Konstrukce lineárního regresního modelu a charakteristika přesnosti kalibrace

Milan Meloun a Jiří Militký,
Katedra analytické chemie,
Univerzita Pardubice, 532 10 Pardubice
a
Katedra textilních materiálů, Technická univerzita Liberec,
461 17 Liberec

Abstract: Postup hledání regresního (např. kalibračního) modelu obsahuje kroky:

(1) **Návrh modelu:** začíná se vždy od nejjednoduššího modelu, lineárního, u kterého vystupují jednotlivé vysvětlující proměnné v prvních mocninách.

(2) **Předběžná analýza dat:** sleduje se proměnlivost jednotlivých proměnných na rozptylových diagramech, indexových grafech. Vyšetřuje se multikolinearita, heteroskedasticita, autokorelace a vlivné body, tj. extrémy a vybočující hodnoty.

(3) **Odhadování parametrů:** odhadování parametrů modelu se provádí klasickou metodou nejmenších čtverců, stejně jako určení základních statistických charakteristik. Následuje testování významnosti jednotlivých parametrů pomocí Studentova t-testu a koeficientu determinace. Je vhodné použít i souhrnné charakteristiky regrese jako je střední kvadratická chyba predikce MEP a Akaikovo informační kritérium AIC.

(4) **Regresní diagnostika:** identifikace vlivných bodů a ověření předpokladů metody nejmenších čtverců jako je homoskedasticita, nepřítomnost vybočujících hodnot autokorelace a normalita rozdělení chyb. Na základě nalezených vlivných bodů se rozhoduje, zda je nutné tyto body z dat eliminovat. V případě více vysvětlujících proměnných se posoudí vhodnost jednotlivých proměnných s využitím parciálních regresních grafů nebo parciálních reziduálních grafů.

(5) **Konstrukce zpřesněného modelu:** Parametry zpřesněného modelu jsou odhadovány s využitím (a) metody vážených nejmenších čtverců (MVNČ) při nekonstantnosti rozptylu, (b) metody zobecněných nejmenších čtverců (MZNČ) při autokorelaci, (c) metody podmínkových nejmenších čtverců (MPNČ) při omezení kladených na parametry, (d) metody racionálních hodností u multikolinearity, (e) metody rozšířených nejmenších čtverců (MRNČ) pro případ, že všechny proměnné jsou zatížené náhodnými chybami, a konečně (f) robustních metod pro jiná rozdělení než normální a data s vybočujícími hodnotami a extrémy.

(6) **Zhodnocení kvality kalibračního modelu:** Pro daný signál y^* se vypočte hodnota x^* spolu s intervalem spolehlivosti. Před vlastním použitím kalibračního modelu je vhodné určit limitu detekce a limitu stanovení, které určují použitelnou dolní hranici kalibračního modelu a odpovídající analytické metody.

Obsah přednášky:

Regresní diagnostika

Kalibrace se v analytické chemii skládá ze dvou fází, z konstrukce kalibračního modelu y = f(**b**, x) a z inverze kalibračního modelu, kdy z naměřeného signálu y^* (např. absorbance) je určována neznámá koncentrace x^* včetně intervalu spolehlivosti.

Regresní triplet = [data, model, metoda odhadu].

Regresní diagnostika = postupy k identifikaci
1. kvality dat pro navržený model - KRITIKA DAT,
2. kvality modelu pro daná data - KRITIKA MODELU,
3. splnění předpokladů MNČ - KRITIKA METODY

Využití průzkumové analýzy dat EDA:
a) statistické zvláštnosti proměnných nebo reziduí,
b) k posouzení "párových" vztahů mezi všemi regresními proměnnými,
c) k ověření předpokladů o rozdělení proměnných nebo reziduí,
d) odhalení ***skryté proměnné***.

EDA - identifikace:
1. **Nevhodnost dat** (malé rozmezí nebo přítomnost vybočujících bodů),
2. **Nesprávnost navrženého modelu** (skryté proměnné),
3. **Multikolinearita,**
4. **Nenormalita,** když jsou proměnné náhodné veličiny.

I. KRITIKA DAT

Výskyt vlivných bodů (VB):
- zdrojem řady problémů,
- zkreslení odhadů a růst rozptylů odhadů parametrů.

Dělení vlivných bodů *dle charakteru:*
1. **Hrubé chyby**: vybočující pozorování, extrémy.
2. **Body s vysokým vlivem** (tzv. golden points): rozšiřují predikční schopnosti modelu.
3. **Zdánlivě vlivné body**: důsledek nesprávného regresního modelu.

Dělení vlivných bodů *dle výskytu:*
1. **vybočující pozorování** (outliers, O): na y se liší,
2. **extrémy** (high leverage points, E): liší se na ose x

Statistická analýza reziduí

1. Klasická rezidua: $\hat{e}_i = y_i - \mathbf{x}_i \mathbf{b}$,

Nesprávné představy o reziduích:

1. Rozdělení reziduí je stejné jako rozdělení chyb,
2. Vlastnosti reziduí jsou shodné s vlastnostmi chyb,
2. Čím je reziduum $\hat{e}_i$ větší, tím je bod vlivnější, a tím spíše by se měl z dat vyloučit. Rozepsáním

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

slovy: každé $\hat{e}_i$ je lineární kombinací všech chyb ε_i

Rozdělení reziduí je závislé:

1. Na rozdělení chyb,
2. Na prvcích projekční matice **H**,
3. Na velikosti výběru n.

Vlastnosti klasických reziduí:

a) Rozptyl reziduí $\quad D(\hat{e}_i) = (1 - H_{ii})\, \hat{\sigma}^2$

je *nekonstantní,* i když rozptyl chyb $\hat{\sigma}^2$ je konstantní.

b) Rezidua jsou korelovaná: existuje *párový korelační koeficient* r_{ij} mezi dvěma rezidui e_i a e_j

$$r_{ij} = \frac{- H_{ij}}{\sqrt{(1 - H_{ii})(1 - H_{jj})}}$$

i když chyby ε_i a ε_j jsou nezávislé.

c) Rezidua *neindikují* silně extrémní hodnoty.

d) Rezidua jsou normálnější než chyby: *(effekt supernormality)*

e) *U malých výběrů* nemusí správně indikovat model.

2. Normovaná rezidua $\hat{e}_{Ni} = \hat{e}_i / \hat{\sigma}$

normálně rozdělené veličiny $\hat{e}_{Ni} \sim N(0, 1)$.

Diagnostika: pravidlo 3σ, tj. rezidua větší než $\pm 3\hat{\sigma}$ indikují vybočující.

3. Standardizovaná rezidua $\hat{e}_{Si} = \dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - H_{ii}}}$

Mají konstantní rozptyl. Maximální hodnota $\hat{e}_{Si}$ je ohraničena velikostí $\sqrt{n - m}$.

Diagnostika: k indikaci heteroskedasticity.

4. Jackknife rezidua ("plně studentizovaná") užijeme místo $\hat{\sigma}$ odhadu směrodatné odchylky $\hat{\sigma}_{(-i)}$,

$$\hat{e}_{Ji} = \hat{e}_{Si} \sqrt{\frac{n - m - 1}{n - m - \hat{e}_{Si}^2}} = \sqrt{n - m}\ \mathrm{cotg}\ \Theta_i$$

mají Student. rozdělení s (n - m - 1) stupni volnosti.
Diagnostika: k identifikaci vybočujících bodů (outliers).

5. Predikovaná rezidua

$$\hat{e}_{Pi} = y_i - \mathbf{x}_i \mathbf{b}_{(i)} = \frac{\hat{e}_i}{1 - H_{ii}}$$

kde $\mathbf{b}_{(i)}$ jsou MNČ odhady ze všech bodů kromě i-tého.
Diagnostika: indikace vybočujících hodnot (outliers).

6. Rekurzívní rezidua dopředná rekurzívní rezidua jsou definována vztahy

$$\hat{e}_{Ri} = 0, \qquad i = 1, \ldots, m$$

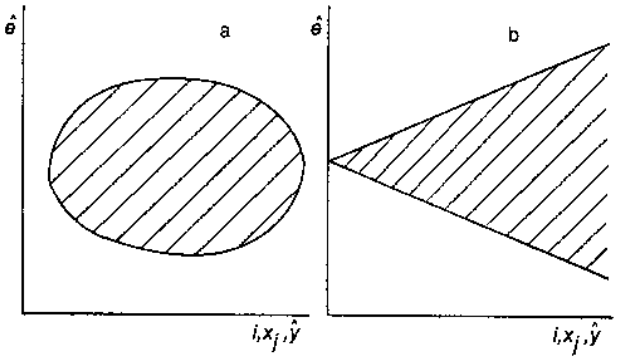
$$\hat{e}_{Ri} = \frac{y_i - \mathbf{x}_i \mathbf{b}_{i-1}}{\sqrt{1 + \mathbf{x}_i (\mathbf{X}_{i-1}^T \mathbf{X}_{i-1})^{-1} \mathbf{x}_i^T}} \qquad i = m + 1, \ldots, n$$

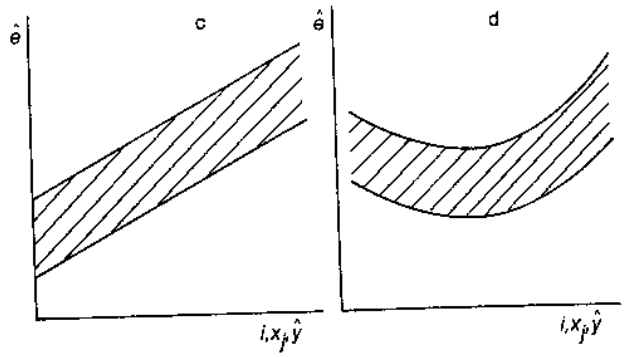
kde $\mathbf{b}_{i-1}$ jsou odhady získané z prvních (i - 1) bodů.
Diagnostika: umožňují identifikovat autokorelaci, nestabilitu modelu, např. v čase.

Rezidua k diagnostice

1. K detekci heteroskedasticity: standardizovaná rezidua $\hat{e}_{Si}$.
2. K detekci vybočujících bodů: Jackknife rezidua $\hat{e}_{Ji}$, predikovaná rezidua $\hat{e}_{Pi}$.
3. K detekci autokorelace: rekurzivní rezidua $\hat{e}_{Ri}$.

Obrazce v diagnostických grafech

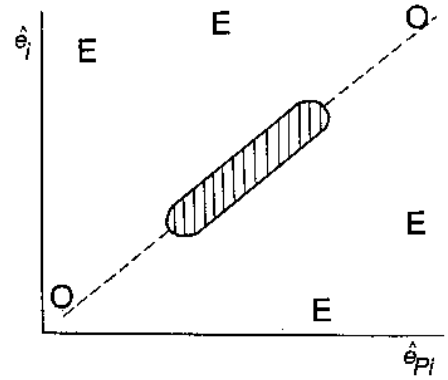




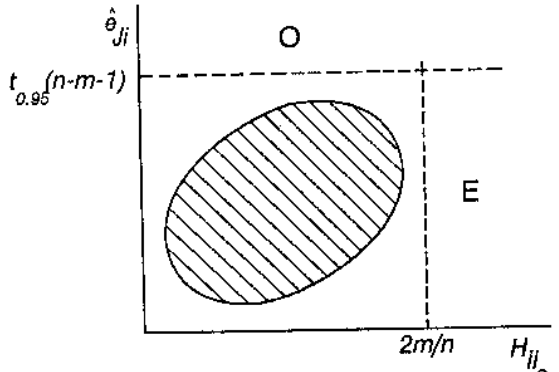
(a) tvar mraku, (b) tvar výseče, (c) tvar pásu a (d) nelineární tvar

Grafy identifikace vlivných bodů

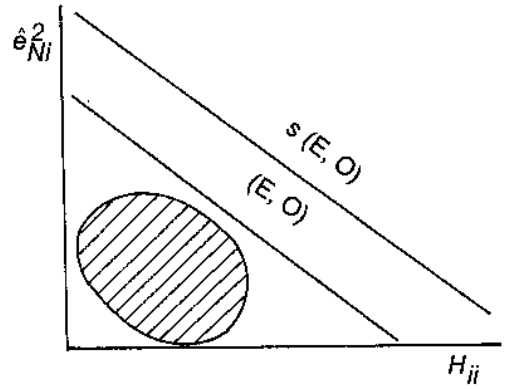
1. Graf predikovaných reziduí (GPR), osa x: $\hat{e}_{Pi}$, osa y: $\hat{e}_i$



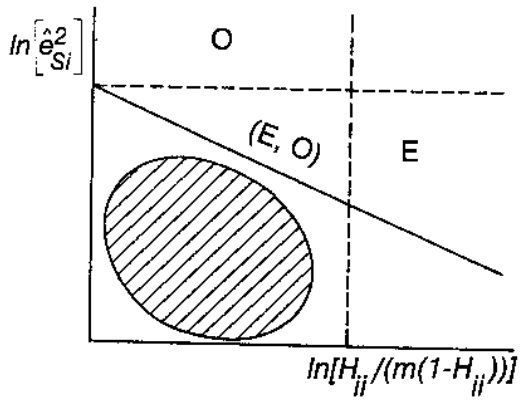
2. Williamsův graf (WG), osa x: prvky H_{ii}, osa y: $\hat{e}_{Ji}$



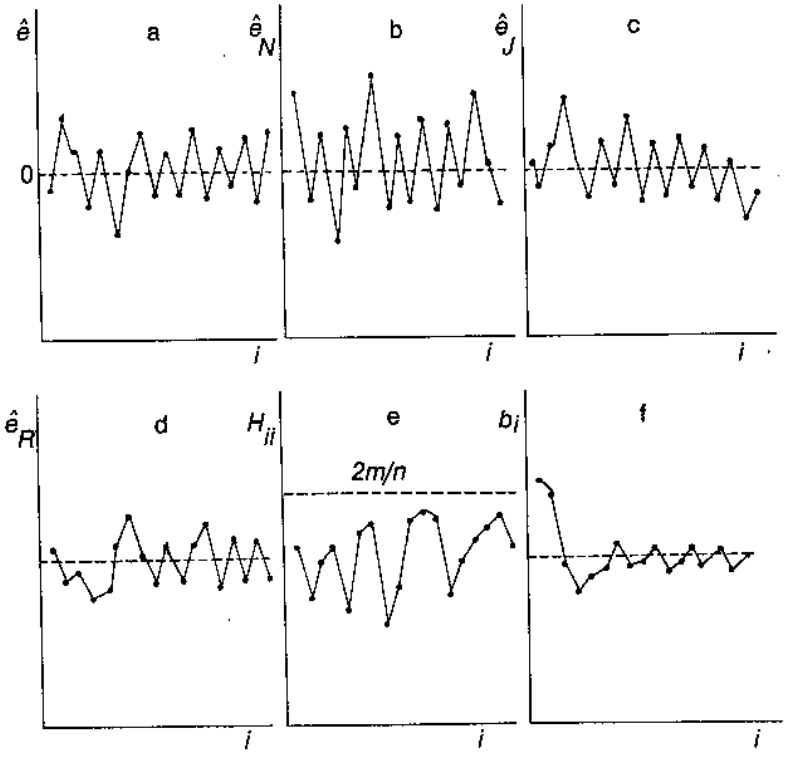
3. Pregibonův graf (PG), osa x: prvky H_{ii}, osa y: $\hat{e}^2_{Ni}$



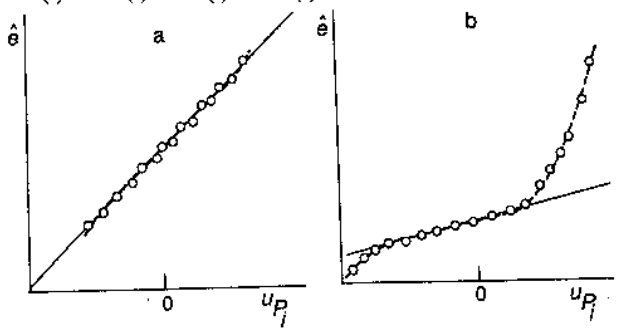
4. McCullohův-Meeterův graf (MMG), osa x: ln $[H_{ii} /(m (1 - H_{ii}))]$, osa y: ln $\hat{e}^2_{Si}$

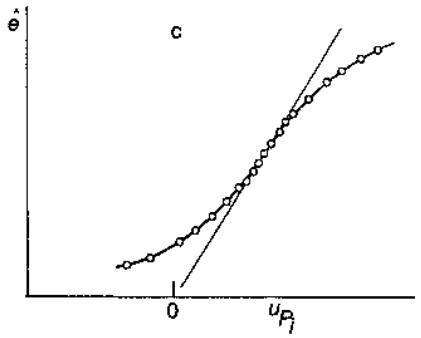


5. Indexové grafy (IG), osa x: index i,
osa y: rezidua $\hat{e}_i$, $\hat{e}_{Si}$, $\hat{e}_{Ni}$, $\hat{e}_{Pi}$, $\hat{e}_{Ji}$, $\hat{e}_{Ri}$, nebo prvky H_{ii} či H^*_{ii}, a odhady $\mathbf{b}_i$



6. Rankitové grafy (Q-Q), osa x: u_{P_i} pro $P_i = i / (n + 1)$,
osa y: $\hat{e}_{(i)}$, $\hat{e}_{S(i)}$, $\hat{e}_{N(i)}$, $\hat{e}_{P(i)}$, $\hat{e}_{J(i)}$, $\hat{e}_{R(i)}$





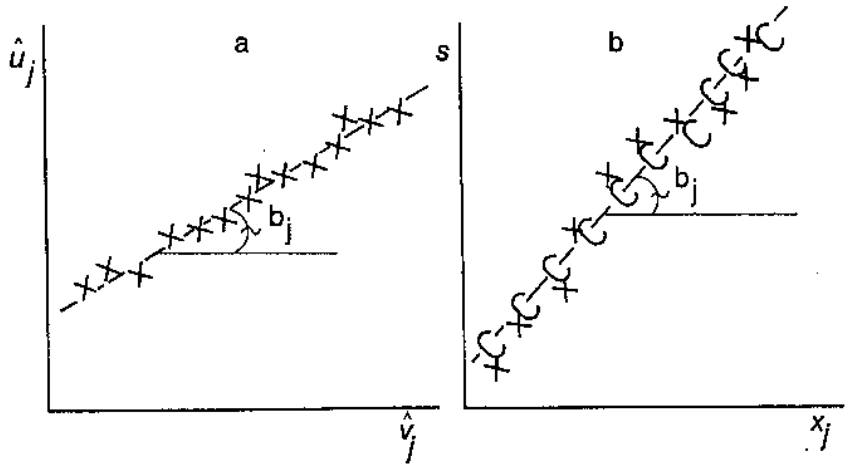
(a) přibližně normální rozdělení, (b) rozdělení s dlouhými konci, a (c) rozdělení s krátkými konci

II. Kritika modelu

1. *Jediná proměnná x:* rozptylový graf y na x.
   Více proměnných x: rozptylové grafy mohou ***mylně indikovat*** nelinearitu.

2. *K posouzení vztahu y a x_j:* parciální regresní grafy a parciální reziduální grafy.

Belseyho parciální regresní grafy (partial regression leverage plots)



umožňují:

1. Posouzení kvality navrženého regresního modelu,
2. Indikují přítomnost vlivných bodů,
3. Nesplnění předpokladů klasické MNČ,
4. Vyjadřují závislost mezi **y** a zvolenou proměnnou $\mathbf{x}_j$ při statisticky neměnném vlivu ostatních $\mathbf{X}_{(j)}$.

Vychází se z regresního modelu

$$\mathbf{y} = \mathbf{X}_{(j)} \boldsymbol{\beta}^* + \mathbf{x}_j c + \boldsymbol{\varepsilon}$$

kde $\boldsymbol{\beta}^*$ má rozměr $(m - 1) \times 1$, c je regresní parametr příslušející j-té proměnné.

Projekcí do prostoru kolmého na prostor sloupců matice $\mathbf{X}_{(j)}$ bude

$\mathbf{P}_{(j)}\,\mathbf{y}$	=	$\mathbf{P}_{(j)}\,\mathbf{X}_{(j)}$	+	$\mathbf{P}_{(j)}\,\mathbf{x}_j\,c$	+	$\mathbf{P}_{(j)}\,\boldsymbol{\varepsilon}$
$\hat{\mathbf{u}}_j$		$= 0$		$\hat{\mathbf{v}}_j$		
vektor reziduí proměnné $\mathbf{x}_j$ na proměnných, které tvoří sloupce $\mathbf{X}_{(j)}$				vektor reziduí proměnné $\mathbf{y}$ na proměnných, které tvoří sloupce $\mathbf{X}_{(j)}$		

Střední hodnota $E(\hat{\mathbf{u}}_j)$ je $E(\hat{\mathbf{u}}_j) = c\,E(\hat{\mathbf{v}}_j)$ a závislost $\hat{\mathbf{u}}_j$ na $\hat{\mathbf{v}}_j$ tvoří *parciální regresní graf*, kde u správného modelu jde o závislost s nulovým úsekem.

Odhad směrnice z MNČ bude

$$\hat{c} = \frac{\hat{\mathbf{u}}_j^{\,T}\,\hat{\mathbf{u}}_j}{\hat{\mathbf{v}}_j^{\,T}\,\hat{\mathbf{v}}_j} = \frac{\mathbf{x}_j^{\,T}\,\mathbf{P}_{(j)}\,\mathbf{y}}{\mathbf{x}_j^{\,T}\,\mathbf{P}_{(j)}\,\mathbf{x}_j}$$

Platí $\hat{\mathbf{e}} = \hat{\mathbf{u}}_j - \hat{\mathbf{v}}_j\,\hat{c}$ a ukazuje na vztah mezi $\hat{\mathbf{e}}$ s parciálními rezidui $\hat{\mathbf{u}}_j$ a $\hat{\mathbf{v}}_j$.

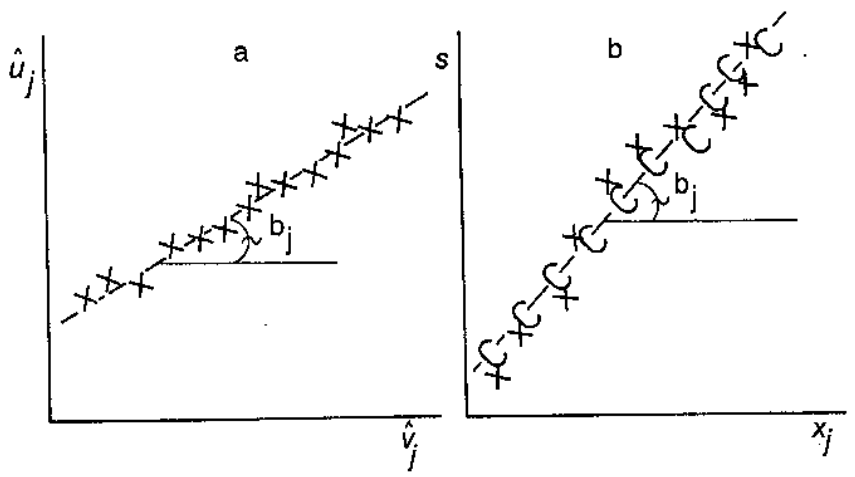
Vlastnosti:

a) směrnice $\hat{c}$ je stejná s b_j v neděleném modelu a úsek je roven nule, když je navržený model správný,
b) korelační koeficient mezi $\hat{\mathbf{u}}_j$ a $\hat{\mathbf{v}}_j$ odpovídá parciálnímu korelačnímu koeficientu $\hat{R}_{yx_j(x)}$,
c) rezidua regresní přímky v parciálním regresním grafu jsou shodná s rezidui $\hat{e}_i$ pro nedělený model,
d) v grafu jsou zvýrazněny vlivné body a některá porušení předpokladů MNČ (heteroskedasticita).

Nevýhody:

a) x-ové souřadnice $\hat{\mathbf{v}}_j$ v parciálních regresních grafech nejsou v původním měřítku proměnné $\mathbf{x}_j$,
b) jsou-li proměnné sloupců matice $\mathbf{X}$ silně korelované, nemusí parciální regresní grafy indikovat správně nelinearitu, a tím i vhodnost navrženého modelu

Parciální reziduální grafy (grafy "komponenta + reziduum")



Rovnici $\hat{\mathbf{e}} = \hat{\mathbf{u}}_j - \hat{\mathbf{v}}_j . b_j$ přepíšeme do tvaru

$$\hat{\mathbf{u}}_j = \hat{\mathbf{e}} + b_j (\mathbf{E} - \mathbf{H}_{(j)}) \mathbf{x}_j$$

a vyjádříme jako závislost veličiny

$$\hat{\mathbf{e}} + b_j (\mathbf{E} - \mathbf{H}_{(j)}) \mathbf{x}_j \qquad \text{na} \qquad (\mathbf{E} - \mathbf{H}_{(j)}) \mathbf{x}_j$$

Parciální reziduální graf je analogií parc. regr. grafu při volbě $\mathbf{H}_{(j)} = \mathbf{0}$.
Jedná se o závislost parciálních reziduí **s** na proměnné $\mathbf{x}_j$.

$$\mathbf{s} = \hat{\mathbf{e}} + \mathbf{b}\, \mathbf{x}_j = \mathbf{y} - \sum_{k \neq j}^{m} \mathbf{x}_k\, b_k$$

Pokud model obsahuje absolutní člen, je možno použít modifikovaná parciální rezidua

$$s_i^* = \hat{e}_i + (x_{ij} - \bar{x}_j) b_j + \bar{y}$$

kde $\bar{x}_j$, $\bar{y}$ jsou aritmetické průměry veličin x_j a y. V grafu se znázorňuje deterministická komponenta

$$c_{ij} = (x_{ij} - \bar{x}_j) b_j \qquad i = 1, ..., n$$

která se značí písmenem "C" a parciální reziduum $s_i = c_{ij} + \hat{e}_j$, i = 1, ..., n, které se označuje křížkem "+".

Vlastnosti:

a) směrnice závislosti **s** na $\mathbf{x}_j$ je rovna $\mathbf{b}_j$ a úsek je nulový. Lineární závislost ukazuje na vhodnost navržené $\mathbf{x}_j$ v modelu;
b) rezidua přímky jsou přímo rezidua $\hat{e}_i$ pro nedělený model;
c) pokud je úhel mezi $\mathbf{x}_j$ a některými sloupci matice $\mathbf{X}_{(j)}$ malý (multikolinearita), ukazuje parciální reziduální graf nesprávně malý rozptyl kolem regresní přímky $b_j\, \mathbf{x}_j$ a dochází i k potlačení efektu vlivných bodů.

Znaménkový test vhodnosti modelu

nenáhodnost reziduí lze testovat znaménkovým testem.

Postup:

1. Určuje se počet sekvencí n_U, kde mají rezidua stejná znaménka.
   (např. pro rezidua -1, -1, 1, -1, 1, 2, 1 je počet sekvencí roven $\hat{n}_U = 4$)
2. Stanoví se počet reziduí kladných (n_+) a záporných (n_-).
3. Teoretický počet sekvencí n_t a jeho rozptyl D_t je

$$n_t = 1 + \frac{2\, n_+\, n_-}{n_+ + n_-} \approx 1 + \frac{n}{2}$$

$$D_t = \frac{2\, n_+\, n_-\, (2\, n_+\, n_- - n_+ - n_-)}{(n_+ + n_-)^2\, (n_+ + n_- - 1)} \approx \frac{n}{4}$$

Pro $n_+ > 10$ a $n_- > 10$ lze využít náhodné veličiny

$$U = \frac{n_u - n_t + C}{\sqrt{D_t}}$$

s normovaným normálním rozdělením N(0, 1).

Testování: Je-li $n_U < (n_t - C\ 1.96\ \sqrt{D_t})$, pak je v reziduích trend a model je nesprávně navržen. (Konstanta C = 0.5 je korekcí na spojitost).

III. Kritika metody - ověření předpokladů MNČ

1. Hetero/homoskedasticita (nekonstantnost rozptylu)

Rozptyl veličiny y_i v itém bodě je popsán $\sigma_i^2 = \sigma^2 \exp(\lambda\, \mathbf{x}_i\, \boldsymbol{\beta})$, kde $\mathbf{x}_i$ je i-tý řádek matice **X**.

Test: ověření H_0: $\lambda = 0$ (homoskedasticita). Cook-Weisbergovo testační kritérium

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

kde $$\hat{y}_P = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$$

Testování: 1. Je-li $S_f < \chi^2(1)$, H_0 (heteroskedasticita) je přijata.

2. Pro homoskedasticitu tvoří diagnostický graf $\hat{e}_{Si}^2$ na $(1 - H_{ii})\ \hat{y}_i$ náhodný mrak bodů. Pro heteroskedasticitu vznikne typický klínový obrazec.

2. Autokorelace: Data časových řad mají chyby ε_i vzájemně korelované. Nejčastější je případ autokorelace prvního řádu

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + u_i$$

kde $u_i \sim N(0, \sigma^2)$.

a) Pro $\rho_1 = 1$ případ kumulativních chyb, který se v chemii vyskytuje často.

b) Pro $\rho_1 \leq 1$ jde o autokorelační koeficient 1. řádu.

Test:

1. Grafická indikace autokorelace:
2. Waldův test pro ρ_1: H_0: $\rho_1 = 0$ *vs.* H_A: $\rho_1 \neq 0$. Je-li Waldovo kritérium

$$W_a = \frac{n \hat{\rho}_1^2}{1 - \hat{\rho}_1^2} < \chi^2(1), \; H_0 \text{ je přijata.}$$

3. Normalita chyb

1. ***Rankitový (Q-Q) graf:*** $\hat{e}_{(i)}$ nebo $\hat{e}_{Ri}$ na u_{Pi} pro $P_i = i / (n + 1)$
2. ***Test normality:*** H_0: normalita *vs.* H_A: nenormalita. Užívá se Jarque-Berrova testační statistika

$$L(\hat{e}) = n \left[\frac{\hat{u}_3^2}{6 \hat{u}_2^3} + \frac{\frac{\hat{u}_4}{\hat{u}_2^2} - 3}{24} \right] + n \left[\frac{3 \hat{u}_1^2}{2 \hat{u}_2} - \frac{\hat{u}_3 \hat{u}_1}{\hat{u}_2^2} \right]$$

kde $\hat{u}_j$ je j-tý výběrový moment reziduí

$$\hat{u}_j = \frac{\sum_{i=1}^{n} \hat{e}_i^{\,j}}{n}$$

Test: je-li $L(\hat{e}) > \chi^2_{1-\alpha}(2) = 5.99$, je H_0 (normalita) zamítnuta. Např. pro lineární modely s absolutním členem je $\hat{u}_1 = 0$ a $L(\hat{e})$ je

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

kde $\hat{g}_1 = \dfrac{\hat{u}_3^2}{\hat{u}_2^3}$ $\qquad$ $\hat{g}_2 = \dfrac{\hat{u}_4}{\hat{u}_2^2}$

(test není vhodný pro malé výběry)

Kalibrační modely

Složitost řešení úlohy kalibrace souvisí zejména s užitým kalibračním modelem. Pro lineární regresní modely lze vyjádřit konfidenční pásy kolem modelu. Složky vektoru **x** jsou funkcemi měřené vlastnosti (obvykle koncentrace). Obyčejně se uvažují polynomické modely, u kterých jednotlivé složky odpovídají mocninám měřené vlastnosti. Při hledání hodnoty $\hat{x}^*$ se pak řeší úloha hledání kořene polynomu. Pro nelineární regresní modely se hledá řešení ve tvaru

$$\hat{x}^* = f^{-1}(y^*)$$

Na základě Taylorova rozvoje této funkce lze nalézt přibližnou formuli pro rozptyl $D(\hat{x}^*)$ ve tvaru

$$D(\hat{x}^*) \approx \left[\frac{\delta f(x, \mathbf{b})}{\delta x}\right]^{-2} \left[\frac{D(y^*)}{M} + D(f(x, \mathbf{b}))\right]$$

kde $D(y^*)$ je rozptyl y^*-ových hodnot, který je obyčejně roven σ^2, a $D(f(x,\mathbf{b})) = D(\hat{y})$ je rozptyl predikce, který se určuje také z Taylorova rozvoje funkce $f(x, \mathbf{b})$. Pro lineární regresní modely je rozptyl predikce roven

$$D(\hat{y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{(y^* - \bar{y})^2}{b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}\right]$$

kde b_1 představuje odhad směrnice regresní přímky. Po dosazení dostaneme

$$D(\hat{x}^*) \approx \frac{\sigma^2}{b_1^2} \left[\frac{1}{M} + \frac{1}{n} + \frac{(y^* - \bar{y})^2}{b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}\right]$$

Problémem je, že rozdělení veličiny $\hat{x}^*$ je obecně nesymetrické. Jedině pro případ kalibrační přímky a malý reziduální rozptyl lze rozdělení veličiny $\hat{x}^*$ považovat za přibližně normální. Za předpokladu, že jak y-ové, tak i y^*-ové hodnoty jsou náhodné proměnné s normálním rozdělením, platí, že rozdíl $\Delta = \bar{y}^* - f(x^*, \mathbf{b})$ bude mít také normální rozdělení. Standardizovaná náhodná veličina $\Delta / \sqrt{D(\Delta)}$ má pak Studentovo rozdělení se stupni volnosti, které byly užity při určení $D(\Delta)$.

Kalibrační přímka

Model kalibrační přímky patří v laboratořích k nejpoužívanějším. Předpokládá se obvykle, že tento model vyhovuje v celém sledovaném rozsahu proměnných x a y. Z hlediska statistického zpracování lze užít rovnice

$$y_i = \beta_2 + \beta_1 x + \varepsilon_i, \qquad i = 1, \ldots, n$$

$$y_j^* = \beta_2 + \beta_1 \varkappa + \varepsilon_j^*, \qquad j = 1, \ldots, M$$

Úlohou kalibrace je potom nalezení odhadu $\hat{x}^*$ parametru $\varkappa$ jako primárního a odhadů parametrů β_1, β_2 jako doplňkových. Odhad $\hat{x}^*$ a jeho odpovídající konfidenční interval je možné určit několika způsoby:

(1) Dosazením dostaneme **přímý odhad** parametru $\varkappa$ ve tvaru

$$\hat{x}^* = \bar{x} + \frac{y^* - \bar{y}}{b_1}$$

kde y^* je měřená hodnota signálu (resp. průměr $\bar{y}^*$ pro $M > 1$ opakovaných měření) a b_1 je odhad směrnice kalibrační přímky. Tento odhad je obecně vychýlený.

(2) Korekci na vychýlení lze provést pomocí **Naszodiho modifikovaného odhadu**

$$\hat{x}_B^* = \bar{x} + \frac{(y^* - \bar{y})\, b_1}{b_1^2 + \dfrac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

(3) Krutchhoft navrhl **inverzní odhad**

$$\hat{x}_I^* = \bar{x} + (y^* - \bar{y}) \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

který vychází z inverzního regresního modelu $E(x/y) = \alpha_1 (y - \bar{y}) + \alpha_2$. Na základě rozboru odhadu $\hat{x}_I^*$ bylo zjištěno, že jde také o vychýlený odhad, který není lepší než přímý odhad $\hat{x}^*$. Navíc se při odhadování parametrů α_1 a α_2 chybně předpokládá, že y-ové hodnoty jsou měřeny se zanedbatelnými chybami vůči x-ovým hodnotám.

(4) V práci Schwartze je navržen **nelineární odhad** typu

$$\hat{x}_N^* = \frac{\sum_{i=1}^{n} x_i \exp\left[\dfrac{-(y^* - b_2 - b_1 x_i)^2}{2\,\hat{\sigma}^2}\right]}{\sum_{i=1}^{n} \exp\left[\dfrac{-(y^* - b_2 - b_1 x_1)^2}{2\,\hat{\sigma}^2}\right]}$$

který je však založen na předpokladu normality reziduí.

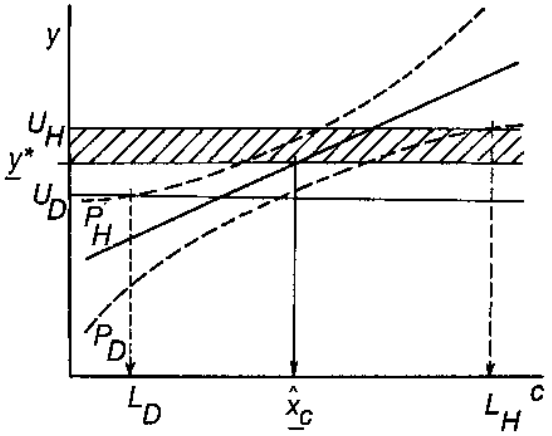
Při konstrukci intervalů spolehlivosti pro odhady $\hat{x}^*$ nebo $\hat{x}_B^*$ u silněji rozptýlených dat je nejjednodušší užít $D(\hat{x}^*)$ a předpokladu asymptotické normality. Meze 95%ního intervalu spolehlivosti se pak vypočtou ze vztahů

$$L_D = \hat{x}^* - 1.96 \sqrt{D(\hat{x}^*)}$$

a

$$L_H = \hat{x}^* + 1.96 \sqrt{D(\hat{x}^*)}$$

Je patrné, že v případě opakování měření signálu y a určení hodnoty $\bar{y}^*$ je třeba stanovit konfidenční přímky U_D a U_H, a řešit úlohu hledání průsečíku U_H s dolní konfidenční parabolou P_D kalibrační přímky (výsledkem je bod L_H), resp. průsečíku přímky U_D s horní konfidenční parabolou P_H kalibrační přímky (výsledkem je bod L_D).



Určení intervalu spolehlivosti veličiny x kalibrační přímky. Šrafovaně je vyznačena poloviční šíře intervalu spolehlivosti signálu

Při znalosti rozptylu měření σ^2 lze snadno definovat $100(1 - \alpha)\%$ní interval spolehlivosti pro signál y^* ve tvaru

$$U_{D,H} = \bar{y}^* \mp u_{1-\alpha/2}\,\sigma$$

kde $u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení. Pokud σ^2 není známo, lze

využít nerovnosti $\sigma^2 < \frac{(n-2)\,\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-2)\,M}$, kde $\chi^2_{\alpha/2}$ je dolní kvantil χ^2 rozdělení.

Interval spolehlivosti signálu $U_{D,H}$ se potom vypočte ze vztahu

$$U_{D,H} = \bar{y}^* \mp u_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{M}} \sqrt{\frac{n-2}{\chi^2_{\alpha/2}(n-2)}}$$

Místo kvantilu $u_{1-\alpha/2}$ se v této rovnici pro $M = 1$ užívá kvantil Studentova rozdělení $t_{1-\alpha/2}(n-2)$ a rozptyl σ^2 je nahrazen odhadem $\hat{\sigma}^2$. Pro celou regresní přímku se vypočtou hraniční $100(1-\alpha)\%$ní paraboloidy

$$P_{D,H} = b_1 x + b_2 \mp \hat{\sigma} \left\{ 2 F_{1-\alpha}(2, n-2) \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \right\}^{1/2}$$

Hraniční hodnota L_H je řešením rovnice $U_H = P_D$ vzhledem k proměnné x. Hraniční hodnota L_D je řešením rovnice opět vzhledem k proměnné x $U_D = P_H$. Obě rovnice jsou vzhledem k proměnné x kvadratické. Kvalita intervalu spolehlivosti kolem parametru je příznivě ovlivněna:

1. Opakováním měření signálu y^*, čili růstem M.
2. Zúžení konfidenčních parabol lze dosáhnout eliminací vlivných bodů.
3. Zmenšením reziduálního rozptylu $\hat{\sigma}^2$, a tedy buď zpřesněním měření, nebo užitím správného kalibračního modelu.

Přesnost kalibrace

K vyjádření přesnosti kalibračních metod se obyčejně definují limitní hodnoty, které souvisejí s úrovní koncentrace, pro kterou je signál ještě statisticky významně odlišný od šumu. V souvislosti s vyjádřením přesnosti a citlivosti kalibračních metod se definují tři specifické úrovně signálu:

1. **Kritická úroveň** y_c představuje horní mez $100(1-\alpha)\%$ního intervalu spolehlivosti predikce signálu z kalibračního modelu pro koncentraci rovnou nule, tzv. *slepý pokus*. Náhradou $\sqrt{2\,F_{1-\alpha}(2, n-2)}$ kvantilem $t_{1-\alpha/2}(n-2)$ a dosazením $x = 0$ dostaneme pro kritickou úroveň y_c vztah

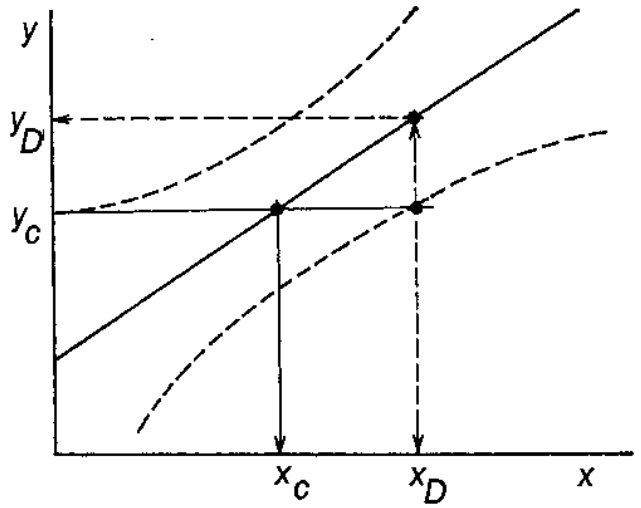
$$y_c = \bar{y} - b_1 \bar{x} + t_{1-\alpha/2}(n-2)\,\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Nad hodnotou y_c lze signál odlišit od šumu. Koncentrace x_c, odpovídající hodnotě kritické úrovně, se určí z kalibračního modelu pomocí vztahu

$$x_c = \frac{y_c - \bar{y}}{b_1} + \bar{x}$$

2. **Limita detekce** y_D odpovídá hodnotě koncentrace, pro kterou je dolní mez $100(1 - \alpha)\%$ního intervalu spolehlivosti predikce signálu z kalibračního modelu rovna y_c. Pro lineární kalibrační model lze psát

$$y_D = y_c + \hat{\sigma}\, t_{1-\alpha/2}(n - 2) \sqrt{1 + \frac{1}{n} + \frac{(x_D - \bar{x}^2)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$



Definice kritické úrovně y_c, limity detekce y_D a jim odpovídající koncentrace x_c a x_D

Odpovídající koncentrace x_D se pak vypočte podle vztahu

$$x_D = \frac{y_D - \bar{y}}{b_1} + \bar{x}$$

Limita detekce udává skutečnou úroveň signálu, která umožňuje ještě *detekci koncentrace*. Velikost x_D udává minimální koncentraci, kterou lze ještě s pravděpodobností $(1 - \alpha)$ odlišit od nulové hodnoty.

3. **Limita stanovení** y_s je nejmenší hodnota signálu, pro kterou je relativní směrodatná odchylka predikce z kalibračního modelu dostatečně malá a rovna číslu C. Pro číslo C se volí obyčejně velikost C = 0.1. Označme predikci v místě x_s výrazem $y(x_s) = \bar{y} + b_1(x_s - \bar{x})$. Podmínka pro určení y_s je pak rovna

$$\frac{\sqrt{D(y(x_s))}}{\hat{y}(x_s)} = C$$

Dosazením a úpravami se určí $y_s = \frac{\hat{\sigma}}{C}\sqrt{1 + \frac{1}{n} + \frac{(x_s - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

K praktickým výpočtům v laboratoři se však často užívá aproximace

$$y_s \approx \frac{\hat{\sigma}}{C}\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Odpovídající koncentrace x_s je pak rovna $x_s = \frac{y_s - \bar{y}}{b_1} + \bar{x}$. Z uvedených čtyř charakteristik lze snadno konstruovat limitu detekce y_D a limitu stanovení y_s i pro nelineární kalibrační modely a pro případy dat, kdy rozptyly měření nejsou konstantní Obecně platí, že $y_c \leq y_D \leq y_s$.

Doporučená literatura a software:

(1) Milan Meloun a Jiří Militký: ***Statistické zpracování experimentálních dat,*** Plus Praha 1994.

(2) Milan Meloun a Jiří Militký: ***Statistické zpracování experimentálních dat - Sbírka úloh***, Univerzita Pardubice 1996.

(3) **ADSTAT 1.25** a verze **2.0**, TriloByte Statistical Software Pardubice, 1992, 1993.