Porovnání rozlišovací schopnosti regresní analýzy spekter a spolehlivosti

Prof. RNDr. Milan Meloun, DrSc.

Katedra analytické chemie,
Chemickotechnologická fakulta, Univerzita Pardubice,
nám. Čs. Legií 565, 532 10 Pardubice,

http: //meloun.upce.cz,
milan.meloun@upce.cz,

Cíl regresní analýzy

Cílem regresní analýzy je nalezení vhodného modelu studované závislosti tak, že se **snažíme nahradit**

každou měřenou (experimentální) **hodnotu závisle proměnné** y_{exp}

hodnotou vypočtenou (predikovanou) y_{vyp}

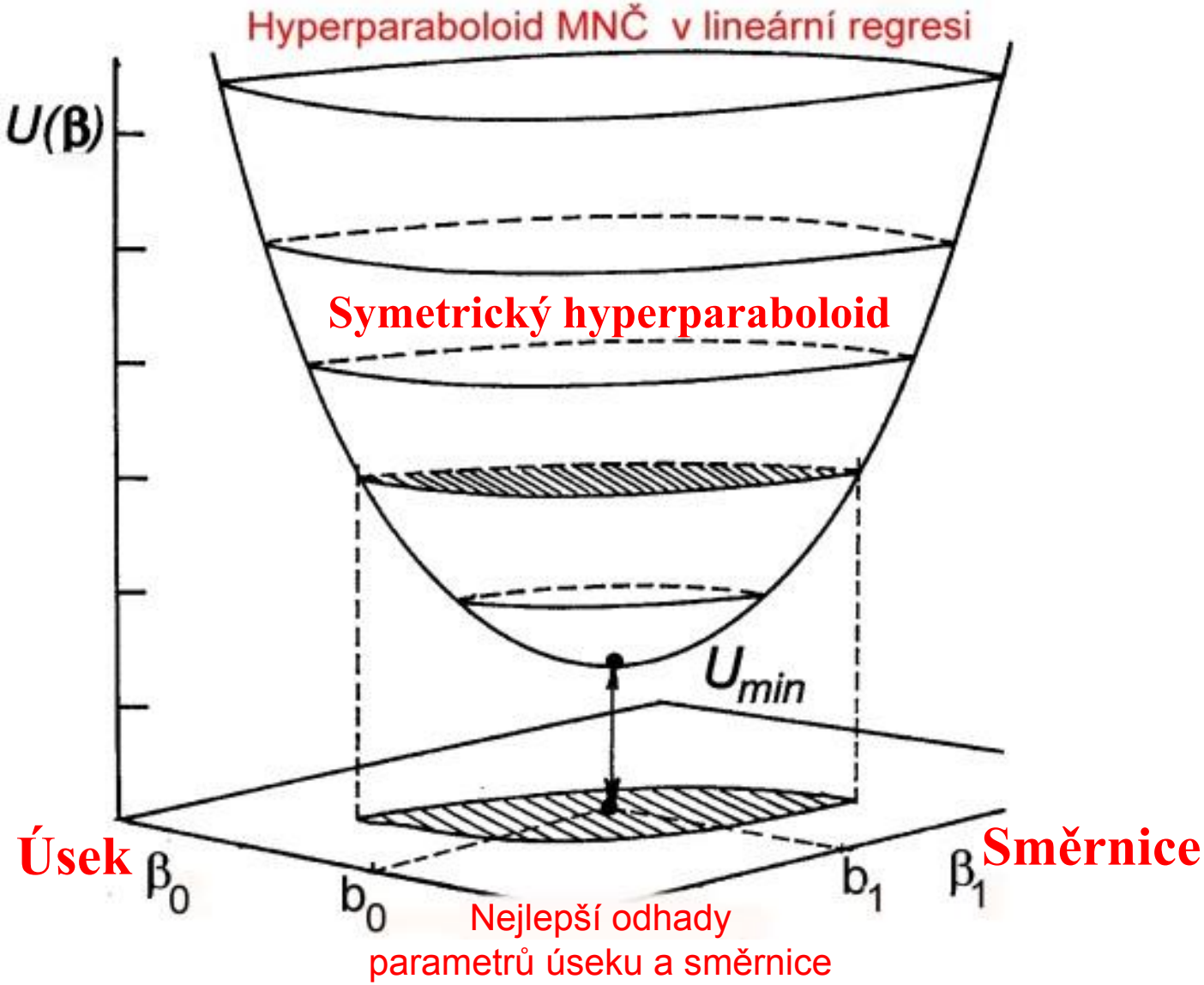
čili **hodnotou** ležící na spojité funkci (modelu) **nezávisle proměnné** x .

FORMULACE REGRESNÍHO MODELU

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix}}_{\mathbf{X}} \cdot \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_m \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

závisle proměnná — **nezávisle proměnná** — **regresní parametry** — **náhodná chyba**

$$\boxed{\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$$



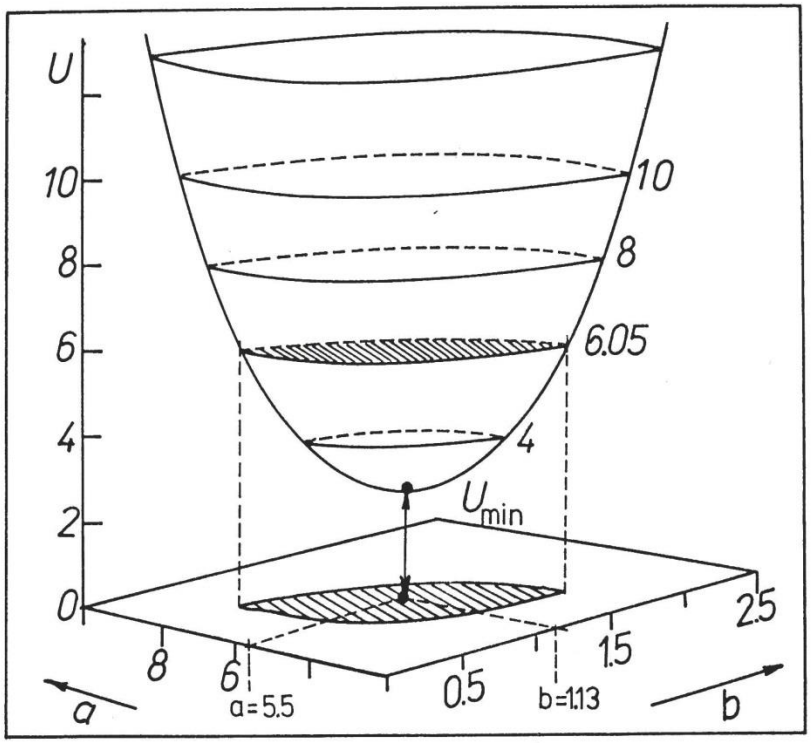
Účelová funkce U dosáhne minima pro nejlepší odhady parametrů úseku a směrnice

Lineární regresní model:

$$U = \sum_{i=1}^{n} w_i \left(y_{\exp,i} - y_{vyp,i}\right)^2 \approx \text{minimum}$$

$$\text{kde} \quad y_{vyp,i} = f(\boldsymbol{x}; \beta_1, \ldots, \beta_m)$$

Symetrický paraboloid v (m + 1)-rozměrném prostoru

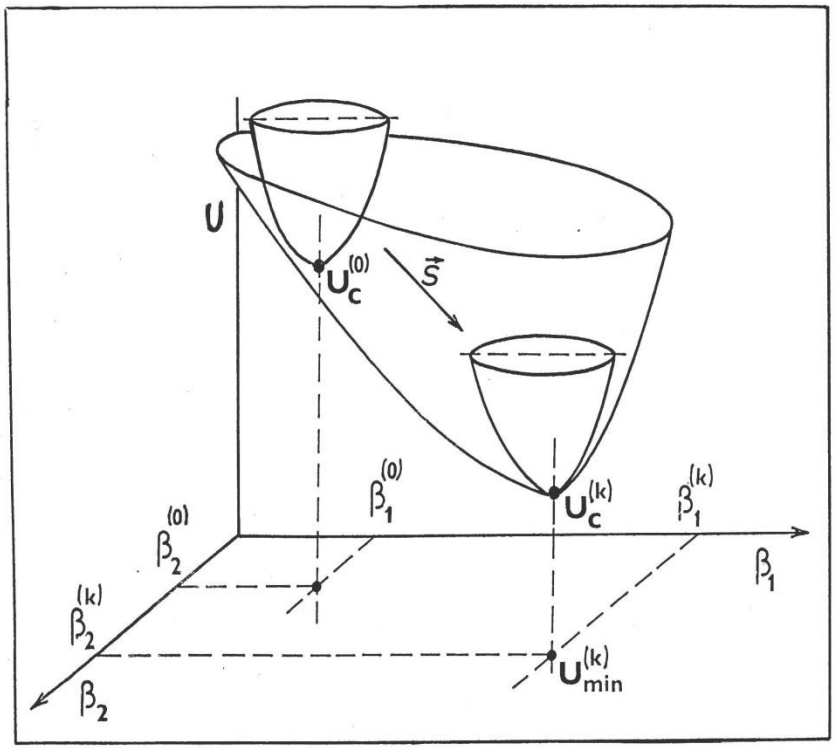


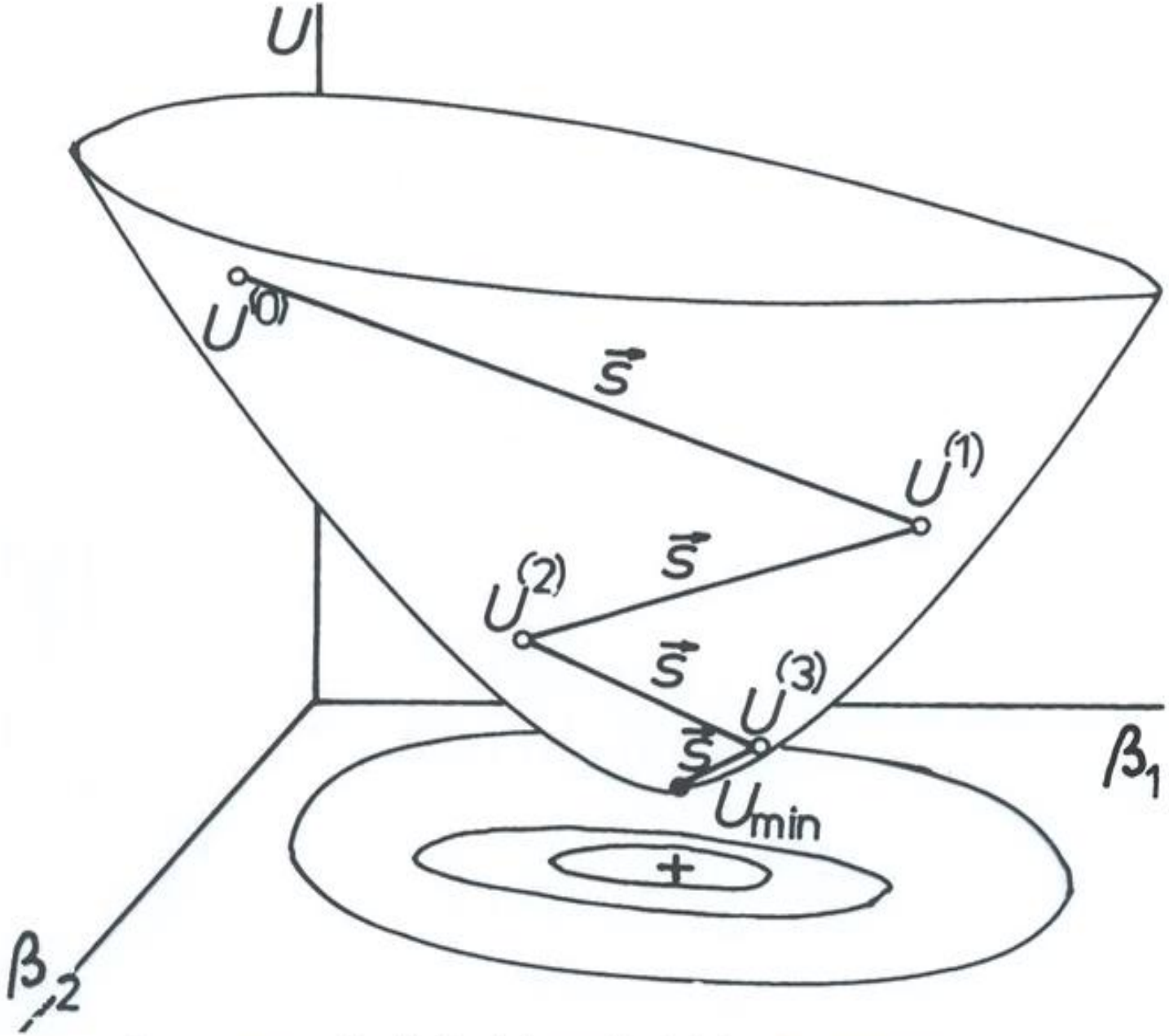
Nelineární regresní model:

$$U = \sum_{i=1}^{n} w_i \left(y_{\exp,i} - y_{vyp,i}\right)^2 \approx \text{minimum}$$

$$\text{kde} \quad y_{vyp,i} = f(\boldsymbol{x}; \beta_1, \ldots, \beta_m)$$

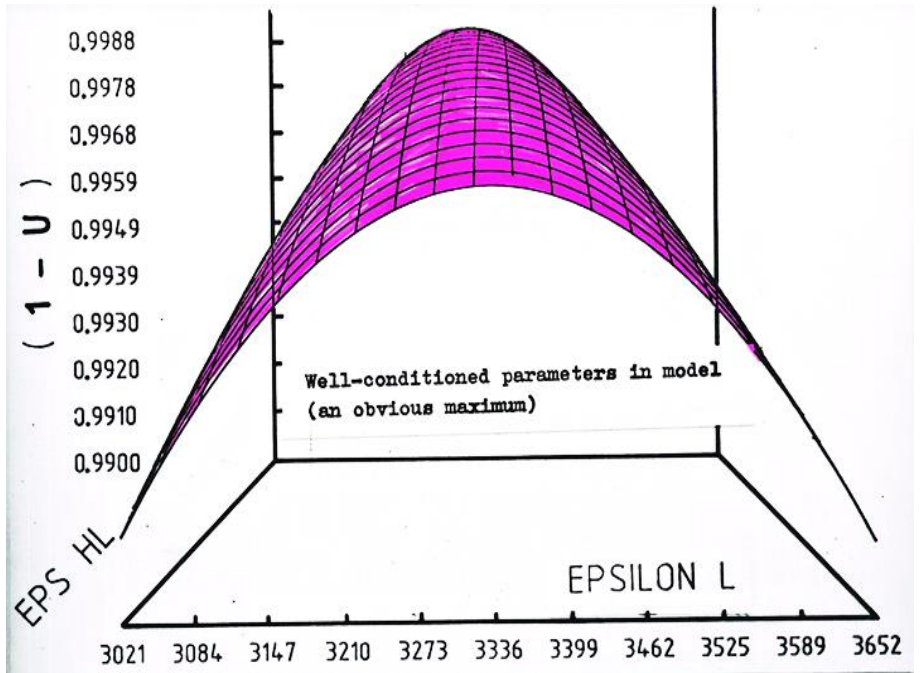
Eliptický hyperparaboloid v (m + 1)-rozměrném Eukleidovském prostoru

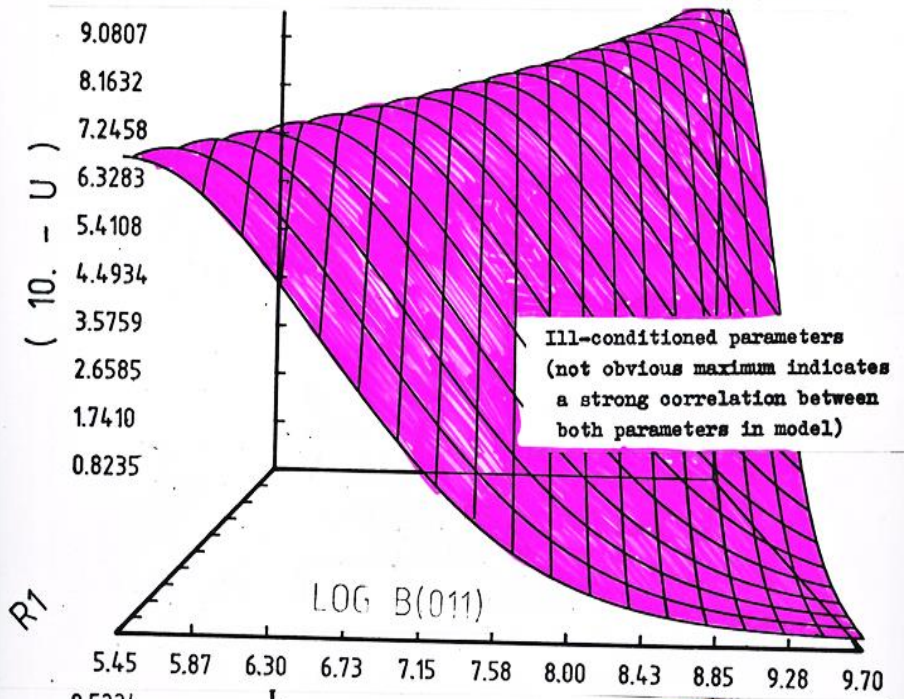


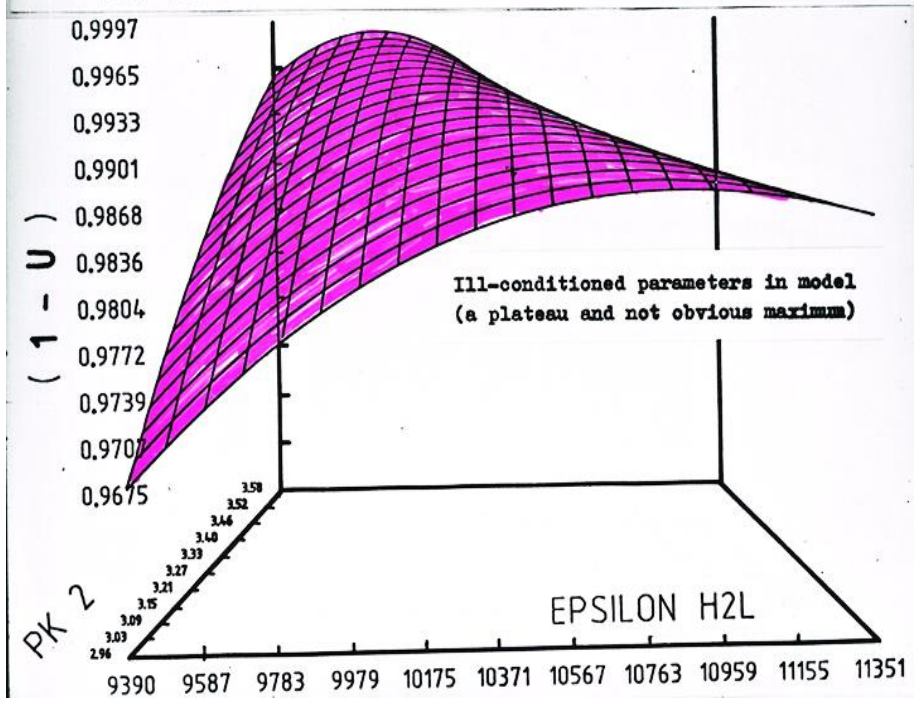


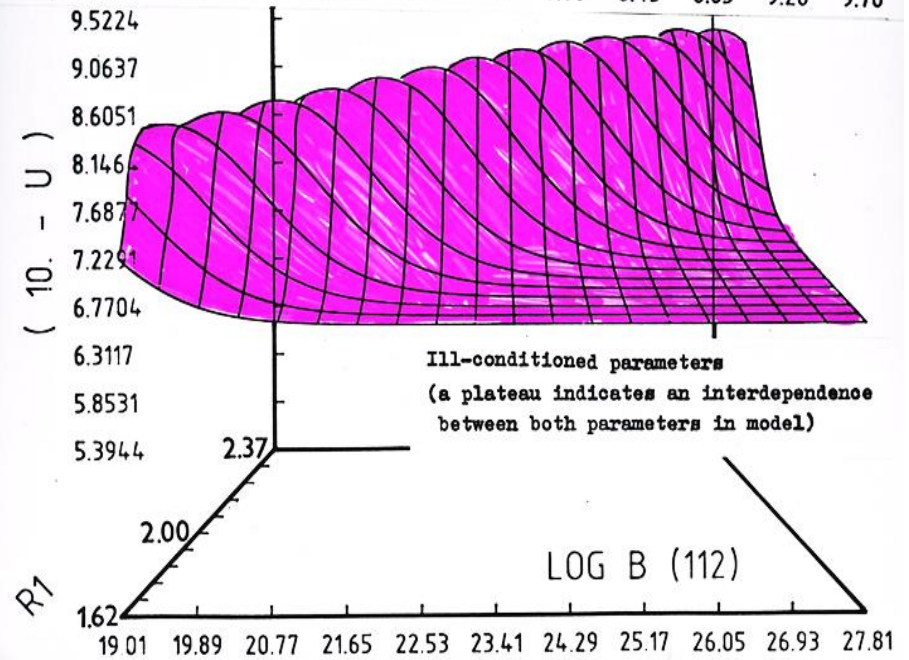
Interpretrace hledání minima eliptického hyperparaboloidu

Residual-square-sum function response surface $U = f(r_1, \log \beta_{011})$



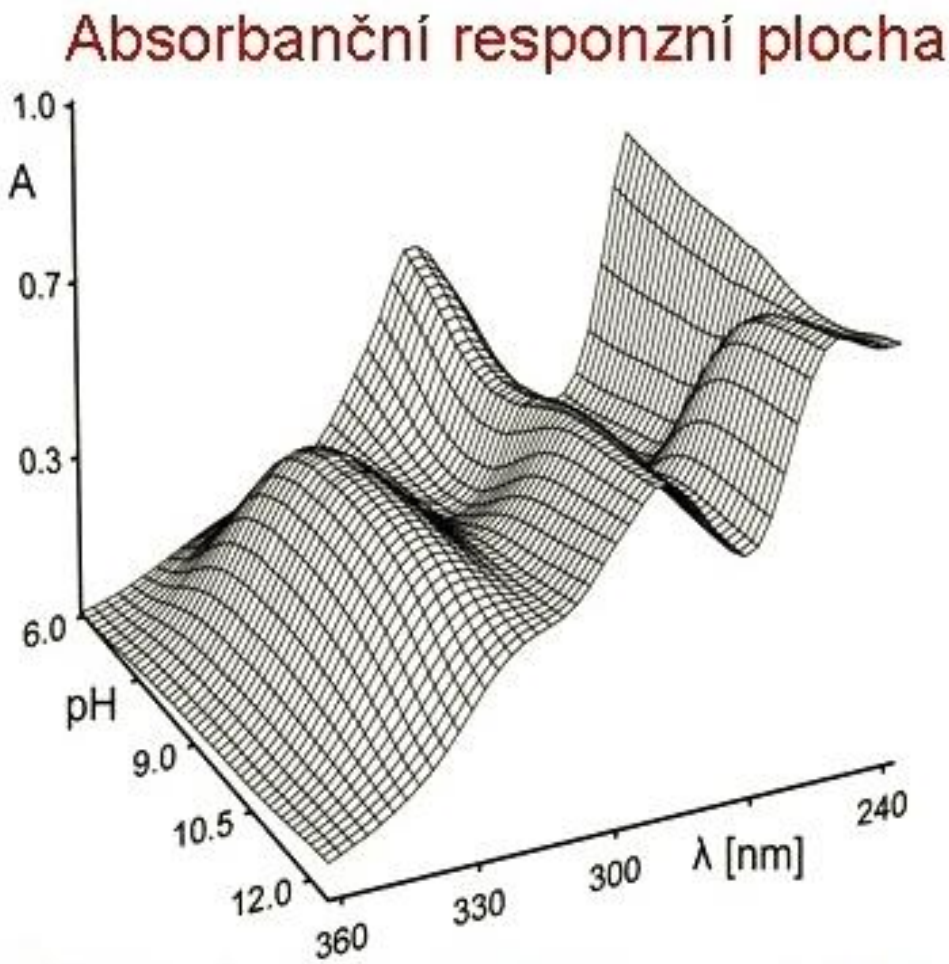


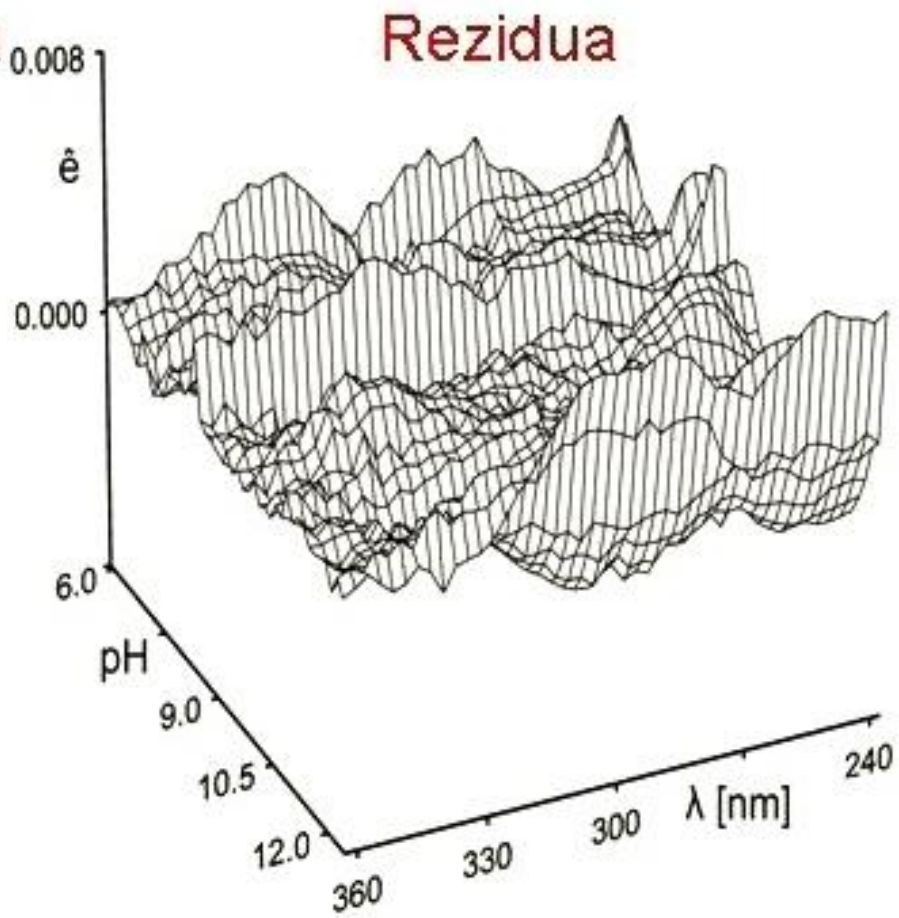




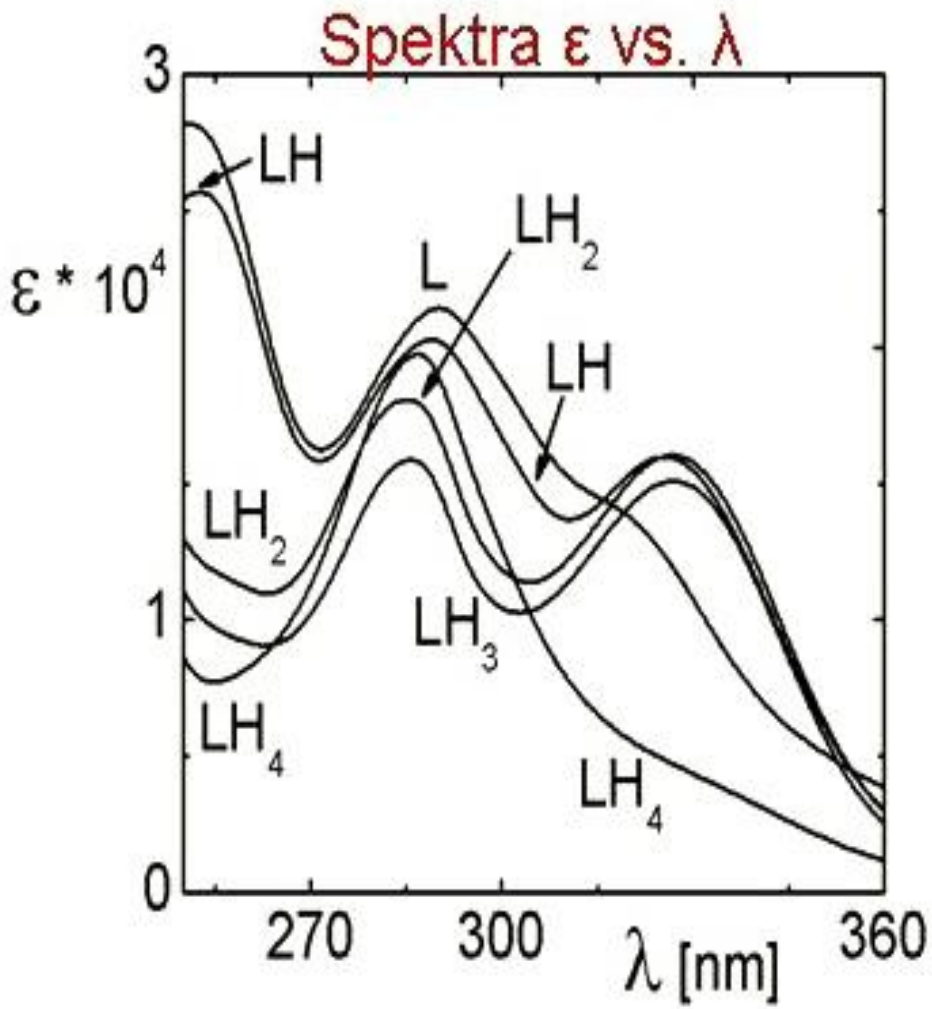
Input dat

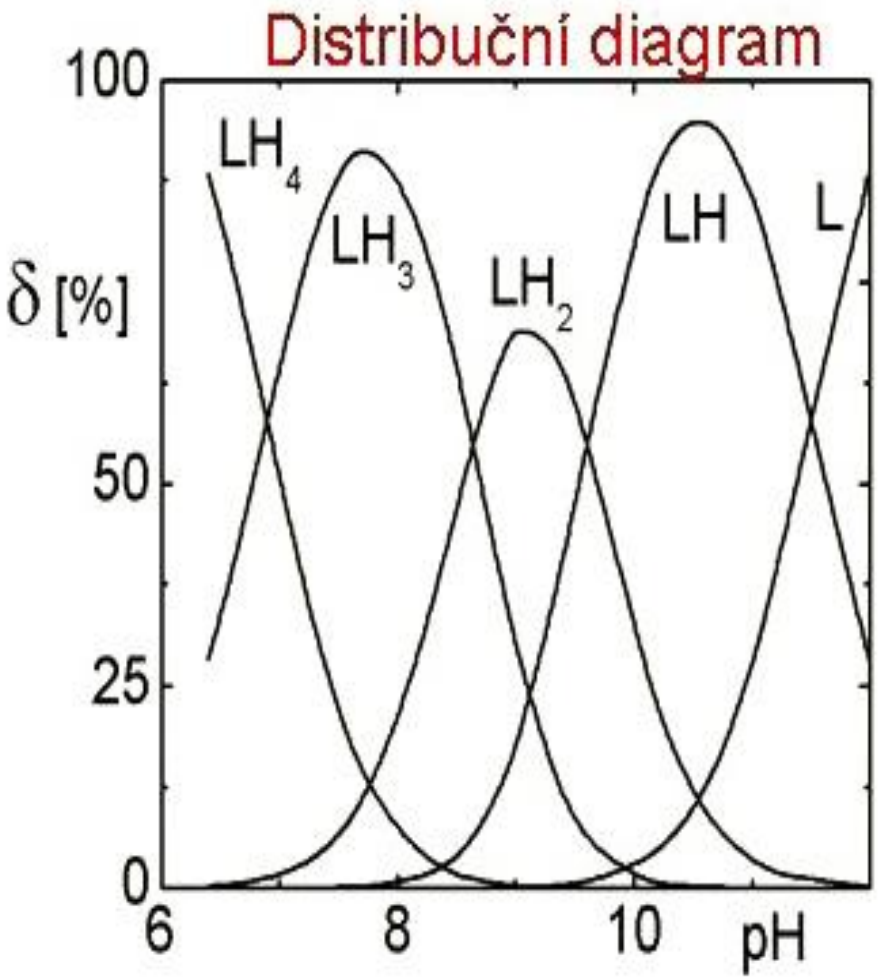
Grafická analýza reziduí

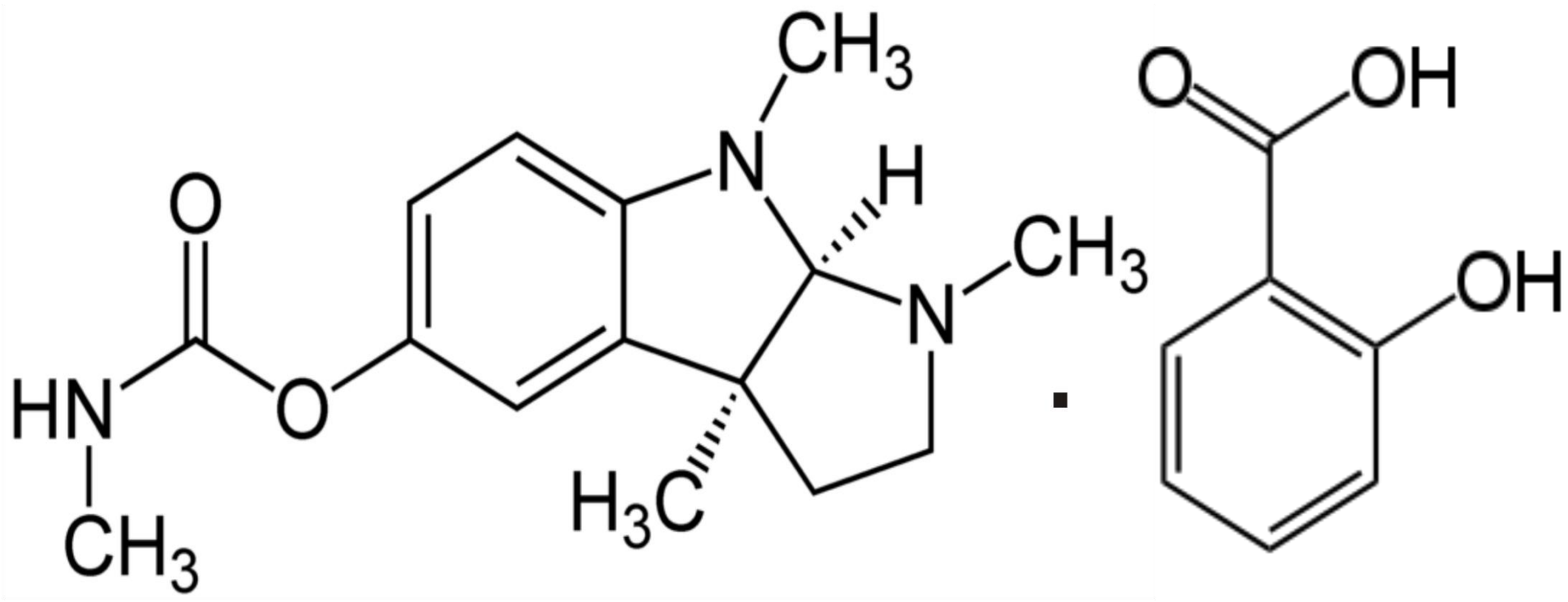




Grafické závěry o nalezených odhadech parametrů z navrženého protonačního (= nelineárního regresního) modelu

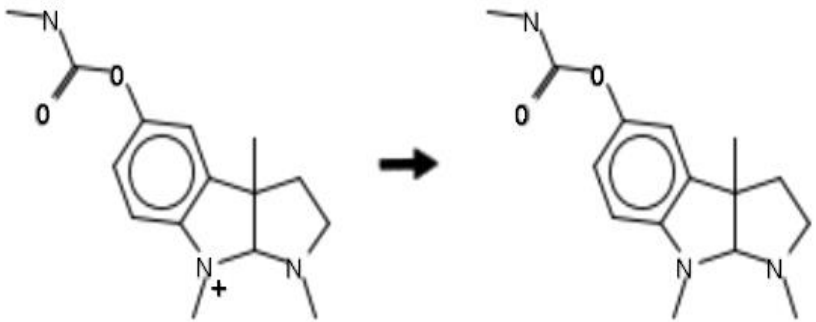




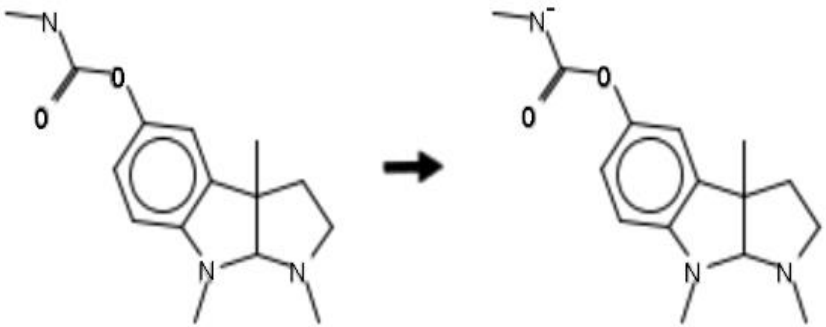


Physostigmine salicylate

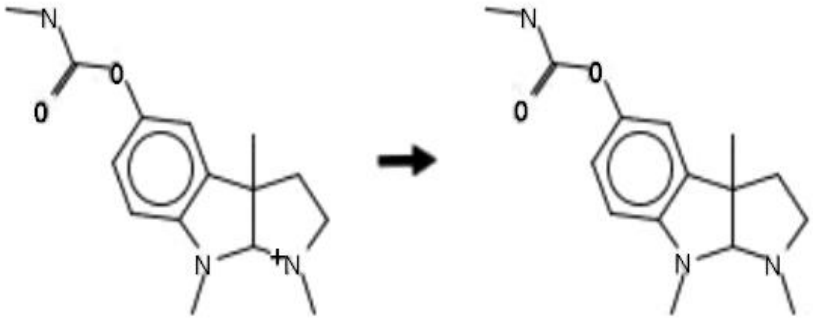
Predikce disociačních konstant kvantově chemickými programy



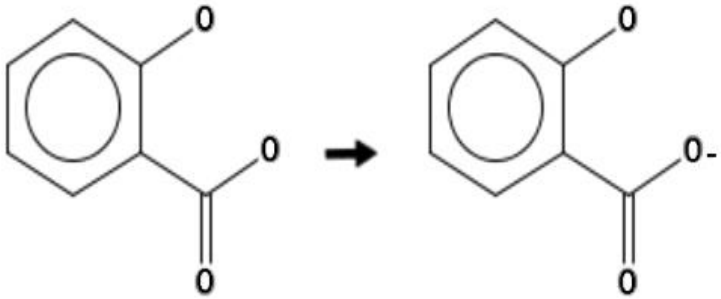
pK_{a2} is 2.71



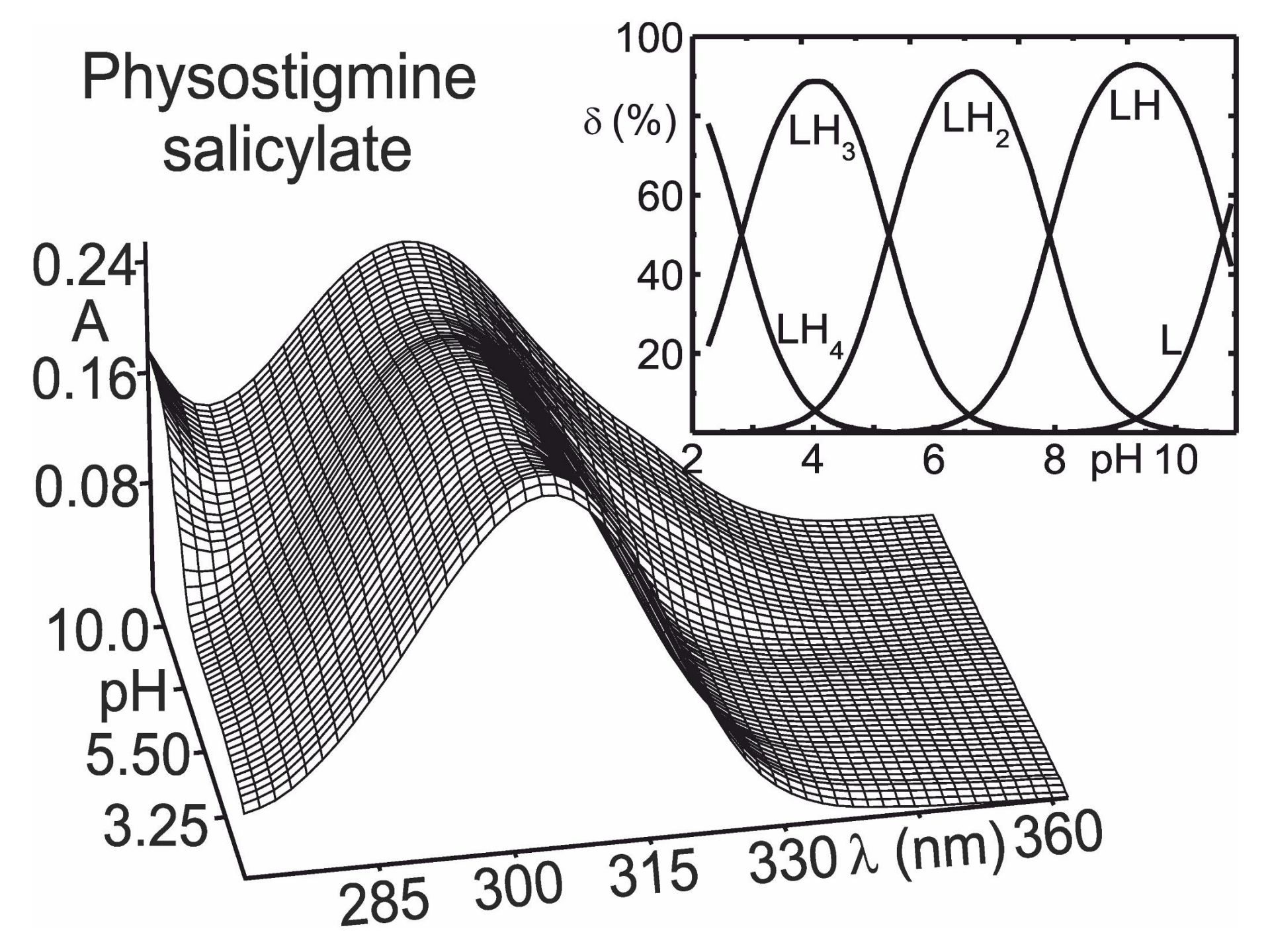
pK_{a4} is 11.42

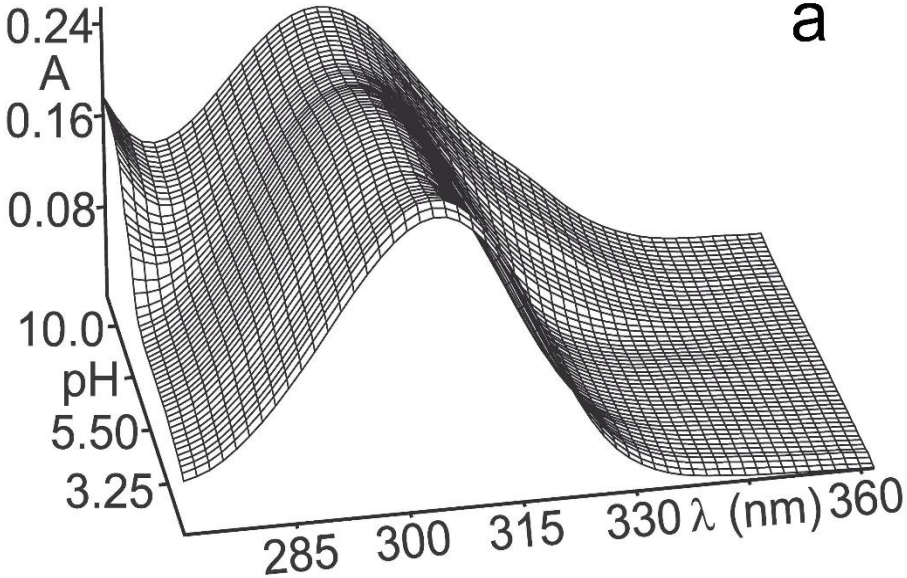


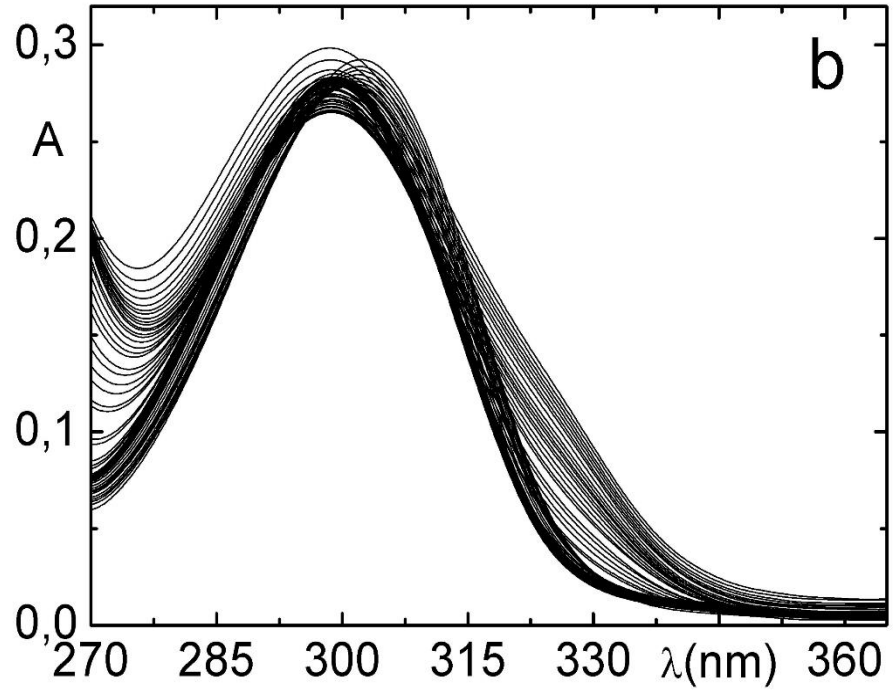
pK_{a3} is 8.23

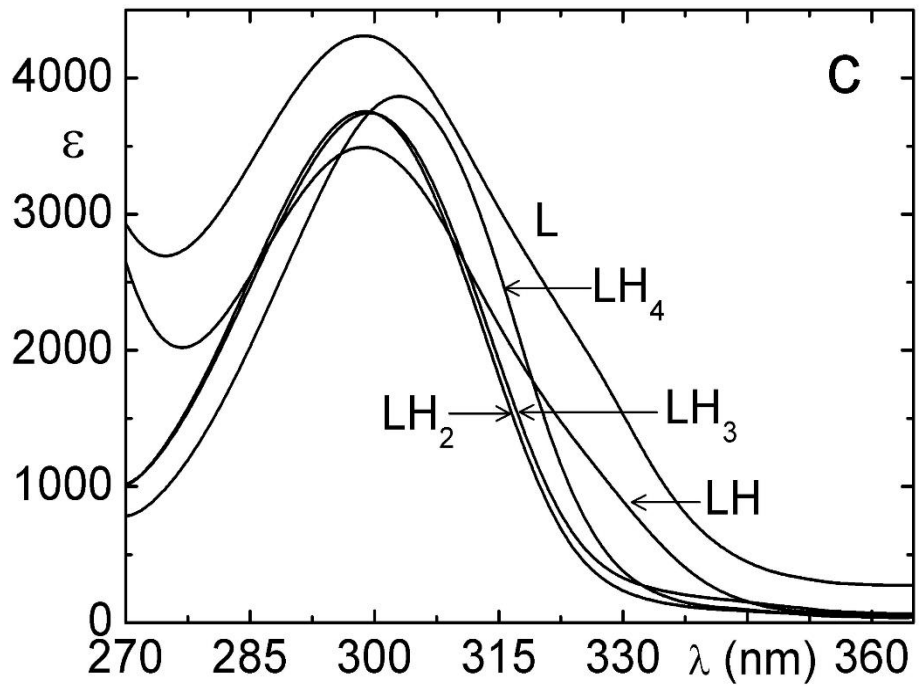


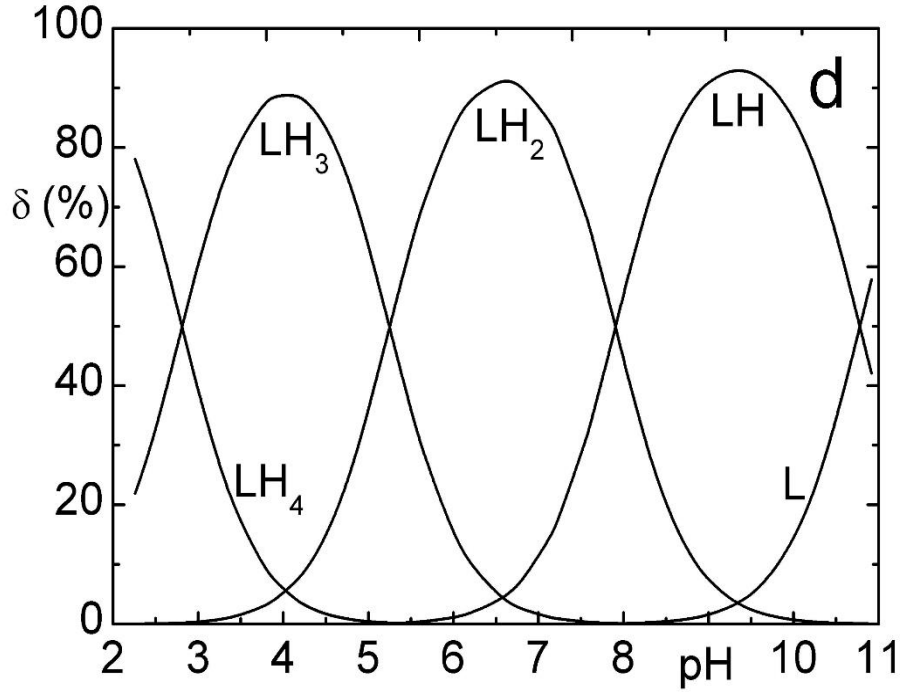
pK_{a1} is 3.08


Physostigmine salicylate
0.24
A
0.16
0.08
10.0
pH
5.50
3.25
285
300
315
330
λ (nm)
360
100
δ (%)
60
40
20
LH3
LH2
LH
LH4
L
2
4
6
8
pH
10


a
0.24
A
0.16
0.08
10.0
pH
5.50
3.25
285
300
315
330
λ (nm)
360


b
0,3
A
0,2
0,1
0,0
270
285
300
315
330
λ(nm)
360


c
4000
ε
3000
2000
1000
0
L
LH4
LH2
LH3
LH
270
285
300
315
330
λ (nm)
360

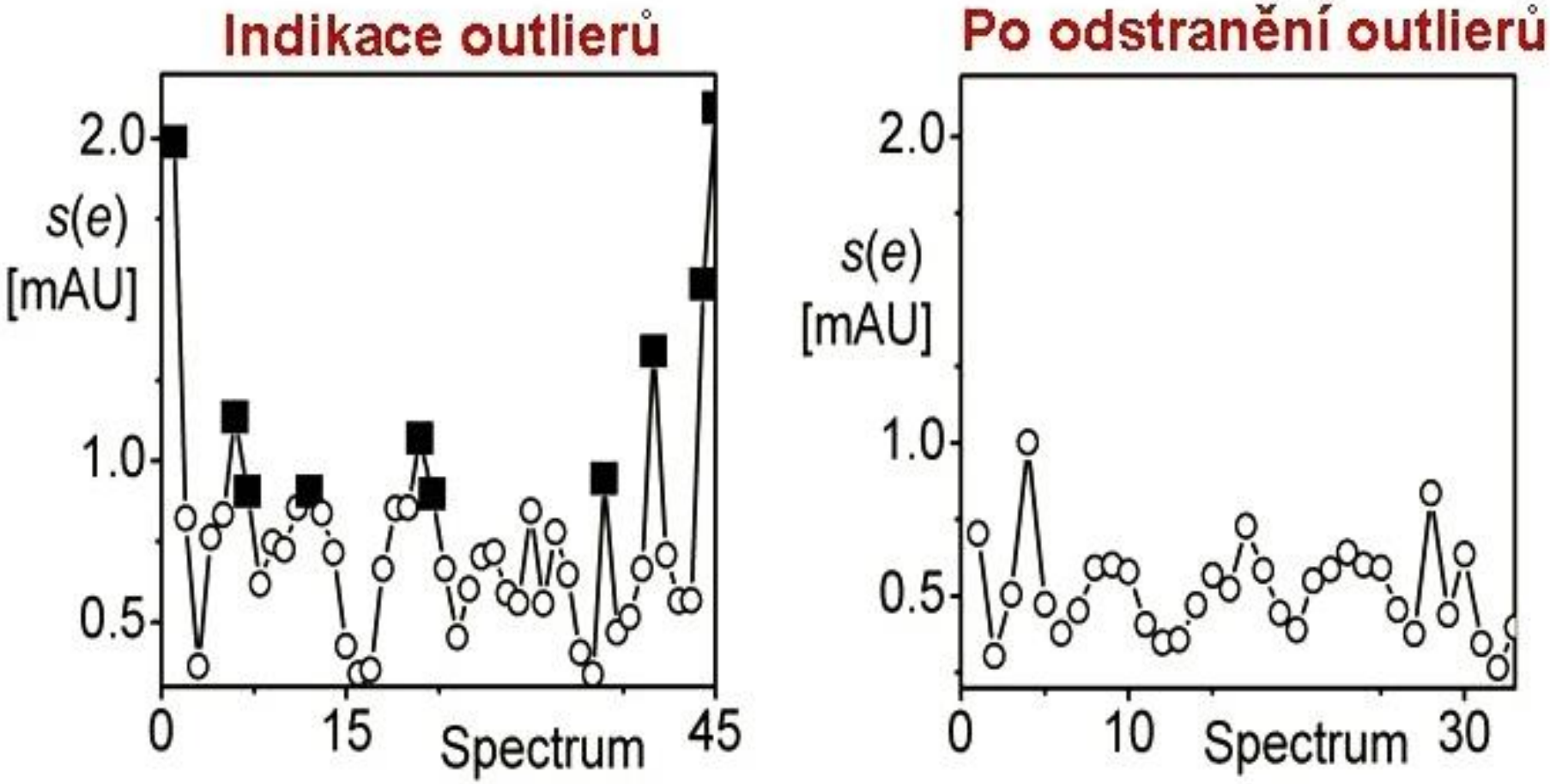

d
100
80
δ (%)
60
40
20
0
LH3
LH2
LH
LH4
L
2
3
4
5
6
7
8
pH
10
11

REGRESNÍ DIAGNOSTIKA

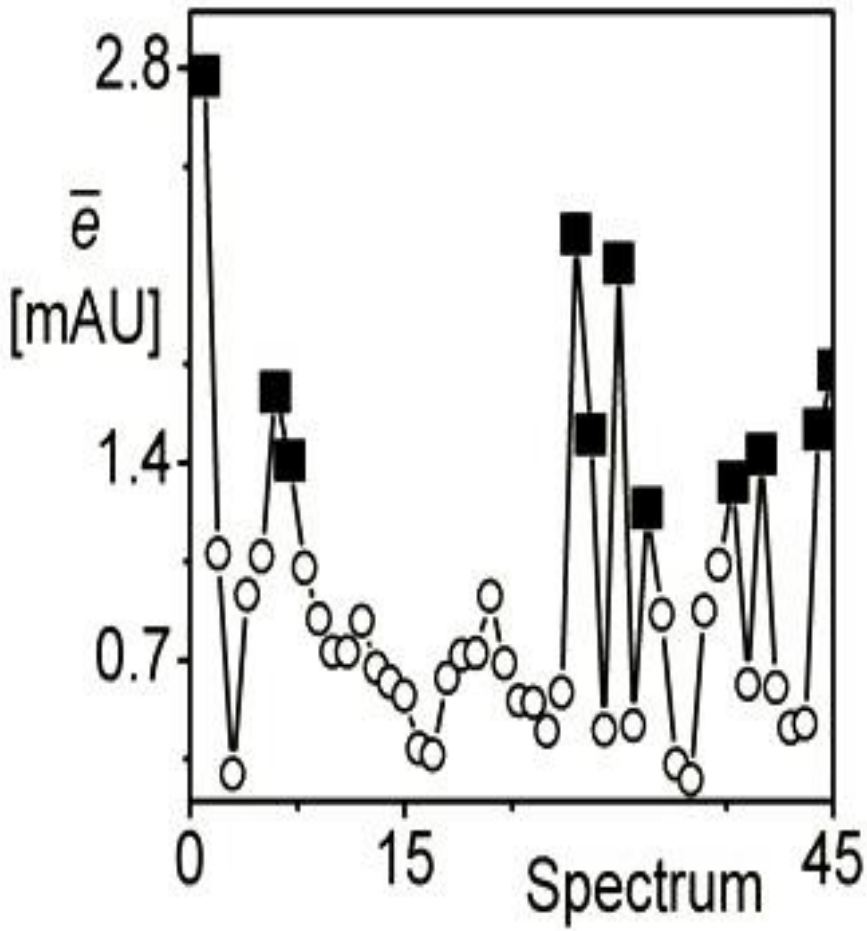
Vyšetřuje **regresní triplet,** což představuje

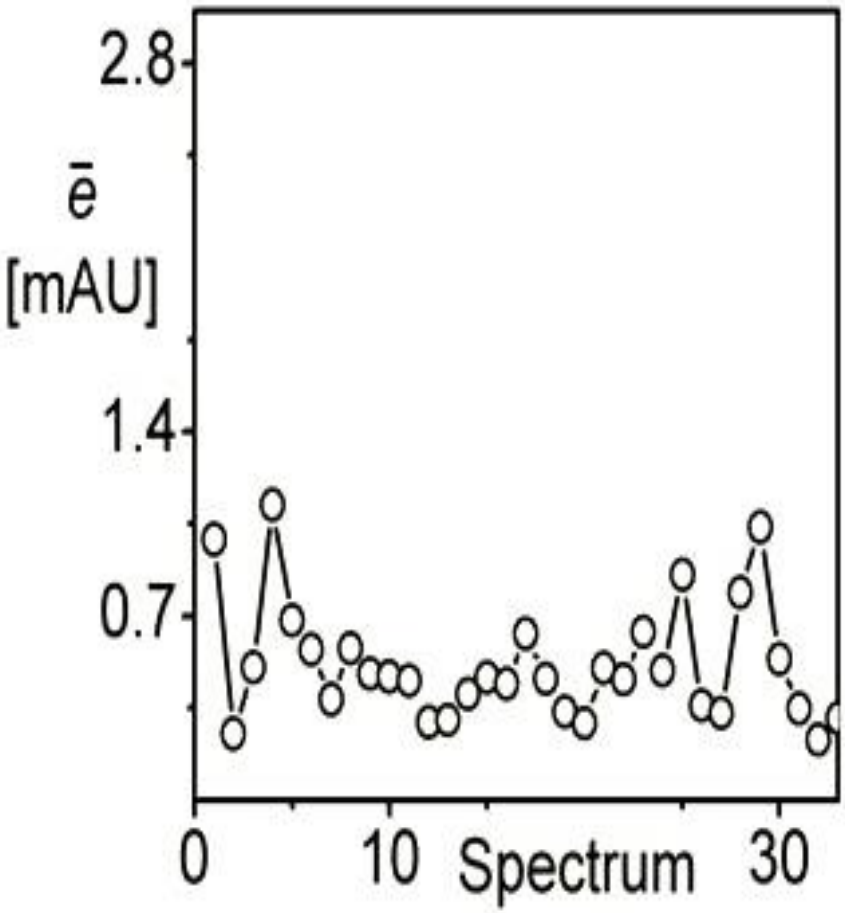
- **Kritiku dat** (zkoumá kvalitu dat pro navržený model)
- **Kritiku modelu** (zkoumá kvalitu modelu pro daná data)
- **Kritiku metody odhadu** (prověřuje splnění všech předpokladů požadovaných metodou MNČ)

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.

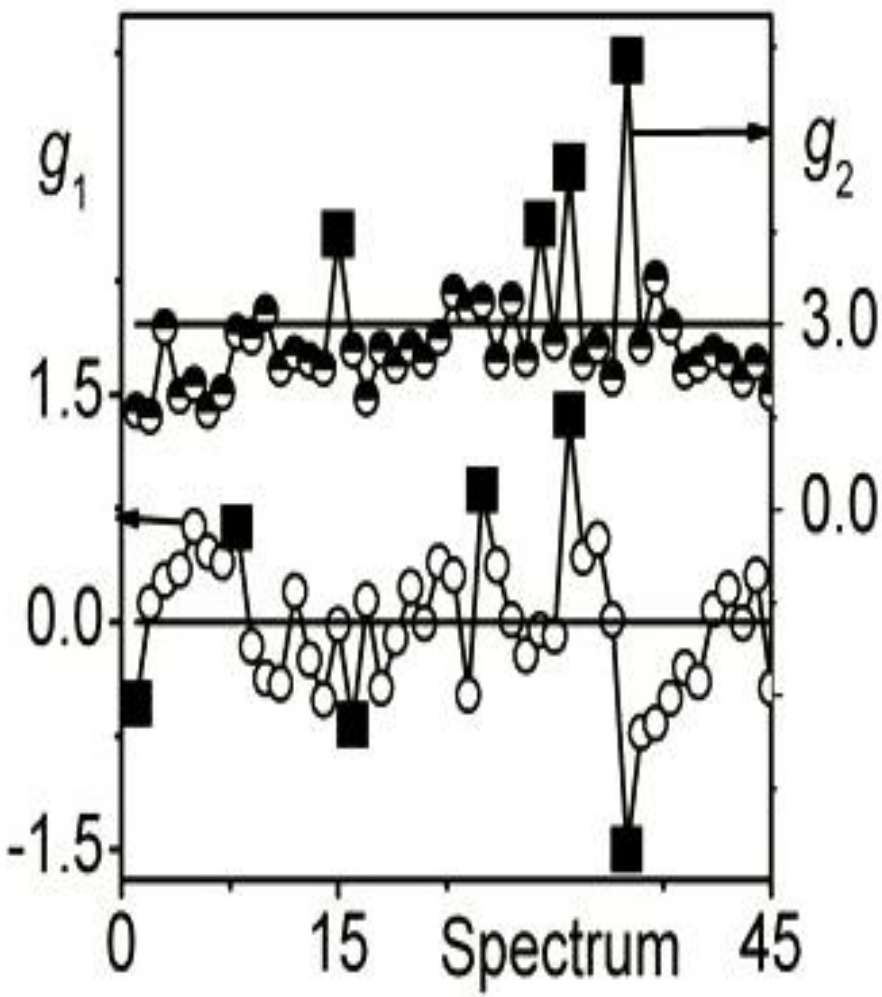


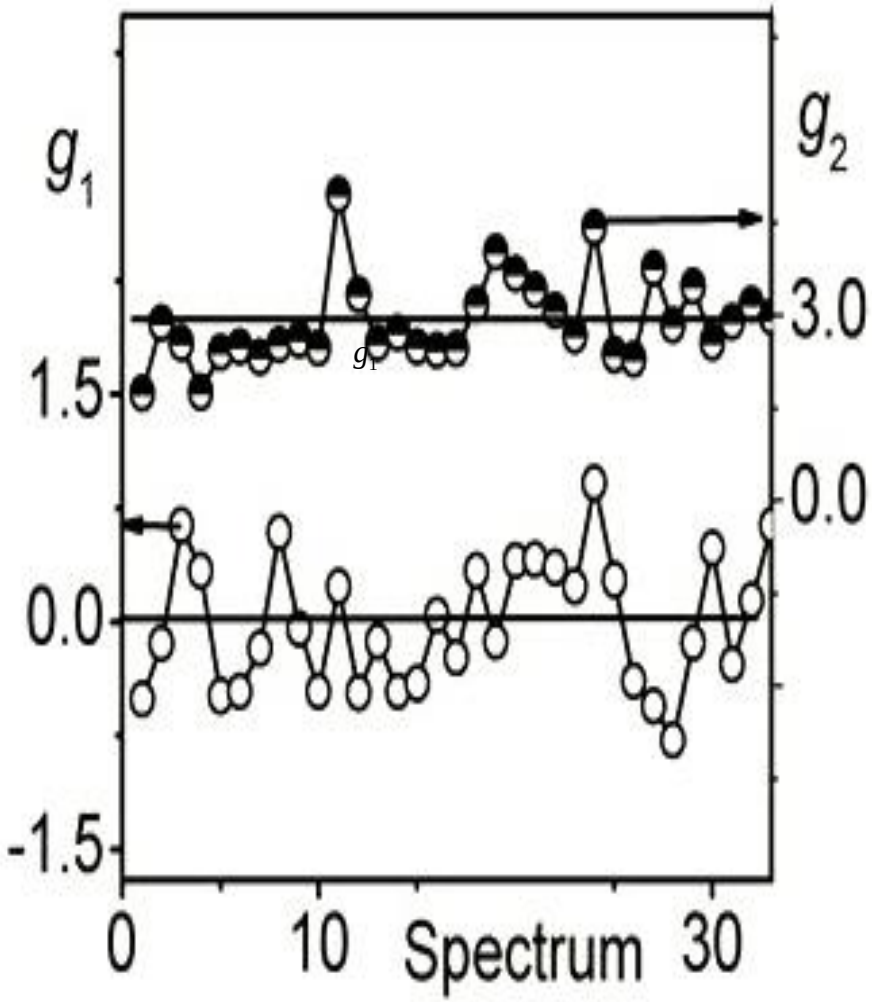
Míru těsnosti proložení je vhodné posuzovat dle velikosti reziduí a jejich rozdělení, a to především dle střední hodnoty reziduí a symetrie rozdělení (vyjádřené koeficientem šikmosti a špičatosti).





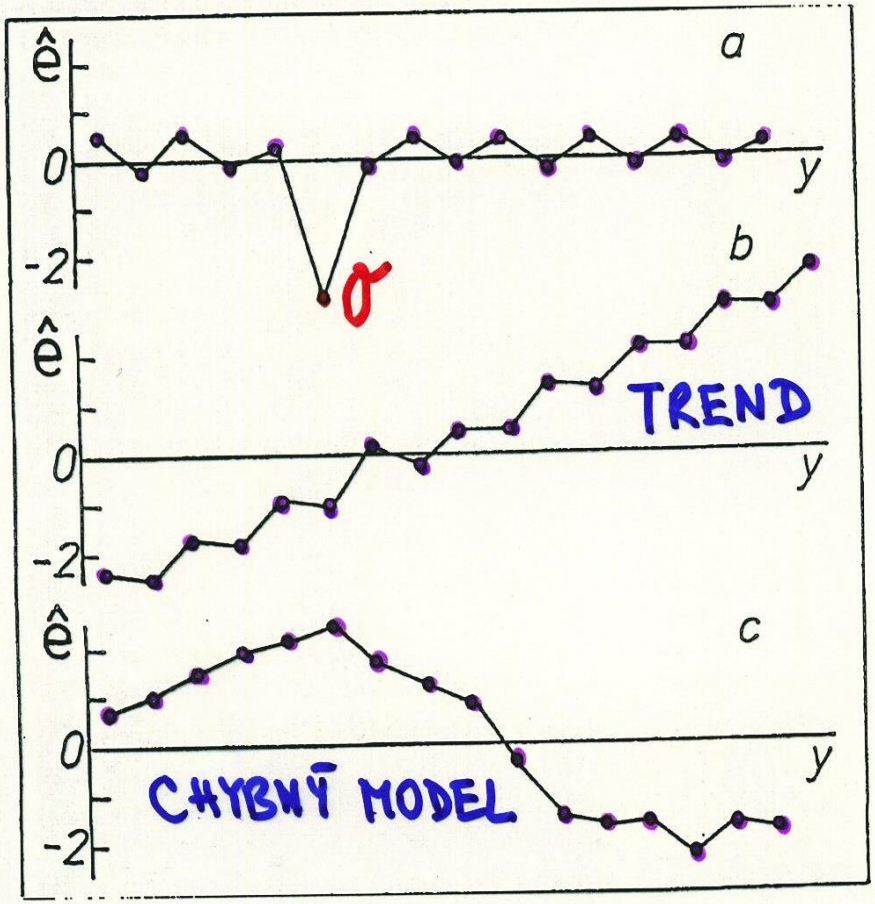
Symetrii a tvar rozdělení vystihuje koeficient šikmosti g_1 a špičatosti g_2) před a po odstranění outlierů.

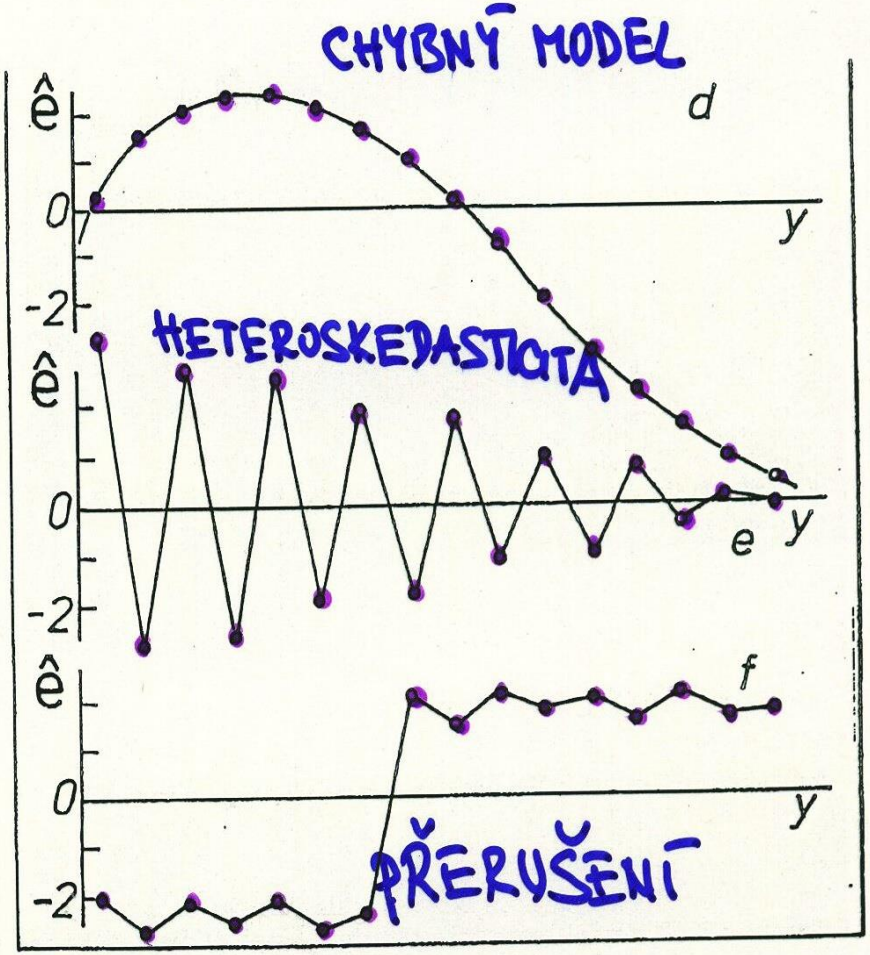




Statistická analýza klasických reziduí

1. Grafická analýza





2. Numerická analýza:

(a) **Střední hodnota reziduí** $E(\hat{e}) = 0$,

(b) **Průměrné reziduum** $|\bar{e}| = \frac{1}{n}\sum |\hat{e}_i| \approx \epsilon$,

(c) **Směrodatná odchylka střední hodnoty reziduí** $s(\hat{e}) \approx \epsilon$,

(d) **Koeficient šikmosti souboru reziduí** $g_1(\hat{e}) \approx 0$,

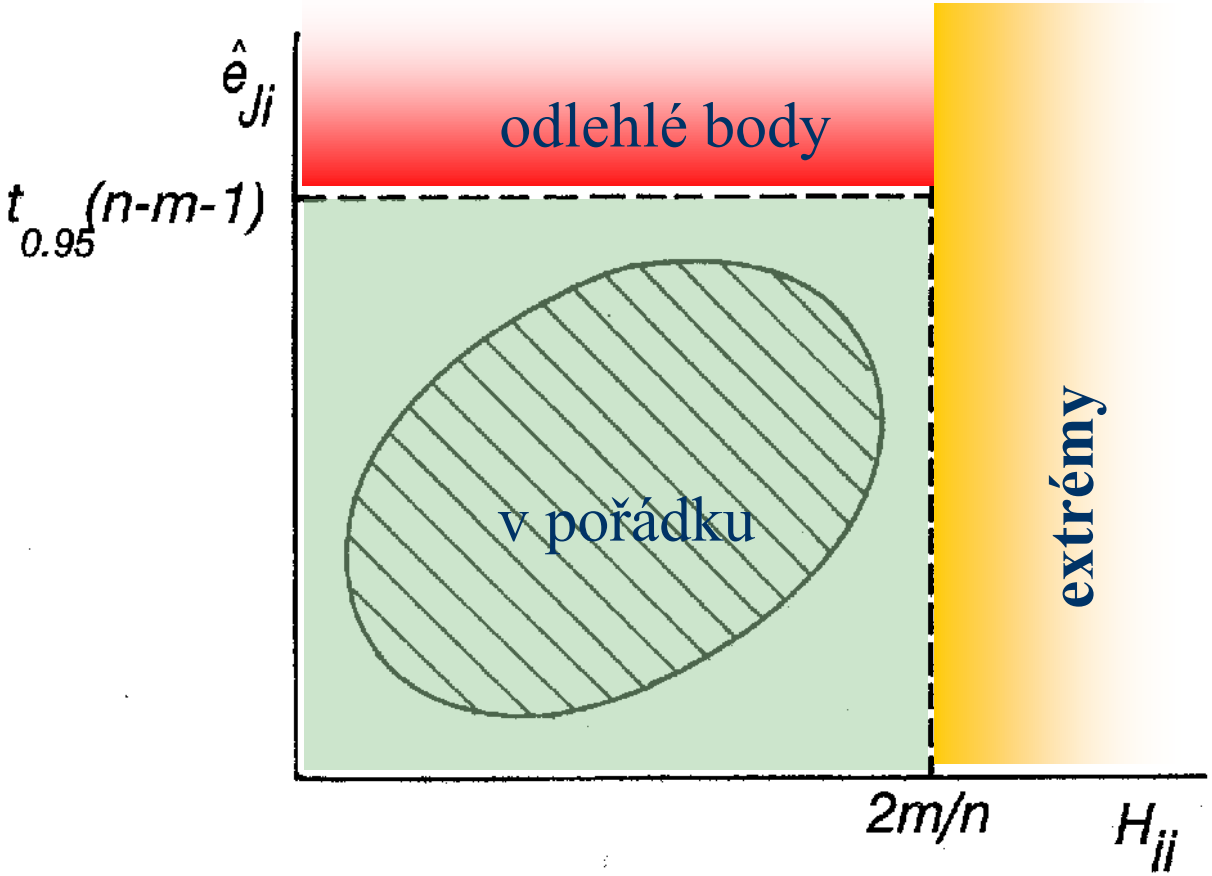
(e) **Koeficient špičatosti souboru reziduí** $g_2(\hat{e}) \approx 3$,

(f) **Pearsonův χ^2- test dobré shody:** χ^2_{exp} vs. $\chi^2_{crit.}$,

(g) **Hamiltonův R-faktor relativní těsnosti:** $R \approx 0.5\ \%$.

Williamsův graf vlivných bodů

$$e_{Ji} = e_{Si} \cdot \sqrt{\frac{n-m-1}{n-m-e_{Si}}}$$



$$e_{Si} = \frac{e_i}{\sigma\sqrt{1-H_{ii}}}$$

***i*-tý diagonální prvek projekční matice H**

KRITÉRIA PRO HLEDÁNÍ A ROZLIŠENÍ NEJLEPŠÍHO REGRESNÍHO MODELU

Střední kvadratická chyba predikce (MEP)

$$MEP = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i^2}{(1-H_{ii})^2}$$

e_i^2 čtverec reziduí modelu

H_{ii} *i*-tý diagonální prvek projekční matice **H**

Akaikovo informační kritérium (AIC)

$$AIC = n \cdot \ln\left(\frac{RSC}{n}\right) + 2m$$

RSC reziduální součet čtverců

m počet parametrů

Pravidlo: Čím je AIC (nebo MEP) menší, tím je model vhodnější.

POSTUP VÝSTAVBY REGRESNÍHO MODELU

1. Kvalita nalezených odhadů parametrů

a) Podle **intervalů spolehlivosti** (čím menší interval spolehlivosti, tím lépe)

$$\beta_j = b_j \pm \sqrt{C_{mm} \cdot m \cdot s^2 \cdot F_{1-\alpha;m;n-m}}$$

b) Podle **rozptylů parametrů**, kde pro kvalitní odhad musí platit

$$2 \cdot \sqrt{D(b_j)} < |b_j|$$

(3) ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

Parametr	Odhad	Směrodatná odchylka	Test H0: B[j] = 0 vs. HA: B[j] <> 0 t-kriterium	hypoteza H0 je	Hlad. výz.
B[0]	-7.2666E-02	1.3791E-01	-5.2692E-01	Akceptována	0.608
B[1]	-6.8505E-01	1.9165E-01	-3.5746E+00	Zamítnuta	0.004
B[2]	8.9619E-01	1.6072E-01	5.5761E+00	Zamítnuta	0.000
B[3]	8.3769E-01	1.3322E-01	6.2879E+00	Zamítnuta	0.000

2. Kvalita dosažené těsnosti proložení

a) Podle **reziduálního rozptylu *s*(*y*)**.

b) Podle **regresního rabatu *D*** (= koeficient determinace v %: čím více se blíží 100 %, tím lepší je proložení).

3. Vhodnost navrženého modelu

Akaikovo informační kritérium *AIC* (čím je menší nebo zápornější, tím vhodnější je navržený model).

Střední kvadratická chyba predikce *MEP* (čím je MEP menší, tím je predikční schopnost navrženého modelu lepší).

(4) STATISTICKÉ CHARAKTERISTIKY REGRESE:

Vícenásobný korelační koeficient, R	: 9.9858E-01
Koeficient determinace, R^2	: 9.9716E-01
Predikovaný korelační koeficient, Rp^2	: 9.9527E-01
Střední kvadratická chyba predikce, MEP	: 2.1014E-01
Akaikeho informační kritérium, AIC	:-3.6180E+01

4. Predikční schopnost modelu

Střední kvadratická chyba predikce *MEP* (čím je MEP menší, tím je predikční schopnost navrženého modelu lepší).

5. Kvalita experimentálních dat

a) Na základě analýzy rozličných druhů **reziduí**.

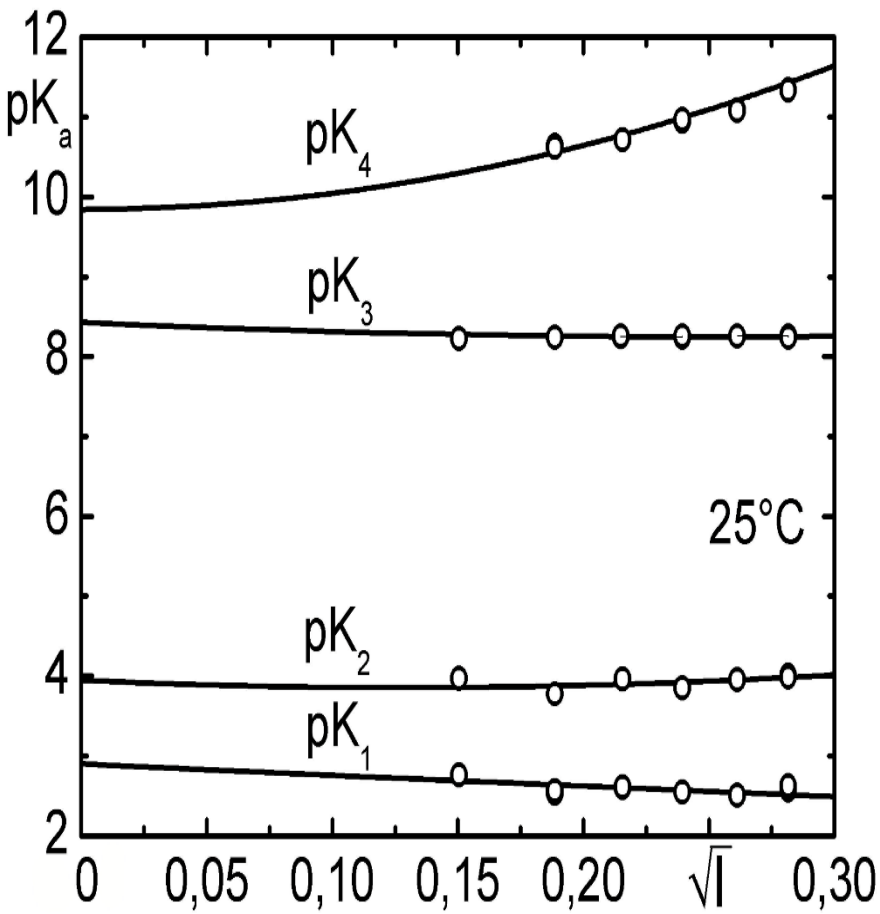
b) Na základě **Indikace vlivných bodů**

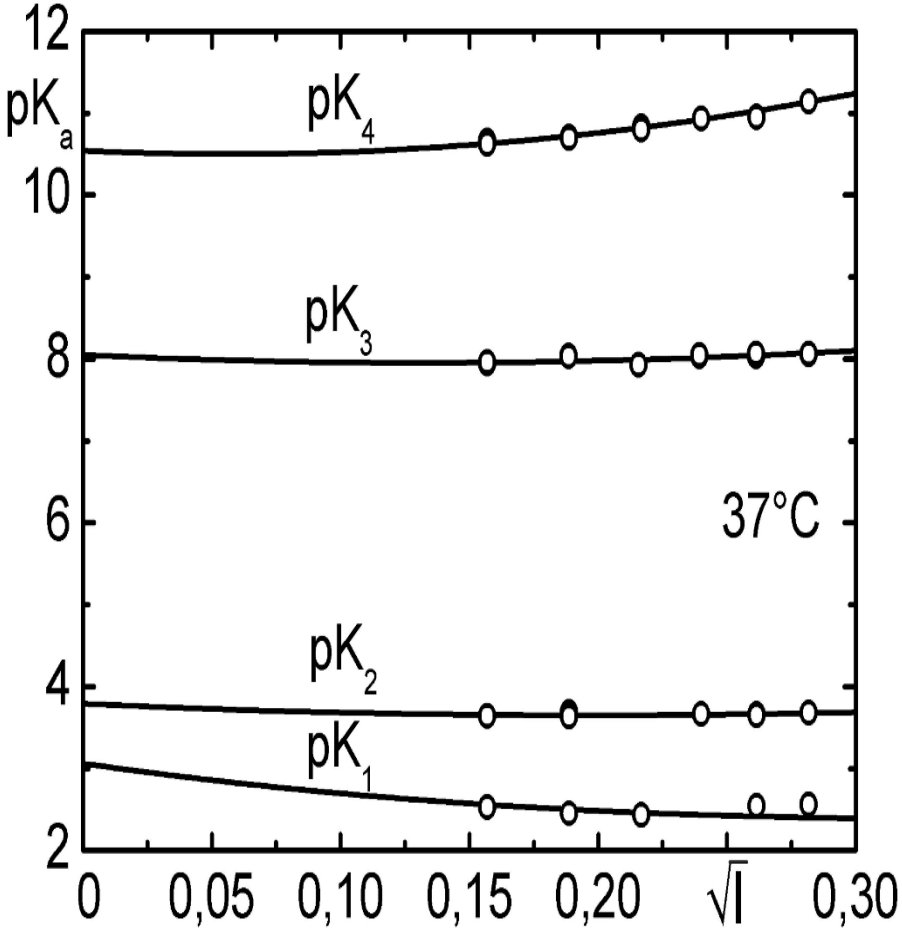
(Jackknife rezidua, standardizovaná rezidua, normovaná rezidua, predikovaná rezidua, rekurzivní rezidua,

Cookova vzdálenost, diagonální prvky projekční matice a věrohodnostní vzdálenosti).

6. Testy regresního tripletu (Data + Model + Metoda):

- 6.1 Fisher-Snedecorův test celkové regrese,
- 6.2 Scottovo kritérium multikolinearity,
- 6.3 Cook-Weisbergův test heteroskedasticity,
- 6.4 Jarque-Berrův test normality reziduí,
- 6.5 Waldův test autokorelace,
- 6.6 Znaménkový test reziduí.





Extrapolací na nulovou iontovou sílu se vyčíslí thermodynamická disociační konstanta, efektivní průměr iontu a vysolovací konstanta dle Debye-Hückelova zákona.

Výsledky z experimentálních operací jsou vždy přibližné pro omezenou přesnost přístrojů.

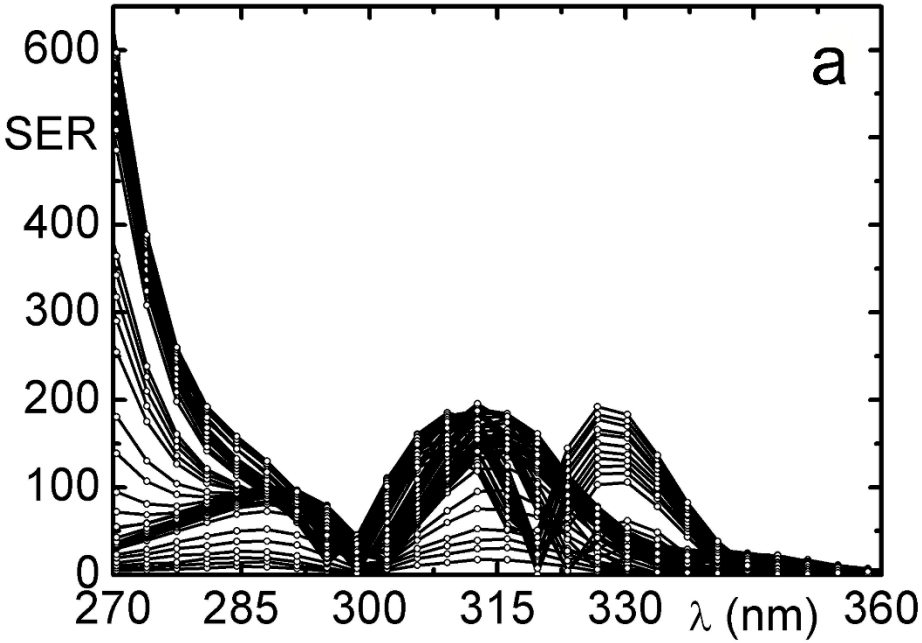
Mírou "experimentálního šumu" v experimentu je poměru **SNR** čili **signálu k šumu:** poměr maximálního signálu vůči maximální hodnotě šumu.

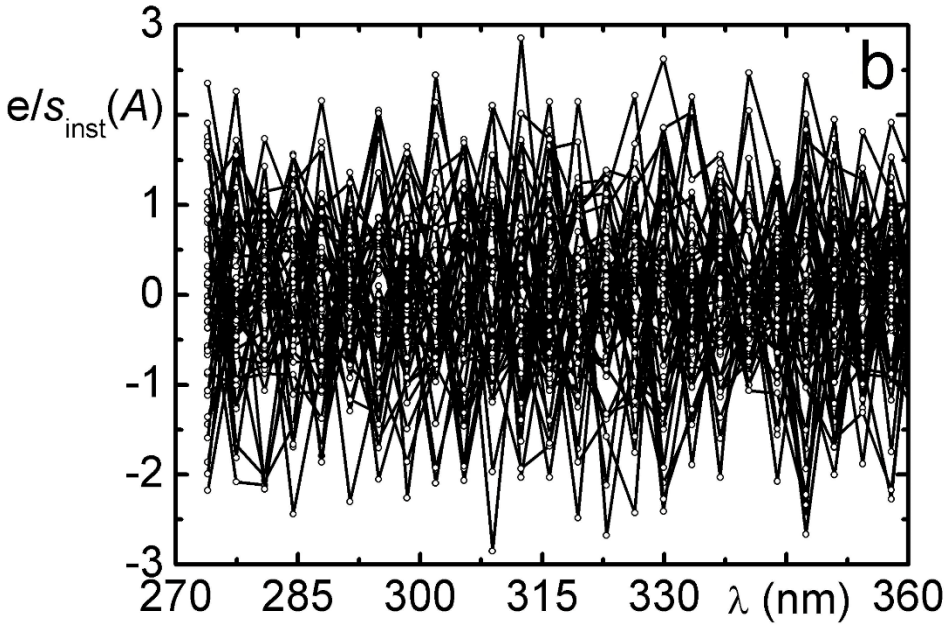
Alternativou k SNR je poměr **SER** čili **signál k chybě:**

$$\mathrm{SER} = \Delta / s_{\mathrm{inst}}(A)$$

čili poměr max. rozdílu absorbancí pro *j*-tou vlnovou délku *i*-tého spektra $\Delta_{ij} = A_{ij} - A_i$ podělený šumem, kde A_i, představuje absorbanci kyselé formy léčiva.

Poměr $\mathrm{SER} = \Delta / s_{\mathrm{inst}}(A)$ v závislosti na vlnové délce pro spektra vyšetří, zda jsou rezidua podobné velikosti jako šum, čímž se prokáže těsnost proložení nejlepším vypočteným spektrem.





Vlevo: Graf malých změn absorbance testuje poměr SER = $(A_{ij} - A_i)/s_{inst}(A)$ v závislosti na vlnové délce λ pro prvky spektra při změně pH léčiva, kde A_i je limitující spektrum kyselé formy léčiva.
SER se porovná s mezní hodnotou SER, zda změny absorbance $A_{ij} - A_i$ při změně pH jsou měřitelné a dostatečně větší než instrumentální šum $s_{inst}(A)$.

Vpravo: Těsnost proložení spektry posoudí graf reziduí vůči instrumentální směrodatné odchylce absorbance

$$e/s_{inst}(A)$$

v závislosti na vlnové délce λ a pro všechna spektra.

Graf ukazuje kolikrát je reziduum větší než šum.

Propagace chyb při měření spekter

V experimentálních spektrech se **náhodné a systematické chyby** propagují v absorbanci, v nastavení vlnové délky a i v koncentraci léčiva a nelze je proto odlišit:

1) Systematické chyby v měření pH jsou známé.

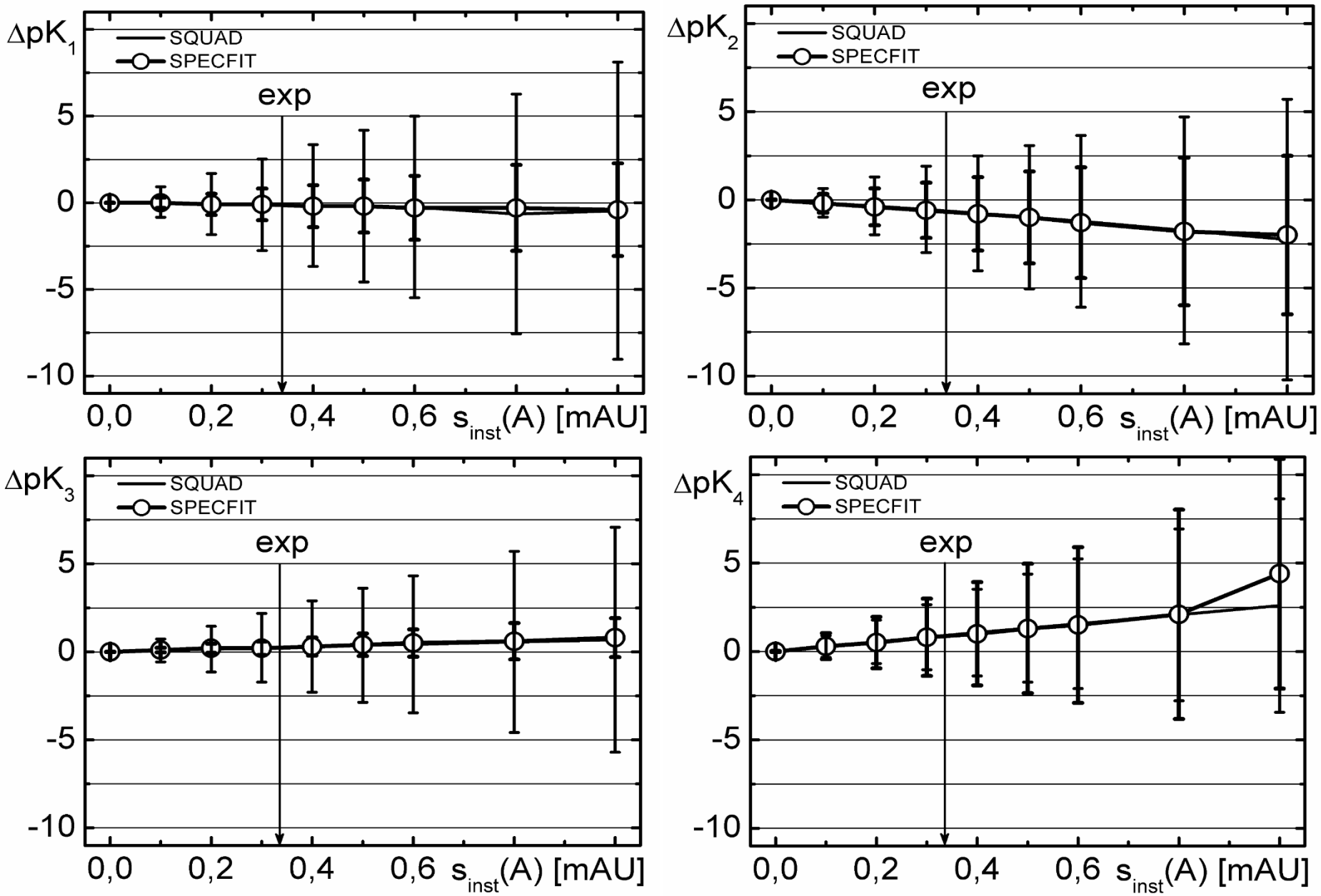
2) Barevné nečistoty v léčivu mohou mít acidobazický charakter.

3) Při kyselém pH lze některé částice oddělit z roztoku.

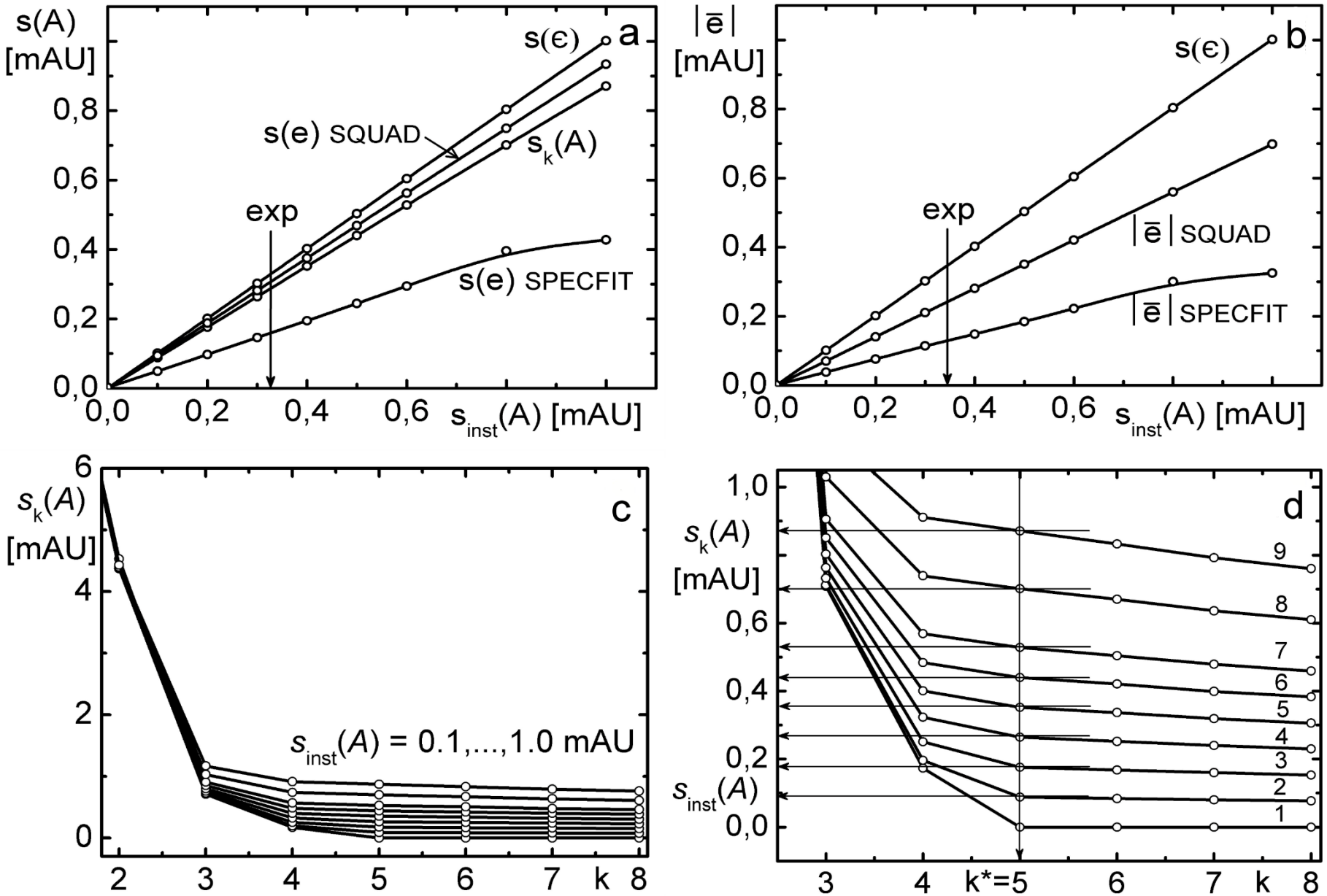
4) Při vyšších koncentracích léčiva se mohou tvořit oligomery a micely, které způsobí systematické chyby.

Všechny statistické testy v software jsou založeny na předpokladu, že **systematické chyby v datech nejsou.**

Správnost a přesnost odhadu pK v závislosti na velikosti šumu ve spektrech



Porovnání reziduí s náhodnými chybami na velikosti šumu ve spektru



Závěr:

1) Systematická chyba je u všech čtyřech odhadů disociačních konstant velmi malá, jedině pro pK_{a4} je vyšší.

2) V intervalu $s_{inst}(A)$ od 0,1 do 1,0 mAU jsou všechny čtyři disociační konstanty dostatečně přesné.

3) Přesnost $s(pK)$ na hladině šumu $s_{inst}(A)$ lze vyjádřit lineárním modelem $s(pK) = \beta_0 + \beta_1 s_{inst}(A)$, a úsek $\beta_0 = 0$ byl vesměs statisticky nevýznamný.

4) Instrumentální chyba $s_{inst}(A)$ mírně ovlivňuje Cattellův indexový graf vlastních čísel $s_k(A) = f(k)$, užívaný k odhadu počtu světlo-absorbujících částic rovnovážné směsi.

5) Program SQUAD(84) vyniká proti SPECFITu bohatší statistickou analýzou reziduí, která slouží k prokázání věrohodnosti nalezených odhadů disociačních konstant.

http://meloun.upce.cz

Domů - Internet Explorer

http://meloun.upce.cz/

Soubor Úpravy Zobrazit Oblíbené položky Nástroje Nápověda

Převést Vybrat

Search Web Upgrade Your Toolbar Now Mail

Navrhované weby

Milan Meloun

Profesor analytické chemie and chemometrie na Univerzitě Pardubice.

Vítejte na osobních stránkách

Prof. RNDr. Milana Melouna, DrSc., profesora analytické chemie a chemometrie na Univerzitě Pardubice

Menu

- Domů
- Aktuality
- Životopis
- Výuka
- Výzkum
- Publikace
- Ke stažení
- Fotogalerie
- Užitečné odkazy

Výuka

Blok obsahuje studijní materiály jako slidy přednášek, sylaby ke zkoušce, zkušební otázky, vzorové semestrální práce studentů a doporučenou literaturu k

Výzkum

Blok obsahuje oddíl **Projekty** s názvy grantů a projektů, dále tituly PhD dizertací a názvy diplomových prací ke studiu rovnováh v roztocích. Oddíl **Rovnováhy**

Publikace

Blok obsahuje 6 oddílů publikační aktivity, jako jsou **Kompendium, Původní práce, Doporučené knihy, Konference, Patenty a Citační index**. Oddíl

Start Adresa Odkazy Plocha

a pak kliknout na Výuka

Výuka - Internet Explorer

http://meloun.upce.cz/vyuka

Výuka

Soubor Úpravy Zobrazit Oblíbené položky Nástroje Nápověda

Převést Vybrat

Search Web Upgrade Your Toolbar Now Mail Bookmarks

Navrhované weby

Milan Meloun

Profesor analytické chemie and chemometrie na Univerzitě Pardubice.

Menu

- Domů
- Aktuality
- Životopis
- Výuka
  - Analytická chemie
  - Chemometrie I.
  - Chemometrie II.
  - Chemometrie III.
  - Metody a vzorové úlohy
  - Licenční studium - ARISTOTELES,GALILEO
  - Kurzy
- Výzkum
- Publikace
- Ke stažení
- Fotogalerie
- Užitečné odkazy
- Návštěvnost stránek
- Hledat
- Kontakt

Výuka (Teaching)

> *Když lidé vyučují, sami se učí.*
> *HOMINES, DUM DOCENT, DISCUNT.*
> — Lucius Annaeus Seneca (4 př. n. l. – 65 n. l.) Epistulae morales ad Lucilium 7, 8.

Kliknutím se fotografie zvětší na celou obrazovku. Click on this photo to enlarge it.

Analytická chemie (Analytical chemistry)

Nabízený učební text **ANALYTICKÁ CHEMIE I** podává uceleně náplň přednášek o teoretických základech chemických metod analytické chemie, kvalitativní a kvantitativní

a pak kliknout na Chemometrie I

Chemometrie I. - Internet Explorer

http://meloun.upce.cz/vyuka/chemometrie-i

Chemometrie I.

Soubor Úpravy Zobrazit Oblíbené položky Nástroje Nápověda

Převést Vybrat

Search Web Upgrade Your Toolbar Now Mail Bookmarks

Navrhované weby

Milan Meloun

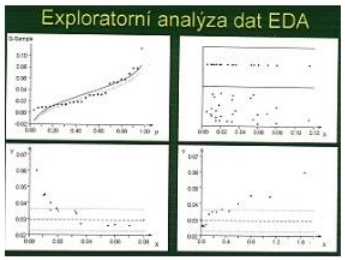
Profesor analytické chemie and chemometrie na Univerzitě Pardubice.

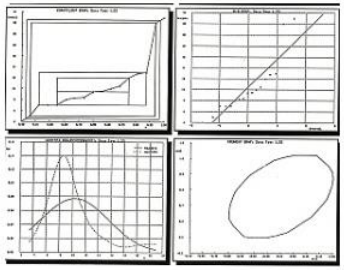
Menu

- Domů
- Aktuality
- Životopis
- Výuka
  - Analytická chemie
  - Chemometrie I.
  - Chemometrie II.
  - Chemometrie III.
  - Metody a vzorové úlohy
  - Licenční studium - ARISTOTELES,GALILEO
  - Kurzy
- Výzkum
- Publikace
- Ke stažení
- Fotogalerie
- Užitečné odkazy
- Návštěvnost stránek
- Hledat
- Kontakt

Chemometrie I - Jednorozměrná data

> *Učený člověk má své bohatství vždy v sobě.*
> *HOMO DOCTUS IN SE SEMPER DIVITIAS HABET.*
> — Phaedrus (kol. 15 př.n.l. - 50 n.l.) Fabulae Aesopiae IV, 22, 1.





A. Potřebné učebnice a vzory semestrálních prací:

1. M. Meloun, J. Militký: **STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT**, Plus, Praha 1994 (1. vydání) nebo East Publishing Praha 1998 (2. vydání), Academia 2004, **INTERAKTIVNÍ STATISTICKÁ ANALÝZA DAT**, Karolinum Praha 2012.
2. M. Meloun, J. Militký: **Kompendium statistického zpracování dat**, Učebnice s CD, Academia Praha 2002 (1. vydání), Academia Praha 2006 (2. vydání), Karolinum Praha 2013 (3. vydání)
3. M. Meloun, J. Militký: **Sbírka úloh - STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT**, Univerzita Pardubice 1997.
4. M. Meloun, J. Militký: **Počítačová analýza vícerozměrných dat v příkladech**, Učebnice s CD, Academia Praha 2002 (1. vydání), **Statistická analýza vícerozměrných dat v příkladech**, Učebnice s CD, Academia Praha 2012 (2. vydání)

Start Adresa Odkazy Plocha

Děkuji za pozornost!

http://meloun.upce.cz