IV.

Jednoduchá lineární regrese

- Lineární závislost $y = \alpha + \beta x$

koeficienty lineární regrese

$$b = \frac{\left(\sum x\right)\left(\sum y\right) - n\sum xy}{\left(\sum x\right)^2 - n\sum x^2}$$

$$a = \frac{\sum y - b\sum x}{n}$$

a, b = odhady koeficientů pomocí MNČ

Regresní rezidua

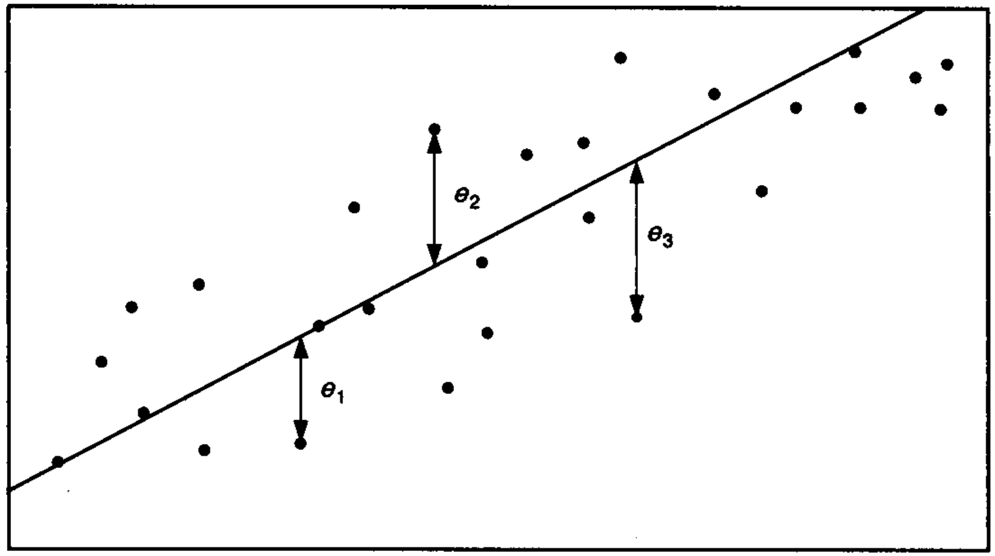
Reziduální hodnota predikce (chyba predikce) – e: $e_i = y_i - a - bx = y_i - \hat{y}_i$

MNČ: minimalizace reziduální sumy čtverců (RSC)

$$RSC = s_r^2 = \sum e_i^2 = \sum (y_i - a - bx_i)^2$$

Reziduální rozptyl

$$s_{yx}^2 = \frac{s_r^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$



Předpoklady použití metody nejmenších čtverců

- Mezi proměnnými x a y je lineární vztah.

- Náhodné chyby e mají normální rozdělení, nulovou střední hodnotu a konstantní rozptyl (homoskedasticita). Náhodné chyby e jsou vzájemně nekorelované:

  $\mathrm{cov}\,(e_i, e_j) = 0$.

- Hodnoty y_i mají pro dané hodnoty přesně určených x_i normální rozdělení. Hodnoty y jsou vzájemně nekorelované: $\mathrm{cov}\,(y_i, y_j) = 0$.

Ověřování předpokladů regresní analýzy

- Grafická analýza reziduí
- Normalita reziduí: histogram, QQplot, PPplot
- Vhodnost modelu (znaménkový test)

Testování regresních koeficientů analýzy

testy významnosti regresních koeficientů

testy shody regresních koeficientů (srovnání dvou přímek)

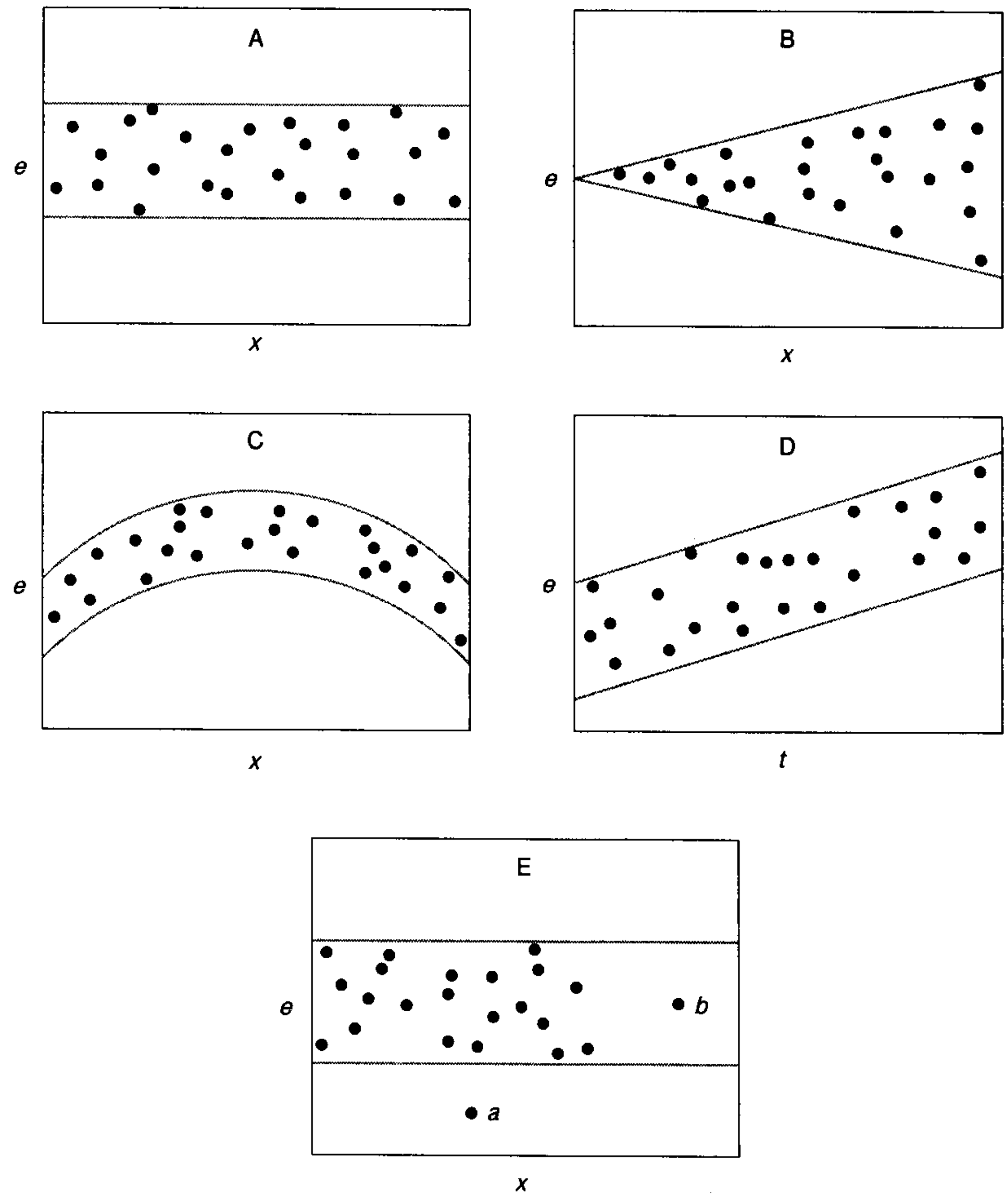
Regresní rezidua

Grafická analýza

Osa x: hodnoty x

Osa y: rezidua (e)

a) Splněn předpoklad lineárního vztahu
b) Heteroskedasticita
c) Nelinearita
d) Časová závislost
e) Odlehlé hodnoty



Testování významnosti regresních koeficientů

Pomocí t-testu

$$L_a = a \pm t_\alpha s_a$$

$$L_b = b \pm t_\alpha s_b$$

$$s_a = s_{yx}\sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - \left(\sum x\right)^2 / n}\right]}$$

$$s_b = \frac{s_{yx}}{\sqrt{\sum x^2 - \left(\sum x\right)^2 / n}}$$

Test shody regresních přímek

- RSC_c reziduální suma čtverců všech M skupin dat

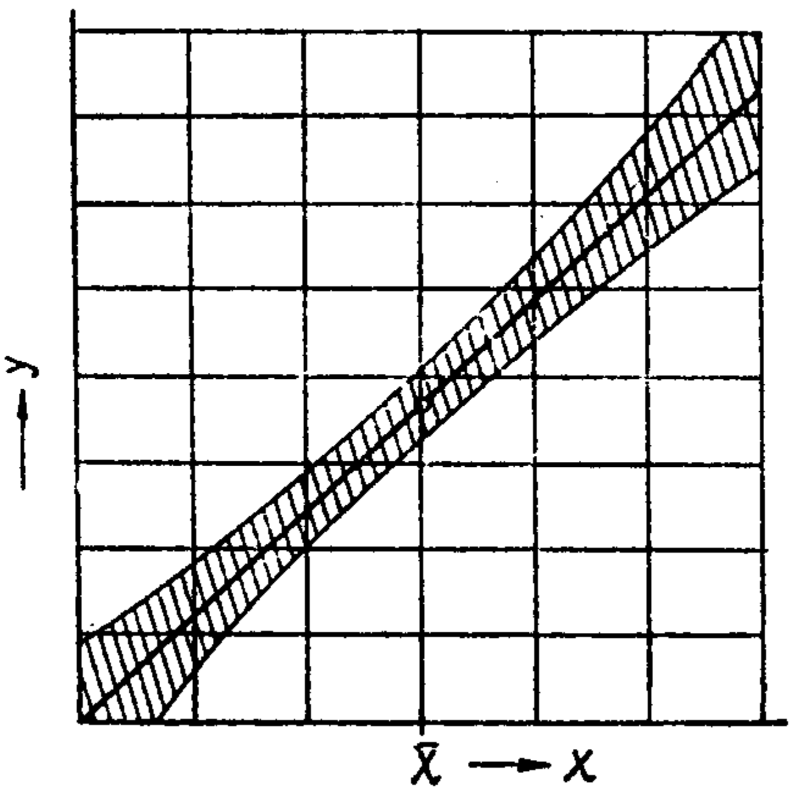
$$RSC_c = \sum_{j=1}^{M} RSC_j$$

$$F_Á = \frac{\dfrac{RSC_k - RSC_c}{2M-2}}{\dfrac{RSC_c}{n-2M}}$$

F rozdělení s 2M-2 a n-2M stupni volnosti.

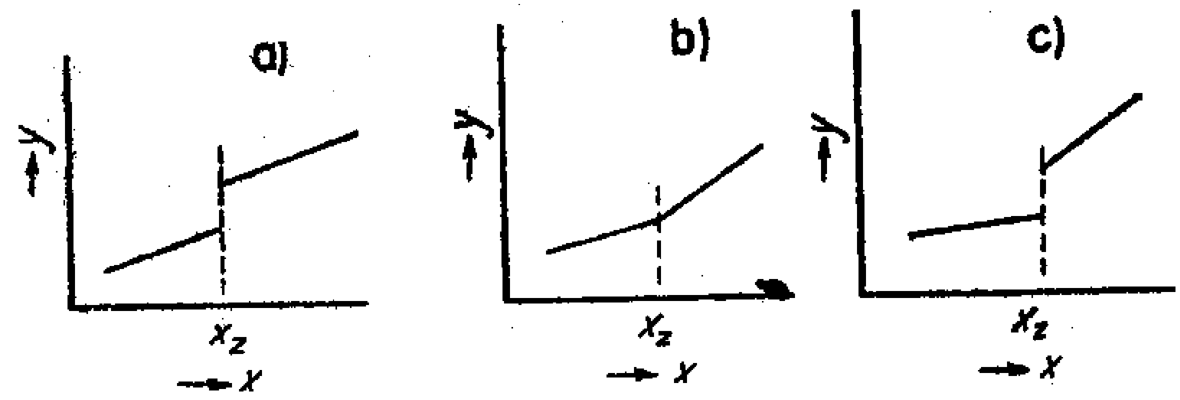
Interval spolehlivosti predikovaných hodnot

$$L_{12} = \hat{y}_i \pm s_{yx} \sqrt{\left[\frac{1}{n} + \frac{(\hat{x}_i - \bar{x})^2}{\sum x^2 - \left(\sum x\right)^2 / n} \right]}$$



Fiktivní proměnné

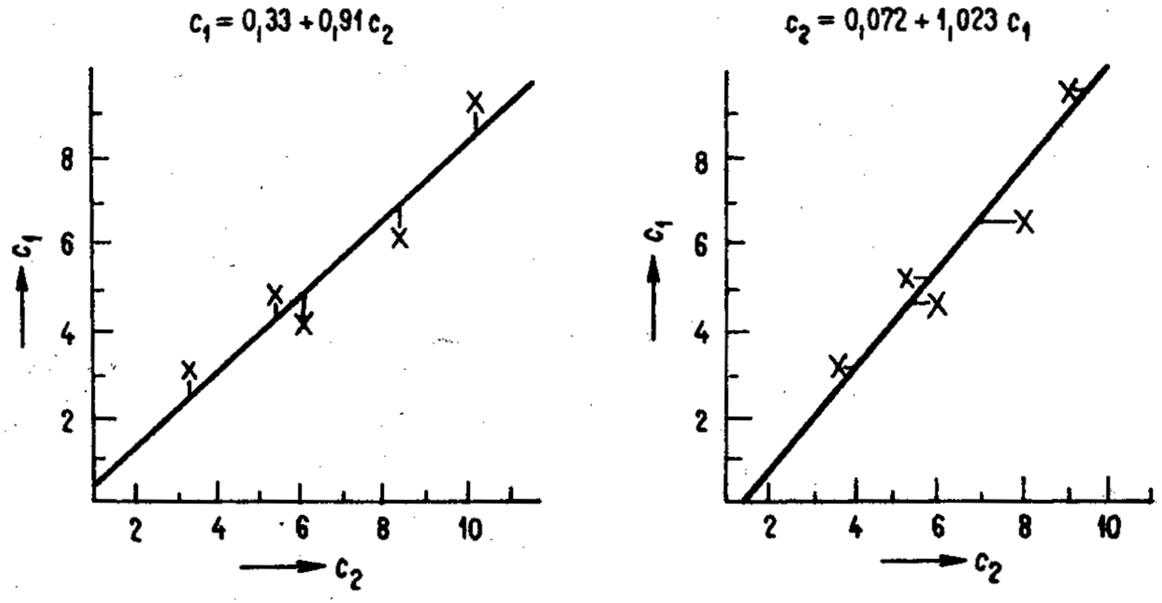
- používají se v případě nespojitostí; jsou zpravidla celočíselné, nejčastěji 0/1



Lineární lomená regrese

Neparametrická regrese

Chyby u „závisle“ i „nezávisle“ proměnné

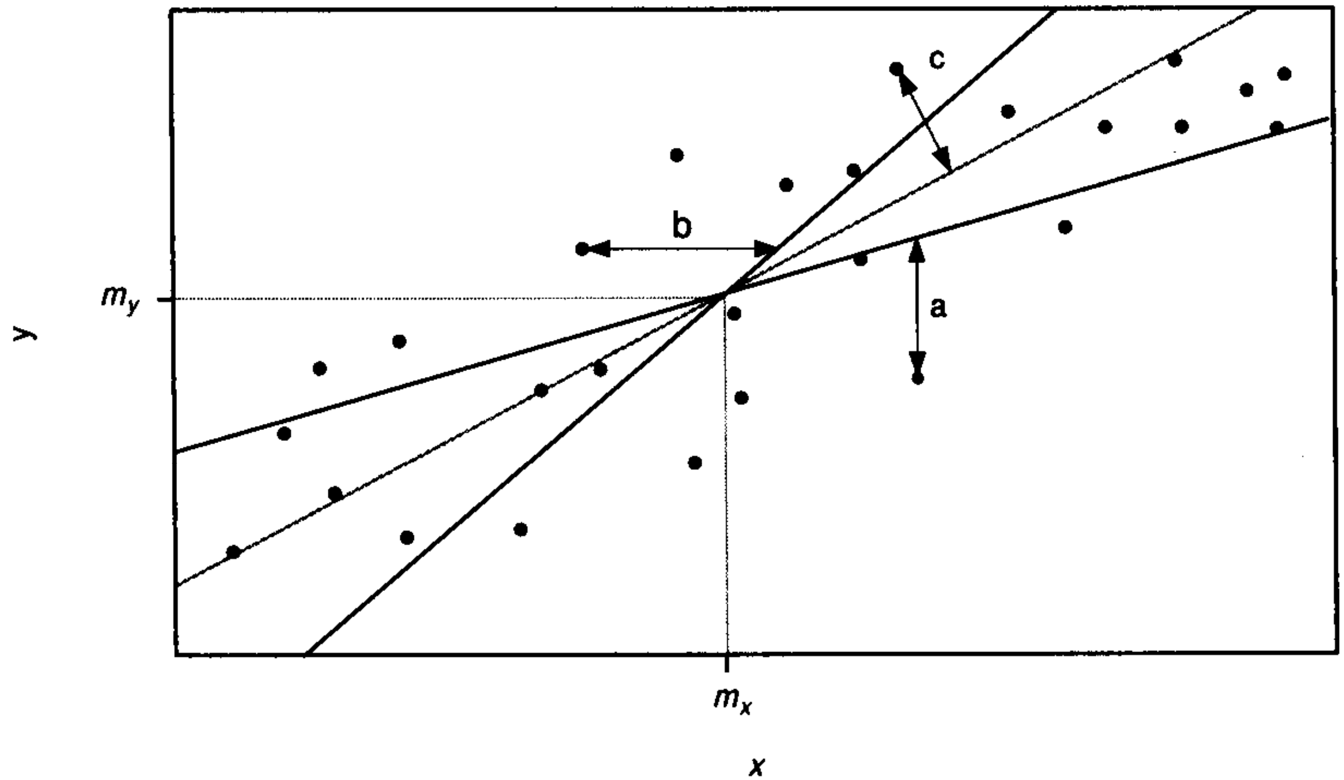


Odhad regresních přímek závislosti c_1 na c_2 a c_2 na c_1

$$r = \sqrt{b_{xy} b_{yx}}$$

Neparametrická regrese

Ortogonální regrese



a) vzdálenost pro regresi, kde *y* je závisle proměnná
b) vzdálenost pro regresi, kde *x* je závisle proměnná
c) vzdálenost pro ortogonální regresi

Neparametrická regrese

Demingova regrese

Vztah mezi dvěma metodami pomocí vážené regrese

$$\min S = \sum w_i (y_i - \hat{y}_i)^2 + v_i (x_i - \hat{x}_i)^2$$

$$x_i = \hat{x}_i + \varepsilon_i$$

$$y_i = \hat{y}_i + \delta_i$$

$$b = G + \sqrt{G^2 + 1/\lambda}$$

$$G = \frac{s_y^2 - (1/\lambda) s_x^2}{2 r s_x s_y}$$

$$a = \overline{y} - b\overline{x}$$

Neparametrická regrese

- Pasing-Bablokova regrese

medián částečných odhadů regresního koeficientu (j<i)

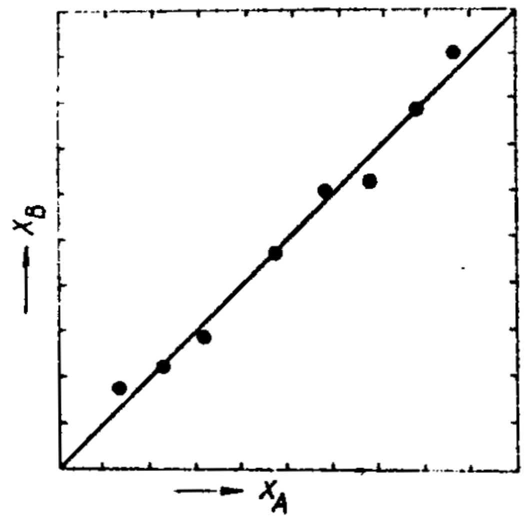
$$bij = \frac{y_i - y_j}{x_i - x_j}$$

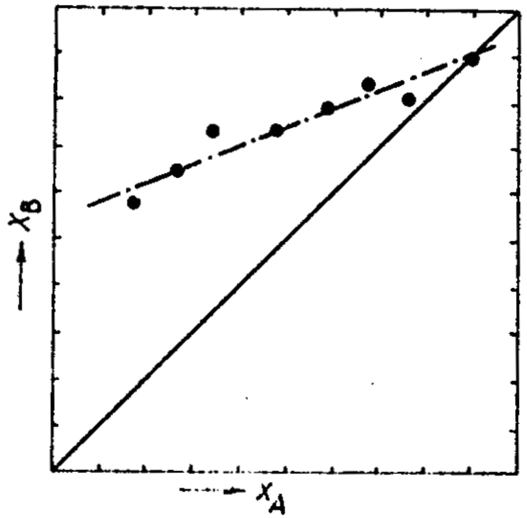
$$a = \overline{y} - b\overline{x}$$

Srovnání dvou metod

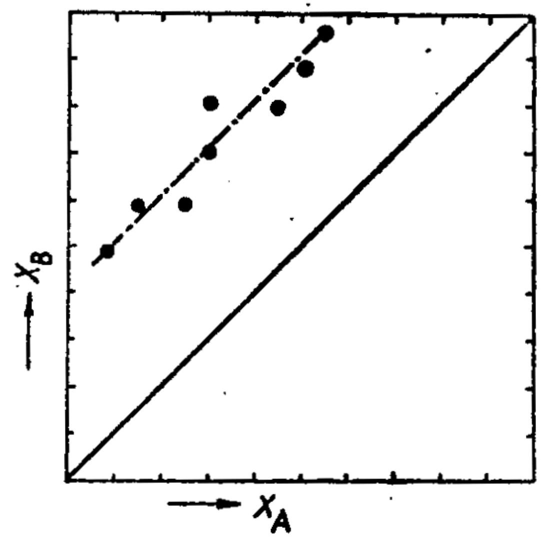
Youdenova metoda

Obr. 9a. Youdenova metoda Výsledky x_A, získané standardní metodou, vynášíme proti x_B, výsledkům získaným na stejných vzorcích testovanou metodou. Správné výsledky leží na úhlopříčce





Obr. 9b. Youdenova metoda Výsledky jsou zatíženy soustavnou proporcionální chybou



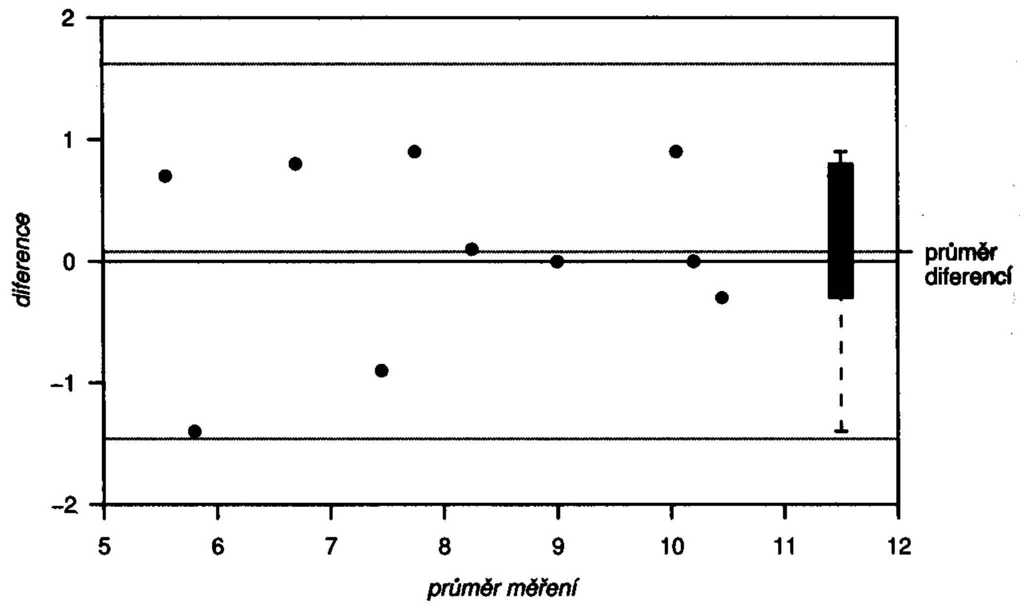
Obr. 9c. Youdenova metoda Výsledky jsou zatíženy konstantní soustavnou chybou

Srovnání dvou metod

- Bland-Altmanův graf

  Osa y: $x_{1i} - x_{2i}$

  Osa x: x_{1i}, x_{2i} nebo $(x_{1i} + x_{2i})/2$



- Poměrový graf

  Osa y: x_{1i} / x_{2i}

  Osa x: x_{1i} nebo x_{2i}

Regulační diagramy

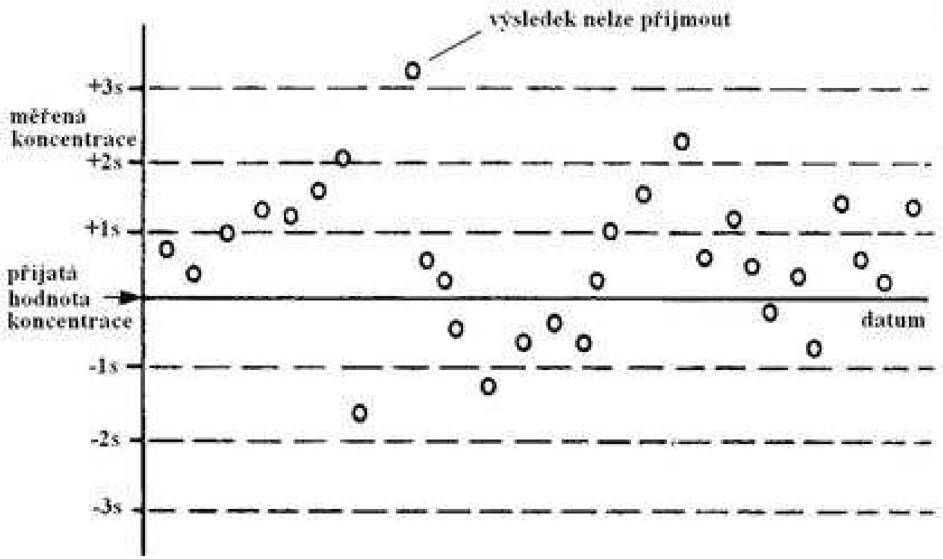
- Diagramy pro posouzení střední hodnoty (správnost): průměr, medián (Shewhart)
- Diagramy pro posouzení variability (přesnost): směr. odchylka, var. rozpětí
- Diagramy kumulativních součtů (CUSUM)
- Diagramy na bázi lokálního vyhlazování:
  pohyblivý geometrický průměr (EWMA).

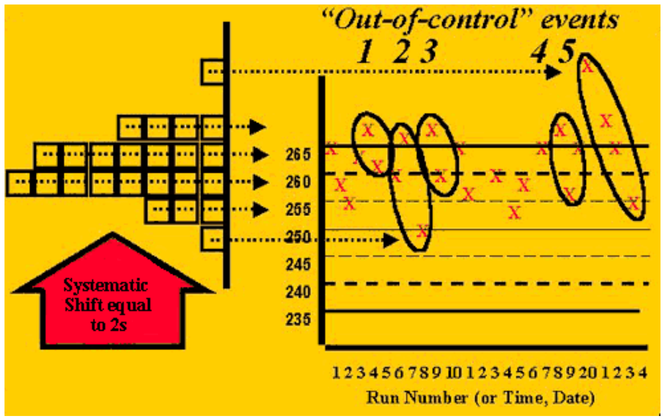
Regulační diagramy - Shewhart

Slouží k dlouhodobému sledování správnosti produkovaných výsledků

Po dlouhou dobu jsou zaznamenávány opakované výsledky analýzy jednoho parametru referenčního materiálu

Regulační diagram ukazuje, zda je celý analytický systém ve statistické rovnováze



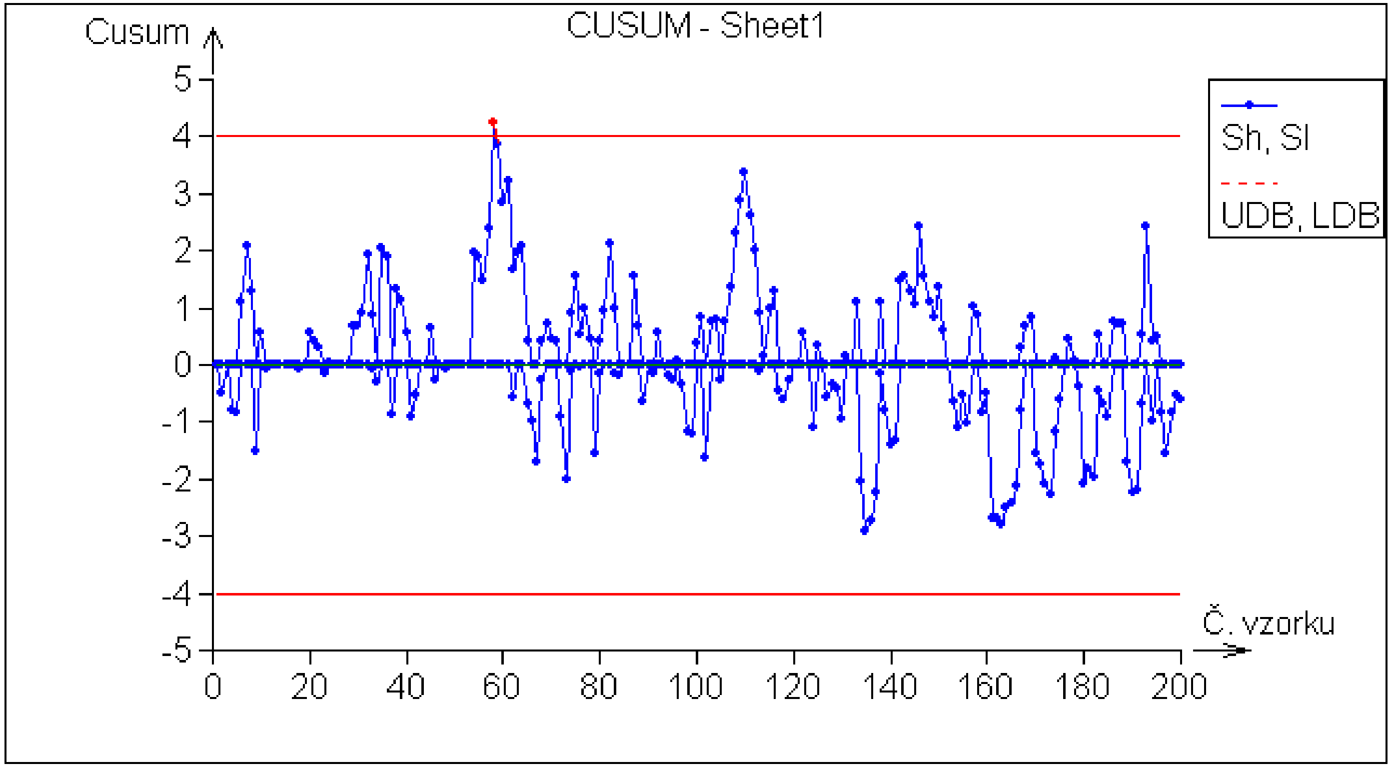


Shewhartův diagram

- **Nenormalita**: transformace dat, empirická distribuční funkce/hustota pravděpodobnosti.
- **Nestejné velikosti výběru**: úprava mezí.
- **Autokorelace**: roste počet případů překročení regulační meze, korekce – zvětšení mezí o $1 / \sqrt{1-\rho^2}$
- **Vybočující měření**: využití kvantilových charakteristik.

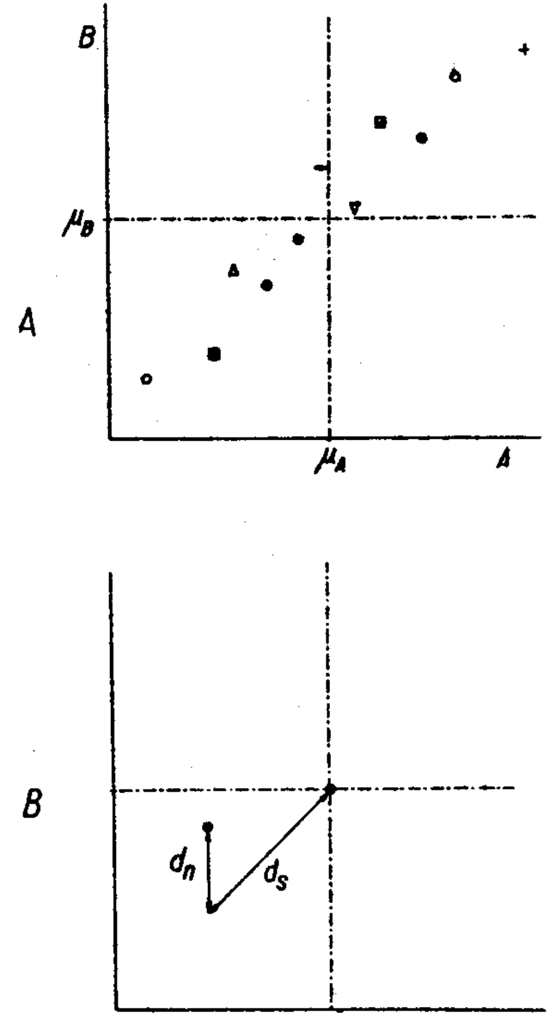
Regulační diagramy - CUSUM

- Postupně se sčítají odchylky sledovaných hodnot od střední hodnoty.
- Citlivější než Shewhart x více ovlivněn odchylkami od normality a autokorelací.



Srovnání více metod pro dva vzorky

- Youdenova metoda



Youdenova metoda
A - soustavná chyba jednotlivých laboratoří, *B* - absolutní soustavná chyba d_s a náhodná chyba d_n

Vícenásobná lineární regrese

- Závislost na několika faktorech

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + ... + \beta_n x_n$$

Interakce mezi x_1 a x_2

Logistická regrese: y = dichotomická proměnná

Vícenásobná lineární regrese

- Předpoklady:
  - Homoskedasticita (konstantní rozptyl reziduí)
  - Nezávislost reziduí (autokorelace = korelace mezi rezidui)
  - Normalita reziduí

Multikolinearita = závislost mezi proměnnými x_i

Grafy: Parciální regresní
Parciální reziduální

Nelineární regrese

- Mocninná funkce
- Exponenciální funkce
- Logaritmická funkce
- Logistická funkce

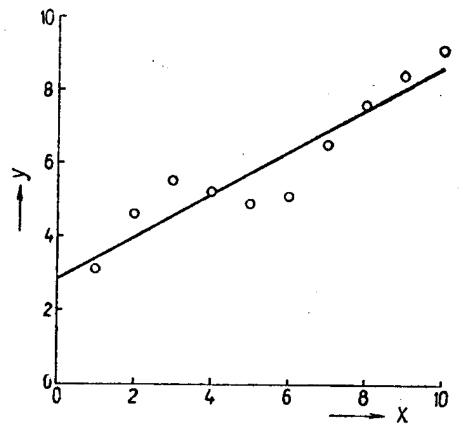
$$y = \frac{c}{1 + be^{-ax}}$$

Linearizace nelineárních závislostí

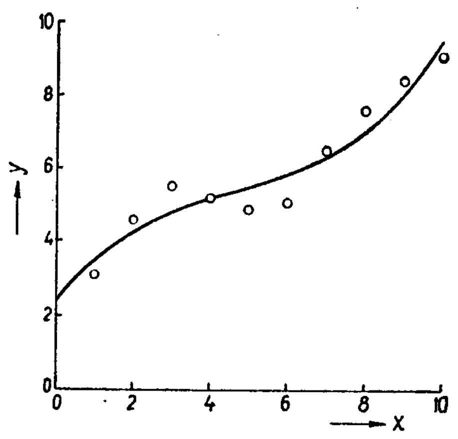
Nelineární vztah	Transformace proměnných Y	X	Transformace koeficientů $\bar{a}$	$\bar{b}$
$y = a + \frac{b}{x}$	y	$\frac{1}{x}$	a	b
$y = \frac{1}{a + bx}$ nebo $\frac{1}{y} = a + bx$	$\frac{1}{y}$	x	a	b
$y = \frac{x}{a + bx}$	$\frac{x}{y}$	x	a	b
$y = ab^x$	$\ln y$	x	$\ln a$	$\ln b$
$y = ae^{bx}$	$\ln y$	x	$\ln a$	b
$y = ax^b$	$\ln y$	$\ln x$	$\ln a$	b

Polynomická regrese

$y = k_0 + k_1x + k_2x^2 + \ldots + k_nx^n$



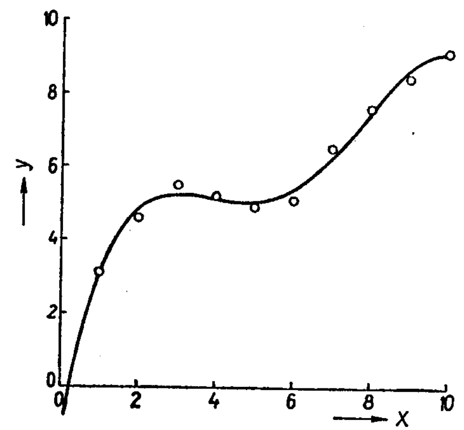
Obr. 13a. Množina deseti bodů proložena přímkou $y = k_0 + k_1x$
Při extrapolaci na $x = 0$ je $y_0 = +2{,}75$



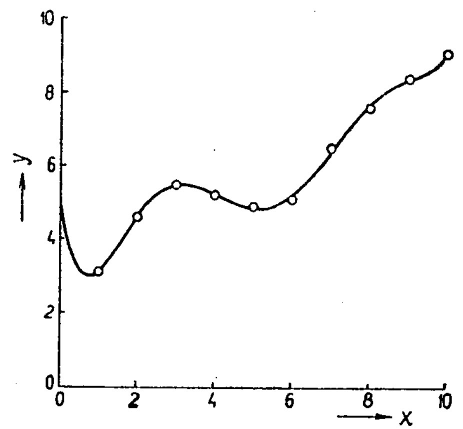
Obr. 13b. Tytéž body z obr. 13a proloženy křivkou, která odpovídá polynomu třetího stupně

$$y = k_0 + k_1x + k_2x^2 + k_3x^3$$

Při extrapolaci na $x = 0$ je $y_0 = +2{,}33$



Obr. 13c. Tytéž body z obr. 13a proloženy křivkou, která odpovídá polynomu čtvrtého stupně

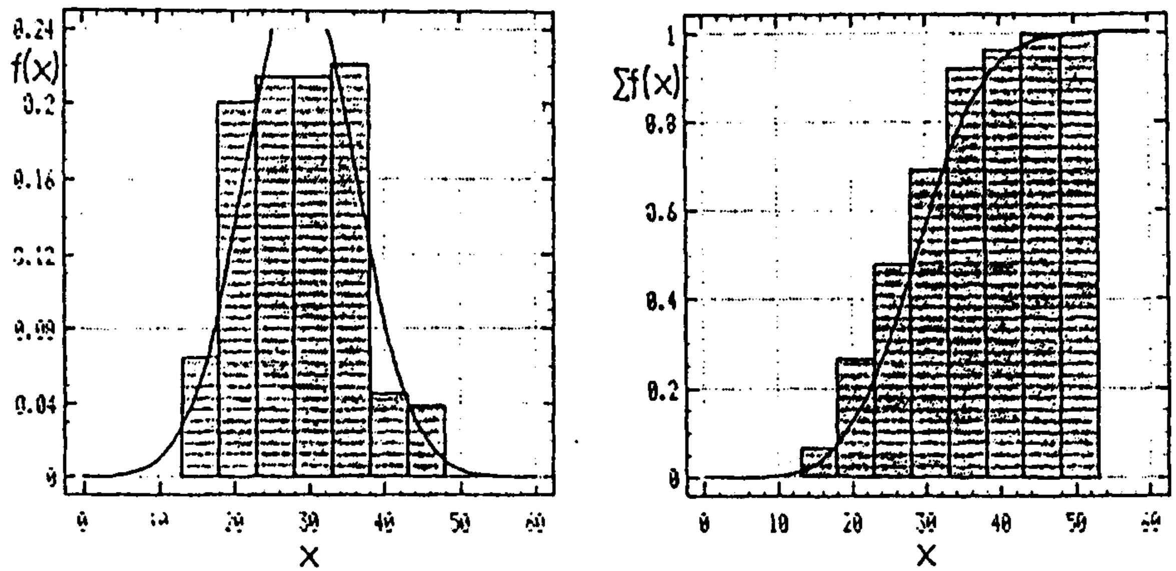


Obr. 13d. Tytéž body z obr. 13a: proloženy křivkou, která odpovídá polynomu šestého stupně

Vyrovnávání dat (v histogramu)

Funkcí rozdělení

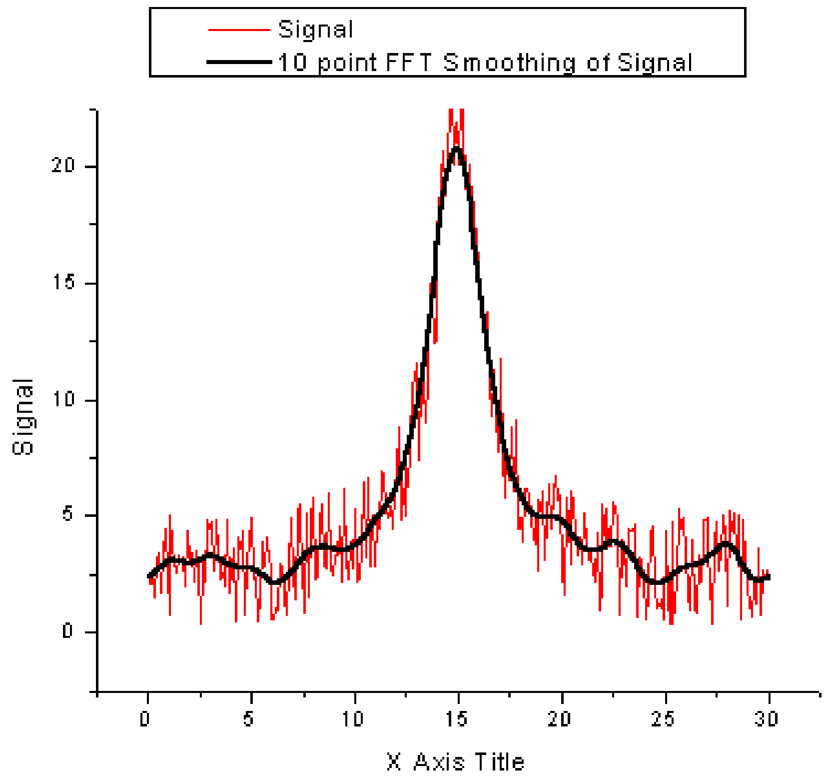
- binomické,
- Poissonovo,
- normální,
- lognormální).



Empirickou funkcí (KDE, ortogonálními polynomy, FFT, Savitzky – Golay)

Vyrovnávání časových závislostí

- Metoda klouzavých průměrů (moving averages)
- Metoda vážených klouzavých průměrů
- Metoda Savitzky - Golay
- Spliny
- Rychlá Fourierova transformace (FFT)
- Jádrové odhady



Klouzavé průměry

Hodnota závisle proměnné je nahrazena dílčím průměrem z dané hodnoty, *n* předchozích a *n* následujících hodnot.