

Těsnost proložení

Statistická analýza reziduí

Pro aditivní modely měření:

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

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

$$\hat{\epsilon} \approx (\mathbf{E} - \mathbf{P}) \epsilon$$

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

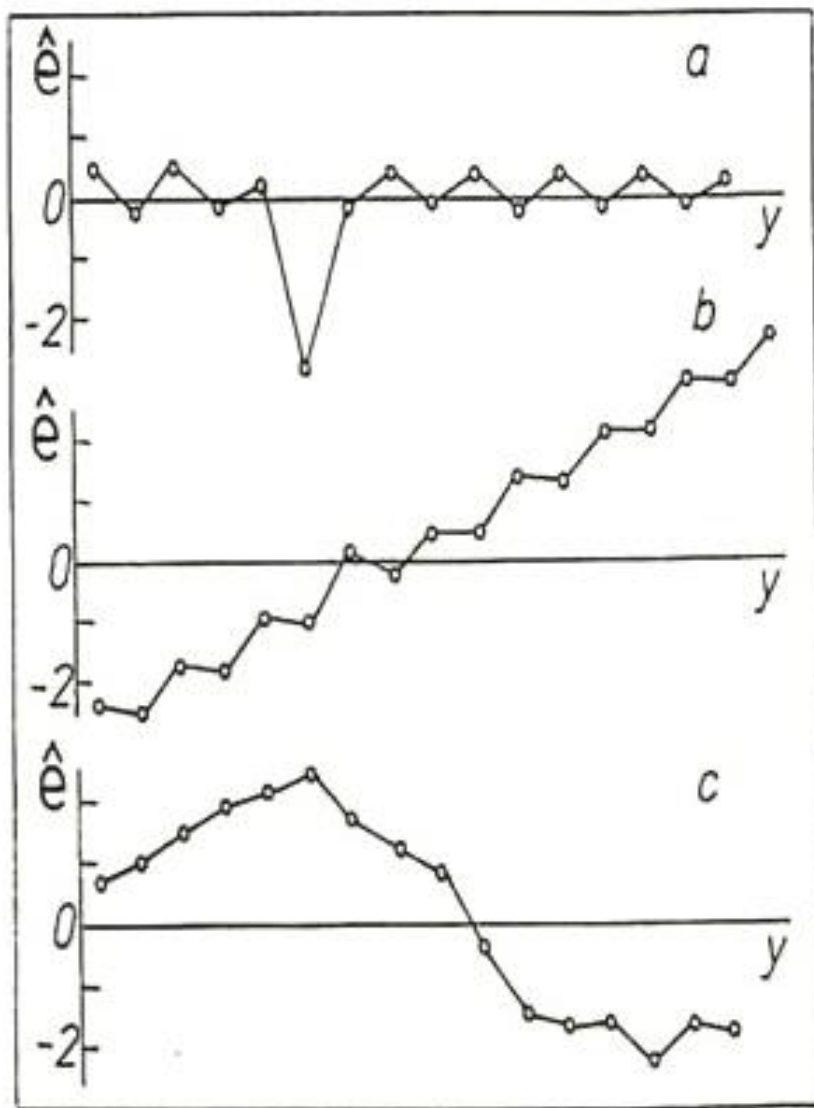
Každé reziduum je přibližně lineární kombinací všech chyb

$$\hat{\epsilon}_i = \epsilon_i - \sum_{k=1}^q P_{ik} \epsilon_k$$

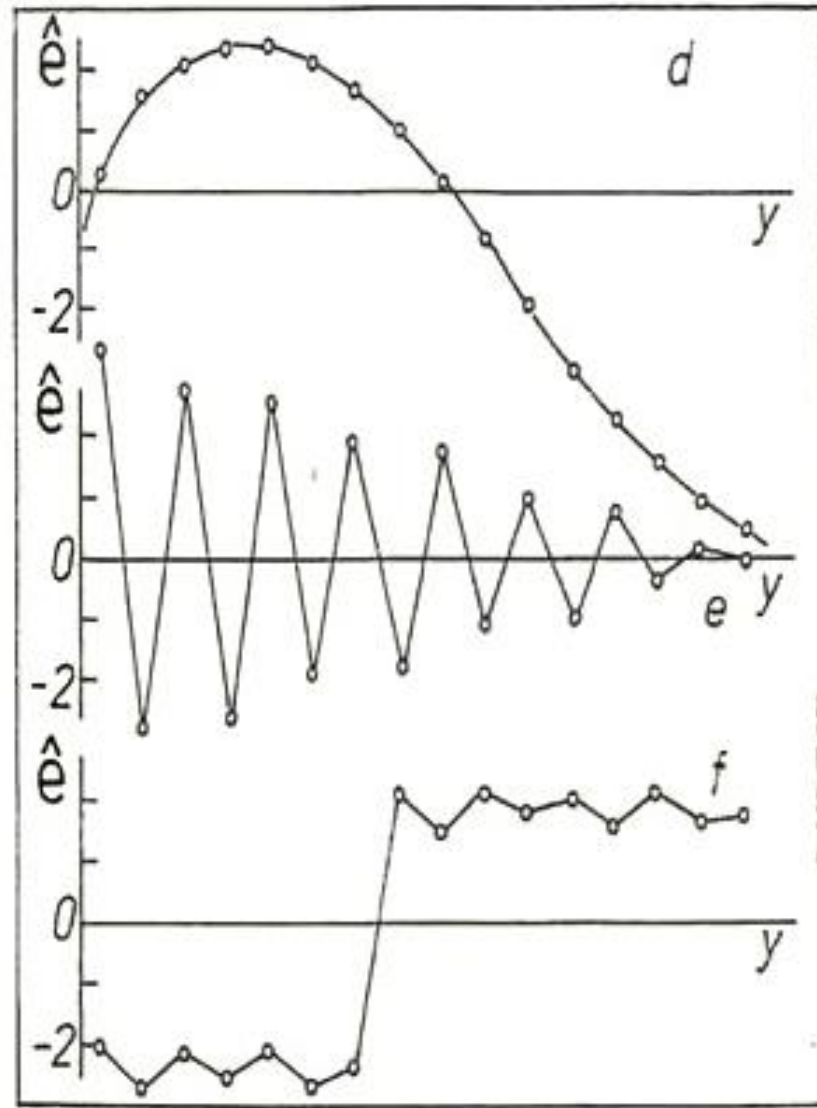
ale u malých výběrů je rušivý *efekt supernormality*,

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*, $E(\hat{\epsilon})$, by se měla rovnat nule,

2. *Průměrné reziduuum* $|\bar{e}| = \frac{1}{n} \sum_{i=1}^n |\hat{\epsilon}_i|$ by se mělo rovnat

náhodné chybě.

3. *Směrodatná odchylka střední hodnoty reziduí* $s(\hat{\epsilon})$ by se měla rovnat náhodné chybě.

4. *Koeficient šikmosti* $g_1(\hat{\epsilon})$ se pro Gaussovo rozdělení rovná nule.

5. *Koeficient špičatosti* $g_2(\hat{\epsilon})$ 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 obvykle platí, že $\hat{\mathbf{e}}^T \mathbf{f}(x_i, \mathbf{b}) = 0$.
- 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 $\mathbf{b}_{(i)}$

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

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

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

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

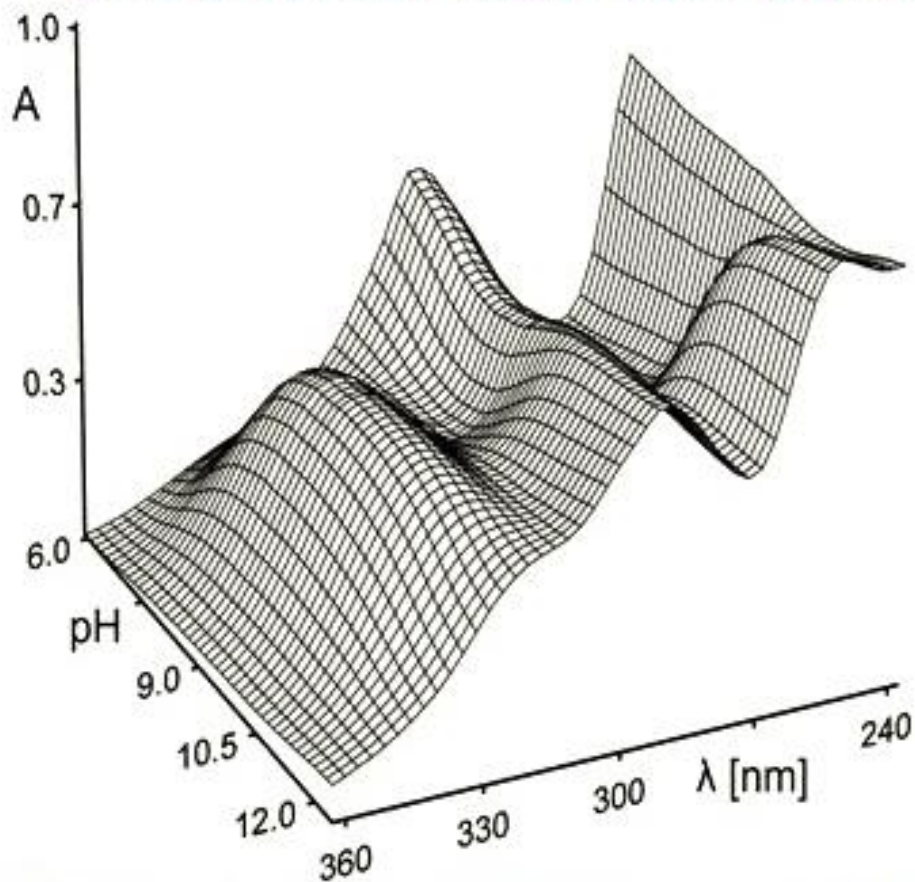
$$LD_i = n \ln \left[\frac{U(\mathbf{b}_{(i)})}{U(\mathbf{b})} \right]$$

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

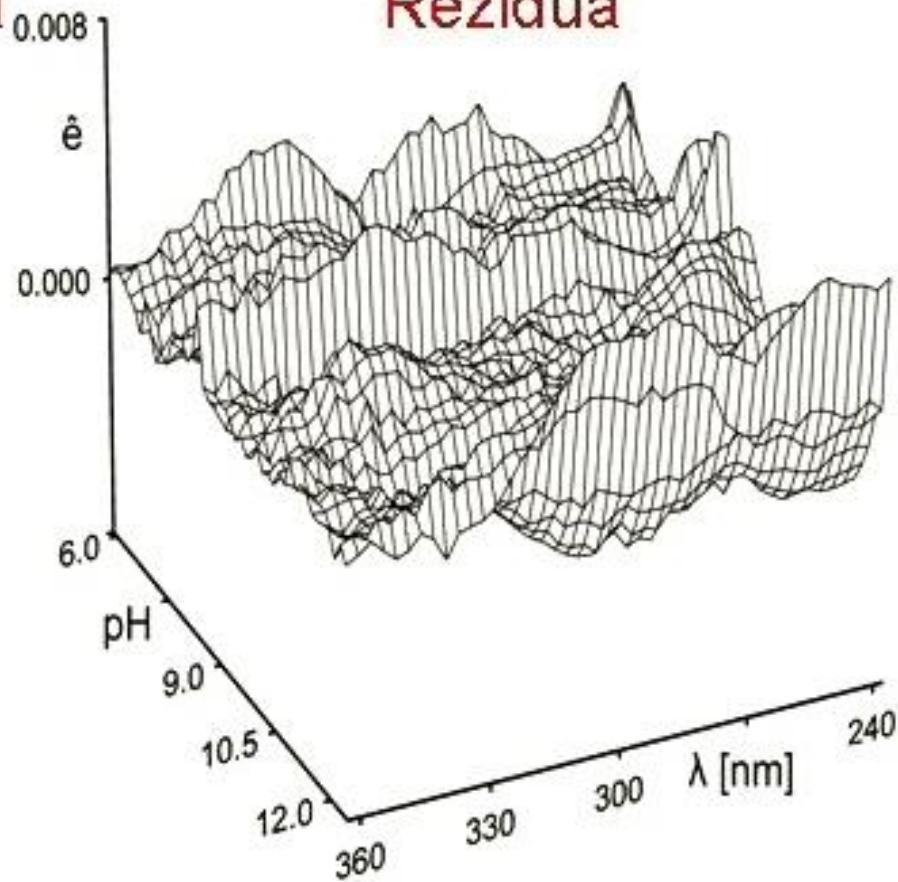
Input dat

Grafická analýza reziduí

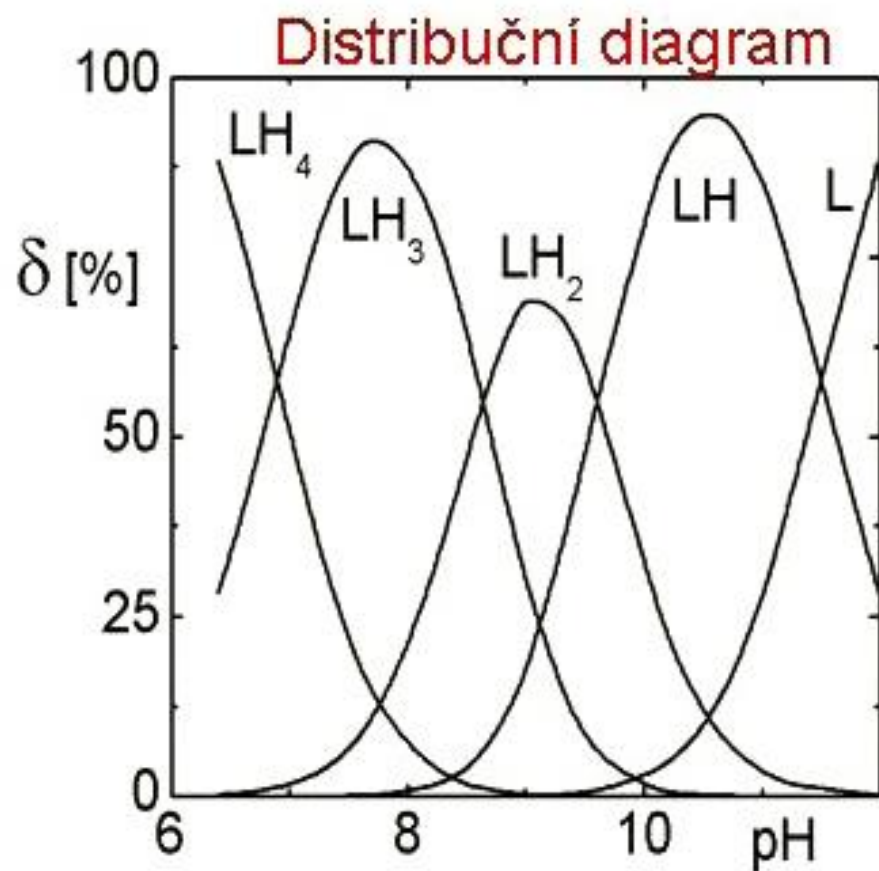
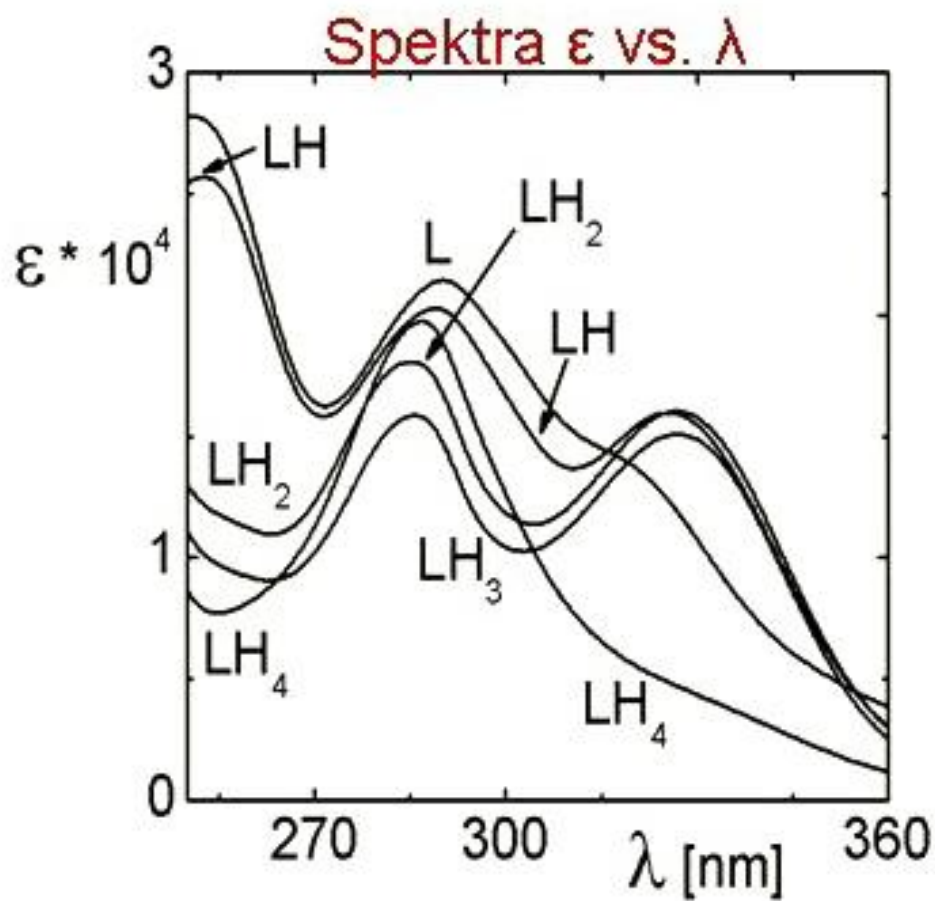
Absorbanční responzní plocha



Rezidua

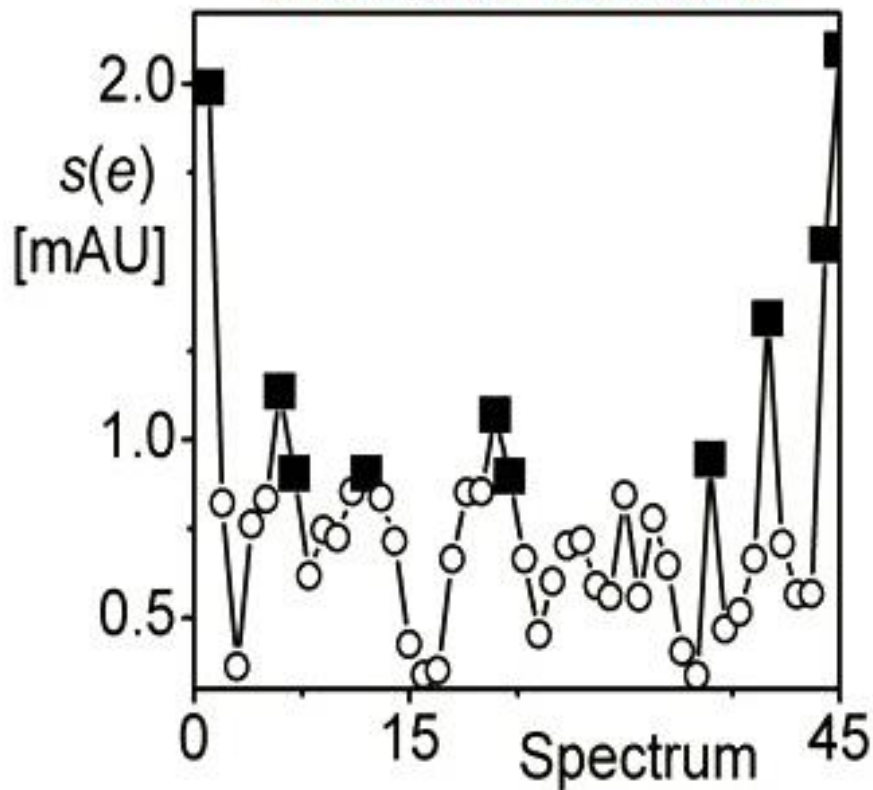


Grafické závěry o nalezených odhadech parametrů z postaveného chemického (=nelineárního regresního) modelu

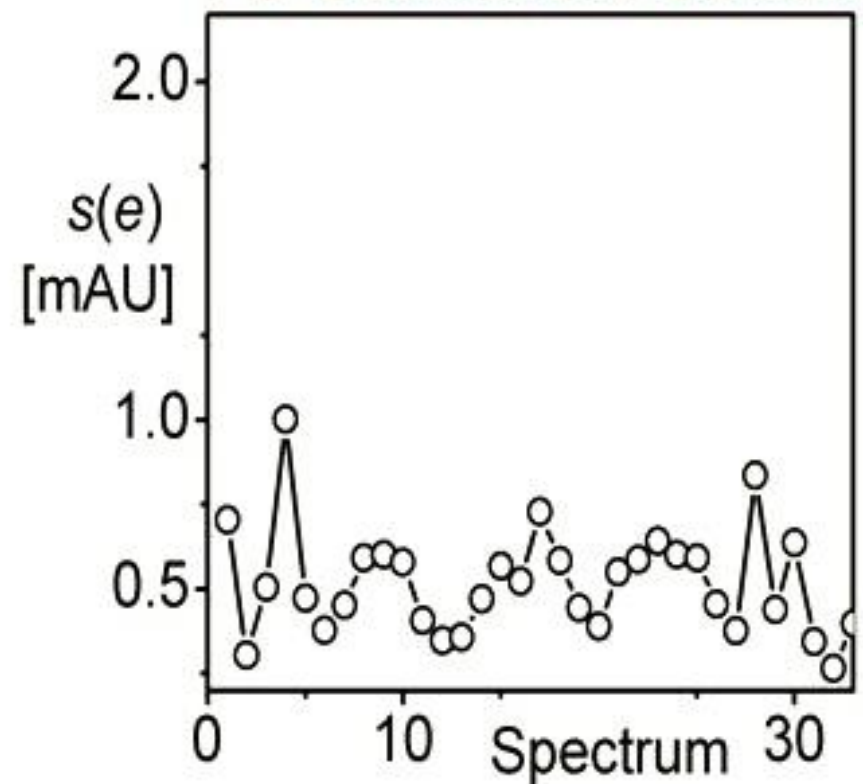


Mírou věrohodnosti nalezených odhadů regresního modelu je vždy míra těsnosti proložení experimentálních bodů vypočtenou regresní křivkou čili statistická analýza reziduí před a po odstranění outlierů.

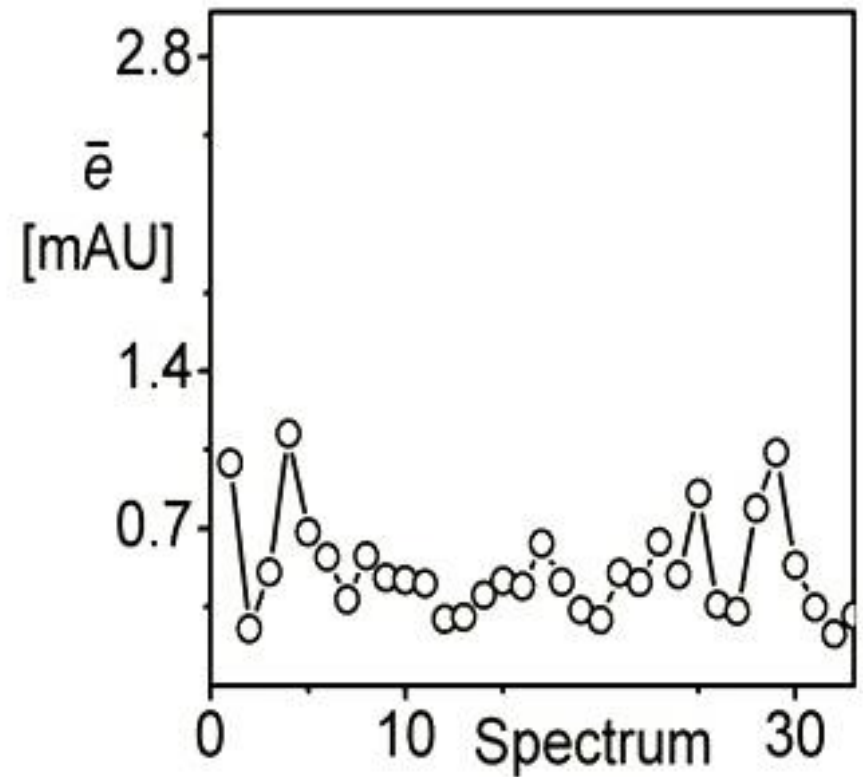
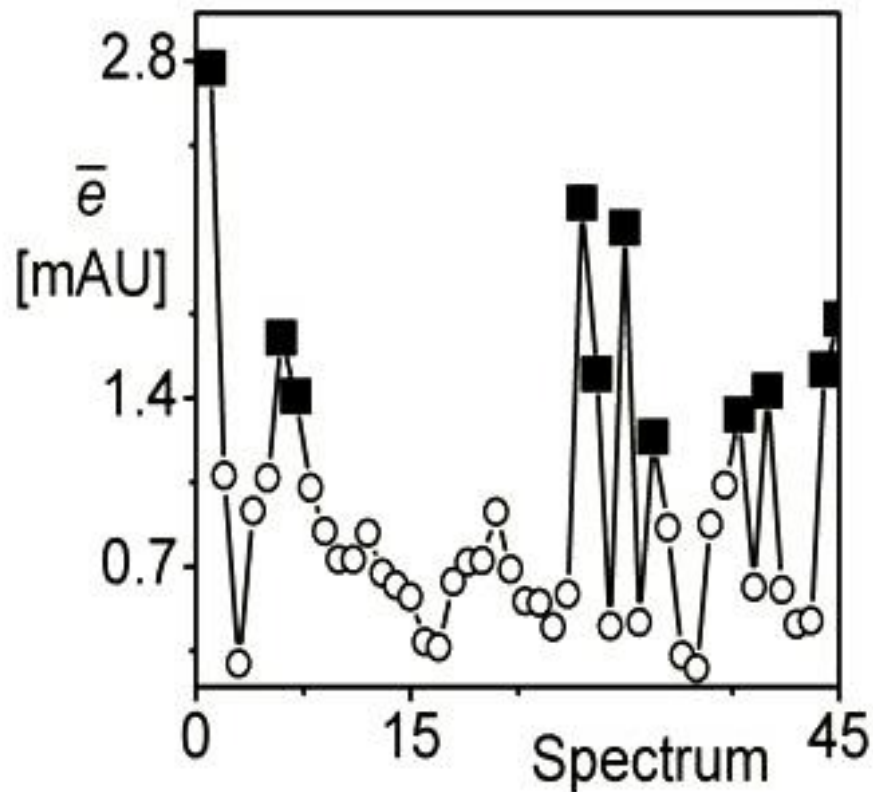
Indikace outlierů



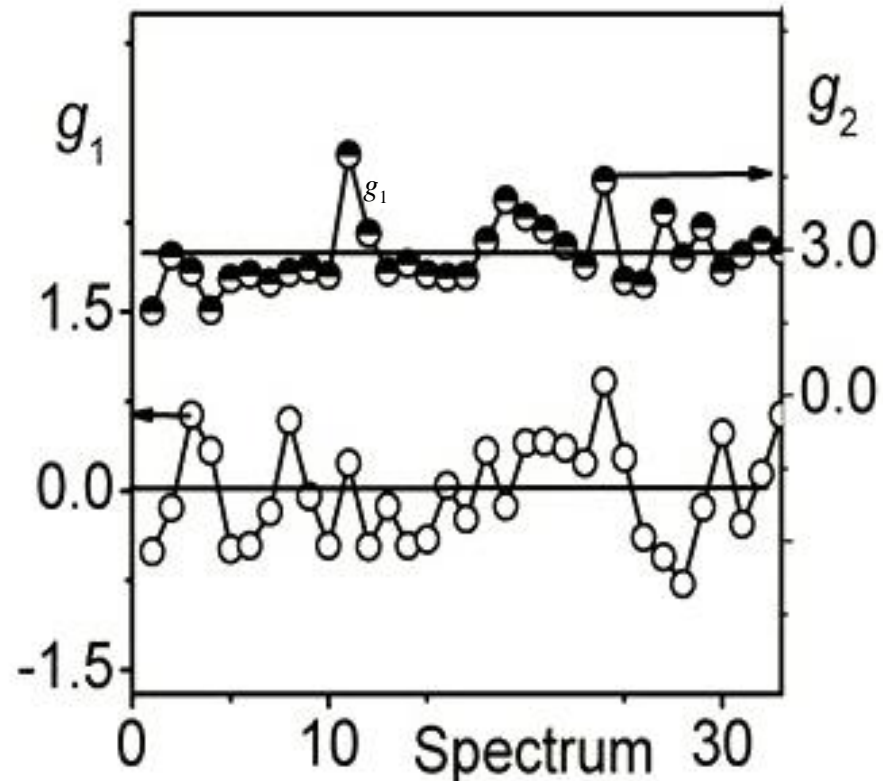
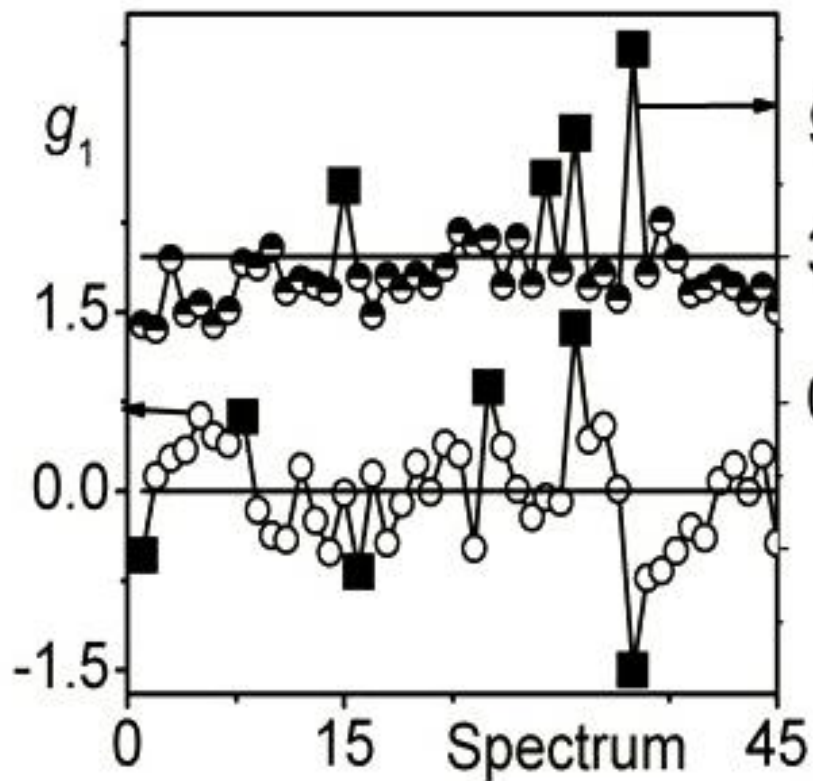
Po odstranění outlierů



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.



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 užijí rankitové grafy a vyčíslení koeficientu šikmosti $g_1(\hat{\epsilon})$ a špičatosti $g_2(\hat{\epsilon})$.

5. Základní statistické charakteristiky.

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

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

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

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

a pro $\bar{y} = 0$ platí, že *R²-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*

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

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

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

6. Regresní diagnostika.

Obsahuje **pomůcky analýzy regresního tripletu**, tj. pro *kritiku dat, kritiku modelu a kritiku metody*.

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

Pro aditivní modely měření a MNC 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

$$P = X (X^T X)^{-1} X^T.$$

U nelineárních regresních modelů: je proto třeba konstruovat matici P vztahem

$$P = J (J^T J)^{-1} J^T$$

kde 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 $\mathbf{b}_{(i)}$

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

kde P_{ii} jsou prvky projekční matice P .

Charakteristika DFS_{ij} vyjadřuje vliv i -tého bodu na odhad j -tého

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

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

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

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

Vlivné body VB lze identifikovat jednokrokovou aproximací *Jackknife*

reziduí 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 $\mathbf{h}_{(i)}$ 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*

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

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

Do obou vztahů lze dosadit buď odhady $\mathbf{b}_{(i)}$, určené regresí při vynechání i -tého bodu, nebo $\mathbf{b}_{(i)}^1$, 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 \mathbf{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 kritériální funkci $U(\beta)$ a
- špatná podmíněnost parametrů v regresním modelu.

Problémy lze indikovat analýzou *normalizovaných citlivostních*

$$\textit{koeficientů} \quad C_{j(i)} = \beta_j \frac{\delta f(x_i, \beta)}{\delta \beta_j} \quad \begin{array}{l} j = 1, \dots, m \\ i = 1, \dots, n \end{array} .$$

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

Citlivost regresních modelů na změnu parametru β_j vyjadřuje

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

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

Označí se:

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

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

Predikční schopnost modelu se vyjádří *kritériem* K

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

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

Kritérium *střední kvadratická chyba predikce*

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

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 $\log \beta_{pqr}$ v oboru čísel 0 až 50 atd.).