Výstavba regresního modelu regresním tripletem

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice

Souhrn: *Postup hledání regresního modelu je popsán obecně a dokumentován na 3 úlohách analytické laboratoře. Skládá se z těchto kroků: 1. Návrh modelu začíná vždy od nejjednoduššího modelu, lineárního. 2. Předběžná analýza dat sleduje proměnlivost proměnných na rozptylových diagramech, indexových grafech. Vyšetřuje se multikolinearita, heteroskedasticita, autokorelace a vlivné body. 3. Odhadování parametrů se provádí klasickou metodou nejmenších čtverců, následuje testování významnosti parametrů Studentovým t-testem. Střední kvadratická chyba predikce MEP a Akaikovo informační kritérium AIC jsou rozhodčí kritéria při hledání modelu. 4. Regresní diagnostika provádí identifikaci vlivných bodů a ověření předpokladů metody nejmenších čtverců. V případě více vysvětlujících proměnných se posoudí vhodnost proměnných pomoci parciálních regresních grafů a parciálních reziduálních grafů. 5. Konstrukce zpřesněného modelu: parametry zpřesněného modelu jsou odhadovány s využitím (a) metody vážených nejmenších čtverců (MVNČ) při nekonstantnosti rozptylu, (b) metody zobecněných nejmenších čtverců (MZNČ) při autokorelaci, (c) metody podmínkových nejmenších čtverců (MPNČ) při omezení kladených na parametry, (d) metody racionálních hodností u multikolinearity, (e) metody rozšířených nejmenších čtverců (MRNČ) pro případ, že všechny proměnné jsou zatížené náhodnými chybami, a konečně (f) robustních metod pro jiná rozdělení než normální a data s vybočujícími hodnotami a extrémy.*

Při výstavbě regresních modelů se běžně užívá metody nejmenších čtverců. Metoda nejmenších čtverců poskytuje postačující odhady parametrů jenom při současném splnění všech předpokladů o datech a o regresním modelu. Pokud tyto předpoklady nejsou splněny, ztrácí metoda nejmenších čtverců své vlastnosti.

Základní předpoklady metody nejmenších čtverců (MNČ): Statistické vlastnosti odhadů $\hat{\boldsymbol{y}}_P, \hat{\boldsymbol{e}}, \boldsymbol{b}$ závisí na splnění jistých předpokladů. Pokud platí předpoklady I až IV, jsou odhady **b** parametrů **β** **n**ejlepší, **n**estranné a **l**ineární (NNLO). Navíc mají asymptoticky normální rozdělení. Pokud platí ještě předpoklad VII, mají odhady **b** normální rozdělení i pro konečné výběry.

I. *Regresní parametry β mohou nabývat libovolných hodnot.* V praxi však často existují omezení parametrů, která vycházejí z jejich fyzikálního smyslu.

II. *Regresní model je lineární v parametrech a platí aditivní model měření.*

III. *Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných* ***X*** *má hodnost rovnou právě m*. To znamená, že žádné její dva sloupce $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ nejsou *kolineární*, tj. rovnoběžné vektory. Tomu odpovídá i formulace, že matice $\boldsymbol{X}^T\boldsymbol{X}$ je symetrická regulární matice, ke které existuje inverzní matice a jejíž determinant je větší než nula.

IV. *Náhodné chyby* ε_i *mají nulovou střední hodnotu* $E(\varepsilon_i) = 0$. To musí u korelačních modelů platit vždy. U regresních modelů se může stát, že $E(\varepsilon_i) = K$, $i = 1, ..., n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení bude $E(\varepsilon_i^{/}) = 0$, kde $\varepsilon_i^{/} = y_i - \hat{y}_{P,i} - K$.

V. *Náhodné chyby* ε_i *mají konstantní a konečný rozptyl* $E(\varepsilon_i^2) = \sigma^2$. Také podmíněný rozptyl $D(y/\boldsymbol{x}) = \sigma^2$ je konstantní a jde o *homoskedastický* případ.

VI. *Náhodné chyby* ε_i *jsou vzájemně nekorelované a platí* $cov(\varepsilon_i\ \varepsilon_j) = E(\varepsilon_i\ \varepsilon_j) = 0$. Pokud mají chyby normální rozdělení, jsou nezávislé. Tento požadavek odpovídá požadavku nezávislosti měřených veličin y.

VII. *Chyby* ε_i *mají normální rozdělení* $N(0, \sigma^2)$. Vektor $\boldsymbol{y}$ má pak vícerozměrné normální rozdělení se střední hodnotou $\boldsymbol{X}\boldsymbol{\beta}$ a kovarianční maticí $\sigma^2\ \boldsymbol{E}$, kde $\boldsymbol{E}$ je jednotková matice.

Regresní diagnostika

Metoda nejmenších čtverců nezajišťuje obecně nalezení přijatelného modelu, a to jak ze statistického, tak i z fyzikálního hlediska. Musí být splněny podmínky, odpovídající složkám tzv. *regresního tripletu* [data, model, metoda odhadu].

Regresní diagnostika obsahuje postupy k identifikaci

a) vhodnosti dat pro navržený regresní model (složka *data*),

b) vhodnosti modelu pro daná data (složka *model*),

c) splnění základních předpokladů MNČ (složka *metoda*).

Základní rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není třeba přesně formulovat alternativní hypotézu. Tímto pojetím se regresní diagnostika blíží spíše k *exploratorní regresní analýze*, která vychází z faktu, že "uživatel ví o analyzovaných datech přece jenom více než počítač". Počítač slouží jako nástroj analýzy dat, modelu a metody odhadu. Model je navrhován v interakci uživatele s programem. Tím by měl být omezen vznik formálních regresních modelů, které nemají fyzikální smysl a jsou v technické praxi obyčejně jen omezeně použitelné.

1. Data: mezi základní techniky diagnostiky patří stanovení rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. K tomu lze využít grafů rozptýlení s kvantily a řady postupů průzkumové analýzy jednorozměrných dat. Přes svoji jednoduchost umožňuje diagnostika identifikovat ještě před vlastní regresní analýzou

a) *nevhodnost dat* (malé rozmezí nebo přítomnost vybočujících bodů),

b) *nesprávnost navrženého modelu* (skryté proměnné),

c) *multikolinearitu*,

d) *nenormalitu* v případě, kdy jsou vysvětlující proměnné náhodné veličiny.

Kvalita dat úzce souvisí s užitým regresním modelem. Při posuzování se sleduje především výskyt *vlivných bodů* (VB), které mohou být hlavním zdrojem řady problémů, jako je zkreslení odhadů a růst rozptylů až k naprosté nepoužitelnosti regresních modelů. Podle toho, kde se vlivné body vyskytují, lze provést dělení na

1. *Vybočující pozorování* (outliers), které se liší v hodnotách vysvětlované (závisle) proměnné y od ostatních, a

2. *Extrémy* (high leverage points), které se liší v hodnotách vysvětlujících (nezávisle) proměnných $\boldsymbol{x}$ nebo v jejich kombinaci (v případě multikolinearity) od ostatních bodů.

Vyskytují se však i body, které jsou jak vybočující, tak i extrémní. K identifikaci vlivných bodů typu vybočujícího pozorování se využívá zejména různých typů reziduí a k identifikaci extrémů pak diagonálních prvků H_{ii} projekční matice $\boldsymbol{H}$.

2. Model: kvalitu regresního modelu lze posoudit v případě jedné vysvětlující proměnné x přímo z rozptylového grafu závislosti y na x. V případě více vysvětlujících proměnných a multikolinearity mohou však rozptylové grafy *mylně indikovat* nelineární trend i u lineárního modelu. Z řady různých grafů k posouzení vztahu y a x_j se omezíme na a) parciální regresní grafy, a b) parciální reziduální grafy.

Parciální regresní grafy byly Belseyem zařazeny mezi základní nástroje počítačové interaktivní analýzy regresních modelů. Umožňují nejenom posouzení kvality navrženého regresního modelu, ale indikují i přítomnost vlivných bodů a nesplnění předpokladů klasické metody nejmenších čtverců. Parciální regresní graf pro posouzení vztahu mezi y a i-tou vysvětlující proměnnou x_i je závislost *reziduí* $\boldsymbol{v}$ regrese y na sloupcích matice $\boldsymbol{X}_{(i)}$ a reziduí u regrese x_i na sloupcích matice $\boldsymbol{X}_{(i)}$. Přitom matice $\boldsymbol{X}_{(i)}$ vznikne z matice $\boldsymbol{X}$ vynecháním i-tého sloupce $\boldsymbol{x}_i$, odpovídajícího i-té vysvětlující proměnné. Parciální regresní grafy mají tyto vlastnosti:

a) Směrnice přímky v parciálním regresním grafu je stejná jako odhad b_j v neděleném modelu a úsek je roven nule. Tato lineární závislost platí pouze v případě, že navržený model je správný.

b) Korelační koeficient mezi oběma proměnnými parciálního regresního grafu odpovídá parciálnímu korelačnímu koeficientu $R_{yx_j(x)}$.

Parciální reziduální grafy se označují také jako grafy *"komponenta + reziduum"*. Parciální reziduální grafy však poskytují poněkud odlišné informace než parciální regresní grafy. Směrnice lineární závislosti je rovna b_j a úsek je nulový. Lineární závislost pak ukazuje na vhodnost navržené proměnné x_j v modelu.

Parciální reziduální grafy se doporučují především k indikaci rozličných typů nelinearity v případě nesprávně navrženého regresního modelu.

3. Metoda: V praxi bývají některé předpoklady MNČ porušeny, což vede k použití jiných kritérií. K porušení předpokladů dochází v těchto základních případech:

a) Na parametry jsou kladena omezení, což vede na užití *metody podmínkových nejmenších čtverců (MPNČ).*

b) Kovarianční matice chyb není diagonální (autokorelace), příp. data nemají stejný rozptyl (heteroskedasticita), což vede na užití *metody zobecněných nejmenších čtverců (MZNČ)*, resp. *metody vážených nejmenších čtverců (MVNČ).*

c) Rozdělení dat nelze považovat za normální nebo se v datech vyskytují vlivné body. V takovém případě se místo kritéria metody nejmenších čtverců užije *robustního* kritéria, které je na porušení předpokladu o rozdělení chyb a na vlivné body málo citlivé. Z robustních kritérií jsou nejznámější *M-odhady*. Jedná se o maximálně věrohodné odhady pro vhodnou hustotu pravděpodobnosti chyb. Pro odhad parametrů $\boldsymbol{b}$ se užívá *iterační metody vážených nejmenších čtverců* (*IVNČ*).

d) Také proměnné x mohou být zatížené náhodnými chybami, což vede na užití *metody rozšířených nejmenších čtverců (MRNČ).* Pro případ regresní přímky je použití metody rozšířených nejmenších čtverců velmi jednoduché. Postačuje znalost poměru rozptylu σ_y^2 (vysvětlovaná proměnná) a σ_x^2 (vysvětlující proměnné), $K = \sigma_y^2/\sigma_x^2$. Pro odhad směrnice regresní přímky $y = a\,x + b$ pak platí

$$a = L + \operatorname{sign}(S_{yx})\sqrt{K + L^2}$$

kde

$$L = \frac{S_{yx} - K\,S_x}{2S_x}$$

a sign S_{yx} je znaménková funkce. Symboly S označují součty čtverců, odpovídajících proměnných

$$S_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \ , \quad S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \ , \quad S_{yx} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Při znalosti odhadu směrnice $\hat{a}$ se snadno určí odhad úseku $\hat{b}$ ze vztahu

$$\hat{b} = \bar{y} - \hat{a}\,\bar{x}$$

Pro případ stejných rozptylů, tj. $K = 1$ vede dosazení do výše uvedených vztahů k odhadům minimalizujícím kolmé vzdálenosti (*orthogonální regrese).* Pro odhady rozptylů odhadů $\hat{a}$, $\hat{b}$ se pak používá speciálních vztahů.

e) Pro špatně podmíněné matice $\boldsymbol{X}^T\boldsymbol{X}$ se používá *metoda racionálních hodností*, vedoucí k systému vychýlených odhadů, kde vychýlení je řízeno jedním parametrem.

Postup výstavby lineárního regresního modelu:

1. Návrh modelu: začíná se vždy od nejjednoduššího modelu, u kterého vystupují jednotlivé vysvětlující proměnné v prvních mocninách a nevyskytují se žádné interakční členy typu $x_j\, x_k$.

2. Předběžná analýza dat: sleduje se proměnlivost jednotlivých proměnných a možné párové vztahy. Užívá se proto rozptylových diagramů závislosti x_j na x_k nebo indexových grafů závislosti x_j na j. Posuzuje se významnost proměnných s ohledem na jejich proměnlivost a přítomnost multikolinearity. Přibližně lineární vztah mezi proměnnými v rozptylových grafech závislosti x_j na x_k indikuje multikolinearitu. Lze rovněž odhalit vlivné body, které způsobují multikolinearitu.

Podle volby uživatele se provedou požadované transformace původních proměnných. Zadává se, zda model obsahuje absolutní člen. Uživatel může volit polynomickou transformaci zadáním stupně polynomu.

Provádí se sestavení korelační matice $\boldsymbol{R}$ a její rozklad na vlastní čísla a vlastní vektory. Jsou vypočteny *VIF* k indikaci multikolinearity a tisknuta setříděná vlastní čísla. K určení inverzní matice $\boldsymbol{R}^{-1}$ se užívá metoda racionálních hodností pro standardně zadávané vychýlení $P = 10^{-15}$. Uživatel může zadat jinou hodnotu parametru vychýlení P, což však vede pro vyšší hodnoty P k vychýleným odhadům. Bývá proto vhodné volit P z tohoto intervalu $10^{-5} \le P \le 10^{-3}$.

3. Odhadování parametrů: odhadování parametrů modelu se provádí metodou racionálních hodností s volbou $P = 10^{-5}$. Ze zobecněné inverzní matice $\boldsymbol{R}^{-1}$ jsou určovány odhady parametrů $\boldsymbol{b}$, jejich směrodatné odchylky $\sqrt{D(b_j)}$ a velikosti testačních statistik Studentova t-testu významnosti pro $\beta_j = 0$. Dále jsou provedeny testy významnosti odhadů b_j, vícenásobného korelačního koeficientu R a koeficientu determinace D. Je vhodné sledovat souhrnné charakteristiky regrese jako je střední kvadratická chyba predikce *MEP* a Akaikovo informační kritérium *AIC*, případně posoudit linearitu modelu.

4. Regresní diagnostika: s využitím pěti rozličných grafů je prováděna identifikace vlivných bodů, a to *grafy Wiliamsovým, Pregibonovým, McCulloh-Meeterovým, L-R,* a *grafem predikovaných reziduí*. Dále pak ověření splnění předpokladů metody nejmenších čtverců jako je homoskedasticita, nepřítomnost autokorelace a normalita rozdělení chyb. Pokud dojde k úpravě dat, je třeba provést znovu regresní diagnostiku se zaměřením na porušení předpokladů metody nejmenších čtverců a posouzení vlivu multikolinearity. V případě více vysvětlujících proměnných se posoudí vhodnost jednotlivých proměnných a jejich funkcí s využitím parciálních regresních grafů nebo grafů "komponenta + reziduum". *Tabulka reziduí*

obsahuje klasická rezidua $\hat{e}_i$, normovaná rezidua $\hat{e}_{Ni}$, standardizovaná rezidua $\hat{e}_{Si}$ a Jackknife rezidua $\hat{e}_{Ji}$. Je uveden odhad autokorelačního koeficientu reziduí prvního řádu $\hat{\varrho}_1$. *Tabulka vlivných bodů* obsahuje veličiny H_{ii}, $\boldsymbol{H}^*_{\boldsymbol{ii}}$, D_i, A_i, DF_i, $LD_i(\boldsymbol{b})$, $LD_i(\hat{\sigma}^2)$ a $LD_i(\boldsymbol{b}, \sigma^2)$. Hvězdičkou jsou označeny hodnoty silně vlivných bodů.

5. Konstrukce zpřesněného modelu: s využitím

a) *metody vážených nejmenších čtverců (MVNČ)* při nekonstantnosti rozptylů,

b) *metody zobecněných nejmenších čtverců (MZNČ)* při autokorelaci,

c) *metody podmínkových nejmenších čtverců (MPNČ)* při omezeních na parametry,

d) *metody racionálních hodností RH* u multikolinearity,

e) *metody rozšířených nejmenších čtverců (MRNČ)* pro případ, že všechny proměnné jsou zatížené náhodnými chybami,

f) *robustní metody* pro jiná rozdělení dat než normální a data s vybočujícími hodnotami a extrémy jsou odhadovány parametry zpřesněného modelu.

6. Zhodnocení kvality modelu: s využitím klasických testů, postupů regresní diagnostiky a doplňkových informací o modelované soustavě se provede posouzení kvality navrženého lineárního regresního modelu.

Vzorová úloha: *Model teplotní závislosti přechodového tlaku bismutu* (J6.01)

Ukážeme postup analýzy jednorozměrného lineárního regresního modelu. Byl studován přechodový tlak bismutu I - II *p* jako funkce teploty *t*. Nalezněte lineární regresní model, který bude adekvátní daným datům. Vyšetřete regresní triplet a indikujte vlivné body.

Data: Teplota t [°C], tlak *p* [bar]:

20.8	25276,	20.9	25256,	21.0	25216,	21.9	25187,	22.1	25217,
22.1	25187,	22.4	25177,	22.5	25177,	24.8	25098,	24.8	25093,
25.0	25088,	34.0	24711,	34.0	24701,	34.1	24716,	42.7	24374,
42.7	24394,	42.7	24384,	49.9	24067,	50.1	24057,	50.1	24057,
22.5	25147,	23.1	25107,	23.0	25077				

Řešení:

1. Odhadování parametrů: klasickou metodou nejmenších čtverců (MNČ) byly nalezeny nejlepší odhady úseku β_0 a směrnice β_1. Studentův *t*-test ukázal, že úsek (absolutní člen) β_0 je statisticky významný a směrnice β_1 je statisticky významná.

	Odhad	Směrodatná odchylka	Test H0: B[j] = 0 vs. HA: B[j] <> 0 t-kriterium	hypoteza H0 je	Hlad. význam.
B[0]	2.6068E+04	1.6169E+01	1.6122E+03	Zamítnuta	0.000
B[1]	-3.9874E+01	5.0419E-01	-7.9084E+01	Zamítnuta	0.000

2. Regresní diagnostika: absolutní hodnota *párového korelačního koeficientu R* ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota *koeficientu determinace* $D = R^2$ (99.67%), představuje procento variability, vysvětlené modelem. *Predikovaný koeficient determinace* R^2_p ukazuje na predikční schopnost modelu, je však vyčíslen jinak než R^2, místo *RSC* se ve vztahu užije *MEP*. *Střední kvadratická chyba predikce MEP* a *Akaikovo informační kritérium AIC* se užívají k rozlišení mezi několika navrženými modely. Za optimální se považuje model, pro který dosahuje *MEP* a *AIC* minimální hodnotu.

Vícenásobný korelační koeficient, R	: 9.9833E-01
Koeficient determinace, *D*	: 9.9665E-01
Predikovaný koeficient determinace, Rp^2	: 9.9804E-01

Střední kvadratická chyba predikce, *MEP*	: 6.8546E+02
Akaikeho informační kritérium, *AIC*	: 1.5054E+02

3. Konstrukce zpřesněného modelu:

(a) Po odstranění bodů č. 23 (*kritika dat*) byly nalezeny nové odhady parametrů zpřesněného modelu. Zpřesněný model (v závorce je uveden vždy odhad směrodatné odchylky parametru) y = 26 078 (13) - 40.1 (0.4) x_1 je doložen statistickými charakteristikami: *párový korelační koeficient* R = 0.9990, *koeficient determinace* D = 99.808% a *predikovaný korelační koeficient* R_P = 0.99885 dosáhly vesměs vysokých hodnot. *Střední kvadratická chyba predikce MEP* = 414.22 a *Akaikeho informační kritérium AIC* = 132.62 dosáhly nižších hodnot než u předešlého modelu, což dokazuje, že zpřesněný model je lepší. Rezidua nyní vykazují normální rozdělení a nevykazují trend, stále však vykazují heteroskedasticitu, a proto lze doporučit použít metodu vážených nejmenších čtverců.

(b) Užitím statistické váhy ($w_i = 1/y_i^2$) kompenzujeme heteroskedasticitu v datech. Obdržíme nové odhady parametrů. Opravený model má tvar, (v závorce je uveden odhad směrodatné odchylky parametru) y = 26 079 (13) - 40.1 (0.4) x_1. Jelikož došlo ke snížení rozhodujících kritérií, tj. *střední kvadratické chyby predikce MEP* = 410.29 a *Akaikeho informačního kritéria AIC* = 132.39, lze považovat tyto odhady za lepší než předešlé.

4. Zhodnocení kvality modelu: porovnáním hodnot regresní diagnostiky lze snadno provést zhodnocení *regresního tripletu* u dosaženého lineárního regresního modelu pro upravená data, zbavená odlehlých hodnot a metodou vážených nejmenších čtverců. Nalezený a prokázaný model teplotní závislosti přechodového tlaku bizmutu má tvar, (v závorce je vždy uveden odhad směrodatné odchylky parametru)

$$y = 26\,079\ (13) - 40.1\ (0.4)\ x_1.$$

Poděkování: Práce vznikla za podpory grantu Ministerstva zdravotnictví NS9831-4/2008 a vědeckých záměrů MSMT0021627502.

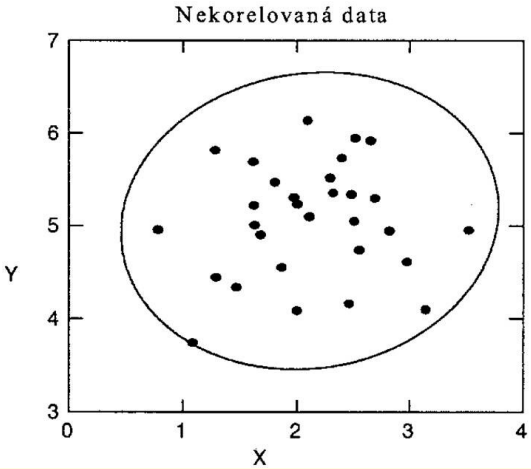
Doporučená literatura

[1] Meloun M., Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994, Academia Praha 2004..

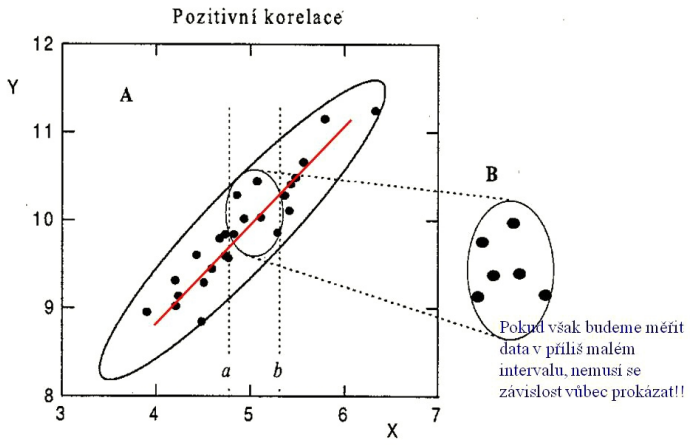
[2] Meloun M., Militký J., Hill M., *Počítačová analýza vícerozměrných dat v příkladech,* Academia Praha 2005.

[3] Meloun M., Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha 2002, 2006.

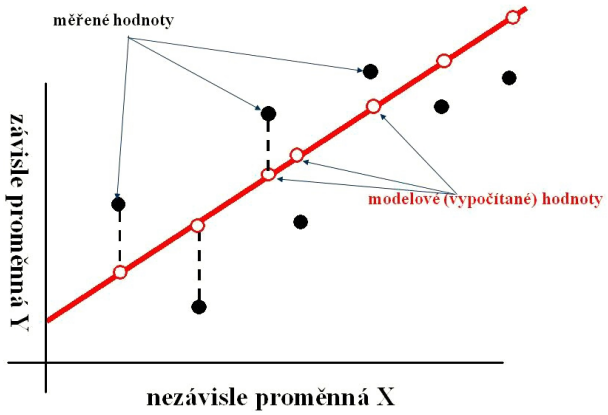
STATISTICKÁ ZÁVISLOST



STATISTICKÁ ZÁVISLOST



Grafické vysvětlení:



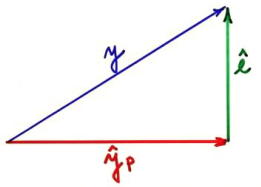
Definice regresního modelu

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

y – závisle proměnná; **X** – nezávisle proměnná; **β** – regresní parametry; **ε** – náhodná chyba

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

Metoda projekčních matic



y vektor proměnné
$\hat{y}_P$ vektor predikce
$\hat{e}$ vektor reziduí

Cílem: minimální délka vektoru $\hat{e} = y - \hat{y}_p$

délku vektoru vyjádříme $D = \sqrt{\langle \hat{e}, \hat{e} \rangle} = \sqrt{\sum_{i=1}^{n} \hat{e}_i^2}$

Odhady modelových parametrů **b** minimalizují čtverec délky vektoru

$$D^2 = \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{m} x_{ij} b_j \right]^2 = \sum_{i=1}^{n} (y_i - \hat{y}_{P,i})^2$$

Důsledek: vektor reziduí $\hat{e}$ je kolmý na všechny sloupce matice **X**, a proto odpovídající skalární součiny jsou nulové

$$\langle \mathbf{x}_j, \hat{\mathbf{e}} \rangle = \sum_{i=1}^{n} x_{ij}\, \hat{e}_i = 0, \qquad j = 1, \ldots, m$$

a maticově $\mathbf{X}^T \hat{\mathbf{e}} = 0$

a dosazením $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\,\mathbf{b}$ bude

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X}\, \mathbf{b}$$

Odhad **b**, minimalizující vzdálenost D, má tvar

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

projekční matice **H**, pro kterou platí

$$\hat{\mathbf{y}}_P = \mathbf{H}\, \mathbf{y}$$

a pomocí vektoru **b**

$$\hat{\mathbf{y}}_P = \mathbf{X}\, \mathbf{b} = \mathbf{X}(\mathbf{X}^T)\mathbf{X}^{-1}\mathbf{X}\, \mathbf{y}$$

Vlastnosti projekční matice H:

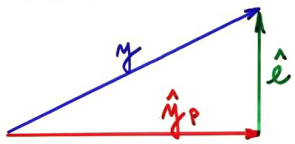
1. Projekční matice $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^T \mathbf{X})^{-1}\, \mathbf{X}^T$ promítne libovolný vektor **V** do roviny L.
2. Pokud vektor **V** již leží v rovině L, platí **H V = V**.
3. Pokud je však vektor **V** kolmý na rovinu L, platí **H V = 0**, kde **0** je nulový vektor.

Vlastnosti projekční matice P:

Projekční matice **P** pro kolmou projekci do nadroviny $L^{\perp}$, kolmé na nadrovinu L, má tvar

$$\mathbf{P} = \mathbf{E} - \mathbf{H}$$

kde **E** je jednotková matice.



Závěr: Rozklad vektoru **y** do dvou složek

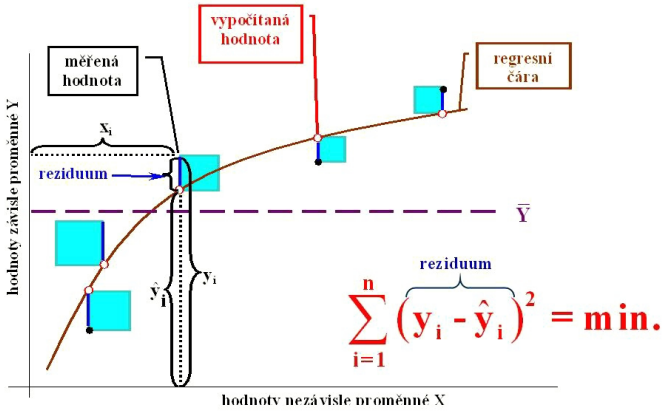
$$\mathbf{y} = \mathbf{H}\, \mathbf{y} + \mathbf{P}\, \mathbf{y} = \hat{\mathbf{y}}_P + \hat{\mathbf{e}}$$

Geometricky: vektror **y** byl rozložen na dva kolmé vektory

Předpoklady metody nejmenších čtverců

1. Parametry $\boldsymbol{\beta}$ mohou nabývat libovolných hodnot. Omezení jsou pouze fyzikálního smyslu.
2. Model je lineární v parametrech $\boldsymbol{\beta}$; platí přitom additivní model měření $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
3. Matice $\mathbf{X}$ je nenáhodná, nastavitelných hodnot nezávisle proměnných. Má hodnost m:
   (a) Žádné dva sloupce x_j a x_k nejsou kolineární (čili paralení).
   (b) $\mathbf{X}^T\mathbf{X}$ je pak symetrická regulární matice.
   (c) Rovina L je m rozměrná a vektory $\mathbf{X}\,\mathbf{b}$ jsou jednoznačné.
4. Náhodné chyby ϵ_i mají nulovou střední hodnotu $E(\epsilon_i) = 0$. Je-li $E(\epsilon_i) = K$, je nutno zavést absolutní člen a pak bude $E(\epsilon_i) = 0$.
5. Náhodné chyby ϵ_i mají konstantní rozptyl, homoskedasticita, $E(\epsilon_i^2) = \sigma^2$.
6. Náhodné chyby ϵ_i jsou vzájemně nekorelované, $\mathrm{cov}(\epsilon_i \,.\, \epsilon_j) = E(\epsilon_i \,.\, \epsilon_j) = 0$.
7. Náhodné chyby mají normální rozdělení $\epsilon \approx N(0, \sigma^2)$.

VYČÍSLENÍ ODHADŮ PARAMETRŮ REGRESNÍHO MODELU METODOU NEJMENŠÍCH ČTVERCŮ (MNČ)



$$\sum_{i=1}^{n} \overbrace{\left(y_i - \hat{y}_i\right)}^{\text{reziduum}}{}^2 = \min.$$

Test homoskedasticity

1. Pomocí testu trendu reziduí

$$D = \sum_{i=1}^{n} \left[R\left(\left|\hat{e}_i\right|\right) - i \right]^2$$

$$\rho_s = 1 - \frac{6}{n^3 - n} \cdot D$$

Testujeme významnost Spearmanova korelačního koeficientu ρ_s

$$t_R = \frac{\rho_s \cdot \sqrt{n-2}}{\sqrt{1-\rho_s^2}}$$

2. Pomocí Cookova - Weisbergova testu

Vycházíme z předpokladu, že rozptyl naměřené hodnoty y_i je určitou funkcí proměnné $x_i\,\beta$ (např. exponenciální funkcí)

$$S_f = \frac{\left[\sum_{i=1}^{n} \left(y_i' - \bar{y}'\right)^2 e_i^2\right]^2}{2 \cdot \sigma^4 \sum_{i=1}^{n} \left(y_i' - \bar{y}'\right)^2}$$

Pravidlo: Pokud v datech není heteroskedasticita, potom platí, že $S_f < \chi^2(1)$

Kompenzace nesplnění předpokladu homoskedasticity

Použití metody **vážených nejmenších čtverců**, kdy se suma reziduí násobí vhodně zvolenými váhami:

$$U(\boldsymbol{b}) = \sum_{i=1}^{n} \left[y_i V_{ii} - \sum_{j=1}^{m} V_{ii} x_{ij} b_j \right]^2$$

Pravidlo: Běžně se za váhy volí hodnoty $1/y_i$ nebo $1/y_i^2$.

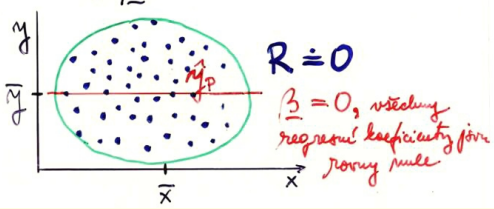
VYUŽITÍ VÍCENÁSOBNÉHO KORELAČNÍHO KOEFICIENTU R

$$R = \sqrt{\frac{TSC}{CSC}} = \sqrt{1 - \frac{RSC}{CSC}} = \cos\alpha$$

$$= \sqrt{1 - \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

DISKUSE: (a) $\hat{R} \to 0$, pak platí $H_0: \hat{y}_p = \underline{J} \cdot \bar{y}$

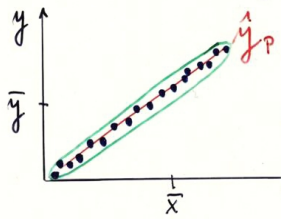
čili $H_0: \underline{\beta} = 0$



$R \doteq 0$

$\underline{\beta} = 0$, všechny regresní koeficienty jsou rovny nule

(b) $\hat{R} \to 1$, pak platí $H_A: \hat{y}_p = \mathbb{X} \cdot \underline{b}$ čili $H_A: \underline{\beta} \neq 0$



$R \doteq 1$

$\underline{\beta} \neq 0$, všechny regresní koeficienty jsou různé od nuly

TESTOVÁNÍ R: $H_0: \underline{\beta} = 0$ čili $\hat{R}^2 = 0$ vs. $H_A: \underline{\beta} \neq 0$ čili $\hat{R}^2 \neq 0$

$$F_R = \frac{(CSC - RSC)(n-m)}{RSC(m-1)} = \frac{\hat{R}^2 (n-m)}{(1-\hat{R}^2)(m-1)}$$

porovnání s $F_{1-\alpha}(m-1, n-m)$

Příklad 6.8 *Omezení klasické analýzy lineárního modelu*
Anscomb[5] uvádí testační data pro čtyři simulované výběry. Testujte statistickou významnost obou parametrů β_1 a β_2 a proveďte grafickou analýzu reziduí.
Data: čtyři simulované výběry vykazují stejné charakteristiky
$b_1 = 0.5$, $b_2 = 3.0$, $D(b_1) = 0.0139$ a $D(b_2) = 1.2656$.

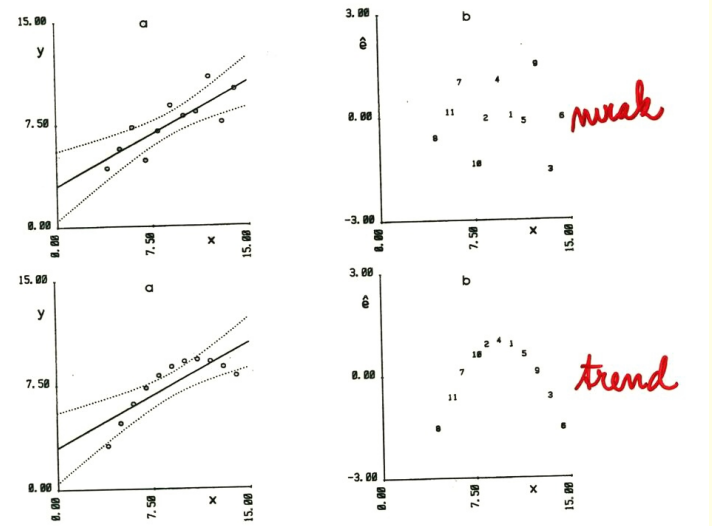
Výběr	A		B		C		D	
Proměnná	x	y	x	y	x	y	x	y
1	10	8.04	8	9.14	8	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	8	8.74	8	12.74	8	7.71
4	9	8.81	8	8.77	8	7.11	8	8.84
5	11	8.33	8	9.26	8	7.81	8	8.47
6	14	9.96	8	8.10	8	8.84	8	7.04
7	6	7.24	8	6.13	8	6.08	8	5.25
8	4	4.26	8	3.10	8	5.39	19	12.50
9	12	10.84	8	9.13	8	8.15	8	5.56
10	7	4.82	8	7.26	8	6.42	8	7.91
11	5	5.68	8	4.74	8	5.73	8	6.89

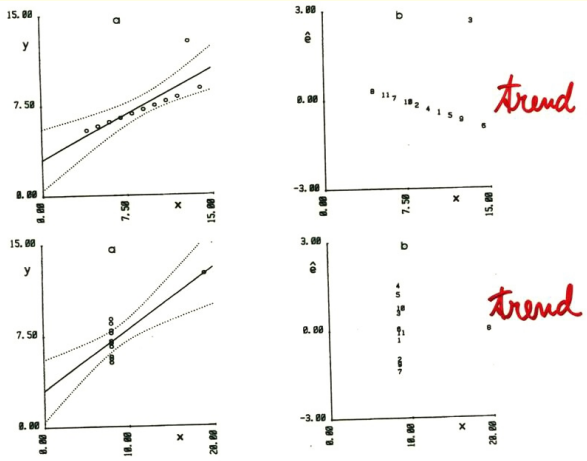
Řešení:
Lineární regresní model $E(y/x) = \beta_1 x + \beta_2$ má pro všechny výběry

a) ***Stejné*** **odhady parametrů:**
$b_1 = 0.5$, $b_2 = 3.0$, $D(b_1) = 0.0139$, $D(b_2) = 1.2656$.

b) ***Stejné*** **testační statistiky významnosti parametrů:**
$T_1 = 2.667$ a $T_2 = 4.241$.

c) ***Stejné*** **testační statistiky:** $F_R = 17.97$, $\hat{R}^2 = 0.66$, $\hat{\sigma} = 1.237$ ukazují,
že β_1 a β_2 jsou významně odlišné od nuly.





Závěr: **Neshodu modelu s daty indikuje grafická analýza reziduí**

Konstrukce intervalů spolehlivosti

Bodové odhady parametrů β mají pro praxi menší význam.

Intervaly spolehlivosti (konfidenční oblasti), ve kterých leží teoretická hodnota β se zvolenou pravděpodobností $(1 - \alpha)$.

1. Interval spolehlivosti parametrů β

Pro hladinu významnosti $\alpha = 0.10$, 0.05 nebo 0.01 odpovídají 90%ní, 95%ní nebo 99%ní intervaly spolehlivosti.

$\alpha = 0.10$ odpovídá
90%ní jistota

$\alpha = 0.05$ odpovídá
95%ní jistota

$\alpha = 0.01$ odpovídá
99%ní jistota

2. Interval spolehlivosti predikce $\hat{y}_{P,i}$:

$100(1 - \alpha)$%ní interval spolehlivosti predikce $\hat{y}_{P,i}$ v místě x_0:

$$b - t_{1-\alpha/2}(n-m)\, \hat{s}_{P,0} \le x_0^T \beta \le x_0^T b + t_{1-\alpha/2}(n-m)$$

kde $\hat{s}^2_{P,0}$ je rozptyl predikce

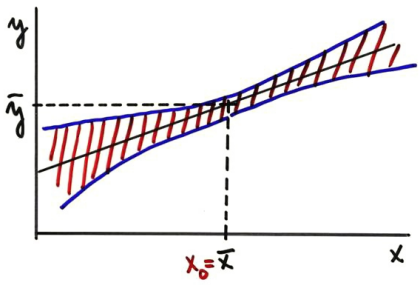
$$D(\hat{y}_{P,0}) = s^2_{P,0} = \hat{\sigma}^2\, x_0^T (X^T X)^{-1} x_0$$

$100(1 - \alpha)$%ní pásy spolehlivosti (resp. konfidenční pásy): jsou nejužší v těžišti nezávisle proměnných, $x_{0j} = \bar{x}_j$.

Scheffého metoda: s pravděpodobností $(1 - \alpha)$ leží teoretická čili "pravdivá" hodnota $x^T \beta$ v oblasti pásů

$$x^T b \pm \sqrt{m\, F_{1-\alpha}(m, n-m)}\; \hat{\sigma} \sqrt{x^T (X^T X)^{-1} x}$$

a konfidenční pásy se nazývají ***pásy Workinga-Hottellinga***.



TEST VÝZNAMNOSTI REGRESNÍHO MODELU – co testujeme?

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_m x_m$$

Testuje se významnost odhadů jednotlivých parametrů:

Test: Jestliže je testovaný odhad parametru statisticky nevýznamný, pak jeho příslušná proměnná x_j **nepřispívá** ke zpřesnění odhadu závisle proměnné y a tato proměnná x_j je v modelu zbytečná.

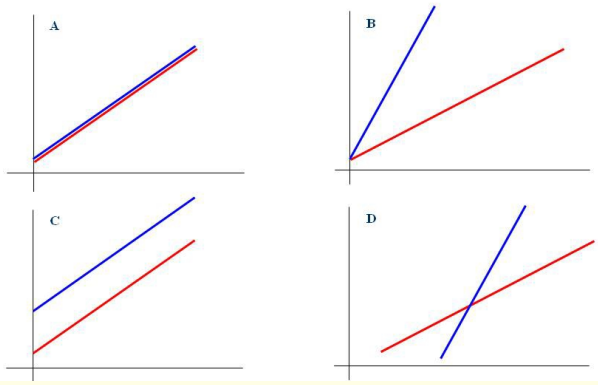
Testuje se model jako celek:
Test: zda příslušná kombinace všech nezávisle proměnných statisticky významně zpřesní odhad závisle proměnné y oproti použití pouhého průměru všech hodnot y.

PODSTATNÉ TESTY VÝZNAMNOSTI V KORELAČNÍ A REGRESNÍ ANALÝZE

- test významnosti korelačního koeficientu
- test významnosti modelu jako celku
- test významnosti jednotlivých regresních parametrů
- test shody lineárních regresních modelů

a mnoho dalších testů.....

Porovnání regresních přímek



Porovnání regresních přímek

Porovnání M navržených regresních modelů

$$y_{ij} = \beta_{2j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \qquad j = 1, \dots, M \quad i = 1, \dots, n_j$$

Postup:

1. krok: vyčíslení b_{2j}, b_{1j}, $\hat{\sigma}_j^2$, $j = 1, \dots, \cancel{n}, M$

2. krok: test homoskedasticity $\hat{\sigma}_j^2 = \hat{\sigma}^2$, $j = 1, \dots, \cancel{n}, M$

3. krok: testování, zda-li
   a) regresní přímky mají společný průsečík,
   b) regresní přímky mají společnou směrnici,
   c) regresní přímky jsou totožné.

Test homoskedasticity (Bartlettův test)

H_0: $\sigma_j^2 = \sigma^2$, $j = 1, \dots, M$,
se stupni volnosti $v_j = (n_j - m)$, pro přímku $m = 2$

Celkový počet stupňů volnosti $V = \sum_{j=1}^{M} v_j$

$$L = 1 + \frac{\sum_{j=1}^{M} v_j^{-1} - V^{-1}}{3M - 3}$$

Sdružený odhad rozptylu $\hat{\sigma}_c^2 = \sum_{j=1}^{M} \frac{v_j \hat{\sigma}_j^2}{V}$ je vážená kombinace odhadů rozptylu jednotlivých modelů,

Testační kritérium Bartlettova testu

$$B = \frac{V \ln \hat{\sigma}_c^2 - \sum_{j=1}^{M} v_j \ln \hat{\sigma}_j^2}{L}$$

Test: Je-li $B < \chi^2_{1-\alpha}(M - 1)$, je H_0 přijmuta (homoskedasticita).

K porovnání dvou přímek (skupin bodů), $M = 2$,

stačí testační kritérium $F_2 = \frac{\max(\hat{\sigma}_1^2, \hat{\sigma}_2^2)}{\min(\hat{\sigma}_1^2, \hat{\sigma}_2^2)}$

Test homogenity úseků

H_0: $\beta_{21} = \beta_{22} = \dots = \beta_{2j} = \dots = \beta_{2M} = \beta_{2c}$,

Sdružený odhad úseku β_{2c} je vážená kombinace odhadů jednotlivých úseků

$$b_{2c} = \frac{\sum_{j=1}^{M} w_{Bj} b_{2j}}{\sum_{j=1}^{M} w_{Bj}}$$

kde j-tý váhový koeficient w_{Bj} úseku j-té přímky

$$w_{Bj} = \frac{n_j \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{n_j} x_{ij}^2}$$

Testační statistika F-testu

$$F_I = \frac{\frac{1}{M-1} \sum_{j=1}^{M} w_{B,j} (b_{2j} - b_{2c})^2}{\frac{1}{n - 2M} \sum_{j=1}^{M} \sum_{i=1}^{n_j} \hat{e}_{ij}^2}, \quad \text{kde } n = \sum_{j=1}^{M} n_j$$

Rezidua $\hat{e}_{ij}$ jsou určována z jednotlivých přímek a platí

$$\sum_{j=1}^{M} \sum_{i=1}^{n_j} \hat{e}_{ij}^2 = \sum_{j=1}^{M} RSC_j ,$$

kde RSC_j je reziduální součet čtverců v j-té skupině.

Test: Je-li $F_I < F_{1-\alpha}(M - 1, n - 2M)$, je H_0 přijata a všechny přímky mají stejný úsek.

Nejlepším odhadem "společného úseku" je b_{2c}
a jeho rozptyl je

$$D(b_{2c}) = \frac{\hat{\sigma}^2}{\sum_{j=1}^{M} w_{Bj}} = \frac{\frac{1}{n - 2M} \sum_{j=1}^{M} \sum_{i=1}^{n_j} \hat{e}_{ij}^2}{\sum_{j=1}^{M} w_{Bj}}$$

Test homogenity směrnic

(test rovnoběžnosti regresních přímek)

H_0: $\beta_{11} = \beta_{12} = \ldots \beta_{1j} = \ldots = \beta_{1M} = \beta_{1c}$,

Sdružený odhad celkové směrnice β_{1c} je vážená kombinace jednotlivých směrnic b_{1j} dle

$$b_{1c} = \frac{\sum_{j=1}^{M} w_{Sj}\, b_{1j}}{\sum_{j=1}^{M} w_{Sj}}, \quad \text{kde} \quad w_{Sj} = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

Testační statistika F-testu

$$F_S = \frac{\frac{1}{M-1} \sum_{j=1}^{M} w_{Sj}\, (b_{1j} - b_{1c})^2}{\frac{1}{n-2M} \sum_{j=1}^{M} \sum_{i=1}^{n_j} \hat{e}_{ij}^2}$$

Test: Je-li $F_S < F_{1-\alpha}(M - 1, n - 2M)$, H_0 je přijata a regresní přímky jsou rovnoběžné.

Nejlepším odhadem celkové směrnice je b_{1c} a její rozptyl je

$$D(b_{1c}) = \frac{\frac{1}{n-2M} \sum_{j=1}^{M} \sum_{i=1}^{n_j} \hat{e}_{ij}^2}{\sum_{j=1}^{M} w_{Sj}}$$

Test shody regresních přímek

H_0: $\beta_{2j} = \beta_{2c}$, $\beta_{1j} = \beta_{1c}$, $j = 1, \ldots, M$,

je kombinací předchozích testů

$$\text{Testační statistika } F_A = \frac{\frac{RSC_K - RSC_c}{2M - 2}}{\frac{RSC_c}{n - 2M}},$$

$$\text{kde } RSC_c = \sum_{j=1}^{M} RSC_j$$

Test: Je-li $F_A < F_{1-\alpha}(2M - 2, n - 2M)$, je H_0 přijata a všechny regresní přímky jsou totožné se společným odhadem úseku b_{2K} a směrnice b_{1K}.

Test shody dvou lineárních modelů

test shody parametrů $\boldsymbol{\beta}_1$ a $\boldsymbol{\beta}_2$ dvou lineárních modelů

$$\mathbf{y}_1 = \mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \qquad \ldots RSC_1$$

$$\mathbf{y}_2 = \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2 \qquad \ldots RSC_2$$

kde $\mathbf{X}_1$ rozměru $(n_1 \times m)$, $\mathbf{y}_1$ rozměru $(n_1 \times 1)$,
$\mathbf{X}_2$ rozměru $(n_2 \times m)$, $\mathbf{y}_2$ rozměru $(n_2 \times 1)$.

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \qquad \ldots RSC$$

Chowův test shody dvou lineárních modelů

H_0: $\beta_1 = \beta_2$ *vs.* H_A: $\beta_1 \neq \beta_2$

Testační kritérium

$$F_c = \frac{(RSC - RSC_1 - RSC_2)\,(n - 2m)}{(RSC_1 + RSC_2)\,(m)}$$

kde $n = n_1 + n_2$

Testování:

(a) Při homoskedasticitě $\sigma_1^2 = \sigma_2^2$: F_c *vs.* $F(m, n - 2m)$,

(b) Při heteroskedasticitě $\sigma_1^2 \neq \sigma_2^2$: F_c *vs.* $F(m, r)$,

$$\text{kde } r = \frac{[(n_1 - m)\,\sigma_1^2 + (n_2 - m)\,\sigma_2^2]^2}{(n_1 - m)\,\sigma_1^4 + (n_2 - m)\,\sigma_2^4}$$

Příklad 6.14 *Porovnání výsledků měření ze dvou laboratoří*

Stanovení volné enthalpie $-\Delta G_i$ par oxidu boritého v závislosti na teplotě T [Kelvin] bylo provedeno paralelně ve dvou laboratořích. Je možné hodnoty naměřené v obou laboratořích považovat za shodné ?

Data: n = 6; m = 2

T [K]	$-\Delta G$ [kcal mol^{-1}] Laboratoř A	Laboratoř B
1409	34.9	34.9
1441	34.6	33.8
1457	31.9	33.4
1492	33.1	32.4
1569	30.1	30.3
1610	29.3	29.1

Řešení: Lineární regresní model pro obě skupiny dat

$$E(-\Delta G/T) = \beta_{1,A}\, T + \beta_{2,A}$$

$$E(-\Delta G/T) = \beta_{1,B}\, T + \beta_{2,B}$$

Chowův test: H_0: $\boldsymbol{\beta}_A = \boldsymbol{\beta}_B$ proti H_A: $\boldsymbol{\beta}_A \neq \boldsymbol{\beta}_B$,
kde $\boldsymbol{\beta}_A = (\beta_{1,A}, \beta_{2,A})^T$ a $\boldsymbol{\beta}_B = (\beta_{1,B}, \beta_{2,B})^T$.

Laboratoř A: $b_{1,A} = -2.768 \cdot 10^{-2}$ $(\pm\, 5.25 \cdot 10^{-3})$
$b_{2,A} = 73.73$ $(\pm\, 7.865)$, $\hat{\sigma} = 0.916$
$RSC_A = 3.358$

Laboratoř B: $b_{1,B} = -2.776 \cdot 10^{-2}$ $(\pm\, 1.57 \cdot 10^{-4})$
$b_{2,B} = 73.82$ $(\pm\, 0.235)$, $\hat{\sigma} = 0.0274$
$RSC_B = 2.992 \cdot 10^{-3}$.

Laboratoř A + B: (označení C = A + B)
$b_{1,C} = -2.772 \cdot 10^{-2}$ $(\pm\, 2.35 \cdot 10^{-3})$
$b_{2,C} = 73.77$ $(\pm\, 3.521)$, $\hat{\sigma} = 0.58$
$RSC_C = 3.364$

Dosazením za $RSC = RSC_C$, $RSC_1 = RSC_A$ a $RSC_2 = RSC_B$ do testačního kritéria Chowova testu

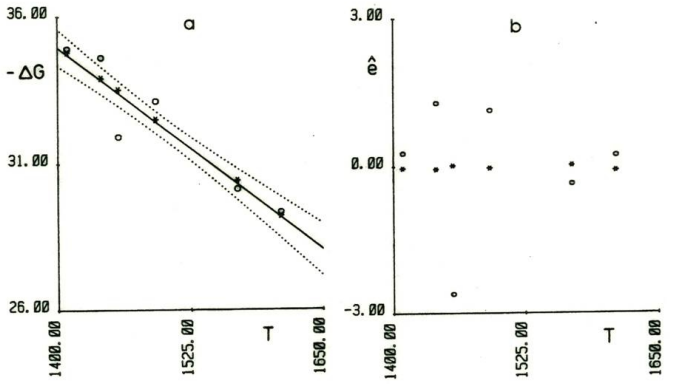
$$F_c = \frac{(RSC - RSC_1 - RSC_2)\,(n - 2m)}{(RSC_1 + RSC_2)\,(m)}$$

bude

$$F_c = \frac{(3.364 - 3.358 - 0.002992)\,(12 - 4)}{(3.358 + 0.002992)\,(2)} = 0.0036$$

Data A se od B liší v rozptylu, budou stupně volnosti dle

$$r = \frac{[4 \times 0.916^2 + 4 \times 0.0274^2]^2}{4 \times 0.916^4 + 4 \times 0.0274^4} = 4.007 \approx 4$$



Test: Jelikož $F_{0.95}(2, 4) = 6.94 > F_c$, je H_0: $\beta_A = \beta_B$ přijata.

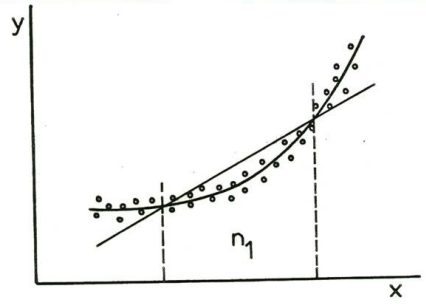
Závěr: Chowovým testem je prokázáno, že výsledky v obou laboratořích A a B jsou shodné.

Testy vhodnosti lineárního modelu

1. Test správnosti lineárního modelu

$f(x, \beta) = X\beta$ dle Uttsové:

H_0: lineární model *vs.* H_A: nelineární model



RSC_1 regresí s využitím n_1 bodů,
RSC regresí s využitím všech n bodů.

Testační kritérium

$$F_U = \frac{(RSC - RSC_1)(n_1 - m)}{RSC_1 (n - n_1)}$$

Uttsová: volit $n_1 \approx n/2$ a body co nejblíže k těžišti

Test: Je-li $F_U < F_{1-\alpha}(n - n_1, n_1 - m)$, je H_0 přijata.

2. Test kvadratického členu

H_0: $\beta_2 = 0$ (test významnosti β_2)

$$E(y/x) = \beta_1 x + \beta_2 x^2 + \beta_3$$

RSC_K pro kvadratický model
RSC_L pro lineární model

Testační kritérium linearity

$$F_L = \frac{(RSC_L - RSC_K)(n - 3)}{RSC_K (1)}$$

Test: Je-li $F_L < F_{1-\alpha}(1, n - 3)$, je H_0 přijata.

3. Linearita testem všech charakteristik

a) Střední kvadratická chyba predikce

$$MEP = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T b_{(i)})^2$$

kde $b_{(i)}$ je odhad, určený ze všech bodů kromě i-tého
Platí vztah

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

pro velké n jsou prvky $H_{ii} \approx 0$ a MEP = RSC/n.

b) Predikovaný koeficient determinace

$$\hat{R}_P^2 = 1 - \frac{n\,MEP}{\sum_{i=1}^{n} y_i^2 - n\bar{y}}$$

c) Akaikovo informační kritérium

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

nejvhodnější model má AIC minimální

HODNOCENÍ KVALITY 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

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

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)

Kritika dat

Vyšetření vlivných bodů:

(1) Odlehlé body

(2) Extrémy

Metoda:

Grafy vlivných bodů

Vybrané diagnostiky vlivných bodů

1. KRITIKA DAT

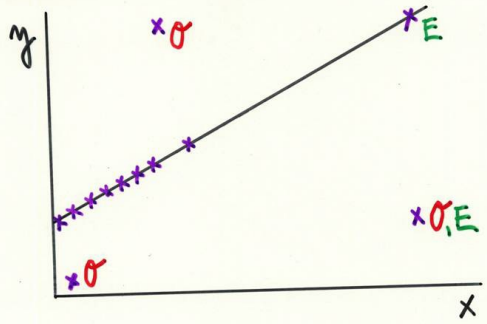
Výskyt *vlivných bodů* (VB):
- zdrojem řady problémů,
- zkreslení odhadů a růst rozptylů odhadů parametrů.

Dělení *vlivných bodů* dle charakteru:

1. ***Hrubé chyby***: vybočující pozorování, ~~extrémy~~.
2. ***Body s vysokým vlivem*** (tzv. golden points): rozšiřují predikční schopnosti modelu.
3. ***Zdánlivě vlivné body***: důsledek nesprávného regresního modelu.

Dělení *vlivných bodů* dle výskytu:

1. ***vybočující pozorování*** (outliers, O): na y se liší,
2. ***extrémy*** (high leverage points, E): liší se na ose x



Statistická analýza reziduí

1. Klasická rezidua: $\hat{e}_i = y_i - \mathbf{x}_i\,\mathbf{b}$,

Nesprávné představy o reziduích:
1. Rozdělení reziduí je stejné jako rozdělení chyb,
2. Vlastnosti reziduí jsou shodné s vlastnostmi chyb,
2. Čím je reziduum $\hat{e}_i$ větší, tím je bod vlivnější, a tím spíše by se měl z dat vyloučit.

$$\hat{\mathbf{e}} = \mathbf{P}\,\mathbf{y} = \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{P}\,\boldsymbol{\varepsilon} = (\mathbf{E} - \mathbf{H})\,\boldsymbol{\varepsilon}$$

Rozepsáním

$$\hat{e}_i = (1 - H_{ii})\,y_i - \sum_{j \neq i}^{n} H_{ij}\,y_j = (1 - H_{ii})\,\varepsilon_i - \sum_{j=i}^{n} H_{ij}\,\varepsilon_j$$

slovy: **každé $\hat{e}_i$ je lineární kombinací všech chyb ε_i**

Rozdělení reziduí je závislé:
1. Na rozdělení chyb,
2. Na prvcích projekční matice **H**,
3. Na velikosti výběru n.

Vlastnosti klasických reziduí:

1. Rozptyl reziduí

$$D(\hat{e}_i) = (1 - H_{ii})\,\sigma^2$$

je ***nekonstantní***, i když rozptyl chyb σ^2 je konstantní.

2. Rezidua jsou korelovaná: existuje ***párový korelační koeficient r_{ij}*** mezi dvěma rezidui e_i a e_j

$$r_{ij} = \frac{-H_{ij}}{\sqrt{(1 - H_{ii})(1 - H_{jj})}}$$

i když chyby ε_i a ε_j jsou nezávislé.

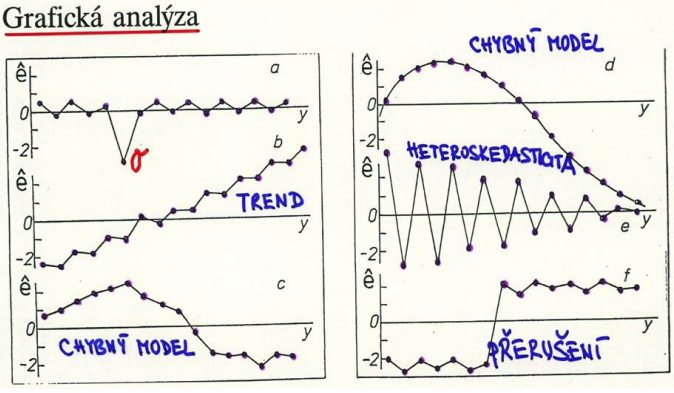
3. Rezidua *neindikují* silně extrémní hodnoty.

4. Rezidua jsou normálnější než chyby: (effekt supernormality)

5. U malých výběrů nemusí správně indikovat model.

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:** $R \approx 0.5\ \%$.

2. Normovaná rezidua $\hat{\sigma}_N$

$$\hat{e}_{Ni} = \hat{e}_i / \hat{\sigma}$$

normálně rozdělené veličiny $\hat{e}_{Ni} \sim N(0, 1)$.
Diagnostika: pravidlo 3σ, tj. rezidua větší než $\pm 3\hat{\sigma}$ indikují vybočující.

3. Standardizovaná rezidua $\hat{\sigma}_S$

$$\hat{e}_{Si} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - H_{ii}}}$$

Mají konstantní rozptyl. Maximální hodnota $\hat{e}_{Si}$ je ohraničena velikostí $\sqrt{n - m}$.
Diagnostika: k indikaci heteroskedasticity.

4. Jackknife rezidua $\hat{\sigma}_J$ ("plně studentizovaná")

užijeme místo $\hat{\sigma}$ odhadu směrodatné odchylky $\hat{\sigma}_{(-i)}$,

$$\hat{e}_{Ji} = \hat{e}_{Si}\sqrt{\frac{n - m - 1}{n - m - \hat{e}_{Si}^2}} = \sqrt{n - m}\ \mathrm{cotg}\,\Theta_i$$

mají Student. rozdělení s (n - m - 1) stupni volnosti.
Diagnostika: k identifikaci vybočujících bodů (outliers).

5. Predikovaná rezidua $\hat{\sigma}_P$

$$\hat{e}_{Pi} = y_i - x_i\,\mathbf{b}_{(i)} = \frac{\hat{e}_i}{1 - H_{ii}}$$

kde $\mathbf{b}_{(i)}$ jsou MNČ odhady ze všech bodů kromě i-tého.
Diagnostika: indikace vybočujících hodnot (outliers).

6. Rekurzívní rezidua $\hat{\sigma}_R$

Dopředná rekurzívní rezidua jsou definována vztahy

$$\hat{e}_{Ri} = 0, \qquad i = 1, \ldots, m$$

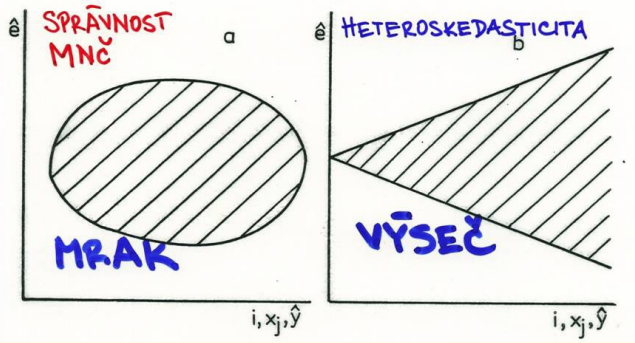
$$\hat{e}_{Ri} = \frac{y_i - \mathbf{x}_i\,\mathbf{b}_{i-1}}{\sqrt{1 + \mathbf{x}_i(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1})^{-1}\mathbf{x}_i^T}} \qquad i = m + 1, \ldots, n$$

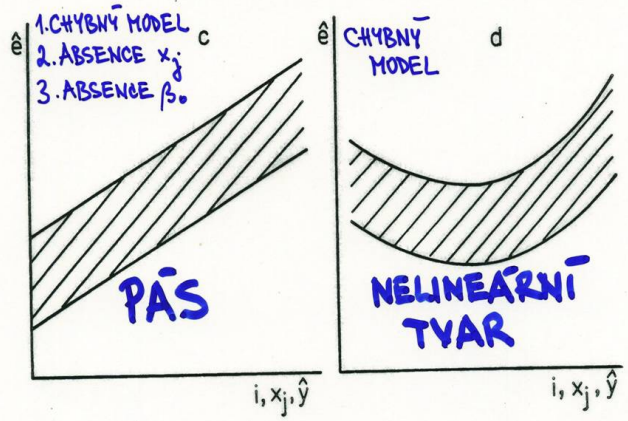
kde $\mathbf{b}_{i-1}$ jsou odhady získané z prvních (i - 1) bodů.

Diagnostika: umožňují identifikovat autokorelaci čili nestabilitu modelu, např. v čase.

Obrazce v diagnostických grafech

Typ I: Graf reziduí $\hat{e}_i$ proti indexu i.
Typ II: Graf reziduí $\hat{e}_i$ proti proměnné x_j.
Typ III: Graf reziduí $\hat{e}_i$ proti predikci $\hat{y}_i$.



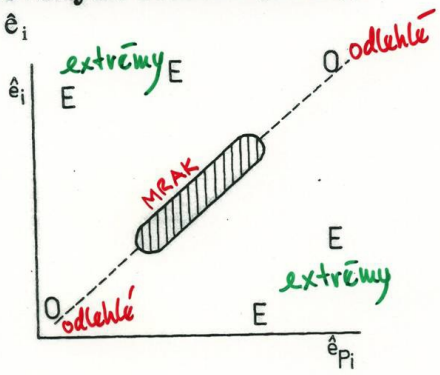


(a) tvar mraku, (b) tvar výseče, (c) tvar pásu a (d) nelineární tvar

Grafy identifikace vlivných bodů

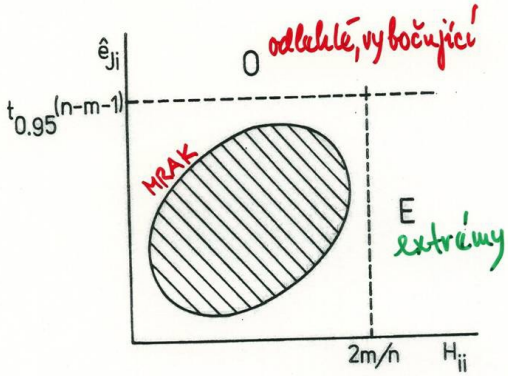
1. Graf predikovaných reziduí (GPR),

osa x: $\hat{e}_{Pi}$, osa y: $\hat{e}_i$



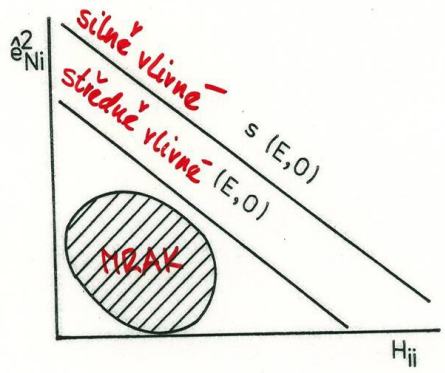
2. Williamsův graf (WG),

osa x: prvky H_{ii}, osa y: $\hat{e}_{Ji}$

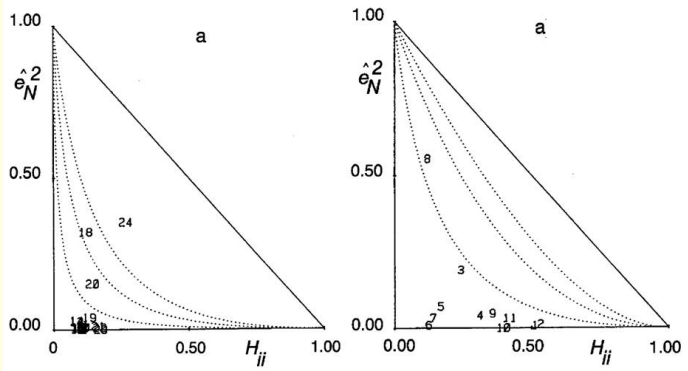


3. Pregibonův graf (PG),

osa x: prvky H_{ii}, osa y: $\hat{e}^2_{Ni}$

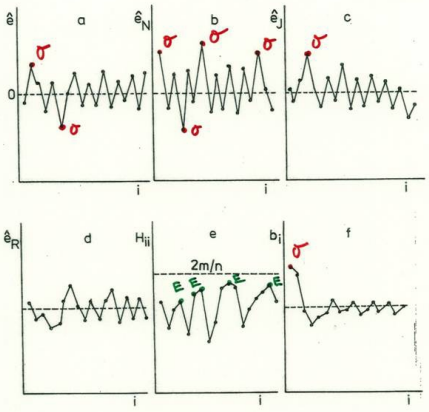


4. L-R graf



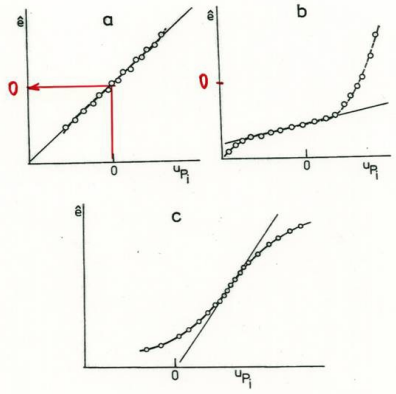
5. Indexové grafy (IG),

osa x: index i,
osa y: rezidua $\hat{e}_i$, $\hat{e}_{Si}$, $\hat{e}_{Ni}$, $\hat{e}_{Pi}$, $\hat{e}_{Ji}$, $\hat{e}_{Ri}$, nebo prvky H_{ii} či H^*_{ii}, a nebo odhady b_i



6. Rankitové grafy (Q-Q),

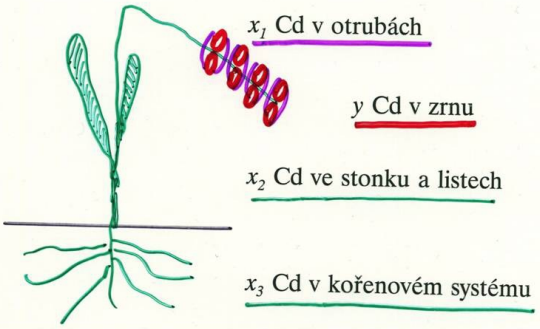
osa x: u_{P_i} pro $P_i = i / (n + 1)$,
osa y: $\hat{e}_{(i)}$, $\hat{e}_{S(i)}$, $\hat{e}_{N(i)}$, $\hat{e}_{P(i)}$, $\hat{e}_{J(i)}$, $\hat{e}_{R(i)}$



(a) přibližně normální rozdělení, (b) rozdělení s dlouhými konci, a (c) rozdělení s krátkými konci

Úloha M619. ***Vliv tří parametrů na obsah kadmia v potravinářské pšenici***

Obsah kadmia v zrnu y [mg/l] v závislosti na obsahu kadmia v otrubách x_1 [mg/l], ve stonku s listy x_2 [mg/l] a v kořenovém systému x_3 [mg/l]. Vyšetřete regresní triplet (data, model, metoda) a nalezněte linearní regresní model.



Výstavba lineárního regresního modelu:

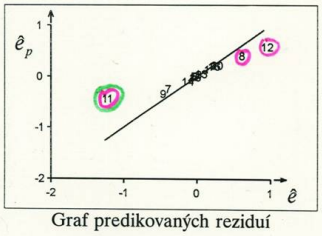
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Kritika dat

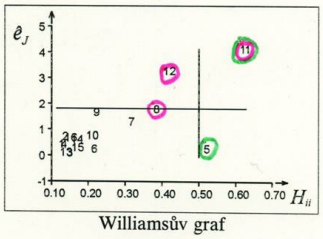
Kritika modelu

Kritika metody

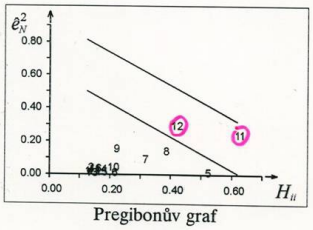
Grafy vlivných bodů (odlehlých O a extrémů E)



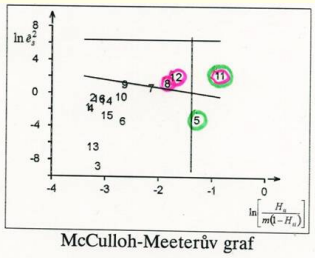
Graf predikovaných reziduí



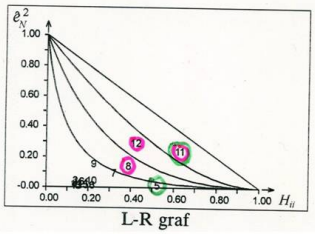
Williamsův graf



Pregibonův graf

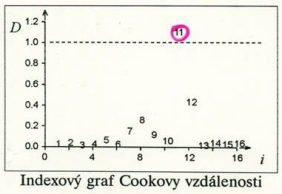


McCulloh-Meeterův graf

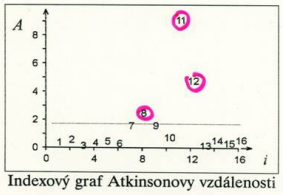


L-R graf

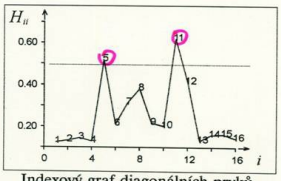
Vybrané regresní diagnostiky vlivných bodů



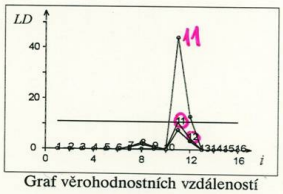
Indexový graf Cookovy vzdálenosti



Indexový graf Atkinsonovy vzdálenosti

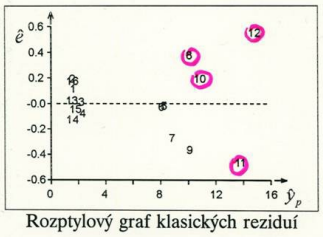


Indexový graf diagonálních prvků **H** projekční matice

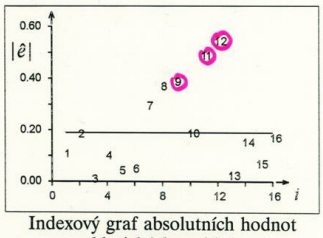


Graf věrohodnostních vzdáleností

Indexové a rozptylové grafy

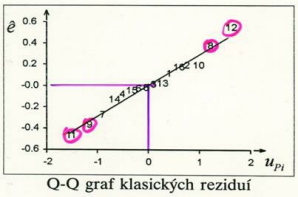


Rozptylový graf klasických reziduí

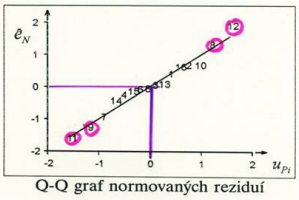


Indexový graf absolutních hodnot klasických reziduí

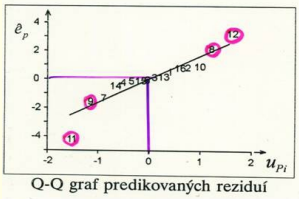
Rankitové Q-Q grafy (normalita reziduí)



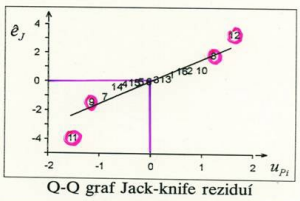
Q-Q graf klasických reziduí



Q-Q graf normovaných reziduí



Q-Q graf predikovaných reziduí



Q-Q graf Jack-knife reziduí

VÝSTUP

(1) PŘEDBĚŽNÁ STATISTICKÁ ANALÝZA:

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočtená hladina výz.
y	6.0125E+00	4.8734E+00	1.0000	-----
x1 OTRUBY	4.8937E+00	3.5692E+00	0.9837	0.000
x2 LISTY	5.7812E+00	4.5296E+00	0.9935	0.000
x3 KOŘEN	5.0812E+00	3.8782E+00	0.9948	0.000

Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných

			Spočtená hladina výz.
x1	versus x2 :	9.9344E-01	0.000
x1	versus x3 :	9.8693E-01	0.000
x2	versus x3 :	9.8847E-01	0.000

(2) INDIKACE MULTIKOLINEARITY:

Č [j]	Vlastní čísla korel. matice l[j]	Čísla podmíněnosti K[j]	Variance inflation factor VIF[j]	Vícenás.korel. koef pro X[j]
1	6.4568E-03	4.6141E+02	8.3272E+01	0.9940
2	1.4307E-02	2.0823E+02	9.4324E+01	0.9947
3	2.9792E+00	1.0000E+00	4.7508E+01	0.9894

Maximální číslo podmíněnosti K : 4.6141E+02

(K[j], K > 1000 indikuje silnou multikolinearitu)

(VIF[j] > 10 indikuje silnou multikolinearitu)

Kritika dat: původní data obsahují vlivné body 8, 11, 12

Kritika modelu $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$: technika MEP, AIC, s(e)) našla, že absolutní člen β_0 statisticky nevýznamný

	Odhad parametru	Směrodatná odchylka	Student t-test $t_{0,95}(16\text{-}4) = 2.18$	Parameter je	Hladina významnosti

A. Původní data a původní model vedou k závěru

$$y = -0.07(0.14) - 0.69(0.19)x_1 + 0.90(0.16)x_2 + 0.84(0.13)x_3$$

(LS: $t_{1-0.05/2}(16\text{-}4) = 2.179$, $D = 99.72\%$, $MEP = 0.21014$, $AIC = -36.18$, $s(e) = 0.290$)

β_0	-0.073	0.138	-0.5269	Nevýznamný	0.608
β_1	-0.685	0.192	-3.5746	Významný	0.004
β_2	0.896	0.161	5.5761	Významný	0.000
β_3	0.838	0.133	6.2879	Významný	0.000

B. Data bez 8, 11, 12 a původní model vedou k závěru

$$y = -0.02(0.10) - 0.80(0.50)x_1 + 0.92(0.34)x_2 + 0.91(0.15)x_3$$

(LS: $t_{1-0.05/2}(13\text{-}4) = 2.262$, $D = 99.83\%$, $MEP = 0.06118$, $AIC = -41.58$, $s(e) = 0.178$)

β_0	-0.017	0.103	-0.1637	Nevýznamný	0.874
β_1	-0.802	0.504	-1.5919	Nevýznamný	0.146
β_2	0.920	0.338	2.7193	Významný	0.024
β_3	0.906	0.152	5.9698	Významný	0.000

C. Data bez 11, 12 a opravený model vedou k závěru

$$y = -1.18(0.38)x_1 + 1.25(0.24)x_2 + 0.92(0.15)x_3$$

(LS: $t_{1-0.05/2}(14\text{-}3) = 2.201$, $D = 99.80\%$, $MEP = 0.06078$, $AIC = -43.47$, $s(e) = 0.193$)

β_0	0.000	---	---	---	---
β_1	-1.181	0.383	-3.0854	Významný	0.010
β_2	1.245	0.236	5.2751	Významný	0.000
β_3	0.917	0.151	6.091	Významný	0.000

D. Data bez 8, 11, 12 a opravený model vedou k závěru

$$y = -0.86(0.37)x_1 + 0.95(0.25)x_2 + 0.92(0.13)x_3$$

(LS: $t_{1-0.05/2}(13\text{-}3) = 2.228$, $D = 99.83\%$, $MEP = 0.05101$, $AIC = -43.55$, $s(e) = 0.170$)

β_0	0.000	---	---	---	---
β_1	-0.855	0.372	-2.3002	Významný	0.044
β_2	0.954	0.251	3.8038	Významný	0.003
β_3	0.916	0.132	6.9263	Významný	0.000

Model: $y = -0.86(0.37)x_1 + 0.95(0.25)x_2 + 0.92(0.13)x_3$

Numerická diagnostika vlivných bodů (odlehlých bodů a extrémů)

Odlehlé body jsou tištěny tučně

i	y_i	$y_{P,i}$	$s(y_i)$	e_i	$e_{M,i}$	$e_{S,i}$	$e_{L,i}$	$e_{P,i}$	H_{ii}	$Hm_{i,i}$	D_i	A_i	AP_i	DF_i	$LD(b)_i$	$LD(s^2)_i$	$LD(b, s^2)_i$
1	1.60	1.50	0.10	0.10	6.21	0.37	0.35	0.11	0.12	0.13	0.00	0.23	0.87	0.13	0.03	0.02	0.05
2	1.60	1.42	0.11	0.18	11.08	0.66	0.64	0.20	0.13	0.16	0.02	0.43	0.84	0.25	0.09	0.01	0.09
3	2.10	2.10	0.11	0.00	-0.14	-0.01	-0.01	0.00	0.14	0.14	0.00	0.01	0.86	0.00	0.00	0.03	0.03
4	2.10	2.19	0.10	-0.09	-4.41	-0.34	-0.33	-0.11	0.13	0.14	0.00	0.22	0.86	-0.13	0.02	0.02	0.05
5	8.10	8.14	0.21	-0.04	-0.44	-0.18	-0.17	-0.07	**0.52**	0.52	0.01	0.30	**0.48**	-0.17	0.04	0.03	0.07
6	7.90	7.94	0.13	-0.04	-0.52	-0.16	-0.15	-0.05	0.21	0.21	0.00	0.14	0.79	-0.08	0.01	0.03	0.04
7	8.40	8.69	0.16	-0.29	-3.42	-1.19	-1.22	-0.42	0.31	0.39	0.16	1.42	0.61	-0.82	0.84	0.03	0.95
8	10.30	9.94	0.18	0.36	3.52	1.59	1.71	0.59	0.38	0.51	**0.39**	**2.32**	**0.49**	**1.34**	1.94	0.25	2.69
9	9.60	9.98	0.14	-0.38	-3.96	-1.48	-1.57	-0.49	0.22	0.36	0.15	1.44	0.64	-0.83	0.80	0.16	1.10
10	10.80	10.62	0.13	0.18	1.66	0.69	0.68	0.22	0.20	0.23	0.03	0.59	0.77	0.34	0.16	0.00	0.16
11	13.10	13.58	0.23	-0.48	-3.67	**-2.68**	**-4.06**	**-1.26**	**0.62**	**0.85**	**2.91**	**8.94**	**0.15**	**-5.16**	**10.86**	**7.83**	**44.18**
12	15.10	14.57	0.19	0.54	3.55	**2.40**	**3.19**	**0.91**	0.41	**0.69**	**1.01**	**4.62**	**0.31**	**2.67**	**4.63**	**3.44**	**13.13**
13	1.30	1.29	0.11	0.01	0.71	0.03	0.03	0.01	0.13	0.13	0.00	0.02	0.87	0.01	0.00	0.03	0.03
14	1.20	1.34	0.12	-0.14	-12.01	-0.54	-0.53	-0.17	0.16	0.18	0.01	0.40	0.82	-0.23	0.07	0.01	0.08
15	1.50	1.56	0.12	-0.06	-3.98	-0.22	-0.22	-0.07	0.16	0.17	0.00	0.16	0.83	-0.10	0.01	0.03	0.04
16	1.50	1.34	0.11	0.16	10.74	0.60	0.58	0.19	0.14	0.17	0.02	0.41	0.83	0.24	0.08	0.01	10.09

Závisle proměnná
Predikce
Směrodatná odchylka predikce
Klasické reziduum
Relativní reziduum
Standardizované reziduum
Jack-knife reziduum
Predikované reziduum
Diagon. prvky **H** projekce
Modif. diagon. prvky **H** projekce
Cookova vzdálenost
Atkinsonova vzdálenost
Andrews-Pregibonova vzdálenost
Belseyho DF statistika
Věrohodnostní vzdálenost
Věrohodnostní vzdálenost
Věrohodnostní vzdálenost

Ostatní charakteristiky vlivných bodů

Vlivné body (VB) ovlivňují:

1. výrazně predikci $\hat{y}_{P,i}$
2. odhady parametrů b_j,
3. rozptyl parametrů σ_j,

Identifikace dle diagnostik:

(a) Sledování změn při **vypuštění** jednotlivých bodů.

(b) Model **zvětšeného rozptylu** (inflated variance): model působení vlivných bodů. Speciální struktura rozptylů chyb:

a) **i-tá chyba ε_i** má normální rozdělení $N(0, \sigma^2/w_i)$,

b) **ostatní chyby ε_j, $j \neq i$,** mají $N(0, \sigma^2)$,

Skalární míry (vzdálenosti):

- vznikají normováním vlivové funkce
- vyjadřují ***relativní vliv*** daného bodu na **všechny** parametry,

1. **Cookova vzdálenost D_i** : posun odhadů vynecháním i-tého bodu. Jde o eukleidovskou vzdálenost mezi vektory $\hat{\mathbf{y}}$ a $\hat{\mathbf{y}}_{(i)}$. Vznikne normováním výběrové vlivové funkce $\mathbf{SC}_i$

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{(i)})^T (\mathbf{X}^T \mathbf{X}) (\mathbf{b} - \mathbf{b}_{(i)})}{m\,\hat{\sigma}^2} = \frac{(\mathbf{y} - \mathbf{y}_{(i)})^T (\mathbf{y} - \mathbf{y}_{(i)})}{m\,\hat{\sigma}^2} =$$

$$= \frac{\hat{e}_{Si}}{m} \frac{H_{ii}}{1 - H_{ii}}$$

Diagnoza: pro $D_i > 1$ posun přesahuje 50%ní konfidenční oblast a daný bod je vlivný.
(D_i vyjadřuje vliv i-tého bodu pouze na odhady **b** nikoliv na odhad reziduálního rozptylu σ^2).

2. **Diagnostika DS_{ij}**: vyjadřuje relativní změnu odhadů parametrů po vynechání i-tého bodu

$$DS_{ij} = \frac{b_j - b_{(i)j}}{\hat{\sigma}_{(i)} \sqrt{V_{ii}}}$$

kde V_{ii} je diagonální prvek matice $\mathbf{X}^T \mathbf{X}$

Diagnoza: pro $DS > 2/\sqrt{n}$ je vliv i-tého bodu na odhad j-tého regresního parametru je významný.

5. **Andrewsova-Pregibonova diagnostika AP_i**: vyjadřuje vliv i-tého bodu na změnu objemu konfid. elipsoidu

$$AP_i = \frac{\det(\mathbf{X}_{(i)}^{*T} \mathbf{X}_{(i)}^{*})}{\det(\mathbf{X}^{*T} \mathbf{X}^{*})}$$

kde $\mathbf{X}^* = (\mathbf{X} \vdots \mathbf{y})$ je matice **X** rozšířená o vektor **y**.

Diagnoza: za výrazně vlivné se považují body, pro které platí $H_{ii}^* = (1 - AP_i) > 2\,(m + 1) / n$.

6. **Cookova-Weisbergova věrohodnostní vzdálenost LD_i** : unifikované vyjádření vlivných bodů

$$LD_i = 2\,(L(\hat{\Theta}) - L(\hat{\Theta}_{(i)}))$$

kde $L(\hat{\Theta})$ je max. log. věrohodn. funkce při užití všech bodů a $L(\hat{\Theta}_{(i)})$ bez i-tého bodu, kde Θ značí odhady **b** a $\hat{\sigma}^2$.

Diagnoza: Pro silně vlivné body je $LD_i > \chi^2_{1-\alpha}(m + 1)$.

(a) **Varianta $LD_i(\mathbf{b}_i)$:** vliv i-tého bodu pouze na odhady **b**

$$LD_i(\mathbf{b}) = n \ln\left[\frac{d_i\, H_{ii}}{1 - H_{ii}} + 1\right]$$

$LD_i(\mathbf{b}_i)$ je monotónní funkcí D_i a nepřináší nové poznatky oproti D_i.

(b) **Varianta $LD_i(\hat{\sigma}^2)$:** sledování citlivosti reziduálního rozptylu $\hat{\sigma}^2$ na přítomnost vlivných bodů

$$LD_i(\hat{\sigma}^2) = n \ln \frac{n}{n - 1} + n \ln(1 - d_i) + \frac{d_i\,(n - 1)}{1 - d_i} - 1$$

$LD_i(\hat{\sigma}^2)$ nezávisí na H_{ii} a není proto ovlivněna extrémními body.

(c) **Varianta $LD_i(\mathbf{b}_i, \hat{\sigma}^2)$:** sledování vlivu i-tého bodu na odhady parametrů i rozptylu

$$LD_i(\mathbf{b}_1, \hat{\sigma}^2) = n \ln\left(\frac{n}{n - 1}\right) + n \ln(1 - d_i) +$$

$$+ \frac{(n - 1)\, d_i}{(1 - d_i)(1 - H_{ii})} - 1$$

kde $d_i = \dfrac{\hat{e}_{si}^2}{n - m}$. Je výhodnější než A_i a DF_i.

7. **Jackknifeovo reziduum $\hat{e}_{Ji}$** : testování vlivu i-tého bodu na statistické charakteristiky

Diagnoza: Pro nepřítomnost vlivných bodů platí

$$\hat{e}_{Ji}^2 \le F_{1-\alpha/n}(1, n-m-1, 0.5)\ .$$

Dále silně vlivné body mají $\hat{e}_{Ji}^2$ větší než 10.

8. **Indexové grafy**: charakteristiky vlivných bodů v závislosti na indexu i monitorují vlivné body.

MULTIKOLINEARITA

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

Vektory matice X musí být skutečně navzájem nezávislé (jejich párové R musí být nulové nebo statisticky nevýznamné). Pokud tomu tak není, dochází k **multikolinearitě**, která způsobuje početní i statistické problémy.

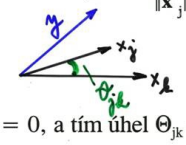
Multikolinearita

M. neznamená porušení předpokladů MNČ,
M. souvisí pouze s pozitivní definitností matice $\mathbf{X}^T\mathbf{X}$,

Dle úhlu Θ_{jk} mohou nastat dva krajní případy:

1. **Ortogonalita**, kdy je kosinus Θ_{jk} nulový, ($\cos\Theta_{jk} = 0$),

$$\cos\Theta_{jk} = \frac{\langle \mathbf{x}_j, \mathbf{x}_k\rangle}{\|\mathbf{x}_j\| \cdot \|\mathbf{x}_k\|} = 0$$



t. zn. že skalární součin $\langle \mathbf{x}_j, \mathbf{x}_k\rangle = 0$, a tím úhel Θ_{jk} mezi vektory $\mathbf{x}_j$ a $\mathbf{x}_k$ je 90°.

(Symbol $\|\mathbf{x}_j\| = \sqrt{\langle \mathbf{x}_j, \mathbf{x}_j\rangle}$ značí délku vektoru $\mathbf{x}_j$).

Jsou-li všechny sloupce matice $\mathbf{X}$ vzájemně ortogonální, je $\mathbf{X}^T\mathbf{X}$ diagonální matice.

2. **Kolinearita**, kdy je kosinus Θ_{jk} roven 1, ($\cos\Theta_{jk} = 1$), protože úhel Θ_{jk} mezi vektory $\mathbf{x}_j$ a $\mathbf{x}_k$ je nulový ($\Theta_{jk} = 0$).

Pak jsou oba vektory $\mathbf{x}_j$ a $\mathbf{x}_k$ rovnoběžné čili lineárně závislé

$c_j\,\mathbf{x}_j + c_k\,\mathbf{x}_k = \mathbf{0}$ a c_j, c_k jsou nenulové konstanty.

- Platí-li tato rovnice pro q dvojic sloupců matice $\mathbf{X}$, je její hodnost rovna rozdílu (m - q) a matice $\mathbf{X}^T\mathbf{X}$ je singulární.

Základní příčiny vzniku:

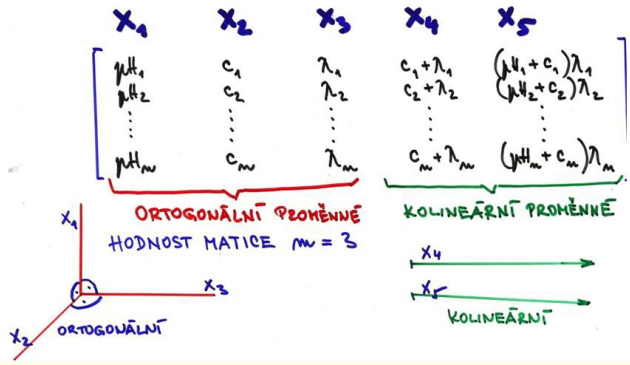
1. ***Přeurčený regresní model*** obsahuje nadměrný počet nezávislých proměnných, které vyjadřují stejné faktory.

2. ***Nevhodné rozmístění bodů***: hodnoty nezávisle proměnné kolísají jen v malém rozmezí a jsou proto kolineární s vektorem odpovídajícím absolutnímu členu.

3. ***Fyzikální omezení v modelu nebo v datech***, kdy vznikají vazby mezi nezávisle proměnnými přímo ve studovaném systému.

Test multikolinearity

Paradoxní situace: F-test je významný a všechny t-testy jsou nevýznamné, protože je silná multikolinearita mezi sloupci matice $\mathbf{X}$, čili existuje rovnoběžnost vektorů $\mathbf{x}_j$ a $\mathbf{x}_k$, $j \neq k$, sloupců matice $\mathbf{X}$.



Statistické obtíže:

1. ***Nestabilita odhadů*** je způsobená citlivostí odhadů na malé změny v datech. Odhady mívají často nesprávné znaménko, což znemožňuje jejich věcnou (fyzikální) interpretaci a jsou co do absolutních hodnot příliš veliké.

2. ***Velké rozptyly*** $D(b_j)$ jednotlivých odhadů způsobují, že t-testy indikují statistickou nevýznamnost β_j .

3. ***Silná korelovanost*** mezi prvky vektoru odhadů $\mathbf{b}$ způsobuje, že odhady b_j nelze interpretovat odděleně.

4. ***Koeficient determinace vysoký*** a regresní model může dobře popisovat experimentální data.

Metoda racionálních hodností

Matici $\mathbf{R}$ (symetrická) vyjádříme:

- pomocí vlastních čísel $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_m$
- a odpovídajících vlastních vektorů $\mathbf{P}_j$, $j = 1, \ldots, m$, ve tvaru

$$\mathbf{R} = \sum_{j=1}^{m} \lambda_j\, \mathbf{P}_j\, \mathbf{P}_j^T$$

a inverzní matici $\mathbf{R}^{-1}$ vztahem

$$\mathbf{R}^{-1} = \sum_{j=1}^{m} \lambda_j^{-1}\, \mathbf{P}_j\, \mathbf{P}_j^T$$

a ***vektor normovaných odhadů parametrů*** bude

$$b_N = \sum_{j=\omega}^{m} \left(\lambda_j^{-1}\, \mathbf{P}_j\, \mathbf{P}_j^T\right) \mathbf{r}$$

a ***kovarianční matice normovaných odhadů*** bude

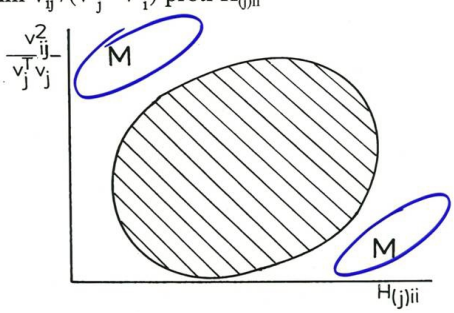
$$D(\mathbf{b}_N) = \hat{\sigma}_N^2 \sum_{j=\omega}^{m} \lambda_j^{-1}\, \mathbf{P}_j\, \mathbf{P}_j^T$$

V případě MNČ bude $\omega = 1$.

Platí důsledek: pokud budou vlastní čísla λ_j malá, budou odhady b_N i jejich rozptyly neúměrně vysoké.

Diagnostika:

1. **Grafická analýza:** vlivné body (multikolinearitu) odhalíme vynesením $v_{ij}^2/(\mathbf{v}_j^T\mathbf{v}_i)$ proti $H_{(j)ii}$



Diagnostika multikolinearity: M značí multikolinearitu

2. Numerická kritéria:

a) **Determinant matice R,** $\det(\mathbf{R}) = \prod_{j=1}^{m} \lambda_j$, kde λ_j jsou vlastní čísla matice **R**. Je-li determinant det(**R**) příliš malý, tj. menší než 10^{-3}, jde o silnou multikolinearitu.

b) **Číslo podmíněnosti K** $= \dfrac{\lambda_{max}}{\lambda_{min}}$, kde λ_{max}, λ_{min} jsou maximální a minimální vlastní číslo matice **R**. Je-li číslo podmíněnosti $K > 10^3$, jde o silnou multikolinearitu.

c) **VIF-faktor** (Variance Inflation Factor) je $VIF_j = \tilde{R}_{jj}$, kde $\tilde{R}_{jj}$ je j-tý diagonální prvek matice $\mathbf{R}^{-1}$. Platí vztah

$$VIF_j = \frac{1}{1 - \hat{R}^2_{x_j}}$$

. Je-li $VIF_j > 10$, jde o silnou multikolinearitu.

MULTIKOLINEARITA - testování

VIF – variance inflation factor – diagonální prvky inverzní matice ke korelační matici nezávisle proměnných (**diag(R^{-1})**)

	A	B	C	D	E	F
1		*X1*	*X2*	*X3*	*X4*	*X5*
2	X1	1	0.23	-0.15	0.07	0
3	X2	0.23	1	0.08	0.25	0.34
4	X3	-0.15	0.08	1	0.73	0.67
5	X4	0.07	0.25	0.73	1	0.98
6	X5	0	0.34	0.67	0.98	1
7						
8						
9		*X1*	*X2*	*X3*	*X4*	*X5*
10	X1	2.25	-1.28	1.51	-10.2	9.44
11	X2	-1.28	2.15	-0.88	9.05	-9.03
12	X3	1.51	-0.88	3.38	-11.3	9.15
13	X4	-10.2	9.05	-11.3	89.6	-83.5
14	X5	9.44	-9.03	9.15	-83.5	79.9

VIF > 10 ⇒ kritická multikolinearita

korelační matice R

=INVERZE(B2..F6)

Ctrl+Shift+Enter

inverzní matice R^{-1}

kriticky vysoké hodnoty VIF

Příklad 6.9 *Aproximace absorpčního spektra polynomem*
Popište závislost molárního absorpčního koeficientu ε na vlnové délce λ, $\varepsilon = f(\lambda)$ polynomem druhého stupně

$$E(\varepsilon/\lambda) = \beta_2 \lambda + \beta_3 \lambda^2 + \beta_1$$

Data: n = 15; m = 3

ε [dm³ mol⁻¹ cm⁻¹]	3	3.4	4.3	5	6	6.8	8.1	9.2	10.7	11.6
λ [nm]	460	470	480	490	500	510	520	530	540	550

ε [dm³ mol⁻¹ cm⁻¹]	12.9	13.6	14.6	15.3	15.5
λ [nm]	560	570	580	590	600

Řešení:

a) Pro $\beta_j = 0$: $F_R = 696 > F_{0.95}(2, 12) = 3.885$, je model jako celek statisticky významný.

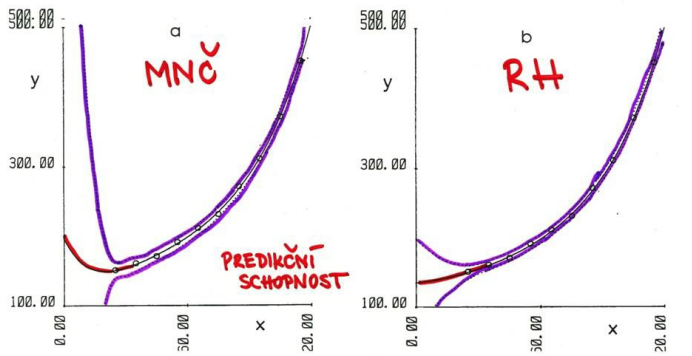
b) $t_{0.975}(12) = 2.179$ je vyšší než T_2 a T_3, a oba parametry β_2 a β_3 vycházejí statisticky nevýznamné.

c) $M_T = 0.989$ ukazuje na silnou multikolinearitu.

Tabulka 6.2 Odhady parametrů polynomu

j	Parametr	Odhad b_j	$\sqrt{D(b_j)}$	T_j
1	1	-43.93	19.38	-2.267
2	2	0.1018	0.0735	1.386
3	3	$-2.505 \cdot 10^{-6}$	$6.925 \cdot 10^{-5}$	-0.0361

Závěr: U polynomických regresních modelů ***nelze*** z výsledků Studentova t-testu usuzovat na významnost jednotlivých členů polynomu. Příčinou je silná multikolinearita.



Závěr:

1. RH poskytne odhady parametrů, které zajišťují průběh modelu odpovídající trendům dat a nemá nadbytečné extrémy či inflexy.
2. Při použití klasické MNČ by úloha byla řešitelná pouze při zavedení omezení na regresní parametry.

REGRESNÍ DIAGNOSTIKA – kvalita modelu

1) **Graf reziduí**
2) **Parciální regresní grafy**

Vyjadřuje závislost **mezi vysvětlovanou proměnnou** (vektorem *y*) a jednou **vysvětlující proměnnou** x_j při statisticky neměnném vlivu ostatních vysvětlujících proměnných, které tvoří matici $\mathbf{X}_{(j)}$ (kdy je vynechaná *j*-tá proměnná).

Jde zde o určitou grafickou obdobu parciálního korelačního koeficientu u korelačních modelů.

HODNOCENÍ 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 by pro statisticky významný odhad mělo platit pravidlo Sillénovy podmínky

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

HODNOCENÍ REGRESNÍHO MODELU

2. Kvalita dosažené těsnosti proložení

a) podle **reziduálního rozptylu.**

b) podle **regresního rabatu**, což je v procentech vyjádřený koeficient determinace (čím více se blíží 100 %, tím lepší je proložení modelem).

3. Vhodnost navrženého modelu

a) **Akaikovo informační kritérium (AIC)**: čím je AIC menší, tím je model vhodnější.

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

HODNOCENÍ REGRESNÍHO MODELU

4. Predikční schopnost modelu

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

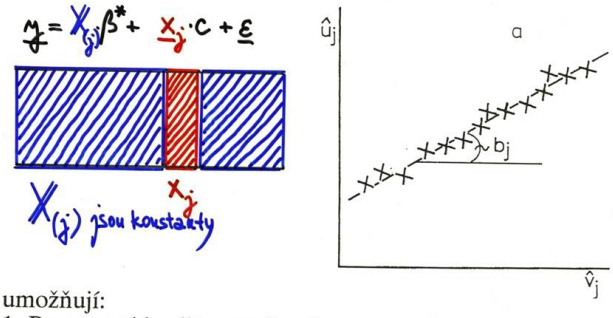
5. Kvalita experimentálních dat

a) na základě analýzy reziduí

b) na základě analýzy vlivných bodů (podle Jackknife reziduí, Cookovy vzdálenosti, diagonální prvky projekční matice a věrohodnostní vzdálenost).

Belseyho parciální regresní grafy

(partial regression leverage plots)



umožňují:

1. Posouzení kvality navrženého regresního modelu,
2. Indikují přítomnost vlivných bodů,
3. Nesplnění předpokladů klasické MNČ,
4. Vyjadřují závislost mezi **y** a zvolenou proměnnou x_j při statisticky neměnném vlivu ostatních $X_{(j)}$

Střední hodnota $E(\hat{u}_j)$ je

$$E(\hat{u}_j) = c\,E(\hat{v}_j)$$

a závislost $\hat{u}_j$ na $\hat{v}_j$ tvoří **parciální regresní graf** (u správného modelu jde o závislost s nulovým úsekem).

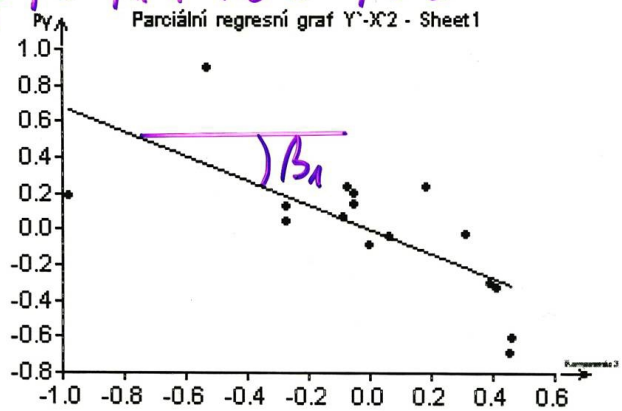
Odhad směrnice z MNČ bude

$$\hat{c} = \frac{\hat{u}_j^T \hat{u}_j}{\hat{v}_j^T \hat{v}_j} = \frac{x_j^T P_{(j)} y}{x_j^T P_{(j)} x_j}$$

Platí $\hat{e} = \hat{u}_j - \hat{v}_j \hat{c}$ a ukazuje na vztah mezi $\hat{e}$ s parciálními rezidui $\hat{u}_j$ a $\hat{v}_j$.

Kritika modelu: parciální regresní graf

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$



Rovnici $\hat{e} = \hat{u}_j - \hat{v}_j . b_j$ přepíšeme do tvaru

$$\hat{u}_j = \hat{e} + b_j (E - H_{(j)}) x_j$$

a vyjádříme jako závislost veličiny

$$\hat{e} + b_j (E - H_{(j)}) x_j \quad \text{na} \quad (E - H_{(j)}) x_j$$

Parciální reziduální graf je analogií parc. regr. grafu při volbě $H_{(j)} = 0$.
Jedná se o závislost parciálních reziduí **s** na proměnné x_j.

$$s = \hat{e} + b\, x_j = y - \sum_{k \neq j}^{m} x_k b_k$$

Kritika metody

Ověření předpokladů metody nejmenších čtverců

Metoda:

- **Vyšetření heteroskedasticity**
- **Vyšetření autokorelace**
- **Vyšetření multikolinearity**
- **Vyšetření normality náhodných chyb**
- **Vyšetření omezení parametrů**
- **Vyšetření trendů reziduí**

REGRESNÍ DIAGNOSTIKA
testy vybraných předpokladů klasické MNČ

Multikolinearita: VIF diagnostika indikuje

Heteroskedasticita: testy heteroskedasticity (např. Cook-Weisberg)

Autokorelace reziduí: test významnosti autokorelačního koeficientu

Normalita reziduí: testy normality

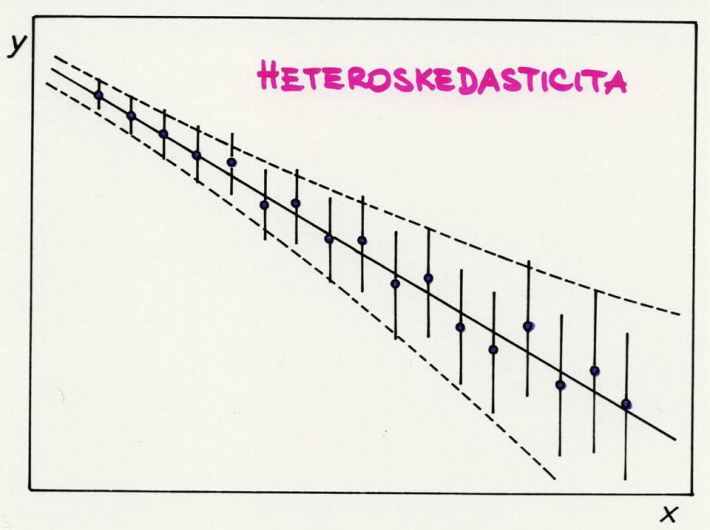
Ověření předpokladů MNČ

1. Hetero/homoskedasticita
(nekonstantnost rozptylu)

Rozptyl veličiny y_i v itém bodě je popsán

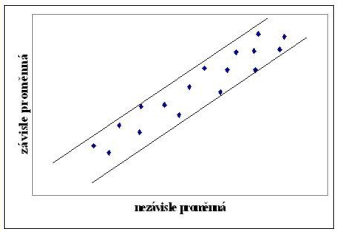
$$\sigma_i^2 = \sigma^2 \exp(\lambda\, x_i \boldsymbol{\beta}),$$

kde x_i je i-tý řádek matice **X**.

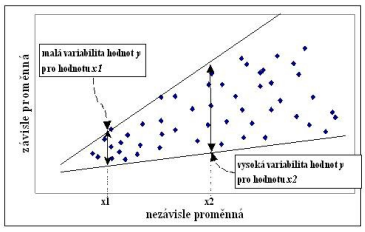


HOMOSKEDASTICITA *vs.* ETEROSKEDASTICITA
Grafické vysvětlení principu

Homoskedasticita znamená, že hodnoty závisle proměnné y mají pro všechny hodnoty nezávisle proměnné x **konstantní rozptyl** (variabilitu, proměnlivost).



Homoskedasticita



Heteroskedasticita

HOMOSKEDASTICITA
Testování

Test trendu reziduí

$$D = \sum_{i=1}^{n} \left[R\left(\left|\hat{e}_i\right|\right) - i \right]^2$$

$$\rho_s = 1 - \frac{6}{n^3 - n} \cdot D$$

Testujeme významnost Spearmanova korelačního koeficientu ρ_s

$$t_R = \frac{\rho_s \cdot \sqrt{n-2}}{\sqrt{1-\rho_s^2}}$$

HOMOSKEDASTICITA
Testování

Vycházíme z předpokladu, že rozptyl naměřené hodnoty y_i je určitou funkcí proměnné $x_i\,\beta$ (např. exponenciální funkcí).

Cookův - Weisbergův test homoskedasticity

$$S_f = \frac{\left[\sum_{i=1}^{n} (y'_i - \bar{y}')^2 e_i^2\right]^2}{2 \cdot \sigma^4 \sum_{i=1}^{n} (y'_i - \bar{y}')^2}$$

Test: Pokud v datech není heteroskedasticita potom platí, že $S_f < \chi^2(1)$.

Test: ověření H_0: $\lambda = 0$ (homoskedasticita).
Cook-Weisbergovo testační kritérium

$$S_f = \frac{\left[\sum_{i=1}^{n} (\hat{y}_i - \hat{y}_p)\, \hat{e}_i^2\right]^2}{2\,\hat{\sigma}^4 \sum_{i=1}^{n} (\hat{y}_i - \hat{y}_P)^2}$$

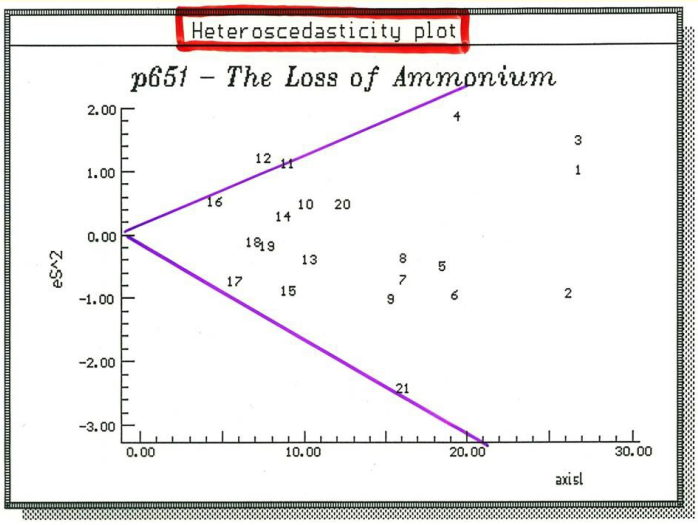
kde

$$\hat{y}_P = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$$

Testování:
1. Je-li $S_f < \chi^2(1)$, H_0 (homoskedasticita) je přijata.
2. Pro homoskedasticitu tvoří diagnostický graf

$$\hat{e}_{Si}^2 \text{ na } (1 - H_{ii})\ \hat{y}_i$$

náhodný mrak bodů. Pro heteroskedasticitu vznikne typický klínový obrazec.



2. Autokorelace

Data časových řad mají chyby ε_i vzájemně korelované.

Nejčastější je případ autokorelace prvního řádu

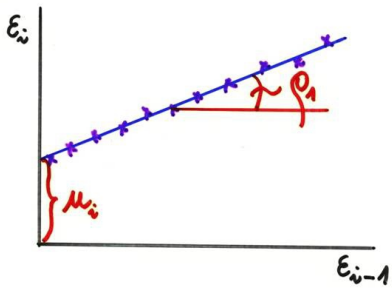
$$\varepsilon_i = \rho_1\, \varepsilon_{i-1} + u_i$$

kde $u_i \sim N(0, \sigma^2)$.

a) Pro $\rho_1 = 1$ případ kumulativních chyb, který se v chemii vyskytuje často.

b) Pro $\rho_1 \leq 1$ jde o autokorelační koeficient 1. řádu.

Test:
1. Grafická indikace autokorelace:



2. Waldův test pro ρ_1: H_0: $\rho_1 = 0$ *vs.* H_A: $\rho_1 \neq 0$

Je-li Waldovo kritérium $W_a = \dfrac{n\,\hat{\rho}_1^2}{1 - \hat{\rho}_1^2} < \chi^2(1)$, H_0 je přijata.

Normalita chyb

1. **Rankitový (Q-Q) graf:** $\hat{e}_{(i)}$ nebo $\hat{e}_{Ri}$ na u_{Pi} pro $P_i = i / (n + 1)$

2. **Test normality:** H_0: normalita *vs.* H_A: nenormalita
   Jarque-Berrova testační statistika

$$L(\hat{e}) = n \left[\frac{\hat{u}_3^2}{6\,\hat{u}_2^3} + \frac{\frac{\hat{u}_4}{\hat{u}_2^2} - 3}{24} \right] + n \left[\frac{3\,\hat{u}_1^2}{2\,\hat{u}_2} - \frac{\hat{u}_3\,\hat{u}_1}{\hat{u}_2^2} \right]$$

kde $\hat{u}_j$ je j-tý výběrový moment reziduí

$$\hat{u}_j = \frac{\sum_{i=1}^{n} \hat{e}_i^{\,j}}{n}$$

Test: je-li $L(\hat{e}) > \chi^2_{1-\alpha}(2) = 5.99$, je H_0 (normalita) zamítnuta.

Např. pro lineární modely s absolutním členem je $\hat{u}_1 = 0$ a $L(\hat{e})$ je

$$L(\hat{e}) = n \left(\frac{\hat{g}_1}{6} + \frac{(\hat{g}_2 - 3)^2}{24} \right)$$

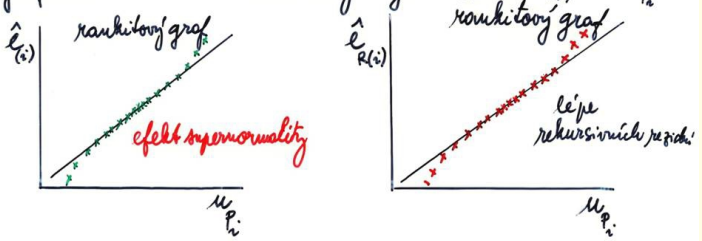
kde

$$\hat{g}_1 = \frac{\hat{u}_3^2}{\hat{u}_2^3} \qquad \hat{g}_2 = \frac{\hat{u}_4}{\hat{u}_2^2}$$

(test není vhodný pro malé výběry)

NORMALITA CHYB

grafická indikace normality chyb: $\hat{e}_{(i)}$ nebo $\hat{e}_{R(i)}$ vs. u_{P_i}



Znaménkový test vhodnosti modelu

nenáhodnost reziduí lze testovat znaménkovým testem.

Postup:

1. Určuje se počet sekvencí n_U, kde mají rezidua stejná znaménka.
   (např. pro rezidua -1, -1, 1, -1, 1, 2, 1 je počet sekvencí roven $\hat{n}_U = 4$)

2. Stanoví se počet reziduí kladných (n_+) a záporných (n_-).
3. Teoretický počet sekvencí n_t a jeho rozptyl D_t je

$$n_t = 1 + \frac{2\,n_+\,n_-}{n_+ + n_-} \approx 1 + \frac{n}{2}$$

$$D_t = \frac{2\,n_+\,n_-\,(2\,n_+\,n_- - n_+ - n_-)}{(n_+ + n_-)^2\,(n_+ + n_- - 1)} \approx \frac{n}{4}$$

Pro $n_+ > 10$ a $n_- > 10$ lze využít náhodné veličiny

$$U = \frac{n_u - n_t + C}{\sqrt{D_t}}$$

s normovaným normálním rozdělením N(0, 1).

Testování: Je-li $n_U < (n_t - C\,1.96\,\sqrt{D_t})$, pak je v reziduích trend a model je nesprávně navržen.
(Konstanta C = 0.5 je korekcí na spojitost).

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), $\beta = (\beta_1, ..., \beta_m)^T$.

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

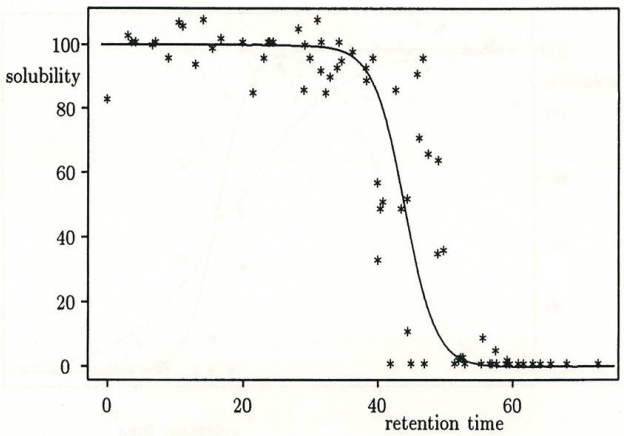


FIGURE 3.7. Peptides example: observed and adjusted solubilities of 75 peptides versus their RP-HPLC retention time

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

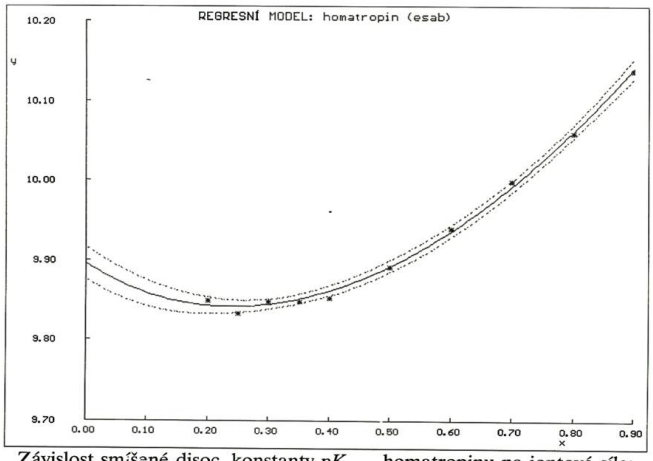
Model: rozšířený Debye-Hückelův zákon

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

Proměnné:

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

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



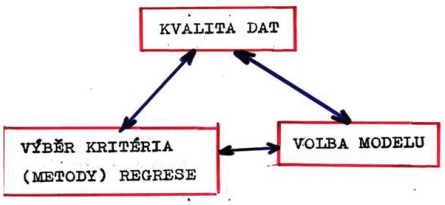
Závislost smíšené disoc. konstanty $pK_{a,smíš}$ homatropinu na iontové síle:

Nulté přiblížení: $pK_a{}^T = 1$, $å = 1$ Å, $C = 1$

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

POSTUP A DIAGNOSTIKA REGRESE:

Regresní triplet:



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

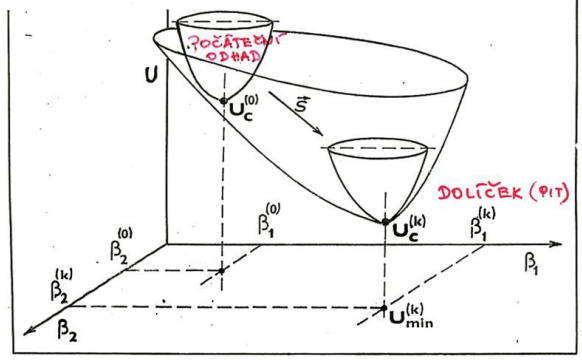
kde $y_{vyp,i} = f(x_i;\ b_1, \ldots, b_m)$

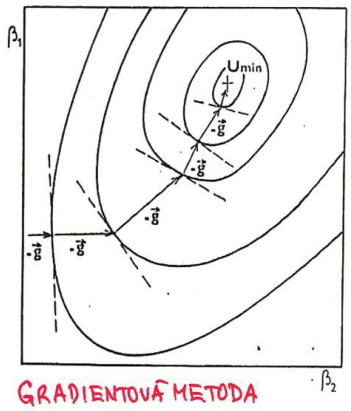
GEOMETRICKÉ ZNÁZORNĚNÍ KRITÉRIA U

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

$y_{vyp,i} = f(x_i;\ \beta_1, \ldots, \beta_m;\ \beta_{m+1}, \ldots, \beta_k)$

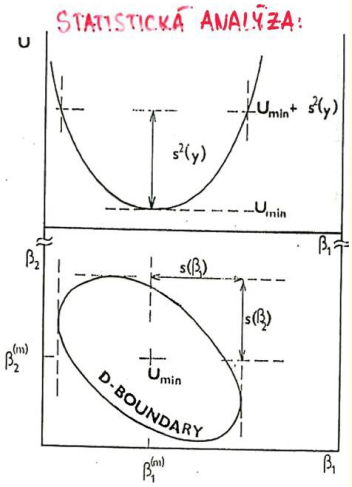
OPTIMALIZAČNÍ PROBLÉM V (m+1) ROZMĚRNÉM PROSTORU





GRADIENTOVÁ METODA HLEDÁNÍ MINIMA

STATISTICKÁ ANALÝZA:



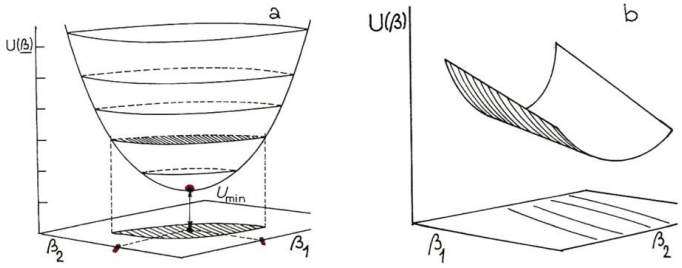
Geometrie nelineární regrese

Kritérium regrese U(β) vektorovým zápisem

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

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

Pro lineární regresní modely:

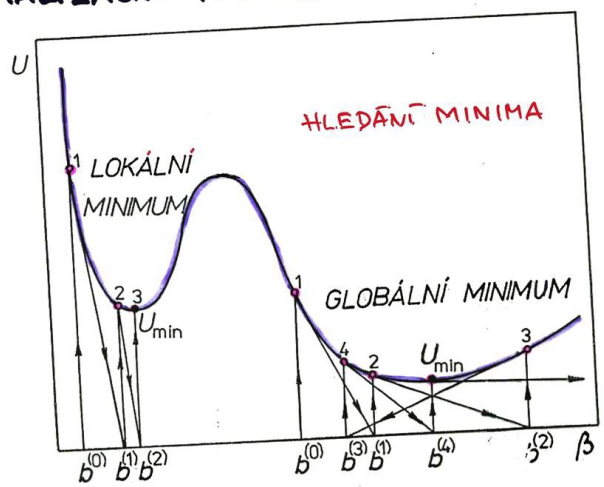
1. Útvar kritéria U(β) v prostoru odhadů je ***eliptický hyperparaboloid*** se středem v bodě [**b**, U(**b**)], kde nabývá minimální hodnoty.

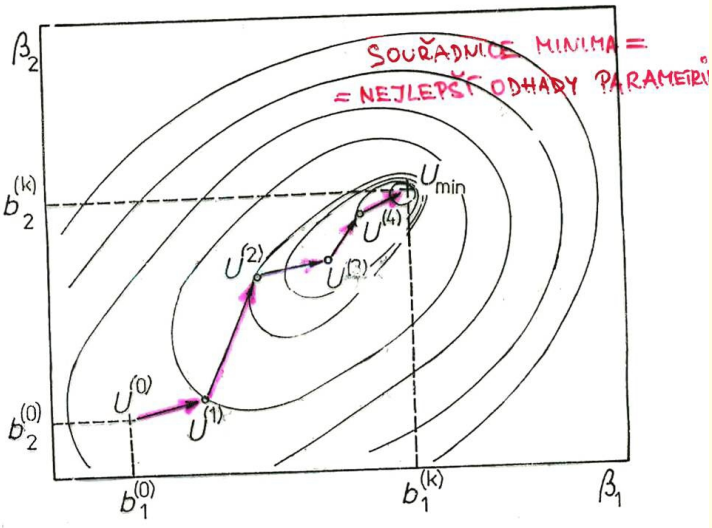
2. Účelová funkce U(β) je vzhledem k β kvadratickou formou β^T $(\mathbf{X}^T\,\mathbf{X})\,\beta$ a matice $\mathbf{X}^T\,\mathbf{X}$ je pozitivně definitní.

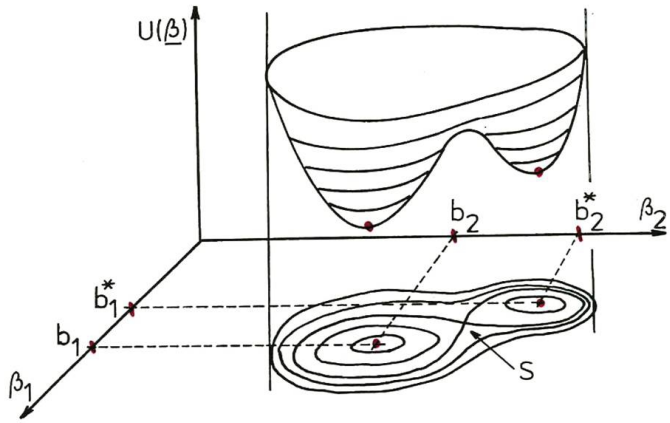
Pro nelineární regresní modely:

1. Složitější útvary vznikají u nelineárních modelů v závislosti na nelinearitě funkce f(**x**, β), počtu extrémů a sedlových bodů.

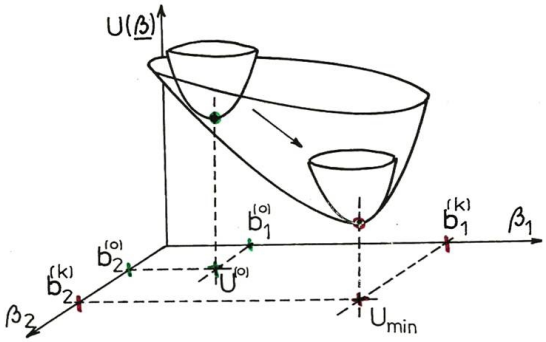
MINIMALIZAČNÍ PROCES:







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, β) 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.

Nederivační metody

Umožňují nalezení extrému G(β) vzhledem k β:

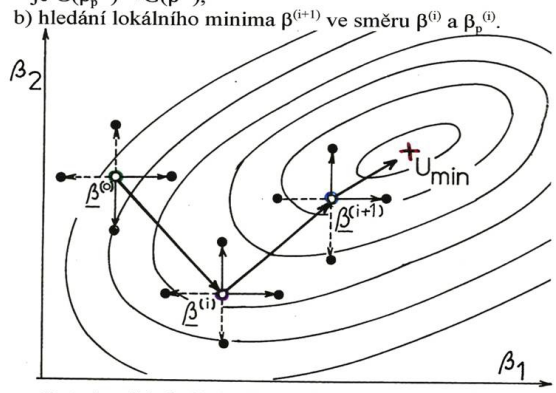
1. metody přímého hledání,
2. simplexové metody,
3. metody využívající náhodných čísel,
4. postupy speciálně vhodné pro MNČ.

1. Metody přímého hledání

Hookův-Jeevsův algoritmus:

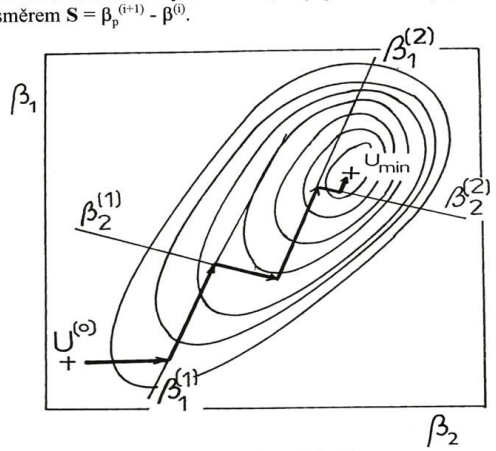
a) krokové posuny a nalezení zlepšeného odhadu $\beta_p^{(i)}$, pro který je $G(\beta_p^{(i)}) < G(\beta^{(i)})$,

b) hledání lokálního minima $\beta^{(i+1)}$ ve směru $\beta^{(i)}$ a $\beta_p^{(i)}$.



Postup koordinátního hledání (čárkovaně znázorněny neúspěšné směry)

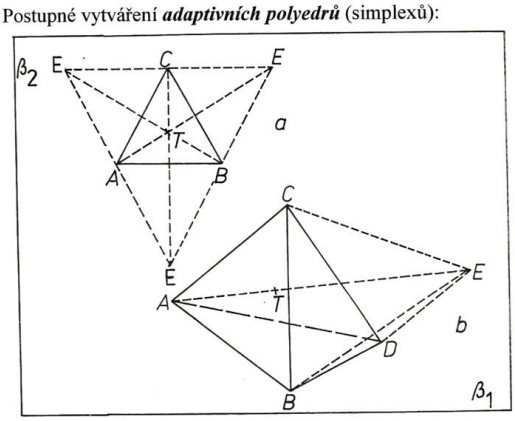
Rosenbrockova metoda: místo krokového hledání ve směru **s** rotaci souřadného systému tak, aby jedna osa splynula se směrem $\mathbf{S} = \beta_p^{(i+1)} - \beta^{(i)}$.



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

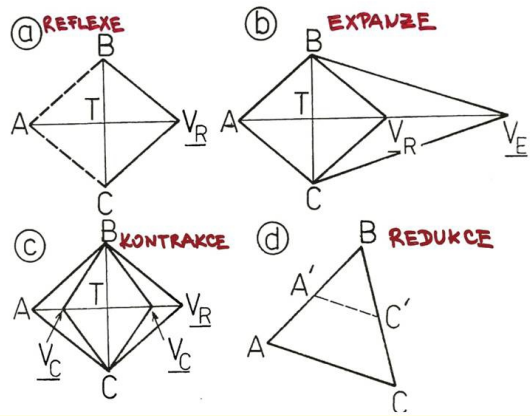
Z počátečního odhadu $\beta^{(0)}$ se postupně tvoří ***simplexy***.
Simplex je konvexní mnohostěn v m-rozměrném prostoru parametrů právě svými (m+1) vrcholy.

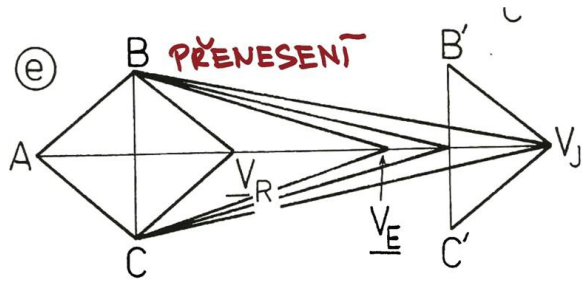
Postup minimalizace:

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

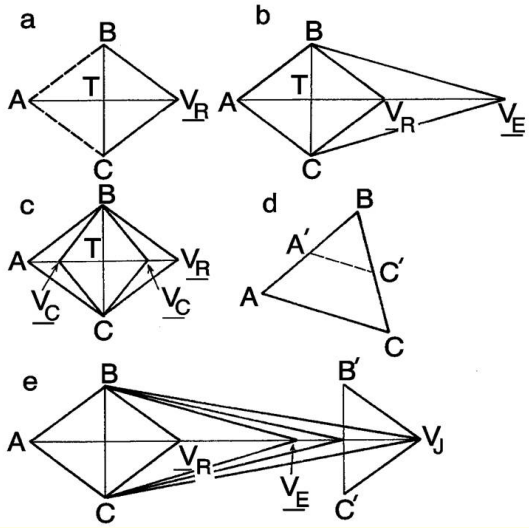
2. Iterativní postup k minimu

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



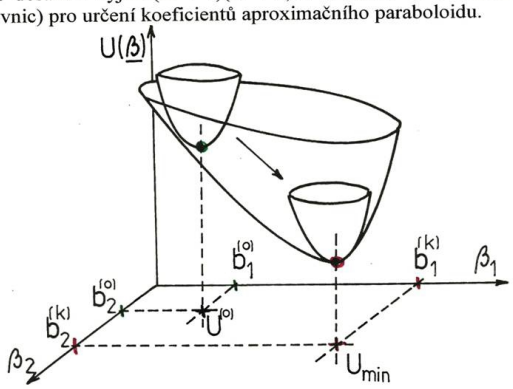


Základní operace simplexu:
a) reflexe, b) expanze, c) kontrakce, d) redukce a e) přenesení



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)

Těsnost proložení

Statistická analýza reziduí

Pro aditivní modely měření:

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

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

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

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

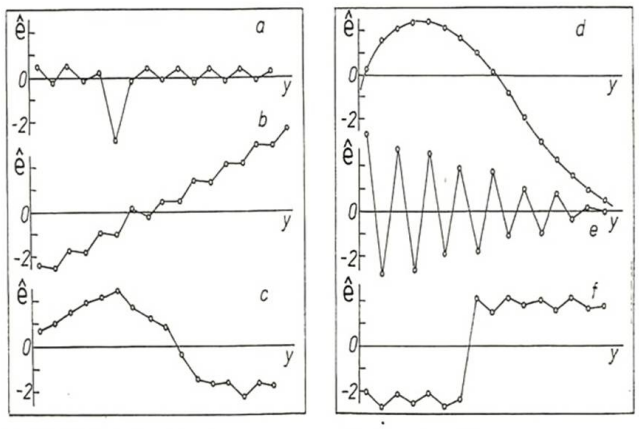
Každé reziduum je přibližně lineární kombinací všech chyb

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

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

1. Grafická analýza reziduí

a) odlehlé (extrémní) hodnoty v souboru reziduí,
b) trend v reziduích,
c) nedostatečné střídání znaménka u reziduí,
d) chybný model nebo vzájemnou závislost reziduí,
e) heteroskedasticitu (nekonstantnost rozptylu) závisle proměnné (měřené) veličiny **y**,
f) náhlou změnu podmínek při měření hodnot y.



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

2. Numerická analýza reziduí

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

3. Analýza vlivných bodů

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

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

kde P_{ii} jsou prvky projekční matice,

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

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

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

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

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

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

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

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

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

c) ***Věrohodnostní vzdálenost***

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

Pro případ MNČ bude věrohodnostní vzdálenost ve tvaru

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

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

Postup tvorby nelineárního regresního modelu

1. Návrh regresního modelu.

Obvykle se používá nějaká fyzikální nebo empirická závislost.

2. Odhadování parametrů.

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

3. Posouzení kvality odhadů.

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

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

4. Grafické posouzení vhodnosti modelu.

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

a) odlehlé hodnoty,
b) trend v reziduích,
c) nedostatečné střídání znaménka u reziduí,
d) heteroskedasticitu.

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

5. Základní statistické charakteristiky.

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

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

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

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

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***

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

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

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

6. Regresní diagnostika.

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

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

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

B. Analýza vlivných bodů:

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

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

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

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

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 $\boldsymbol{V} = (\boldsymbol{J}^T \boldsymbol{J})^{-1}$.

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

Vlivné body VB lze identifikovat jednokrokovou aproximací ***Jackknife 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 $\boldsymbol{h}_{(i)}$ 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.

8. Predikční schopnost modelu "cross-validation":

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

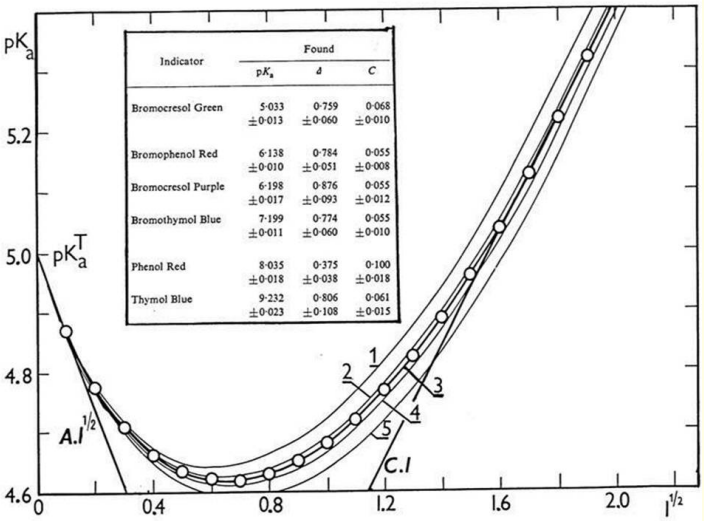
Na odhady parametrů jsou kladena omezení, vycházející z fyzikálního smyslu. Odhady musí ležet v jisté předpokládané oblasti (např. koncentrace v oblasti kladných čísel, molární absorpční koeficienty ε v oboru čísel 10 až 10^6, konstanty stability $\log \beta_{pqr}$ v oboru čísel 0 až 50 atd.).

DHLET

Model: rozšířený Debye-Hückelův zákon

$$pK_i = pK_0 - \frac{A z_i^2 \sqrt{I_i}}{1 + B å \sqrt{I_i}} + C I_i$$

Parametry: pK_0, å, C



Indicator	Found pK_a	a	C
Bromocresol Green	5·033 ±0·013	0·759 ±0·060	0·068 ±0·010
Bromophenol Red	6·138 ±0·010	0·784 ±0·051	0·055 ±0·008
Bromocresol Purple	6·198 ±0·017	0·876 ±0·093	0·055 ±0·012
Bromothymol Blue	7·199 ±0·011	0·774 ±0·060	0·055 ±0·010
Phenol Red	8·035 ±0·018	0·375 ±0·038	0·100 ±0·018
Thymol Blue	9·232 ±0·023	0·806 ±0·108	0·061 ±0·015

Úloha 8.1 *Odhad parametrů Debyeova-Hückelova vztahu*

Stanovte termodynamickou disociační konstantu pK^T_a (parametr β_1), efektivní průměr iontů *å* (parametr β_2) a vysolovací konstantu *C* (parametr β_3) závislosti smíšené disociační konstanty *y* na iontové síle *x* podle rozšířeného Debyeova-Hückelova vztahu pro vybrané sulfonftaleiny. Mají-li oba ionty L^{Z-1} a HL^Z zhruba stejnou velikost *å* [10^{-10} m] a je-li celkový vysolovací koeficient $C = C^Z_{HL} - C^{Z-1}_L$, lze formulovat Debyeův-Hückelův vztah tvarem $y = \beta_1 - \frac{(1 - 2Z)\, A \sqrt{x}}{(1 + B\, \beta_2 \sqrt{x})} + \beta_3\, x$, kde $A = 0.5112$ $mol^{-1/2}.\, l^{1/2}.\, K^{3/2}$, $B = 0.3291$ $mol^{-1/2}.\, m^{-1}.\, l^{1/2}.\, K^{1/2}$ jsou pro 25 °C. Předpokládejte aditivní model měření a normalitu chyb závisle proměnné *y*.

Data: Bromkrezolová zeleň: *Z* = -1, *{x, y}*.

0.010 4.901	0.022 4.871	0.040 4.834	0.060 4.808	0.116 4.765	0.232 4.709
0.392 4.691	0.594 4.677	0.923 4.664	1.330 4.662	2.050 4.686	3.720 4.785

Řešení:

1. Návrh modelu: označme
parametr pK^T_a jako první parametr β_1 s jeho odhadem b_1,
parametr *å* jako druhý parametr β_2 s jeho odhadem b_2 a
parametr *C* třetí parametr β_3 se odhadem b_3.

Pro nulté přiblížení odhadovaných parametrů je voleno
$b_1^{(0)} = 1.0,\ b_2^{(0)} = 1.0,\ b_3^{(0)} = 1.0.$

2. Odhadování parametrů:

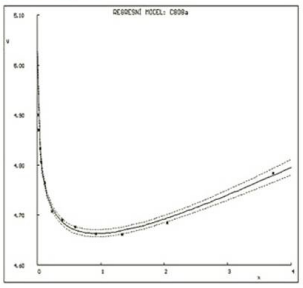
Bodové odhady parametrů:

Parametr β_j	Bodový odhad b_j	Směrodatná odchylka $s(b_j)$	Absolutní vychýlení h_j	Relativní vychýlení $h_{R,j}$ [%]
β_1	5.0336E+00	4.2468E-03	-1.9583E-05	-3.8905E-04
β_2	7.6010E+00	2.1432E-01	4.2364E-03	5.5735E-02
β_3	6.8337E-02	3.4380E-03	1.7995E-05	2.6332E-02

Intervalové odhady parametrů:

Parametr β_j	Bodový odhad b_j	Poloviční délka intervalu spolehlivosti spočtená z délky poloos	maxim
β_1	5.0336E+00	+-1.1915E-02	+-1.4456E-02
β_2	7.6010E+00	+-7.2957E-01	+-7.2957E-01
β_3	6.8337E-02	+-9.9546E-03	+-1.1703E-02

3. Graf regresní křivky:



Obr. 8-1a Rozptylový graf, ***ADSTAT.***

4. Korelační matice parametrů:

Ukazuje, že korelace mezi parametry je výrazná.

	b_1	b_2	b_3
$x[1, i]$	1.00000	-0.82415	0.52400
$x[2, i]$	-0.82415	1.00000	-0.85058
$x[3, i]$	0.52400	-0.85060	1.00000

5. Základní statistické charakteristiky:

Bylo dosaženo výtečného proložení, regresní rabat 99.49% ukazuje, že vysoké procento bodů vyhovuje navrženému modelu Debyeovy-Hückelovy závislosti.

Regresní rabat *D* [%]	: 99.487
Akaikeho informační kriterium *AIC*	: -117.59
Střední kvadratická chyba predikce *MEP*	: 1.5516E-04
První statistický moment residuí m_1	: -1.9709E-07
Třetí statistický moment residuí m_3	: 1.1306E-07
Čtvrtý statistický moment residuí m_4	: 2.0362E-09
Parametr Γ^T	: 8.4501E-02
Parametr Γ^N	: 8.6108E+02

6. Regresní diagnostika:

Obsahuje pomůcky pro kritiku dat, kritiku modelu a kritiku metody.

A. Analýza klasických reziduí:

Směrodatná odchylka reziduí dosahuje hodnoty stejné velikosti, jako je odhad náhodných chyb (šumu) proměnné *y*, tj. $\varepsilon(pK_{a,i} \approx 0.01)$. Rozdělení reziduí je mírně asymetrické, sešikmené k nižším hodnotám, protože *odhad šikmosti* dosahuje záporné hodnoty. Rozdělení se blíží rovnoměrnému, protože *odhad špičatosti* je blízký hodnotě 1.80. S ohledem na malý počet dat nelze z výběrové šikmosti a špičatosti usuzovat na nenormalitu.

Bod *i*	Měřená hodnota y_i	Predikovaná hodnota $y_{vyp,i}$	Směrodatná odchylka $s(y_{vyp,i})$	Vychýlení h_i	Klasické reziduum e_i
1	4.9010	4.9115	3.7122E-03	-9.6193E-06	-1.0524E-02
2	4.8710	4.8691	3.3094E-03	-3.8063E-06	1.9192E-03
3	4.8340	4.8318	2.9072E-03	1.0885E-06	2.2400E-03
4	4.8080	4.8046	2.6222E-03	4.1321E-06	3.3797E-03
5	4.7650	4.7593	2.2938E-03	7.4433E-06	5.6887E-03
6	4.7090	4.7142	2.3949E-03	7.2323E-06	-5.2232E-03
7	4.6910	4.6860	2.7365E-03	4.1517E-06	5.0140E-03
8	4.6770	4.6703	2.9978E-03	3.8001E-07	6.7445E-03
9	4.6640	4.6635	3.1121E-03	-3.8612E-06	5.2314E-04
10	4.6620	4.6689	3.0511E-03	-6.3855E-06	-6.9445E-03
11	4.6860	4.6941	3.1750E-03	-6.2046E-06	-8.1348E-03
12	4.7850	4.7797	6.1717E-03	5.4489E-06	5.3175E-03

Reziduální součet čtverců *RSC*	: 4.0410E-04
Směrodatná odchylka reziduí *s(e)*	: 5.8030E-03
Odhad šikmosti g_1	: -0.579
Odhad špičatosti g_2	: 1.796
Hamiltonův *R-faktor* [%]	: 0.122

B. Tabulka vlivných bodů:

Obsahuje základní charakteristiky k odhalení vlivných bodů.

Bod *i*	Jackknife reziduum $e_{J,i}$	Cookova vzdálenost D_i	Diagonální prvky $H_{i,i}$	Normalizovaná vzdálenost *FDA*	Věrohodnostní vzdálenost *LDA*
1	-2.2875E+00	5.2536E-01	3.0692E-01	1.2144E-01	2.6567E-01
2	3.1245E-01	1.1668E-02	2.4392E-01	3.8138E-03	2.8886E-03
3	3.5252E-01	1.0641E-02	1.8824E-01	4.6230E-03	7.9056E-03
4	5.2560E-01	1.8108E-02	1.5314E-01	9.3211E-03	8.6827E-03
5	8.9337E-01	3.6124E-02	1.1718E-01	2.1943E-02	1.4799E-02
6	-8.1923E-01	3.4006E-02	1.2774E-01	1.9626E-02	1.3011E-02
7	8.0345E-01	4.4839E-02	1.6679E-01	2.1813E-02	1.4450E-02
8	1.1447E+00	1.0565E-01	2.0015E-01	4.4082E-02	3.4816E-02
9	8.3151E-02	7.1250E-04	2.1571E-01	2.7009E-04	7.3859E-03
10	-1.1908E+00	1.1815E-01	2.0734E-01	4.5560E-02	3.7563E-02
11	-1.4634E+00	1.8341E-01	2.2452E-01	6.5850E-02	7.1201E-02
12	2.6179E+00	7.7430E+00	8.4834E-01	4.1027E-03	2.7646E-03

7. Mapa citlivostní funkce: Citlivostní funkce vyjadřuje změnu regresního modelu při změně parametru o ±5% ukazují, že parametr b_1 a b_3 jsou dobře podmíněny v modelu, jejich změna způsobí změnu účelové funkce 8 až 9 řádů. Parametr b_2 je ve srovnání s předchozími parametry méně citlivý, hůře podmíněný v modelu, změna je podstatně menší.

Parametr	Relativní změna	Souhrnná citlivost	Relativní změna
j	C_{jR}(-5%), [%]	C_j, [%]	C_{jR} (+5%), [%]
1	8.0486E-09	1.0000	-7.2821E-09
2	1.5937E+01	1.0301E-03	-1.3244E+01
3	-3.3799E-07	1.7701	9.5080E-08

8. Predikční schopnost modelu:

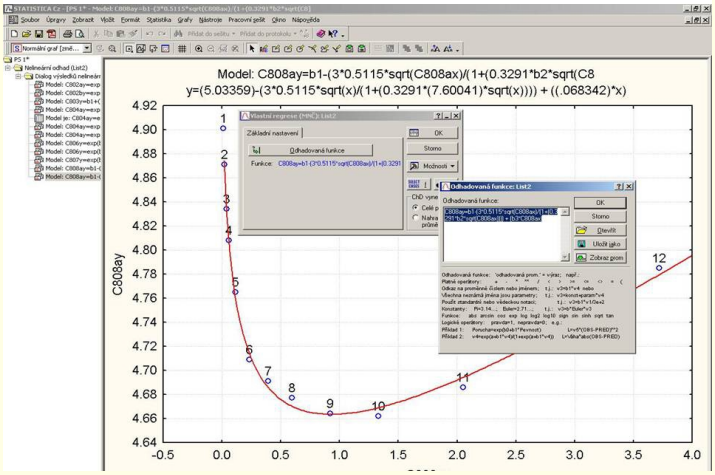
Pro $n = 12$ bude $M_1 = 1$ až 6, $M_2 = 7$ až 12, $RSC(M_1) = 2.8534E-05$, $RSC(M_2) = 5.3549E-05$, $U(\boldsymbol{b}) = 40.410E-05$, a proto $K = 4.923$.

Jelikož se K neblíží k 1, je predikční schopnost modelu slabší.

9. Souhlas s požadavky fyzikálního smyslu:

Termodynamickou disociační konstantu $pK^T_a = 5.034$ a *vysolovací konstanta* $C = 0.068$ mají fyzikální smysl, *efektivní průměr iontu* $\mathring{a} = 7.6 \times 10^{-10}$ m je v souhlase s hodnotami Kielandových tabulek.

Úloha C8.08 *Odhad tří parametrů rozšířeného Debyeova-Hückelova vztahu*



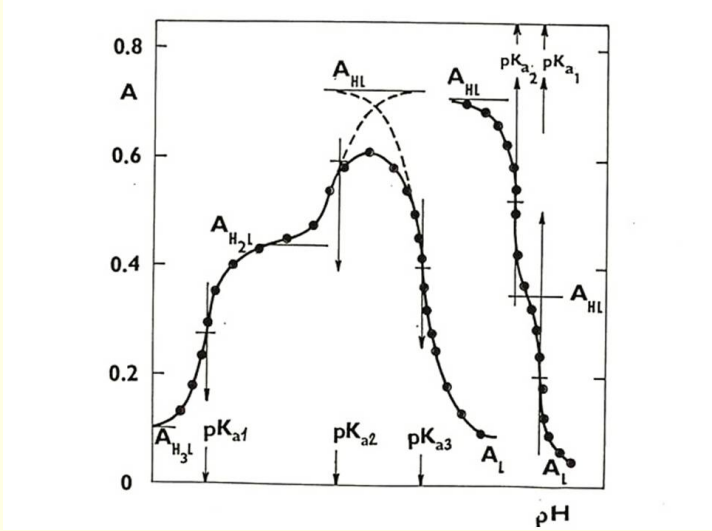
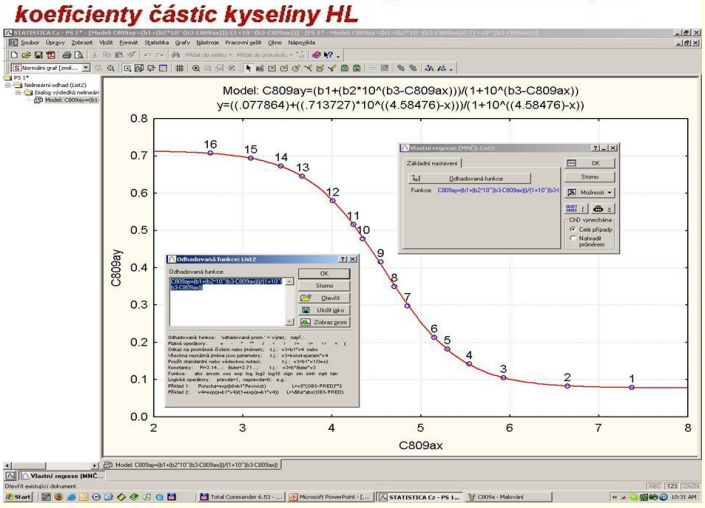
DCLET

Model: rovnice A-pH křivky pro rovnováhu

$$L^{3-} + H^+ = LH^{2-} + H^+ = LH_2^- + H^+ = LH_3$$

$$A = c_L\, d\, \frac{\varepsilon_L + \sum_{j=1}^{J} \varepsilon_{LH_j} . 10^{j.\log a_{H^+} + \log\beta_{01j}}}{1 + \sum_{j=1}^{J} 10^{j.\log a_{H^+} + \log\beta_{01j}}}$$

Parametry: $\log \beta_{01j}$, ε_L, ε_{LHj}, j = 1, ..., J



Úloha C8.09 *Disociační konstanty a molární absorpční koeficienty částic kyseliny HL*