vodní hospodářství®

www.vodnihospodarstvi.cz

ročník 57
6
2007

KUNST
spol. s r.o. Hranice

JSME TU PRO VÁS

ČERPACÍ STANICE
čerpací stanice

ÚPRAVNY VODY
úpravny vody

ČISTÍRNY
čistírny čistírn

KONZULTACE
konzultace kon

PROJEKTY
projekty projek

DODÁVKY
dodávky dodáv

MONTÁŽE
montáže montáže montáže montá

Více informací na str. 221

1997-2007
EWA
10 let členství

aqua®
19. - 21. 6. 2007

Radi Vás privítame
v stánku redakcie časopisu
Vodní hospodářství,
v pavilóne 1 - prízemie, č. 5

PŘÍLOHA
•VTEI•

Statistické zpracování vodohospodářských dat

5. Výstavba regresního modelu diagnostikami regresního tripletu

Milan Meloun

Klíčová slova
Regresní triplet - kritika dat - kritika modelu - kritika metody - regresní metoda - metoda nejmenších čtverců - regresní diagnostika - analýza reziduí - vlivné body - odlehlé body - extrémy - validace nové analytické metody - vícenásobná regresní analýza -

Souhrn

Postup hledání regresního modelu je popsán na dvou úlohách vodohospodářské kontrolní laboratoře a obsahuje: 1. Návrh modelu začíná vždy od nejjednoduššího modelu, lineárního. 2. Předběžná analýza dat sleduje proměnlivost proměnných na rozptylových diagramech a indexových grafech. Vyšetřuje se multikolinearita, heteroskedasticita, autokorelace a vlivné body. 3. Odhadování parametrů se provádí klasickou metodou nejmenších čtverců, následované testem významnosti parametrů Studentovým *t*-testem. Střední kvadratická chyba predikce *MEP* a Akaikovo informační kritérium *AIC* představují rozhodčí kritéria při hledání nejlepšího regresního modelu. 4. Regresní diagnostika provádí identifikaci vlivných bodů a ověření předpokladů metody nejmenších čtverců. V případě více vysvětlujících proměnných se posoudí vhodnost jednotlivých proměnných pomocí parciálních regresních grafů a parciálních reziduálních grafů. 5. 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 (RH) 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.

Úvod

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

Základní předpoklady metody nejmenších čtverců (MNČ): Statistické vlastnosti odhadů $\hat{y}_P, \hat{e}$, $\boldsymbol{b}$ závisí na splnění jistých sedmi předpokladů. Pokud platí předpoklady I až IV, představují odhady $\boldsymbol{b}$ parametrů β **n**ejlepší, **n**estranné a **l**ineární **o**dhady (značeno metoda NNLO). Navíc mají asymptoticky normální rozdělení. Pokud platí ještě předpoklad VII, mají odhady $\boldsymbol{b}$ 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* $\boldsymbol{X}$ *má hodnost rovnou právě m.* To znamená, že žádné její dva sloupce $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ nejsou *kolineární* čili 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 $E(\varepsilon_i') = K$, $i = 1, \ldots, n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení bude $E(\varepsilon_i') = 0$, kde $\varepsilon_i' = y_i - \hat{y}_{Pi} - K$.

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 $\boldsymbol{X}\beta$ a kovarianční maticí $\sigma^2\boldsymbol{E}$, kde $\boldsymbol{E}$ je jednotková matice.

Regresní diagnostika: Metoda nejmenších čtverců MNČ nezajišťuje ještě nalezení přijatelného modelu, a to jak ze statistického, tak i z fyzikálního hlediska. Musí být totiž splněny podmínky, odpovídající složkám tzv. *regresního tripletu* [data, model, metoda odhadu]. Regresní diagnostika obsahuje pomůcky a postupy k identifikaci a) vhodnosti dat pro navržený regresní model (*kritika dat*), b) vhodnosti modelu pro daná data (*kritika modelu*), c) splnění základních předpokladů MNČ (*kritika metody*). 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*. 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.

I. KRITIKA DAT

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* čili malé rozmezí nebo přítomnost vybočujících bodů, b) *nesprávnost navrženého modelu* čili 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 jejich dělení na *a) vybočující pozorování* (outliers), které se liší v hodnotách vysvětlované (závisle) proměnné *y* od ostatních, a *b) 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 $\boldsymbol{H}$, bližší lze nalézt v citaci [1]. Výskyt vlivných bodů (VB) je zdrojem řady problémů a způsobuje zkreslení odhadů a růst rozptylů odhadů parametrů. Vlivné body se dělí dle charakteru: a) *hrubé chyby*, což jsou vybočující pozorování, b) *body s vysokým vlivem* (tzv. golden points), které rozšiřují predikční schopnosti modelu. c) *zdánlivě vlivné body*, jež jsou důsledkem nesprávného regresního modelu.

Statistická analýza reziduí dělí rezidua na následujících několik druhů:

1. Klasická rezidua $\hat{e}_i = y_i - \boldsymbol{x}_i\,\boldsymbol{b}$. Existují nesprávné představy o reziduích, že a) rozdělení reziduí je stejné jako rozdělení chyb, b) vlastnosti reziduí jsou shodné s vlastnostmi chyb, c) čím je reziduum $\hat{e}_i$ větší, tím je bod vlivnější, a tím spíše by se měl z dat vyloučit. Rozepsáním definičního vztahu

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

je zřejmé, že každé reziduum $\hat{e}_i$ je vlastně lineární kombinací všech chyb ε_i. Rozdělení reziduí je proto závislé na rozdělení chyb, na prvcích projekční matice $\boldsymbol{H}$, na velikosti výběru *n*. Klasická rezidua mají vlastnosti: a) Rozptyl reziduí je *nekonstantní*, i když rozptyl chyb $\hat{\sigma}^2$ je konstantní. b) Rezidua jsou korelovaná: existuje *párový korelační koeficient* r_{ij} mezi dvěma rezidui e_i a e_j formulovaný jako $r_{ij} = \dfrac{-H_{ij}}{\sqrt{(1 - H_{ii})\,(1 - H_{jj})}}$, i když chyby ε_i a ε_j jsou nezávislé. c) rezidua *neindikují* extrémní hodnoty, d) rezidua jsou normálnější než chyby (effekt supernormality), e) u malých výběrů nemusí správně indikovat model.

2. Normovaná rezidua $\hat{e}_N$ definovaná vztahem $\hat{e}_{iNi} = \hat{e}_i/\hat{\sigma}$ jsou normálně rozdělené veličiny $\hat{e}_{Ni} \sim N(0, 1)$. Platí u nich diagnostické pravidlo 3σ, které říká, že rezidua větší než $\pm 3\hat{\sigma}$ indikují vybočující body.

3. Standardizovaná rezidua $\hat{e}_{Si}$ definovaná vztahem $\hat{e}_{Si} = \dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - H_{ii}}}$ mají konstantní rozptyl a maximální hodnota $\hat{e}_{Si}$ je ohraničena velikostí $\sqrt{n - m}$. Platí u nich diagnostické pravidlo, že jsou vhodná především k indikaci heteroskedasticity.

4. Jackknife rezidua $\hat{e}_J$ čili **„plně studentizovaná“,** jsou definovaná

vztahem, ve kterém se užije místo $\hat{\sigma}$ odhadu směrodatné odchylky $\hat{\sigma}_{(-i)}$, a dostane se $\hat{e}_{Ji} = \hat{e}_{Si}\sqrt{\dfrac{n-m-1}{n-m-\hat{e}_{Si}^2}} = \sqrt{n-m}\,\mathrm{cotg}\,\Theta_i$.

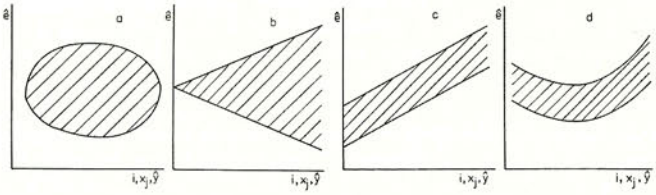
Rezidua mají Studentovo rozdělení s $(n - m - 1)$ stupni volnosti. Platí u nich diagnostické pravidlo, že jsou vhodná především k identifikaci vybočujících bodů (outliers).

5. Predikovaná rezidua $\hat{e}_P$ jsou definovaná vztahem $\hat{e}_{Pi} = y_i - x_i \boldsymbol{b}_{(i)} = \dfrac{\hat{e}_i}{1 - H_{ii}}$, kde $\boldsymbol{b}_{(i)}$ jsou MNČ odhady ze všech bodů kromě i-tého. Platí u nich diagnostické pravidlo, že indikují vybočující hodnoty (outliers).

6. Rekurzívní rezidua $\hat{e}_R$ jsou definována vztahy $\hat{e}_{Ri} = 0,\ i = 1, \ldots, m,$

$$\hat{e}_{Ri} = \frac{y_i \boldsymbol{x}_i \boldsymbol{b}_{i-1}}{\sqrt{1 + \boldsymbol{x}_i (\boldsymbol{X}_{i-1}^{\mathrm{T}} \boldsymbol{X}_{i-1})^{-1} \boldsymbol{x}_i^{\mathrm{T}}}},\ i = m + 1, \ldots, n$$ kde $\boldsymbol{b}_{i-1}$ jsou odhady získané z prvních $(i - 1)$ bodů. Platí pro ně diagnostické pravidlo, že indikují autokorelaci a nestabilitu regresního modelu v čase.

Obrazce v diagnostických grafech: Pokud se v diagnostických grafech reziduí objeví tvar "mraku" bodů (**obr. 1.**), je detekována správnost metody nejmenších čtverců. Jiné obrazce bodů v grafu reziduí indikují vesměs nesprávnost v datech či nesprávnost modelu: tvar výseče odhaluje nekonstantnost rozptylu čili heteroskedasticitu v datech, tvar pásu indikuje chybu ve výpočtu nebo nepřítomnost x_j v modelu, tvar pásu může být i důsledkem vybočujících hodnot nebo indikuje chybný výpočet, kdy v regresním modelu chybí absolutní člen. Nelineární tvar ukazuje na nesprávně navržený model.

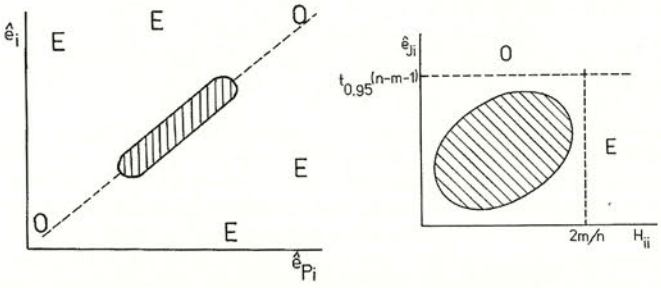


Obr. 1. Nejčastější tvary obrazce bodů v grafické analýze reziduí: a) tvar mraku, b) tvar výseče, c) tvar pásu a d) nelineární tvar.

Grafy identifikace vlivných bodů: Existuje pět grafických diagnostik vlivných bodů, které současně testují charakter vlivných bodů, zatímco ostatní diagnostické grafy spíše odhalují podezřelé body.

1. Graf predikovaných reziduí GPR, (osa x: $\hat{e}_{Pi}$, osa y: $\hat{e}_i$). V grafu jsou extrémy snadno identifikovány svou polohou, neboť leží mimo přímku $y = x$. Vybočující body leží sice na přímce $y = x$ nebo v její blízkosti, jsou však dostatečně vzdáleny od mraku, shluku ostatních bodů (**obr. 2a**).

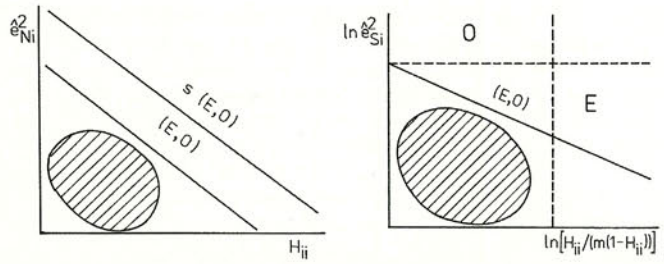
2. Williamsův graf WG, (osa x: prvky H_{ii}, osa y: $\hat{e}_{Ji}$). Do tohoto grafu se zakreslují jednak mezní linie pro vybočující body, $y = t_{0.95}(n - m - 1)$, a jednak mezní linie pro extrémy, $x = 2\,m / n$. Zde $t_{0.95}(n - m - 1)$ je 95% kvantil Studentova rozdělení s $n - m - 1$ stupni volnosti (**obr. 2b**).



Obr. 2. Grafy vlivných bodů: a) Graf predikovaných reziduí (GPR), b) Williamsův graf: E značí extrém a O značí vybočující bod.

3. Pregibonův graf PG, (osa x: prvky H_{ii}, osa y: $\hat{e}_{Ni}^2$). Protože platí, že $E(H_{ii} + \hat{e}_{Ni}^2) = (m + 1)/n$, lze do tohoto grafu zakreslit dvě hraniční přímky $y = -x + 2\,(m + 1)/n$ a $y = -x + 3\,(m + 1)/n$. K rozlišení mezi body platí pravidlo: bod je silně vlivný, leží-li nad horní přímkou. Bod je pouze vlivný, leží-li mezi oběma přímkami. Může jít ale jak o extrém, tak o vybočující bod (**obr. 3a**).

4. McCullohův-Meeterův graf MMG, (osa x: $\ln [H_{ii} / (m\,(1 - H_{ii}))]$, osa y: $\ln \hat{e}_{Si}^2$). Linie v tomto grafu představují přímky stejného vlivu přímky se směrnicí -1. Například pro 90% konfidenční čáru je $y = -x - \ln [F_{0.9}(n - m, m)]$. Omezující čára pro extrémy je $x = \ln \dfrac{2}{n - 2m}$ a omezující čára

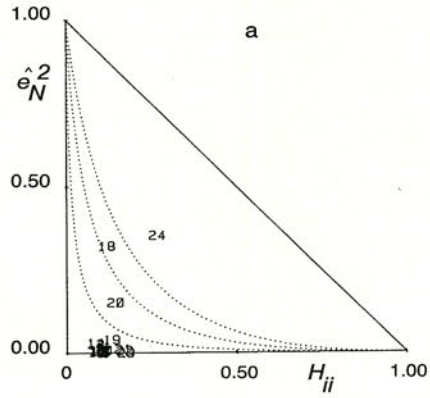


Obr. 3. Grafy vlivných bodů: a) Pregibonův graf (PG), kde (E, O) značí vlivné body, s(E,O) značí silně vlivné body, b) McCullohův-Meeterův graf (MMG), kde E značí extrém a O vybočující bod, (E, O) obojí vlivné body.

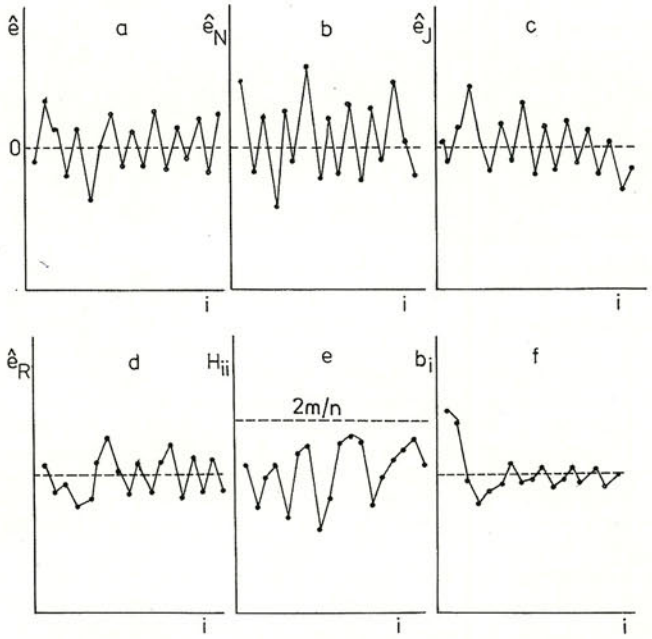
pro vybočující body je $y = \ln [(n - m)\, t^2_{0.95}(t^2_{0.95} + n - m)]$, kde $t_{0.95}$ je 95% kvantil Studentova rozdělení s $n - m - 1$ stupni volnosti a $F_{0.9}\,(n - m - 1)$ je 90% kvantil F-rozdělení s $n - m$ a m stupni volnosti (**obr. 3b**).

5. L-R grafy, (osu y: $\hat{e}_{Ni}^2 = \hat{e}_i^2 / RSC$, na osu x: prvky H_{ii}), **obr. 4.** Body nad jednotlivými lemniskátami jsou vlivné, a to ve směru horního rohu trojúhelníka jde o vybočující body a ve směru pravého rohu trojúhelníka jde o extrémy.

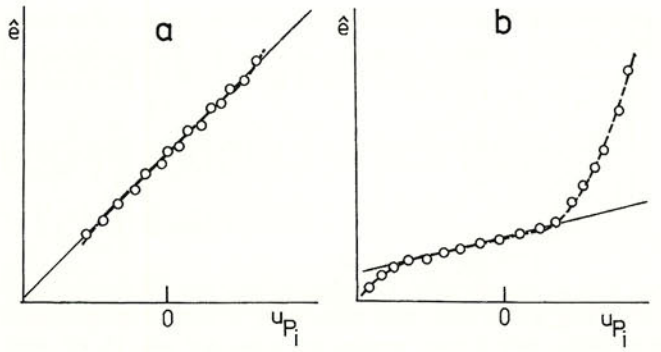
6. Indexové grafy IG, (osa x: index i, osa y: rezidua $\hat{e}_i$, $\hat{e}_{Si}$, $\hat{e}_{Ni}$, $\hat{e}_{Pi}$, $\hat{e}_{Ji}$, $\hat{e}_{Ri}$, nebo prvky H_{ii} či H^*_{ii} a nebo odhady $\boldsymbol{b}_i$). Mohou zde být i rekurzivní odhady regresních parametrů $\boldsymbol{b}_i$ (**obr. 5.**). Tyto grafy indikují pouze podezřelé body, nejsou schopny nikterak testovat vybočující body jako je tomu v pěti předešlých grafech.



Obr. 4. Grafy vlivných bodů: L-R graf.



Obr. 5 Rozličné typy indexových grafů znázorňujících indexovou závislost: a) $\hat{e}_i$, b) $\hat{e}_{Ni}$, c) $\hat{e}_{Ji}$, d) $\hat{e}_{Ri}$, e) H_{ii}, f) b_i na i.



Obr. 6. Rankitový graf *Q-Q* reziduí indikuje: a) přibližně normální rozdělení, b) rozdělení s dlouhými konci.

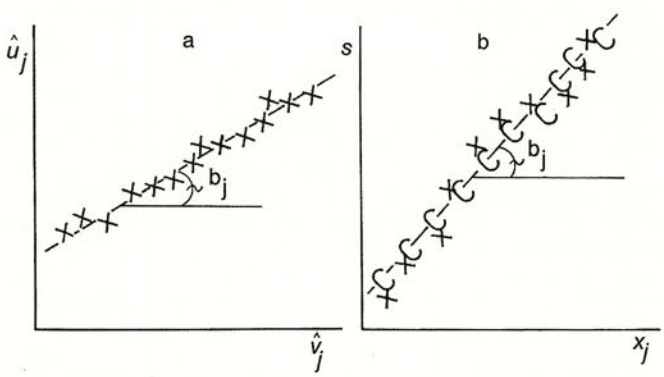
7. Rankitové grafy Q-Q, (osa *x*: u_{P_i} pro $P_i = i / (n + 1)$, osa *y*: $\hat{e}_{(i)}$, $\hat{e}_{S(i)}$, $\hat{e}_{N(i)}$, $\hat{e}_{P(i)}$, $\hat{e}_{J(i)}$, $\hat{e}_{R(i)}$). Na osu *x* se vynášejí kvantily normovaného normálního rozdělení u_P pro $P_i = i / (n + 1)$ a na osu *y* pořádkové statistiky čili vzestupně setříděné hodnoty reziduí, (**obr. 6.**). Cílem je ověřit normalitu rozdělení reziduí.

II. KRITIKA MODELU

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 rozptylové grafy *mylně indikovat* nelineární trend i u lineárního modelu. Z řady různých grafů k posouzení vztahu *y* a x_j se zde omezíme na a) parciální regresní grafy, a b) parciální reziduální grafy.

Belseyho parciální regresní grafy umožňují posouzení kvality navrženého regresního modelu, indikují přítomnost vlivných bodů, nesplnění předpokladů klasické MNČ a konečně vyjadřují závislost mezi $\boldsymbol{y}$ a zvolenou proměnnou $\boldsymbol{x}_j$ při statisticky neměnném vlivu ostatních $\boldsymbol{X}_{(j)}$. Přitom matice $\boldsymbol{X}_{(j)}$ vznikne z matice $\boldsymbol{X}$ vynecháním *j*-tého sloupce $\boldsymbol{x}_j$, odpovídajícího *j*-té vysvětlující proměnné. Grafy mají vlastnosti: a) Směrnice přímky v parciálním regresním grafu je stejná jako odhad b_j v neděleném modelu a úsek je roven nule. Lineární závislost platí pouze v případě, že navržený model je správný. b) Korelační koeficient mezi oběma proměnnými parciálního regresního grafu odpovídá parciálnímu korelačnímu koeficientu $\hat{R}_{yx_j(x)}$. Vychází se z regresního modelu $y = \boldsymbol{X}_{(j)}\beta^* + \boldsymbol{x}_j c + \varepsilon$, kde β^* má rozměr $(m - 1) \times 1$, *c* je regresní parametr příslušející *j*-té proměnné.

Parciální reziduální grafy jsou zvané také grafy „komponenta + reziduum". Rovnici $\hat{\boldsymbol{e}} = \boldsymbol{u}_j - \boldsymbol{v}_j . b_j$ lze přepsat do tvaru $\hat{\boldsymbol{u}}_j = \hat{\boldsymbol{e}} + b_j(E - H_{(j)})x_j$ jako závislost $\hat{\boldsymbol{e}} + b_j(E - H_{(j)})x_j$ na $(E - H_{(j)})x_j$. Jedná se o závislost parciálních reziduí $\boldsymbol{s}$ na proměnné $\boldsymbol{x}_j$, kde $s = \hat{\boldsymbol{e}} + bx_j = y - \sum_{k \neq j}^{m} x_k b_k$. V grafu se znázorňuje deterministická komponenta $c_{ij} = (x_{ij} - \bar{x}_j)b_j, i = 1,\ldots, m$, která se v grafu značí písmenem „C" a parciální reziduum $s_i = c_{ij} + \hat{e}_i$, $i = 1, \ldots, n$, které se zde označuje křížkem „+". Uvedená diagnostika má vlastnosti: a) Směrnice závislosti $\boldsymbol{s}$ na $\boldsymbol{x}_j$ je rovna $\boldsymbol{b}_j$ a úsek je nulový. Lineární závislost ukazuje na vhodnost navržené $\boldsymbol{x}_j$ v modelu. b) Rezidua přímky jsou přímo rezidua $\hat{e}_i$ pro nedělený model. c) Pokud je úhel mezi $\boldsymbol{x}_j$ a některými sloupci matice $\boldsymbol{X}_{(j)}$ malý (multikolinearita), ukazuje parciální reziduální graf nesprávně malý rozptyl kolem regresní přímky $b_j\boldsymbol{x}_j$ a dochází i k potlačení efektu vlivných bodů.



Obr. 7. Kritika modelu pomocí a) parciálního regresního grafu, b) parciálního reziduálního grafu.

Znaménkový test vhodnosti modelu. Nenáhodnost reziduí čili trend v reziduích lze testovat znaménkovým testem postupem: 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. Počet sekvencí n_t a jeho rozptyl D_t se vyčíslí dle $n_t = 1 + \frac{2n_+n_-}{n_+ + n_-} \approx 1 + \frac{n}{2}$

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

Vlastní testování spočívá ve vyšetření podmínky: je-li $n_U < n_t - C1.96\sqrt{D_t}$, existuje v reziduích trend a model je nesprávně navržen. Konstanta $C = 0.5$ je korekcí na spojitost.

III. KRITIKA METODY

V praxi bývají některé předpoklady MNČ porušeny, což vede k použití modifikací metody MNČ. K porušení předpokladů dochází v základních případech: a) Na parametry jsou kladena omezení, což vede na užití metody podmínkových nejmenších čtverců (MPNČ). b) Kovarianční matice chyb není diagonální (autokorelace), případně data mají nestejný rozptyl (heteroskedasticita), což vede na užití *metody zobecněných nejmenších čtverců* (MZNČ), respektive *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. Pro odhad parametrů *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 stejných rozptylů vede metoda k odhadům minimalizujícím kolmé vzdálenosti (*orthogonální regrese*). e) Pro špatně podmíněné matice X^TX se používá *metoda racionálních hodností* (RH), vedoucí k systému vychýlených odhadů, kde vychýlení je řízeno jedním parametrem.

1. Hetero/homoskedasticita čili nekonstantnost rozptylu lze popsat následovně: rozptyl veličiny y_i v *i*-tém bodě je popsán vztahem $\sigma_i^2 = \sigma^2 \exp(\lambda\, \boldsymbol{x}_i\, \beta)$, kde $\boldsymbol{x}_i$ je *i*-tý řádek matice $\boldsymbol{X}$. Předmětem tohoto testu homoske-dasticity je ověření nulové hypotézy H_0: $\lambda = 0$ za použití Cookova-Weisbergova testačního kritéria

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

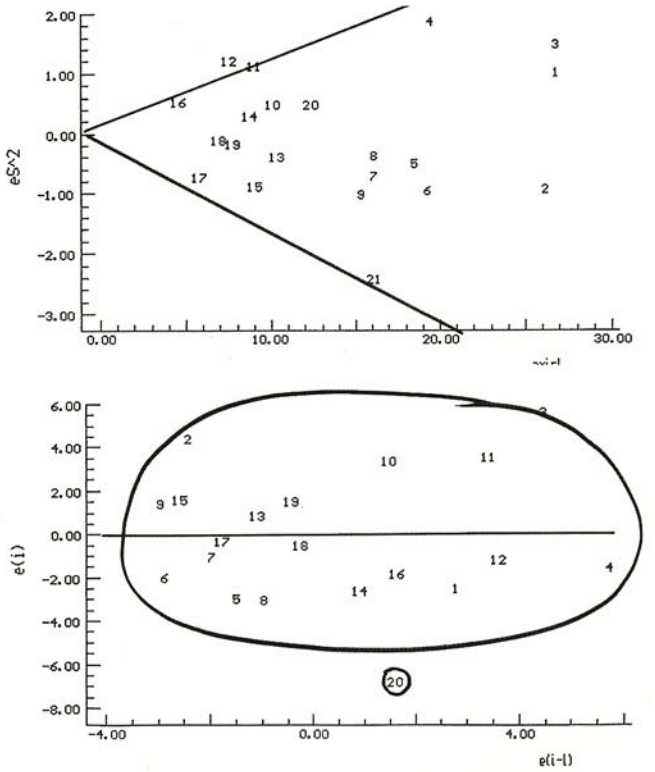
kde $\hat{y}_P = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$. Výsledkem testování je závěr: a) Je-li $S_f < \chi^2(1)$, H_0 (homoskedasticita) je přijata. b) Pro homoskedasticitu tvoří diagnostický graf $\hat{e}_{Si}^2$ v závislosti na $(1 - H_{ii})\hat{y}$ náhodný elipsovitý mrak bodů. Pro heteroskedasticitu vznikne v tomto grafu typický klínový obrazec, (**obr. 8a**).

2. Autokorelace: Data časových řad mají náhodné 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 laboratorních datech vyskytuje často. b) Pro $\rho_1 \leq 1$ jde o autokorelační koeficient 1. řádu. Waldův test pak testuje nulovou hypotézu H_0: $\rho_1 = 0$ *vs.* H_A: $\rho_1 \neq 0$. Je-li Waldovo testační kritérium $W_a = \frac{n\,\hat{\rho}_1^2}{1 - \hat{\rho}_1^2} < \chi^2(1)$, je H_0 o homoskedasticitě přijata a v datech není prokázána autokorelace, (**obr. 8b**).

3. Normalita chyb: normalitu náhodných chyb lze testovat dvojím způsobem, a to rankitovým grafem a numericky: a) *Rankitový (Q-Q) graf*, ($\hat{e}_{(i)}$ nebo $\hat{e}_{Ri}$ na u_{Pi} pro $P_i = i / (n + 1)$), b) *Test normality* testuje nulovou hypotézu H_0: normalita ***vs.*** H_A: nenormalita. Slouží k tomu Jarque-Berrova testační statistika, která je formulována vztahem

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

kde $\hat{u}_j$ je *j*-tý moment reziduí $\hat{u}_j = \frac{\sum_{i=1}^{n}\hat{e}_i^j}{n}$. Test normality spočívá ve vyšetření nerovnosti: je-li $L(\hat{e}) > \chi^2_{1-\alpha}(2) = 5.99$, je H_0 (normalita) zamítnuta. Je třeba podotknout, že test není vhodný pro malé výběry.



Obr. 8. a) Klínový obrazec indikuje heteroskedasticitu v datech, b) elipsa indikuje data bez autokorelace.

Postup výstavby lineárního regresního modelu [1, 4]

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. Uživatel může volit polynomickou transformaci zadáním stupně polynomu. Provádí se sestavení korelační matice $\boldsymbol{R}$ a její rozklad na vlastní čísla a vlastní vektory.

3. Odhadování parametrů: Odhadování parametrů modelu se provádí metodou nejmenších čtverců MNČ nebo metodou racionálních hodností RH. 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 $\beta_j = 0$. Jsou provedeny testy významnosti odhadů b_j, vícenásobného korelačního koeficientu R a koeficientu determinace D. Je vhodné sledovat rozhodčí kritéria hypotézy regresního modelu jako je střední kvadratická chyba predikce *MEP* a Akaikovo informační kritérium *AIC*.

4. Regresní diagnostika: S využitím pěti rozličných grafů je prováděna identifikace vlivných bodů, a to *grafem predikovaných reziduí, grafy Wiliamsovým, Pregibonovým, McCulloh-Meeterovým*, a *L-R grafem*. Pak následuje ověření předpokladů metody nejmenších čtverců jako jsou homoskedasticita, nepřítomnost autokorelace a normalita rozdělení chyb. V případě více vysvětlujících proměnných se posoudí vhodnost jednotlivých proměnných a jejich funkcí s využitím parciálních regresních grafů nebo grafů „komponenta + reziduum". Je uveden odhad autokorelačního koeficientu reziduí prvního řádu $\hat{\rho}_1$. *Tabulka vlivných bodů* obsahuje rozličná rezidua, u kterých jsou hvězdičkou označeny hodnoty silně vlivných bodů, které by měly být z dat odstraněny.

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.

Úloha 1: *Validace analytické metody stanovení formaldehydu ve vodách* (**V6.16** v cit. [4])

Postup validace nové analytické metody bude ukázán na stanovení formaldehydu ve vzorcích fenolových vod polarografickou metodou. Laboratoř fenoplastů však navrhla jednodušší redox-titraci. Rozptyl obou metod je prakticky stejný. Je proto třeba validovat, zda navržená metoda bude poskytovat správné a reprodukovatelné výsledky. Obsah formaldehydu ve vodě [mg/l] byl stanoven polarograficky *x*, redox-titrací *y*: 76.8 80.5, 117 112.6, 129.1 128, 160.1 152.2, 236.4 239.4, 258 250, 284.2 287, 303.2 307.8, 386.2 391.7, 474.3 480.2, 532.4 530.8, 937.6 934.2, 2654.3 2647.2.

Řešení: **1. Odhadování parametrů:** Metodou nejmenších čtverců MNČ byly nalezeny odhady dvou parametrů, úseku β_0 a směrnice β_1. Studentův *t*-test ukázal, že úsek (absolutní člen) β_0 je statisticky nevýznamný čili nulový, zatímco směrnice β_1 je vůči nule statisticky významná a blízká jedničce.

	Odhad	Směrodatná odchylka	Test H_0: $b_j = 0$ *vs.* H_A: $b_j <> 0$ *t*-kritérium	hypotéza H0 je	Hladina významnosti
b_0	1.0432E-02	5.6588E-03	1.843	Akceptována	0.081
b_1	9.2170E-01	1.9331E-02	47.681	Zamítnuta	0.000

2. Regresní diagnostika: *Párový korelační koeficient* $R = 0.99585$ ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota *koeficientu determinace* $D = R^2$ (99.17%) představující procento bodů, vyhovujících regresnímu modelu ukazuje, že všechny body výtečně korespondují s modelem přímky. *Střední kvadratická chyba predikce MEP* = 4.14E-04 a *Akaikovo informační kritérium AIC* = -166.78 jsou vhodné k rozlišení mezi několika navrženými modely. Za optimální se považuje takový model, pro který dosahuje *MEP* a *AIC* nejmenších hodnot. *Graf predikovaných reziduí* a *Williamsův graf* ukazují na odlehlé body 14, 17 a 21.

3. Konstrukce zpřesněného modelu: Po odstranění bodů 14, 17, 21 byly nalezeny odhady parametrů zpřesněného modelu (v závorce je vždy uveden odhad směrodatné odchylky parametru) $y = 0.00639\ (0.00241) + 0.9403\ (0.0094)\ x$, který je doložen statistickými charakteristikami: *střední kvadratická chyba predikce MEP* = 6.15E-05 a *Akaikeho informační kritérium AIC* = -174.03 dosáhly nyní nižších hodnot, čímž dokazují kvalitnější model než model předešlý. Rezidua nyní vykazují normální rozdělení a nevykazují trend, stále však vykazují heteroskedasticitu, a proto lze doporučit užití metody vážených nejmenších čtverců ($w_i = 1/y_i^2$) ke kompenzaci heteroskedasticity v datech. Opravený model má potom tvar, $y = 0.00213\ (0.00197) + 0.9462\ (0.0731)\ x$. Jelikož došlo k dalšímu snížení rozhodčích kritérií, *střední kvadratické chyby predikce MEP* = 5.12E-05 a *Akaikeho informačního kritéria AIC* = -181.63, lze považovat tyto nalezené odhady za lepší než předešlé.

4. Zhodnocení kvality modelu: Intervalový odhad parametrů úseku b_0 a směrnice b_1 bude pro $n = 20$ a $m = 2$

$$b_0 - t_{1-\alpha/2}(n-m)\sqrt{D(b_0)} \le \beta_0 \le b_0 + t_{1-\alpha/2}(n-m)\sqrt{D(b_0)}$$

dosazením $0.00213 - 2.12 \times 0.00197 \le \beta_0 \le 0.00213 + 2.12 \times 0.00197$ ve tvaru $\mathbf{-0.00205 \le \beta_0 \le 0.00630}$.

Tento interval spolehlivosti úseku regresní přímky zahrnuje nulu, takže lze úsek β_0 považovat za nulový.

Analogicky dosazením do intervalu spolehlivosti směrnice se obdrží nerovnost

$$0.9462 - 2.12 \times 0.0731 \le \beta_1 \le 0.9462 + 2.12 \times 0.0731$$

a vyčíslením $\mathbf{0.7912 \le \beta_1 \le 1.1012}$. Jelikož *interval spolehlivosti směrnice obsahuje jedničku, lze považovat směrnici β_1 za jednotkovou.*

Závěr: Lze uzavřít, že úsek regresní přímky lze považovat za nulový $\beta_2 = 0$ a směrnice β_1 není významně odlišná od jedničky. Výsledky získané nově navrženou titrační metodou se proto statisticky významně neliší od výsledků metodou standardní polarografickou.

Úloha 2: *Lineární regresní model k posuzování účinnosti čistíren odpadních vod* (**M642** v cit [4])

Při posuzování účinnosti čistíren odpadních vod ČOV se sleduje ve výtoku několik parametrů: vyšetřením regresního tripletu je třeba nalézt lineární regresní model a vyšetřit významnost jednotlivých regresních parametrů při sledování závislosti biologické spotřeby kyslíku za 5 dní BSK_5 y na chemické spotřebě kyslíku CHSK-Cr x_1, na množství rozpuštěných látek RL x_2 a na množství amoniakálního dusíku N-NH_4^+ x_3 za jeden měsíc CHSK-Cr x_1 [mg/l], RL x_2 [mg/l], N-NH_4^+ x_3 [mg/l], BSK5 y [mg/l]: 320.0, 363.0, 23.0, 66.0; 370.0, 416.0, 32.0, 70.0; 460.0, 677.0, 17.0, 50.0; 125.0, 270.0, 33.0, 23.0; 200.0, 123.0, 20.0, 32.0; 340.0, 156.0, 35.0, 155.0;

230.0, 87.0, 35.0, 62.0; 320.0, 240.0, 19.0, 55.0; 180.0, 330.0, 36.0, 50.0; 270.0, 136.0, 24.0, 74.0.

1. Návrh modelu: začíná se vždy od nejjednoduššího modelu, u kterého vystupují x_1, x_2, x_3 v prvních mocninách a nevyskytují se interakční členy. Na začátku analýzy je vždy zařazen absolutní člen β_0, takže pro daná data bude navržený regresní model tvaru $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

2. Předběžná analýza dat: polohu a rozptýlení všech proměnných y, x_1, x_2, x_3 charakterizuje *průměr* a *směrodatná odchylka* hodnot každé proměnné. *Párový korelační koeficient* y vs. x_1, y vs. x_2, y vs. x_3 ukazuje na korelaci, všechny tři nezávislé proměnné x_1, x_2, x_3 jsou se závisle proměnnou y spjaty lineární závislostí. *Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných* ukazují na korelaci i mezi nezávisle proměnnými. Lineární vztah existuje mezi x_1 vs. x_2 a u x_1 vs. x_3 a x_2 vs. x_3 vykazuje korelaci. Blok INDIKACE MULTIKOLINEARITY vykazuje ve všech kritériích, že multikolinearita v datech není prokázána.

Prom.	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočtená hladina významnosti
y	6.3700E+01	3.5904E+01	1.0000	-----
x_1	2.8150E+02	1.0017E+02	0.4177	0.230
x_2	2.7980E+02	1.7801E+02	-0.2055	0.569
x_3	2.7400E+01	7.5011E+00	0.3173	0.372

Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		Spočtená hladina významnosti
x_1 versus $x2$:	6.1180E-01	0.060
x_1 versus $x3$:	-4.4304E-01	0.200
x_2 versus $x3$:	-3.2587E-01	0.358

INDIKACE MULTIKOLINEARITY:

j	Vlastní čísla korel. matice l_j	Čísla podmíněnosti K_j	Variance inflation faktor VIF_j	Vícenás. korel. koef. pro X_j
1	3.7065E-01	5.2094E+00	1.7880E+00	0.6639
2	6.9847E-01	2.7645E+00	1.6078E+00	0.6148
3	1.9309E+00	1.0000E+00	1.2517E+00	0.4484

Maximální číslo podmíněnosti K: 5.2094E+00

Nápověda: K[j], K > 1000 indikuje silnou multikolinearitu, VIF[j] > 10 indikuje silnou multikolinearitu.

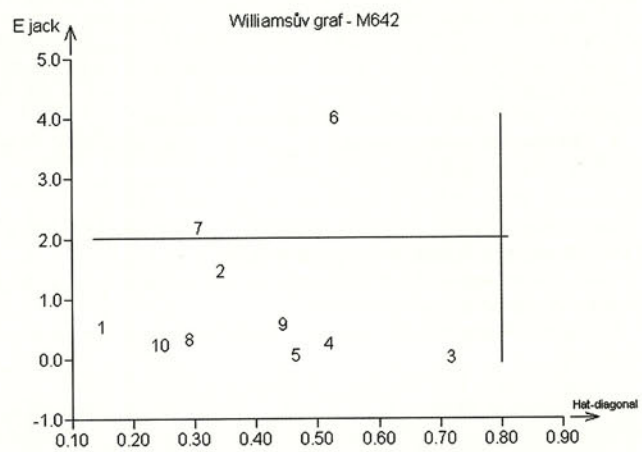
3. Odhadování parametrů: metodou nejmenších čtverců (MNČ) byly nalezeny odhady parametrů β_0, β_1, β_2, β_3.

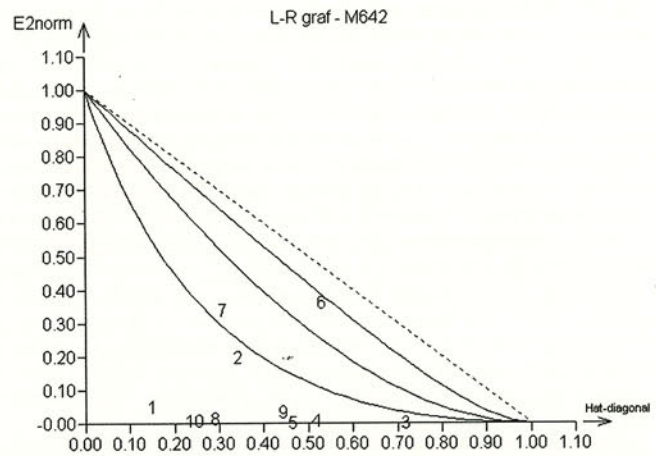
	Odhad	Směrodatná odchylka	Test H_0: $b_j = 0$ vs. H_A: b_j 0 *t*-kriterium	hypotéza H_0 je	Hladina významnosti
b_0	-83.774	42.642	-1.96	Akceptována	0.097
b_1	0.39202	0.09150	4.28	Zamítnuta	0.005
b_2	-0.13839	0.04882	-2.83	Zamítnuta	0.030
b_3	2.7680	1.0224	2.7075	Zamítnuta	0.035

Studentův *t*-test ukázal, že absolutní člen β_0 je statisticky nevýznamný, zatímco ostatní parametry jsou statisticky významné. To je v souladu i s fyzikální interpretací: β_0 se týká zbytkové biologické spotřeby kyslíku BSK_5 za podmínek, když je chemická spotřeba kyslíku CHSK-Cr nulová ($x_1 = 0$) stejně jako jsou nulové i oba obsahy jak rozpuštěných látek RL ($x_2 = 0$) tak i amoniakálního dusíku $N\text{-}NH_3$ ($x_3 = 0$). Tato skutečnost však neodpovídá fyzikální realitě, a proto lze uzavřít, že absolutní člen β_0 nemá fyzikální smysl. Parametr β_1 značí míru, jak hodně ovlivňuje chemická spotřeba kyslíku CHSK-Cr hodnotu biologické spotřeby kyslíku BSK_5. Parametr β_2 značí míru ovlivnění BSK_5 množstvím rozpuštěných látek. Záporné znaménko korelačního koeficientu naznačuje, že čím více bude rozpuštěných látek, tím menší bude hodnota BSK_5. Numericky největší hodnotu má odhad β_3, který značí, že čím více bude ve vodě amoniakálního dusíku $N\text{-}NH_3$, tím větší bude hodnota BSK_5. Statisticky nevýznamný absolutní člen β_0 je nutno ve zpřesněném modelu vynechat.

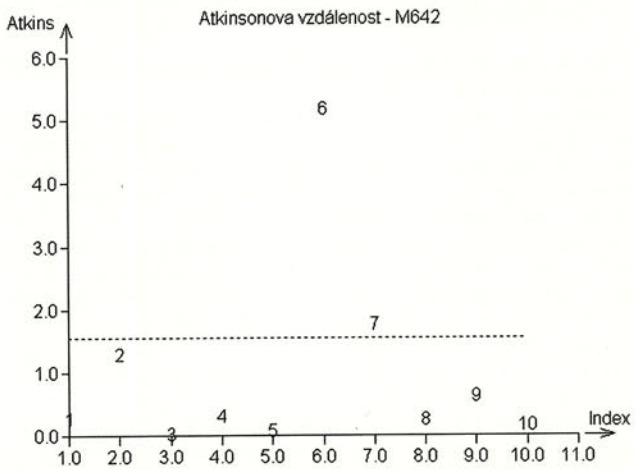
4. Základní statistické charakteristiky: *vícenásobný korelační koeficient* $R = 0.8839$ ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota *koeficientu determinace* $D = R^2 = 78.13\%$ ukazuje, že většina bodů koresponduje s modelem. *Střední kvadratická chyba predikce MEP* = 762.4 a *Akaikovo informační kritérium AIC* = 63.36 se užívají k rozlišení mezi několika navrženými modely. Za optimální se považuje model, pro který dosahuje *MEP* a *AIC* minimální hodnotu.

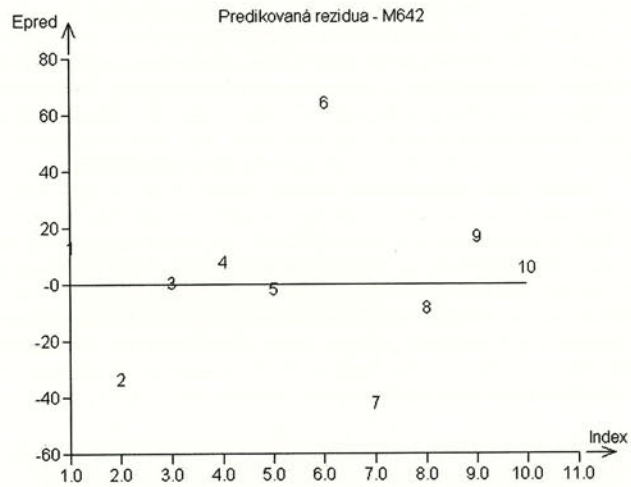
5. Regresní diagnostika: obsahuje pomůcky a postupy pro interaktivní analýzu při (a) kritice dat, (b) kritice modelu, (c) kritice metody, což jsou složky tzv. *regresního tripletu*.



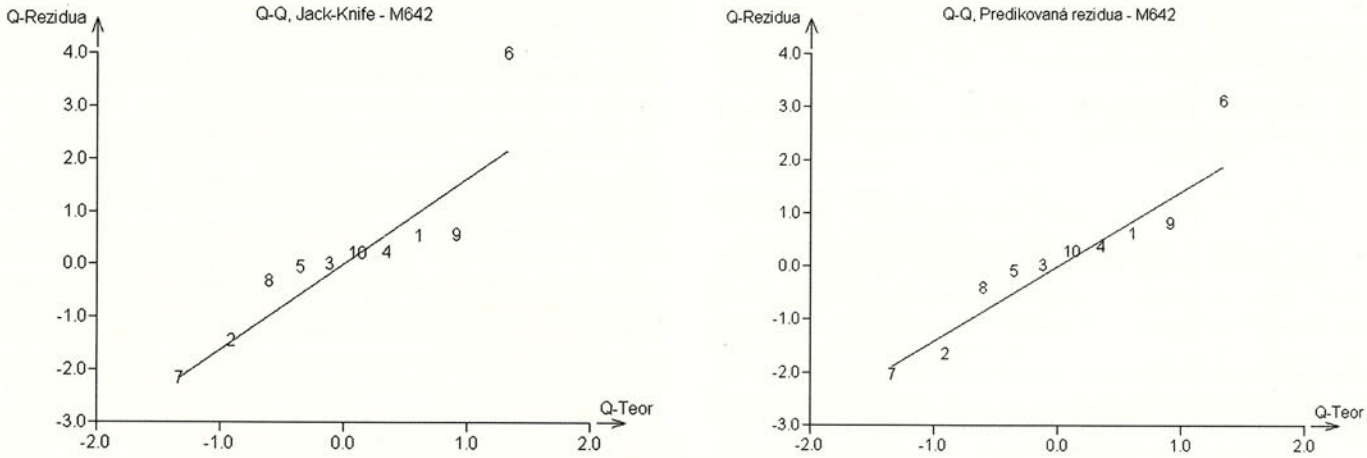


Obr. 9. Grafy vlivných bodů: a) Williamsův graf, b) L-R graf.





Obr. 10. Indexové grafy podezřelých bodů: a) Atkinsonovy vzdálenosti, b) predikovaných reziduí.



Obr. 11. Rankitový Q-Q graf k indikaci normality: a) Jackknife reziduí, b) predikovaných reziduí.

5.1 Kritika dat: skládá se z analýzy několika druhů grafických diagnostik a tabulek různých druhů reziduí.

(a) ***Analýza klasických reziduí*** není příliš spolehlivá, protože klasická rezidua jsou korelovaná, s nekonstantním rozptylem, jeví se normálnější než náhodné chyby (*efekt supernormality)* a nemusí indikovat silně odlehlé hodnoty. Grafická analýza $\hat{e}$ vs. $\hat{y}_P$ je však schopna indikovat podezřelé body, trend, a nekonstantnost podmíněného rozptylu tj. heteroskedasticitu. Míry polohy a rozptýlení klasických reziduí by měly dosahovat hodnot blízkých experimentálnímu šumu. *Odhad směrodatné odchylky s(e)* = 20.6 by se měl blížit svou velikostí experimentální chybě, kterou je zatížena závisle proměnná. *Odhad šikmosti* $\hat{g}_1(e)$ = 0.03 a *odhad špičatosti* $\hat{g}_2(e)$ = 2.85 dokazují normální rozdělení reziduí, normalitu.

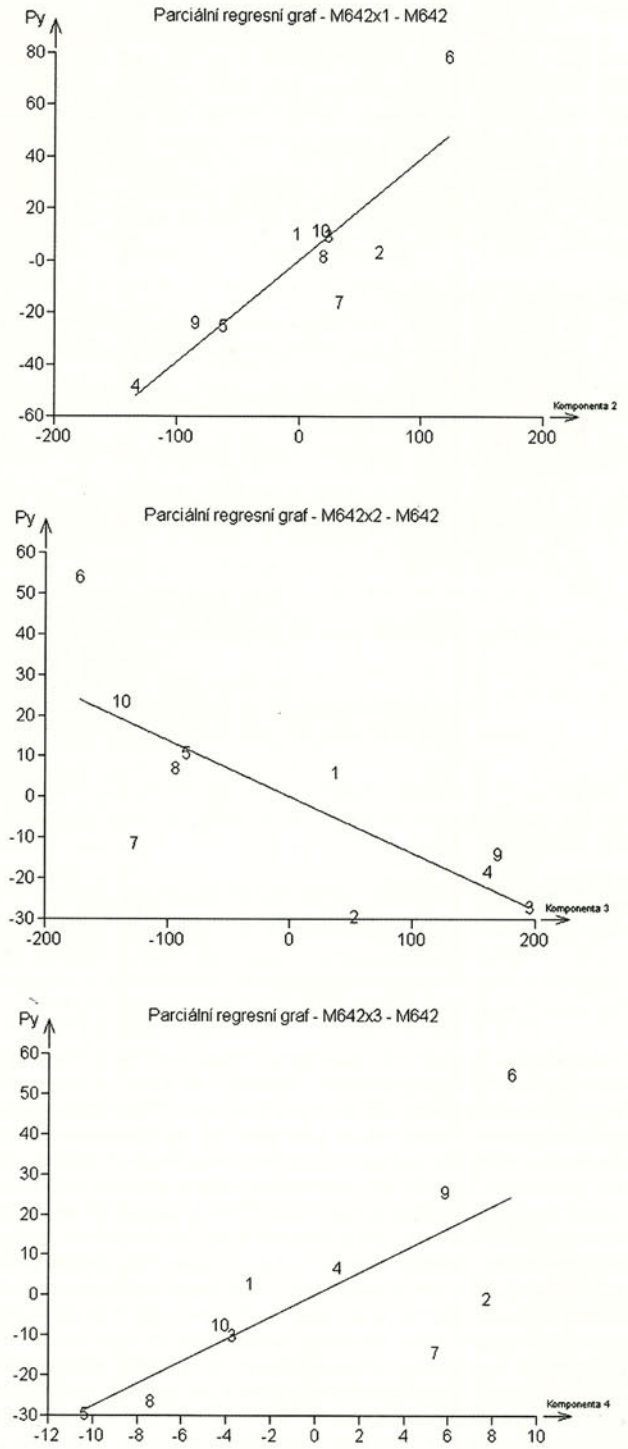
(b) ***Grafy vlivných bodů*** jsou schopny indikovat a současně i testovat, a tak dokazovat přítomnost odlehlých hodnot a extrémů. *Graf predikovaných reziduí* ukazuje na odlehlé body 2, 6 a 7 a mírný extrém 6. *Pregibonův graf* neukazuje na vlivné body. *Williamsův graf* indikuje 6 a 7 jako odlehlé body a žádné extrémy. Konečně *L-R graf* dokazuje odlehlé body 6 a 7 a extrémy 3 a 4, (**obr. 9.**). Lze uzavřít, že body 6 a 7 jsou většinou diagnostik prokázány za odlehlé, a proto je třeba je dále prověřit nebo z výběru vyloučit.

(d) ***Indexové grafy*** upozorňují pouze na podezřelé body. *Andrewsův indexový graf* a *graf normovaných reziduí* ukazují na podezřelé body 6 a 7, které by mohly být vlivné, (**obr. 10.**).

(e) ***Rankitové grafy*** ukazují vedle normality rozdělení dotyčných reziduí i na vlivné (zde odlehlé) body, (**obr. 11.**). *Graf Jackknife reziduí* 6 a 7 jako odlehlé. *Graf predikovaných reziduí* ukazuje na začátku 7 a na konci 6 jako odlehlé body. Po odstranění dvou odlehlých bodů 6 a 7 lze konstatovat, že zbytek dat nevykazuje odchylky od normality.

5.2 Model: *Parciální regresní grafy* nebo obdobně také *parciální reziduální grafy* ukazují na čisté lineární závislosti jednotlivých nezávisle proměnných. Vedle posouzení závislosti navrženého regresního modelu umožňují také indikovat vlivné body, a to 2, 6 a 7, (**obr. 12.**). Navržený model se jeví stran členů $\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ správný, pouze β_0 je nadbytečné.

5.3 Metoda: do této části regresního tripletu patří vyšetření základních předpokladů metody nejmenších čtverců (MNČ), za kterých by tato numerická metoda měla vést k nejlepším nestranným lineárním odhadům regresních parametrů a jsou to: *Fisher-Snedecorův test významnosti regrese* potvrdil, že navržený model je přijat jako významný, protože testační statistika F_{exp}= 7.146 nabývá vyšší hodnoty než kvantil 4.757, jinými slovy: závisle proměnná *y* a nezávisle proměnné x_1, x_2, x_3 jsou v lineární závislosti. *Scottovo testační kritérium multikolinearity* -0.136 ukazuje svou nízkou hodnotou, že navržený model je korektní s ohledem na vazby mezi proměnnými a v datech není multikolenearita. *Cook-Weisbergův test heteroskedasticity* nízkou hodnotou testačního kritéria 2.487 ve srovnání s kvantilem 3.841 dokazuje, že rezidua vykazují homoskedasticitu (konstantnost rozptylu). *Jarque-Berraův test normality reziduí* vykazuje testační kritérium 0.053 podstatně nižší než kvantil 5.991, a tak dokazuje, že klasická rezidua vykazují Gaussovo rozdělení. Nižší hodnota *Waldova testačního kritéria autokorelace* 0.740 vůči kvantilu 3.841 ukazuje, že klasická rezidua nejsou autokorelována. *Znaménkový test* vede na hodnotu testačního kritéria 1.194 nižší



Obr. 12.
Parciální regresní grafy pro proměnnou
a) x_1, b) x_2, c) x_3.

než kvantil 1.960 a tím potvrzuje, že znaménko klasických reziduí se dostatečně střídá a rezidua nevykazují žádný trend. *Graf autokorelace* vykazuje přibližně mrak bodů reziduí, což opět značí důkaz neexistence autokorelace.

6. Konstrukce zpřesněného modelu: Po odstranění bodů 6 a 7 v kritice dat a absolutního členu β_0 v kritice modelu byly nalezeny nové odhady parametrů zpřesněného modelu.

	Odhad	Směrodatná odchylka	Test H_0: $b_j = 0$ vs. H_A: b_j 0 *t*-kriterium	hypotéza H_0 je	Hladina významnosti
b_0	0.0000	0.0000	0.0000		
b_1	0.1982	0.0447	4.43	Zamítnuta	0.007
b_2	-0.0780	0.0356	-2.19	Akceptována	0.080
b_3	0.8800	0.3116	2.83	Zamítnuta	0.037

Zpřesněný model (v závorce je uveden odhad směrodatné odchylky parametru)

$y = 0.198\ (0.045)\ x_1 + 0.880\ (0.312)\ x_3$

je doložen statistickými charakteristikami: *vícenásobný korelační koeficient* $R = 0.8371$, *koeficient determinace* $D = 70.07\%$ dosáhly vesměs dostatečných hodnot. *Střední kvadratická chyba predikce MEP* = 237.53 a *Akaikeho informační kritérium AIC* = 41.46 dosáhly nižších hodnot, což dokazuje lepší model než předešlý.

7. Závěr zhodnocení kvality nalezeného modelu: porovnáním hodnot regresní diagnostiky lze snadno provést zhodnocení *regresního tripletu* dosaženého lineárního regresního modelu pro upravená data, zbavená odlehlých hodnot a upravený regresní model bez absolutního členu. Nalezený model má tvar (v závorce je vždy uveden odhad směrodatné odchylky parametru) $y = 0.198\ (0.045)\ x_1 + 0.881\ (0.312)\ x_3$ čili hodnota biologické spotřeby kyslíku BSK_5 je kladně ovlivněna pouze hodnotou chemické spotřeby kyslíku CHSK-Cr a množstvím amoniakálního dusíku ve vodě $N\text{-}NH_3$.

Doporučená literatura:

[1] M. Meloun, J. Militký: *Statistické zpracování experimentálních dat,* Plus Praha 1994, (1. vydání), East Publishing Praha 1998 (2. vydání), Academia Praha 2004 (3. vydání).

[2] M. Meloun, J. Militký: *Statistické zpracování experimentálních dat - Sbírka úloh,* Univerzita Pardubice 1996.

[3] ADSTAT 1.25, 2.0 a verze 3.0, TriloByte Statistical Software Pardubice, 1992, 1993, 1999.

[4] M. Meloun, J. Militký: *Kompendium statistického zpracování experimentálních dat,* Academia Praha 2002 (1. vydání), Academia Praha 2006 (2. rozšířené vydání).

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Chemickotechnologická fakulta,
Univerzita Pardubice,
Čs. Legií 565, 532 10 Pardubice,
http: //meloun.upce.cz, email: milan.meloun@upce.cz,
telefon: 466037026, fax: 466037068, ICQ: 224-001-003

Computer-Assisted Statistical Data Analysis - 5. Regression Model Building and Testing Using Diagnostic of Regression Triplet (Meloun, M.)

Keywords

Regression triplet - data - regression model - regression method - least-squares - regression diagnostics - residual analysis - influential points - outliers - high-leverages - validation of new method - multivariate regression analysis

Regression models are extremely useful devices for extracting and understanding the essential features of a set of data. These models, however, are nearly always approximate descriptions of more complicated processes. There are a number of common difficulties associated with real datasets. The main one involves the detection and elimination of outliers in the original data. a problem with outliers is that they can strongly influence the model, especially when using least squares criteria, so a several step procedure is required, first to identify whether there are any samples that are atypical of the dataset, then to remove them, and finally to reformulate the model. The method of least-squares LS ensures the acceptable regression model when all three components of *regression triplet* are fulfilled. When some assumptions are not fulfilled, the LS method is inconvenient. Regression diagnostics introduced in this paper represent the procedure for identification of (a) the data quality for a proposed model, (b) the model quality for a given set of data, and (c) fulfilment of all least-squares assumptions. The main difference between the use of regression diagnostics and classical statistical tests is that there is no necessity for an alternative hypothesis, but all types of deviations from an ideal regression triplet are discovered. Our concept of exploratory regression analysis is based on the fact that "the computert user knows more about the data than the computer". The personal computer serves us as an efficient tool for interactive diagnostics of data, model, and estimation method. The procedure of model building with the help of a PC involves interactive co-operation between the user and computer program. Therefore, formal models that do not have physical meaning should not be proposed and analysed. Generally, data quality has a strong influence on any proposed regression model. This paper describes a series of powerful general diagnostics for detecting observations that differ from the bulk of the data These may be individual observations that do not belong to the general model, i. e. influential points or outliers. We think of data as being divided into two classes (1) good observations (the majority of data) reflecting population scatter of data and (2) the outliers (if any) being a part of the so-called influential points. The goal of any outlier detection is to find this true partition and, thus, separate good from outlying observations.

Univerzita spojených národov na Slovensku

Pod hlavičkou vlády Slovenskej republiky sa 13. – 15. mája 2007 zišla na pôde Ministerstva zahraničných vecí SR v Bratislave vedecká komunita zo Slovenska, Česka, Japonska, Namíbie, Nemecka a Holandska, predstavitelia vlády SR a slovenských ministerstiev školstva, zahraničných vecí, poľnohospodárstva a životného prostredia, aby spoločne posúdili možnosti zriadenia **výskumného a vzdelávacieho centra pre zdravú krajinu a ekosystémy Univerzity Organizácie spojených národov (UNU)** na Slovensku. Naliehavosť riešenia tejto problematiky inicioval rektor UNU prof. Hans J. A. van Ginkel. Záujmom UNU je nielen rozšírenie odborného zamerania existujúcej siete 14 univerzít pôsobiacich v rôznych častiach sveta, ale tiež vytvorenie takéhoto pracoviska UNU pre región strednej a východnej Európy.

Odborníci vypracovali širokospektrálne argumenty potreby uvedenej inštitúcie spojených národov a v plnej miere odporučili jej vytvorenie na Slovensku. Základným cieľom centra by malo byť vyhodnocovanie a monitoring kvality krajiny a ekosystémov, hľadanie riešení ako zamedziť postupnému zhoršovaniu ich stavu, ochrana a revitalizácia.

Konferencia o zriadení výskumného a vzdelávacieho centra UNU bola nesporne dobrým miestom pre výmenu vedomostí a skúseností medzi slovenskými a zahraničnými odborníkmi. Začiatok fungovania prestížneho pracoviska UNU sa predpokladá od roku 2008.

- šá -

VODNÍ DÍLA - TBD a.s.®

nabízí odbornou inženýrskou pomoc v oboru bezpečnosti vodních děl i ochrany před povodněmi. Vypracujeme manipulační a provozní řády, povodňové plány objektů, obcí a územních celků, odborné posudky, povodňové a jiné studie. projekty oprav aj.

Ředitelství Praha:
Hybernská 40,110 00 Praha 1
telefon: 221 408 111*
fax: 224 212 803
e-mail: praha@vdtbd.cz

Pracoviště Brno:
Okružní 29a, 638 00 Brno
telefon: 545 222 434
fax: 545 222 642
e-mail: holomeks@vd-tbd.cz

PÖYRY

INŽENÝRSKÁ A PROJEKTOVÁ ČINNOST VE VŠECH OBORECH VODNÍHO HOSPODÁŘSTVÍ

Pöyry Environment a. s.
Botanická 834/56, 602 00 BRNO
tel.: 541 554 111, fax: 541 211 205 E-mail: trade.wecz@poyry.com, www.poyry.cz

Pobočky: Praha, Bezová 1658, 147 14 Praha 4, tel.: 244 062 353
Ostrava, Varenská 49, 701 00 Ostrava, tel.: 596 657 206
Břeclav, Růžičkova 5, 690 39 Břeclav, tel.: 519 322 304
Organizační složka Trenčín, Jesenského 3175, 911 01 Trenčín, tel.: +421 326 522 600