Výstavba a použití regresních modelů v hutní analytice

Prof. RNDr. Milan Meloun, DrSc.,
Univerzita Pardubice, 532 10 Pardubice,
email: milan.meloun@upce.cz

Ing. Roman Lisztwan,
Třinecké železárny, a. s., 739 70 Třinec,
email: roman.lisztwan@trz.cz

Prof. Ing. Jiří Militký, CSc.,
Technická univerzita Liberec, 461 17 Liberec,
email: jiri.militky@vslib.cz

Souhrn: *Při výstavbě regresních modelů se analyzuje **regresní triplet** [data, model, metoda odhadu]. Metoda nejmenších čtverců poskytuje optimální výsledky jenom při splnění předpokladů o datech a o regresním modelu. Pokud předpoklady nejsou splněny, je metoda nejmenších čtverců nevhodná. Regresní diagnostika obsahuje postupy k identifikaci kvality dat pro navržený model, kvality modelu pro daná data, splnění základních předpokladů metody nejmenších čtverců. Průzkumová analýza zde identifikuje 1. nevhodnost dat a malé rozmezí nebo přítomnost vybočujících bodů, 2. nesprávnost navrženého modelu, 3. multikolinearitu, 4. nenormalitu v případě, kdy jsou vysvětlující proměnné náhodné veličiny. Častou úlohou je porovnání několika regresních modelů, zda regresní přímky mají společný průsečík, zda regresní přímky mají společnou směrnici, a zda regresní přímky jsou totožné. Prvním krokem statistické analýzy je vždy odhad parametrů úseku, směrnice a rozptylu regresní přímky pro všechna data zvlášť s využitím metody nejmenších čtverců. Na základě těchto informací se nejdříve ověřuje, zda se rozptyly $\hat{\sigma}_j^2$ významně liší, protože testování tří hypotéz předpokládá konstantnost a totožnost rozptylů ve všech skupinách. Určení odhadů **b** lineárního regresního modelu se zdá na první pohled jednoduchou úlohou. V některých případech, zejména u polynomického modelu, vycházejí často odhady bez fyzikálního smyslu. Regresní křivka sice prochází v těsné blízkosti experimentálních bodů, ale buď mezi nimi silně osciluje u polynomů vysokých stupňů, nebo je systematicky posunutá. Příčiny numerických potíží při počítačovém odhadu parametrů **b** modelu jsou 1. Zanedbání omezené přesnosti počítače při sestavování matice $\boldsymbol{X}^T\boldsymbol{X}$. 2. Nevhodné postupy invertace či řešení soustav lineárních rovnic. 3. Multikolinearita vedoucí ke špatné podmíněnosti matice $\boldsymbol{X}^T\boldsymbol{X}$. 4. Lineární závislost některých sloupců matice $\boldsymbol{X}^T\boldsymbol{X}$, vedoucí k její neinvertovatelnosti z důvodů singularity. V tomto sdělení ukážeme analýzu regresního tripletu při testování shodnosti dvou regresních závislostí a při výstavbě polynomického modelu v hutnických datech.*

Při vyhodnocení regresních modelů se často užívá metody nejmenších čtverců. Tato metoda však ještě nezajišťuje nalezení přijatelného modelu, a to jak ze statistického, tak i z fyzikálního hlediska. Zdrojem problémů jsou složky tzv. ***regresního tripletu*** [data, model, metoda odhadu]. Metoda nejmenších čtverců poskytuje optimální výsledky jenom při současném splnění předpokladů o datech a o regresním modelu. Pokud tyto předpoklady nejsou splněny, je metoda nejmenších čtverců nevhodná. *Regresní diagnostika* obsahuje postupy k identifikaci a) kvality dat pro navržený model, b) kvality modelu pro daná data, c) splnění základních předpokladů metody nejmenších čtverců. V literatuře[5] se pod pojmem regresní diagnostika objevily metody k identifikaci vlivných bodů a multikolinearity. Atkinson[6] do regresní diagnostiky zahrnuje i způsoby navrhování vhodného regresního modelu, a to i s využitím transformace proměnných. Weisberg[7] sem zařazuje soubor speciálních postupů, umožňujících a) ověření předpokladů užitých k odhadu parametrů, b) statistickou analýzu parametrů ("kritika modelu"), c) identifikaci vlivných bodů ("kritika dat").

Mezi základní techniky průzkumové analýzy patří i stanovení volby rozsahu a rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. Přes svoji jednoduchost umožňuje průzkumová analýza identifikovat ještě před vlastní regresní analýzou 1. *nevhodnost dat* (malé rozmezí nebo přítomnost vybočujících bodů), 2. *nesprávnost navrženého modelu* (skryté proměnné), 3. *multikolinearitu*, 4. *nenormalitu* v případě, kdy jsou vysvětlující proměnné náhodné veličiny.

V tomto sdělení ukážeme analýzu regresního tripletu při testování dvou regresních závislostí a při výstavbě polynomického modelu v hutnických datech.

Porovnání regresních přímek

Častou úlohou je porovnání M navržených regresních modelů

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

na základě M skupin experimentálních dat $((x_{ij}, y_{ij}), i = 1, ..., n_j), j = 1, ..., M$. Předmětem testování je, zda a) regresní přímky mají společný průsečík, b) regresní přímky mají společnou směrnici, c) regresní přímky jsou totožné. Prvním krokem statistické analýzy je vždy odhad parametrů b_{2j}, b_{1j} a $\hat{\sigma}_j^2$ pro všechna data zvlášť s využitím metody nejmenších čtverců. Na základě těchto informací se nejdříve ověřuje, zda se rozptyly $\hat{\sigma}_j^2$ významně liší, protože testování hypotéz (a), (b) a (c) předpokládá konstantnost a totožnost rozptylů ve všech skupinách.

Mezi nejpoužívanější testy shody rozptylů patří *Bartlettův test*, který testuje M nezávislých odhadů rozptylu $\hat{\sigma}_j^2$, $j = 1, ..., M$, se stupni volnosti $(n_j - m)$. Testuje se nulová hypotéza H_0: $\sigma_j^2 = \sigma^2$, $j = 1, ..., M$. Tedy u modelů regresní přímky platí $v_j = (n_j - 2)$. Označme

$$V = \sum_{j=1}^{M} v_j , \qquad \hat{\sigma}_c^2 = \sum_{j=1}^{M} \frac{v_j \hat{\sigma}_j^2}{V} , \qquad L = 1 + \frac{\sum_{j=1}^{M} v_j^{-1} - V^{-1}}{3M - 3}$$

Testační kritérium je formulováno vztahem

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

a má při platnosti nulové hypotézy H_0 rozdělení χ^2 s $(M - 1)$ stupni volnosti. Proto pokud je $B < \chi^2_{1-\alpha}(M - 1)$, kde $\chi^2_{1-\alpha}(M - 1)$ je $100(1 - \alpha)$%ní kvantil χ^2 rozdělení, považuje se hypotéza H_0 za přijatelnou a odhadem rozptylu σ^2 je tzv. *sdružený odhad rozptylu* $\hat{\sigma}_c^2$. Bartlettův test je citlivý na odchylky reziduí od normality. K porovnání dvou skupin bodů, $M = 2$, lze testovat shodu dvou

rozptylů dle nulové hypotézy H_0: $\sigma_1^2 = \sigma_2^2$ a pomocí testační statistiky

$$F_2 = \max(\hat{\sigma}_1^2, \hat{\sigma}_2^2)/\min(\hat{\sigma}_1^2, \hat{\sigma}_2^2) ,$$

která má za předpokladu platnosti nulové hypotézy H_0 *F*-rozdělení s (n_1 - 2) a (n_2 - 2) stupni volnosti, pokud je $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$. V opačném případě se pouze mění pořadí stupňů volnosti. Obecně se užívá stupňů volnosti, které byly užity při výpočtu $\hat{\sigma}_i^2$, i = 1, 2.

Test homogenity úseků

Platí-li nulová hypotéza H_0: $\beta_{21} = \beta_{22} = ... = \beta_{2j} = ... = \beta_{2M} = \beta_{2c}$, lze získat ***sdružený odhad*** úseku β_{2c} jako váženou kombinaci odhadů jednotlivých úseků b_{2j} podle vztahu, ve kterém j-tý váhový koeficient w_{Bj} odpovídá odhadu úseku j-té přímky $b_{2c} = \left(\sum_{j=1}^{M} w_{Bj}\ b_{2j} \right) / \left(\sum_{j=1}^{M} w_{Bj} \right)$. Vahový koeficient je dán vztahem $w_{Bj} = \left(n_j \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \right) / \sum_{i=1}^{n_j} x_{ij}^2$. K vlastnímu testování se užívá odhadu rozptylu chyb σ^2 z rozptylu jednotlivých odhadů b_{2j} kolem jejich váženého průměru b_{2c} a z kombinace rozptýlení všech bodů kolem regresní přímky uvnitř jednotlivých skupin dat. Testační statistika má tvar[8]

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

kde $n = \sum_{j=1}^{M} n_j$. Platí-li nulová hypotéza H_0, má testační statistika F_I *F*-rozdělení s (M - 1) a (n - $2M$) stupni volnosti. Rezidua $\hat{e}_{ij}$ jsou určována na základě jednotlivých regresních přímek. Lze psát, že $\sum_{j=1}^{M} \sum_{i=1}^{n_j} \hat{e}_{ij}^2 = \sum_{j=1}^{M} RSC_j$, kde RSC_j je reziduální součet čtverců v j-té skupině. Vyjde-li při testování $F_I < F_{1-\alpha}(M - 1, n - 2M)$, mají na hladině významnosti α všechny přímky stejný úsek a jeho odhad je pak vyčíslen. Rozptyl tohoto úseku se vypočte dle vztahu

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

a odhad úseku b_{2c} má normální rozdělení a je nevychýleným odhadem parametru β_{2c}.

Test homogenity směrnic

Test homogenity směrnic je znám jako test rovnoběžnosti regresních přímek. Platí-li nulová hypotéza H_0: $\beta_{11} = \beta_{12} = ... \beta_{1j} = ... = \beta_{1M} = \beta_{1c}$, lze určit sdružený odhad celkové směrnice β_{1c} jako váženou kombinaci jednotlivých odhadů směrnic b_{1j} vztahem

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

Analogicky jako u testu homogenity úseků lze i zde sestavit testační statistiku

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

která má za předpokladu platnosti nulové hypotézy H_0 *F*-rozdělení s (*M* - 1) a (*n* - *2M*) stupni volnosti. Bude-li proto při testování $F_S < F_{1-\alpha}(M - 1, n - 2M)$, lze považovat regresní přímky na hladině významnosti α za rovnoběžné. Nejlepším odhadem celkové směrnice je b_{1c} a její rozptyl lze odhadnout ze vztahu

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

Platí-li nulová hypotéza H_0, má odhad směrnice b_{1c} normální rozdělení a je nevychýleným odhadem parametru β_{1c}.

Test shody regresních přímek

Test nulové hypotézy H_0: $\beta_{2j} = \beta_{2c}$, $\beta_{1j} = \beta_{1c}$, $j = 1, ..., M$, je vlastně kombinace předchozích testů. Vlastní test spočívá v porovnání reziduálního součtu čtverců RSC_K, získaného po proložení všech *M* skupin dat jedinou společnou přímkou s odhady parametrů b_{1K} a b_{2K}, a reziduálního součtu čtverců $RSC_c = \sum_{j=1}^{M} RSC_j$ z jednotlivých skupin dat odděleně. Testační statistika má tvar

$$F_A = \frac{\frac{RSC_K - RSC_c}{2M - 2}}{\frac{RSC_c}{n - 2M}} \text{ .}$$

Platí-li nulová hypotéza H_0, má testační statistika F_A *F*-rozdělení s (2*M* - 2) a (*n* - *2M*) stupni volnosti. Pokud platí, že $F_A < F_{1-\alpha}(2M - 2, n - 2M)$, je možné na hladině významnosti α považovat všechny ověřované regresní přímky za totožné se společným odhadem úseku b_{2K} a směrnice b_{1K}. Jednotlivé skupiny dat se potom slučují do jednoho společného výběru o velikosti n. V případě, že nulová hypotéza H_0 nebyla prokázána, je obyčejně možné nalézt podskupiny dat, které již jsou homogenní.

Test shody dvou lineárních modelů

Popsaný simultánní test složené hypotézy lze upravit i k testování shody parametrů $\boldsymbol{\beta_1}$ a $\boldsymbol{\beta_2}$ dvou lineárních modelů $\boldsymbol{y_1} = \boldsymbol{X_1}\,\boldsymbol{\beta_1} + \boldsymbol{\varepsilon_1}$, a $\boldsymbol{y_2} = \boldsymbol{X_2}\,\boldsymbol{\beta_2} + \boldsymbol{\varepsilon_2}$. Zde $\boldsymbol{X}_1$ je matice rozměru $(n_1 \times m)$, $\boldsymbol{y_1}$ je vektor rozměru $(n_1 \times 1)$, $\boldsymbol{X}_2$ je matice rozměru $(n_2 \times m)$ a $\boldsymbol{y_2}$ je vektor rozměru $(n_2 \times 1)$. Označme RSC_1 reziduální součet čtverců prvního modelu, RSC_2 reziduální součet čtverců druhého modelu a RSC reziduální součet čtverců odpovídající modelu složenému

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \end{bmatrix} .$$

Chowův test hypotézy H_0: $\beta_1 = \beta_2$ proti alternativní H_A: $\beta_1 \neq \beta_2$ je založen na testačním kritériu

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

kde $n = n_1 + n_2$. Za předpokladu shodných rozptylů obou výběrů (homoskedasticity), $\sigma_1^2 = \sigma_2^2$, má statistika F_c pak F-rozdělení s m a $(n - 2\,m)$ stupni volnosti. Pokud však nejsou rozptyly obou souborů shodné (heteroskedasticita), $\sigma_1^2 \neq \sigma_2^2$, užije se Fisherovo-Snedecorovo rozdělení, ale s m a r stupni volnosti, kde

$$r = ([(n_1 - m)\,\sigma_1^2 + (n_2 - m)\,\sigma_2^2]^2)/((n_1 - m)\,\sigma_1^4 + (n_2 - m)\,\sigma_2^4)$$

dle cit.[9]. Místo testačních statistik lze použít k ověření linearity i všech charakteristik umožňujících porovnání vhodnosti různých modelů. Mezi často užívané charakteristiky patří *střední kvadratická chyba predikce* $MEP = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^T\,\boldsymbol{b_{(i)}})^2$, kde $\boldsymbol{b_{(i)}}$ je odhad parametrů regresního modelu určený ze všech bodů kromě i-tého a $\boldsymbol{x}_i$ je i-tý řádek matice $\boldsymbol{X}$. Statistika MEP využívá predikce $\hat{y}_{P,i}$ z odhadu, při jehož konstrukci byla informace o i-tém bodu vypuštěna. Užije-li se charakteristiky MEP místo RSC ve výpočtu koeficientu determinace, bude výsledkem *predikovaný koeficient determinace*

$$\hat{R}_P^2 = 1 - [n\,MEP\,/(\sum_{i=1}^{n} y_i^2 - n\,\bar{y})] \quad .$$

Univerzální použití mají také rozličná kritéria vycházející z teorie informace a entropie[12]. Mezi nejznámější patří *Akaikovo informační kritérium*

$$AIC = n\,\ln(RSC/n) + 2\,m \ .$$

Za nejvhodnější je považován takový model, pro který je AIC minimální.

Některé problémy ve výstavbě lineárního regresního modelu

Určení odhadů $\boldsymbol{b}$ lineárního regresního modelu se zdá na první pohled jednoduchou úlohou. Zejména jsou-li v knihovně programů k dispozici podprogramy pro maticové operace, je formální řešení snadné. Problémy vznikají, když se matice $\boldsymbol{X}^T\,\boldsymbol{X}$ jeví z hlediska strojové přesnosti a užitého algoritmu jako singulární. V některých případech, zejména u polynomického modelu, vycházejí často odhady bez fyzikálního smyslu. Regresní křivka sice prochází v těsné blízkosti experimentálních bodů, ale buď mezi nimi silně osciluje (u polynomů vysokých stupňů), nebo je systematicky posunutá. Příčiny numerických potíží při počítačovém odhadu parametrů $\boldsymbol{b}$ jsou

1. Zanedbání omezené přesnosti počítače při sestavování matice $\boldsymbol{X}^T \boldsymbol{X}$.
2. Nevhodné postupy invertace či řešení soustav lineárních rovnic.
3. Multikolinearita vedoucí ke špatné podmíněnosti matice $\boldsymbol{X}^T \boldsymbol{X}$.
4. Lineární závislost některých sloupců matice $\boldsymbol{X}^T \boldsymbol{X}$, vedoucí k její neinvertovatelnosti z důvodů singularity.

Kvalitní programy lineární regrese překonávají tyto obtíže a poskytují řešení vždy. Mezi nejefektivnější patří algoritmy, které nesestavují matici $\boldsymbol{X}^T \boldsymbol{X}$, ale řeší přeurčenou soustavu n lineárních rovnic o m neznámých $\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{b}$. Příkladem je algoritmus SVD (singular value decomposition)[10], který pracuje i na počítači s malou přesností zobrazení dat. Z řady technik numerického řešení úlohy nejmenších čtverců se v omezíme na dva případy:

1. **Metodu ortogonálních funkcí OF**, která je jednoduchá a vhodná pro polynomické modely.
2. **Metodu racionálních hodností RH**, která je užita v programu ADSTAT a bude použita i zde. Přehled dalších algoritmů je obsažen v práci[11].

Metoda racionálních hodností

K identifikaci špatné podmíněnosti matice $\boldsymbol{X}^T \boldsymbol{X}$ nebo její standardizované formy $\boldsymbol{R}$ se využívá rozklad na vlastní čísla a vlastní vektory. Jelikož je matice $\boldsymbol{R}$ symetrická, lze ji vyjádřit pomocí vlastních čísel $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_m$ a odpovídajících vlastních vektorů $\boldsymbol{P}_j$, $j = 1, \ldots, m$, ve tvaru

$$\boldsymbol{R} = \sum_{j=1}^{m} \lambda_j \boldsymbol{P}_j \boldsymbol{P}_j^T \quad \text{a inverzní matici} \quad \boldsymbol{R}^{-1} = \sum_{j=1}^{m} \lambda_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^T ,$$

která se přepíše do tvaru $b_N = \sum_{j=\omega}^{m} \left(\lambda_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^T\right) \boldsymbol{r}$. Kovarianční matici normovaných odhadů lze zapsat ve tvaru $D(\boldsymbol{b}_N) = \hat{\sigma}_N^2 \sum_{j=\omega}^{m} \lambda_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^T$. V případě metody nejmenších čtverců se volí $\omega = 1$. Z obou rovnic plyne, že pokud budou vlastní čísla λ_j malá, budou odhady b_N i jejich rozptyly neúměrně vysoké. Podle velikosti vlastních čísel λ_j se dělí regresní úlohy do tří skupin:

I. Všechna vlastní čísla jsou výrazně vyšší než nula. Použití metody nejmenších čtverců v tomto případě nečiní žádné obtíže.

II. Některá vlastní čísla jsou blízká nule. Jde o typický příklad multikolinearity, kdy běžné metody zcela selhávají.

III. Některá vlastní čísla jsou rovna nule. Pak je matice $\boldsymbol{X}^T \boldsymbol{X}$ nebo $\boldsymbol{R}$ singulární a nelze ji proto invertovat.

Odstranění problémů skupin II. a III. je možno docílit užitím metody racionálních hodností, kdy se zanedbají sčítance (resp. jejich části) o malých hodnotách vlastních čísel λ_j, cit[12]. Kritériem pro vypuštění sčítanců, kterým odpovídají příliš malá vlastní čísla, je $abs\left(\sum_{j=1}^{\omega} \lambda_j / \sum_{j=1}^{m} \lambda_j\right) = P$, kde P je zvolená přesnost (obyčejně 10^{-5}). Číslo ω určuje také spodní mez, od které se provádí sčítání. Označme $W = \sum_{j=1}^{\omega} \lambda_j$ a $E = \sum_{j=1}^{m} \lambda_j$. Pokud vyjde W/E > P (tj. ω by mělo být necelé), provádí se sumace od (ω - 1) a vlastní číslo $\lambda_{\omega-1}$ se "váží" faktorem

$u = (W - E P)/\lambda_\omega$. Tím je zajištěno, že lze spojitě v závislosti na růstu přesnosti *P* snižovat délku odhadů $\|\boldsymbol{b}_N\|$ a jejich rozptyly. To je však doprovázeno růstem ***vychýlení odhadů*** a poklesem vícenásobného korelačního koeficientu. Vychýlení odhadů je zde způsobeno zanedbáním sčítanců při $\omega > 1$.

Optimální velikost *P* je možné určit z požadavku minima střední kvadratické chyby predikce *MEP*. V programu ADSTAT si uživatel přesnost *P* volí, nebo je standardně deklarována $P = 10^{-32}$.

Ilustrační úloha 1. *Porovnání regresních přímek závislosti obsahu uhlíku OES a Leco*

Při výrobě automatových ocelí dané jakosti byla porovnávána závislost obsahu uhlíku v posledním zkušebním vzorku, odebraném z mezipánve na ZPO a analyzovaném termoevoluční metodou na Leco analyzátoru s obsahem uhlíku v předposledním zkušebním vzorku, odebraném na vakuovací stanici a analyzovaném na automatickém analyzátoru OES a Leco analyzátoru. Mezi oběma odběrovými místy již nedocházelo k úpravě chemického složení. Uhlík v posledním a předposledním zkušebním vzorku je analyzován na rozdílných Leco analyzátorech. Cílem experimentu bylo ověřit, zda se obě varianty stanovení uhlíku, tj. na OES a Leco1, v předposlední zkoušce liší a zda jsou ve shodě se stanovením uhlíku na Leco2 v poslední zkoušce.

Data: Data1: Leco2, uhlík v poslední zkoušce *x* [%], Leco1, uhlík v předposlední zkoušce y_1 [%]:
0.052 0.056, 0.045 0.053, 0.047 0.053, 0.048 0.054, 0.047 0.051, 0.061 0.061,
0.055 0.056, 0.061 0.065, 0.054 0.060. 0.059 0.064, 0.053 0.055, 0.049 0.049,
0.046 0.052, 0.046 0.049, 0.065 0.070. 0.057 0.060. 0.062 0.064, 0.066 0.070,
0.064 0.072, 0.059 0.066, 0.067 0.073, 0.066 0.072, 0.060 0.067, 0.054 0.057,
0.054 0.058, 0.055 0.055, 0.052 0.060

Data 2: Leco2, uhlík v poslední zkoušce *x* [%], OES, uhlík v předposlední zkoušce y_2 [%]:
0.052 0.053, 0.045 0.050. 0.047 0.051, 0.048 0.051, 0.047 0.049, 0.061 0.061,
0.055 0.061, 0.061 0.066, 0.054 0.059, 0.059 0.065, 0.053 0.053, 0.049 0.048,
0.046 0.046, 0.046 0.049, 0.065 0.068, 0.057 0.060. 0.062 0.064, 0.066 0.070.
0.064 0.068, 0.059 0.066, 0.067 0.071, 0.066 0.072, 0.060 0.062, 0.054 0.057,
0.054 0.054, 0.055 0.059, 0.052 0.058

Program: ADSTAT 2.0: Lineární regrese

Řešení:

1) Testování úseku a směrnice: Metodou nejmenších čtverců byly určeny odhady parametrů úseků a směrnic a zároveň určeny jejich 95%ní intervaly spolehlivosti pro oba modely regresních přímek.

Tabulka 1. Odhadnuté 95%ní intervaly úseku a směrnice přímek

	Úsek		Směrnice	
Model	L_D	L_H	L_D	L_H
Leco1	-0.00439	0.01205	0.86322	1.15618
OES	-0.00868	0.00559	0.95849	1.21271

1. Jelikož intervaly spolehlivosti úseku obou regresních přímek obsahují nulu, lze úseky považovat za nulové.
2. Jelikož intervaly spolehlivosti směrnic obou regresních přímek obsahují jedničku, lze směrnice obou přímek považovat za jednotkové.

1. Identifikace vlivných bodů: byla provedena pomocí grafů vlivných bodů.

a) *Graf predikovaných reziduí,* osa *x:* e_{Pi}, osa *y:* e_i

Data 1: všechny body leží na přímce rovnoměrně rozmístěné, a tudíž jsou bez odlehlých

bodů a extrémů.

data 2: bez odlehlých bodů a extrémů.

b) *Pregibonův graf*, osa *x*: prvky H_{ii}, osa *y:* e_{Ni}

Data 1: všechny body leží pod spodní přímkou, a tudíž odlehlé body a extrémy nejsou identifikovány.

Data 2: bez odlehlých bodů a extrémů.

c) *Williamsův graf*, osa *x:* prvky H_{ii}, osa *y:* e_{Ji}

Data 1: body 6, 12, 26 leží na horní testační osou a proto jsou odlehlé.

Data 2: body 6, 12, 20 leží na horní testační osou a proto jsou odlehlé.

d) *L-R graf* , osa *x:* H_{ii}, osa *y:* e^2_{Ni}

Data 1: všechny body leží pod spodní izolinii, a proto bez odlehlých bodů a extrémů.

data 2: bez odlehlých bodů.

e) *Graf Atkinsonovy vzdálenosti,* osa *x*: index *i*, osa *y*: $|\hat{e}_{ji}| \sqrt{\frac{n-m}{m} \frac{H_{ii}}{1-H_{ii}}}$.

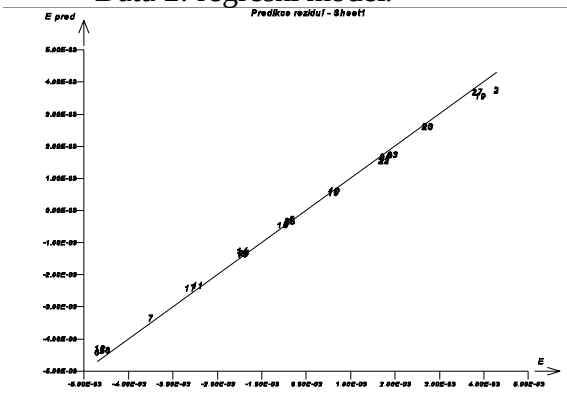
Data 1: indikuje jediný odlehlý bod 2.

Data 2: bez odlehlých bodů.

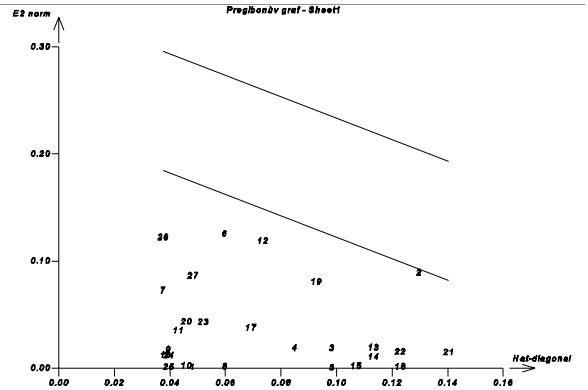
f) *Rozptylový graf regresního modelu,* osa *x:* hodnoty Leco 2, osa *y:* hodnoty Leco1 nebo OES.

Data 1: regresní model.

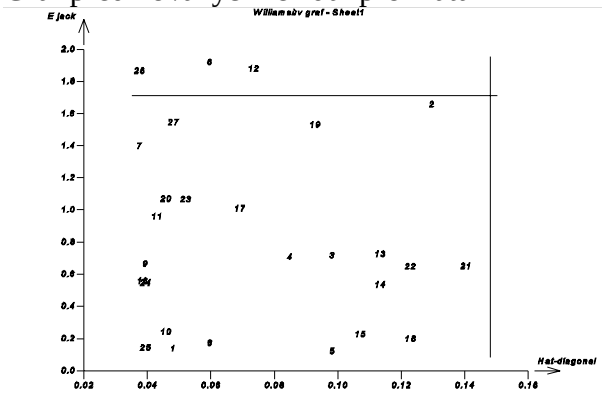
Data 2: regresní model.



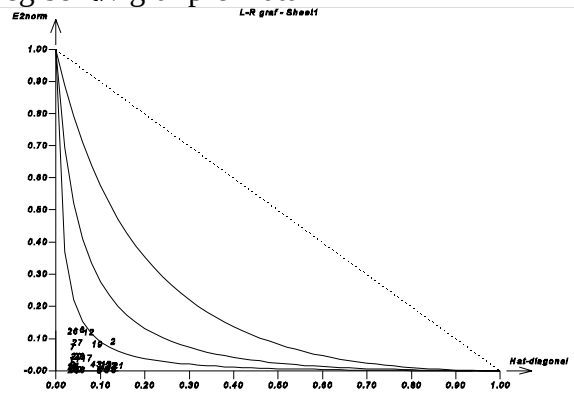
Graf predikovaných reziduí pro Data 1



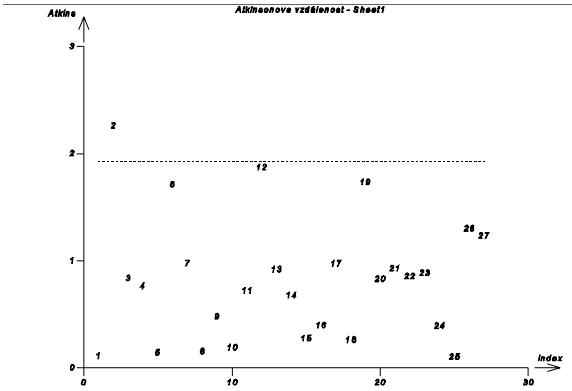
Pregibonův graf pro Data 1



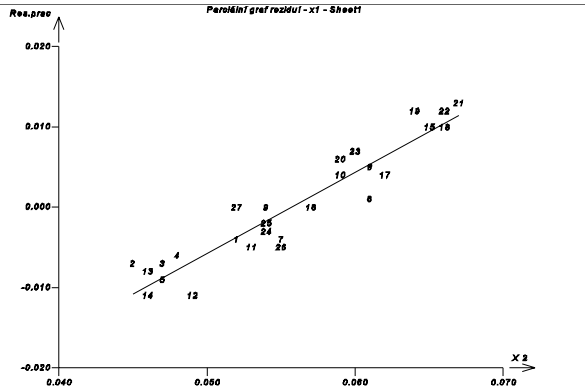
Williamsův graf pro Data 1



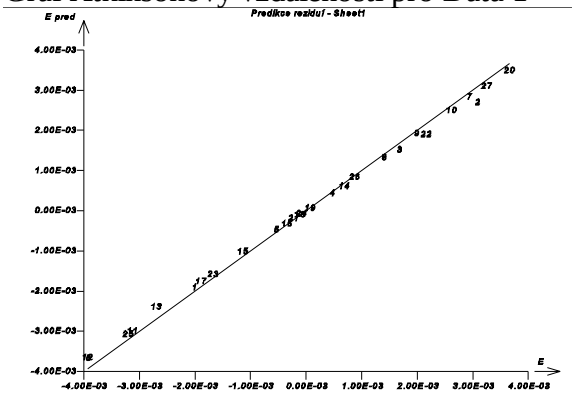
L-R graf pro Data 1



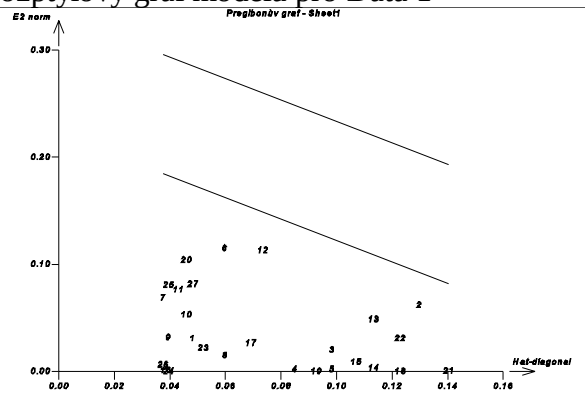
Graf Atkinsonovy vzdálenosti pro Data 1



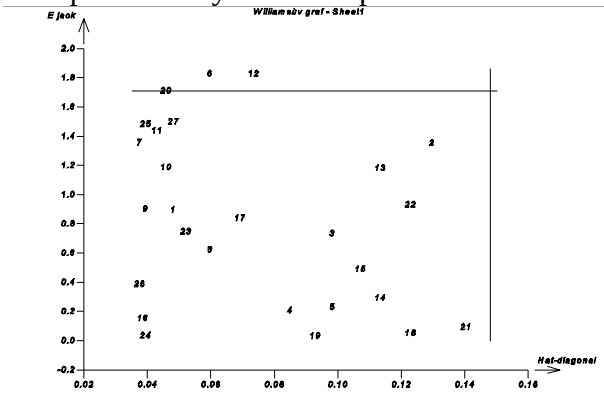
Rozptylový graf modelu pro Data 1



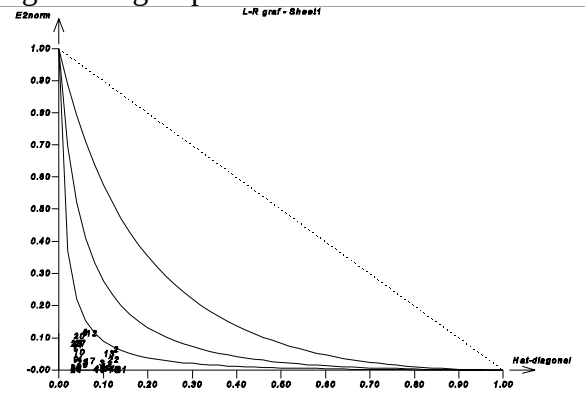
Graf predikovaných reziduí pro Data 2



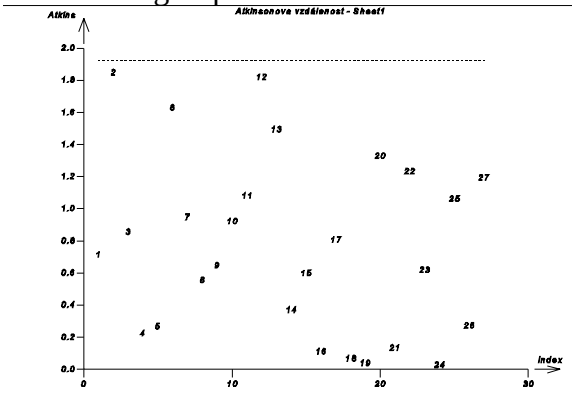
Pregibonův graf pro Data 2



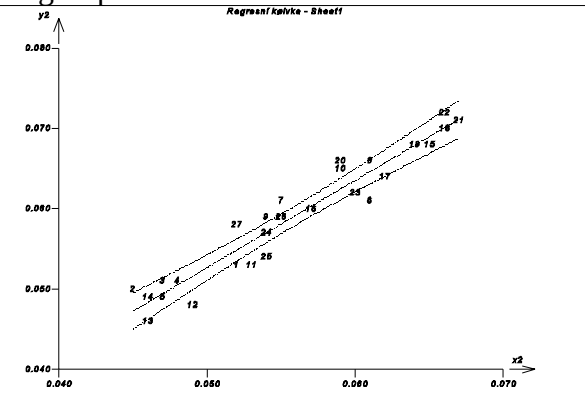
Williamsův graf pro Data 2



L-R graf pro Data 2



Graf Atkinsonovy vzdálenosti pro Data 2



Rozptylový graf modelu pro Data 2

Závěr: Na základě analýzy regresního tripletu můžeme tvrdit, že navržené modely regresních přímek jsou správné.

2. Test shodnosti dvou přímek:

Cílem testování je ověřit zda regresní přímky a) mají společný úsek čili průsečík, b) mají společnou směrnici, a c) jsou totožné. Před samotným testováním hypotéz a), b) a c) je nutno ověřit shodnost a konstantnost rozptylu ve všech skupinách. K tomu se využívá Bartletova testu, kdy se testuje nulová hypotéza H_0: $\sigma_j^2 = \sigma^2$, $j = 1, ..., M$. Pokud je $B < \chi^2_{1-\alpha}(M - 1)$, kde $\chi^2_{1-\alpha}(M - 1)$ je $100(1 - \alpha)$%ní kvantil χ^2 rozdělení, pak je σ^2 považovat za tzv. sdružený odhad rozptylu σ^2_c.

ad a) K testování homogenity úseků se využívá testovací statistiky F_I. Vyjde-li při testování $F_I < F_{1-\alpha}(M - 1, n - 2M)$, mají na hladině významnosti α přímky stejný úsek .

ad b) K testování homogenity směrnic se využívá testovací statistiky F_S, obdobně jako u testu homogenity úseku. Bude-li proto při testování $F_S < F_{1-\alpha}(M - 1, n - 2M)$, lze považovat regresní přímky na hladině významnosti α za rovnoběžné.

ad c) Tento test spočívá v porovnání reziduálního součtu čtverců RSC_K, který se získá proložením všech skupin dat jedinou přímkou a reziduálního součtu čtverců RSC_c. Pokud platí, že $F_A < F_{1-\alpha}(2M - 2, n - 2M)$, je možné na hladině významnosti α považovat všechny ověřované regresní přímky za totožné se společným odhadem úseku b_{2K} a směrnice b_{1K}.

Tabulka 2. Vyčíslené odhady pro testy shodnosti úseků, směrnic a totožnosti dvou přímek

Varianta	*Úsek* b_{0j}	*Směrnice* b_{1j}	*Test* b_{0j}	*Test* b_{1j}	$s(b_{0j})$	$s(b_{1j})$	RSC_j	$s(e)$	*o*	*e*
1. Leco1	0.00383	1.0097	A	Z	0.00399	0.07111	0.00016	0.0025	0	0
2. OES	-0.0015	1.0856	A	Z	0.00346	0.06170	0.00012	0.0022	0	0
1. + 2.	0.00114	1.0476	A	Z	0.00269	0.04793	0.00030	0.0024	0	0

Tabulka 3. Testy shodnosti úseků, směrnic a totožnosti dvou přímek: (a) Homogenita rozptylu

Homogenita rozptylu			
$\chi^2_{1-\alpha}(M - 1)$	B	$F_{1-\alpha}(M - 1, n - 2M)$	F_2
3.84	0.49	3.18	1.33
Závěr testu: Rozptyly jsou shodné			

Tabulka 3. Testy shodnosti úseků, směrnic a totožnosti dvou přímek: (b) Shodnost přímek

Test shody úseků		Test shody směrnic		Test shody přímek	
$F_{1-\alpha}(M - 1, n - 2M)$,	F_I	$F_{1-\alpha}(M - 1, n - 2M)$	F_S	$F_{1-\alpha}(2M - 2, n - 2M)$	F_A
4.03	1.04	4.03	0.65	3.18	1.95
Závěr testu: Úseky jsou shodné		Směrnice jsou shodné		Přímky jsou shodné	

Závěr: Pomocí grafů vlivných bodů nebyly nalezeny žádné významné odlehlé body ani extrémy, Jednotlivé varianty stanovení uhlíku (OES a Leco1) se statisticky významně neliší od stanovení uhlíku na Leco2. Testováním shodnosti obou regresních přímek vyšly testy shody úseků a směrnic

pozitivně, stejně tak i test shody dvou regresních přímek. Obě varianty stanovení uhlíku nelze považovat za statisticky významně odlišné.

Model Leco1: r = 0.9432, D = 88.97%, 0 odlehlých hodnot, 0 extrémů.

$$y = 0.00383\ (0.00399) + 1.0097\ (0.07111)\ x,$$

Model OES: r = 0.9619, D = 92.53%. 0 odlehlých hodnot, 0 extrémů.

$$y = -0.00155\ (0.00346) + 1.0856\ (0.06170)\ x.$$

Ilustrační úloha 2. *Polynomická závislost průtoku argonu jiskřištěm na vstupním tlaku Ar v OES automatickém analyzátoru ocelí*

Byla sledována závislost průtoku argonu v l/min. jiskřištěm na vstupním tlaku argonu v [Bar] během analýzy zkušebních vzorků ocelí na automatickém analyzátoru OES. V průběhu analýzy zkušebních vzorků (~40s) protéká jiskřištěm argon, který vytváří v prostoru mezi analyzovaným vzorkem a elektrodou inertní atmosféru. Metodou nejmenších čtverců MNČ a racionálních hodností RH byl hledán optimální stupeň polynomu a testována statistická významnost jednotlivých parametrů polynomu.

Data: tlak argonu x [Bar], průtok argonu y [l/min]

0.5 1.10, 0.6 1.3, 0.7 1.5, 0.8 1.7, 0.9 2.00, 1.0 2.20, 1.1 2.45, 1.2 2.70, 1.3 2.85, 1.4 3.1, 1.5 3.3, 1.6 3.5, 1.7 3.65, 1.8 3.85, 1.9 3.95, 2.0 4.07,

Program: ADSTAT 2.0: Lineární regrese

Řešení: Metoda racionálních hodností RH se používá v případě, kdy vlastní čísla matice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ jsou blízká nule nebo rovna nule, svědčící o silné multikolinearitě. Pak je matice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ singulární a metodou nejmenších čtverců MNČ nelze odhady neznámých parametrů $\boldsymbol{\beta}$ určit. V těchto případech se zanedbávají sčítance o malých hodnotách vlastních čísel β_j. Kritériem vypuštění je zadávaný parametr přesnost P. Optimální hodnota P se určí z minima střední kvadratické chyby predikce MEP. Růst P je doprovázen růstem vychýlení odhadů a poklesem r. Odhady určené metodou racionálních hodností RH jsou sice vychýlené, ale přesnější a zajišťují, že průběh modelu odpovídá trendům dat. V tabulce 4 jsou uvedeny všechny významné parametry obou metod při různých stupních regresní závislostí. Optimální hodnota P pro metodu RH byla určena 0.027 a dále byl nalezen 4. stupeň polynomu, pro který je hodnota MEP minimální. Testováním regresního tripletu vyšly všechny testy přijatelně pro 4. stupeň polynomu. Na základě t-testů vycházejí na hladině významnosti $\alpha = 0.05$ všechny parametry statisticky významné až na parametr β_3.

Tabulka 4. Hledání nejlepších odhadů parametrů při výstavbě polynomického modelu: A značí, že parametr není statisticky významný, Z značí statisticky významný.

	m = 1		m = 2		m = 3		m = 4		m = 5	
	MNČ	RH	MNČ	RH	MNČ	RH	MNČ	RH	MNČ	RH
Nejlepší odhady parametrů polynomu pro zadaný stupeň polynomu										
b_0	0.119A	0.119A	-0.354Z	0.333Z	0.296Z	-0.121Z	0.54 Z	0.046A	1.76A	0.173Z
b_1	2.066Z	2.066Z	2.942 Z	2.902Z	1.047Z	2.268Z	0.10 A	1.892Z	-5.99A	1.663Z
b_2			-0.350Z	-0.335Z	1.299Z	0.233Z	2.60 A	0.343Z	13.98A	0.358Z
b_3					-0.440Z	-0.155Z	-1.18A	-0.003 A	-11.21A	0.050Z
b_4							0.15 A	-0.07Z	4.35A	-0.02Z

b_5									-0.672A	-0.03Z
Odhady směrodatné odchylky parametrů polynomu										
$s(b_0)$	0.059	0.059	0.084	0.083	0.126	0.050	0.364	0.032	1.028	0.042
s(b1)	0.045	0.045	0.145	0.144	0.352	0.065	1.395	0.035	5.002	0.066
s(b2)			0.057	0.057	0.300	0.003	1.867	0.004	9.186	0.019
s(b3)					0.080	0.012	1.045	0.003	8.004	0.011
s(b4)							0.208	0.003	3.329	0.002
s(b5)									0.532	0.003
Statistické charakteristiky regrese a rozhodčí kritéria k výstavbě regresního modelu										
R	0.997	0.997	0.999	0.999	1.000	1.000	1.000	1.000	1,000	1,000
R^2	0.994	0.994	0.998	0.998	1.000	0.999	1,000	1.000	1,000	0,999
R_p^2	0.995	0.995	0.998	0.998	1.000	0.999	1.000	1.000	0,999	1,000
MEP	0.008	0.008	0.003	0.003	0.001	0.002	0.001	0.001	0,001	0,001
AIC	-78.1	-78.1	-97.7	-97.6	-116.0	-104,3	-114.7	-111,6	-115.1	-107.5
RSC	0.095	0.095	0.025	0.025	0.007	0.014	0.007	0.008	0,006	0.009
s(e)	0,082	0.082	0.043	0.044	0.024	0.035	0.025	0.027	0,024	0.030

Závěr: Byl nalezen stupeň polynomu $m = 4$. Metodou RH byly získány nejlepší odhady parametrů, které zajišťují optimální průběh polynomického modelu experimentálními daty než metoda MNČ.

Poděkování: Předložený projekt byl vypracován za finační podpory Vědeckého záměru MŠMT č. MSM253100002.

Doporučená literatura

[1] Draper N. R. a Smith H.: *Applied Regression Analysis*. 2nd Ed., Wiley, New York 1981.
[2] Seber G. A. F.: *Linear Regression Analysis*. Wiley, New York 1977.
[3] Guttman I.: *Linear Models - An Introduction*. Wiley, New York 1982.
[4] Searle S. R.: *Linear Models*. Wiley, New York 1971.
[5] Belsey D. A., Kuh E. a Welsch R. E.: *Regression Diagnostics*. Wiley New York 1980.
[6] Atkinson A. C.: *Plot, Transformation, Regression*. Claredon Press, Oxford 1986.
[7] Weisberg S.: *Technometrics* **25**, 219 (1983).
[8] Green J. R. a Margerison D.: *Statistical Treatment of Experimental Data*. Elsevier, Amsterdam 1978.
[9] Utts J.: *Commun. Statist.* **11**, 2801 (1982).
[10] Nash J. C.: *Compact Numerical Algorithms for Computer*, A. Hilger, Bristol, 1979.
[11] Lawson Ch. a Hanson R.: *Solving Least-Squares Problems*. Englewood Cliffs, New Jersey, 1974.
[12] Marquardt D. M.: *Technometrics* **12**, 591 (1970).
[13] Rice J. A.: *Mathematical Statistics and Data Analysis*, Wadsworth & Brooks, California 1988.
[14] Cyhelský L. a kol.: *Úlohy k základům statistiky*, SNTL Praha 1988.

[15] Potocký R. a kol.: *Zbierka úloh z pravdepodobnosti a matematickej štatistiky*, ALFA Bratislava 1986.
[16] Kleinbaum D. G. a kol.: *Applied Regression Analysis and Other Multivariate Methods,* PWS-KENT Publishing Comp., Boston, 1988.
[17] Ebel S., G. Herold: *Z. Anal. Chem.* **270**, 20 (1974).
[18] Anderson R. L.: *Practical Statistics for Analytical Chemists,* van Nostrand Reinhold Company, New York 1987.
[19] Miller J. C., Miller J. N.: *Statistics for Analytical Chemistry,* Ellis Horwood, Chichester, 1984.
[20] Meloun M., Militký J.: *Statistické zpracování experimentálních dat,* East Publishing, Praha 1998.