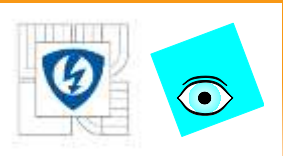


Statistická analýza jednorozměrných dat

Prof. RNDr. Milan Meloun, DrSc.
Univerzita Pardubice, Pardubice

31.ledna 2011

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky.



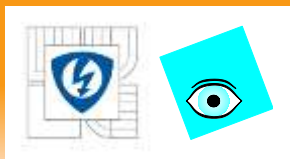
Kapitola 8.1

NELINEÁRNÍ REGRESNÍ MODELY

31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





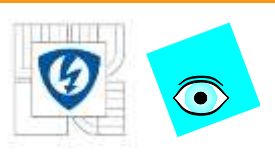
Nelineární regresní modely

Základní úlohy:

1. konstrukce kalibračních modelů,
2. ověření teoretických modelů fyzikálně-chemické zákonitosti,
3. tvorba empirických modelů.

Tvorba regresního modelu $f(c, \beta)$ čili funkce

a) Vektoru nastavovaných proměnných (deterministických, kontrolovatelných, vysvětlujících, nezávislých) x , tj. bodů $\{x_i^T, y_i\}, i = 1, \dots, n$

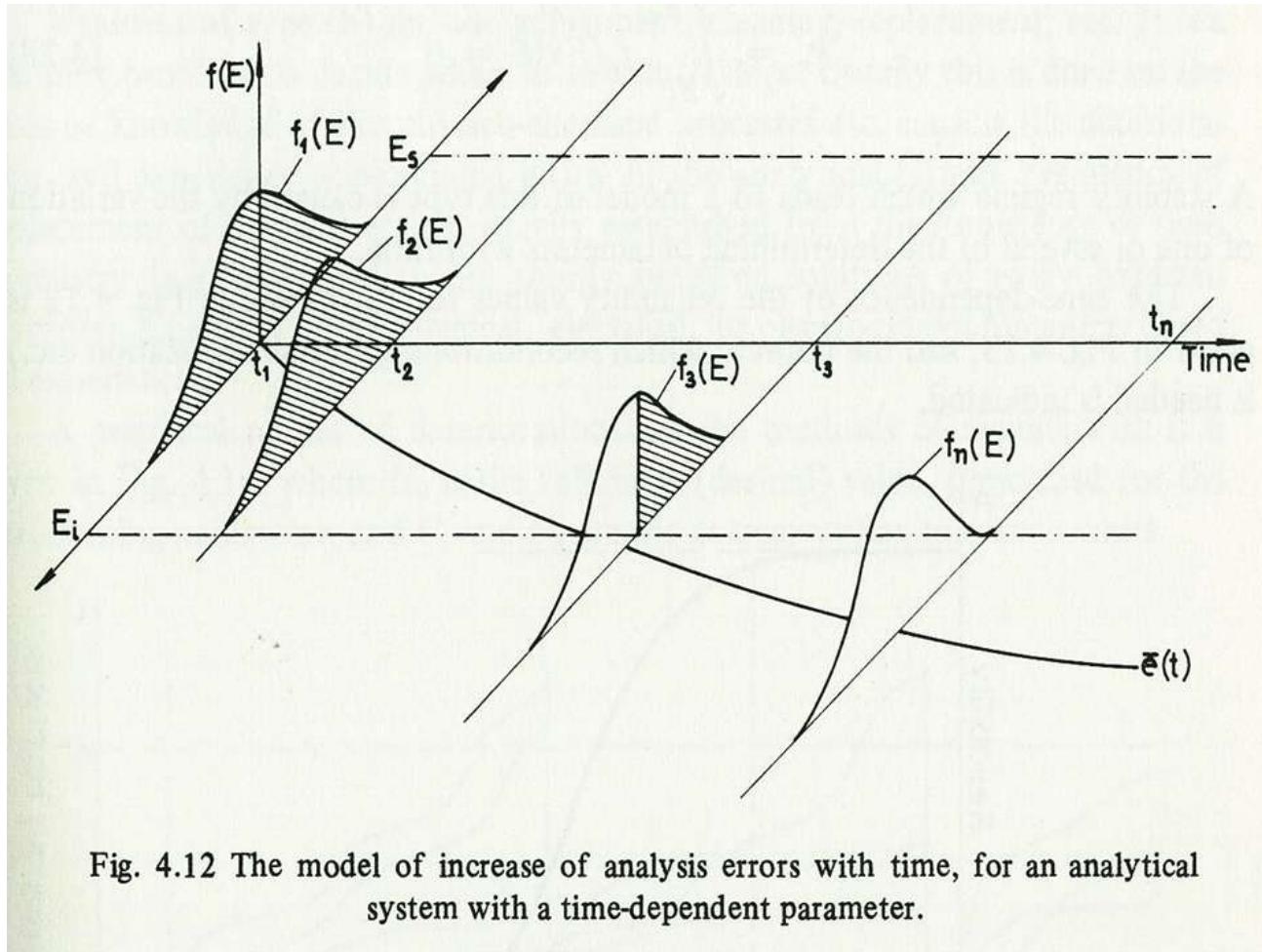


Nelineární regresní modely

- b) vektoru parametrů β o rozměru $(m \times 1)$, $\beta = (\beta_1, \dots, \beta_m)^T$.
- c) y je vysvětlovaná proměnná (závisle p., odezva, měření, pozorování) na zvolenou kombinaci nastavovaných veličin x_i

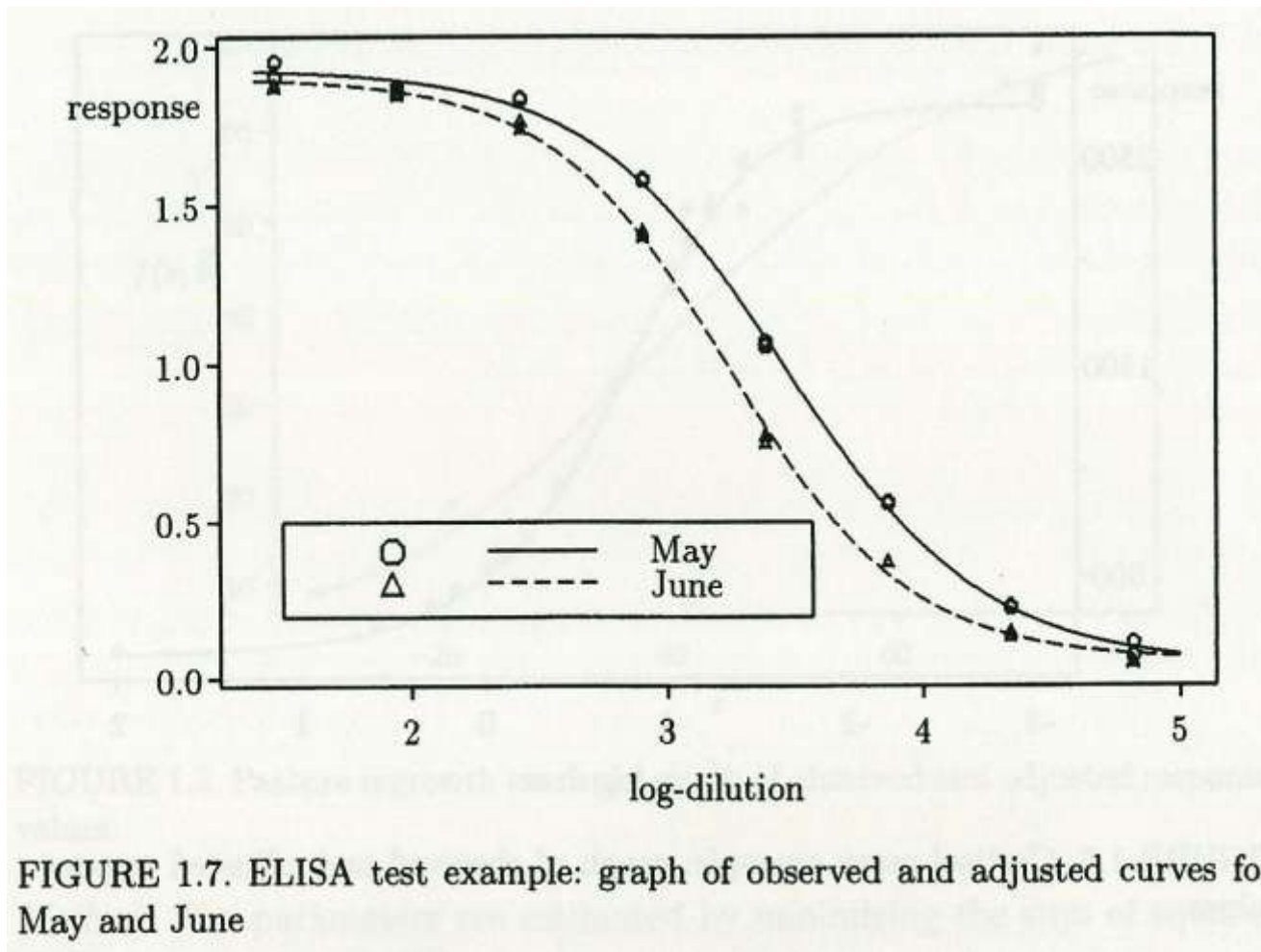


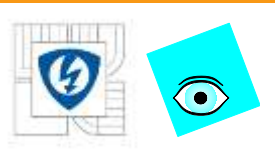
Nelineární regresní modely





Nelineární regresní modely





Nelineární regresní modely

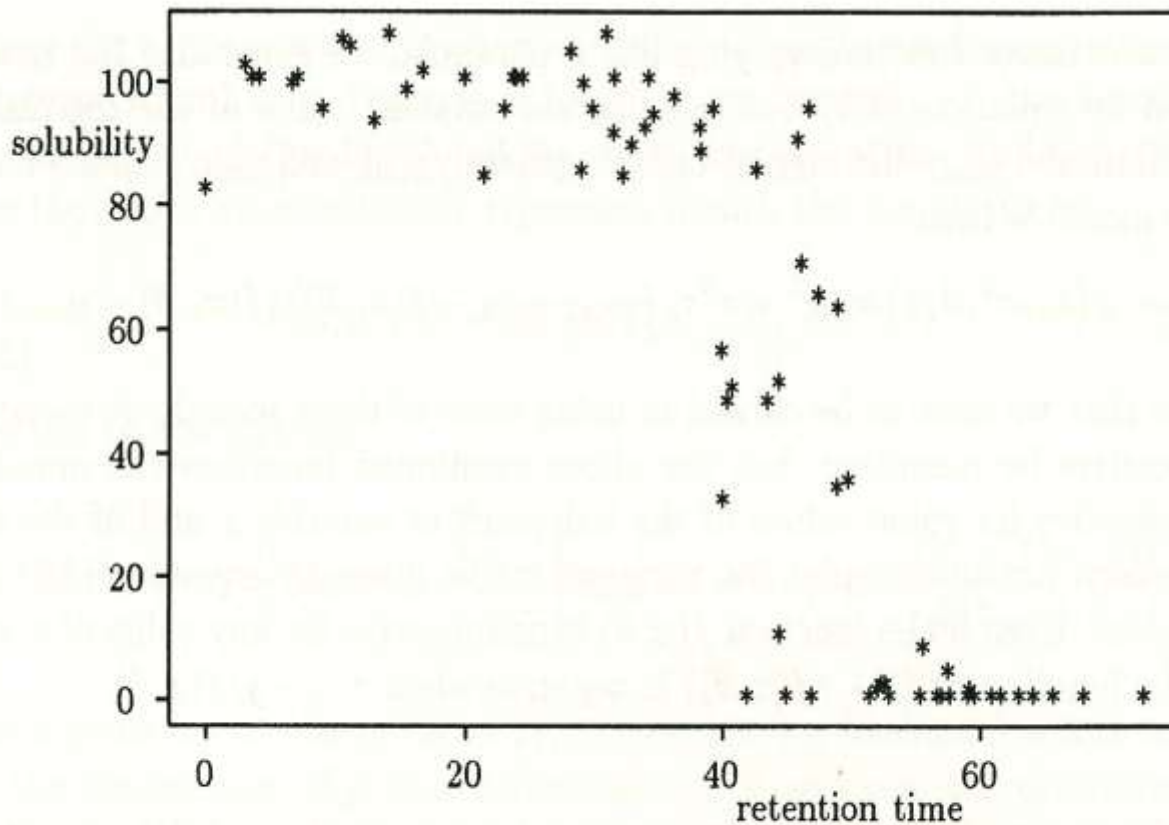


FIGURE 3.3. Peptides example: observed solubility of 75 peptides versus their RP-HPLC retention time



Nelineární regresní modely

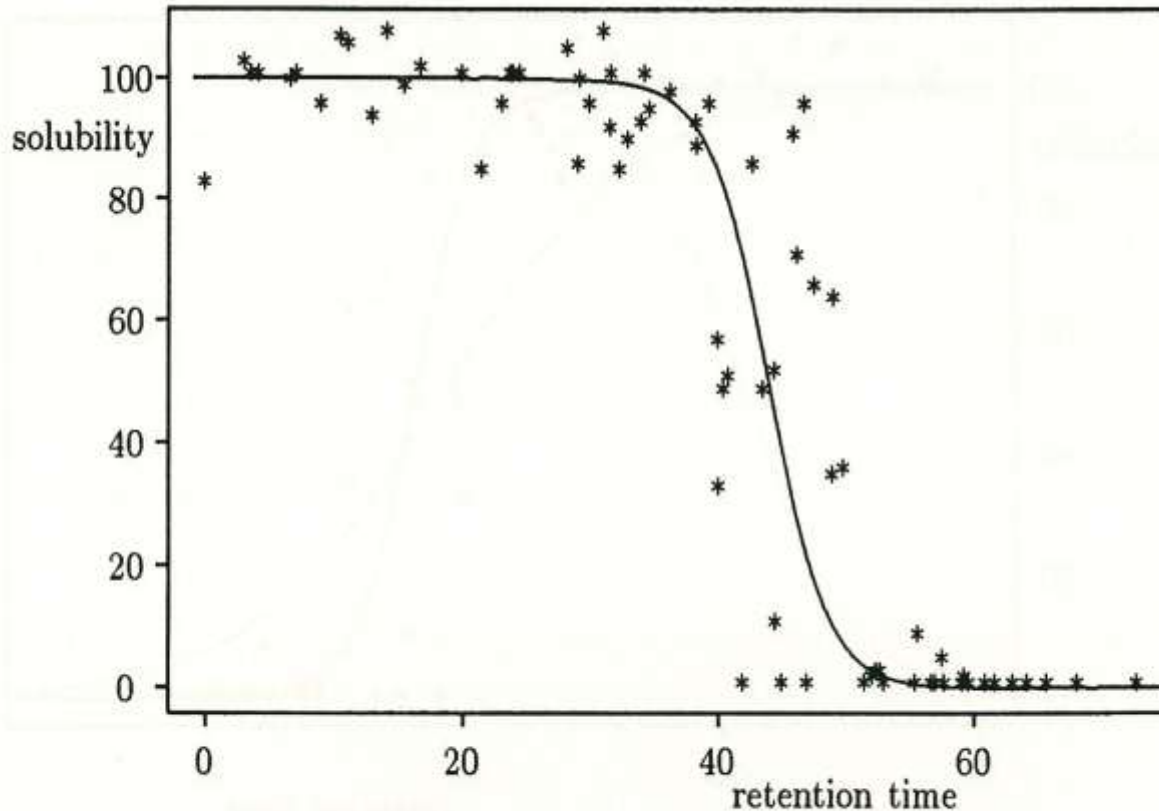


FIGURE 3.7. Peptides example: observed and adjusted solubilities of 75 peptides versus their RP-HPLC retention time



Nelineární regresní modely

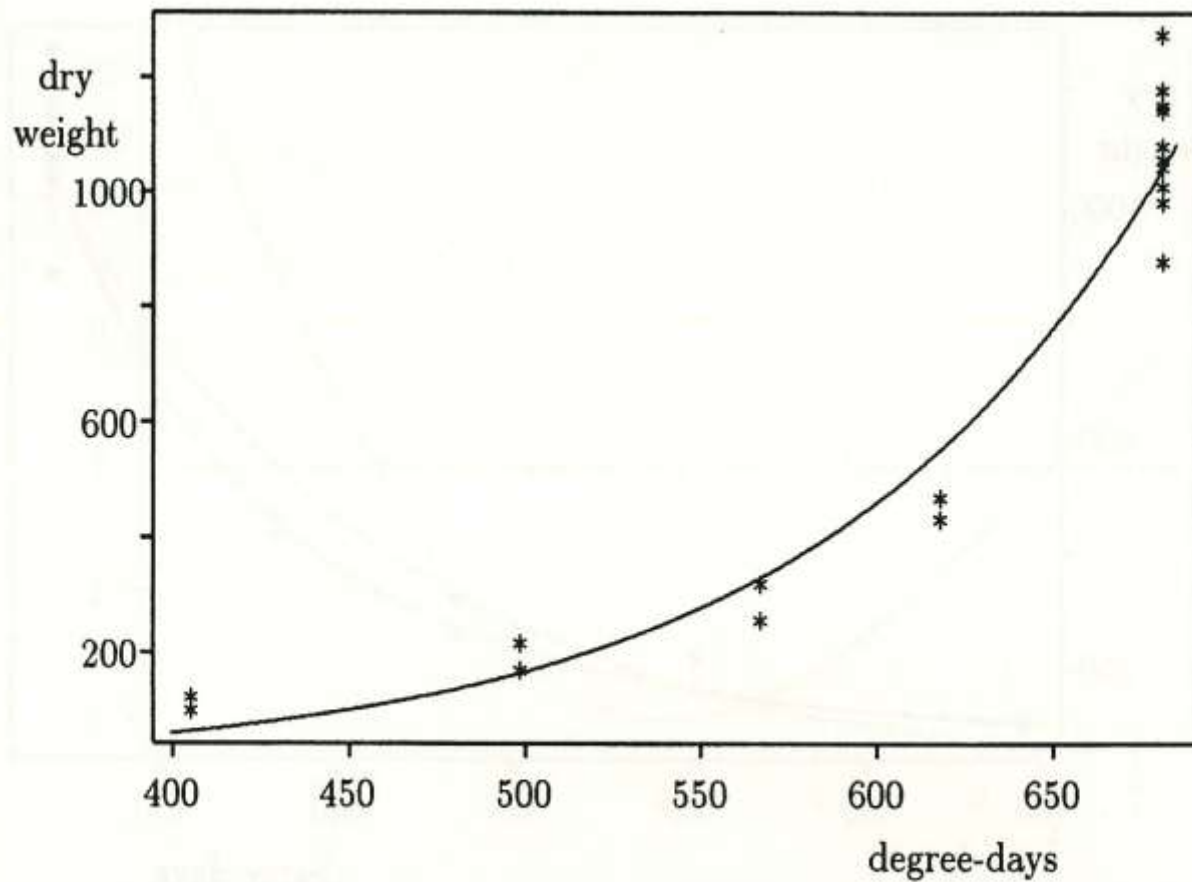


FIGURE 3.4. Weight of tillers example: graph of observed and adjusted curve



Nelineární regresní modely

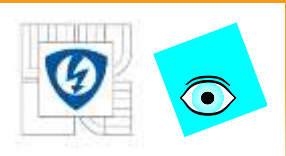
Tvorba se formuluje s ohledem na regresní triplet

1. zadaná data,
2. navržený model,
3. kritérium regrese.

Dělení regresních modelů:

1. Lineární modely: parametry nemají fyzikální smysl (koeficienty),
2. Nelineární modely: parametry mají přesný fyzikální význam. (např. rovnovážné konstanty (disociační, stability, součiny rozpustnosti) reakčních produktů, rychlostní konstanty u kinetických modelů, neznámé koncentrace, atd.

Model: rozšířený Debye-Hückelův zákon

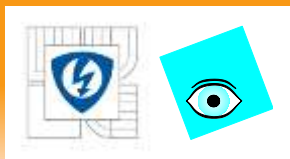


$$pK_{a,smíš} = pK_{a,T} - \frac{0.5115\sqrt{I}}{1 + 3.29 \cdot 10^{10} \dot{a} \sqrt{I}} + CI$$

Proměnné:

Závisle proměnná y : $pK_{a,smíš}$, nezávisle proměnná x : I ,

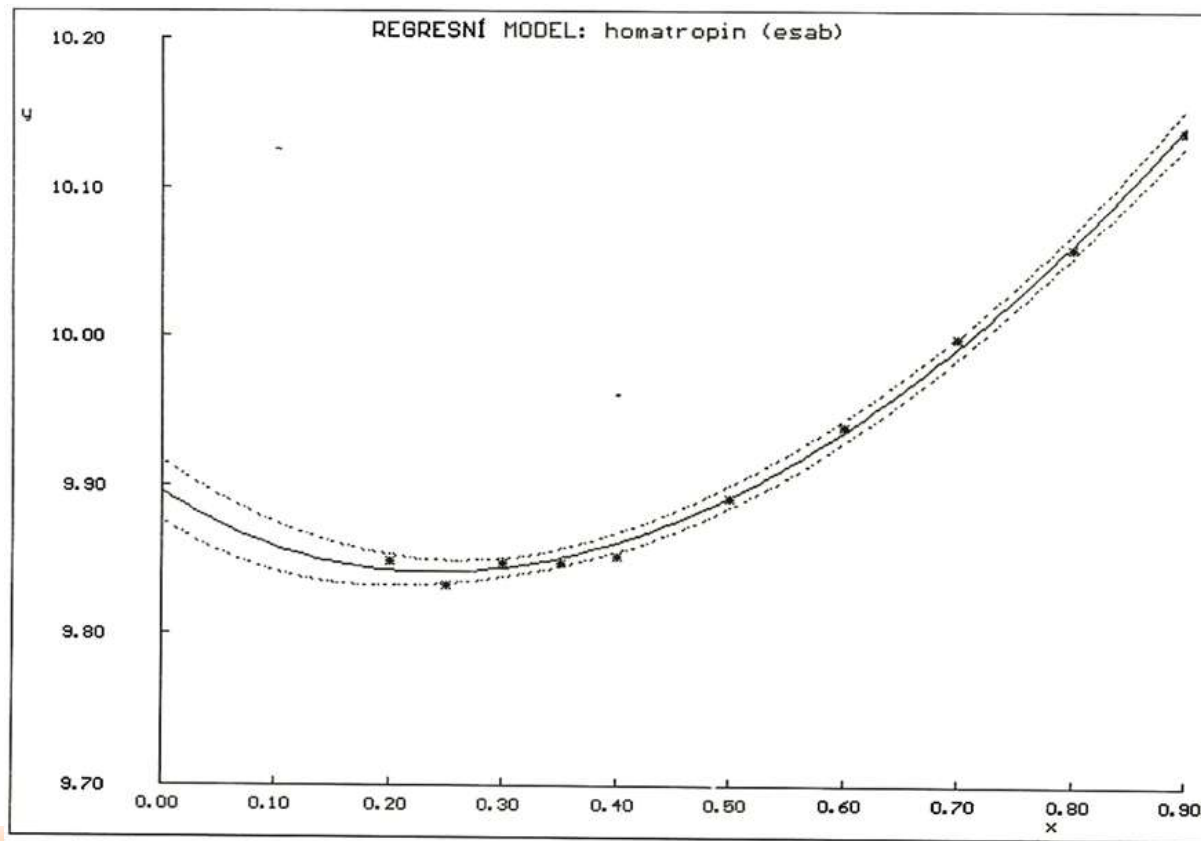
Neznámé parametry: β_1 : $pK_{a,T}$; β_2 : $\dot{a}$; β_3 : C

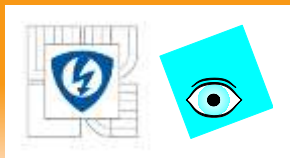


Závislost smíšené disoc. konstanty $pK_{a,smíš}$ homotropinu na iontové síle

Nulté přiblížení: $pK_a^T = 1, \dot{a} = 1, C = 1$

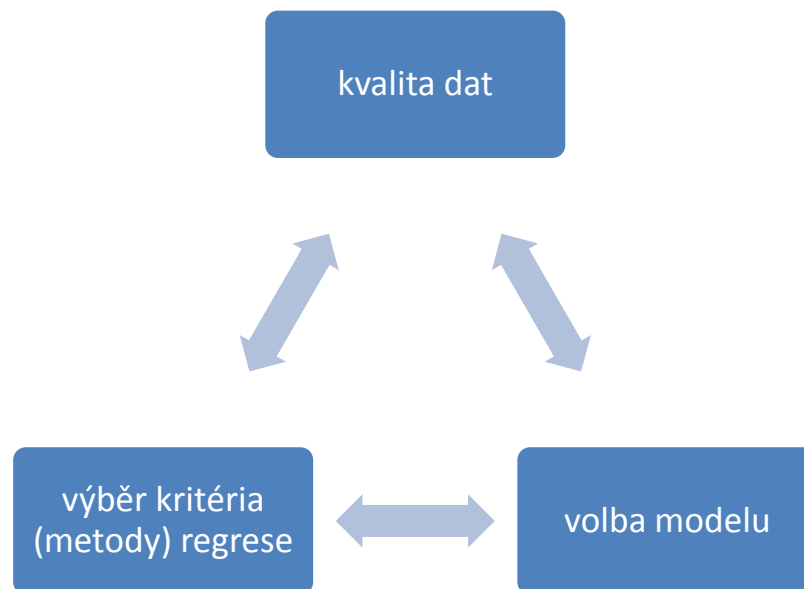
Nalezeno: $pK_a^T = 9.90(1), \dot{a} = 6(2), C = 0.51(3)$





Postup a diagnostika regrese

Regresní triplet:



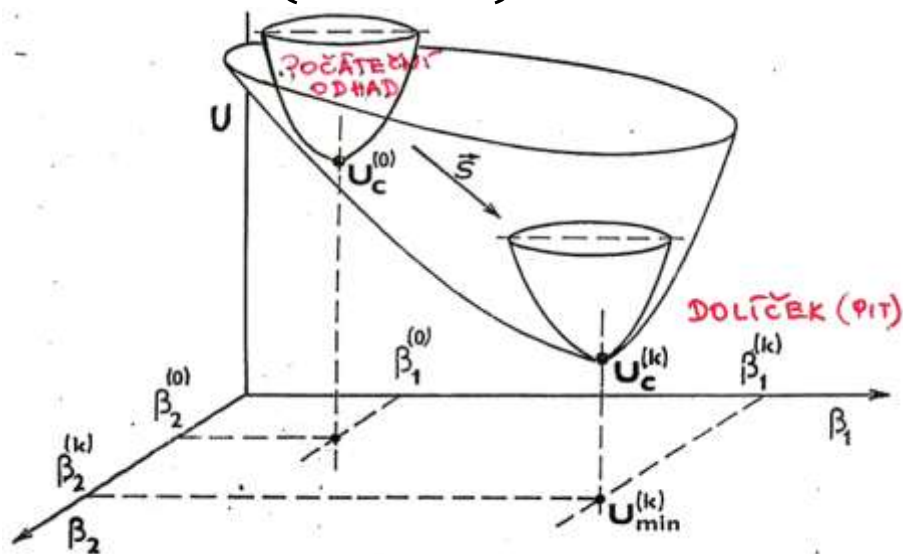
$$U = \sum_{i'+}^n w_i (y_{exp,i} - y_{vyp,i})^2 = \text{minimum}$$

kde $y_{vyp,i} = f(x_i; \beta_1, \dots, \beta_m)$

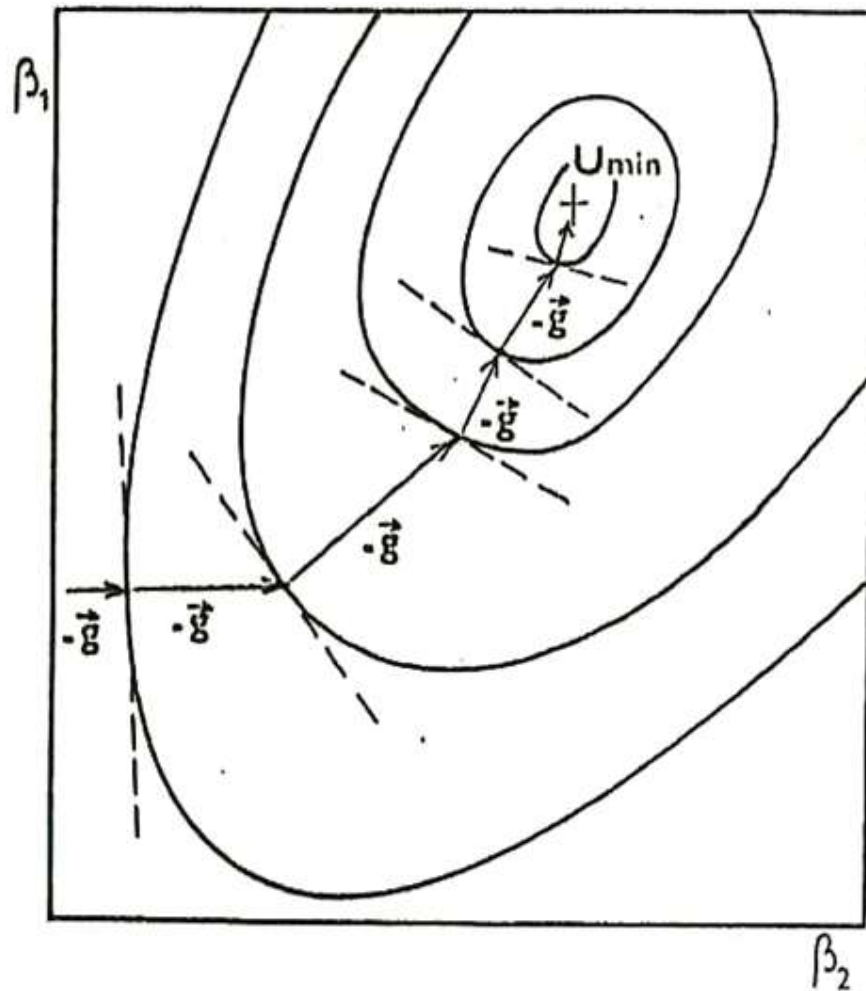
$$U = \sum_{i'+}^n w_i (y_{exp,i} - y_{vyp,i})^2 \approx \text{minimum}$$

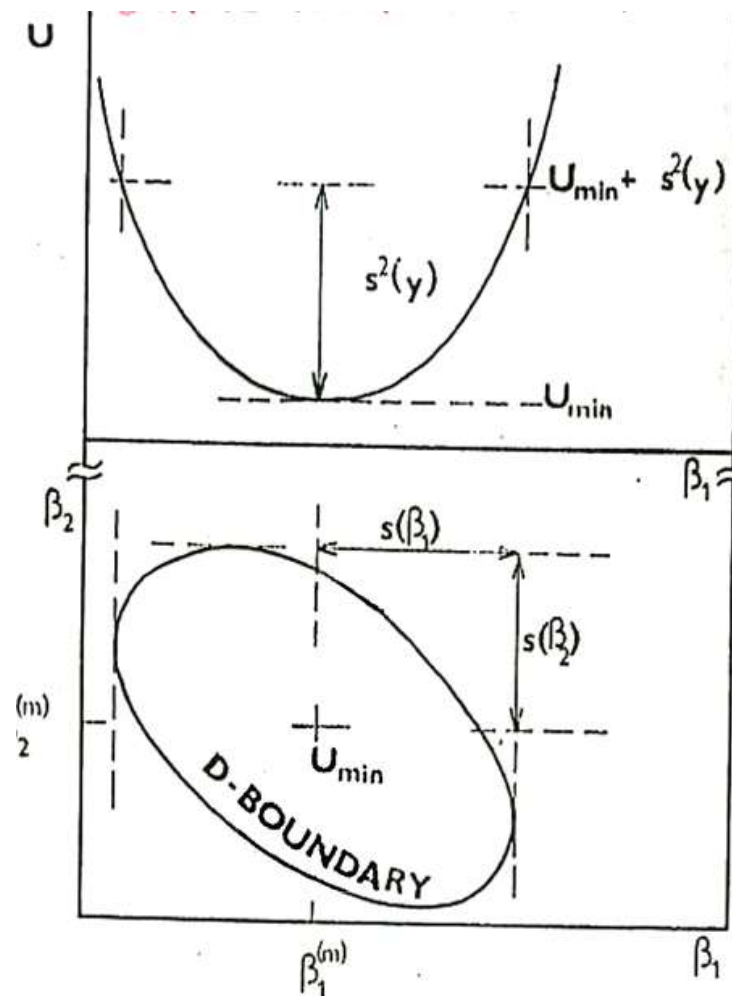
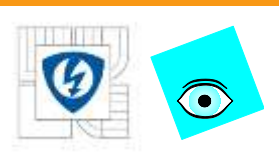
kde $y_{vyp,i} = f(x_i; \beta_1, \dots, \beta_m; p_1, \dots, p_m)$

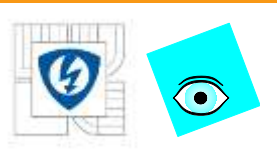
Optimalizační problém v $(m + 1)$ rozměrném prostoru



Gradientová metoda hledání minima

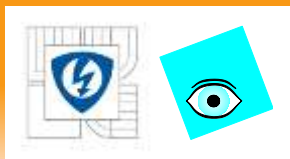






Formulace modelu

1. **Lineární regresní model:** je lineární kombinací parametrů, např. model $f(x, \beta) = \beta_1 + \beta_2 \sin(x)$. Pro lineární regresní modely platí podmínka
$$g_j = \frac{\delta f(x, \beta)}{\delta \beta_j} = konst. \text{ kde } j = 1, \dots, m.$$
2. **Nelineární regresní model:** alespoň pro jeden parametr β_j je parciální derivace g_j jeho funkcí,



Formulace modelu

Dělí se:

1. **Neseeparabilní modely**, kdy podmínka parciální derivace g_j neplatí pro žádný parametr modelu, např.

$$f(x, \beta) = \exp(\beta_1 x) + \exp(\beta_2 x)$$

2. **Separabilní modely**, kdy podmínka parciální derivace g_j platí pro jeden modelový parametr, např.

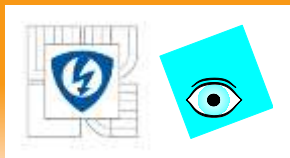
$$f(x, \beta) = \beta_1 + \beta_2 \exp(\beta_3 x)$$

který je nelineární pouze vzhledem k parametru β_3 .

3. **Vnitřně lineární modely**, jsou nelineární, ale lze je reparametrizací převést na lineární regresní model, např.

$$f(x, \beta) = \beta^2 x$$

který reparametrizací $\gamma = \beta^2$ přechází na lineární model



Reparametrizace

Transformace parametrů β do nových γ , např.
reparametrizací Arrheniovy rovnice

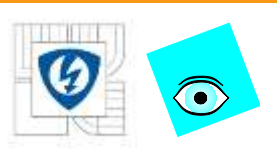
$$f(x, \gamma) = \exp(\gamma_1 + \gamma_2/x)$$

vznikne neseparabilní model: $\gamma_1 = \ln \beta_1$ a $\gamma_2 = \beta_2$.

Linearizace modelu: transformací převést na lineární
regresní model, např. Arrheinův regresní model lze při
zanedbání chyb transformovat do tvaru

$$\ln y = \gamma_1 + \gamma_2 z$$

kde $\gamma_1 = \ln \beta_1$, $\gamma_2 = \beta_2$ a $z = 1/x$.

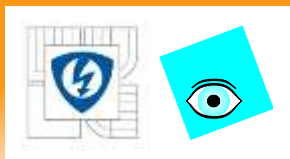


Linearizace modelu

Výsledný model je lineární, ale transformace není správná, protože se projeví vznikem heteroskedasticity:

- naměřené k_i měly konstantní rozptyl $\sigma^2(k)$, a hodnoty $\ln k_i$ mají nekondantní rozptyl

$$\sigma^2(\ln k_i) = \frac{\sigma^2(k_i)}{(\ln k_i)^2}, \text{ tj. konstantní relativní chybu.}$$



Míra citlivosti parametru β_1 v modelu

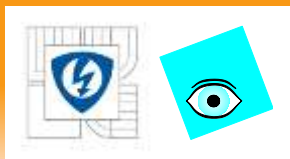
$$g_j = \frac{\delta f(x, \beta)}{\delta \beta_j}, \quad j = 1, \dots, m$$

Redundance (přeurčení): nadbytečný počet parametrů.

- model neobsahuje nadbytečné parametry, když jsou míry citlivosti g_j pro analyzovaná data lineárně nezávislé, tj. nelze určit nenulové koeficienty $v_j, j = 1, \dots, m$, aby bylo splněno kritérium redundance

$$\sum_{j=1}^m g_j v_j = 0$$

- existuje-li alespoň jeden $v_j \neq 0$, je regresní model přeuročeny a je nutné ho zjednodušit

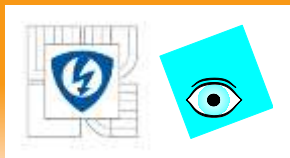


Potíže redundance (přeurčení modelu)

1. Přeurčení vede vždy k singularitě matice $J^T J$.
2. Lokální přeurčení se neprojeví při použití pseudoinverze matice $J^T J$.
3. Špatná podmíněnost nelineárních modelů (redundance), je kvůli špatné podmíněnosti matice $J^T J$. Známe-li přibližné velikosti parametrů $\mathbf{b}^{(0)}$, můžeme sestavit matici $\mathbf{L} = n^{-1}(\mathbf{J}^T \mathbf{J})$ s prvky

$$L_{jk} = \frac{1}{n} \sum_{i=1}^n \frac{\delta f(x_i, \mathbf{b})}{\delta b_j} \frac{\delta f(x_i, \mathbf{b})}{\delta b_k} \Big|_{\mathbf{b}=\mathbf{b}^{(0)}}$$

(Matice $\mathbf{L}$ odpovídá matici $n^{-1}(\mathbf{X}^T \mathbf{X})$ pro lineární modely).



Potíže redundance (přeurčení modelu)

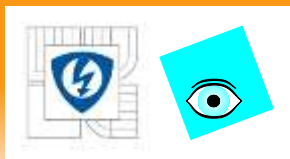
Pro stanovení podmíněnosti se matice L převádí do standardizované formy L^*

$$L_{ij}^* = \frac{L_{ij}}{\sqrt{L_{ii}L_{jj}}}$$

Dle podmíněnosti matice L^* se usuzuje na podmíněnost daného modelu.

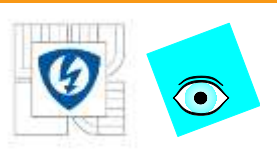
Míra podmíněnosti:

Mírou špatné podmíněnosti je determinant matice L^* : $\det(L^*) < 0.01$, nelineární model špatně podmíněný a je nutné hledat jeho zjednodušení.



Druhy chyb: celková chyba ε závisle proměnné je kombinací

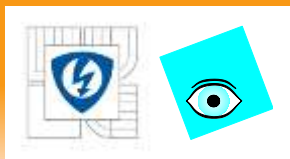
- chyby měření ε_M ,
- chyby formulace modelu ε_A ,
- chyby nastavování nezávisle proměnné ε_x ,
- chyby náhodného kolísání podmínek experimentu ε_N ,
- má nulovou střední hodnotu pro všechny hodnoty $y_i, i = 1, \dots, n, E(\varepsilon_i) = 0$.



Platí

1. pro $E(\varepsilon_i) = konst.$, chybí v modelu absolutní člen,
2. pro $E(\varepsilon_i) \neq konst$ je model nevhodně navržen.

Obecný model regrese: zahrnuje vlastní regresní model a i model měření $y_i = Z_i(x_i, \varepsilon_i, \beta), i = 1, \dots, n$

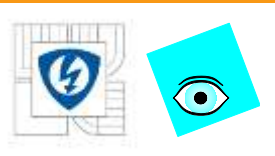


Modely chyb měření

1. **Aditivní model měření:** pro výsledky experimentů y_i platí $Z_i = f(x_i, \beta) + \varepsilon_i$
2. **Multiplikativní model měření:** pro výsledky experimentů y_i platí $Z_i = f(x_i, \beta) \exp \varepsilon_i$
3. **Smíšený model měření:** pro výsledky experimentů y_i platí $Z_i = f(x_i, \beta) \exp(v_i) + \varepsilon_i$

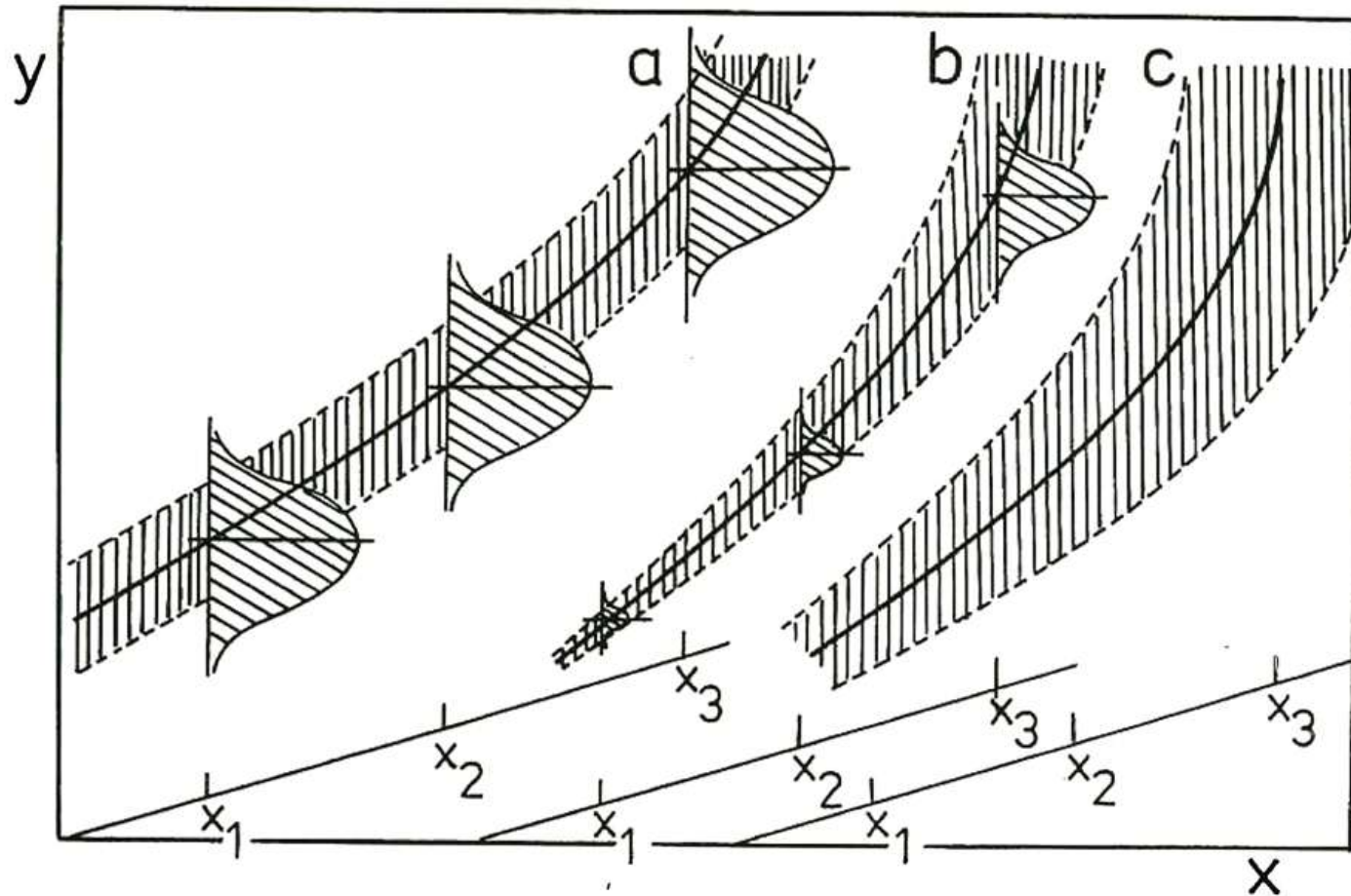
kde chyby v_i a ε_i jsou vzájemně nezávislé.

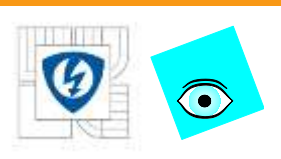
Měření na přístrojích jsou zatížena konstantní relativní chybou $v(y) = \sigma(y)/f(x, \beta)$ a rozptyl měřené veličiny $\sigma^2(y) \approx f^2(x, \beta)$ je úměrný čtverci hodnoty modelové funkce



Modely chyb měření

(a) aditivní, (b) multiplikativní, (c) smíšený model





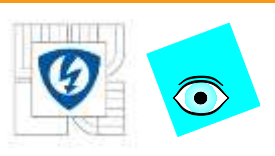
Celková chyba ε_i je pak

$$\varepsilon_i = \sum_{j=1}^i u_j + v_i$$

kde v_i jsou chyby měření a u_j jsou procesní chyby.

Procesní chyby jsou způsobeny fluktuací exper. podmínek, (tj. teploty, tlaku a čistoty chemikálií) a mají kumulativní charakter. Předpokládá se, že chyby ε_i působí aditivně

.



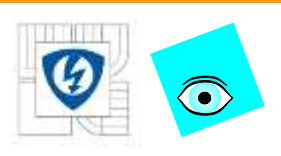
Modely chyb měření

Rozdělení náhodných veličin y_i : bude úzce souviset s rozdělením chyb ε_i , které je dáno sdruženou hodnotou pravděpodobnosti $p_\varepsilon(\varepsilon)$, která je funkcí **distribučních parametrů**, jako je rozptyl σ^2 , atd.

Pro případ nezávislých náhodných chyb ε platí vztah

$$p(y_i) = p_\varepsilon[Z_i^{-1}(x_i, y_i \varepsilon)] \left| \frac{\delta Z_i^{-1}(\cdot)}{\delta y_i} \right|$$

kde symbol $Z_i^{-1}(\cdot)$ označuje inverzi k funkci $Z(\cdot)$.

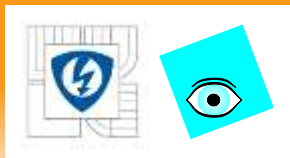


$$Z_i^{-1}(\cdot) = y_i - f(x_i, \beta)$$

a derivace $\left| \frac{\delta Z_i^{-1}(\cdot)}{\delta y_i} \right| = 1$.

Dosazením vyjde $p(y_i) = p_\varepsilon(y_i - f(x_i, \beta))$

Aditivní model měření nezpůsobuje žádné deformace rozdělení měřených veličin vzhledem k rozdělení chyb.



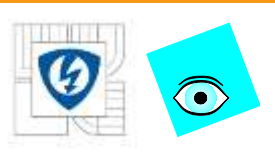
Multiplikativní model měření

$$Z_i^{-1}(\cdot) = \ln y_i - \ln f(x_i, \beta)$$

a derivace $\left| \frac{\delta Z_i^{-1}(\cdot)}{\delta y_i} \right| = \frac{1}{y_i}$, kde se předpokládají pouze kladné hodnoty měřených veličin. Po dosazení bude

$$p(y_i) = \frac{1}{y_i} p_\varepsilon(\ln y_i - \ln f(x_i, \beta))$$

Tato hustota pravděpodobnosti již neodpovídá hustotě pravděpodobnosti chyb $p_\varepsilon(\cdot)$.



Geometrie nelineární regrese

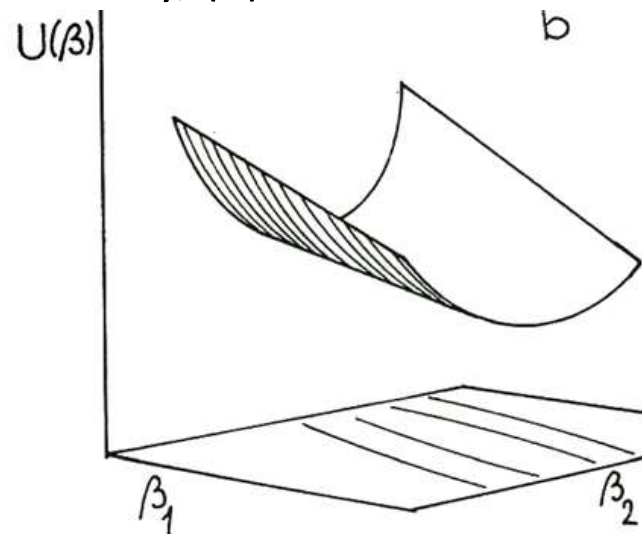
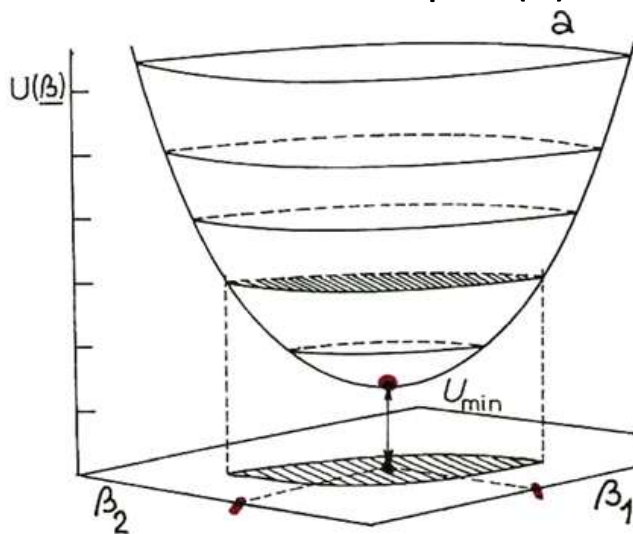
Kritérium regrese $U(\boldsymbol{\beta})$ vektorovým zápisem

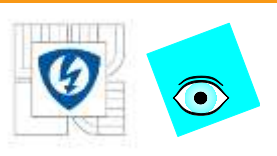
$$U(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{f}\|^2$$

kde $\mathbf{y} = (y_1, \dots, y_n)^T$, $\mathbf{f} = (f(x_1, \boldsymbol{\beta}), \dots, f(x_n, \boldsymbol{\beta}))^T$

a symbol $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$ označuje euklidovskou normu.

Minimum účelové funkce u pro (a) lineární modely, (b) nelineární modely





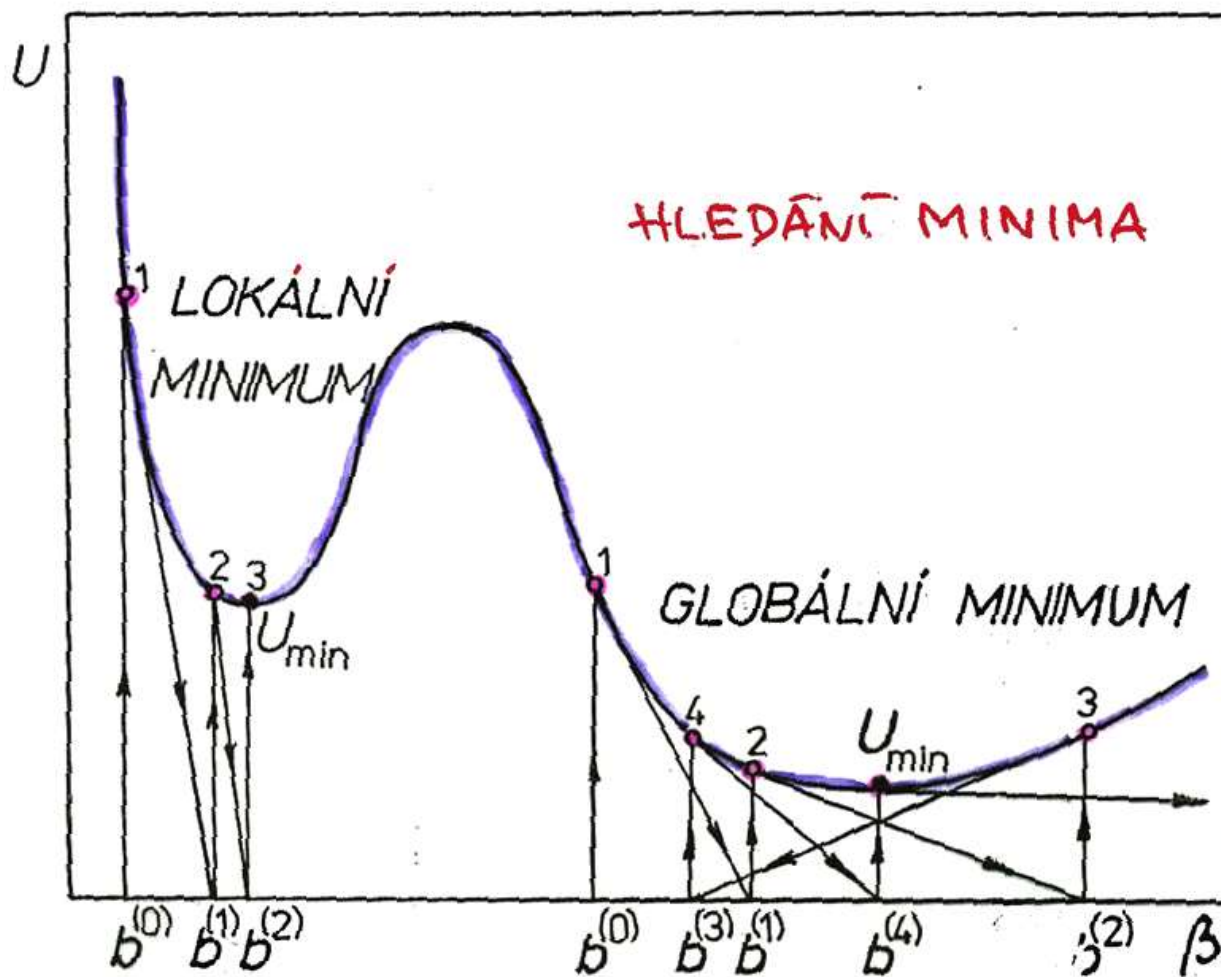
Pro lineární regresní modely

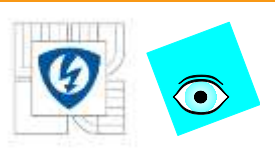
1. Útvar kritéria $U(\boldsymbol{\beta})$ v prostoru odhadů je eliptický hyperparaboloid se středem v bodě $[\boldsymbol{\beta}, U(\boldsymbol{\beta})]$, kde nabývá minimální hodnoty.
2. Účelová funkce $U(\boldsymbol{\beta})$ je vzhledem k $\boldsymbol{\beta}$ kvadratickou formou $\boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X}) \boldsymbol{\beta}$ a matice $\mathbf{X}^T \mathbf{X}$ je pozitivně definitní.

Pro nelineární regresní modely:

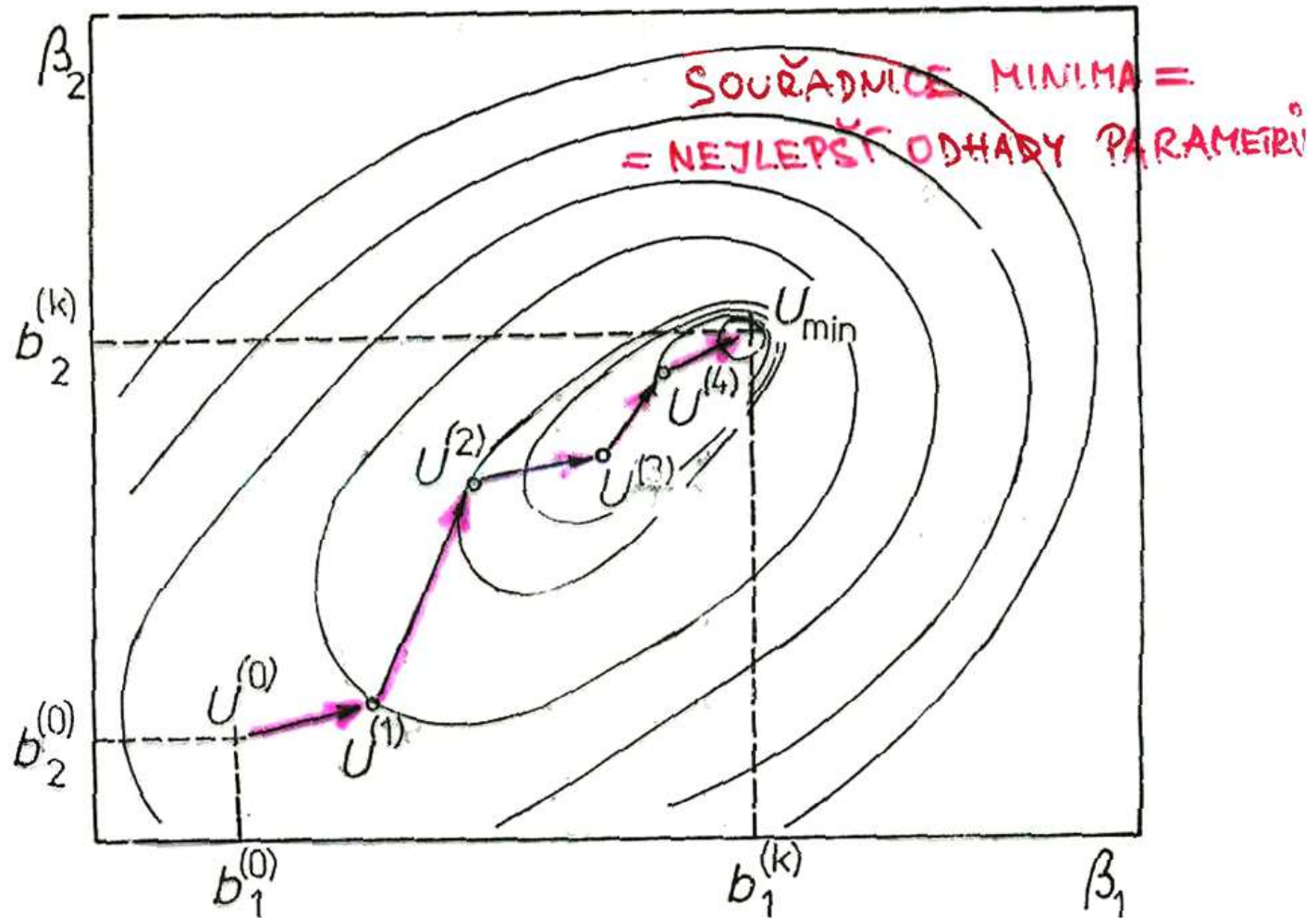
1. Složitější útvary vznikají u nelineárních modelů v závislosti na nelinearitě funkce $f(\mathbf{x}, \boldsymbol{\beta})$, počtu extrémů a sedlových bodů.

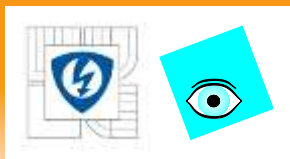
Minimalizační proces



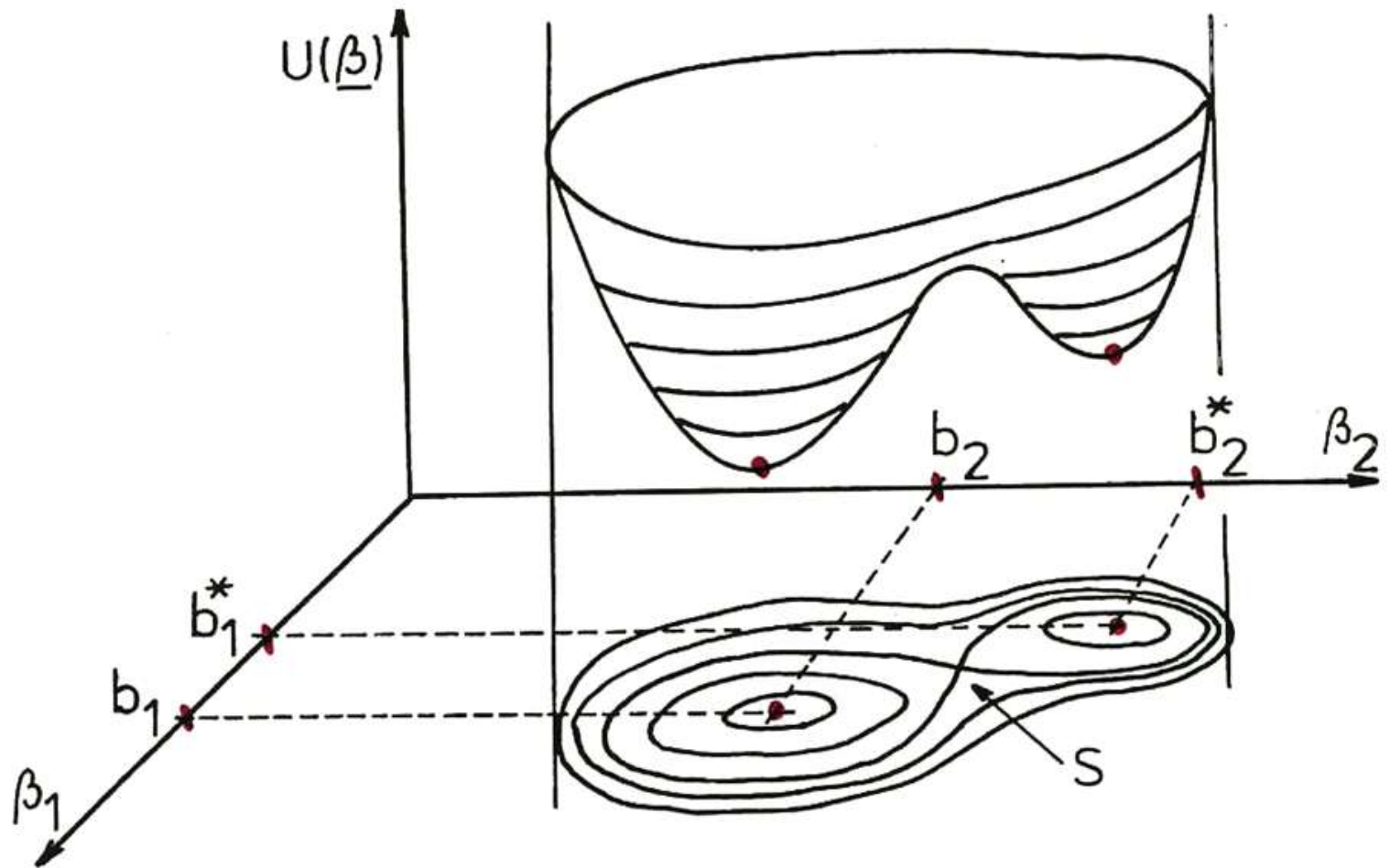


Souřadnice minima = nejlepší odhady parametrů





Případ dvou lokálních minim β a β^* a sedlového bodu S

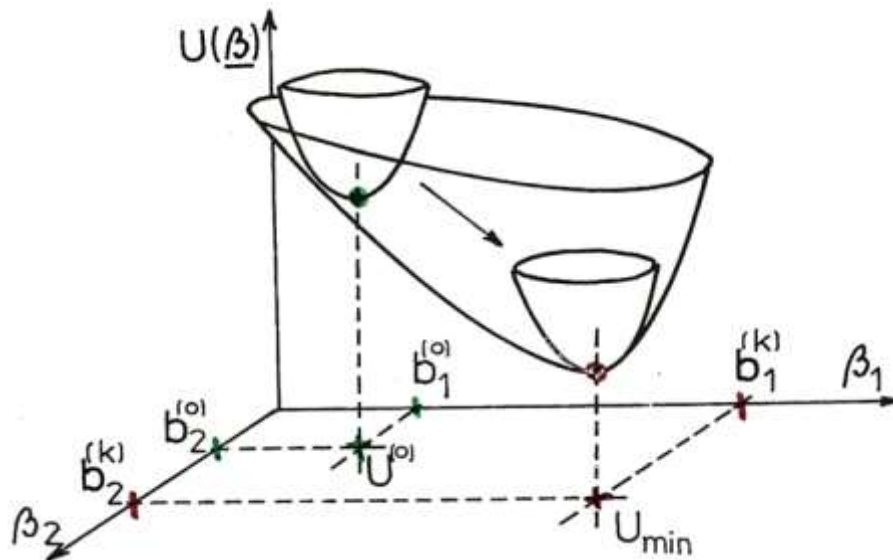


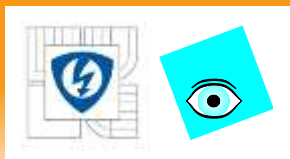


Geometrická interpretace hledání minima $U(\boldsymbol{\beta})$

O lokálním chování funkce $U(\boldsymbol{\beta})$ v okolí libovolného bodu $\boldsymbol{\beta}_j$ lze získat kvantitativní informace z jejího Taylorova rozvoje

$$U(\boldsymbol{\beta}) = U(\boldsymbol{\beta}_j) + \Delta\boldsymbol{\beta}_j^T \mathbf{g}_j + \frac{1}{2} \Delta\boldsymbol{\beta}_j^T \mathbf{H}_j \Delta\boldsymbol{\beta}_j$$





Numerické postupy

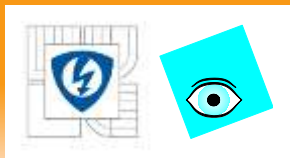
Nelineární minimalizace: model $f(x, \beta)$ je nelineární vzhledem k parametru β_r

Nelineární maximalizace: použití maximální věrohodnosti

Extremalizace: užití libovolného kritéria regrese, kde „proměnné“ jsou regresní parametry β

Optimalizační metody:

- ***hledání volného extrému***, pokud nejsou na regresní parametry kladena žádná omezení,
- ***hledání vázaného extrému***, jestliže regresní parametry musí splňovat jisté omezující podmínky.

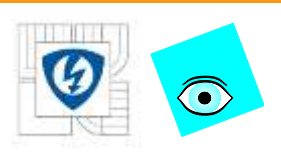


Programy obsahují:

1. hledání extrému v zadaném směru (jednorozměrná optimalizace),
2. invertaci matic,
3. numerické derivace (u derivačních metod),
4. způsoby překonávání lokálních oblastí divergence.

Dělení programů:

- a) nederivační optimalizační metody,
- b) derivační metody pro kritérium metody MNČ,
- c) obecné derivační metody,
- d) algoritmy pro speciální případy.

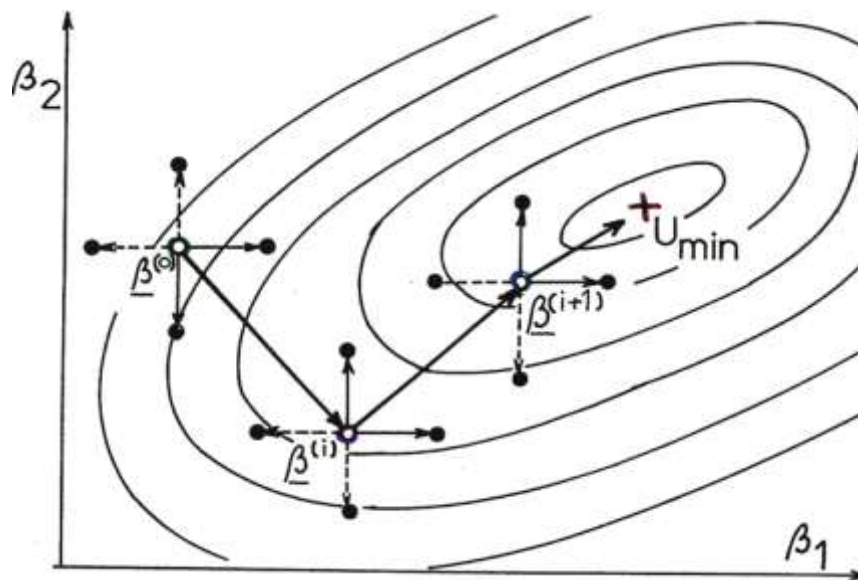


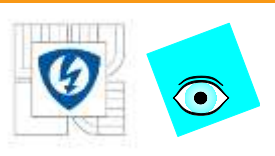
Umožňují nalezení extrému $G(\beta)$ vzhledem k β :

1. metody přímého hledání,
2. simplexové metody,
3. metody využívající náhodných čísel,
4. postupy speciálně vhodné pro MNČ

Hookův-Jeevův algoritmus

- a) krokové posuny a nalezení zlepšeného odhadu $\beta_p^{(i)}$, pro který je $G(\beta_p^{(i)}) < G(\beta^{(i)})$,
- b) hledání lokálního minima $\beta^{(i+1)}$ ve směru $\beta^{(i)}$ a $\beta_p^{(i)}$.

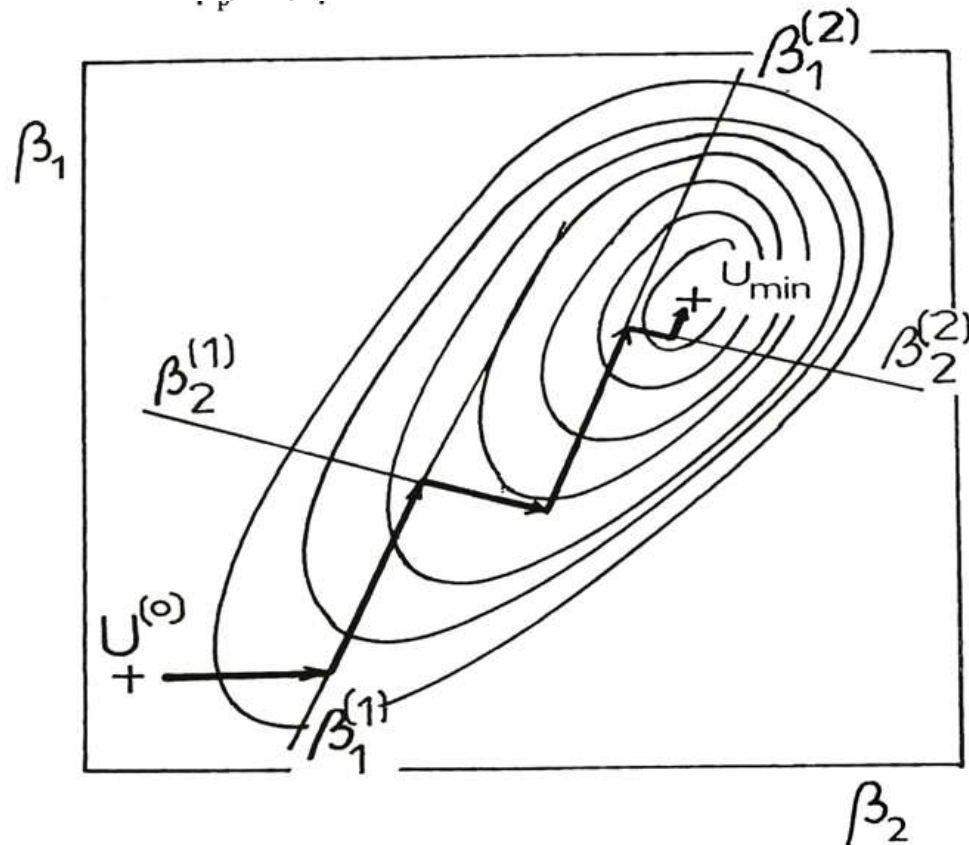




Rosenbrockova metoda

Místo krokového hledání ve směru s rotaci souřadného systému tak, aby jedna osa splynula se směrem

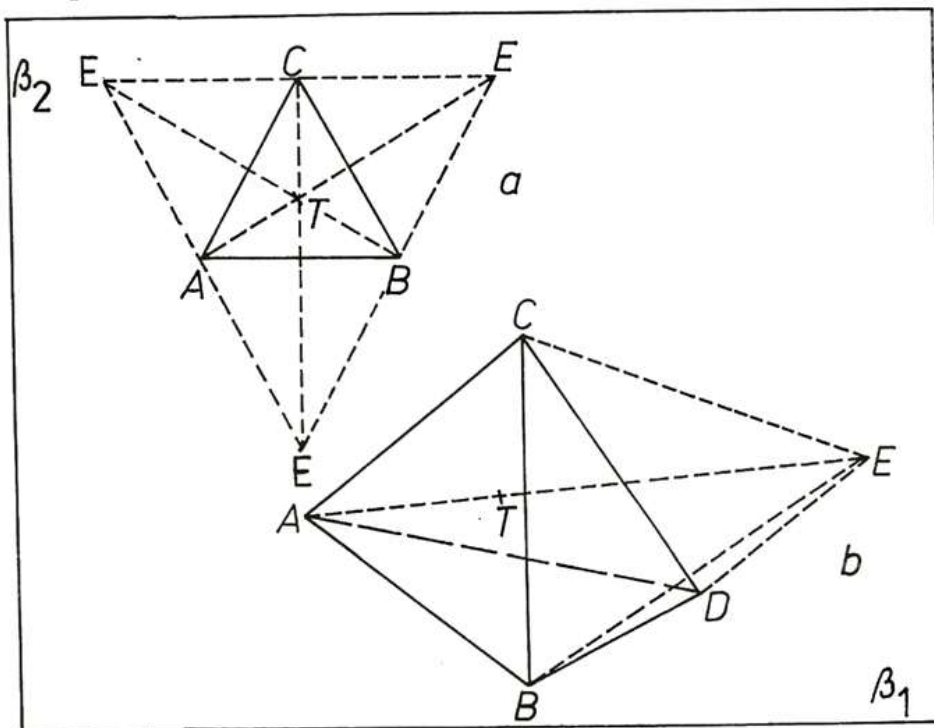
$$S = \beta_p^{(i+1)} - \beta^{(i)}.$$

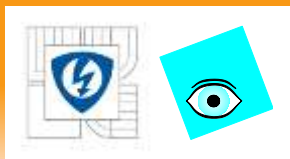


Simplexové metody

Postupné vytváření **adaptivních polyedrů** (simplexů)

Simplex pro (a) $m = 2$, a (b) $m = 3$ parametry. Simplex A, B, C, lze převrátit do tří poloh CBE, ABE, ACE.





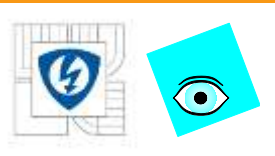
Simplexové metody

Z počátečního odhadu $\beta^{(0)}$ se postupně tvoří ***simplexy***.

Simplex je konvexní mnohostěn v m -rozměrném prostoru parametrů právě svými $(m + 1)$ vrcholy.

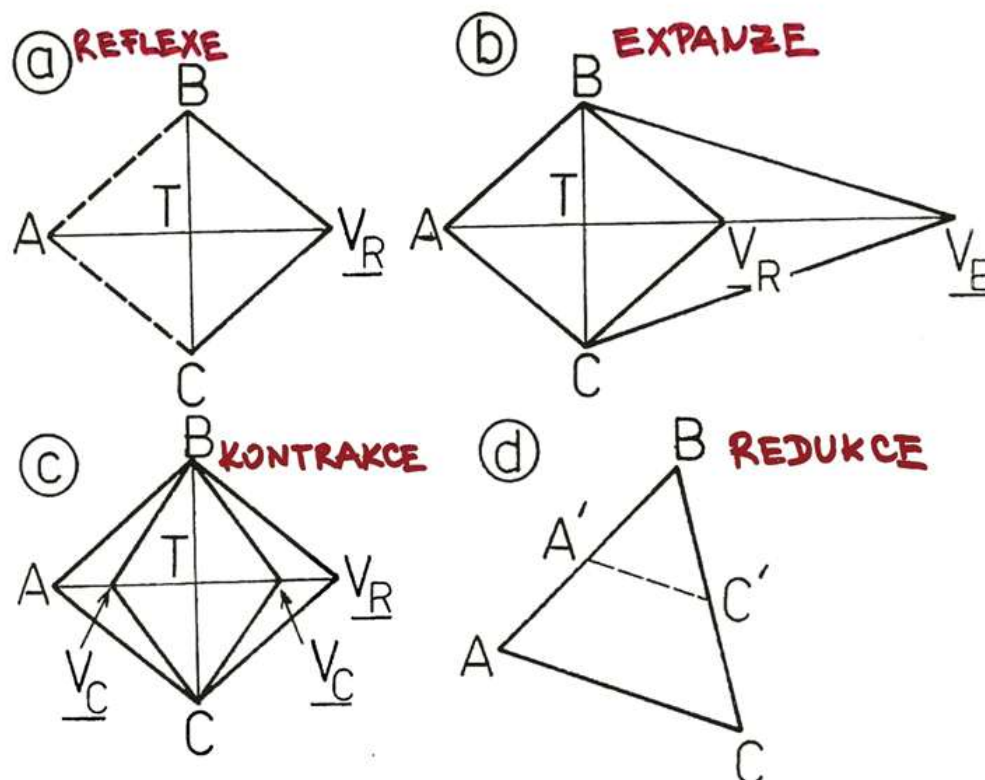
Postup minimalizace:

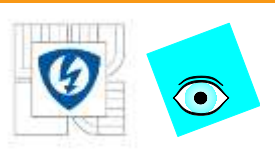
1. konstrukce výchozího simplexu,
2. iterativní postup k minimu,
3. identifikace ukončení hledání terminace.



Iterativní postup k minimu

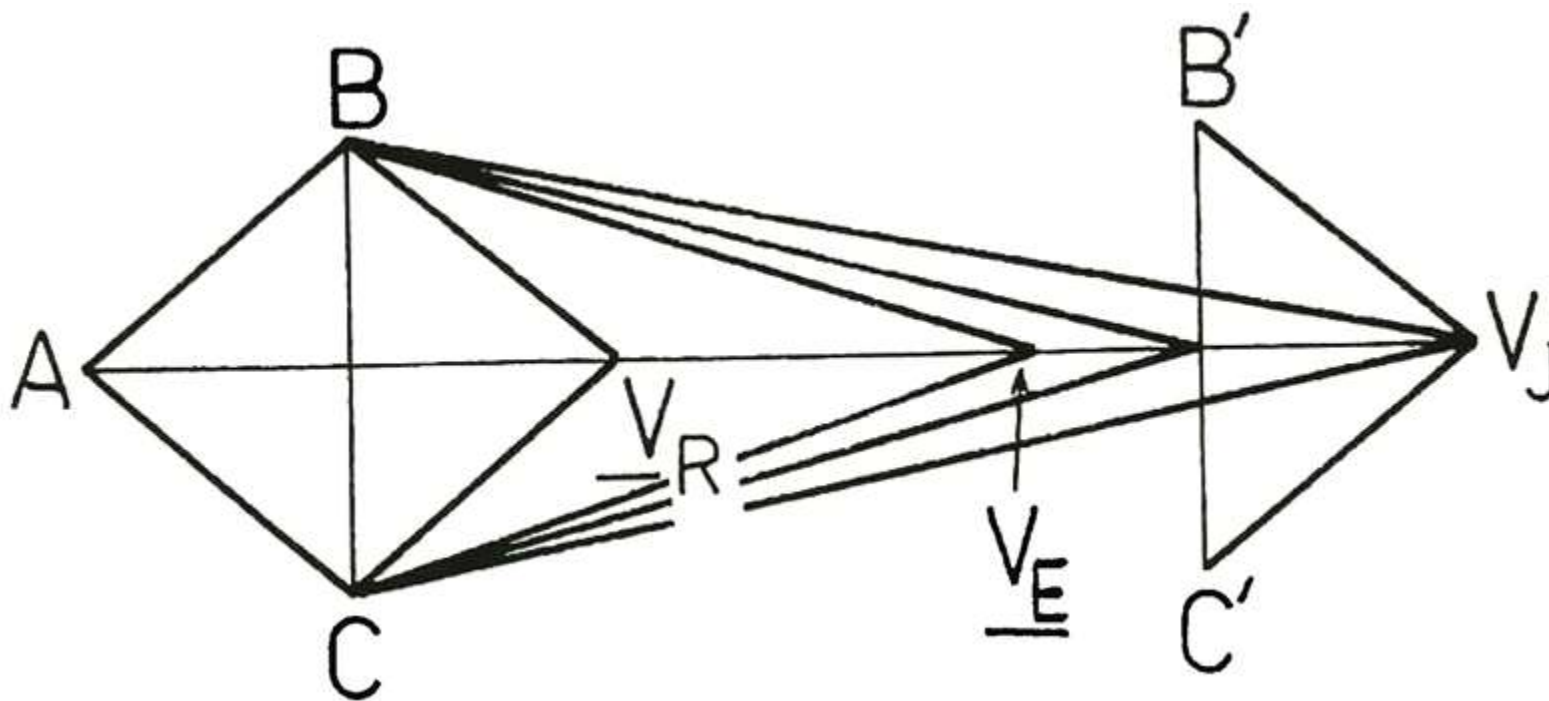
Na spojnici mezi V_H a jeho zrcadlovým obrazem pět ***operací reflexe, expanze, kontrakce, redukce a přenesení.***

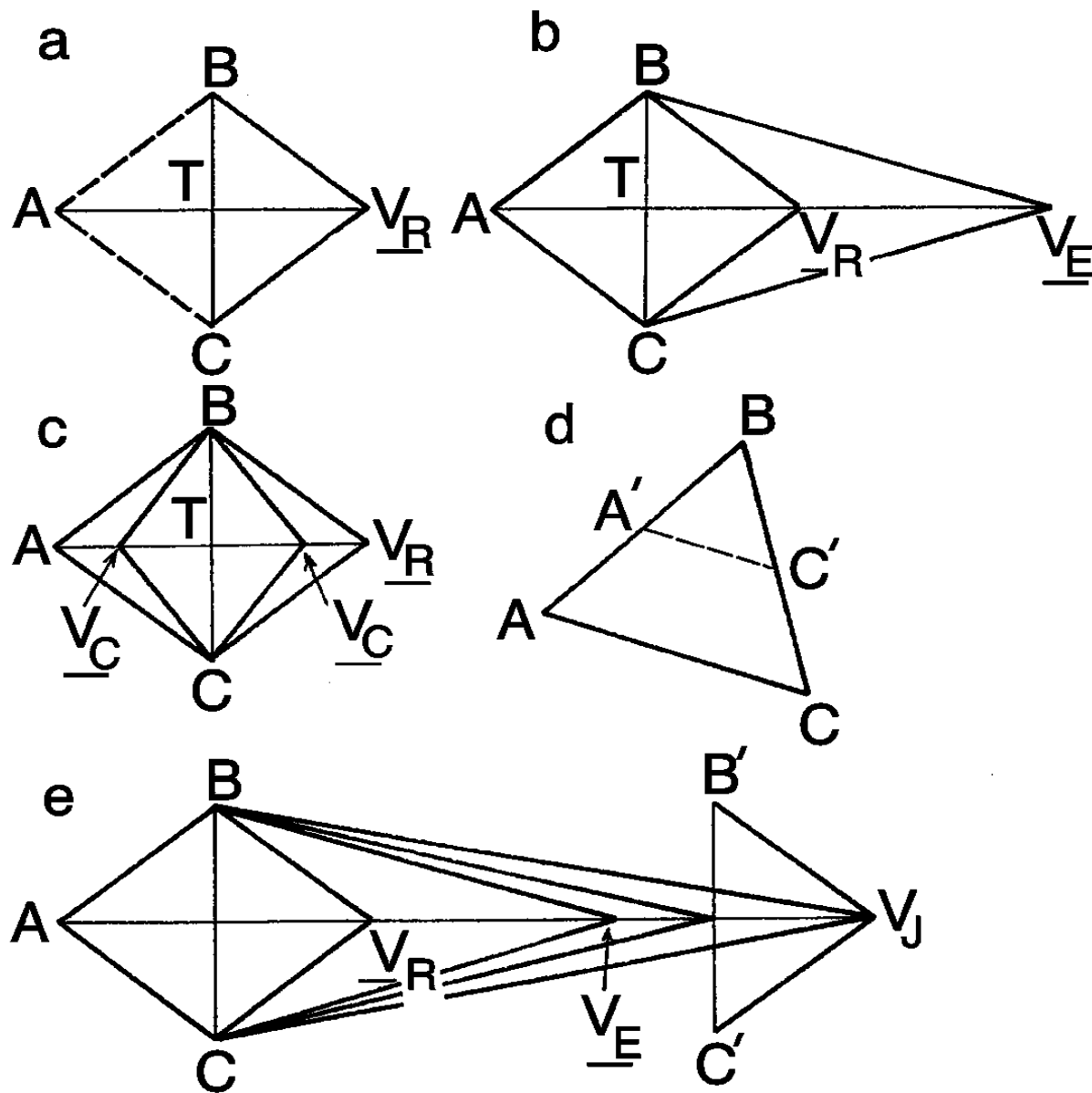
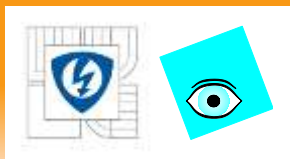


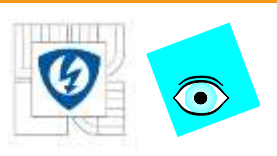


Základní operace simplexu

Přenesení

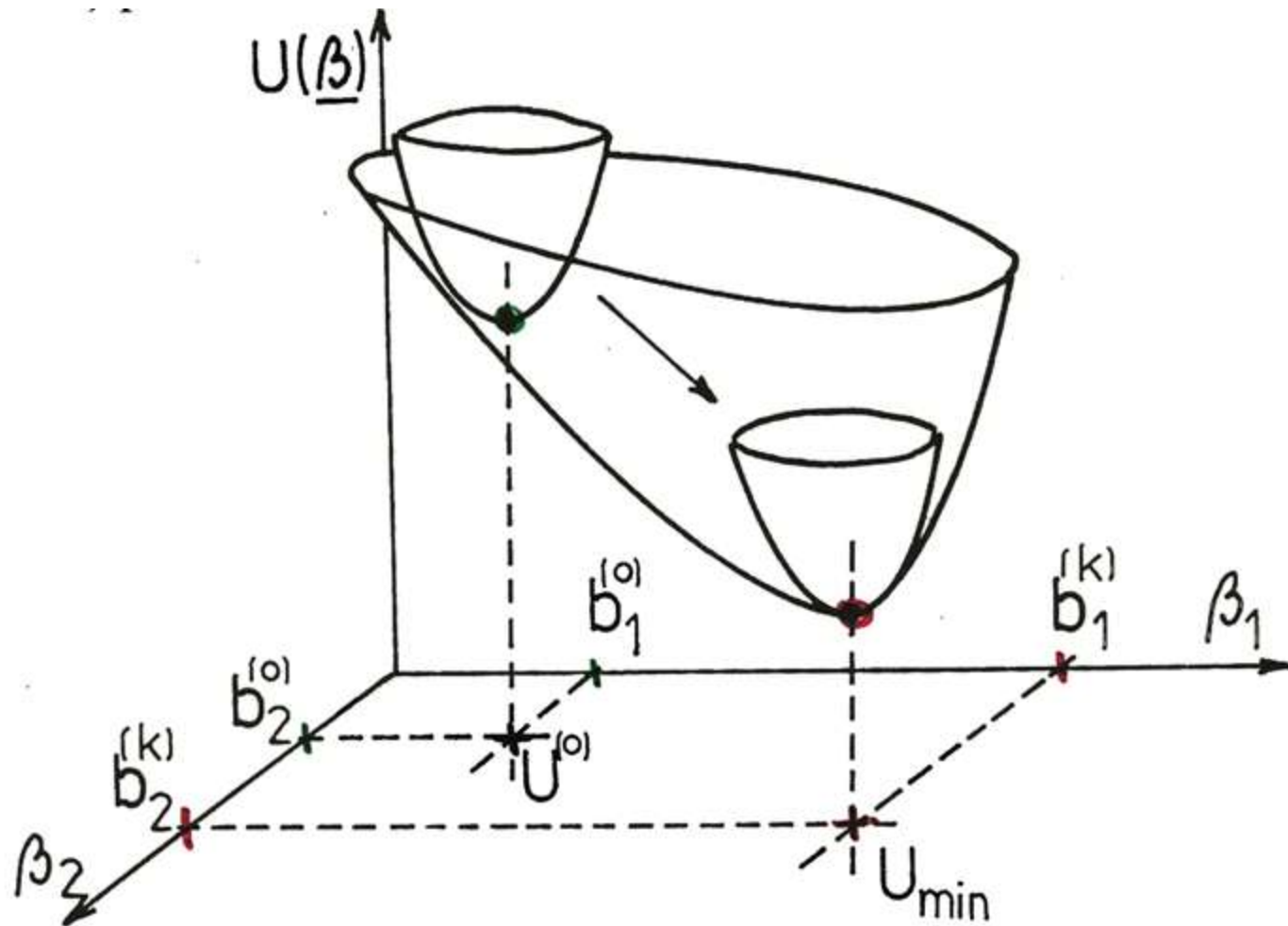






Algoritmus LETAGROP: „mapování dolíčku“ m -rozměrným eliptickým hyperparaboloidem.
Po dosazení vyjde $(m + 1)(m + 2)/2$ „nástřelů“ (=lineárních rovnic) pro určení koeficientů aproximačního paraboloidu.

Geometrická interpretace hledání minima $U(\beta)$ (LETAGROP)



31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



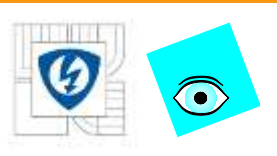
Kapitola 8.2

TĚSNOST PROLOŽENÍ

31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Statistická analýza reziduí

Pro aditivní modely měření: $\hat{e}_i = y_i - f(x_i, \beta)$
vektor reziduí $\hat{e}_i$ souvisí s vektorem chyb ε podle

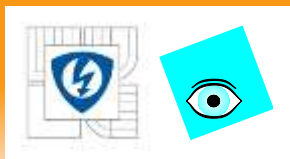
$$\hat{e} \approx (E - P)\varepsilon$$

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

Každé reziduum je přibližně lineární kombinací všech chyb

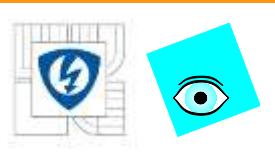
$$\hat{e}_i = \varepsilon_i - \sum_{k=1}^q P_{ik} \varepsilon_k$$

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



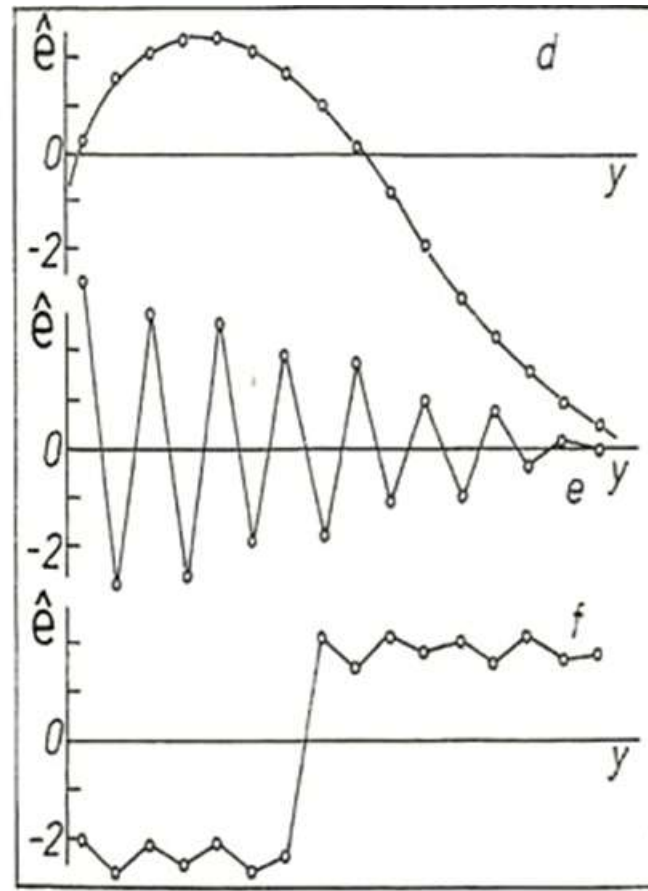
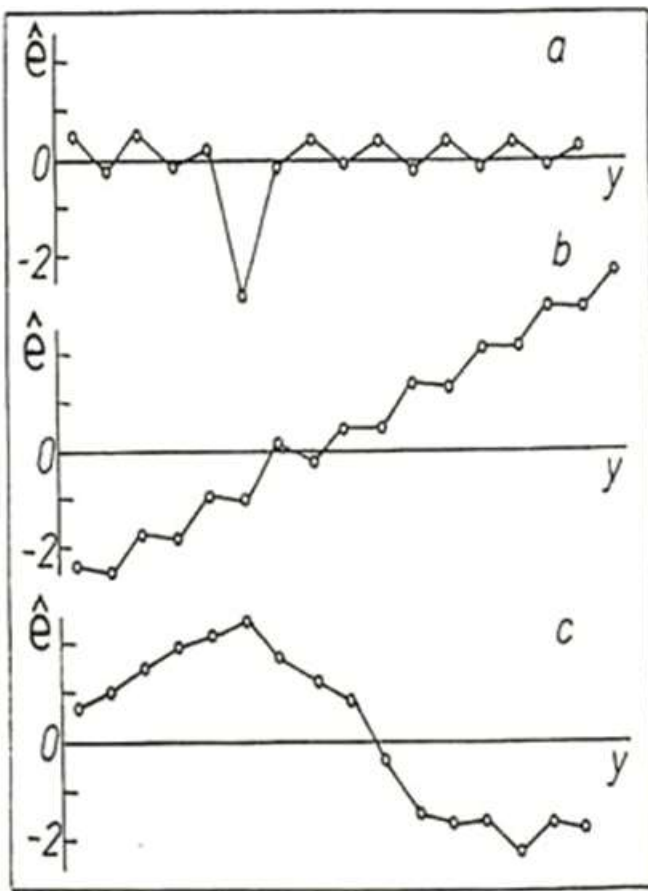
Grafická analýza reziduí

- a) odlehlé (extrémní) hodnoty v souboru reziduí,
- b) trend v reziduích,
- c) nedostatečné střídání znaménka 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 .



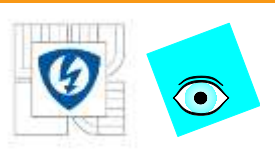
Grafická analýza reziduí

(a) odlehlá hodnota v datech, b) trend v reziduích, c) nedostatečné střídání znaménka reziduí, d) chybný model, e) heteroskedasticita, f) náhlá změna podmínek



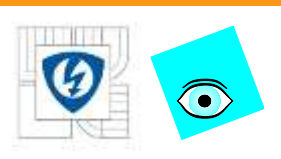
31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



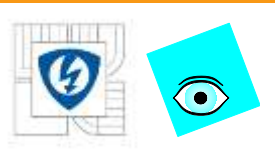
Numerická analýza reziduí

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.



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

- a) **Testační charakteristika $T_{1,1}$** by měla být (přibližně) rovna nule, protože obvykle platí, že $\hat{e}^T f(x_i, \beta) = 0$.
- b) **Testační charakteristika $T_{2,1}$** ukazuje na heteroskedasticitu
- c) **Testační charakteristika $T_{1,2}$** ukazuje na chybně navržený model
- c) **Testační charakteristika $T_{1,0}$** by měla být přibližně rovna nule.



Analýza vlivných bodů

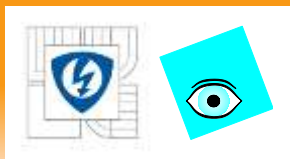
Vychází z jednokrokové aproximace odhadu $b_{(i)}$

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

kde P_{ii} jsou prvky projekční matice,

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

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



Analýza vlivných bodů

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

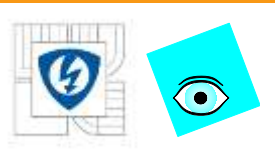
$$\hat{s}_{(i)}^2 = \frac{U(\beta) - \frac{\hat{e}_i^2}{1 - P_{ii}}}{n - m - 1}$$

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

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

b) **Jackknife rezidua** $\hat{e}_{ji} = \frac{\hat{e}_i}{\hat{s}_{(i)}\sqrt{1-P_{ii}}}$

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



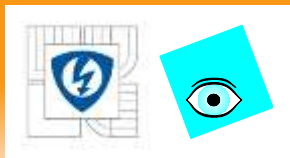
c) Věrohodnostní vzdálenost

$$LD_i = 2[\ln L(\beta) - \ln L(\beta_{(i)})]$$

Pro případ MNČ bude věrohodnostní vzdálenost ve tvaru

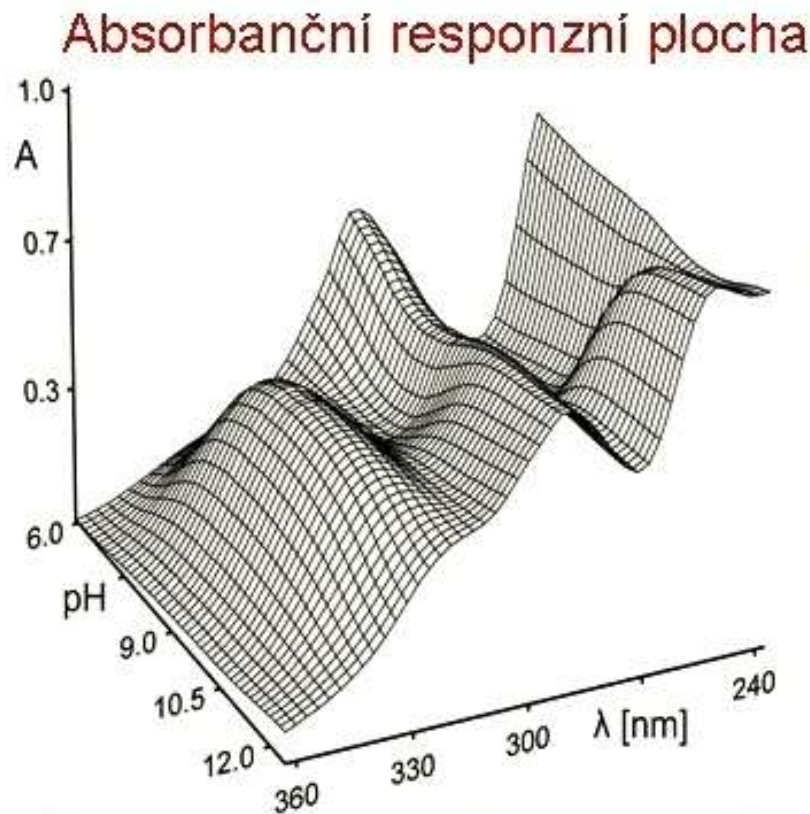
$$LD_i = n \cdot \ln \left[\frac{U(\beta_{(i)})}{U(\beta)} \right]$$

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

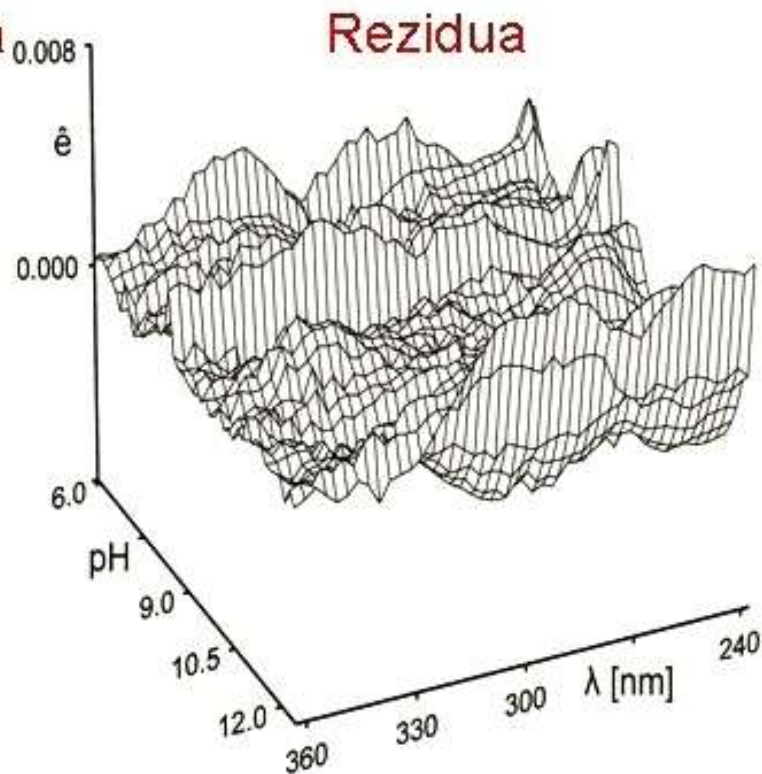


Input dat, grafická analýza reziduí

Input dat

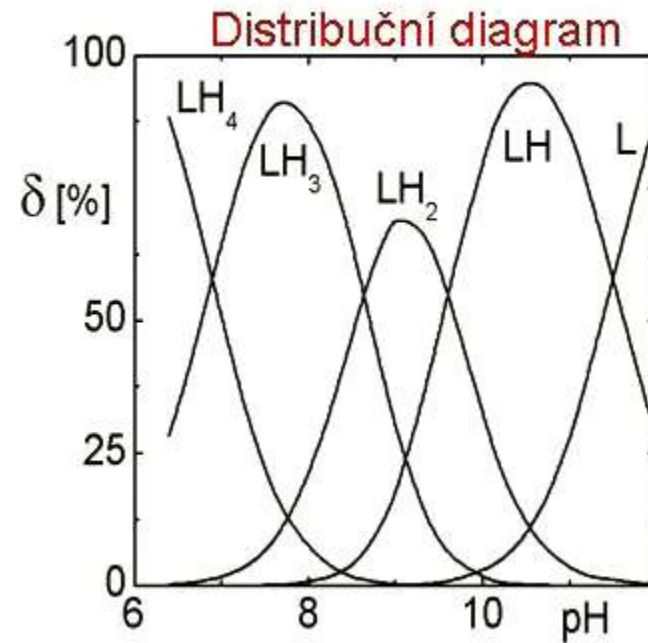
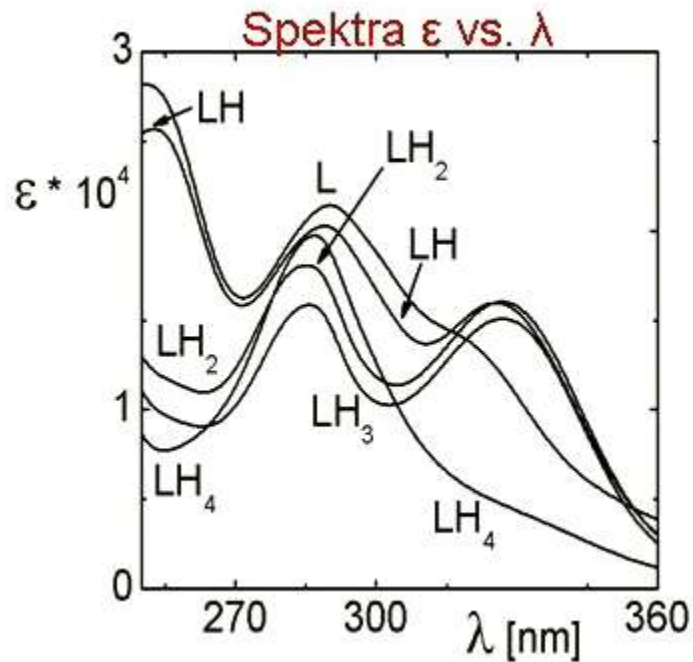


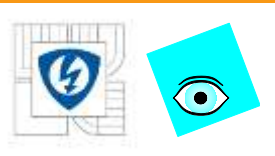
Grafická analýza reziduí





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

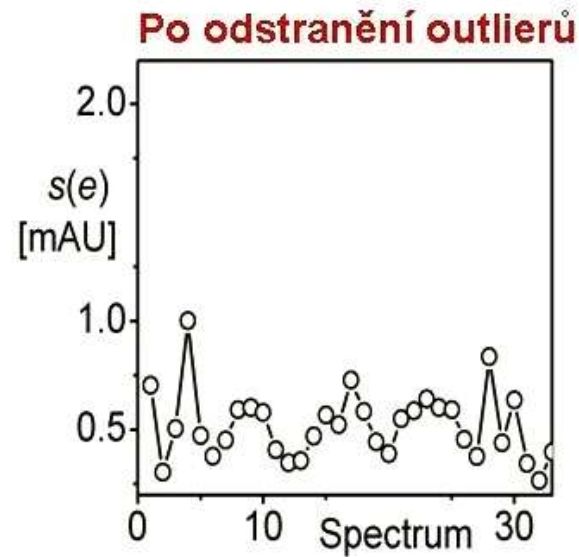
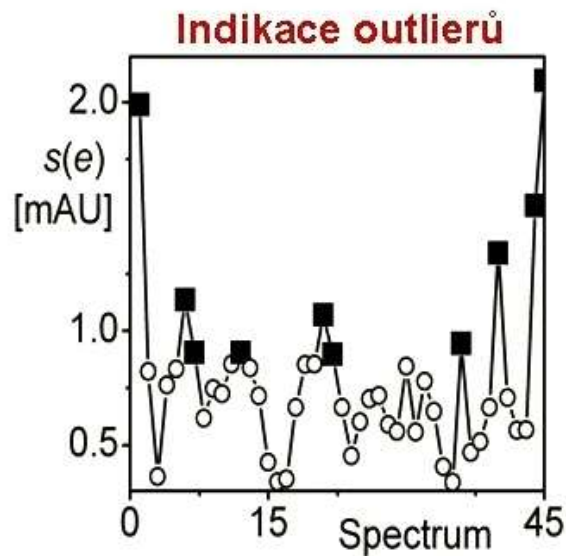


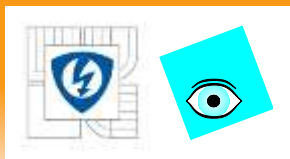


Outliery

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.

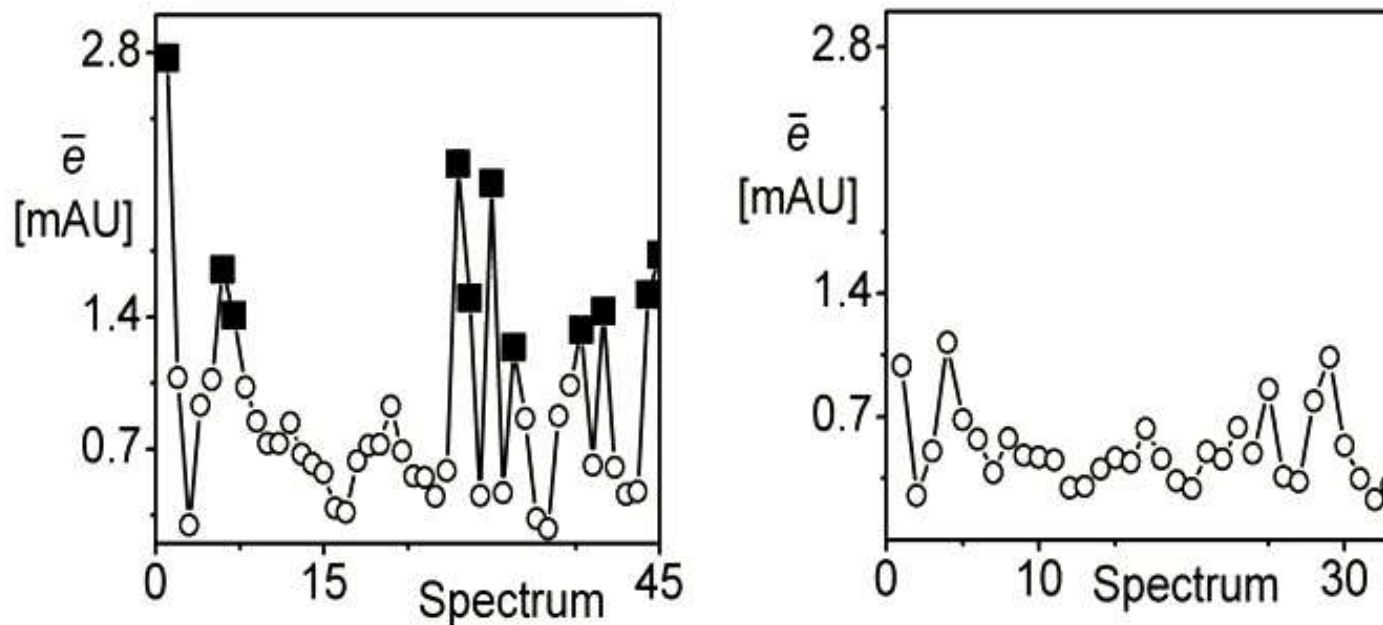
Těsnost proložení statistickou analýzou reziduí

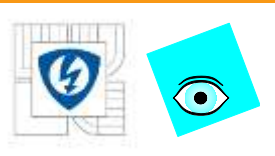




Míra těsnosti proložení

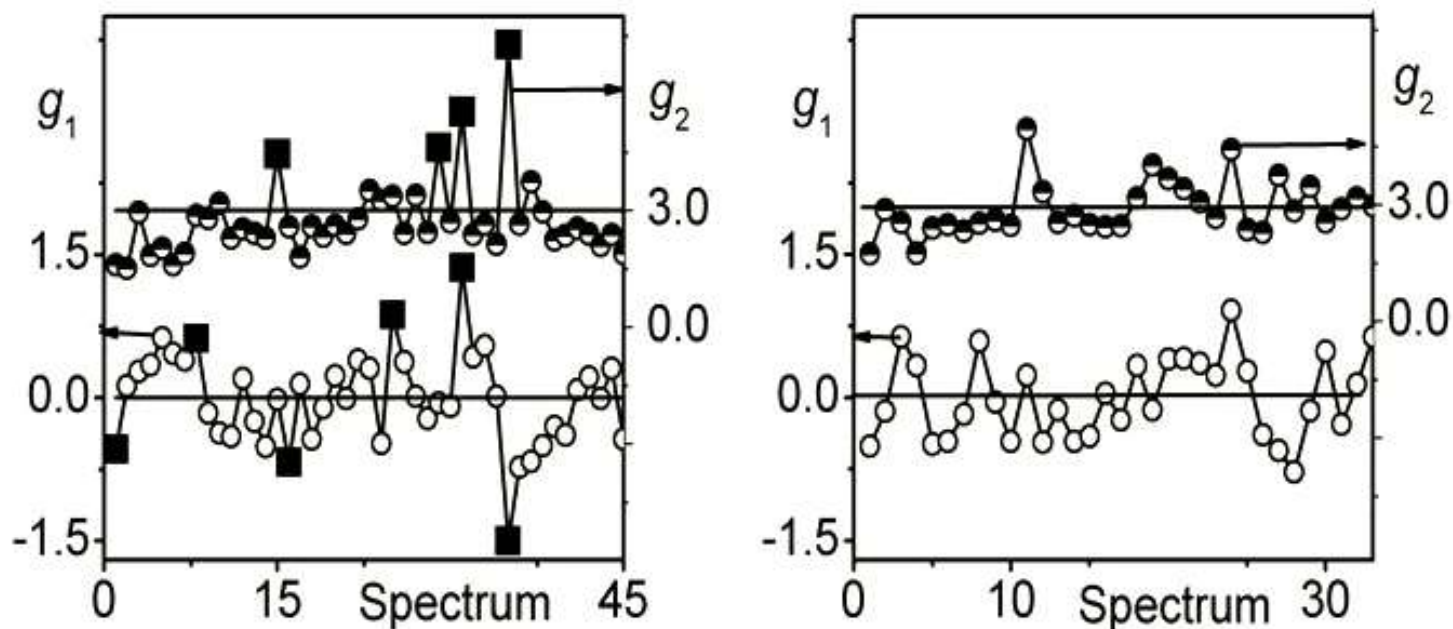
Míru těsnosti proložení je vhodné posuzovat dle velikosti reziduí a jejich rozdělení, a to především dle střední hodnoty reziduí a symetrie rozdělení (vyjádřené koeficientem šikmosti a špičatosti).

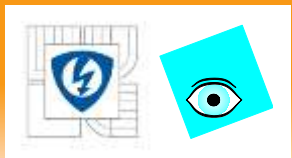




Koeficient šikmosti

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



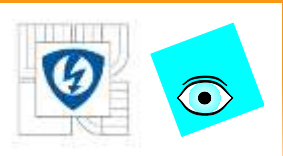


SOUHRN

31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





1. Kvalita nalezených odhadů parametrů

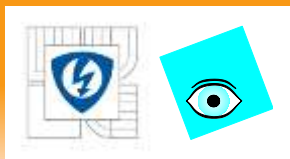
Kvalita se posuzuje podle

- a) jejich intervalů spolehlivosti
- b) rozptylů $D(\beta_j)$

Pravidlo „3 sigma“: $3\sqrt{D(\beta_j)} < |\beta_j|$

2. Kvalita dosažené těsnosti proložení

- a) Reziduální rozptyl $\hat{\sigma}^2 = U(\beta)/(n - m)$
- b) Koeficient determinace D , který je pro lineární modely čtvecem vícenásobného korelačního koeficientu



Postup při testování navrženého modelu

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

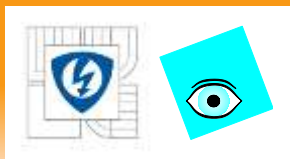
Stonásobek koeficientu determinace je **regresní rabat** $100D$.

c) K rozlišení mezi modely užijeme **Akaikova informačního kritéria** AIC

$$AIC = -L(\beta) + 2m$$

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

$$AIC = n \cdot \ln \left[\frac{U(\beta)}{n} \right] + 2m$$



3. Predikční schopnost modelu

Postup: data se rozdělí na dvě podskupiny

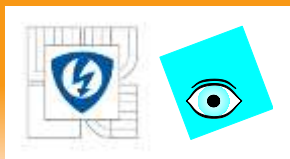
M_1 (s indexy $i = 1, \dots, \text{int}(n/2)$) s odhady $\beta(M_1)$

M_2 (s indexy $i = \text{int}(\frac{n}{2}) + 1, \dots, n$) s odhady $\beta(M_2)$

a) Predikční schopnost modelu

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

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



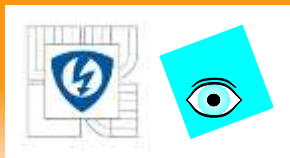
b) Střední kvadratická chyba predikce je definována

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

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

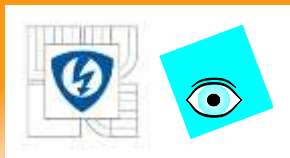
Whitův test

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

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

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

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

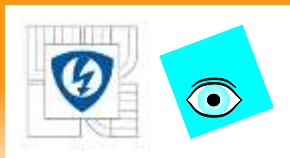


Postup při testování navrženého modelu

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.

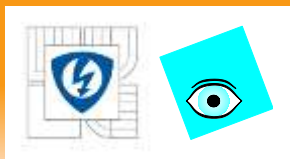


Postup tvorby nelineárního regresního modelu

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 nalezení odhadů se posuzuje dle jejich intervalů spolehlivosti nebo pouze jejich rozptylů $D(\beta_j)$

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

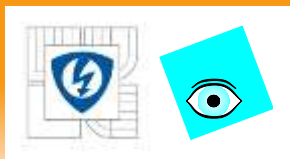


Grafické posouzení vhodnosti modelu

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 $g_1(\hat{e})$ a špičatosti $g_2(\hat{e})$.



Základní statické charakteristiky

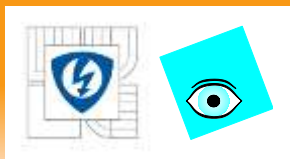
O přiblížení modelu experimentálními daty informuje suma čtverců reziduí v minimu $U(\beta)$, ze které se vyčíslí reziduální rozptyl $\hat{\sigma}^2 = U(\beta)/(n - m)$.

Z $U(\beta)$ je odvozen koeficient determinace D a regresní rabat $100D[\%]$

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

Hamiltonův R – faktor je definován

$$R - \text{faktor} = \sqrt{\frac{U(\beta)}{\sum_{i=1}^n y_i}}$$



Základní statické charakteristiky

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

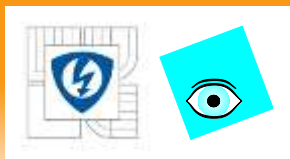
$$R - faktor = \sqrt{(1 - D) - \frac{(1 - D)n\bar{y}^2}{\sum_{i=1}^n y_i^2}}$$

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

Mezi modely rozliší **Akaikovo informační kritérium**

$$AIC = -L(\beta) + 2m$$

Optimální je model, pro který je AIC minimální. Platí přitom $AIC = n \cdot \ln \left[\frac{U(\beta)}{n} \right] + 2m$.



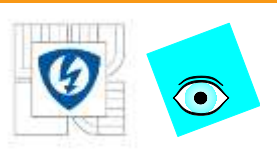
Regresní diagnostika

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

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



Analýza vlivných bodů

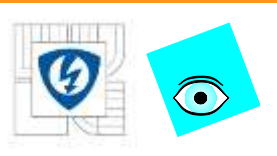
U lineárních regresních modelů: je odhalení vlivných bodů VB pomocí reziduí $\hat{e}_i$ a prvků P_{ii} projekční matice

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

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

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

kde J je Jakobián, čili derivace 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.

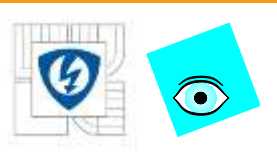


Analýza vlivných bodů

Vychází se proto z jednokrokové aproximace odhadu $\beta_{(i)}$

$$\beta_{(i)}^1 = \beta - \frac{(J^T J)^{-1} J_i \hat{e}_i}{1 - P_{ii}}$$

kde P_{ii} jsou prvky projekční matice P .



Analýza vlivných bodů

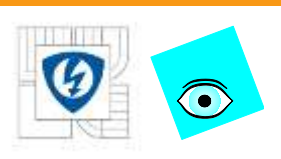
Charakteristika DFS_{ij} vyjadřuje vliv i -tého bodu na

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

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

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

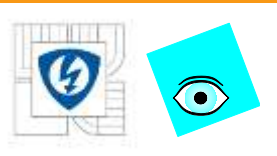
Test: i -tý bod je vlivný, když je $DFS_{ij} > 2/\sqrt{n}$.



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 $h_{(i)}$ při vynechání i -tého bodu nebo změny střední hodnoty i -tého rezidua při vynechání i -tého bodu



Analýza vlivných bodů

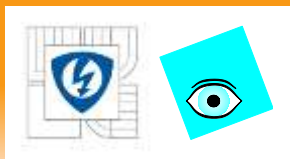
Mezi nelineární míry vlivu i -tého bodu na odhady parametrů patří **věrohodnostní vzdálenost**

$$LD_i = 2[\ln L(\beta) - \ln L(\beta_{(i)})]$$

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

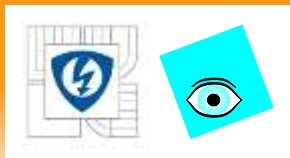
Do obou vztahů lze dosadit buď odhady $\beta_{(i)}$, určené regresí při vynechání i -tého bodu, nebo $\beta_{(i)}^1$, určené z jednokrokové aproximace.

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



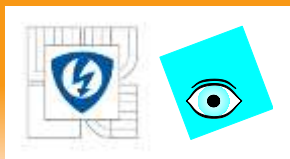
Analýza vlivných bodů

- (a) VB ovlivňují odhady parametrů a relativní vychýlení 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“.



Mapa citlivostní funkce

- U nelineárních modelů existuje celá řada komplikací
- neodhadnutelnost některých parametrů,
 - existence minima funkce $U(\beta)$ jen pro některé regresní modely,
 - výskyt kokálních minim,
 - existence sedlových bodů, ovlivňujících kritériální funkci $U(\beta)$ a
 - špatná podmíněnost parametrů v regresním modelu.

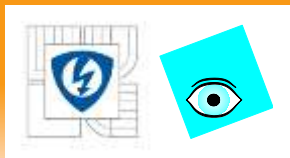


Mapa citlivostní funkce

Problémy lze indikovat analýzou normovaných citlivostních koeficientů

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

Pro vizuální posouzení špatné podmíněnosti např. multikolinearitou mezi parametry β_j, β_h , se konstruuují ***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 .



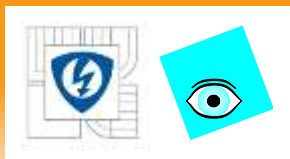
Mapa citlivostní funkce

Citlivost regresních modelů na změnu parametru β_j vyjadřuje *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 β_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.



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 $i = \text{int}(\frac{n}{2}), \dots, n$).

Označí se:

- odhady parametrů z bodů podskupiny M_1 jako $\beta(M_1)$ a
- odhady parametrů z bodů podskupiny M_2 jako $\beta(M_2)$.

Predikční schopnost modelu se vyjádří kritériem K

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

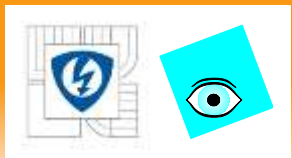
Test: Predikční schopnost modelu je tím vyšší, čím víc se K blíží k 1.



Kritérium střední kvadratická chyba predikce

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

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



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.).