Těsnost proložení

Statistická analýza reziduí

Pro aditivní modely měření:

$$\hat{e}_i = y_i - f(x_i, \mathbf{b})$$

vektor reziduí $\hat{\mathbf{e}}$ souvisí s vektorem chyb ε podle

$$\hat{\mathbf{e}} \approx (\mathbf{E} - \mathbf{P})\, \varepsilon$$

kde a P_{ik} jsou prvky projekční matice $\mathbf{P}$,

Každé reziduum je přibližně lineární kombinací všech chyb

$$\hat{e}_i = \varepsilon_i - \sum_{k=1}^{q} P_{ik}\, \varepsilon_k$$

ale u malých výběrů je rušivý ***efekt supernormality***,

1. Grafická analýza reziduí

a) odlehlé (extrémní) hodnoty v souboru reziduí,

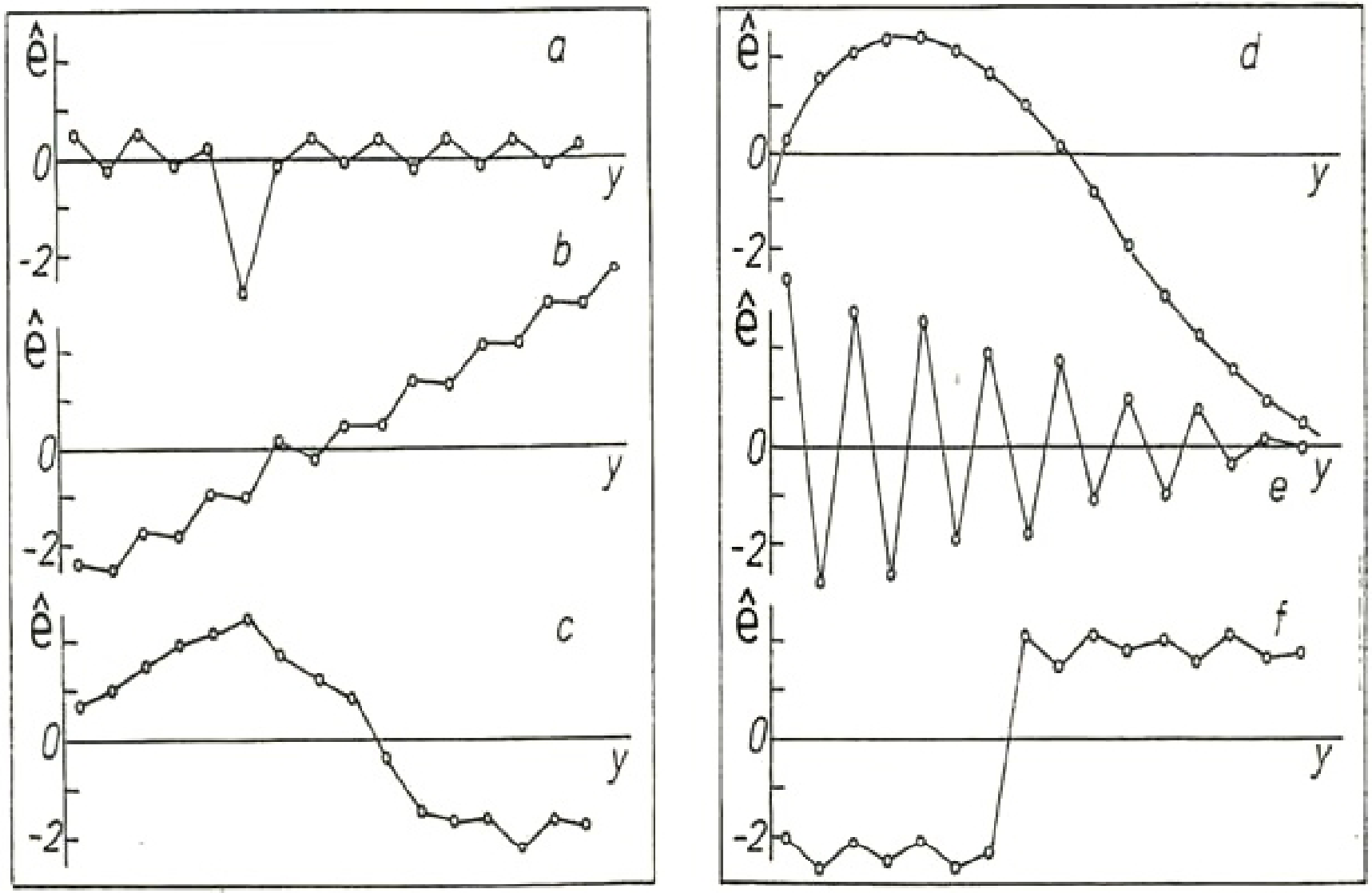
b) trend v reziduích,

c) nedostatečné střídání znaménka u reziduí,

d) chybný model nebo vzájemnou závislost reziduí,

e) heteroskedasticitu (nekonstantnost rozptylu) závisle proměnné (měřené) veličiny **y**,

f) náhlou změnu podmínek při měření hodnot y.



a) odlehlá hodnota v datech;
b) trend v reziduích;
c) nedostatečné střídání znaménka reziduí;

d) chybný model;
e) heteroskedasticita;
f) náhlá změna podmínek

2. Numerická analýza reziduí

1. ***Střední hodnota reziduí***, $E(\hat{e})$, by se měla rovnat nule,
2. ***Průměrné reziduum*** $|\bar{e}| = \frac{1}{n}\sum_{i-1}^{n}|\hat{e}_i|$ by se mělo rovnat náhodné chybě.
3. ***Směrodatná odchylka střední hodnoty reziduí*** $s(\hat{e})$ by se měla rovnat náhodné chybě.
4. ***Koeficient šikmosti*** $g_1(\hat{e})$ se pro Gaussovo rozdělení rovná nule.
5. ***Koeficient špičatosti*** $g_2(\hat{e})$ se pro Gaussovo rozdělení rovná třem.

Obecné testační charakteristiky:

$$T_{p,q} = \sum_{i=1}^{n} \hat{e}_i^{\,p} \, [f(x_i, \mathbf{b})]^q$$

a) ***Testační charakteristika*** $T_{1,1}$ by měla být (přibližně) rovna nule, protože obyčejně platí, že $\hat{\mathbf{e}}^T f(x_i, \mathbf{b}) = 0$.

b) ***Testační charakteristika*** $T_{2,1}$ ukazuje na heteroskedasticitu.

c) ***Testační charakteristika*** $T_{1,2}$ ukazuje na chybně navržený model,

d) ***Testační charakteristika*** $T_{1,0}$ by měla být přibližně rovna nule,

3. Analýza vlivných bodů

Vychází se z jednokrokové aproximace odhadu $\mathbf{b}_{(i)}$

$$\mathbf{b}_{(i)}^{1} = \mathbf{b} - \frac{(\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}_i \hat{e}_i}{1 - P_{ii}}$$

kde P_{ii} jsou prvky projekční matice,

a) ***Charakteristika DFS***$_{ij}$ vyjadřuje vliv i-tého bodu na odhad j-tého parametru

$$DFS_{ij} = \frac{b_j - b_{j(i)}^{1}}{\hat{s}_{(i)} \sqrt{V_{ii}}}$$

kde $\hat{s}^2_{(i)}$ je odhad rozptylu vyčíslený při vynechání i-tého bodu

podle vztahu

$$\hat{s}^2_{(i)} = \frac{U(\mathbf{b}) - \dfrac{\hat{e}_i^2}{1 - P_{ii}}}{n - m - 1}$$

a symbol V_{ii} značí prvky matice $\mathbf{V} = (\mathbf{J}^T \mathbf{J})^{-1}$.

Test: i-tý bod se považuje za vlivný, pokud je $DFS_{ij} > 2/\sqrt{n}$.

b) ***Jackknife rezidua*** $\hat{e}_{Ji}$

$$\hat{e}_{Ji} = \frac{\hat{e}_i}{\hat{s}_{(i)} \sqrt{1 - P_{ii}}}$$

Test: silně vlivné body mají $\hat{e}_{Ji}^2 > 10$.

c) ***Věrohodnostní vzdálenost***

$$LD_i = 2\ [\ln L(\mathbf{b}) - \ln L(\mathbf{b}_{(i)})]$$

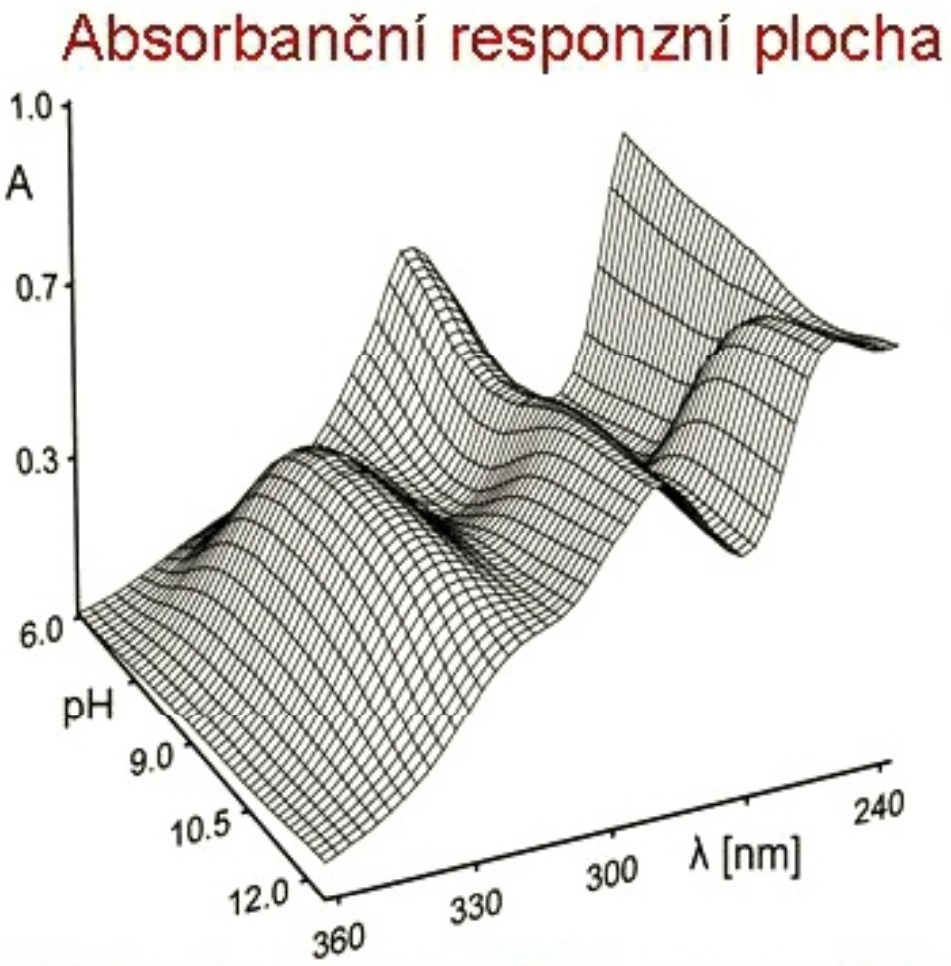
Pro případ MNČ bude věrohodnostní vzdálenost ve tvaru

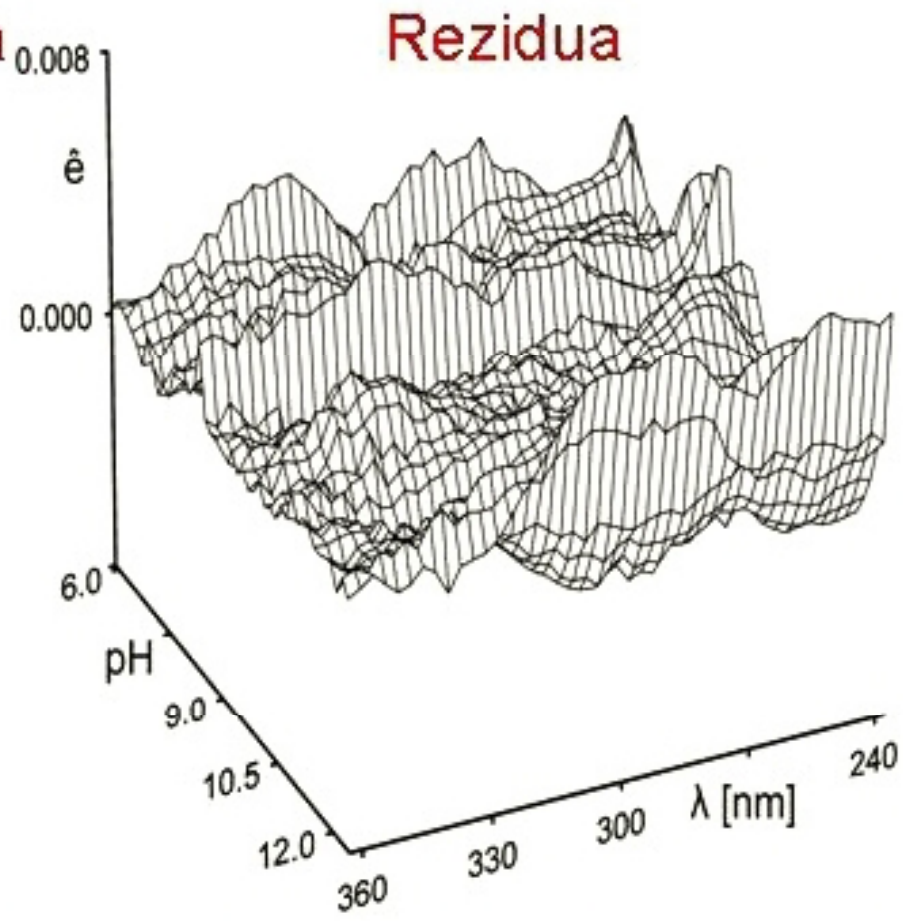
$$LD_i = n \ln\left[\frac{U(\mathbf{b}_{(i)})}{U(\mathbf{b})}\right]$$

Test: Je-li $LD_i > \chi^2_{1-\alpha}(2)$, je daný bod silně vlivný, obyčejně se volí $\alpha = 0.05$.

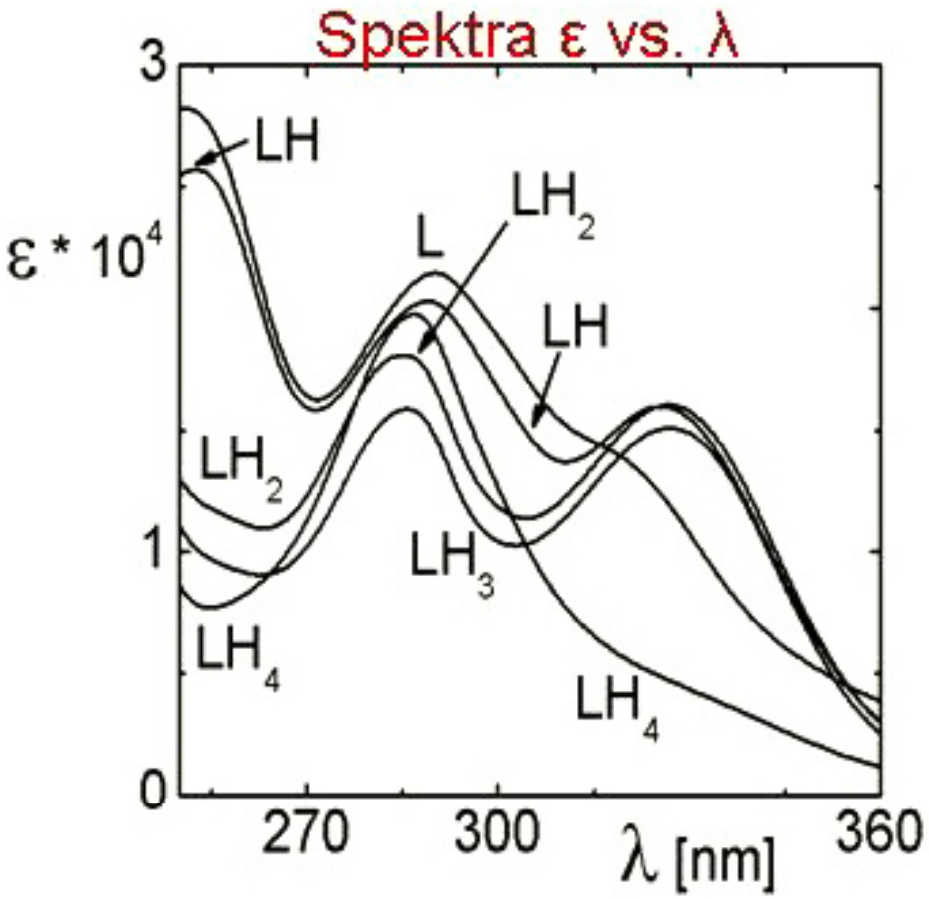
Input dat

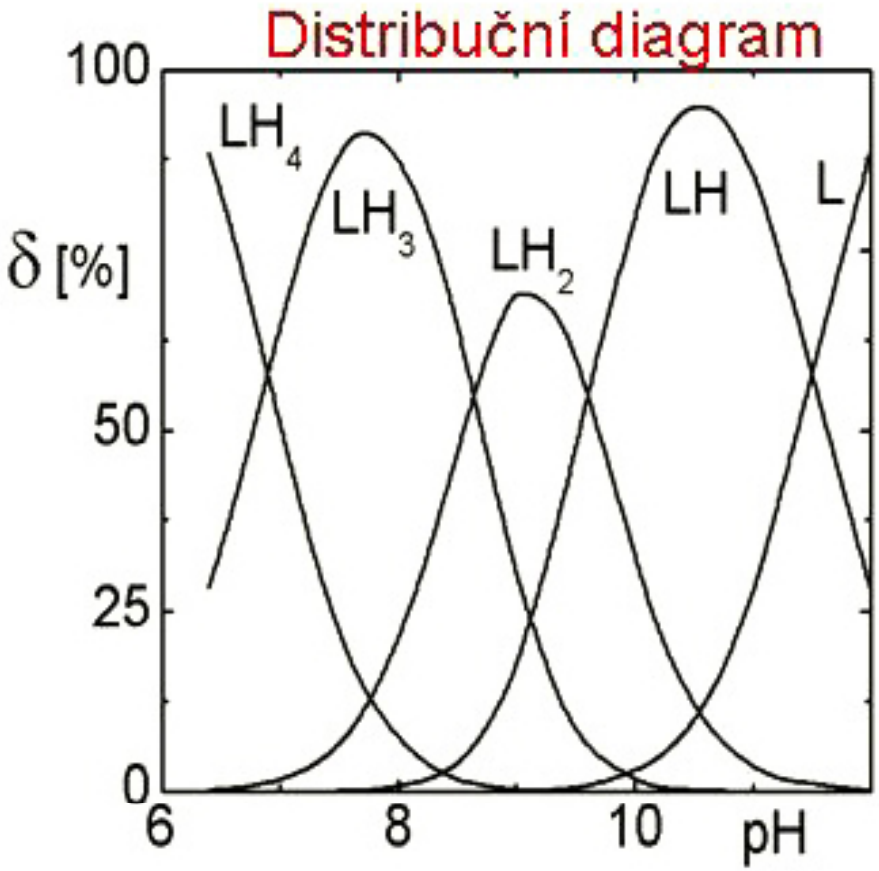
Grafická analýza reziduí



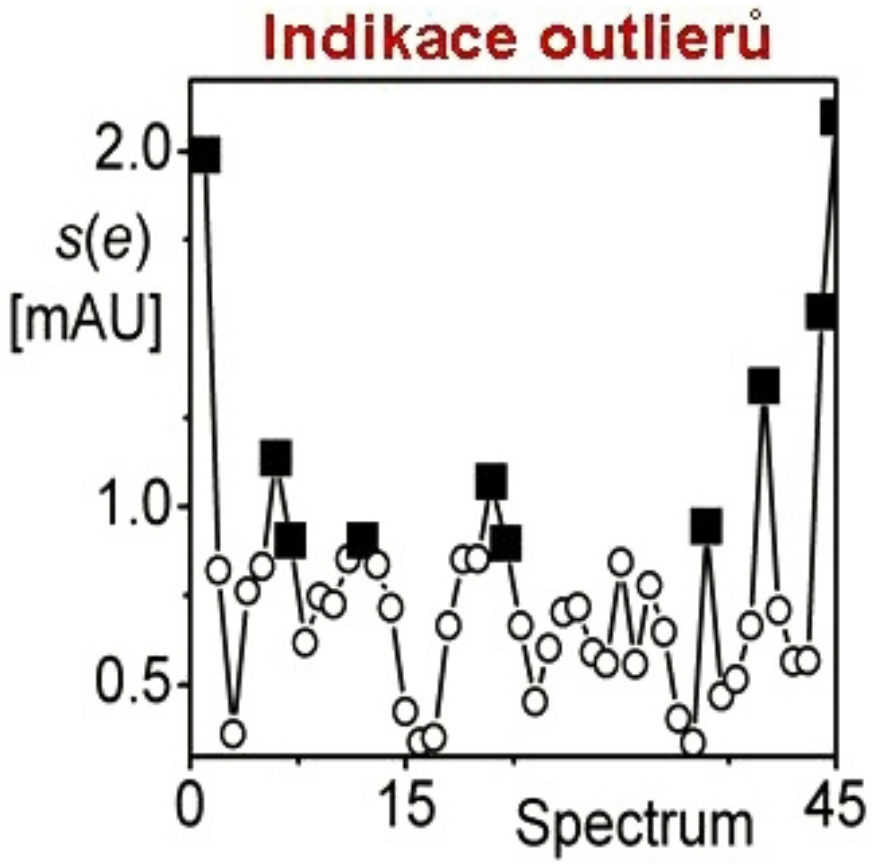


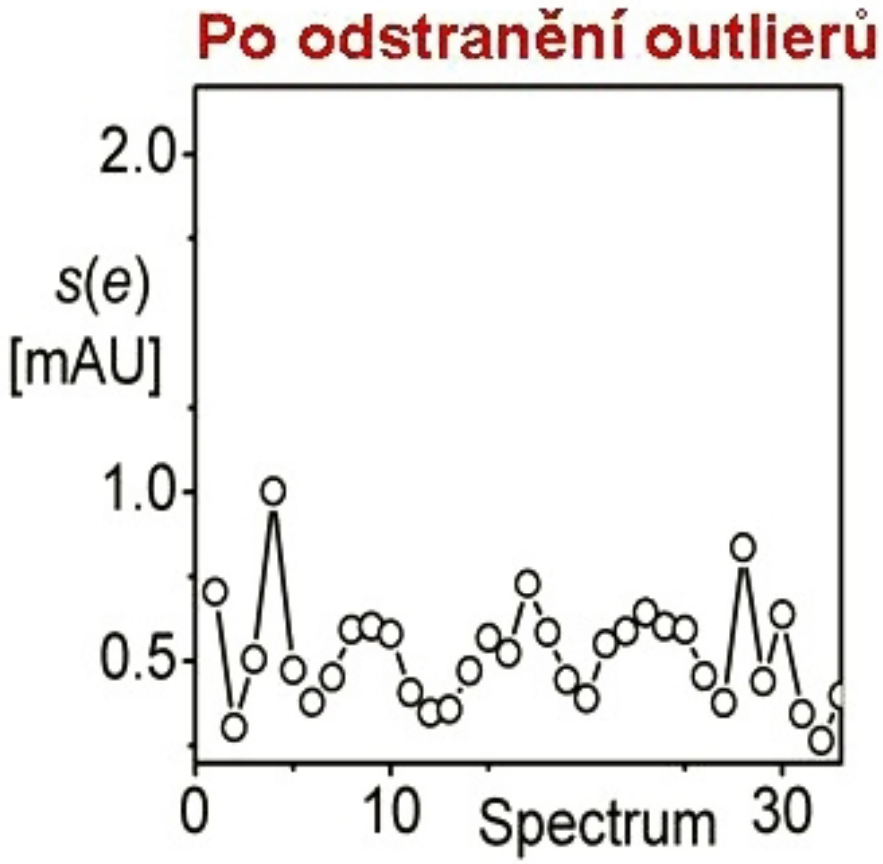
Grafické závěry o nalezených odhadech parametrů z postaveného chemického (=nelineárního regresního) modelu



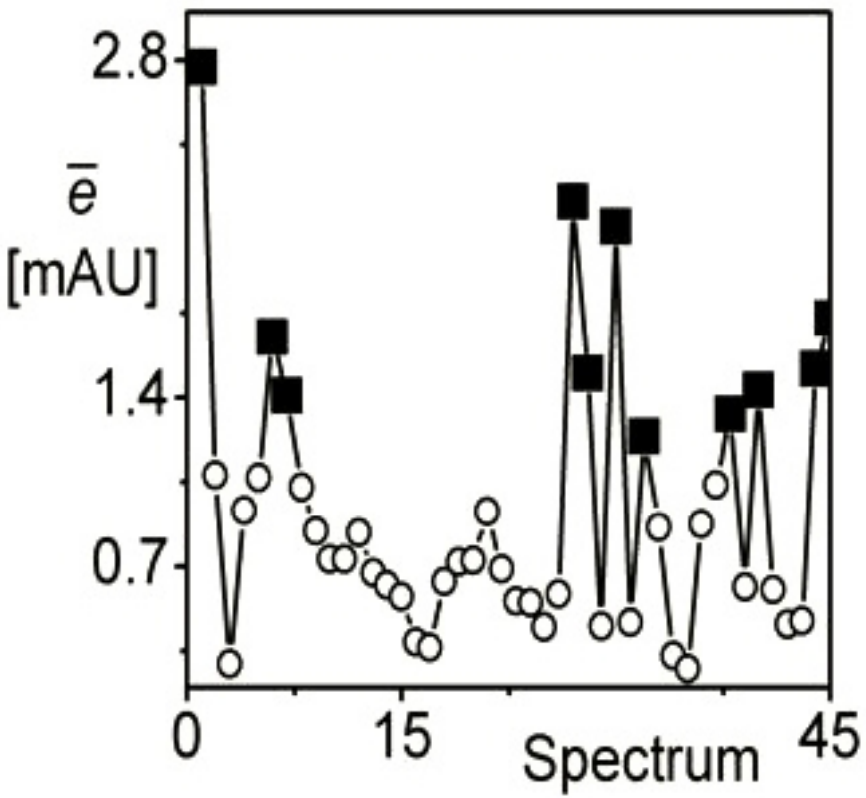


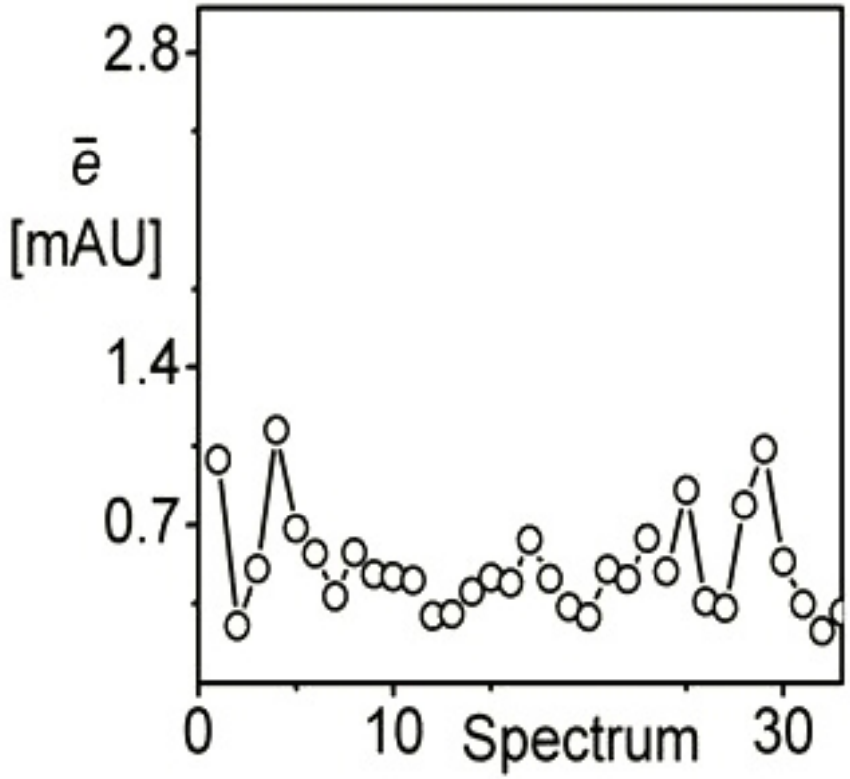
Mírou věrohodnosti nalezených odhadů regresního modelu je vždy míra těsnosti proložení experimentálních bodů vypočtenou regresní křivkou čili statistická analýza reziduí před a po odstranění outlierů.



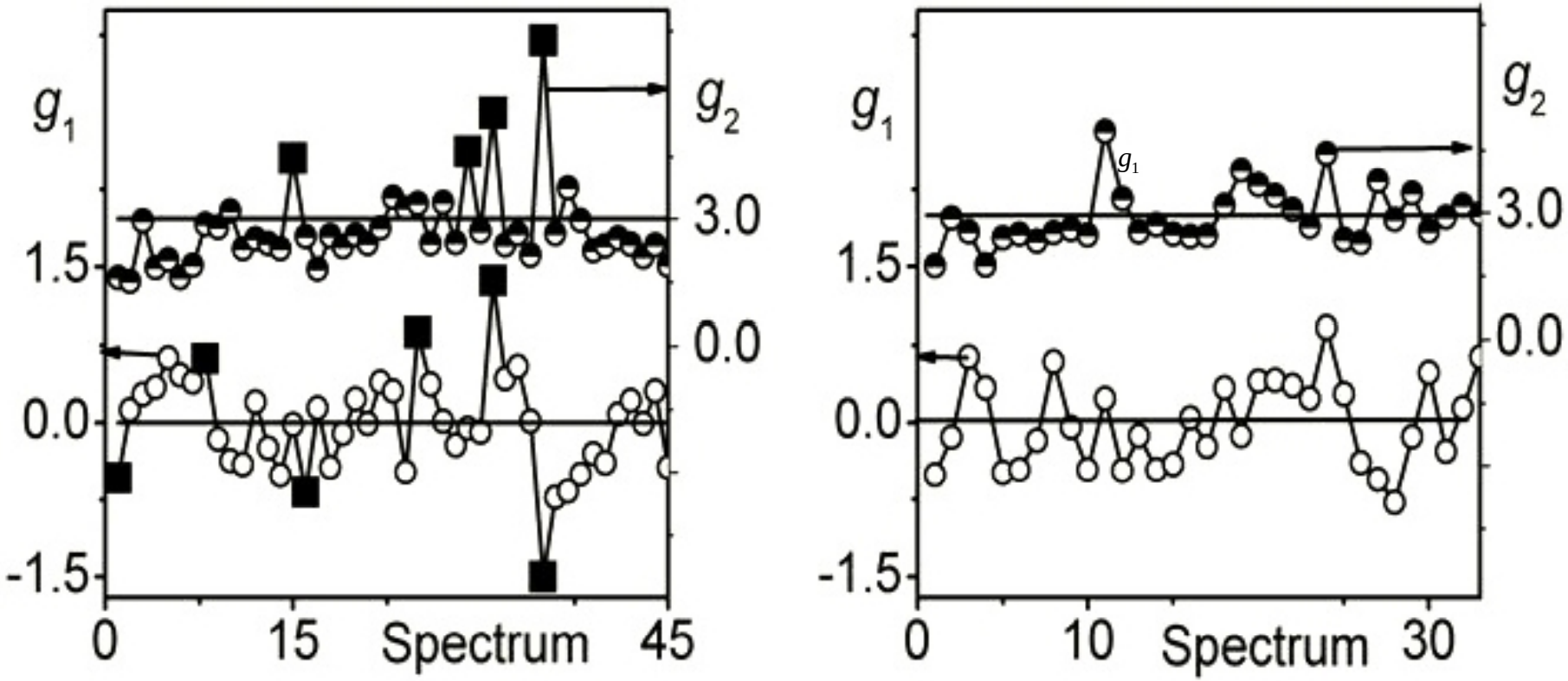


Míru těsnosti proložení je vhodné posuzovat dle velikosti reziduí a jejich rozdělení, a to především dle střední hodnoty reziduí a symetrie rozdělení (vyjádřené koeficientem šikmosti a špičatosti).





Symetrii a tvar rozdělení vystihuje koeficient šikmosti g_1 a špičatosti g_2) před a po odstranění outlierů.



Souhrn:

Postup při testování navrženého modelu

1. Kvalita nalezených odhadů parametrů

Kvalita se posuzuje podle:

a) jejich intervalů spolehlivosti,

b) rozptylů $D(b_j)$.

Pravidlo "3 sigma": $3\sqrt{D(b_j)} < |b_j|$

2. Kvalita dosažené těsnosti proložení

a) ***Reziduální rozptyl*** $\hat{\sigma}^2 = U(\mathbf{b})/(n - m)$.

b) ***Koeficient determinace*** D, který je pro lineární modely čtvercem vícenásobného korelačního koeficientu

$$R^2 = D = 1 - \frac{U(\mathbf{b})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \text{ kde } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Stonásobek koeficientu determinace je ***regresní rabat***, 100 D.

d) K rozlišení mezi modely užijeme

Akaikova informačního kritéria AIC

$$\mathrm{AIC} = -\,L(\mathbf{b}) + 2\,m$$

Za optimální se považuje model, pro který dosahuje AIC minimální hodnoty, a platí

$$\mathrm{AIC} = n\,\ln\left[\frac{U(\mathbf{b})}{n}\right] + 2\,m$$

3. Predikční schopnost modelu

Postup: data se rozdělí na dvě podskupiny

M_1 (s indexy i = 1, ..., int(n/2)) s odhady $\mathbf{b}(M_1)$

M_2 (s indexy i = int(n/2) + 1, ..., n) a s odhady $\mathbf{b}(M_2)$.

a) ***Predikční schopnost modelu*** lze vyjádřit kritériem

$$K = \frac{U(\mathbf{b})}{\sum_{i \in M_1} [y_i - f(x_i, \mathbf{b}(M_2))]^2 + \sum_{i \in M_2} [y_i - f(x_i, \mathbf{b}(M_1))]^2}$$

Predikční schopnost modelu je tím vyšší, čím víc se hodnota K blíží k jedničce.

b) ***Střední kvadratická chyba predikce*** je definována

$$\mathrm{MEP} = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, \mathbf{b}_{(i)})^2$$

Čím je MEP nižší, tím má model lepší predikční schopnost.

4. Kvalita experimentálních dat

K posouzení kvality experimentálních dat se užívá analýzy reziduí a vlivných bodů.

5. Správnost navrženého modelu

Whitův test

$$C = (n - 1) \sum_{i=1}^{n} \frac{\delta f(x_i, \mathbf{b})}{\delta \beta_j} \frac{\delta f(x_i, \mathbf{b})}{\delta \beta_k} \left[\hat{e}_i - \frac{U(\mathbf{b})}{n} \right]$$

Pro správný model je C = 0. Při výpočtu se postupuje tak, že se vyčíslí m(m + 1)/2 umělých proměnných

$$W_s = \frac{\delta f(x_i, \mathbf{b})}{\delta \beta_j} \frac{\delta f(x_i, \mathbf{b})}{\delta \beta_k}$$

pro s = 1, ..., m(m + 1)/2.

Za testační charakteristiku se považuje korelační koeficient regrese veličiny $\hat{e}^2$ na vektoru proměnných **W**. V případě správného modelu má tato charakteristika $\chi^2(m(n + 1)/2)$ rozdělení.

6. Souhlas s požadavky fyzikálního smyslu

Parametry by měly mít fyzikální smysl.

Postup tvorby nelineárního regresního modelu

1. Návrh regresního modelu.

Obvykle se používá nějaká fyzikální nebo empirická závislost.

2. Odhadování parametrů.

K hledání minima kritéria regrese se užívá iterativních algoritmů, kritérium minima součtu čtverců reziduí.

3. Posouzení kvality odhadů.

Kvalita nalezených odhadů se posuzuje dle jejich intervalů spolehlivosti nebo pouze jejich rozptylů $D(b_j)$.

Příčinou vysokých rozptylů parametrů bývá také předčasné ukončení minimalizačního procesu před dosažením minima.

4. Grafické posouzení vhodnosti modelu.

Grafická analýza reziduí využívá rozptylového grafu reziduí vs. predikce a odhalí:

a) odlehlé hodnoty,

b) trend v reziduích,

c) nedostatečné střídání znaménka u reziduí,

d) heteroskedasticitu.

K ověření normality rozdělení reziduí se užíjí rankitové grafy a vyčíslení koeficientu šikmosti $g_1(\hat{e})$ a špičatosti $g_2(\hat{e})$.

5. Základní statistické charakteristiky.

O přiblížení modelu experimentálním datům informuje suma čtverců reziduí v minimu $U(\boldsymbol{b})$, ze které se vyčíslí *reziduální rozptyl* $\hat{\sigma}^2 = U(\boldsymbol{b})/(n - m)$.

Z $U(\boldsymbol{b})$ je odvozen ***koeficient determinace D*** a ***regresní rabat***, 100 D [%].

$$D = 1 - \frac{U(\boldsymbol{b})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad kde\ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i ,$$

Hamiltonův R-faktor je definován $R\text{-}faktor = \sqrt{\dfrac{U(\boldsymbol{b})}{\sum_{i=1}^{n} y_i^2}}$,

a pro $\bar{y} = 0$ platí, že R^2-*faktor* = 1 - D a pro $\bar{y} \neq 0$ pak platí vztah

$$R\text{-}faktor = \sqrt{(1 - D) - \frac{(1 - D)\, n\, \bar{y}^2}{\sum_{i=1}^{n} y_i^2}} .$$

Hamiltonův R-faktor ukazuje na rozdíl mezi modelem $y = f(x, \beta)$ a modelem $y = 0$.

Mezi modely rozliší ***Akaikovo informační kritérium***

$$AIC = -L(\boldsymbol{b}) + 2m.$$

Optimální je model, pro který je AIC minimální. Platí přitom

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

6. Regresní diagnostika.

Obsahuje **pomůcky analýzy regresního tripletu**, tj. pro

kritiku dat, kritiku modelu* a *kritiku metody.

Analýzou vlivných bodů (vybočující pozorování a extrémy) se identifikují body, které silně ovlivňují odhadované parametry v modelu.

Pro aditivní modely měření a MNČ jsou rezidua definována vztahem $\hat{e}_i = y_i - f(x_i, \mathbf{b})$.

B. Analýza vlivných bodů:

U lineárních regresních modelů: je odhalení vlivných bodů VB pomocí reziduí $\hat{e}_i$ a prvků P_{ii} projekční matice

$$\boldsymbol{P} = \boldsymbol{X}\,(\boldsymbol{X}^{\mathrm{T}}\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\mathrm{T}}.$$

U nelineárních regresních modelů: je proto třeba konstruovat matici $\boldsymbol{P}$ vztahem

$$\boldsymbol{P} = \boldsymbol{J}\,(\boldsymbol{J}^{\mathrm{T}}\,\boldsymbol{J})^{-1}\,\boldsymbol{J}^{\mathrm{T}}$$

kde $\boldsymbol{J}$ je Jakobián čili derivaci modelové funkce podle jednotlivých parametrů v daných bodech. Komplikace je v tom, že již nelze vyjádřit odhady parametrů a rezidua jako lineární kombinaci experimentálních dat.

Vychází se proto z jednokrokové aproximace odhadu $\boldsymbol{b}_{(i)}$

$$\boldsymbol{b}_{(i)}^{1} = \boldsymbol{b} - \frac{(\boldsymbol{J}^{T}\,\boldsymbol{J})^{-1}\,\boldsymbol{J}_i\,\hat{e}_i}{1 - P_{ii}},$$

kde P_{ii} jsou prvky projekční matice $\boldsymbol{P}$.

Charakteristika $\boldsymbol{DFS_{ij}}$ vyjadřuje vliv i-tého bodu na odhad j-tého parametru dle vztahu

$$DFS_{ij} = \frac{b_j - b^1_{j(i)}}{\hat{s}_{(i)} \sqrt{V_{ii}}} ,$$

kde $\hat{s}^2_{(i)}$ je odhad rozptylu vyčíslený při vynechání i-tého bodu

$$\hat{s}^2_{(i)} = \frac{U(\boldsymbol{b}) - \dfrac{\hat{e}_i^2}{1 - P_{ii}}}{n - m - 1}$$, kde V_{ii} značí prvky matice $\boldsymbol{V} = (\boldsymbol{J}^{\mathrm{T}} \boldsymbol{J})^{-1}$.

Test: i-tý bod je vlivný, když je $DFS_{ii} > 2/\sqrt{n}$.

Vlivné body VB lze identifikovat jednokrokovou aproximací ***Jackknife rezidui*** dle vztahu

$$\hat{e}_{Ji} = \frac{\hat{e}_i}{\hat{s}_{(i)}\sqrt{1 - P_{ii}}} .$$

K vyjádření vlivu jednotlivých bodů na odhady parametrů lze použít změny vektoru vychýlení $\boldsymbol{h}_{(i)}$ při vynechání i-tého bodu nebo změny střední hodnoty i-tého rezidua při vynechání i-tého bodu:

Mezi nelineární míry vlivu i-tého bodu na odhady parametrů patří ***věrohodnostní vzdálenost***

$$LD_i = 2\,[\ln L(\boldsymbol{b}) - \ln L(\boldsymbol{b}_{(i)})]$$

a pro MNČ je ve tvaru $$LD_i = n \ln\left[\frac{U(\boldsymbol{b}_{(i)})}{U(\boldsymbol{b})}\right].$$

Do obou vztahů lze dosadit buď odhady $\boldsymbol{b}_{(i)}$, určené regresí při vynechání i-tého bodu, nebo $\boldsymbol{b}^1_{(i)}$, určené z jednokrokové aproximace.

Test: Je-li $LD_i > \chi^2_{1-\alpha}(2)$, je daný bod silně vlivný a $\alpha = 0.05$.

(a) VB ovlivňují odhady parametrů a relativní vychýlení $\boldsymbol{h}_R$.

(b) Charakteristiky založené na aproximaci nelineárního modelu neindikují vždy správně přítomnost VB.

(c) Nejlepší indikaci VB poskytuje LD_i, protože umožňuje indikaci celé skupiny vlivných bodů, kde může dojít k jejich vzájemnému "maskování".

7. Mapa citlivostní funkce.

U nelineárních modelů existuje řada komplikací:

-neodhadnutelnost některých parametrů,

-existence minima funkce $U(\beta)$ jen pro některé regresní modely,

-výskyt lokálních minim,

-existence sedlových bodů, ovlivňujících kriteriální funkci $U(\beta)$ a

-špatná podmíněnost parametrů v regresním modelu.

Problémy lze indikovat analýzou ***normalizovaných citlivostních koeficientů***

$$C_{j(i)} = \beta_j \frac{\delta f(x_i, \beta)}{\delta \beta_j} \qquad j = 1, \ldots, m \; . \qquad i = 1, \ldots, n$$

Pro vizuální posouzení špatné podmíněnosti např. multikolinearitou mezi parametry β_j, β_h , se konstruují ***citlivostní grafy závislosti*** $C_{j(i)}$ a $C_{h(i)}$ na x_i , $i = 1, \ldots, n$ nebo závislost normalizovaných citlivostních koeficientů přímo na indexu i.

Citlivost regresních modelů na změnu parametru β_j vyjadřuje

$$\textit{celková citlivostní funkce} \quad C_{cj} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta f(x_i, \beta)}{\delta \beta_j} \right]^2 ,$$

která je nekonstantní pro nelineární parametry β_j, v modelu $f(x, \beta)$.

Interpretace citlivostních grafů parametrů: když jsou závislosti C_{cj} na β_j v okolí bodů $\beta_j^{(0)}$ nebo b_j přibližně konstantní, indikuje to malou citlivost regresního modelu ke změnám j-tého parametru, nebo je model $f(x, \beta)$ vzhledem k parametru β_j *lineární*.

8. Predikční schopnost modelu "cross-validation":

Data se rozdělí na dvě podskupiny M_1 (s indexy $i = 1, ..., \text{int}(n/2)$) a M_2 (s indexy $i = \text{int}(n/2) + 1, ..., n$).

Označí se:

odhady parametrů z bodů podskupiny M_1 jako $\boldsymbol{b}(M_1)$ a

odhady parametrů z bodů podskupiny M_2 jako $\boldsymbol{b}(M_2)$.

Predikční schopnost modelu se vyjádří ***kritériem K***

$$K = \frac{U(\boldsymbol{b})}{\sum_{i \in M_1} [y_i - f(x_i, \boldsymbol{b}(M_2))]^2 + \sum_{i \in M_2} [y_i - f(x_i, \boldsymbol{b}(M_1))]^2} .$$

Test: Predikční schopnost modelu je tím vyšší, čím víc se K blíží k 1.

Kritérium ***střední kvadratická chyba predikce***

$$MEP = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{b}_{(i)}))^2 .$$

Test: Čím je *MEP* nižší, tím je model věrohodnější a má lepší predikční schopnost.

9. Souhlas s požadavky fyzikálního smyslu.

Na odhady parametrů jsou kladena omezení, vycházející z fyzikálního smyslu. Odhady musí ležet v jisté předpokládané oblasti (např. koncentrace v oblasti kladných čísel, molární absorpční koeficienty ε v oboru čísel 10 až 10^6, konstanty stability log β_{pqr} v oboru čísel 0 až 50 atd.).