters β.

Diagnostics based on the diagonal elements of the hat matrix

- $\mathrm{tr}\,\boldsymbol{H} = \sum_{i=1}^{n} h_{ii} = p$,
- the mean value of the diagonal elements is therefore $\frac{p}{n}$
- obviously, this cut-off point will fail to nominate any observation when

$$h_{ii} > \frac{2p}{n}, \qquad \text{ev.}\frac{2(p+1)}{n}, \qquad \text{ev.}\frac{3p}{n}$$

Cook's distance

$$D_i = \frac{(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-i)})^T(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-i)})}{p \cdot s^2} = \frac{e_i^2}{s^2\, p} \cdot \frac{h_{ii}}{(1-h_{ii})^2} = \frac{\left(e_i^{(S)}\right)^2}{p} \cdot \frac{h_{ii}}{1-h_{ii}}$$

- The Cook measure D_i directly expresses the relative influence of the ith point on all parameter estimates.
- It is related to the confidence ellipsoid of the estimates but is really a shift of estimates when the ith point is left out. It is approximately true that when $D_i > 1$, the shift is greater than the 50% confidence region, so the relevant point is rather influential.
- Another interpretation of D_i is based on the Euclidean distance between the prediction vector $\hat{y}$ and the prediction vector $\hat{y}_{(-i)}$ estimated when the ith point is left out.

Welsch - Kuh distance

$$WK_i = \frac{\left(\widehat{Y}_i - \widehat{Y}_{i(-i)}\right)^T \left(\widehat{Y}_i - \widehat{Y}_{i(-i)}\right)}{s_{(-i)}\sqrt{h_{ii}}} = |e_i^{(T)}|\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

- $DFFITS_i$ measure,
- is suggested that the ith point is considered to be significantly influential on prediction when

$$2\sqrt{\frac{p}{n}}, \qquad 2\sqrt{\frac{p}{n-p}}.$$

Andrews-Pregibon diagnostic

$$AP_i = \frac{\det\left((\boldsymbol{X}^*_{(-i)})^T \boldsymbol{X}^*_{(-i)}\right)}{\det\left((\boldsymbol{X}^*)^T \boldsymbol{X}^*\right)}$$

where $\boldsymbol{X}^* = \boldsymbol{X}|\boldsymbol{y}$ is the matrix having as least column the vector $\boldsymbol{y}$.

- expresses the influence of the ith point on the volume of the confidence ellipsoid
- The diagnostic AP_i is related to the elements of the extended projection matrix $\boldsymbol{X}^*$ by the expression

$$AP_i = 1 - h_{ii} - \frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{e}} = (1 - h_{ii})\left(1 - \frac{(e_i^{(s)})^2}{n-p}\right)$$

- A point is considered to be influential if $AP_i < 1 - \frac{2(p+1)}{n}$

Partial influence of ith rows of matrix $\boldsymbol{X}$

- $DFBETAS_{ij} = \frac{b_j - b_{j(-i)}}{\sqrt{\mathrm{var}\,(b_j)}} = e_i^T \frac{c_{ij}}{\sqrt{(1-h_{ii})\boldsymbol{c}_j^T \boldsymbol{c}_j}}$ for $i = 1, 2, \ldots, n$a $j = 1, 2, \ldots, p$ and where $\boldsymbol{c}_j$ is jth column of matrix $\boldsymbol{C} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- the critical value is usually $\frac{2}{\sqrt{n}}$

- $COVRATIO_i = \frac{\det\,(s^2_{(-i)}(\boldsymbol{X}^T_{(-i)}\boldsymbol{X}_{(-i)})^{-1})}{\det\,(s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})} = \left(\frac{n-p-e_i^{(s)}}{n-p-1}\right)^p \cdot \frac{1}{1-h_{ii}}$
- the critical value is usually $|COVRATIO_i - 1| \geq \frac{3p}{n}$.

Procedure for regression model building

- Step 1. Proposal of a model: the procedure usually starts from the simplest model, with individual explanatory controllable variables not raised to powers other than the first, and with no interaction terms of the type $x_j x_k$ included.
- Step 2. Exploratory data analysis in regression: the scatter plot of individual variables and all the possible pair combinations are examined.
- Step 3. Parameter estimation: the parameters of the proposed regression model and the corresponding basic statistical characteristics of this model are determined by the classic least-squares method (LS). Individual parameters are tested for significance by way of Student's t-test.
- Step 4. Analysis of regression diagnostics: different diagnostic graphs and numerical measures are used to examine influential points, namely outliers and high-leverages. If influential points are found, it has to be decided whether these points should be eliminated from the data. If points are eliminated, the whole data treatment must be repeated.
- Step 5. Construction of a more accurate model.