NELINEÁRNÍ REGRESNÍ MODELY

Základní úlohy:

1. Konstrukce ***kalibračních modelů***,
2. Ověření ***teoretických modelů*** fyzikálně-chemické zákonitosti,
3. Tvorba ***empirických modelů***,

Tvorba regresního modelu f(x, β) čili funkce

a) vektoru **nastavovaných proměnných** (***deterministických, kontrolovatelných, vysvětlujících, nezávislých***) x,
tj. bodů $\{\mathbf{x}_i^T, y_i\}$, i=1, ..., n,

b) vektoru **parametrů β** o rozměru (m × 1), $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)^T$.

c) y je **vysvětlovaná proměnná** (závisle p., odezva, měření, pozorování) na zvolenou kombinaci nastavovaných veličin $\mathbf{x}_i$.

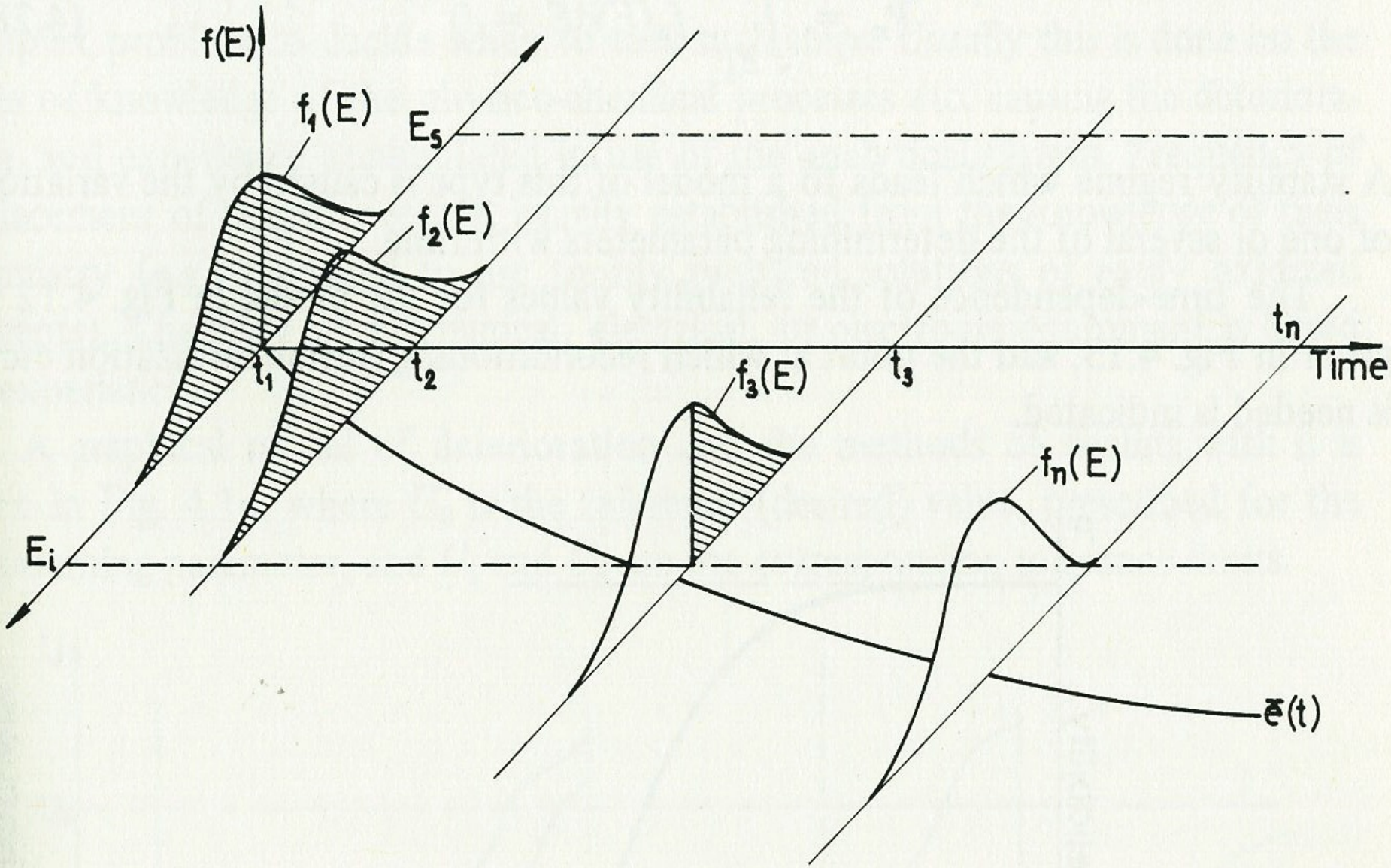


Fig. 4.12 The model of increase of analysis errors with time, for an analytical system with a time-dependent parameter.

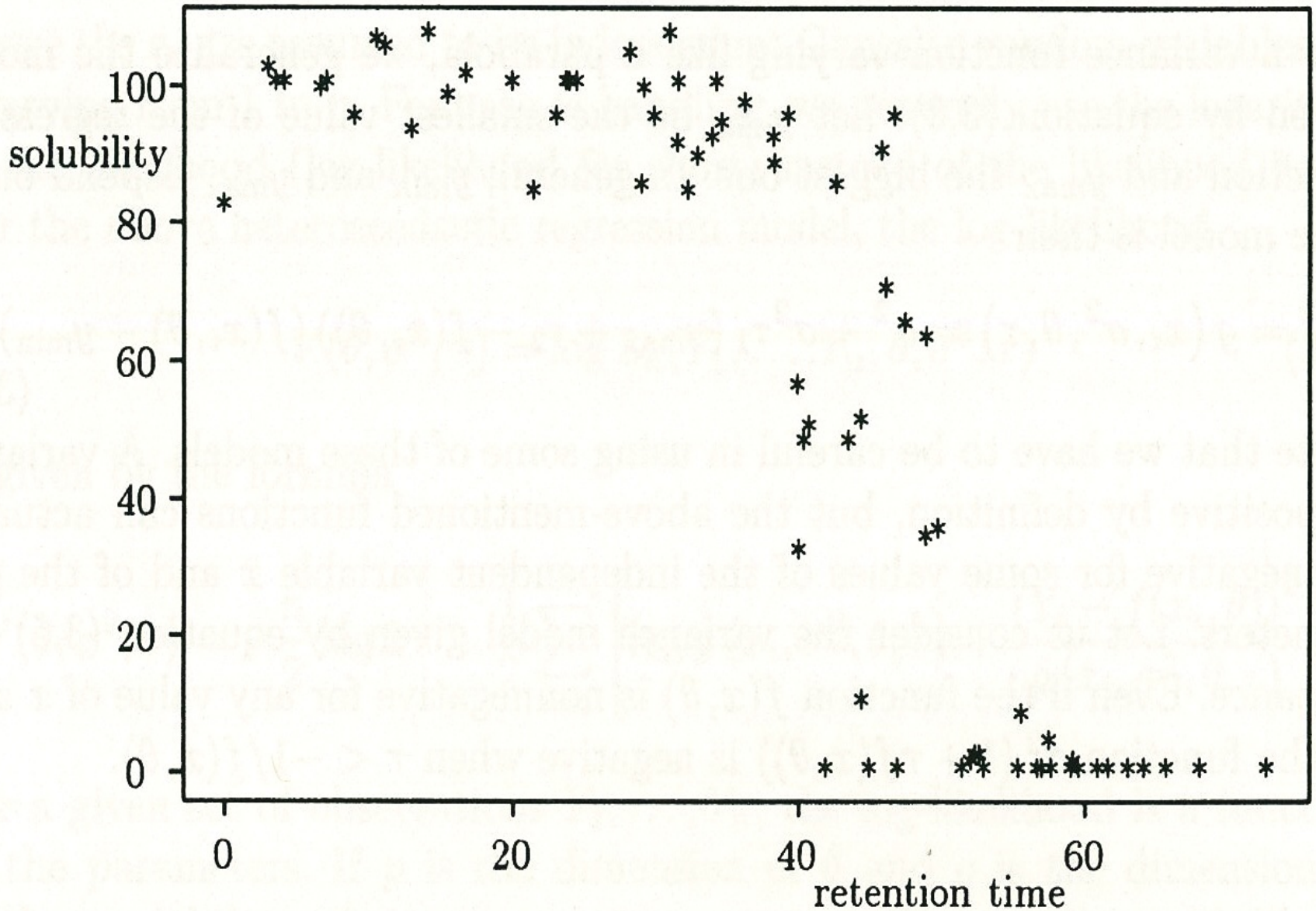


FIGURE 3.3. Peptides example: observed solubility of 75 peptides versus their RP-HPLC retention time

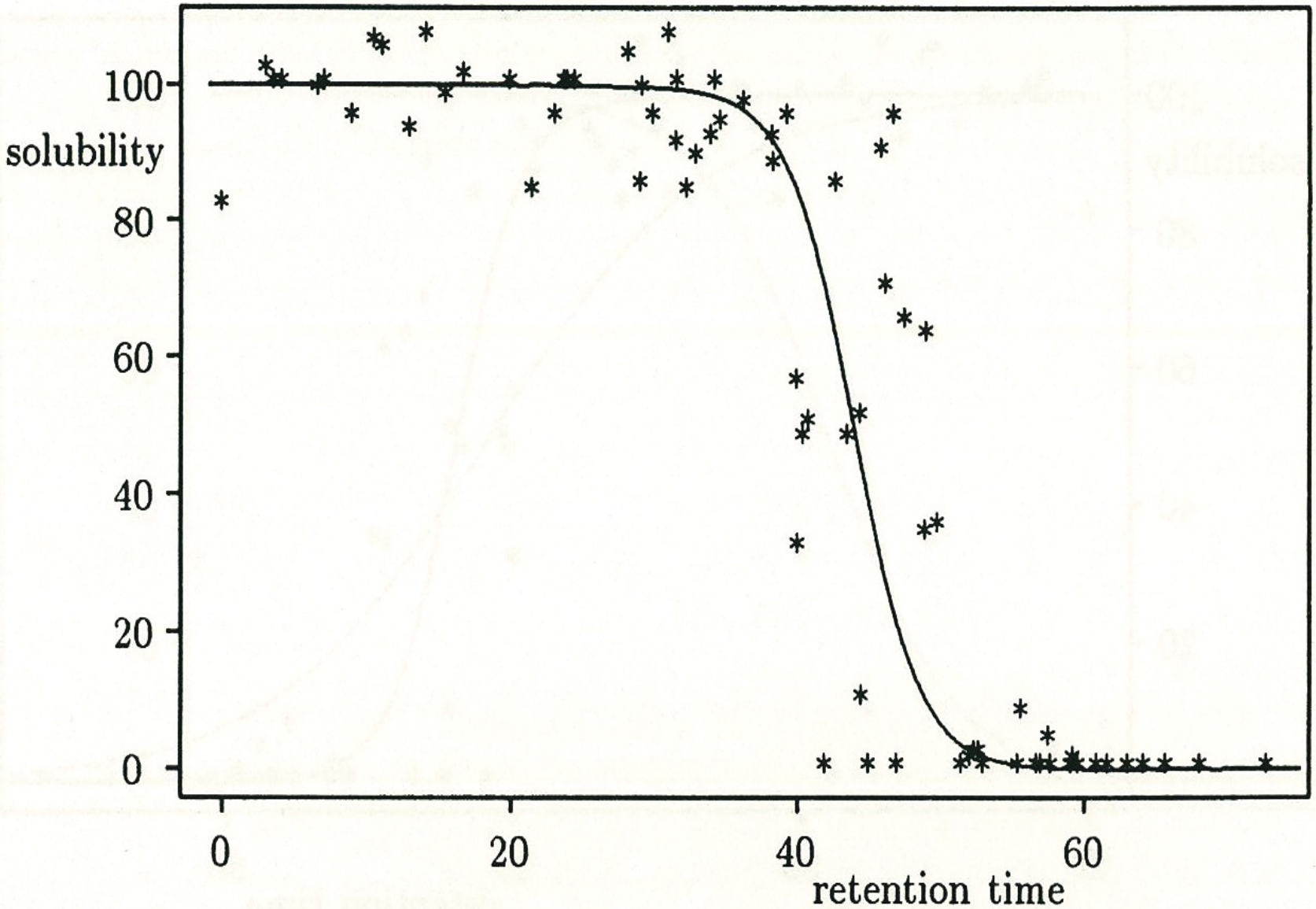


FIGURE 3.7. Peptides example: observed and adjusted solubilities of 75 peptides versus their RP-HPLC retention time

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.
   (nepř. rovnovážné konstanty (disociační, stability, součiny rozpustnosti) reakčních produktů, rychlostní konstanty u kinetických modelů, neznámé koncentrace, atd.

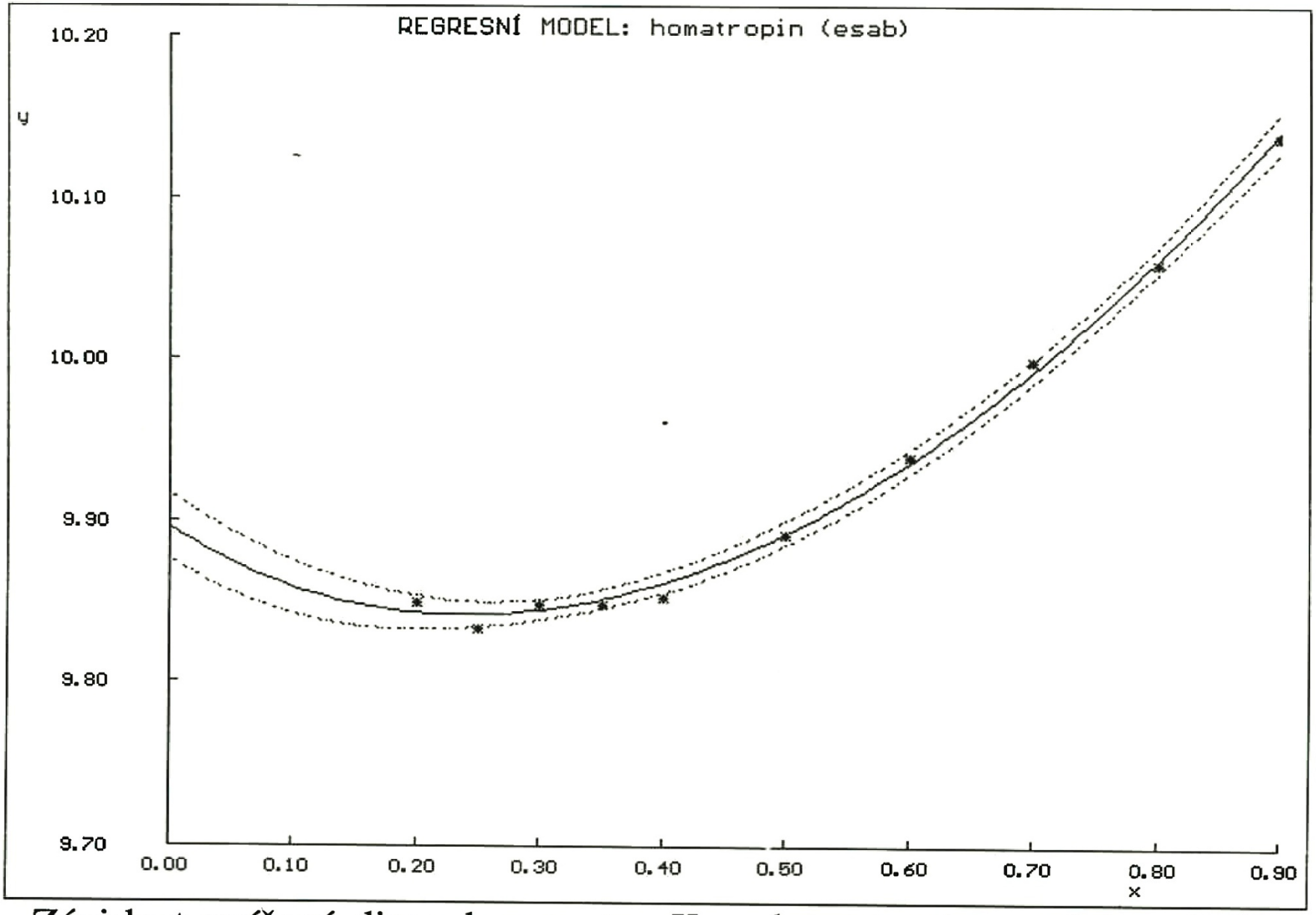
Model: rozšířený Debye-Hückelův zákon

$$pK_{a,smíš} = pK_{a,T} - \frac{0.5115\sqrt{I}}{1 + 3.29\times10^{10}\,å\sqrt{I}} + C\,I$$

Proměnné:

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

Neznámé parametry: β_1: $pK_{a,T}$, β_2: $å$, β_3: C



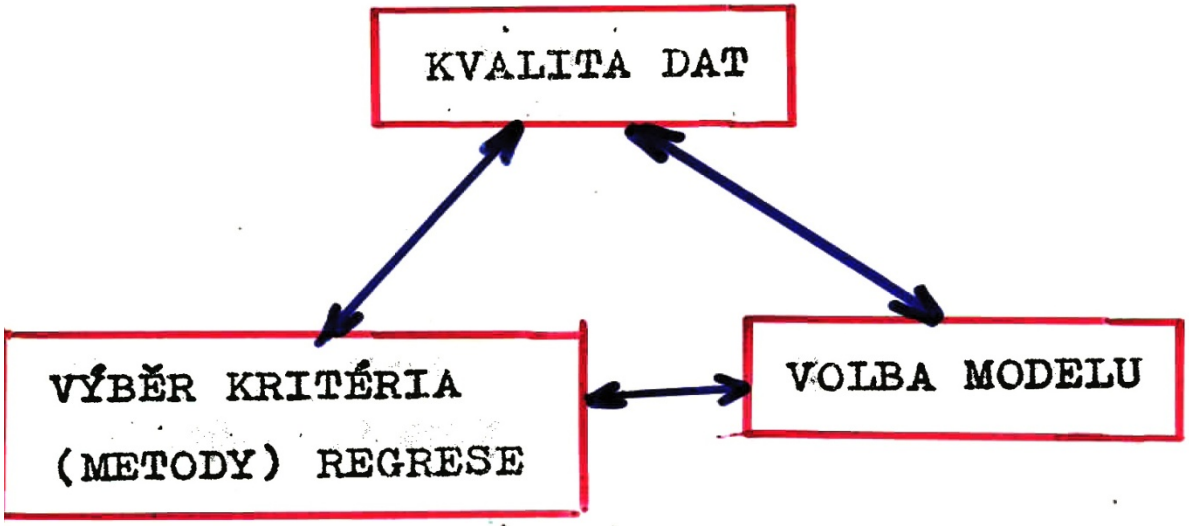
Závislost smíšené disoc. konstanty $pK_{a,smíš}$ homatropinu na iontové síle:

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

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

POSTUP A DIAGNOSTIKA REGRESE:

Regresní triplet:



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

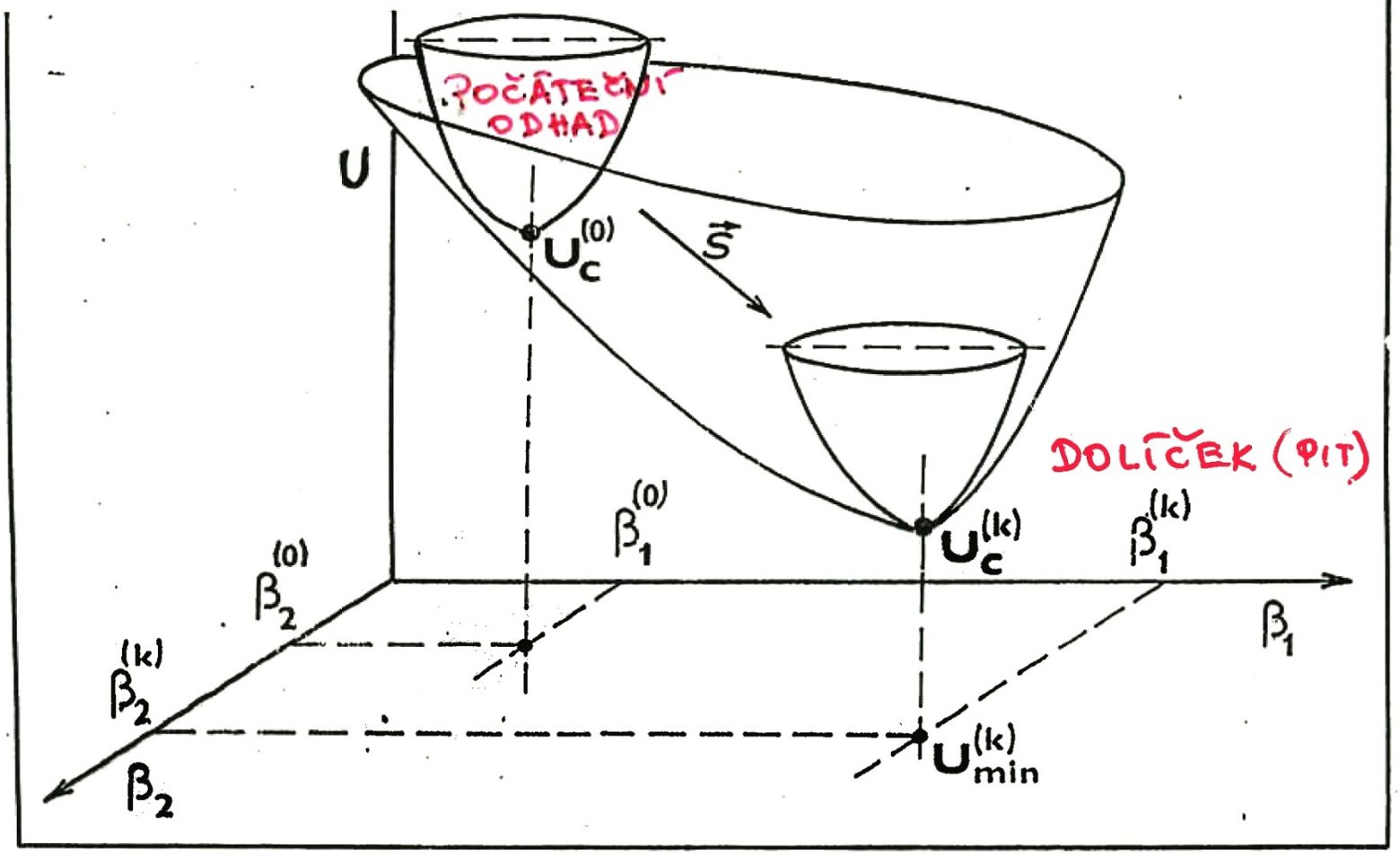
kde $y_{vyp,i} = f(x_i; b_1, \ldots, b_m)$

GEOMETRICKÉ ZNÁZORNĚNÍ KRITÉRIA U

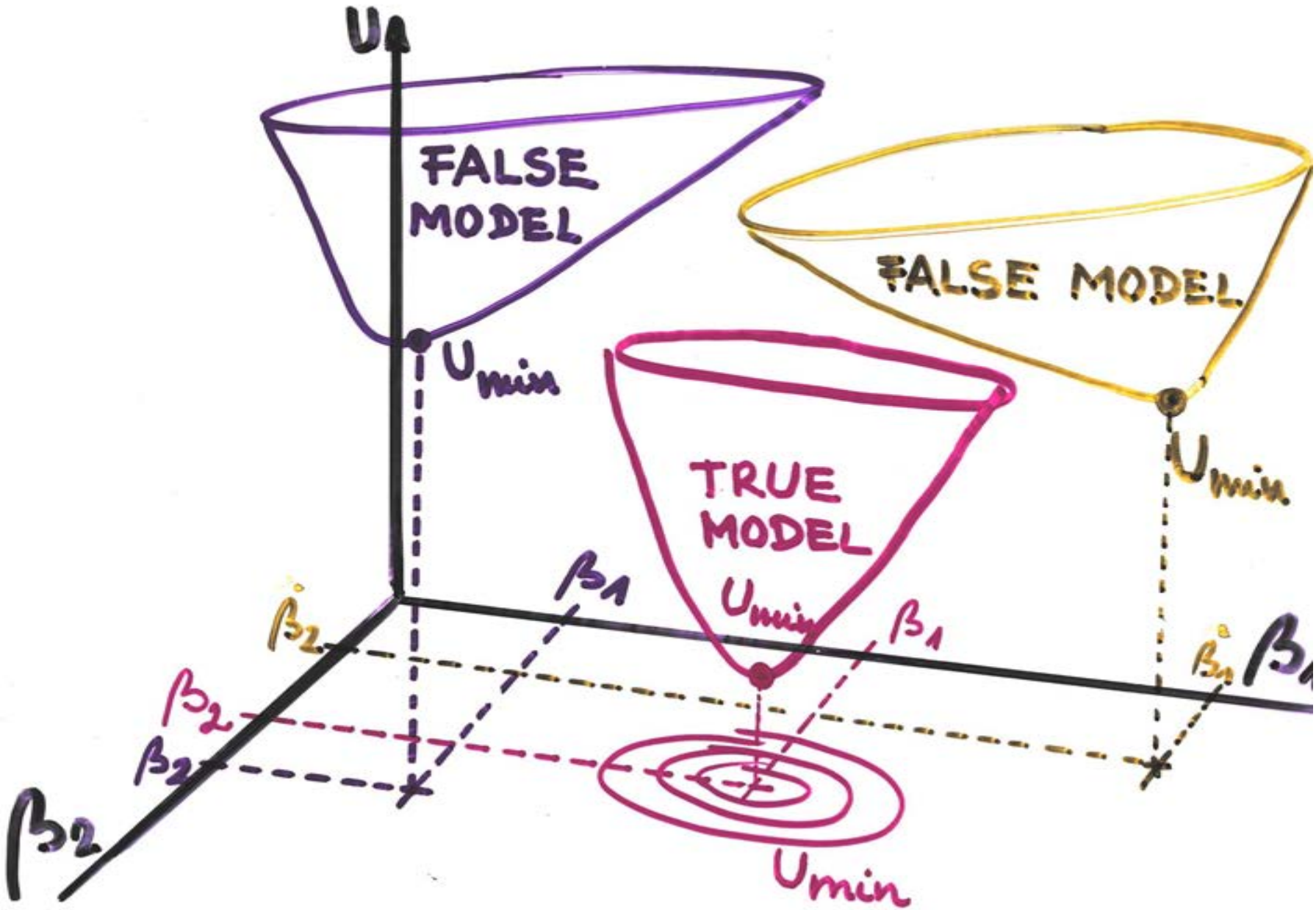
$$U = \sum_{i=1}^{n} w_i \left(y_{exp,i} - y_{vyp,i} \right)^2 \approx \text{minimum}$$

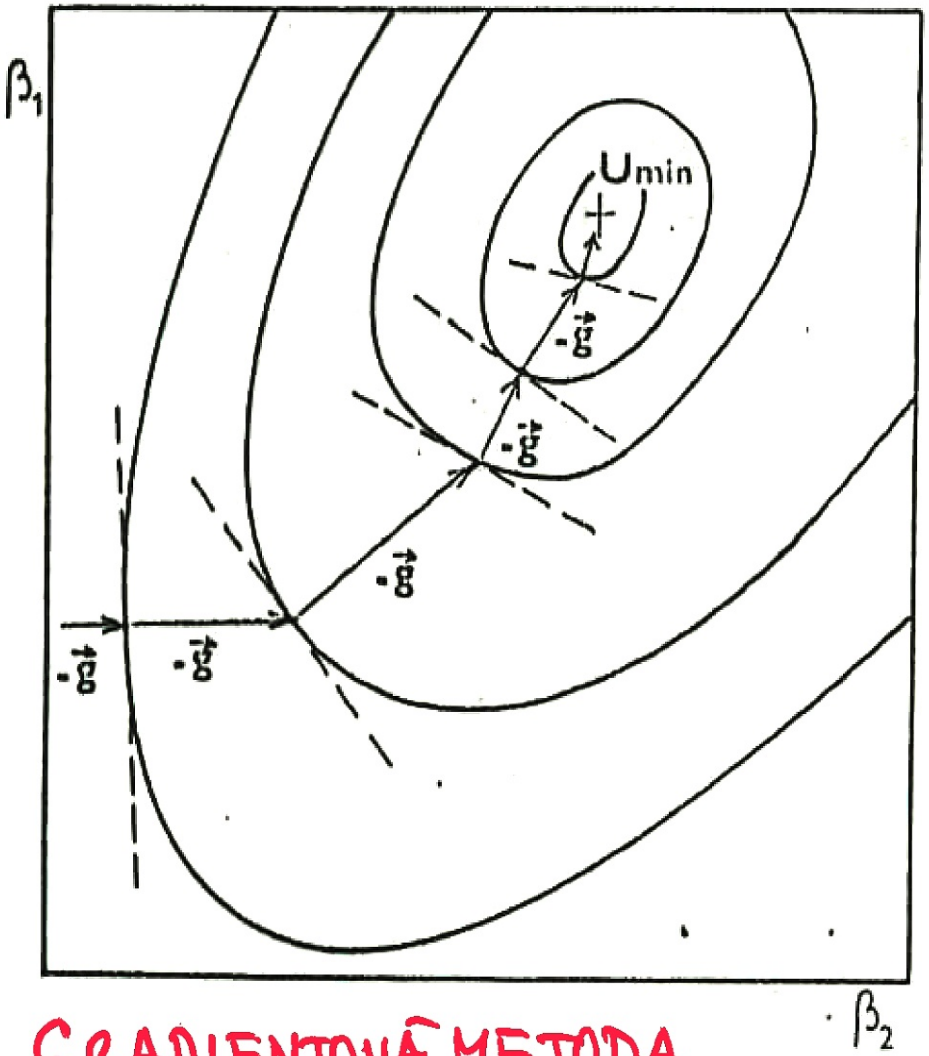
$$y_{vyp,i} = f(x_i; \beta_1, \dots, \beta_m; p_1, \dots, p_l)$$

OPTIMALIZAČNÍ PROBLÉM V (m+1) ROZMĚRNÉM PROSTORU



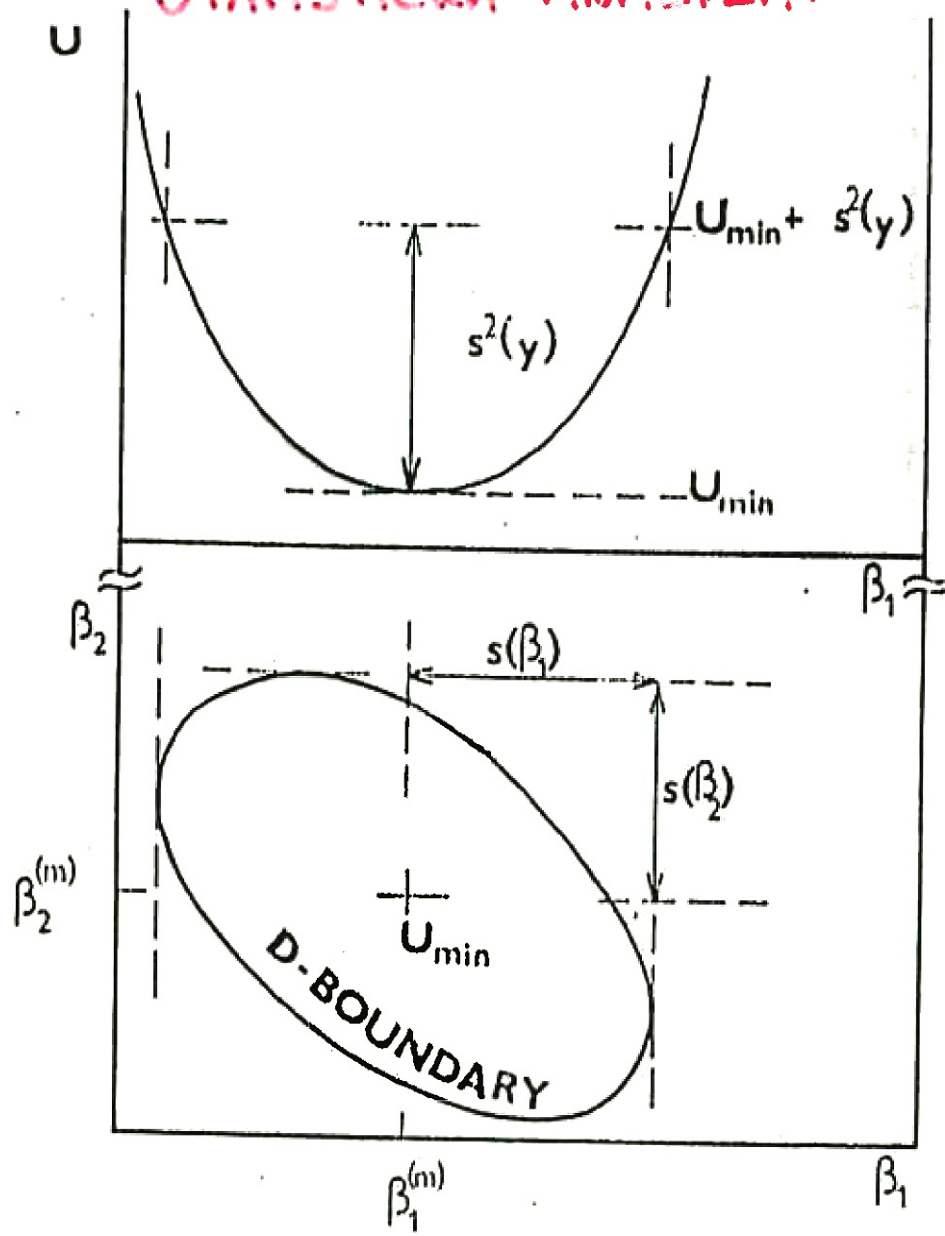
Hledání modelu (Regression Model Building)





GRADIENTOVÁ METODA HLEDÁNÍ MINIMA

STATISTICKÁ ANALÝZA:



Intervaly spolehlivosti parametrů

1. Vyčíslí se 100(1 - α)%ní interval spolehlivosti parametru β_j

$$b_j - \hat{\sigma}\sqrt{V_{jj}}\, t_{1-\alpha/2}(n - m) \leq \beta_j \leq b_j + \hat{\sigma}\sqrt{V_{jj}}\, t_{1-\alpha/2}(n - m)$$

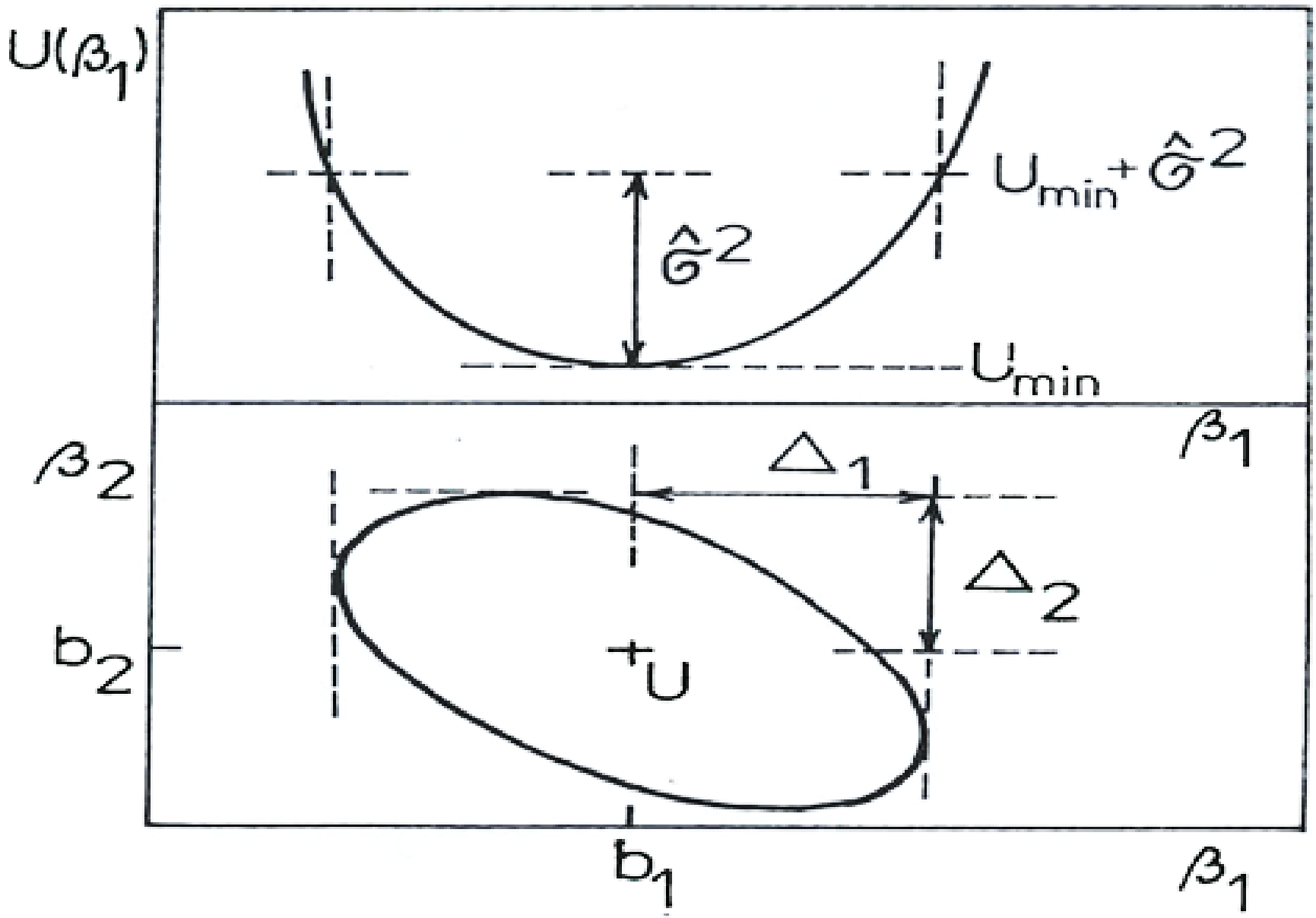
2. Vhodnější je určovat intervaly spolehlivosti parametru β_k na základě maximální délky Δ_k průmětu Δ_{jk} do osy parametru β_k:

$$b_k - \Delta_k \leq \beta_k \leq b_k + \Delta_k$$

Pro intervaly spolehlivosti pak platí

$$b_k - p\sqrt{V_{kk}} \leq \beta_k \leq b_k + p\sqrt{V_{kk}}$$

Pro m = 1 jsou všechny tyto intervaly totožné.



Odhady chyb Δ_1 a Δ_2 parametrů β_1 a β_2

Formulace modelu

1. Lineární regresní model: je lineární kombinací parametrů, např. model

$$f(\mathbf{x}, \beta) = \beta_1 + \beta_2 \sin(x)$$

Pro lineární regresní modely platí podmínka

$$g_j = \frac{\delta f(\mathbf{x}, \beta)}{\delta \beta_j} = \text{konst.} \qquad j = 1, \ldots, m$$

2. Nelineární regresní model: alespoň pro jeden parametr β_j je parciální derivace g_j jeho funkcí,

Dělí se:

1. ***Neseparabilní 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 alespoň **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,

$$\gamma = g(\beta)$$

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. Arrheniův 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 $z = 1/x$.

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 nekonstantní rozptyl

$$\sigma^2(\ln k_i) = \sigma^2(k_i) / (\ln k_i)^2$$

tj. konstantní relativní chybu.

Míra citlivosti parametru β_j v modelu

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

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, ..., 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 ***přeurčený*** a je nutné ho zjednodušit.

Potíže redundance (přeurčení modelu):

1. Přeurčení vede vždy k singularitě matice $\mathbf{J}^T \mathbf{J}$.

2. Lokální přeurčení se neprojeví při použití pseudoinverze matice $\mathbf{J}^T \mathbf{J}$.

3. Špatná podmíněnost nelineárních modelů (redundance), je kvůli špatné podmíněnosti matice $\mathbf{J}^T \mathbf{J}$. Známe-li přibližné velikosti parametrů $\mathbf{b}^{(0)}$, můžeme sestrojit matici $\mathbf{L} = n^{-1} (\mathbf{J}^T \mathbf{J})$ s prvky

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

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

Pro stanovení podmíněnosti se matice **L** převádí do standardizované formy $\mathbf{L}^*$

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

Dle podmíněnosti matice $\mathbf{L}^*$ se usuzuje na podmíněnost daného modelu.

Míra podmíněnosti:

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

Modely chyb měření

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, $i = 1, \ldots, n$, $E(\varepsilon_i) = 0$.

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 $y_i = Z_i(\mathbf{x}_i, \varepsilon_i, \beta) \qquad i = 1, \ldots, n$

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

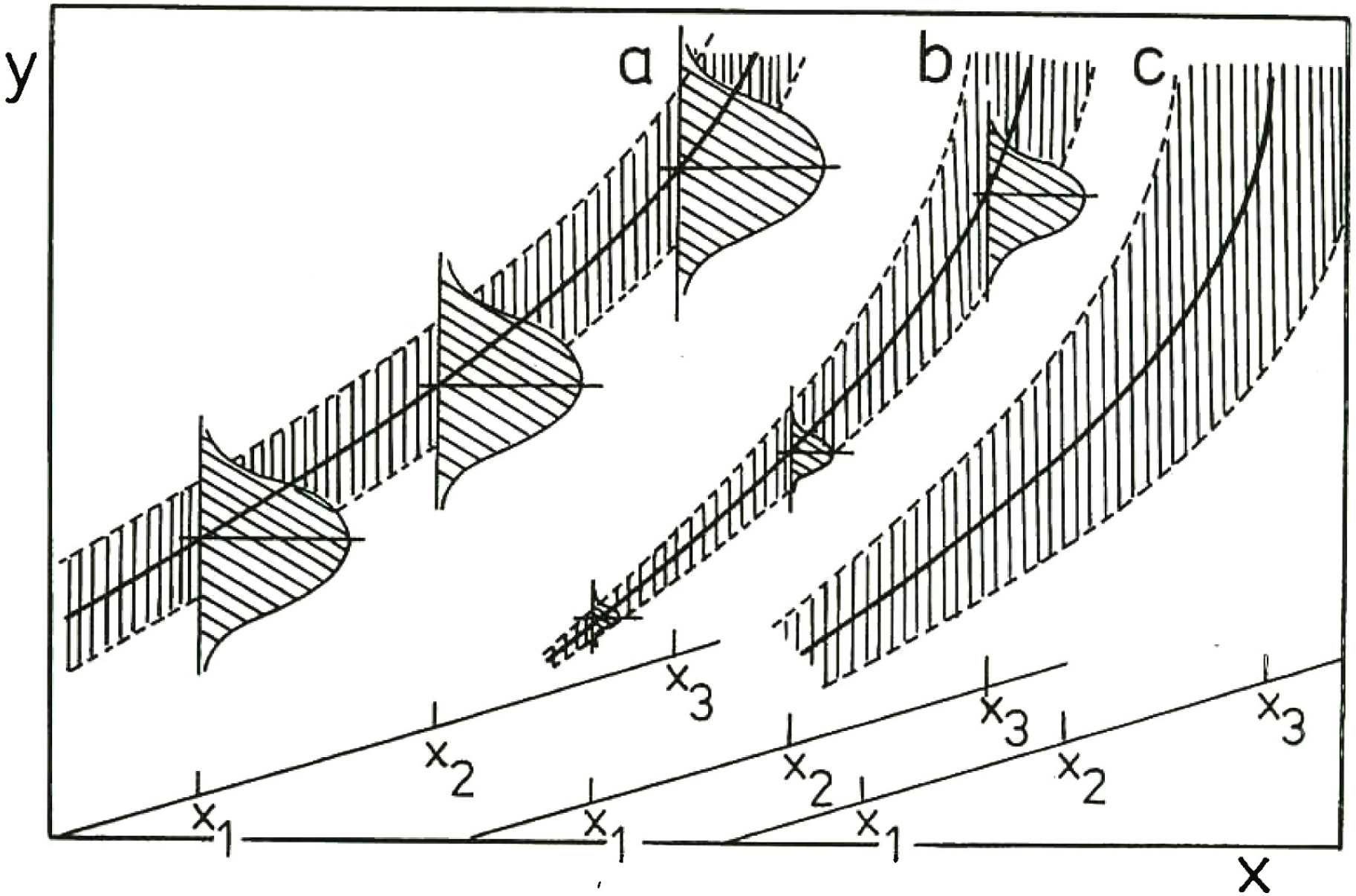
$$Z_i = f(x_i, \beta) \exp(\varepsilon_i)$$

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: (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.

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

Pro případ nezávislých náhodných chyb ε platí vztah

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

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

1. Pro ***aditivní model měření*** je

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

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

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

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

2. Pro ***multiplikativní model měření*** je

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

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

kde se předpokládají pouze kladné hodnoty měřených veličin. P
dosazení bude

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

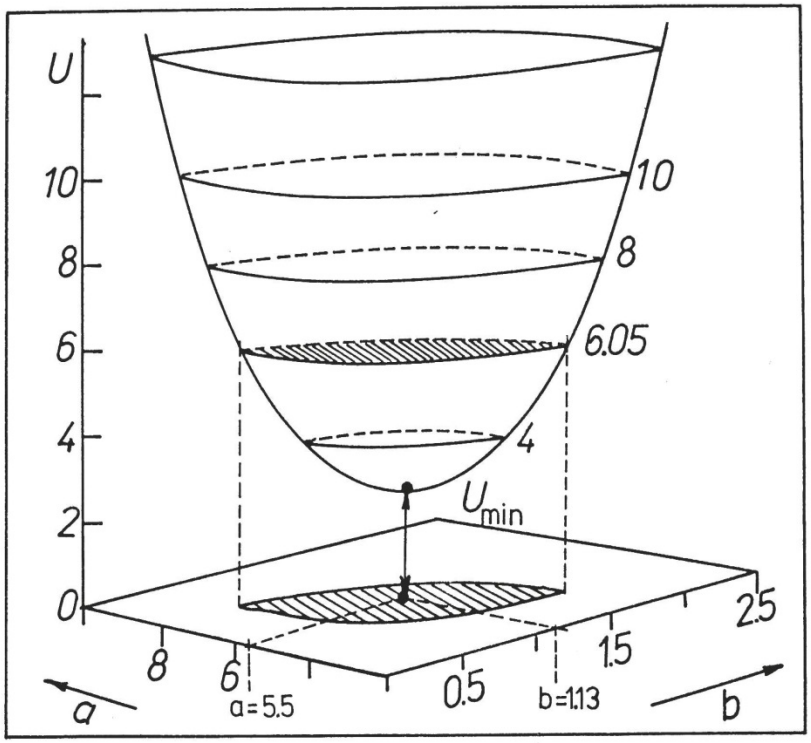
Tato hustota pravděpodobnosti již neodpovídá husto
pravděpodobnosti chyb $p_\varepsilon(.)$.

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

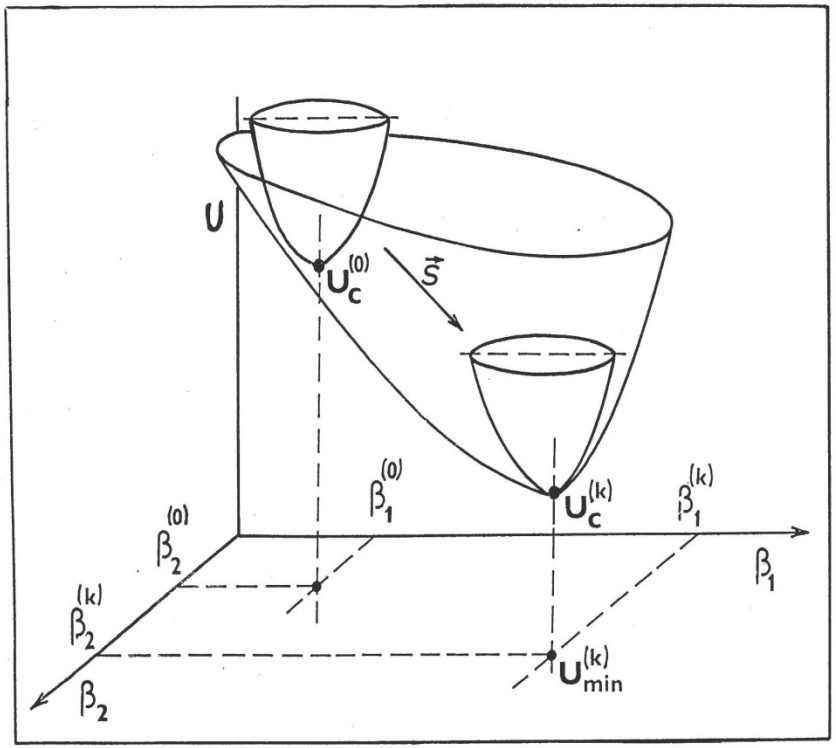


Nelineární regresní model:

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

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

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



Pro lineární regresní modely:

1. Útvar kritéria U(β) v prostoru odhadů je ***eliptický hyperparaboloid*** se středem v bodě [**b**, U(**b**)], kde nabývá minimální hodnoty.

2. Účelová funkce U(β) je vzhledem k β kvadratickou formou β^T $(\mathbf{X}^T \mathbf{X})$ β 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(**x**, β), počtu extrémů a sedlových bodů.

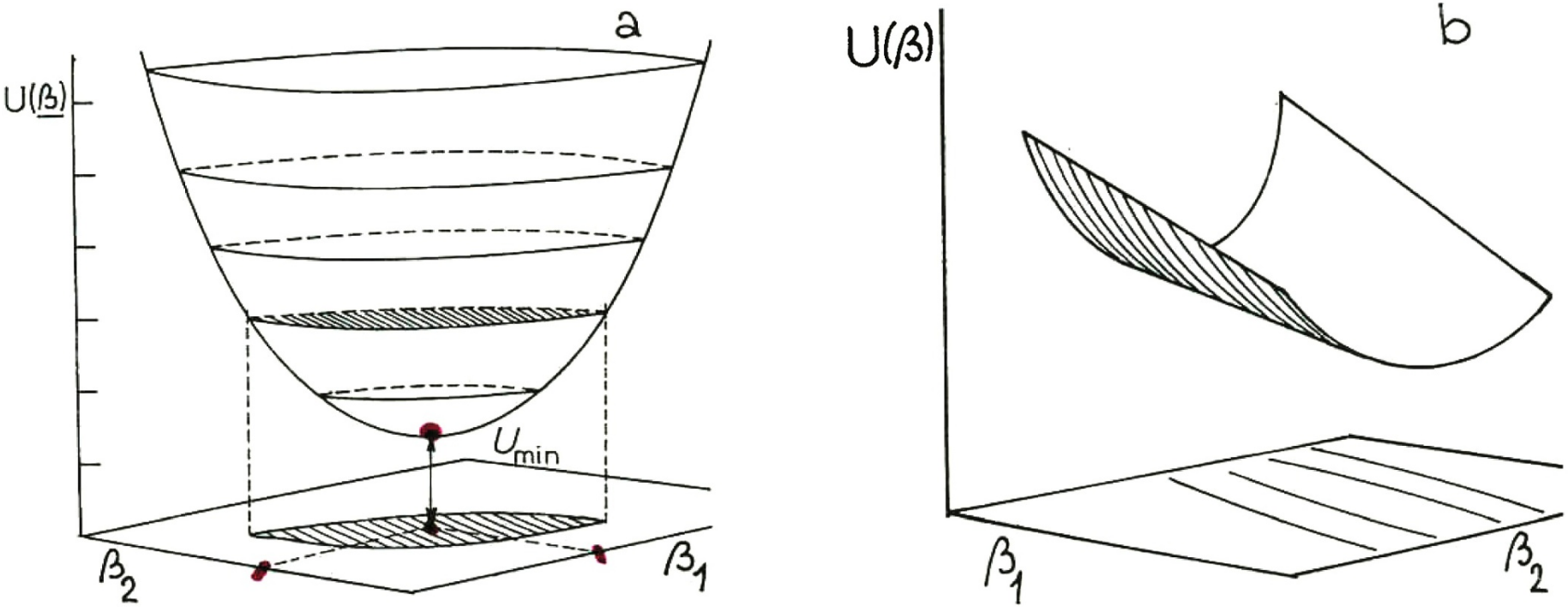
Geometrie nelineární regrese

Kritérium regrese U(β) vektorovým zápisem

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

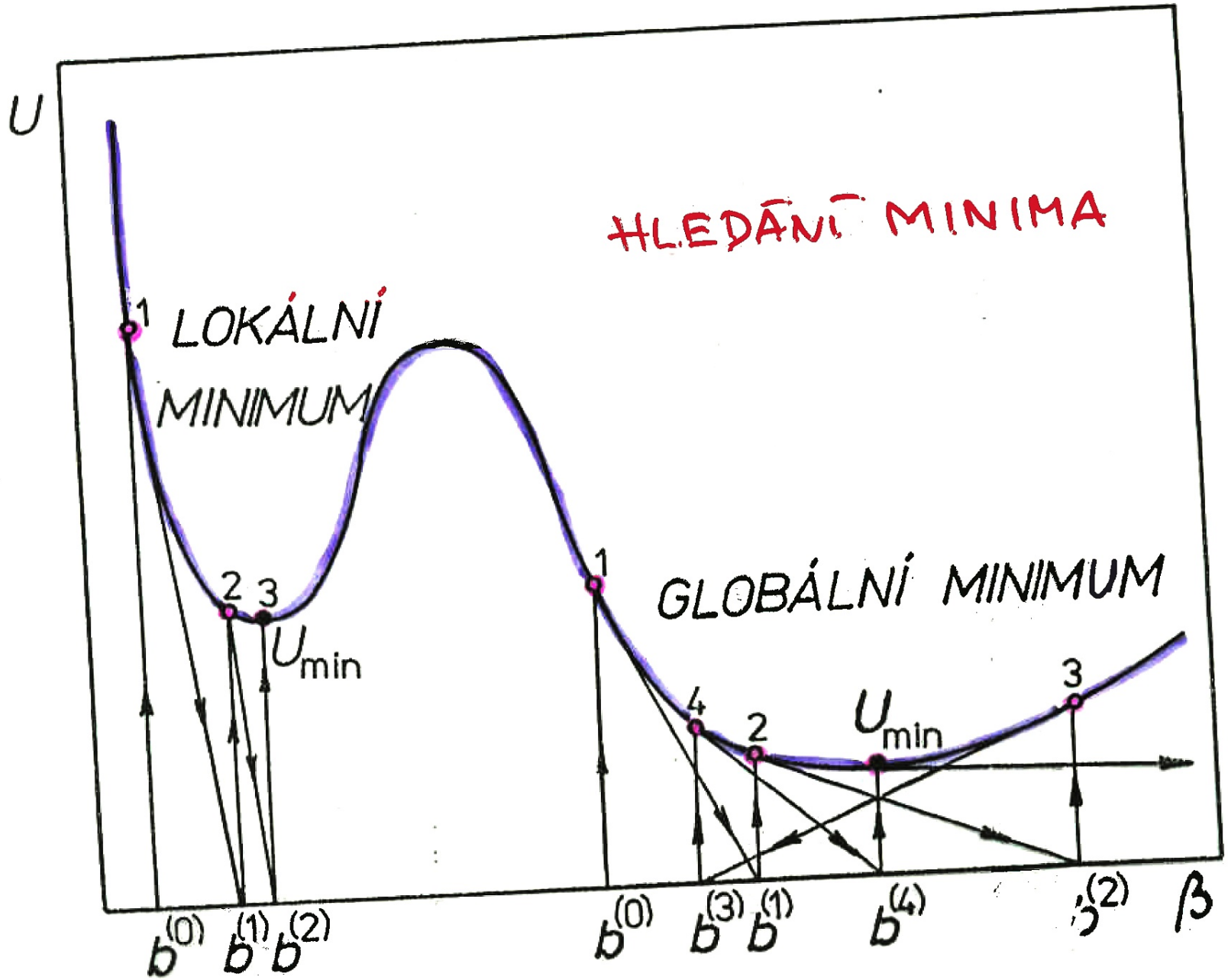
kde $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{f} = (f(x_1, \beta), \ldots, f(x_n, \beta))^T$,

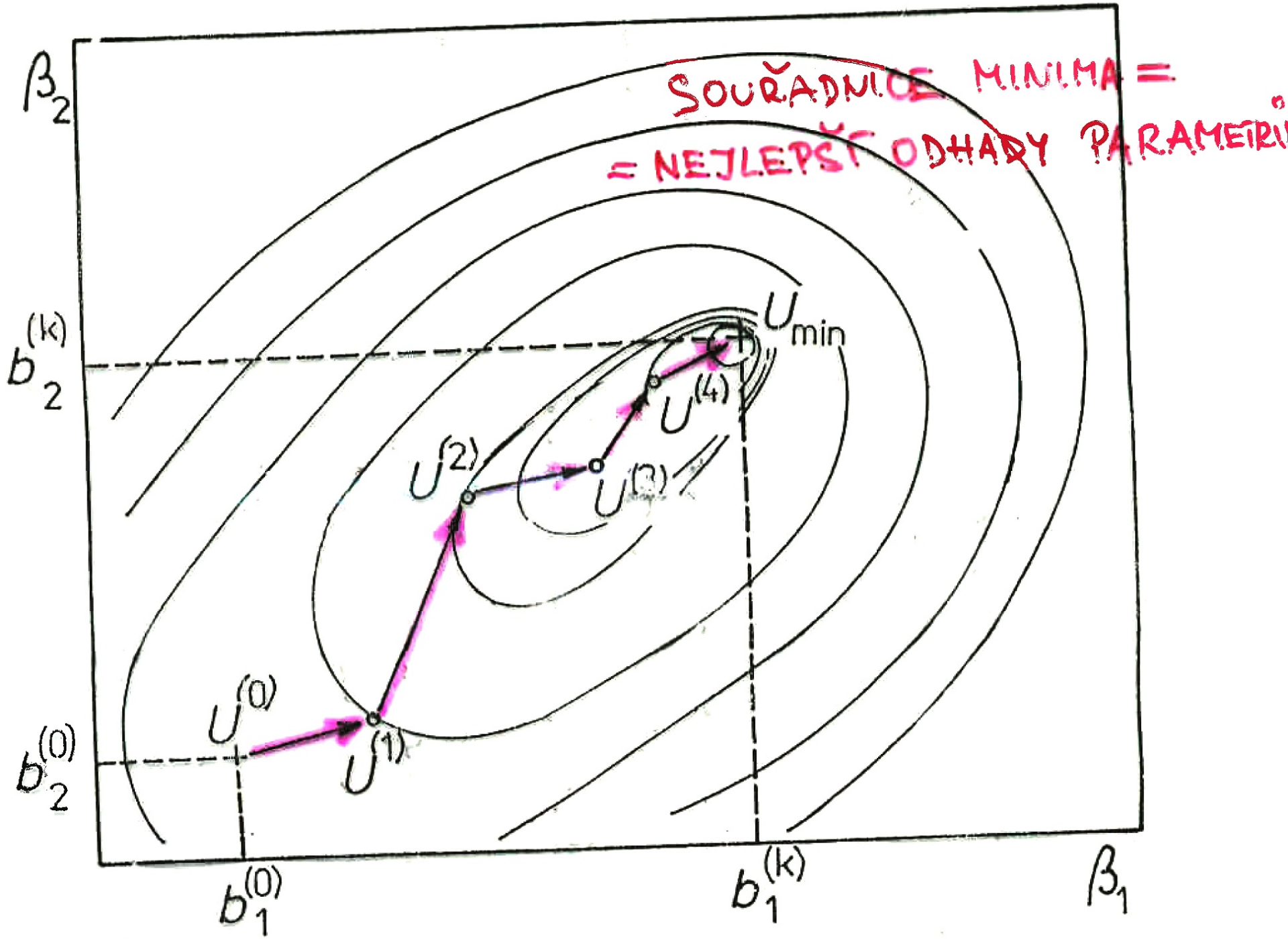
a symbol $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$ označuje euklidovskou normu.

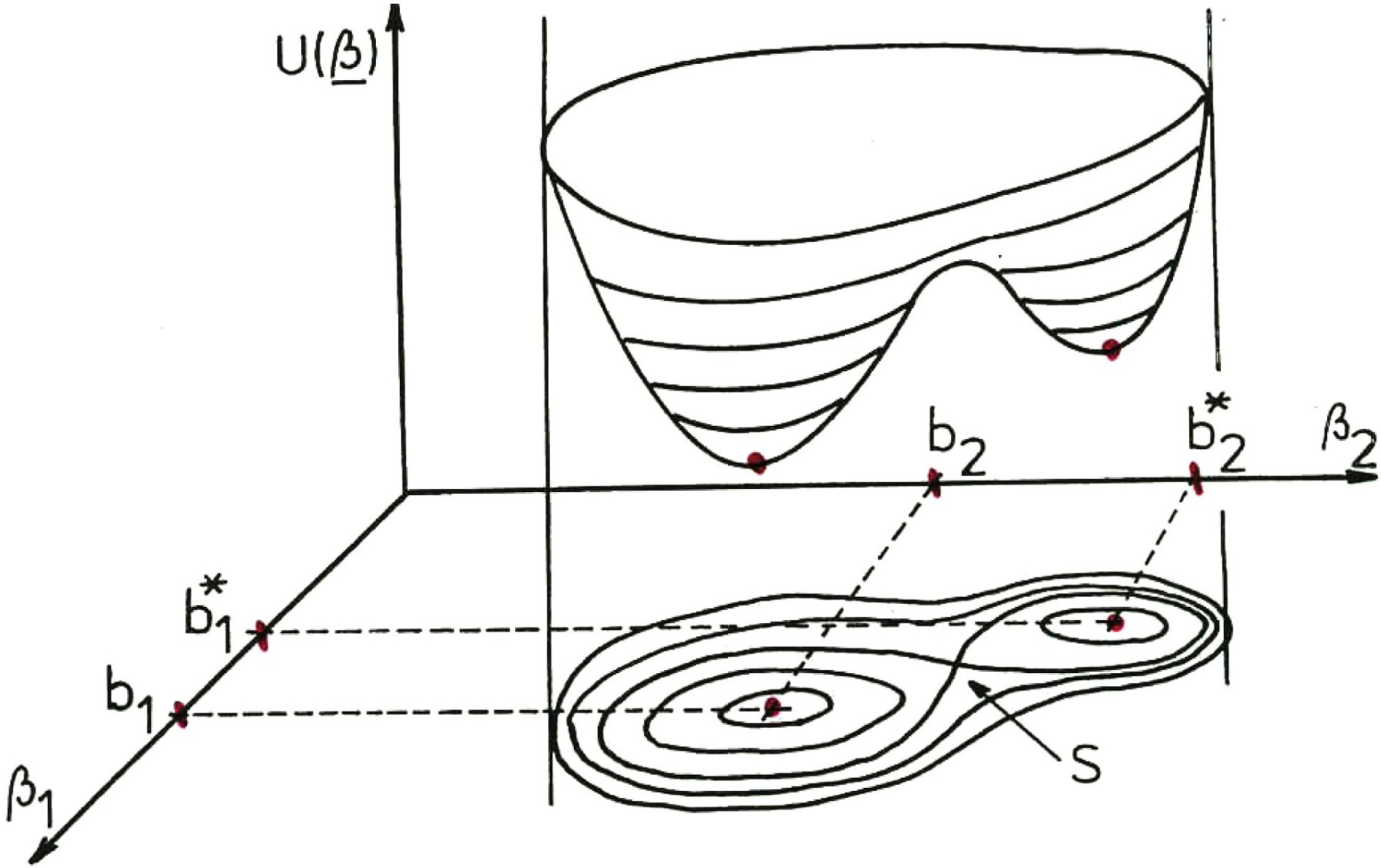


Minimum účelové funkce U(β) pro (a) lineární modely, (b) nelineární modely

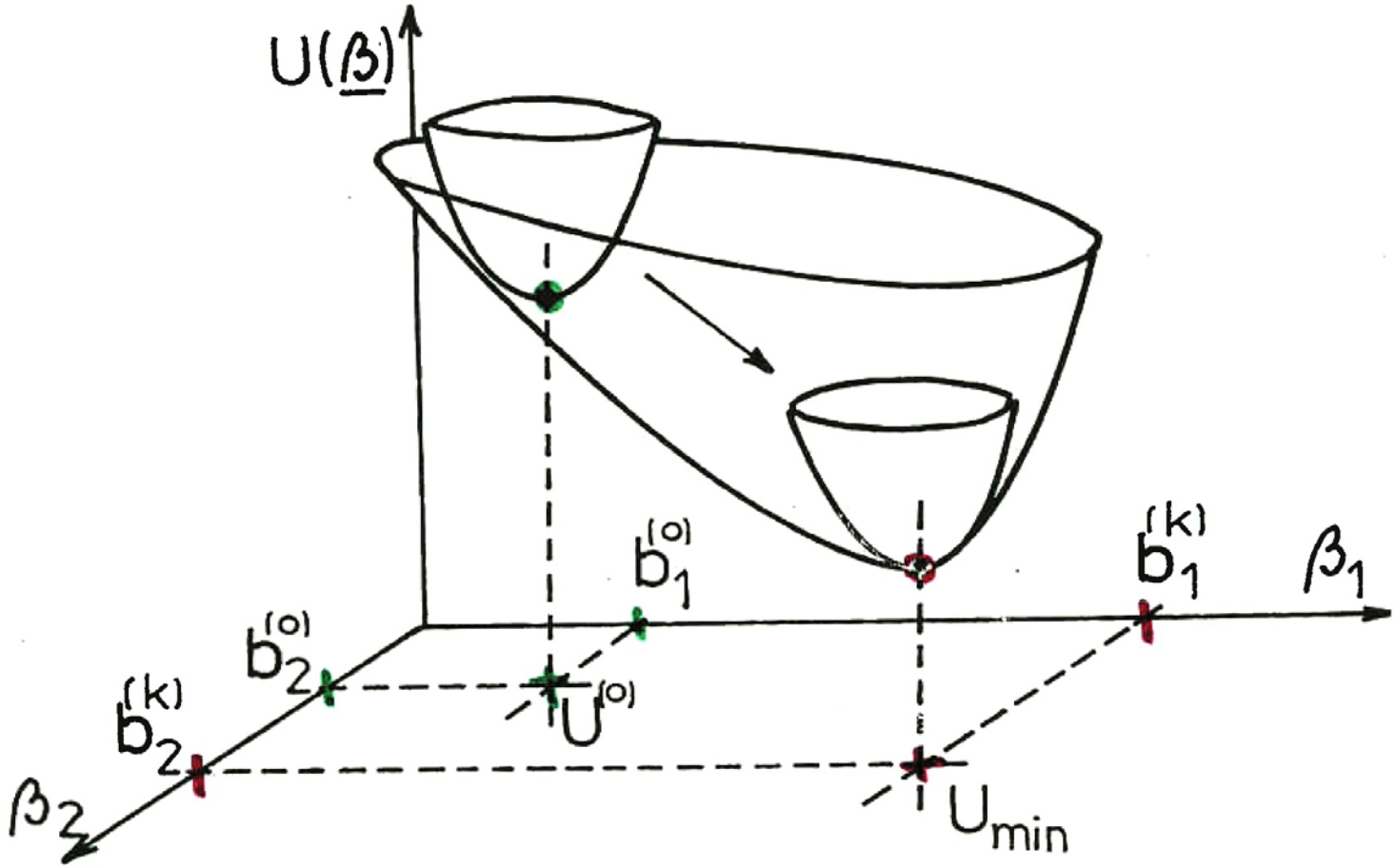
MINIMALIZAČNÍ PROCES:




SOUŘADNICE MINIMA =
= NEJLEPŠÍ ODHADY PARAMETRŮ
β_2
$b_2^{(k)}$
$b_2^{(0)}$
U_{min}
$U^{(0)}$
$U^{(1)}$
$U^{(2)}$
$U^{(3)}$
$U^{(4)}$
$b_1^{(0)}$
$b_1^{(k)}$
β_1



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



Geometrická interpretace hledání minima U(β)

2. O lokálním chování funkce U(β) v okolí libovolného bodu β_j lze získat kvantitativní informace z jejího Taylorova rozvoje

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

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.

Nederivační metody

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

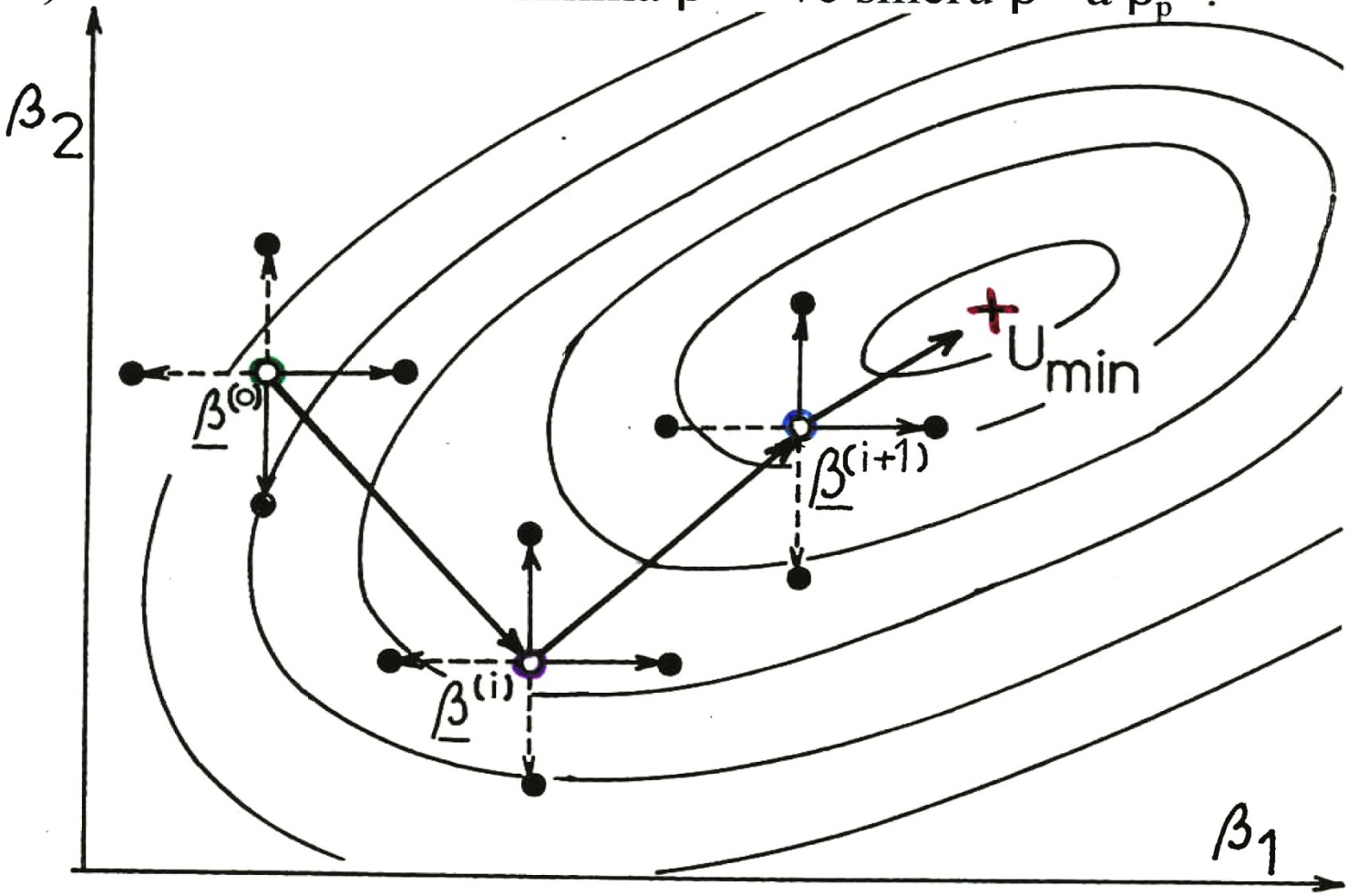
1. metody přímého hledání,
2. simplexové metody,
3. metody využívající náhodných čísel,
4. postupy speciálně vhodné pro MNČ.

1. Metody přímého hledání

Hookův-Jeevsův 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 $\beta_p^{(i)}$.



Postup koordinátního hledání (čárkovaně znázorněny neúspěšné směry)

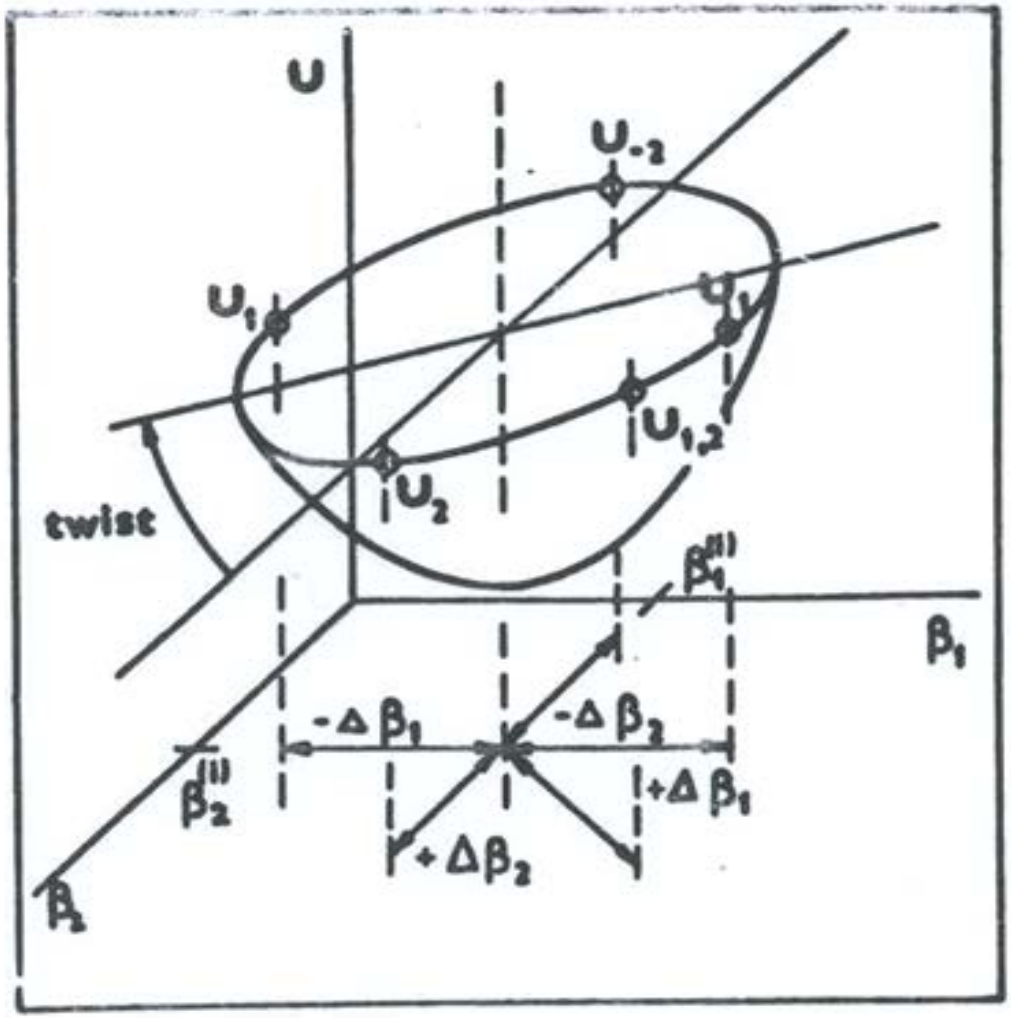
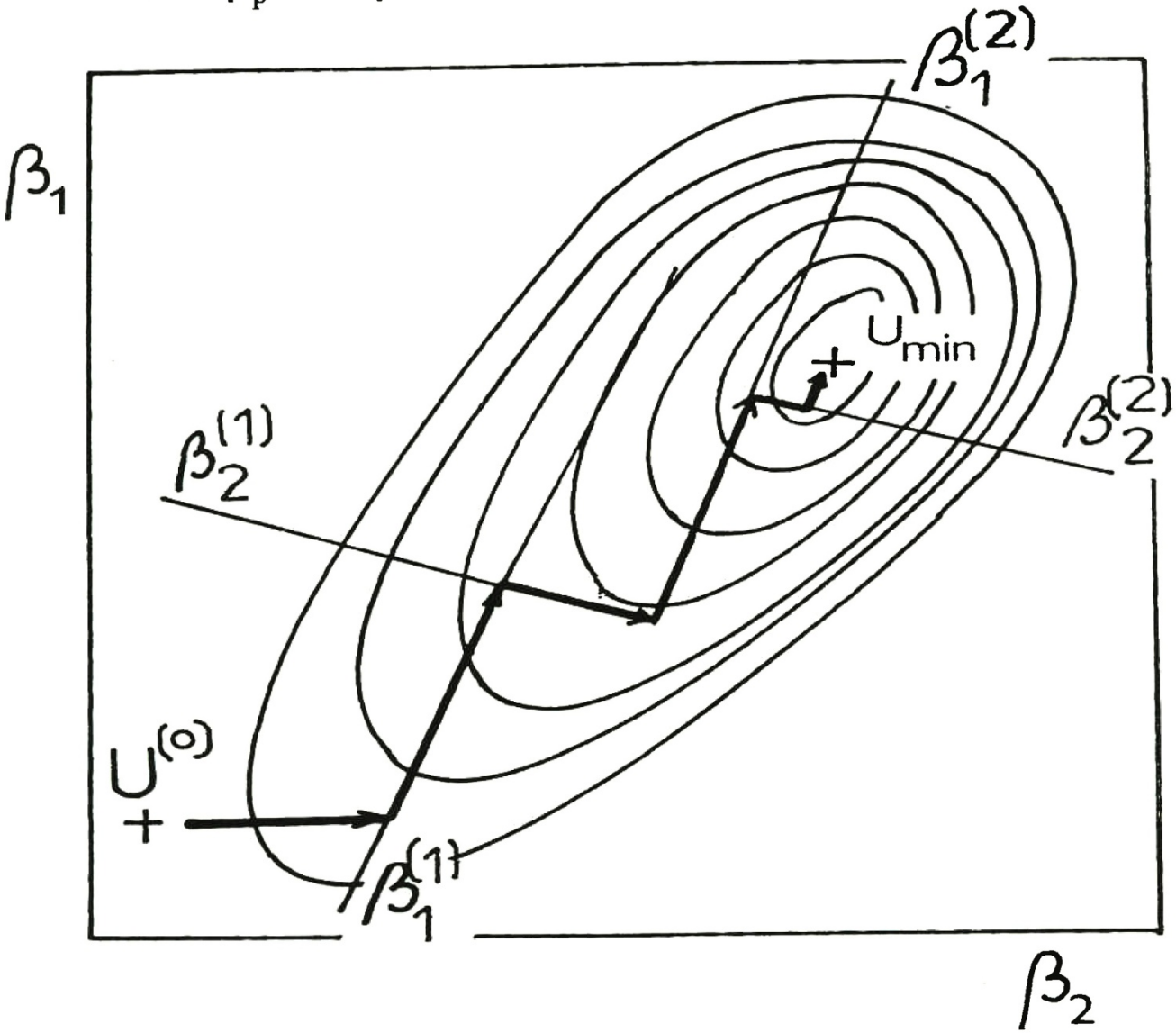


Fig. 5.5 — Graphical description of a systematic search for a pit, U_{min}, of an elliptic hyperparaboloid in $(m + 1)$-dimensional space (here $m = 2$, i.e., a three-dimensional space), by a series of 'shots'. In each shot the effect of a systematic variation is made in parameters β_1 and β_2, by steps $\pm\Delta\beta_1$ and $\pm\Delta\beta_2$ around the central point is calculated and on the basis of the result the next shot is made, etc. Iterations are repeated until a minimum (the pit) is reached.

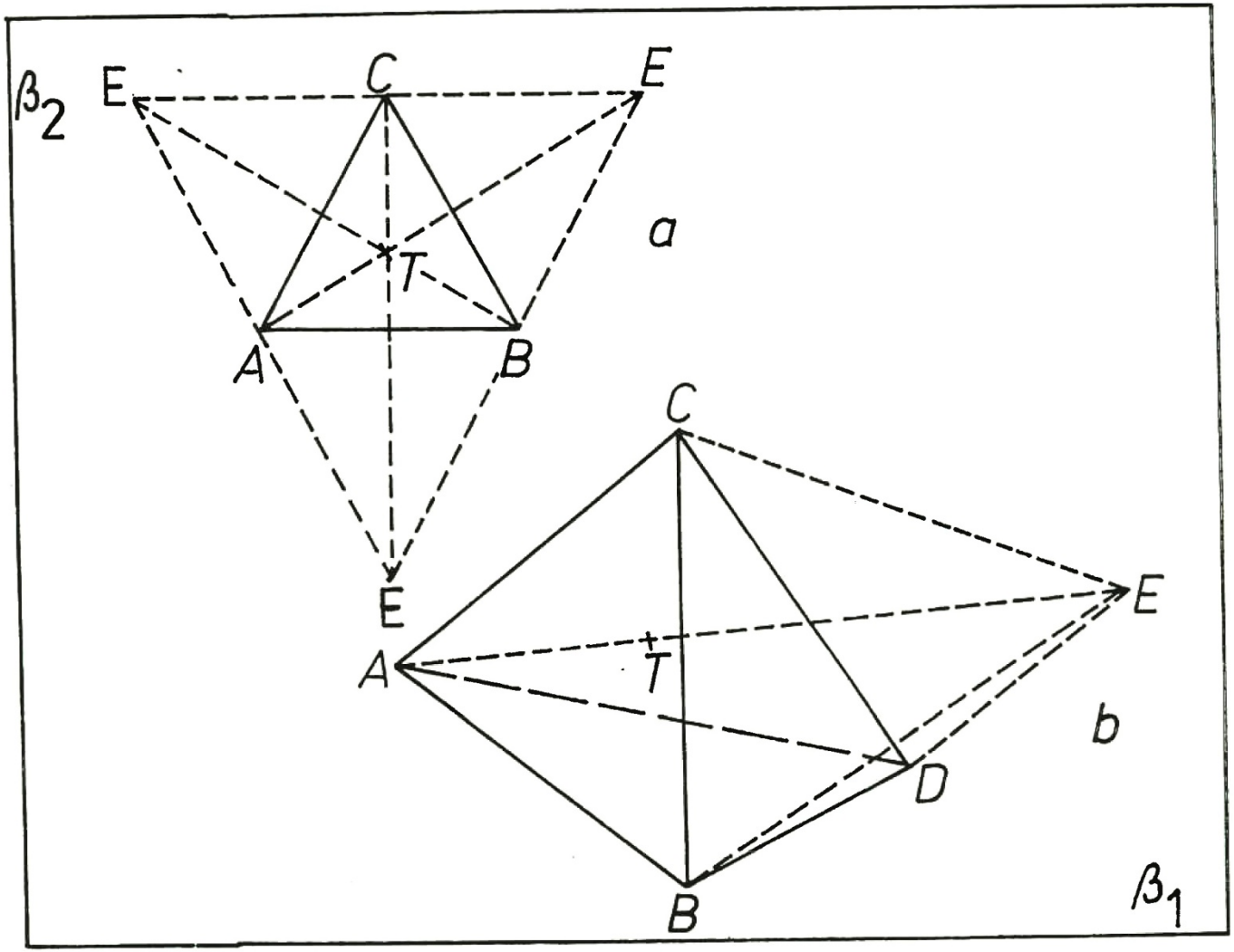
Rosenbrockova metoda: místo krokového hledání ve směru **s** rotaci souřadného systému tak, aby jedna osa splynula se směrem $\mathbf{S} = \beta_p^{(i+1)} - \beta^{(i)}$.



Postup Rosenbrockovy minimalizace

2. 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

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

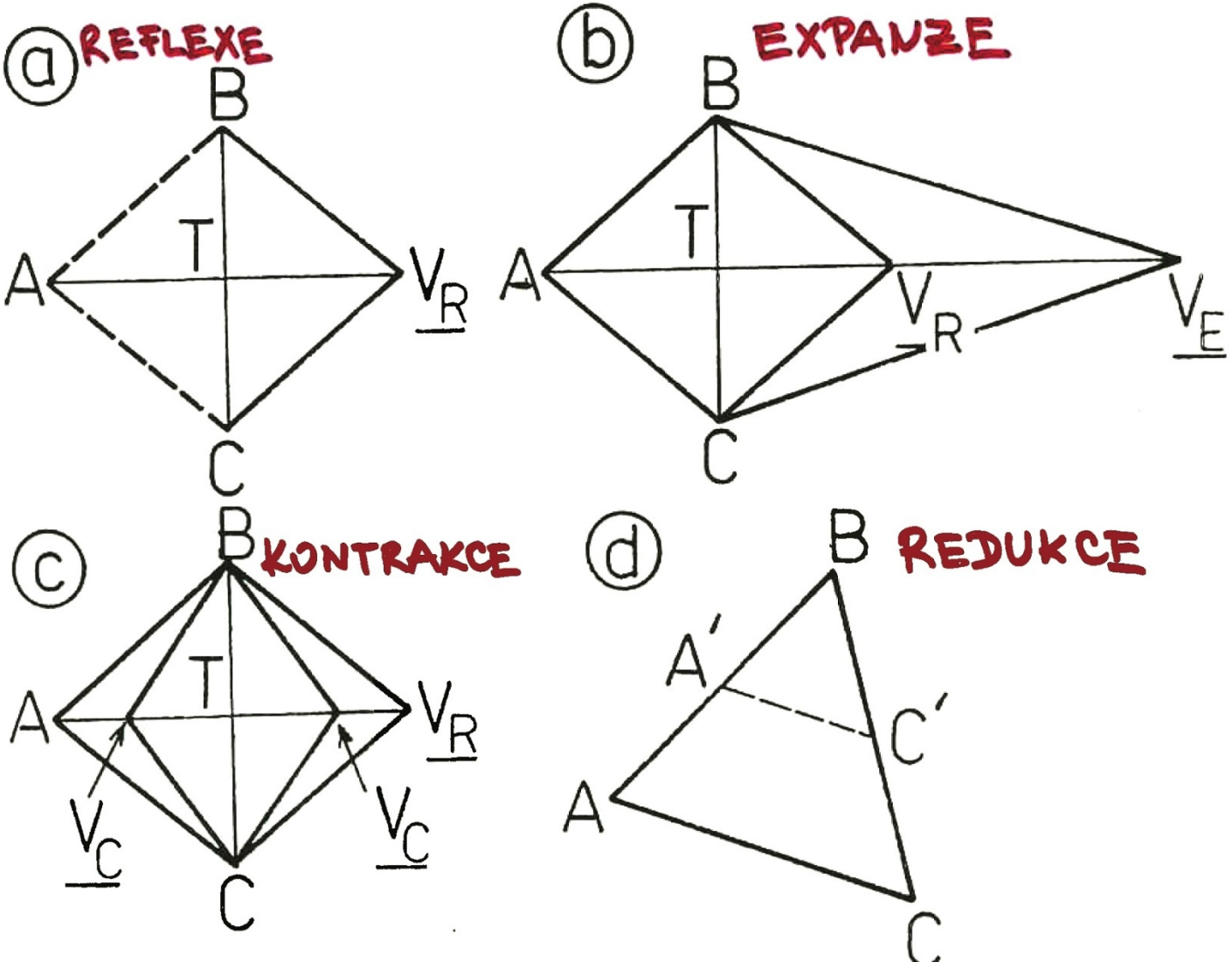
Simplex je konvexní mnohostěn v m-rozměrném prostoru parametrů právě svými (m+1) vrcholy.

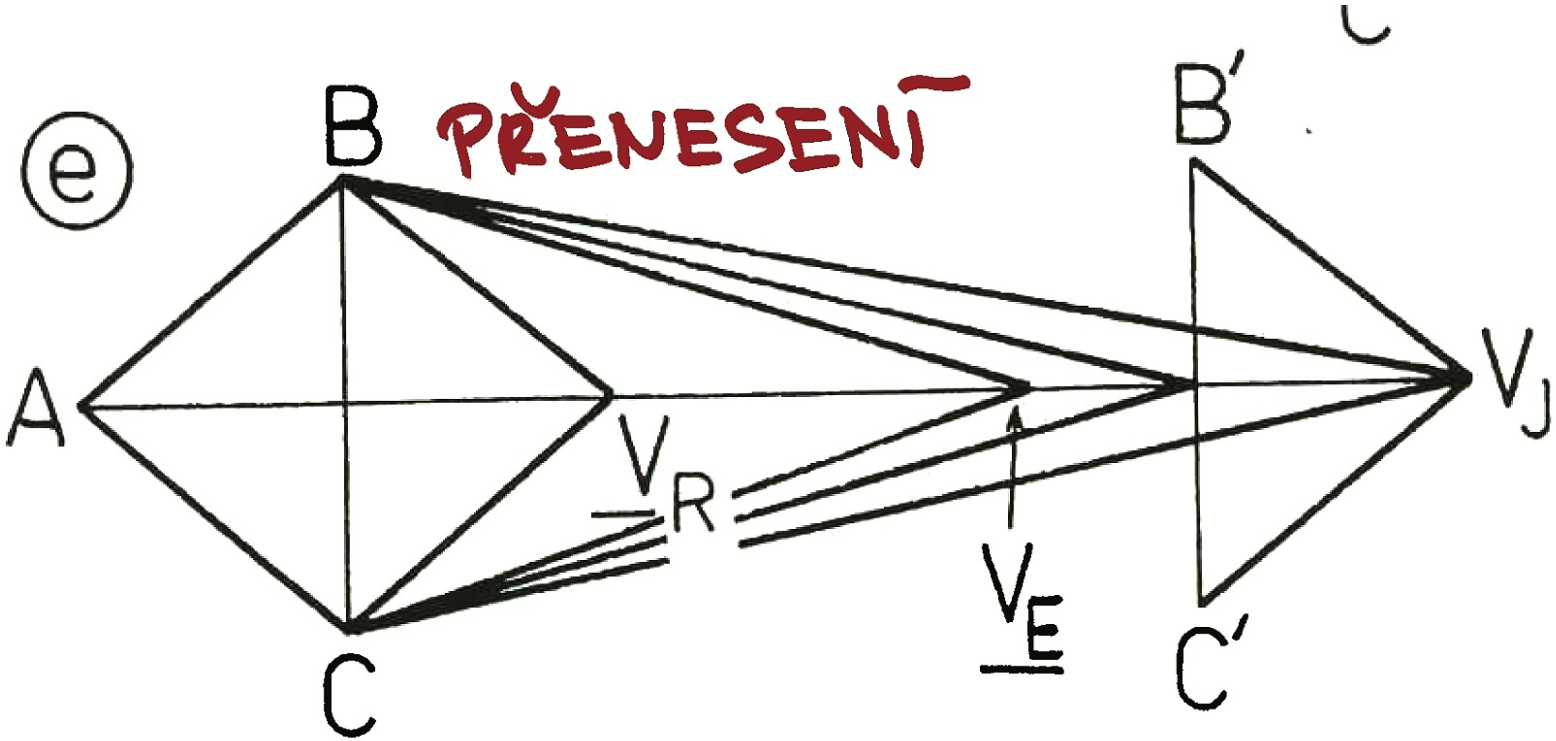
Postup minimalizace:

1. Konstrukce výchozího simplexu.
2. Iterativní postup k minimu.
3. Identifikace ukončení hledání terminace.

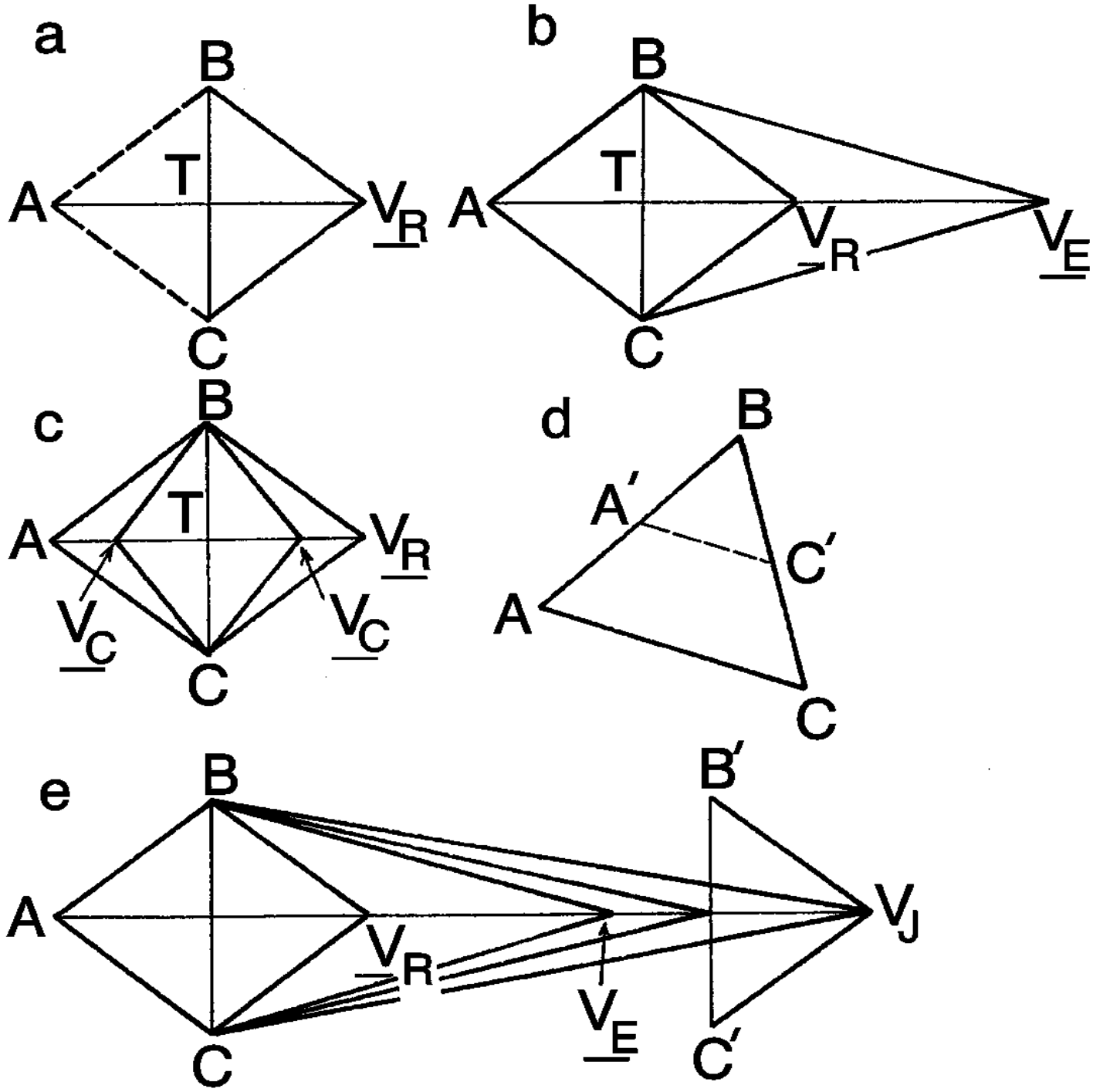
2. *Iterativní postup k minimu*

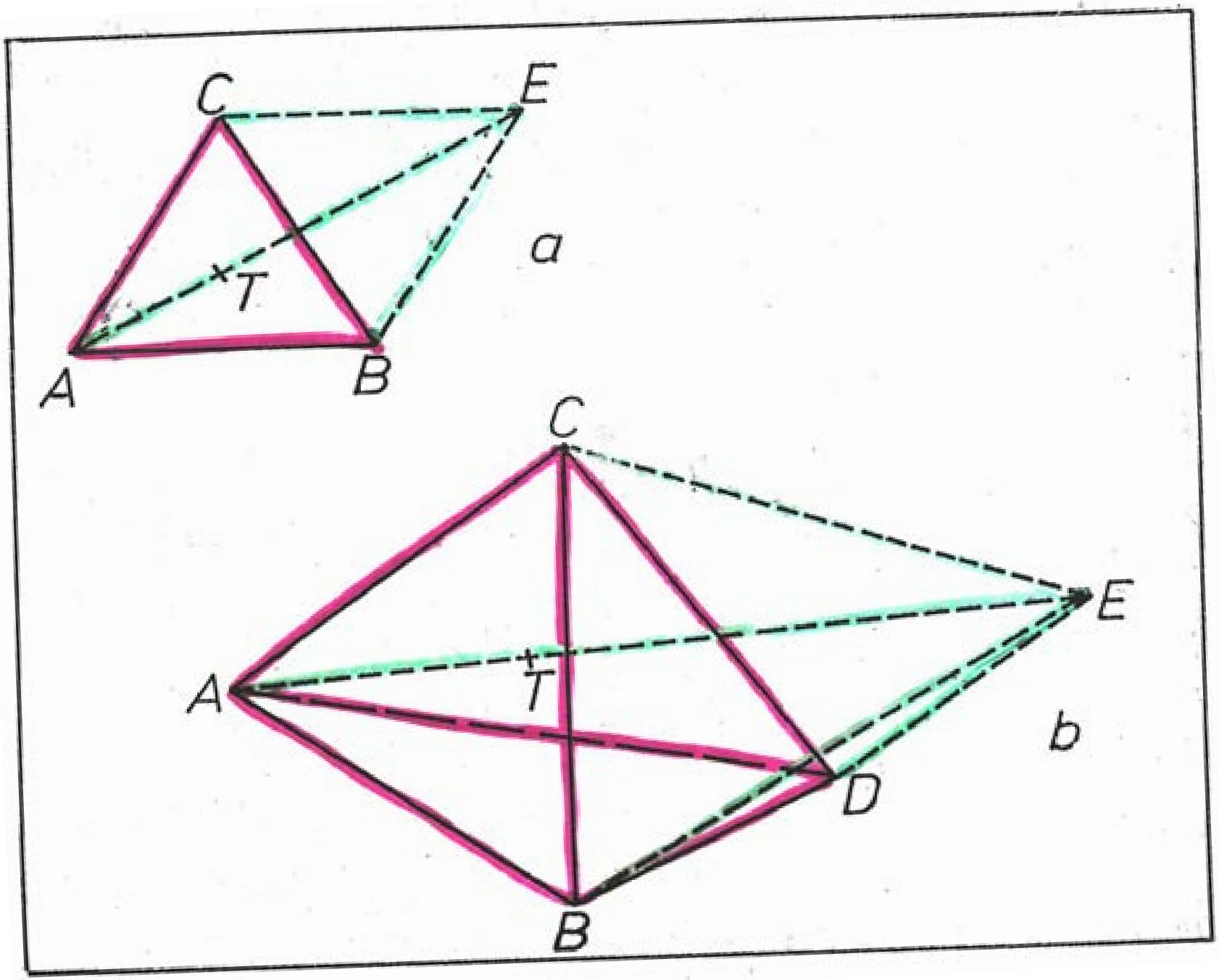
Na spojnici mezi $\mathbf{V}_H$ a jeho zrcadlovým obrazem pět operací: ***reflexe***, ***expanze***, ***kontrakce***, ***redukce*** a ***přenesení***.

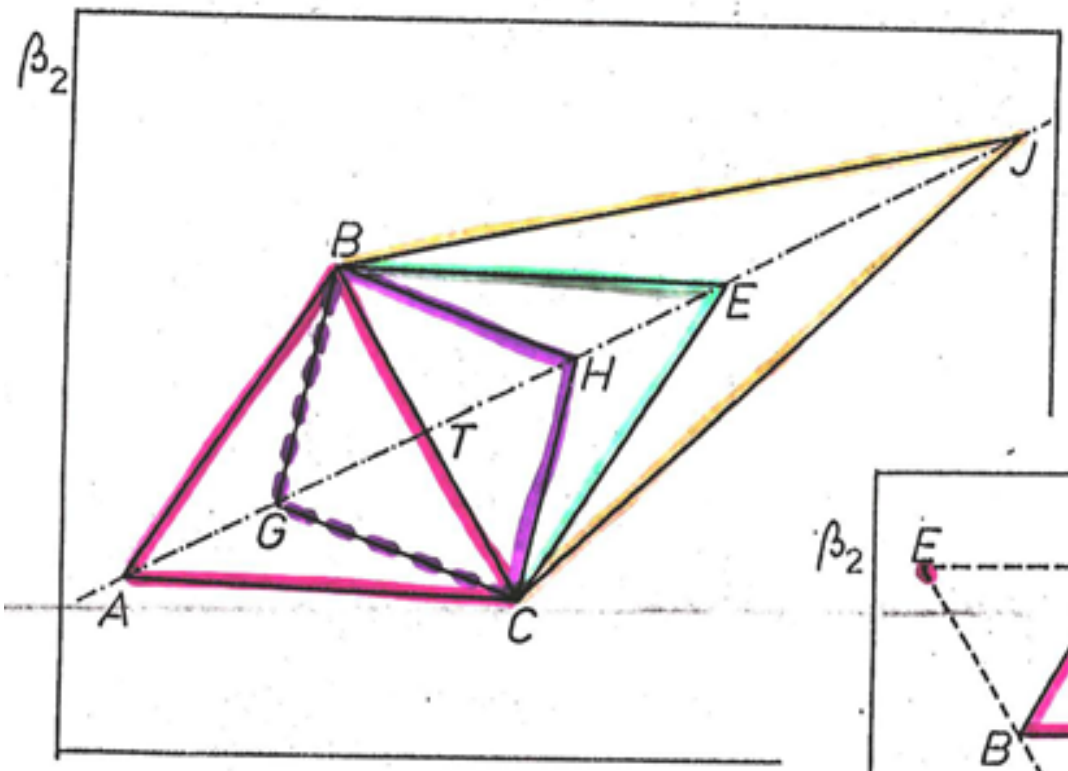


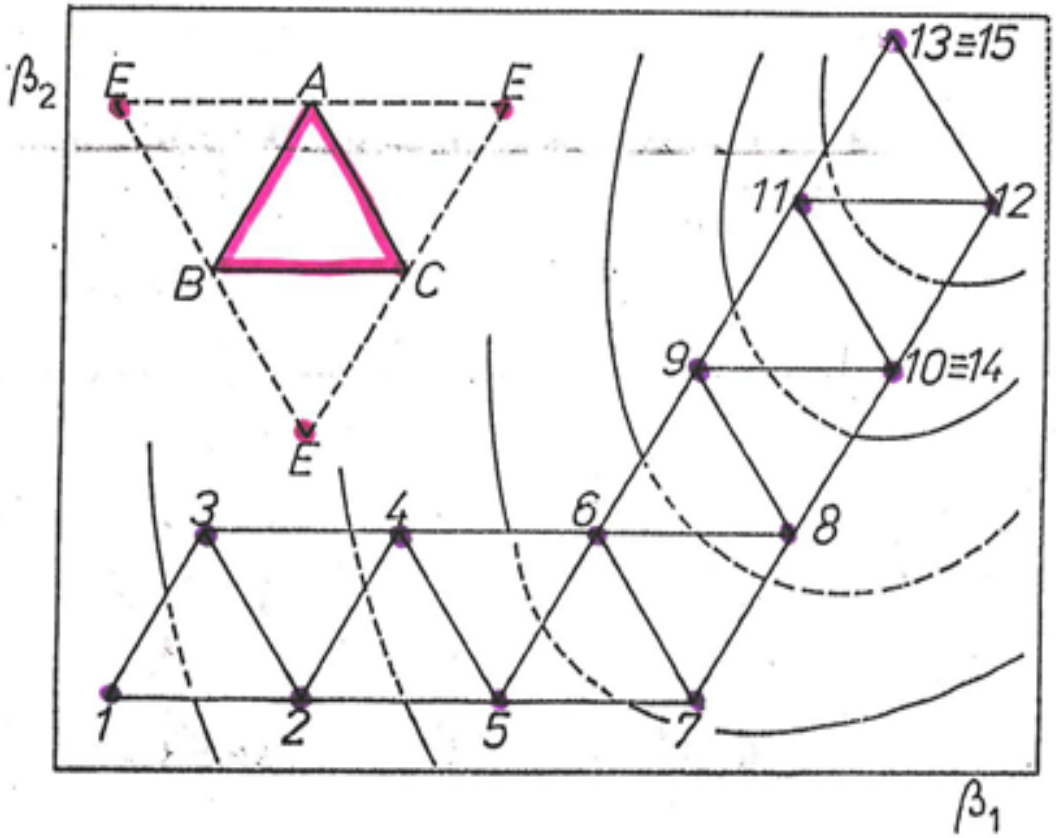


Základní operace simplexu:
a) reflexe, b) expanze, c) kontrakce, d) redukce a e) přenesení


a
B
T
A
V_R
C
b
B
T
A
V_R
V_E
C
c
B
T
A
V_R
V_C
V_C
C
d
B
A′
C′
A
C
e
B
B′
A
V_J
V_R
V_E
C
C′


C
E
T
A
B
a
C
E
A
T
b
D
B

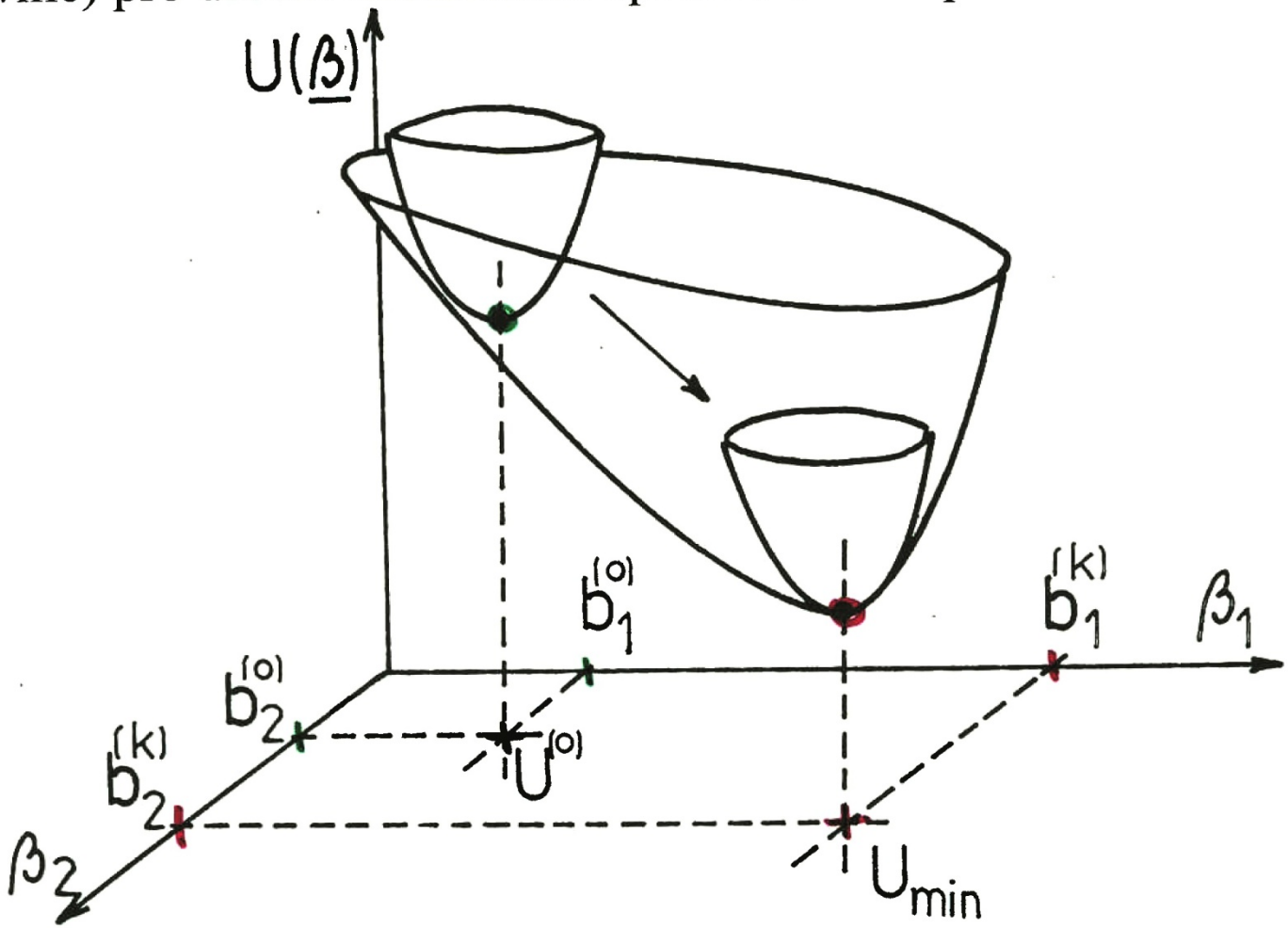

β_2
J
B
E
H
T
G
A
C


β_2
E
A
E
B
C
E
13≡15
11
12
9
10≡14
3
4
6
8
1
2
5
7
β_1

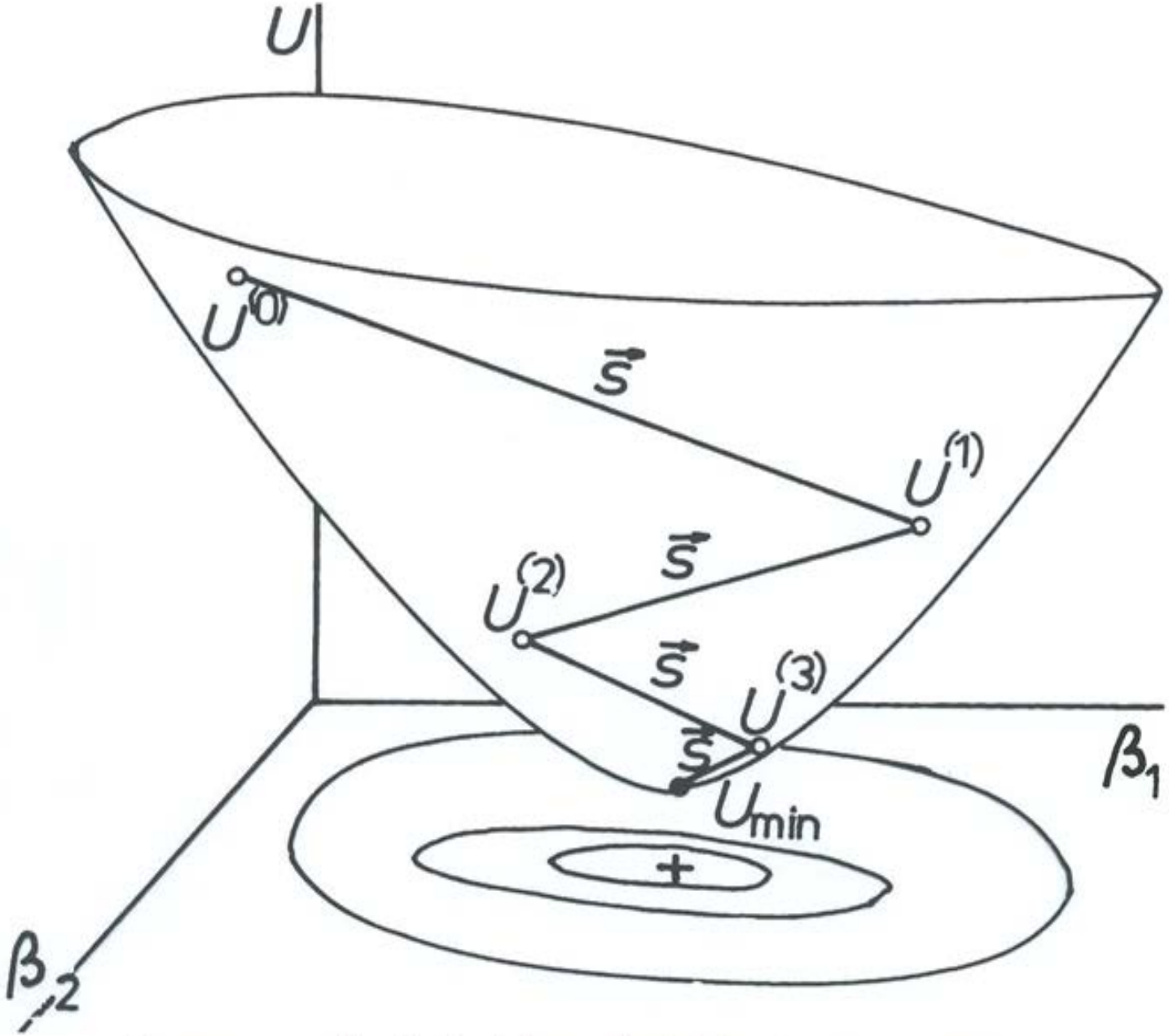
4. Postupy speciálně pro MNC

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(β) (LETAGROP)



Interpretrace hledání minima eliptického hyperparaboloidu

Derivační metody

V i-té iteraci se k odhadu $\beta^{(i)}$ přičítá přírůstkový vektor Δ_i

$$\beta^{(i+1)} = \beta^{(i)} + \Delta_i$$

Postup hledání minima $U(\beta)$ lze rozdělit do čtyř fází:

I. Stanovení prvních odhadů $\beta^{(0)}$.

II. Nalezení vhodného směrového vektoru $\mathbf{V}_i$.

III. Určení α_i, aby přírůstkový vektor $\Delta_i = \alpha_i \mathbf{V}_i$ byl přijatelný.

IV. Ověření, zda bylo dosaženo minima.

Vektor Δ_i se považuje za přijatelný, když

$$U(\beta^{(i)} + \Delta_i) < U(\beta^{(i)})$$

I. Stanovení prvních odhadů

- Z dobrých odhadů $\beta^{(0)}$ konvergují i nenáročné metody.
- Je možno porovnávat průběh $f(x, \mathbf{b}^{(0)})$ se zadanými daty a ověřovat tak kvalitu prvních odhadů.

II. Určení směrového vektoru

Pro derivaci $U(\beta)$ v místě $\Delta = \beta + \alpha\, \mathbf{V}$ podle skaláru α bude

$$\frac{\delta\, U(\mathbf{b})}{\delta\, \alpha} = \left[\frac{\delta\, U(\beta)}{\delta\, \beta}\right]^T \frac{\delta\, \beta}{\delta\, \alpha} = \mathbf{g}^T(\beta)\ \mathbf{V}$$

kde gradient $\mathbf{g}$ je uvažován v místě Δ.

Pro $\alpha \to 0$ získáme tzv. ***směrovou derivaci***

$$S_D = \left.\frac{\delta\, U(\beta)}{\delta\, \alpha}\right|_{\alpha \to 0} = \mathbf{g}^T\ \mathbf{V}$$

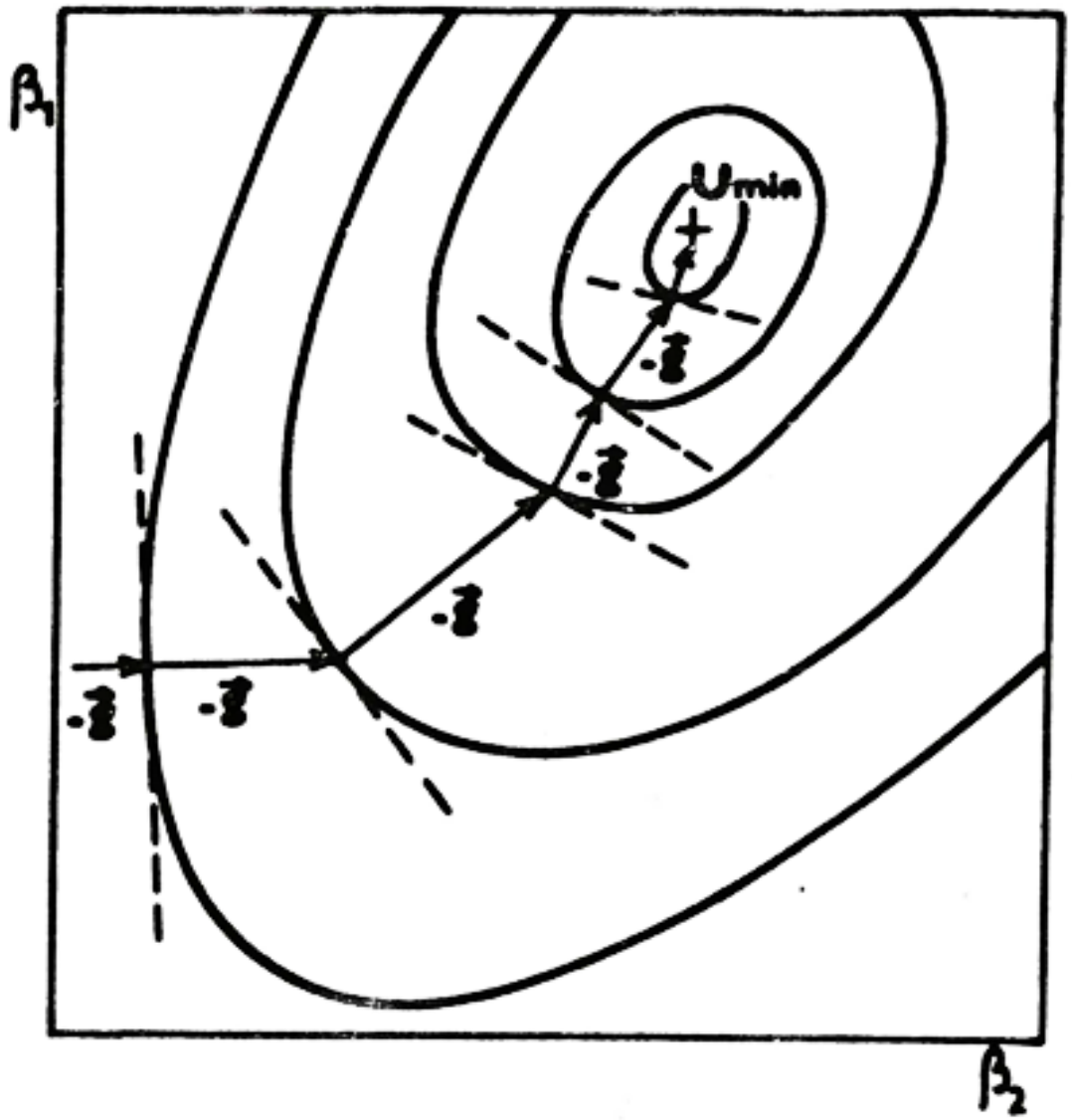


Fig. 5.6 — Seeking a minimum in the direction of the negative gradient vector by the method of steepest descent.

kde $\Delta\beta_j = \beta - \beta_j$ a $\mathbf{g}_j$ je gradient

$$g_k = \frac{\delta\, U(\beta)}{\delta\, \beta_k} \qquad k = 1, \ldots, m$$

Matice $\mathbf{H}_j$ rozměru $(m \times m)$ je symetrická ***<u>Hessova matice</u>*** druhých derivací účelové funkce podle parametrů se složkami

$$H_{lk} = \frac{\delta^2\, U(\beta)}{\delta\, \beta_l\ \delta\, \beta_k} \qquad l, k = 1, \ldots, m$$

a gradient má tvar

$$\mathbf{g}_j = -2\ \mathbf{J}^T\ \mathbf{e}$$

kde $\mathbf{e}$ je vektor reziduí s prvky $e_i = y_i - f(x_i, \beta)$, $i = 1, \ldots, n$.

Matice **J** rozměru (n × m) se nazývá ***Jakobiho matice*** s prvky, odpovídajícími derivacím regresního modelu dle jednotlivých parametrů v jednotlivých bodech

$$J_{ik} = \frac{\delta f(x_i, \beta)}{\delta \beta_k} \qquad i = 1, \ldots, n$$

$$k = 1, \ldots, m$$

a pro Hessovu matici

$$\mathbf{H}_j = 2\,[\mathbf{J}^T \mathbf{J} + \mathbf{B}]$$

kde **B** je matice druhých derivací modelové funkce s prvky

$$B_{jk} = \sum_{i=1}^{n} e_i \frac{\delta^2 f(x_i, \beta)}{\delta \beta_j \, \delta \beta_k} \qquad j, k = 1, \ldots, m$$

V okolí lokálních minim **b** je gradient **g** roven nule:

a) vektor chyb $\hat{\mathbf{e}}$ je zde kolmý na sloupce matice **J**, které tvoří souřadnice m-rozměrného prostoru,

b) funkce $U(\beta)$ je zde úměrná $\Delta \beta_j^T \mathbf{H}_i \Delta \beta_j$.

Dosazením dostaneme kritérium ***"linearizované" metody nejmenších čtverců***

$$U_L(\beta) = \mathbf{e}^T \mathbf{e} - 2\,\Delta\beta\ \mathbf{J}^T \mathbf{e} + \Delta\beta^T (\mathbf{J}^T \mathbf{J})\,\Delta\beta$$

První člen: je roven $U(\beta_j)$,
Druhý člen: je roven $\Delta\beta_j^T\,\mathbf{g}$.
Místo matice **H** je matice $2\,\mathbf{J}^T\,\mathbf{J}$.
Pro malé hodnoty chyb e_i lze matici **B** zanedbat a rovnice jsou totožné.

U nelineárních regresních modelů odpovídá matice $\mathbf{J}^T\,\mathbf{J}$ matici $\mathbf{X}^T\,\mathbf{X}$ u lineárních modelů.
(Bude-li linearizace dostatečně přesná, lze provádět statistickou analýzu stejně jako u lineárních regresních modelů).

Komplikace procesu nelineární regrese

1. Neodhadnutelnost některých parametrů.
2. Existence minima funkce $U(\beta)$ jen pro některé regresní modely.
3. Výskyt lokálních minim a sedlových bodů na kriteriální funkci $U(\beta)$.
4. Špatná podmíněnost parametrů v regresním modelu.
5. Malé rozmezí experimentálních dat.

1. Neodhadnutelnost některých parametrů

Normalizované citlivostní koeficienty

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

$$i = 1, \ldots, n$$

Vizuální posouzení špatné podmíněnosti:
k vyšetření multikolinearity mezi parametry β_j, β_h, se konstruují ***citlivostní grafy*** $C_{j(i)}$ a $C_{h(i)}$ na x_i, $i = 1, \ldots, n$.

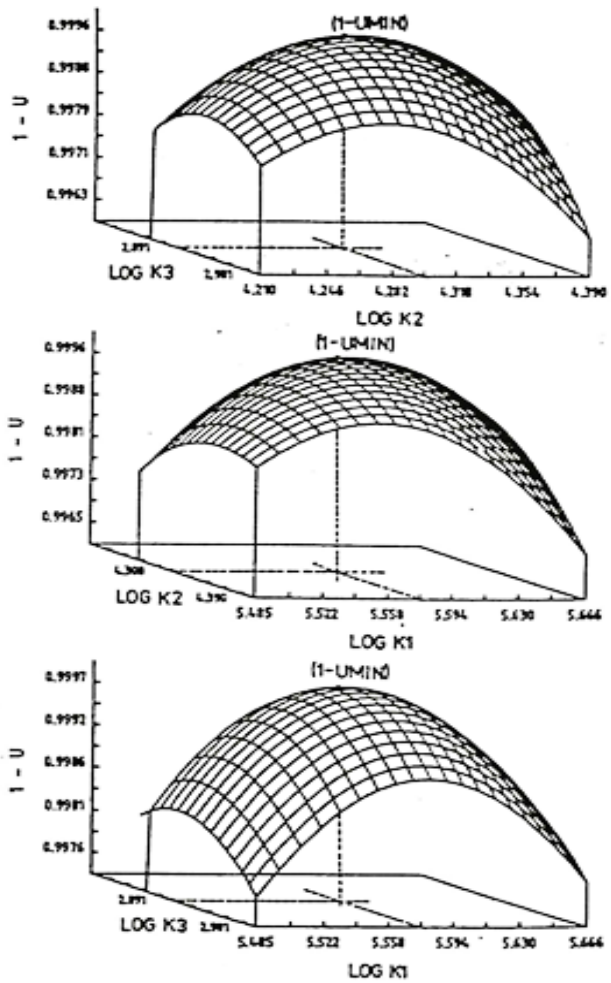
Citlivostní funkce: pro vyjádření citlivosti regresních modelů na změnu parametru β_j

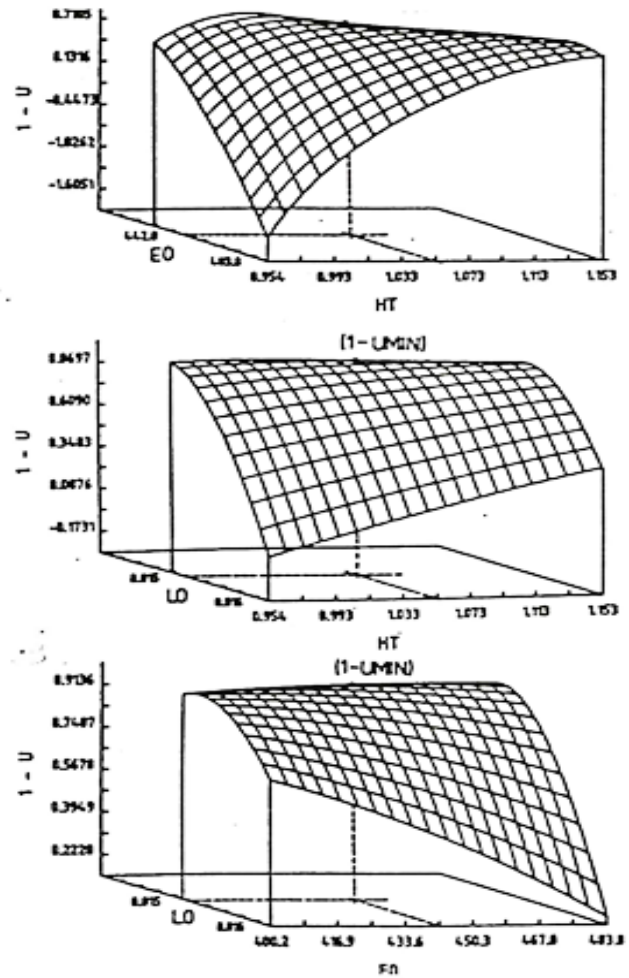
$$C_{cj} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta f(x_i, \beta)}{\delta \beta_j} \right]^2$$

pro nelineární parametry β_j: C_{cj} není konstantní.
pro lineární parametry β_j: C_{cj} je konstantní.

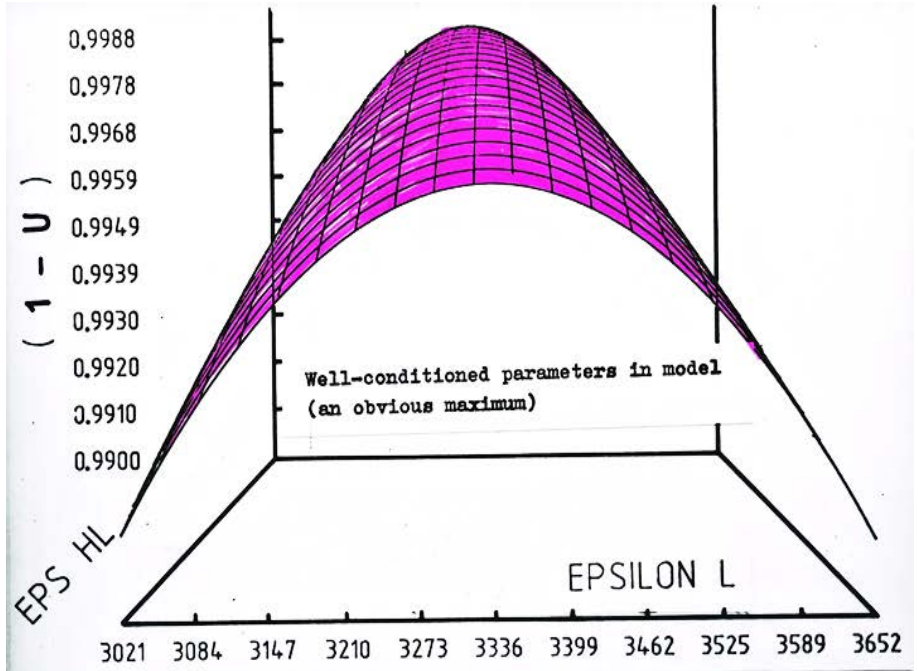
Citlivost parametrů ovlivňuje kvalitu odhadu, a proto jsou **parametry v modelu**

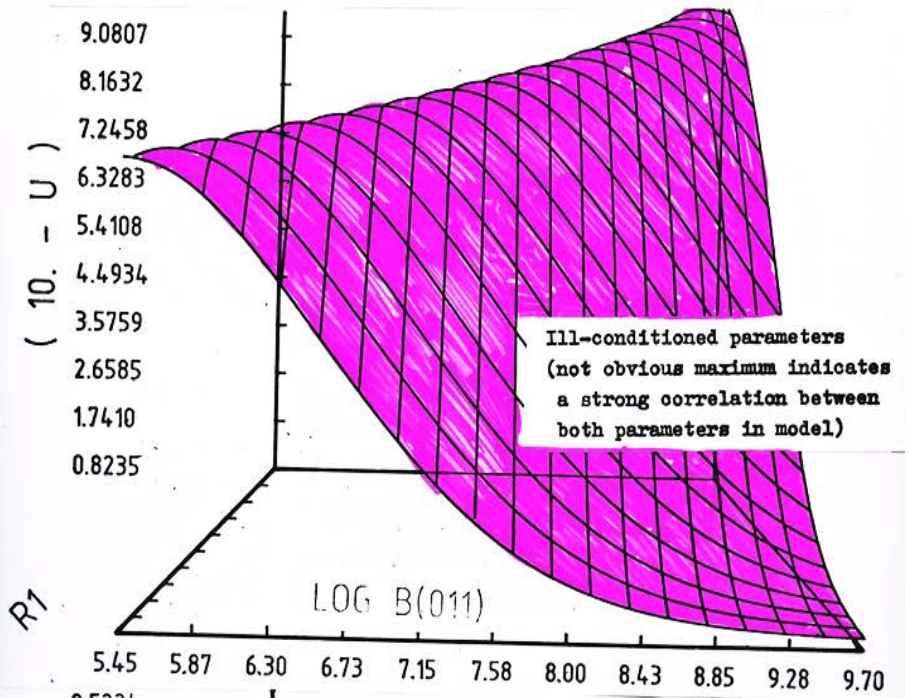
dobře podmíněny **špatně podmíněny**

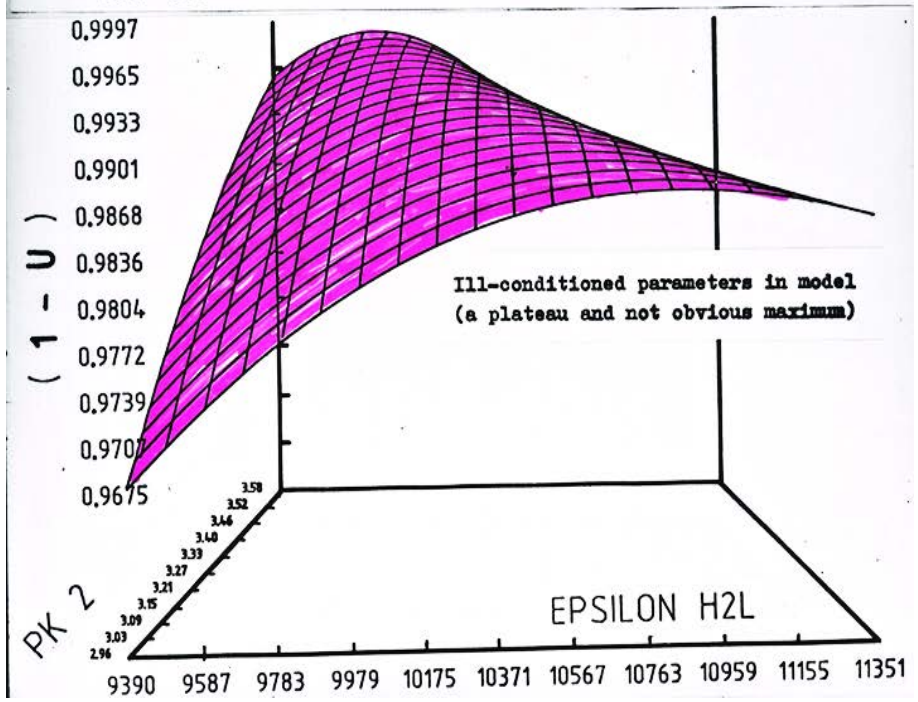


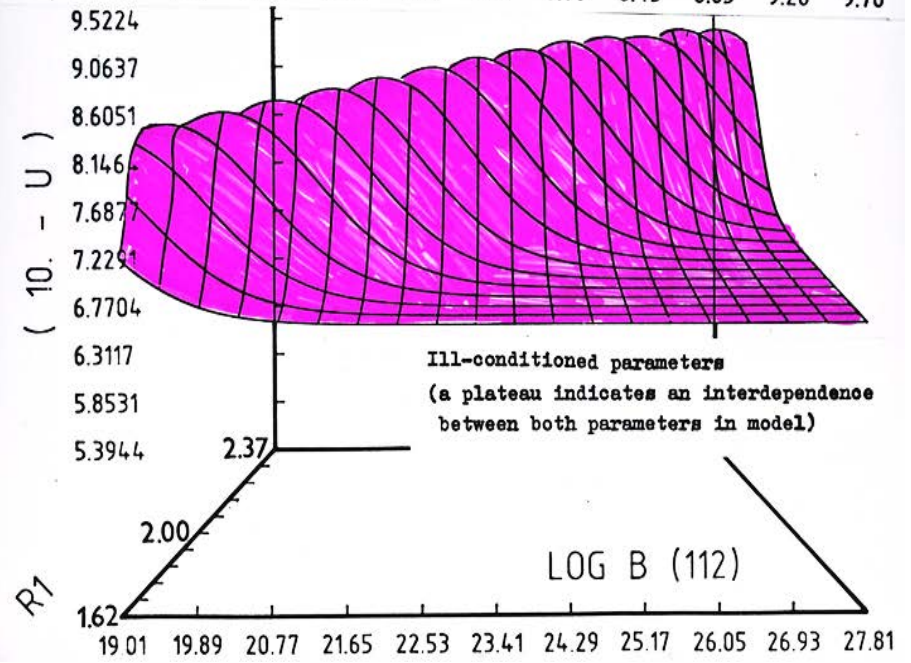


Residual-square-sum function response surface $U = f(r_1, \log \beta_{011})$









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 **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,

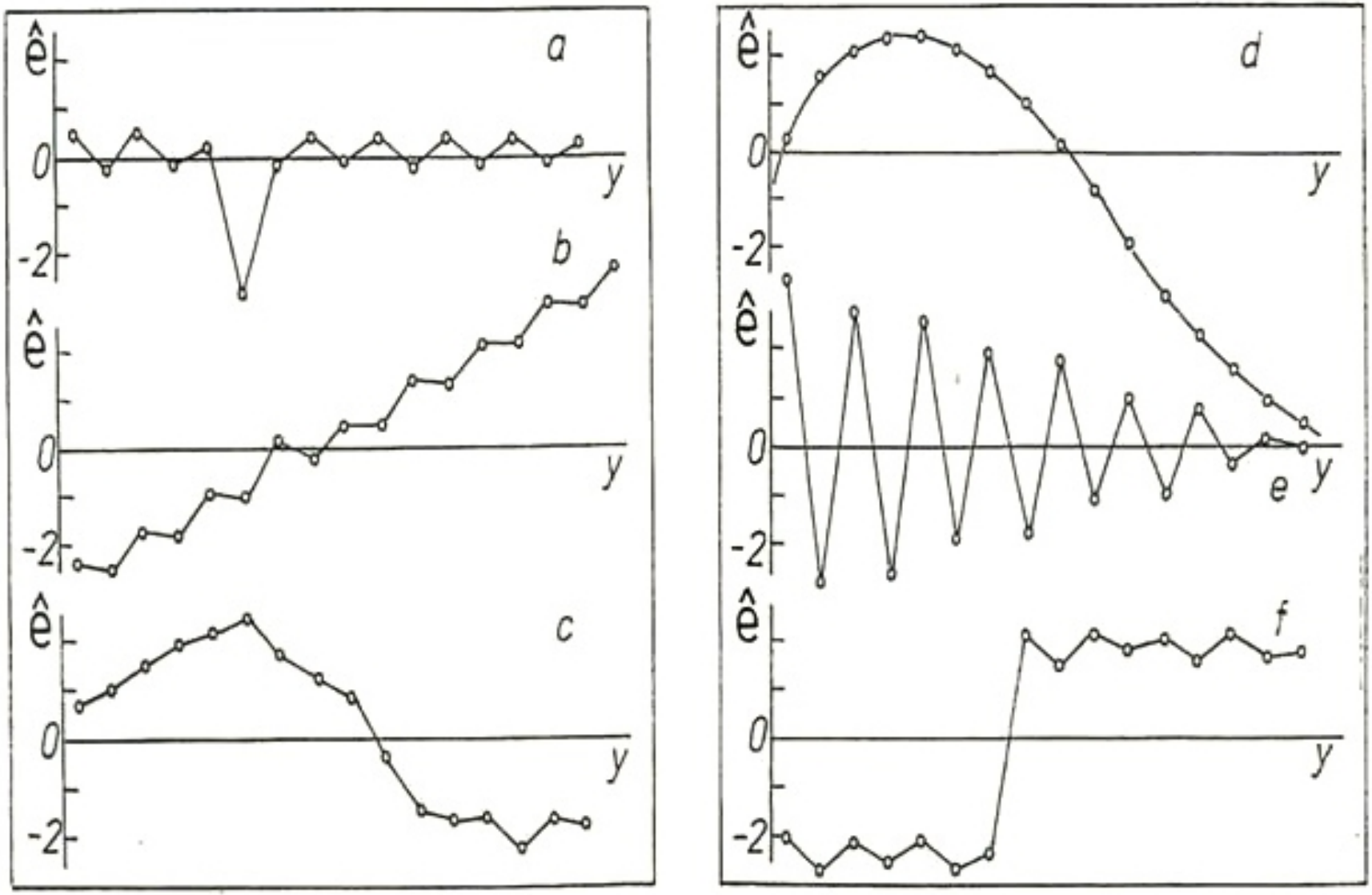
b) trend v reziduích,

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

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

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 $\hat{s}^2_{(i)}$ 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***

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

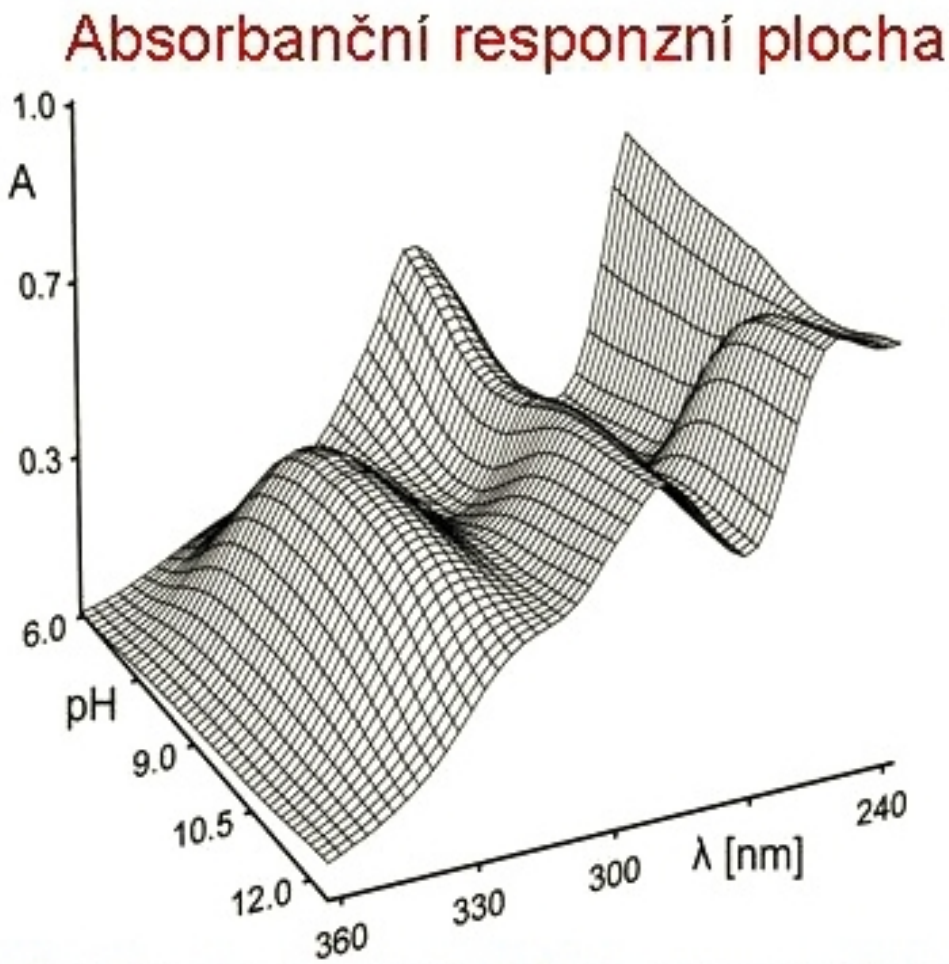
Pro případ MNČ bude věrohodnostní vzdálenost ve tvaru

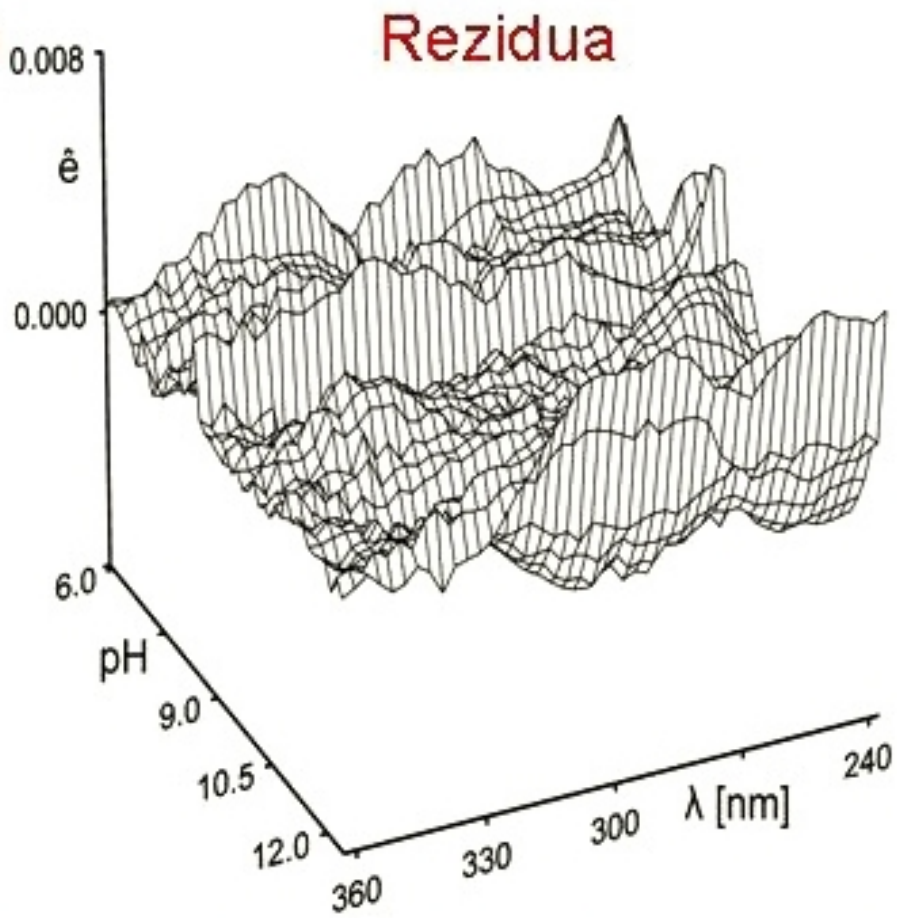
$$LD_i = n \ln\left[\frac{U(\mathbf{b}_{(i)})}{U(\mathbf{b})}\right]$$

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

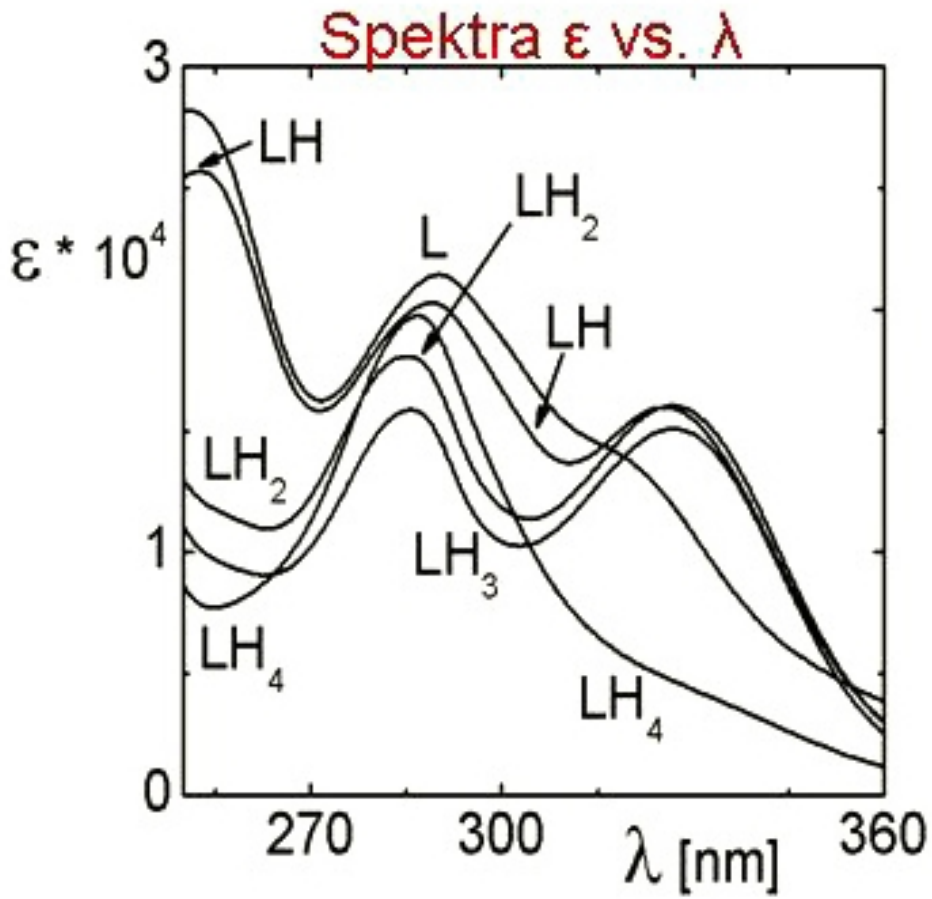
Input dat

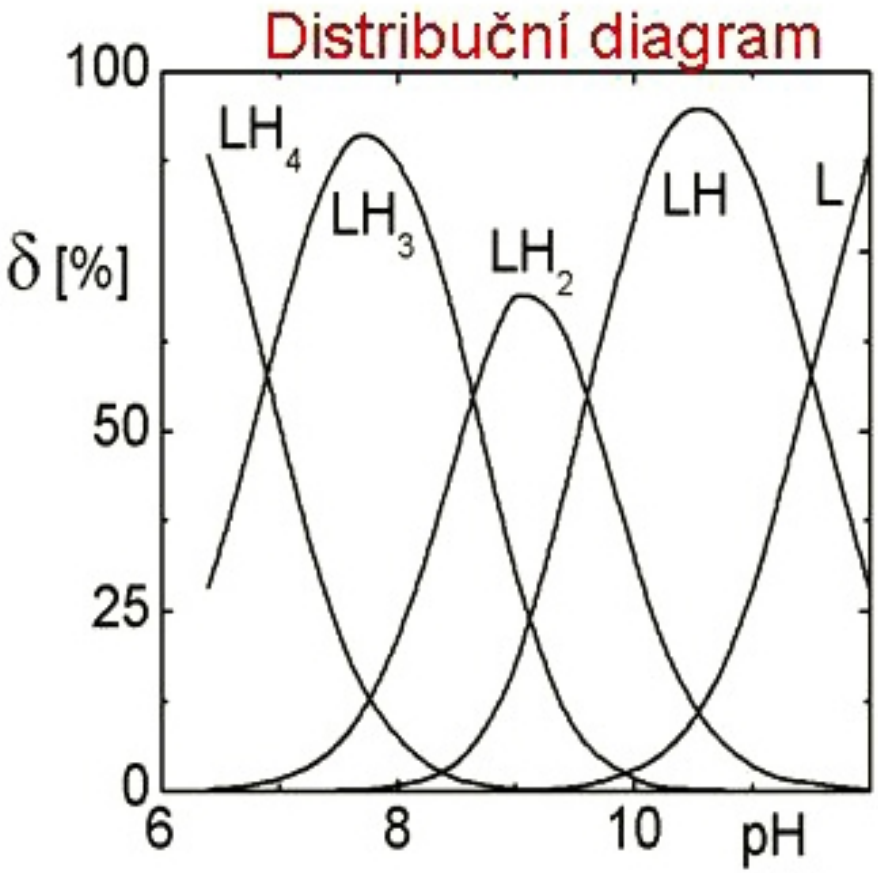
Grafická analýza reziduí



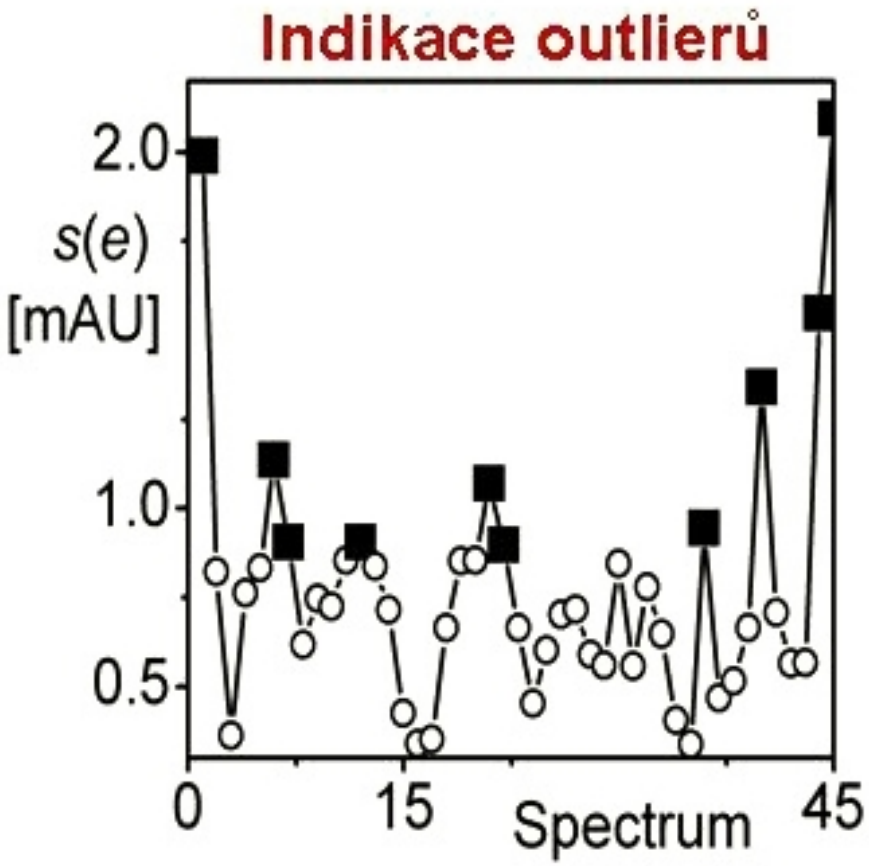


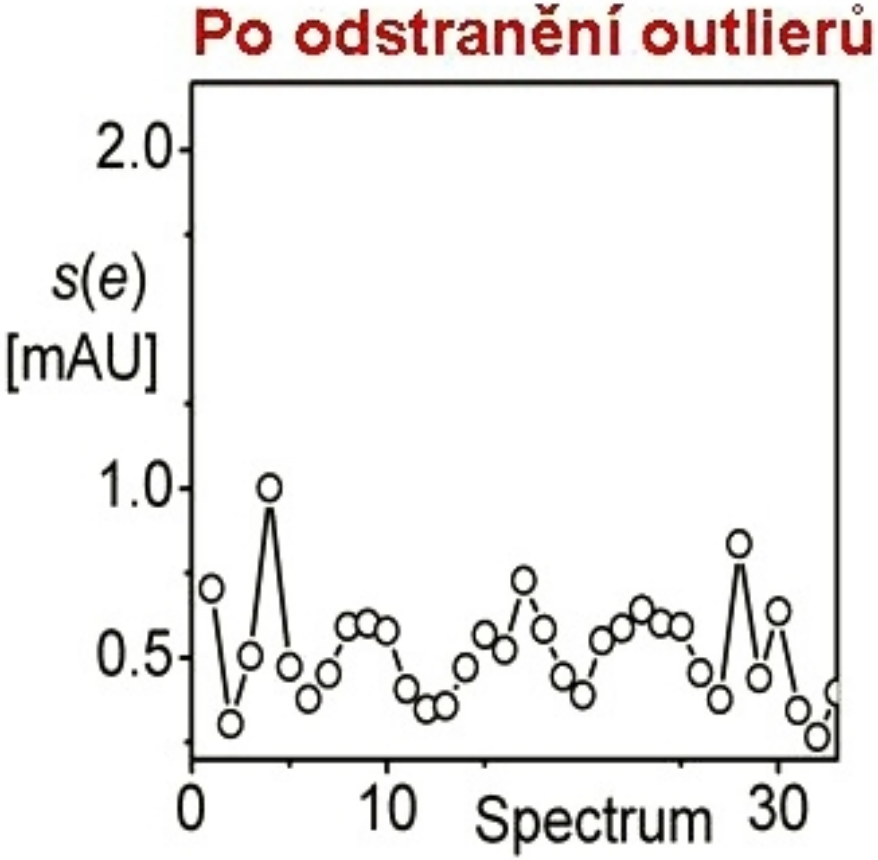
Grafické závěry o nalezených odhadech parametrů z postaveného chemického (=nelineárního regresního) modelu



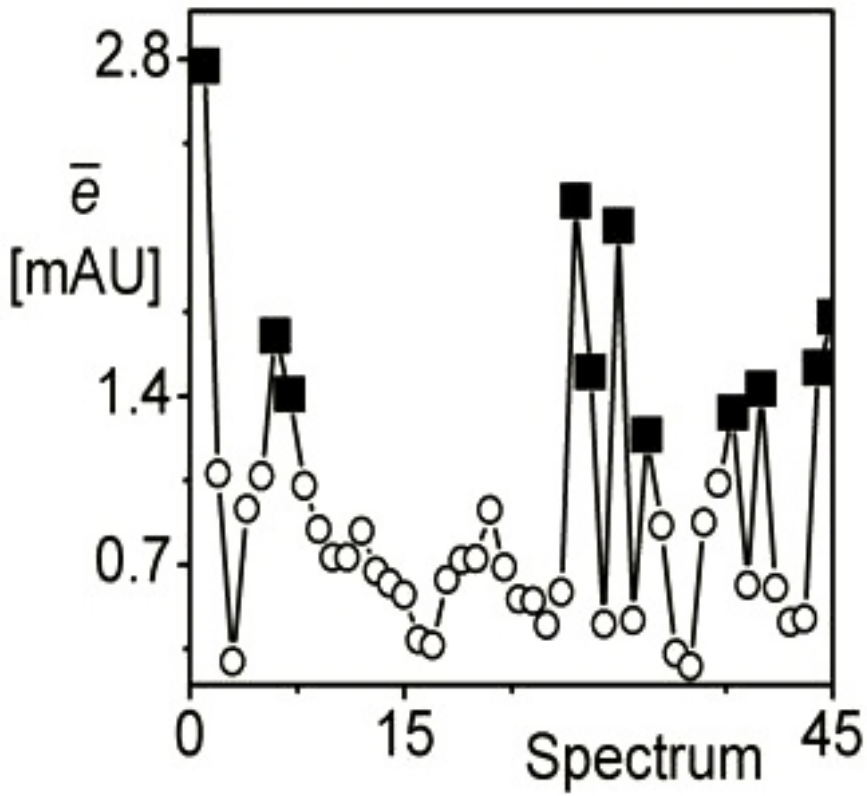


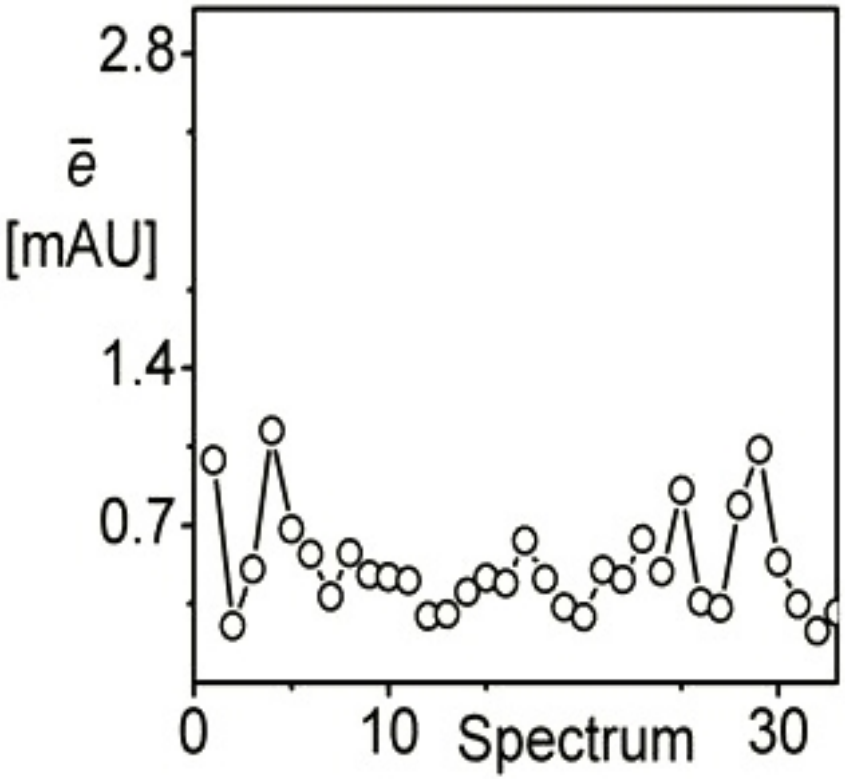
Mírou věrohodnosti nalezených odhadů regresního modelu je vždy míra těsnosti proložení experimentálních bodů vypočtenou regresní křivkou čili statistická analýza reziduí před a po odstranění outlierů.



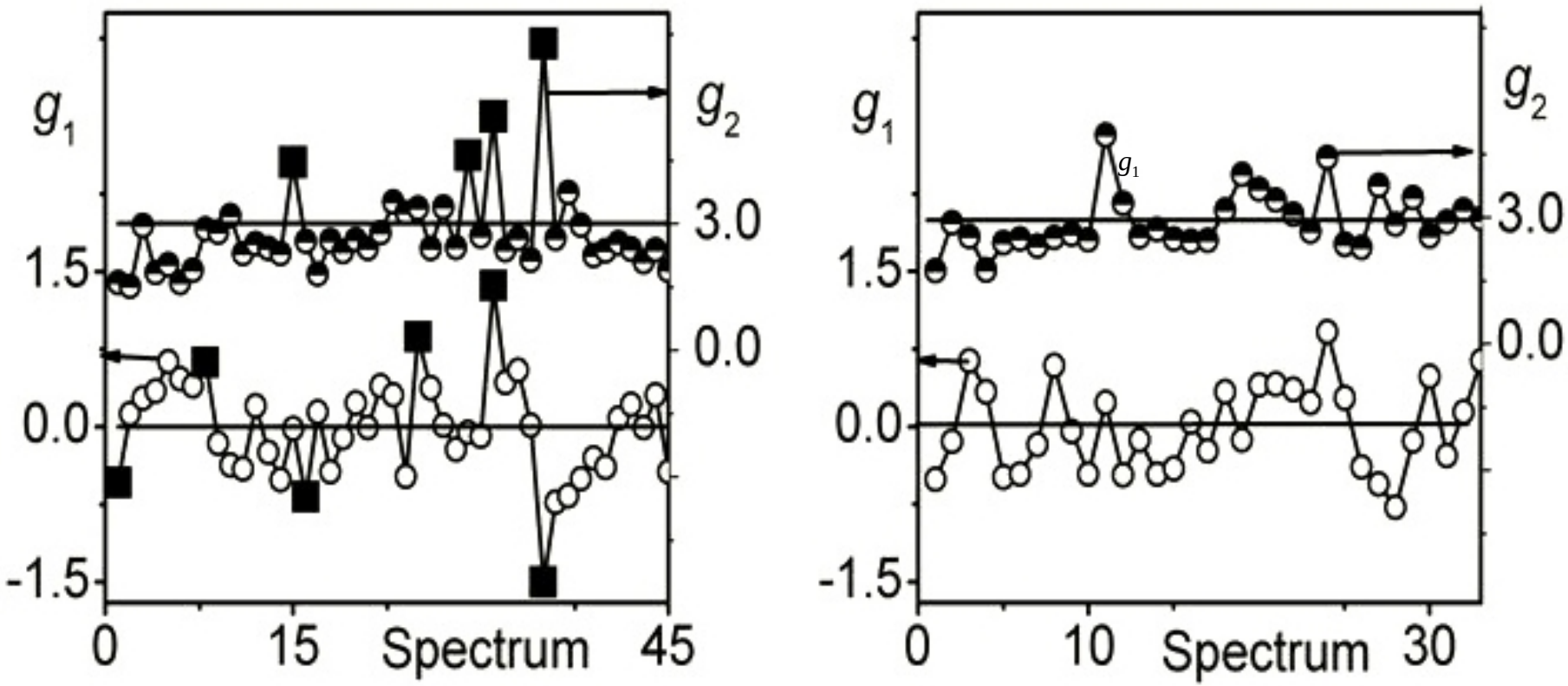


Míru těsnosti proložení je vhodné posuzovat dle velikosti reziduí a jejich rozdělení, a to především dle střední hodnoty reziduí a symetrie rozdělení (vyjádřené koeficientem šikmosti a špičatosti).





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 užijí 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 $\hat{\boldsymbol{e}}_i$ 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

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

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

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

, 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

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 $i = 1, ..., \text{int}(n/2)$) 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.).