Výber vhodných regresných modelov z množiny existujúcich: kvalitatívne, kvantitatívne, grafické a logické kritériá

Ľubica Floreková [1] *a Marta Benková* [1]

A selection of proper regression models from their existing set: qualitative, quantitative, graphical and logical criteria

The contribution presents an approach to the solution of the problem of processing experimental data of various origin using methods of regression and correlation analysis for two- and threedimensional relations between variables. It concentrates on calculation procedures, based on the lastsquare method and other possibilities of obtaining continual information about the quality of processed data as well as of resultant regression models.

***Key words**: regression model, dependent/independent variable, korrelation.*

Úvod

V technickej, ekonomickej, sociologickej praxi je veľmi častá situácia, že sa na danom objekte, systéme, sleduje dvojica alebo k-tica premenných, o ktorých na základe apriórnych informácií vieme, resp. predpokladáme ich vzájomný vzťah, súvislosť, ovplyvňovanie sa. Môžeme potom s využitím vhodného matematicko - štatistického aparátu vyjadriť závislosť medzi premennými pomocou *regresného modelu* a silu, tesnosť väzby medzi premennými pomocou hodnotenia korelácie. Daná úloha má teda dve časti:

1. Vytvorenie skupiny (možných) vhodných regresných modelov - RM.
2. Hodnotenie regresných modelov pomocou korelačných charakteristík a výber najvhodnejšieho regresného modelu.

Tvorba regresných modelov

Regresný model vyjadruje vzťah pre závislú - vysvetľovanú premennú Y pomocou k-tice nezávislých - vysvetľujúcich premenných $x_j, j = 1,\ldots,k$, na základe znalosti (disponibilnosti) empirických súborov dát $\{y_i, x_{ij}\} i = 1,\ldots,n$.

V prípade $j = 1$, hovoríme o jednoduchom, dvojrozmernom RM, v prípade $j = 2,\ldots,k$ jedná sa o model viac/mnohorozmerný.

Za optimálnu transformáciu teoretického/hľadaného vyjadrenia závislosti medzi premennými považujeme jeho tvar

$$Y_{opt}(X) = f(x) \quad , \text{ resp.} \quad Y_{opt}(X) = f(x_j),$$

ktorý vyhovuje najrozšírenejšej Gaussovej metóde najmenších štvorcov - MNŠ, definujúcej kriteriálnu funkciu Z ako sumu štvorcov rozdielov medzi teoretickými a empirickými hodnotami závisle premennej, teda

$$Z = E\left(Y - f(x)\right)^2, \text{ resp.} \quad \sum_{i=1}^{n} \left(Y_i - f(x_i)\right)^2 \rightarrow \min,$$

$$Z = E\left(Y - f(x_j)\right)^2, \text{ resp.} \quad \sum_{i=1}^{n} \left(Y_i - f(x_{ij})\right)^2 \rightarrow \min, \text{ alebo v tvare}$$

$$Z = E\left(e\right)^2 \quad , \text{ resp.} \quad \sum_{i=1}^{n} \left(e_i\right)^2 \rightarrow \min, \text{ vyjadrujúcom minimálny zvyšok - rezíduum - chybu}$$

regresného modelu.

Vzhľadom na skutočnosť, že regresné modely sú stochastického typu, nehovoríme pri ich vyjadrovaní o funkciách, ale o závislostiach (možných), pretože každá n-tica dát môže byť nahradená spravidla nie iba jednou, ale viacerými závislosťami. Aj v prípade viacnásobných regresných modelov je vhodné sledovať

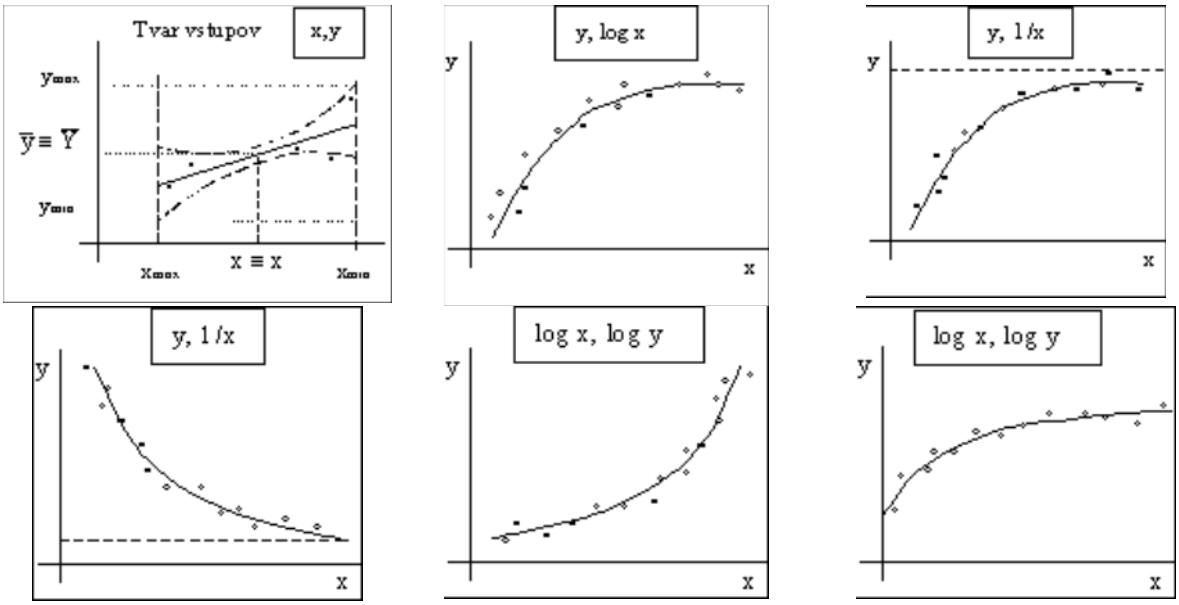
[1] *Doc. Ing. Ľubica Floreková, CSc. a Ing. Marta Benková, CSc., Katedra* riadenia výrobných procesov F BERG TU v Košiciach, Boženy Němcovej 3, 040 01 Košice

(Recenzovali: *Doc. Ing. Michal Leško, CSc.* a *Ing. Dagmar Bednárová*)

závislosti medzi dvojicami premenných a až potom ich spoločný vplyv na závisle premennú.Umožní to pohľad „z druhej strany“ na individuálne vplyvy a na interakcie medzi jednotlivými nezávisle premennými.

Dvojrozmerné regresné modely

Skúsenosť ukazuje, že pri tvorbe jednoduchých regresných modelov je najlepšie vychádzať z grafického vyjadrenia - z *bodového diagramu*, vyneseného v príslušnej mierke. To umožňuje na základe *katalógu kriviek* vybrať skupinu možných typov RM pre spracovávané dáta, $f(x_i), i = 1,\ldots,m$.



Pri *výpočte parametrov* týchto navrhovaných regresných modelov pre zvolené $f(x_i)$ si treba uvedomiť obmedzenia MNŠ. Dovoľuje totiž spracovávať iba také závislosti, ktoré sú v parametroch lineárne, resp. linearizovateľné (transformáciou napr. logaritmovaním, pri rôznych exponenciálnych závislostiach, resp substitúciou pri hyperbolických závislostiach). Vhodné typy úprav sú prehľadne spracované v tab.1.

Tabuľka 1.Typické dvojrozmerné RM (L - lineárne v parametroch, LZ - linearizovateľné v parametroch, t - vhodná transformácia).

Názov RM	Základný tvar Y =	Typ	Transformácia	Základný tvar kriteriálnej funkcie Z
Lineárny	$a_0 + a_1 x$	L	-	$\sum_{i=1}^{n} (a_0 + a_1 x_i - y_i)^2$
Polynomický p_{max}=5	$a_0 + a_1 x_1 + \ldots + a_p x^p$	L	-	$\sum_{i=1}^{n} (a_0 + a_1 x_i + \ldots + a_p x_i^{\,p} - y_i)^2$
Lineárny lomený	$a_0 + a_1 x^{-1}$	L	$t=x^{-1}$,pre $x \neq 0$	$\sum_{i=1}^{n} (a_0 + a_1 t_i - y_i)^2$
Polynomický Lomený p_{max}=5	$a_0 + a_1 x^{-1} + \ldots + a_p x^{-p}$	L	$t=x^{-1}$,pre $x \neq 0$	$\sum_{i=1}^{n} (a_0 + a_1 t_i + \ldots + a_p t_i^{\,p} - y_i)^2$
Hyperbolický	$\frac{1}{a_0 + a_1 x + (a_2 x^2)}$	LZ	$t=y^{-1}$,pre $y \neq 0$	$\sum_{i=1}^{n} (a_0 + a_1 x_i + (a_2 x_i^2) - t_i)^2$

Názov RM	Základný tvar Y =	Typ	Transformácia	Základný tvar kriteriálnej funkcie Z
Hyperbolický	$\frac{a_0}{a_1 + a_2 x}$	LZ	$t=a_0 y^{-1}$,pre $y\neq 0$ odhad a_0 z bod. diagramu	$\sum_{i=1}^{n}\left(a_1 + a_2 x_i - t_i\right)^2$
Hyperbolický	$\frac{x}{a_0 + a_1 x}$	LZ	$t=x.y^{-1}$,pre $y\neq 0$	$\sum_{i=1}^{n}\left(a_0 + a_1 x_i - t_i\right)^2$
Hyp. = exp.	$\frac{1}{a_0 + a_1 e^{-x}}$	LZ	$t=y^{-1}$,pre $y\neq 0$	$\sum_{i=1}^{n}\left(a_0 + a_1 e^{-x_i} - t_i\right)^2$
Mocninný	$a_0 + a_1 x^m$	L	m - dané	$\sum_{i=1}^{n}\left(a_0 + a_1 x_i^m - y_i\right)^2$
Mocninný	$a_0 x^{1/2}$	L	-	$\sum_{i=1}^{n}\left(a_0 x_i^{1/2} - y_i\right)^2$
Exponenciálny	$a_0 x^{a_1}$	LZ	t=log y, r=log x, pre $y\neq 0$, $x\neq 0$	$\sum_{i=1}^{n}\left(\log a_0 + a_1 r_i - t_i\right)^2$
Exponenciálny	$a_o a_1^{x}$	LZ	t=log y ,pre $y\neq 0$	$\sum_{i=1}^{n}\left(\log a_0 + \log a_1 x_i - t_i\right)^2$
Exponenciálny	$a_0 e^{a_1 x + \left(a_2 x^2\right)}$	LZ	t=ln y ,pre $y\neq 0$	$\sum_{i=1}^{n}\left(\ln a_0 + a_1 x_i + \left(a_1 x_i^2\right) - t_i\right)^2$

Pri súčasnej softvérovej podpore je vhodné okrem klasického výpočtu parametrov RM pomocou sústavy normálnych rovníc použiť pre maticové operácie prostriedky,ktoré dávajú všetky typy tabuľkových procesorov, t.j. vytvoriť rozšírenú dátovú maticu (v takom tvare,ktorý zodpovedá vybranému RM) a najlepší odhad parametrov RM získať ako vektor $v_a = \left(X^T . X\right)^{-1}$.

Poznámka: Pri výpočte parametrov lineárneho regresného modelu - LRM ako jediného sa využíva predpoklad tzv. dvojrozmerného normálneho rozdelenia a model sa počíta „obojstranne“, so vzájomnou zámenou premenných, t.j. určujú sa parametre združených regresných priamok, ktoré v grafe vytvárajú tzv. korelačné nožnice.

Pri výpočte parametrov polynomických RM sa neodporúča stupeň polynómu vyšší ako $p_{max}=5$ preto, že potom vypočítaný polynóm umele kopíruje všetky empirické dáta.

Mnohorozmerné regresné modely

Základné skupiny mnohonásobných RM sú:

- **MLRM** - mnohonásobný lineárny regresný model, v základnom tvare $Y = a_0 + \sum_{j=1}^{k} a_j x_j$,
- **MLRMI** - mnohonásobný lineárny regresný model s interakciami,

  v základnom tvare $Y = a_0 + \sum_{j=1}^{k} a_j x_j + \sum_{r=j+1} a_r x_r x_s, \quad pre\ r \neq s$
- **MNLRMI** - mnohonásobný nelineárny regresný model,

  v základnom tvare $Y = a_0 + \sum_{j=1}^{k} a_j x_j + \sum_{r=j+1} a_r x_r x_s + \sum_{j=k+1}^{2k} a_j x_j^2, \quad pre\ r \neq s$.

RM sú výsledkom plánovaných experimentov, ktorými sa zabezpečuje ortogonalita matice X. Pre výpočet MLRM v lineárnom tvare $Y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$,

alebo v linearizovateľnom tvare $Y = a_0 x_1^{a_1} x_1^{a_2} \cdots x_1^{a_k}$,

resp. $Y = a_0 e^{a_1 x_1} e^{a_2 x_2} \cdots e^{a_k x_k}$

sa používajú niektoré úpravy účelovej funkcie pre MNŠ, a to:

- **Klasický tvar** $Z = \sum_{i=1}^{n} (Y_i - y_i)^2 = \sum_{i=1}^{n} (a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_k x_{ik} - y_i)^2$,

  vedúci ku (k+1) rozmernej sústave normálnych rovníc,

$$\begin{vmatrix} n & \sum x_{i1} & \ldots. & \sum x_{ik} \\ \sum x_{i1} & \sum x^2_{i1} & \ldots. & \sum x_{i1}x_{ik} \\ . & . & & \\ . & . & & \\ . & . & & \\ \sum x_{ik} & \sum x_{i1}x_{ik} & \ldots. & \sum x^2_{ik} \end{vmatrix} . \begin{vmatrix} a_0 \\ a_1 \\ \\ \\ a_k \end{vmatrix} = \begin{vmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \\ \\ \sum x_{ik}y_i \end{vmatrix} .$$

  Jej riešením sa vypočíta vektor $a = [a_0, a_1, \ldots, a_k]$ najlepšieho odhadu parametrov MLRM.

- **Klasický tvar, riešený pomocou informačnej matice**

$$X = \begin{vmatrix} 1 & x_{11} & x_{12} & \ldots. & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots. & x_{2k} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_{n1} & x_{n2} & \ldots. & x_{nk} \end{vmatrix}, \quad \text{rozmeru n x (k + 1)}$$

  a vektora $y^T = | y_1 \; y_2 \ldots y_n |$, sa vektor $a = [a_0, a_1, \ldots, a_k]$ vypočíta z maticovej rovnice
  $\mathbf{a} = (\mathbf{X}^T.\mathbf{X})^{-1}.(\mathbf{X}^T.\mathbf{y})$,

  kde $\mathbf{X}^T.\mathbf{X} = \mathbf{M}$ - informačná matica rozmeru (k+1) x (k+1),
  $(\mathbf{X}^T.\mathbf{X})^{-1} = \mathbf{Q}$ - variančno - kovariančná matica.

- **Centrovaný tvar** $Z = \sum_{i=1}^{n} \left(\frac{Y_i - \bar{y}}{n} - \frac{y_i - \bar{y}}{n} \right)^2$, riešený pomocou variančno-kovariančnej matice

$$C = \begin{vmatrix} s^2_{x1} & cov_{x1x2} & \ldots. & cov_{x1xk} \\ cov_{x1x2} & s^2_{x1} & \ldots. & cov_{x2xk} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ cov_{x1xk} & cov_{x2xk} & \ldots. & s^2_{xk} \end{vmatrix}$$

  a kovariančného vektora $\mathbf{c}^T = | cov_{x1y} \; cov_{x2y} \ldots cov_{xky} |$,z ktorých sa vypočíta vektor $\mathbf{a} = \mathbf{C}^{-1}.\mathbf{c}$, najlepších odhadov parametrov $a = [a_1, \ldots, a_k]$.

Absolútny člen $a_0 = \bar{y} - \sum_{j=1}^{k} a_j \bar{x}_j$,

kde $\bar{y}$ - aritmetický priemer hodnôt y_i , i = 1, ..., n ,

$\bar{x}_j$ - aritmetický priemer hodnôt x_{ij} , j = 1, ..., k .

- **Normovaný tvar** $Z = \sum_{i=1}^{n} \left(\frac{Y_i - \bar{y}}{s_y} - \frac{y_i - \bar{y}}{s_y} \right)^2$, riešený pomocou korelačnej matice

$$R = \begin{vmatrix} 1 & r_{x1x2} & \dots & r_{x1xk} \\ r_{x1x2} & 1 & \dots & r_{x2xk} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ r_{x1xk} & r_{x2xk} & \dots & 1 \end{vmatrix}$$

a korelačného vektora $r^T = | r_{x1y} \; r_{x2y} \dots \; r_{xky} |$, z ktorých sa vypočíta vektor $\mathbf{b} = \mathbf{R}^{-1}.\mathbf{r}$

$b = [b_1, \dots, b_k]$ .Tieto hodnoty b_j , j = 1, ..., k, sa transformujú na vektor a prepočtom $a_j = b_j . \frac{s_y}{s_{xj}}$,

kde s_{xj} - smerodajná odchýlka hodnôt príslušnej nezávisle premennej, pričom výpočet a_0 je rovnaký ako u centrovaného tvaru.

Na tomto mieste považujeme za potrebné, uviesť aspoň krátke porovnanie výhod a nevýhod metód výpočtu MLRM.

Klasická MNŠ je zdĺhavá pri príprave jednotlivých súm. Existuje tu možnosť ovplyvnenia presnosti výpočtu parametrov v závislosti na rôznych rozmeroch závisle a nezávisle premenných, pričom vypovedacia hodnota sústavy normálnych rovníc nie je vysoká. Výsledná maticová rovnica v konečnom dôsledku je rovnaká ako pri modifikovanej MNŠ riešenia pomocou informačnej matice.

O dátovej matici sa spravidla predpokladá, že je regulárna a že nie je úplne ortogonálna, t.j. že nezávisle premenné nie sú úplne nezávislé. A ak aj sú, uvažujeme s náhodnými chybami, ktoré ortogonalitu narušujú.

Pri použití všetkých metód, t.j. riešenia pomocou **informačnej matice, variančno-kovariančnej matice** a **korelačnej matice** predpokladáme, že dátam vyhovuje rovnica **Y = X.a + e**, (kde **e** - vektor hodnôt náhodnej (nepozorovanej) zložky s $N(0,\delta^2)$) a teda riešenie platí pre predpoklad rovnakých rozptylov a nulových kovariancií (klasický MLRM) alebo rôznych rozptylov a nenulových kovariancií (všeobecný MLRM). Kovariančná matica $\mathbf{C} = \delta^2\mathbf{I}$ je krajným prípadom kovariančnej matice $\mathbf{C} = \delta^2\mathbf{W}$, kde v prípade heteroskedasticity je matica váh **W** diagonálna a v prípade modelov so závislými poruchami je úplná - autoregresné modely časových radov.

Potom $\mathbf{Q} = \mathbf{M}^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}$ by mala spĺňať niektoré z kritérií optimálnosti, ako ich definoval Fisher, Fedorov, Kiefer (Draper et al., 1981).

Pri D - optimalite jedná sa o minimalizáciu det **Q** „elipsoidu“ odhadu parametrov (riziko nepomeru dĺžky jednotlivých osí).

Pri E - optimalite je kritériom minimálna hodnota maximálneho charakteristického čísla λ matice **Q**.

Pri A - optimalite sa jedná o minimalizáciu stopy matice **Q**, teda o minimalizáciu stredného rozptylu najlepšieho lineárneho odhadu parametrov.

Pri G - optimalite sa minimalizuje maximálny rozptyl v oblasti spracovávaných dát.

Ideálne je, ak má model vlastnosť I - optimality, t.j. keď matica **Q** je jednotková, resp. aspoň ak $\mathbf{Q} = k\mathbf{I}$, teda ak nezávisle premenné sú naozaj nezávislé a majú rovnaký rozptyl.

Pri skúmaní matice $\mathbf{X}^T\mathbf{X}$ je potrebné venovať pozornosť aj multikolinearite, t.j. silnej vzájomnej závislosti nezávislých premenných (determinat matice $\mathbf{X}^T\mathbf{X}$ je blízky nule a matica je skoro singulárna). To má za následok zlú podmienenosť matice, čo sa prejaví tým, že „malá zmena vo vstupoch spôsobí neúmerne veľkú zmenu vo výstupoch“. Multikolinearita sa v kovariančnej matici **C** prejavuje značne vysokými a rôznymi hodnotami rozptylov na jej diagonále, v korelačnej matici **R** vysokými hodnotami koeficientov korelácie mimo hlavnej diagonály, významnou odlišnosťou matice **R** od jednotkovej matice **I**, teda informáciou o skutočnosti, že sledovaný model nie je MLRM.Tu je potom potrebné „zastaviť“ výpočet MLRM a riešiť problém vzťahu medzi premennými pomocou metód plánovania experimentov, ktoré zabezpečujú ortogonalitu matice plánu Q pre modely MLRMI a MNLRMI.

Dôsledkom multikolinearity je veľká šírka intervalov spoľahlivosti, a teda malá presnosť odhadov parametrov **a**, ale aj skutočnosť, že nie sme schopní naozaj vyjadriť samostatný vplyv jednotlivých nezávisle premenných na závisle premennú, pretože tu pôsobia aj ich interakcie.

Sledovanie multikolinearity sa dá testovať, napr. Farrar-Glauberovým testom. Je možné použiť tiež niektorú z metód výberu najlepšej podmnožiny nezávisle premenných, alebo najlepšieho regresného modelu (stepwise

regresion, backward elimination procedure, finite intersection test, method based on all possible regressions), ktoré bývajú súčasťou veľkých štatistických programových balíkov.

Pri hodnotení **variančno-kovariančnej** matice **C** je dôležité si uvedomiť, že pri centrovaní sa znižuje riziko vplyvu multikolinearity. Test diagonality slúži na overenie heteroskedasticity a test sféričnosti rozdelenia na overenie hypotézy, či $\mathbf{C} = \delta^2\mathbf{I}$, teda, či všetky nezávisle premenné sú získané s rovnakou presnosťou, pretože pre presnosť odhadu vektora parametrov a je dôležité, aby všetky rozptyly boli aspoň rádove približne rovnaké, čo svedčí o homoskedasticite.

Kedže kovariancia je mierou intenzity závislosti medzi veličinami v prípade ich lineárneho vzťahu, čím silnejší je tento vzťah, tým sú hodnoty kovariancií vyššie.

Pri takých dátach, ktoré sa vzťahujú na čas, sa v matici C prejavuje často vplyv **autokorelácie**, teda závislosti dát na sebe. Vtedy vzniká problém dát, ktoré za sebou nasledujú, pretože predchádzajúce dáta môžu ovplyvňovať nasledujúce dáta. Autokorelácia teda znamená, že efekt náhodných chýb nie je okamžitý, ale sa prejaví až v budúcnosti. Riešenie tejto problematiky sa spája s Durbinon a Watsonom, so štatistikou D pre odhad autokorelačného koeficienta a s rôznymi výpočtovými autoregresnými schémami.

Pri sledovaní vlastností **korelačnej matice R** sa často používa koeficient korelácie v medziach intervalu< -1,1 > aj ako všeobecná miera závislosti, jej intenzity.

Významnosť jednoduchých koeficientov korelácie možno určiť t-testom. V korelačnej matici **R** očakávame, že **R** = **I**. Ak takáto situácia nenastane, je zrejmé, že treba namiesto MLRM hľadať iný, vhodnejší typ. Hodnoty koeficientov korelácie r_{yxj} nesú informáciu o individuálnej sile závislosti medzi y a jednotlivými x_j, resp. aj o type tejto párovej závislosti.

Určovanie kvality - vhodnosti regresného modelu - korelačné charakteristiky

Miera stochastickej závislosti vychádza najskôr z porovnávania hodnôt súm štvorcov odchýliek, a to:

- $\sum_i \left(y_i - \bar{y}\right)^2 = S_1 \Rightarrow$ celkový rozptyl $s_y^2 = \frac{S_1}{n-1}$,
- $\sum_i \left(Y_i - y_i\right)^2 = S_2 \Rightarrow$ reziduálny, zvyškový rozptyl $s_{yx}^2 = \frac{S_2}{n-p}$,

  kde p je počet koeficientov RM,
- $\sum_i \left(Y_i - \bar{y}\right)^2 = S_3 \Rightarrow$ teoretický, vyrovnaný rozptyl $s_Y^2 = \frac{S_3}{n-1}$.

Pri približnej platnosti $S_1 = S_2 + S_3$, sa potom pre:

- LRM vypočíta *Pearsonov koeficient korelácie*: $r = r_{yx} = r_{xy} = \frac{\mathrm{cov}_{xy}}{s_x . s_y}$, resp. pri malom počte hodnôt Romanovského poradový koeficient korelácie $R = 1 - 6\frac{\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}$, ktoré sú platné v intervale $\langle -1,1 \rangle$ a ktorých významnosť sa testuje,
- NLRM vypočíta *index korelácie* $I = \sqrt{1 - \frac{s_{yx}^2}{s_y^2}}$, resp. $I = \sqrt{\frac{s_Y^2}{s_y^2}}$,ktorý platí v intervale $\langle 0,1 \rangle$ a ktorého významnosť sa tiež testuje.
- MLRM vypočíta mnohonásobný koef.korelácie $r_{y,x_1,\ldots x_k} = \sqrt{1 - \frac{s_{y,x_1,\ldots x_k}^2}{s_y^2}}$.

Koeficient, resp. index korelácie určujú stupeň rozptýlenia teoretických modelových hodnôt Y okolo regresnej čiary, resp. regresnej hyperroviny .

Okrem týchto základných korelačných charakteristík sa odporúča počítať koeficient determincie $R^2 = \frac{s_{rez}^2}{s_{celk}^2}$, so snahou o jeho minimalizáciu, resp. Fisherovu hodnotu adekvátnosti regresného modelu $F = \frac{s_{celk}^2}{s_{rez}^2}$, so snahou o jeho maximalizáciu pri ich testovaní, $H_a : s_{celk}^2 - s_{rez}^2 \neq 0$.

Testovanie významnosti parametrov RM Studentovým testom, prípadne výpočet spoľahlivostného, resp. tolerančného pásu okolo regresných čiar sa uvádza spravidla len ako príklad.

Pri MLRM sa upozorňuje na možnosť autokorelácie hodnôt, t.j. možnosť vzniku časového radu a jeho kontrolu Durbin-Watsonovým testom.

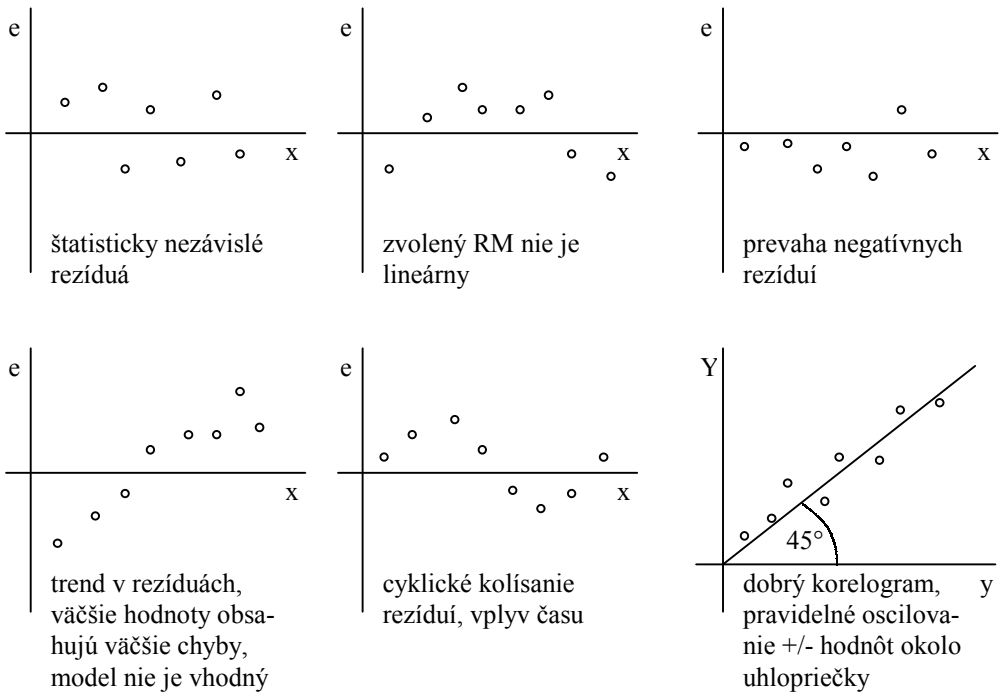
S ohľadom na mnohokrát sa vyskytujúce problémy so získavaním najmä dát navzájom súvisiacich v stabilných podmienkach, je pri mnohonásobnej regresnej analýze vhodné upozorňovať na tzv. MALTHUS nebezpečenstvo (Newbold, 1990) (Multicolinearity, Autocorrelation, Lack of Data, Time and Cost Constraint, Heteroskedasticity, Underidentification, Specification), teda na uvedené riziko prítomnosti multikolinearity, autokorelácie, nedostatku dát pre „dobrý“ model, časové a nákladové obmedzenia, ovplyvňujúce získanie reprezentatívnych empirických dát, heteroskedasticitu, nevhodnú identifikáciu dôležitých premenných a zlú špecifikáciu najmä zložitejších modelov, čo práve uvedené postupy výpočtov umožňujú.

Okrem číselných korelačných charakteristík sa pri výbere najvhodnejšieho RM vykresľujú okrem grafického priebehu RM v bodovom diagrame aj grafy rezíduí, a to spravidla voči x_i a Y_i, ako aj korelogram y_i / Y_i. Pri MLRM sa odporúča vykresliť aspoň jednoduché párové závislosti y na jednotlivých nezávisle premených x_i.

Náhodnosť rezíduí možno sledovať napr.:

- rozložením e_i alebo e_i/x_i na číselnej osi,
- bodovým diagramom závislosti rezíduí e_i na teoretických hodnotách Y_i,
- bodovými diagramami závislosti rezíduí e_i na empirických hodnotách x_{ij},
- bodovým diagramom časovej postupnosti rezíduí (pri časových radoch).

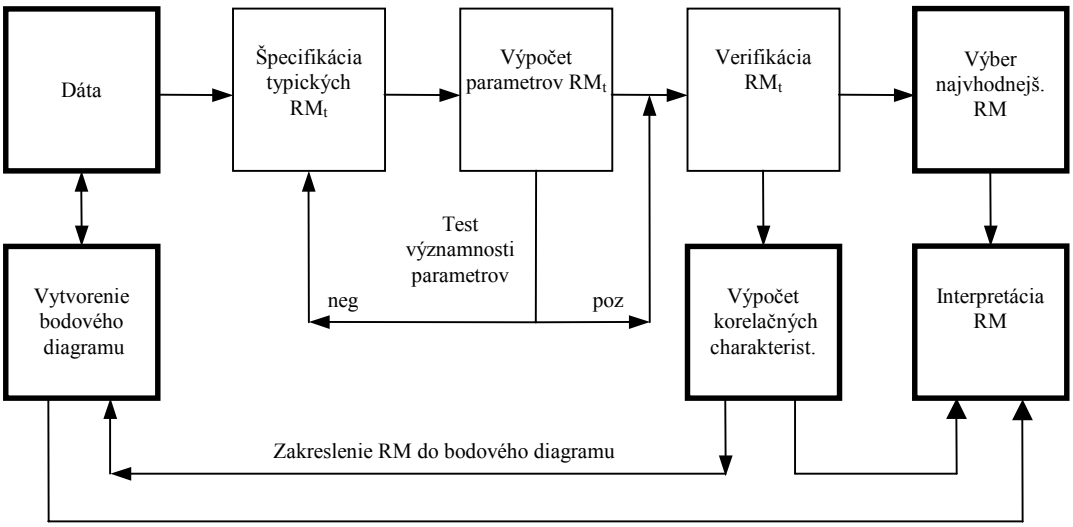
(Vzhľadom na korelovanosť empirických hodnôt y_i a rezíduí e_i sa ich grafy nerobia.)



Zhluk rezíduí do tvaru „mraku“ potvrdzuje správnu voľbu modelu. „Výseč“, ktorú vytvoria rezíduá informuje o heteroskedasticite v dátach a „pás“ informuje o vybočujúcich dátach. Ak je „pás“ zakrivený, model je nevhodne navrhnutý (Hustopecký et al., 1990).

Okrem toho, pri počte K=2, resp. K=3 nezávisle premenných je možné vypočítaný MLRM vizualizovať zobrazením pomocou kužeľosečiek v ^2D alebo kvadrík v ^3D priestore, čo umožňuje ich hlbšie skúmanie(Floreková et al., 1999).

Pre konečný výber RM odporúčame nasledovný postup:



Záver

Vypovedacia schopnosť RM závisí na kvalite spracovávaných dát, získaných za stabilizovaných podmienok, a je ju možné využiť iba pri plnom rešpektovaní všetkých informácií vecného, logického, technologického, výrobného, ekonomického charakteru o danej problematike. Interpretácia RM umožňuje hlbšie spoznanie skúmaných javov, súvislostí medzi premennými z kybernetického modelovacieho pohľadu na objekt ako na čiernu skrinku. RM nedávajú informácie o vnútornom správaní a fungovaní. Sú modelmi typu vstup / výstup.

Literatúra

Draper, N., Smith, H.: Applied Regression Analysis. *J.Wiley, New York, 1981.*
Floreková, Ľ a Benková, M.: Štatistické metódy. *FPP-F BERG, TU v Košiciach, 1999.*
Hustopecký, J. a Hebák, P.: Průvodce moderními statistickými metodami. *SNTL Praha, 1990.*
Krishnaiah, P.R., ed.: Handbook of statistics, 1,2. *North-Holland, 1980, 1982.*
Newbold, P.: Statistics for Business and Economics. *J.Wiley, New York, 1990.*
Rao, R. C.: Lineární metody statistické indukce a jejich aplikace (preklad z angl.). *Academia Praha, 1978.*