Regresná a korelačná analýza ako kognoskačná metóda základného a aplikovaného výskumu

Regressive and correlation analysis as a method of cognition of basic and applied research, Vol. 29, No. 1, 42–44, 1995.

Regressive and correlation analysis represent very important and at present a very well-prepared field of statistical simulation of subordination among investigated quantities. In the last twenty years, statistical methods have become understandable for a large number of people due to the intensive development of computer techniques. Regressive and correlation analysis examine the relationship between two or more variables. The two problems are always solved simultaneously - the regressive one quantifying the relationship among variables and the correlation one which informs about the closeness of a found relationship considering the experimental points.

Classic regressive methods operate with a stable functional model being held constant during computation. The choice of a functional model encounters difficulties especially with non-linear and multi-dimensional problems. Its incorrect choice brings into regressive calculation smaller to larger mistakes. Therefore, this development led to the so called step-methods. The step-methods solve the problems of the functional form of the regressive model. The so called combined procedure operating in three algorithmic steps belongs to these methods. The combined procedure operates using the method of the least squares and it gives impartial calculations. Presently it is one of the most progressive methods of statistical simulation.

Regresná a korelačná analýza predstavuje veľmi dôležitú a v súčasnosti aj veľmi rozpracovanú oblasť štatistického modelovania závislosti medzi pozorovanými veličinami. V poslednom dvadsaťročí, práve vďaka ohromnému rozvoju výpočtovej techniky, stávajú sa štatistické metódy dostupnými širokému okruhu užívateľov. Regresná a korelačná analýza skúma väzby medzi dvoma a viacerými premennými. V podstate sa súčasne riešia vždy dve úlohy, a to regresná úloha, kvantifikujúca hľadanú väzbu medzi premennými a korelačná úloha, informujúca o tesnosti nájdeného vzťahu vzhľadom na experimentálne body.

Najjednoduchšou a zároveň aj najpoužívanejšou formou regresnej závislosti je lineárny regresný model, najmä preloženie regresnej priamky cez sústavu bodov dvoch meraných veličín podľa rovnice:

$$Y = a + b \cdot x \qquad (1)$$

kde Y je závisle premenná, a je konštantný člen, b je regresný koeficient a x nezávisle premenná veličina.

Konštantný člen a regresný koeficient sa vypočítajú pomocou metódy najmenších štvorcov, ktorá zaručí také preloženie regresnej priamky cez experimentálne body, aby prechádzala čo najbližšie ku všetkým bodom.

Zároveň treba riešiť korelačné charakteristiky, ako napríklad koeficient korelácie, test jeho významnosti oproti kritickej hodnote, koeficient determinácie, posúdiť tesnosť a použiteľnosť nájdenej závislosti. Kritické hodnoty koeficienta korelácie obvykle uvádzajú štatistické tabuľky pre rôzne hladiny významnosti.

Podobne možno vytvoriť lineárny regresný model pre tri a viac premenných.

Závislosti však majú niekedy zjavne nelineárny priebeh. Zo vzťahu $Y = a + b * x$ môžeme získať krivkovú závislosť napríklad v tvare:

$$Y = a + b \cdot x^2 \qquad (2)$$
$$Y = a + b \cdot \ln(x) \qquad (3)$$
$$Y = a \cdot x^b \qquad (4)$$
$$Y = a \cdot e^{bx} \qquad (5)$$

Pri týchto závislostiach je význam a a b podobný ako v prvom vzťahu, okrem toho ln je prirodzený logaritmus, e je základ prírodzeného logaritmu.

V prípade závislosti (4) a (5) pre možnosť použitia metódy najmenších štvorcov treba upraviť funkčný model tak, aby regresné koeficienty boli vo vzájomnom lineárnom vzťahu, čo dosiahneme logaritmovaním. Podobným spôsobom možno rozvíjať regresné modely na báze lineárnej rovnice s dvoma alebo viacerými nezávisle premennými pre riešenie nelineárnych úloh. Príkladom je vzťah:

$$Y = a \cdot x^b \cdot e^{cx} \qquad (6)$$

Logaritmovaním dostaneme lineárnu funkčnú formu

$$\ln (Y) = \ln (a) + b \cdot \ln (x) + cx \qquad (7)$$

pre ktorú už možno použiť aproximačnú metódu najmenších štvorcov na stanovenie regresných koeficientov.

Takto sa dá vytvárať široká paleta všeobecných (hypotetických) regresných modelov, od najjednoduchších plošných foriem až po modely v n-rozmernom priestore a vzápätí overiť, či sa potvrdí predpokladaná závislosť alebo nie.

Príklady, ktoré sme uviedli, predstavujú regresné modely s pevným funkčným vzťahom, pretože ešte pred začatím výpočtu regresnej rovnice a korelačných charakteristík musíme určiť konkrétny tvar regresnej funkcie, ktorý sa už v priebehu výpočtu nemení. Túto voľbu vyhovujúceho funkčného vzťahu môžeme uskutočniť odhadom očakávaného priebehu hľadanej závislosti, prípadne pri párovej regresii grafickým zobrazením korelačného poľa a podľa jeho konfigurácie voliť matematickú funkciu, ktorá má predpoklady vykresliť hľadanú závislosť.

Výpočet regresných koeficientov zo sústavy normálnych rovníc nenaráža na ťažkosti, avšak stanovenie správneho funkčného tvaru regresného modelu, najmä pri viacrozmerných úlohách, kde strácame možnosť grafického overenia priebehu hľadanej závislosti, je stále značným problémom. Nesprávny odhad funkčného tvaru regresného modelu môže vniesť do výpočtu aj značné nepresnosti, prejavujúce sa v zhoršení odhadu.

Ďalším nedostatkom regresnej a korelačnej analýzy s pevným funkčným modelom, najmä v prípade zložitých modelov je to, že sa dostanú do modelu aj premenné, ktoré nevýznamne participujú na vysvetlení rozptylu závisle premennej.

Na preklenutie ťažkostí so stanovením funkčného tvaru regresného modelu sa vypracovalo niekoľko procedúr regresnej a korelačnej analýzy, ktoré sa súhrne nazývajú „krokové metódy". Sú to:
- procedúra všetkých regresií,
- procedúra etapovej regresie,
- priama procedúra,
- spätná procedúra,
- kombinovaná procedúra.

Spoločným znakom týchto procedúr je snaha riešiť i problém funkčného tvaru regresného modelu. Každá ho rieši inak, iným usilím a úspechom. Kombinovaná procedúra je zo všetkých najprogresívnejšia, pretože zhŕňa prednosti priamej a spätnej procedúry. Budeme sa preto venovať len tejto procedúre a označíme ju *„Mnohonásobná kroková regresná a korelačná analýza"*. Jej nespopornou výhodou je to, že zlučuje dve funkcie, a to výber štatisticky významných premenných zo všetkých do úvahy prichádzajúcich premenných, ako aj stanovenie konkrétneho regresného modelu, t. j. výpočet hodnôt, parametrov a charakteristík. Práve tento princíp umožňuje zo súboru pozorovaní vyťažiť prakticky maximum v prospech hľadanej závislosti. Metóda vychádza z lineárneho regresného modelu:

$$Y_i = a + \sum_{j=0}^{n} b_j x_{ij} + e_i \qquad (8)$$

kde index i označuje jednotlivé pozorovania i=1, 2, ..., m, indexom j sú označené jednotlivé nezávisle premenné j=1, 2, ..., n, a, b_1, ..., b_n sú parametre regresnej funkcie, pričom a je absolútny člen a b sú regresné koeficienty, e_i je normálne rozdelená náhodná veličina s nulovou očakávanou hodnotou a konštantným rozptylom σ^2.

Lineárny regresný model zavedením transformácií pôvodných nezávisle premenných možno použiť aj v celom rade nelineárnych úloh. Pri zadaní výpočtu sa programu ponúknu skupiny transformácií nezávisle premenných typu: X^n, SQRT(X), $\ln^n(X)$, e^X, sin(X), párové súčiny $X_1.X_2$, $X_1.X_3$, ..., $X_2.X_3$ ..., párové podiely X_1/X_2, X_1/X_3, ..., X_2/X_3, ... a ich recipročné hodnoty. Ako kritérium pre zaradenie alebo vyradenie premennej z regresného modelu sa používa F-test, T-test alebo priamo hladina významnosti.

Algoritmus mnohonásobnej krokovej regresnej a korelačnej analýzy pozostáva z troch krokov:

- **1. krok** – do regresného modelu sa zaradí premenná, ktorá je so závisle premennou najtesnejšie korelovaná. To znamená, že sa zo všetkých základných nezávisle premenných a ich transformácií (ďalej len premenných) vyberie tá, ktorej rozptyl najviac vysvetľuje rozptyl závisle premennej. Prvý krok sa vo výpočte viackrát neopakuje. Kritériom pre výber prvej premennej sú párové koeficienty korelácie medzi závisle premennou a jednotlivými nezávisle premennými. Pre vstupujúcu premennú je vypočítaná hodnota F-testu, ktorý má Fisher-Snedecorovo rozdelenie a je stanovená zodpovedajúca hladina významnosti, ktorá sa porovná so zadanou hladinou pre vstup premennej do modelu. Ak toto kritérium nie je splnené, výpočet sa ukončí a regresný model zostane prázdny. Inak výpočet pokračuje krokom 2.
- **2. krok** – robí výber premennej z tých, ktoré dosiaľ neboli zahrnuté do regresnej rovnice. Testuje ich prínos pre vysvetlenie rozptylu závisle premennej. Kritériom výberu ďalšej premennej sú parciálne koeficienty korelácie medzi závisle premennou a nezávisle premennými s vylúčením vplyvu tých, ktoré sa momentálne nachádzajú v regresnej

rovnici. Pre vybranú premennú sa vypočíta hodnota veličiny F i zodpovedajúca hladina významnosti. Ak je vypočítaná hladina významnosti väčšia než zadaná, prínos vybranej premennej je štatisticky nevýznamný a výpočet končí. Výslednou rovnicou je rovnica získaná v predchádzajúcom kroku.

● **3. krok** – postupne prehodnocuje významnosť všetkých premenných, ktoré sú v regresnej rovnici. Rovnako ako druhý krok, je to všeobecný krok a v priebehu výpočtu sa môže niekoľkokrát opakovať. Z premenných nachádzajúcich sa v regresnej rovnici sa vyberie v tomto kroku tá, ktorá má najmenší parciálny koeficient korelácie so závisle premennou. K nej sa dopočíta hodnota parciálneho F-testu a hladina významnosti. Ak je menšia alebo rovnaká ako zadaná pre výstup z modelu, prínos premennej je významný a premenná zostane v modeli. Výpočet v tomto prípade pokračuje krokom 2. Ak je však vypočítaná hladina významnosti väčšia než zadaná, prínos premennej je nevýznamný. Premenná sa vylúči z regresnej rovnice a vypočítajú sa parametre a charakteristiky novej regresnej rovnice, ktorá vznikne po vylúčení tejto premennej.

Výpočet pokračuje opakovaním kroku 3 dovtedy, kým sa nevylúčia všetky nevýznamné premenné, potom pokračuje krokom 2. Kroky 2 a 3 sa striedajú, kým sa na kroku 2 nepresunú všetky významné premenné do regresného modelu.

Z algoritmu vyplýva, že výpočet regresného modelu sa v rámci zadania optimalizuje a výsledný regresný model obsahuje len také členy, ktorých rozptyl významne participuje na vysvetlení rozptylu závisle premennej.

Ak sa pre výpočet nezadá žiadna skupina transformácií zo štandardného polynómu, program realizuje lineárnu regresiu medzi pôvodnými premennými, ale s rešpektovaním uvedených kritérií. Program je preto vhodný aj na predbežnú selekciu významných premenných.

Pri mnohonásobnej krokovej regresii v prípade výskytu významnej multikolinearity, t. j. silnejšej vzájomnej väzby medzi nezávisle premennými (napr. genetickej alebo technologickej väzbe), môže sa stať, že výsledný regresný model nebude správne špecifikovaný. Štandardný polynóm (súbor transformácií) preto obsahuje aj kombinované premenné (párové súčiny, podiely a ich recipročné hodnoty). Ich zaradením do hypotetického modelu sa obvykle dosiahne požadované zloženie výsledného regresného modelu.

Protokol nášho programu mnohonásobnej krokovej regresnej a korelačnej analýzy pozostáva z dvoch častí.

- **Úvodná časť s údajmi** - názov úlohy, názov súboru, test nenulovosti a nezápornosti vstupných dát, výber transformačných skupín premenných pre tvorbu modelu, slovník premenných (pôvodných a transformovaných), minimum, priemer, maximum a smerodajnú odchýlku jednotlivých premenných, párové korelácie, počty pozorovaní, volené hladiny významnosti pre vstup a výstup premenných z modelu.

- **Vlastný výpočet** - po každom kroku s výsledkami riešenia korelačnej úlohy, obsahujúcimi reziduálnu chybu, reziduálny variačný koeficient, koeficient korelácie linearizovaného regresného modelu, koeficient determinácie ako percento odôvodnenej regresie, F-test koefecienta korelácie a hladinu významnosti, na ktorej je vypočítaný koeficient korelácie štatistický významný (práve kritický) a regresnej úlohy, obsahujúcej číslo premenných zaradených do regresného modelu podľa slovníka premenných, hodnotu regresných koeficientov, ich smerodajnú odchýlku, parciálne F-testy a hladiny významnosti, na ktorých sú jednotlivé regresné koeficienty štatisticky významné. Protokol predstavuje vyčerpávajúcu informáciu o spracúvanom súbore a jednotlivých krokoch tvorby regresného modelu, takže ďalšiu správu zadávateľ úlohy obvykle nepotrebuje.

Počet vstupných údajov pre regresiu obmedzuje len kapacita vonkajšej pamäte. Požiadavkou na vstupné dáta z hľadiska výpočtu závislosti je ich homogenita (vznikli za rovnakých, dátami nedefinovaných podmienok) a rovnomernosť rozloženia experimentálnych bodov v korelačnom poli, ale aj vzájomná nekorelovanosť nezávisle premenných.

Práve pre spomínané vlastnosti sa mnohonásobná kroková regresná a korelačná analýza zaraďuje medzi najpokrokovejšie metódy štatistického modelovania, ktorá dokáže zo súboru pozorovaní vyťažiť maximum v prospech hľadanej závislosti.

Matematická analýza prenikala do rôznych oblastí života rôznou rýchlosťou. Niektoré odvetvia, napr. poľnohospodárstvo, zdravotníctvo a pod., nie sú ešte dnes dostatočne pripravené na aplikáciu výpočtových metód. Ale aj v rámci nich sa uskutočňuje veľmi seriózny výskum, ktorého výsledky treba vyhodnotiť a zovšeobecniť, eventuálne vysloviť ich s určitou pravdepodobnosťou. V tom práve môžu pomôcť metódy regresnej a korelačnej analýzy, ktoré kvantifikujú závislosti a skúmajú ich štatistickú významnosť. Nie je jedno, ktorú zo širokej palety ponúkaných metód použijeme na riešenie konkrétneho problému. Práve preto sme poukázali na výhody a nedostatky najpoužívanejších foriem regresnej a korelačnej analýzy ako kognoskačnej metódy základného a aplikovaného výskumu.

Literatúra

Čiško, V., 1972: Použitie regresnej a korelačnej analýzy v rudnom úpravníctve. Zborník prednášok X. celoštátnej úpravníckej konferencie, Košice.

Miesch, A. T., Connor, J. J., 1968: Stepwise regression and nonpolynomial models in trend analysis. The university of Kansas, Lawrence, USA.

Vachuška, V., 1972: Mnohonásobná regresná analýza. Publikácia VZUP Kamenná, Příbram.