Ракописот примен: јуни, '94

Зборник на трудови на РГФ Штип 1, стр. 113-118(1994)

UDC: 550.83.001.573

стручен труд

Метода на корелација и линеарна регресиона анализа

Тодор Делипетров, Јордан Живановиќ, Рударско-геолошки факултет Штип

Клучни зборови : корелација, регресија, модел, параметар

ВОВЕД

Современит тренд на развој на истражувањата во геологијата бара се поголема примена на егзактни физички, хемиски и други инструментални методи, како и масовна примена на компјутерската техника. Самите геолошки проблеми како и зголемените можности на моделирањето бараат примена на соодветни математички техники во разрешувањето на комплексните геолошки задачи.

Методата на регресиона анализа е типичен пример на математичка метода која може да има широка примена во решавањето на поставените геолошки задачи. Од тој аспект овој труд то прикажува начинот на нејзината примена.

Определувањето на корелационите параметри, за дефинирање на соодветен модел, успешно може да се користи методата на регресиона анализа. Во овој труд детално ќе биде прикажана методата на линеарна регресиона анализа помеѓу параметрите Y и X, односно објаснет е начинот на моделирање од линеарен тип, даден со формулата :

$$Y = a + bx$$

Треба да се има предвид дека регресионата анализа може да биде изразена и со друга зависност, а не само линеарна зависност како и повеќе параметарска зависност.

Сакајќи да ја откриеме, дефинираме и анализираме врската помеѓу два или повеќе параметри, мора истовремено, симултано да се пратат и споредуваат вариациите на двата (или повеќе) параметра и да се мерат односите помеѓу тие вариации. Заради тоа, оваа метода се вика метода на корелација (метода на истражување на взаемните односи).

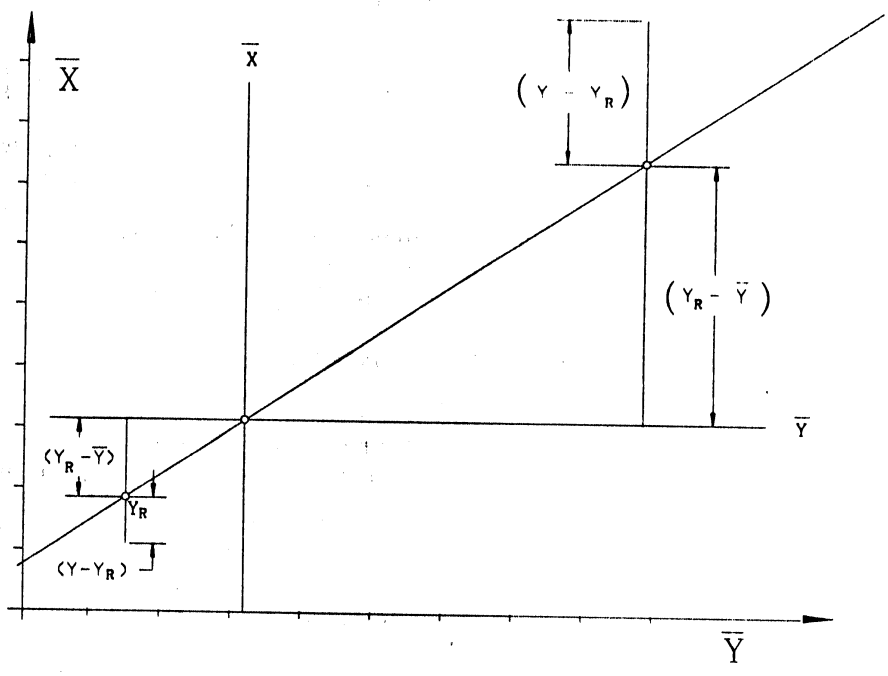
Со методата на корелација се определуваат:

- Знакот на корелацијата, кој може да биде позитивен (+) или негативен (-). Ако вредностите на едниот параметар *(Y)* растат а исто така растат и вредностите на другиот параметар *(x)* или ако вредностите на параметарот *(x)* опаѓаат аналогно на тоа опаѓаат и вредностите на параметарот *(Y)* тогаш корелацијата има позитивен знак (+). Ако вредностите на параметарот (x) растат а во исто време на параметарот (*Y)* опаѓаат, односно за поголеми вредности на *(Y)* одговараат помали вредности на *(x)*, тогаш корелационата врска е со негативен предзнак (-).

- Покрај знакот на корелацијата важен елемент е и јачината на меѓузависноста на корелираните параметри, изразена со вредноста на корелациониот коефициент. Најјака врска е функционалната зависност, односно на секоја вредност на едната појава (параметар - *y*) одговара точно одредена вредност на другата појава (параметар - *x*). Интензитетот на корелациониот коефициент (по апсолутна вредност) се движи од *0 - 1*. Ако корелациониот коефеценит има вредност 0 тоа значи врската помеѓу разгледуваните параметри не постои, колку е поголем копрелациониот коефициент, односно колку повеке се доближува до вредноста 1 толку врската помеѓу тие параметри е појака.

- Покрај знакот и интензитетот на врската, важен елемент е и обликот на врската. Во нашите истражувања е користена линеарна зависност *($y = a + bx$)*.

Нека се зададени N парови од (x_i, y_i) $i = 1,2,....,N$, каде x_i и y_i се параметрите кои се цел на нашите истражувања. Испитувањето на врската почнува со графичко прикажување на параметарската серија на парови. Ова прикажување се врши во Декартов координатен систем, параметарските парови на истражуваните појави (x_i, y_i) $i =1,2,....,N$, се нанесуваат во дадениот координатен систем. Еден параметарски пар (x_k, y_k), во координатниот систем представува една точка, со апциса x_k и ордината y_k, на тој начин во координатниот систем се добиваат N точки,



Сл.1

За полесна ориентација ќе ги повлечиме ординатата и апсцисата на точката (x, y). Овие координати се определени со изразите :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

На база на дијаграмот на дисперзија, односно внесените N точки во координатниот систем, потребно е да се повлече права која "најмногу " одговара на дадените влезни податоци. Најдобра може да биде само една права, задача е, значи да ја определиме равенката на тој правец на регресија:

$$y_R = a + bx$$

Вредноста на правата е означена со y_R за разлика од оригиналните (зададени) вредности на променливата y. Правата која најдобро се прилагодува на оригиналните вредности, треба да биде поставена помеѓу точките на дијаграмот на дисперзија, така да сумата на квадратите на одстапувањата на оригиналните вредности (y) од вредностите на правата мерени вертикално по ординатата е најмала. Овој услов може да се изрази на следниот начин:

$$\sum_{i=1}^{N}\left(y_i - y_{R_i}\right)^2 = min. \qquad (1)$$

Со помош на овој услов равенката на правата може да се определи со методата на најмали квадрати. Оваа метода е објективна метода во која процесот на минимализација на квадратите на одстапувања во изразот (1) дава оптимални прилагодувања на правата во однос на оригиналните вредности. Во наредниот текст ке биде изложена методата на најмали квадрати.

За правата $y_R = a + bx$ најдобро да се прилагоди на оригиналните вредности, мора да биде исполнет условот (1) односно

$$\sum_{i=1}^{N}\left(y_i - a - bx_i\right)^2 = min \tag{2}$$

Параметрите a и b треба да се определат така да горниот израз е минимален. Според правилото за барање на екстремни вредности на функции со две променливи параметрите a и b се определуваат така што горниот израз (2) го диференцираме по a и b :

$$\frac{d}{da}\sum_{i=1}^{N}\left(y_i - a - bx_i\right)^2 = 0$$

$$\frac{d}{db}\sum_{i=1}^{N}\left(y_i - a - bx_i\right)^2 = 0 \tag{3}$$

и добиените изрази ги прирамниме на нула :

$$-2\sum_{i=1}^{N}\left(y_i - a - bx_i\right) = 0$$

$$-2\sum_{i=1}^{N}\left(y_i - a - bx_i\right)x_i = 0 \tag{4}$$

овие равенки можат да се напишат во вид на нормални равенки:

$$\sum_{i=1}^{N} y_i = Na + b\sum_{i=1}^{N} x_i$$

$$\sum_{i=1}^{N} x_i y_i = a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 \tag{5}$$

Ако тие две равенки ги решиме ќе се добие:

$$b = \frac{\begin{vmatrix} N & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i y_i \end{vmatrix}}{\begin{vmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{vmatrix}} = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i} \tag{6}$$

Бидејќи е:

$$\sum_{i=1}^{N} y_i = N\bar{y} \qquad \sum_{i=1}^{N} x_i = N\bar{x}$$

тогаш е :

$$b = \frac{N\sum_{i=1}^{N} x_i y_i - -N\bar{x}\bar{y}}{N\sum_{i=1}^{N} x_i^2 - N(\bar{x})^2} = \frac{\sum_{i=1}^{N} x_i y_i - \bar{x}\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - N(\bar{x})^2} = \frac{\sum_{i=1}^{N} x_i y_i - \bar{y}\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 - \bar{x}\sum_{i=1}^{N} x_i} \tag{7}$$

Ако првата нормална равенка од (5) се подели со N, се добива следното :

$$a = \frac{\sum_{i=1}^{N} y_i}{N} - b\frac{\sum_{i=1}^{N} x_i}{N} \quad \text{односно} \quad a = \bar{y} - b\bar{x} \tag{8}$$

Со равенките (7) и (8) ги дефинираме коефициентите (a и b) на регресионата права, која го задоволува условот за оптималност, кој порано е поставен.

При проценката на вредноста на зависно променливата со помош на регресионата права ни служи стандардната девијација, пресметана врз основа на одстапувањата на оригиналните вредности на зависно променлива, од вредностите на таа променлива пресметани со помош на регресионата права. Тие одстапувања ја претставуваат грешката на проценката. На Сл. 1 може јасно да се види дека секое одстапување на оригиналните вредности на зависно променливара *(y)* од аритметичката средина y се состои од два дела: од одстапување на вредноста пресметана со помош на регресионата права y_R од аритметичката средина y, односно $y_R - \bar{y}$, и од одстапувањата на оригиналната вредност *(y)* од пресметаната вредност y_R, односно $(y - y_R)$, односно:

$$(y - \bar{y}) = (y_R - \bar{y}) + (y - y_R) \tag{9}$$

Првиот дел на одстапувањето во изразот (9) се препишува на врската која постои помеѓу појавите x и y, па тој дел од одстапувањето може да се објасни со постоењето на таа врска. Другиот дел на одстапувањето во изразот (9) останува необјаснет со врската помеѓу појавите. Тоа одстапување се вика резидуално одстапување. Ако ова одстапување е еднакво на *0*, тогаш оригиналните вредности на зависно променливата ќе се наоѓаат на регресионата права, а тоа значи дека регресионата права ќе биде потполно прецизна процена на зависно променливата. Ако резидуалното одстапување се зголеми, тогаш оригиналните вредности на зависно променливата се повеќе ќе одстапуваат од регресионата права. Со други зборови, резидуалното одстапување може да биде добра мера на прецизноста на проценатa на зависно променливата со помош на регресионата права. Но, такви одстапувања има толку колку што има парови на оригинални вредности. За да се дојде до показател изразен со еден број, мора да се земе просекот на тие одстапувања. Аритметичка средина не може да се земе, бидејќи таа секогаш ќе е *0*, поради тоа треба да се земе просекот на тие одстапувања. Ако равенката (9) ја квадрираме и сумираме се добива

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(y_{Ri} - \bar{y})^2 + \sum_{i=1}^{N}(y_i - \bar{y}_{Ri})^2 \tag{10}$$

Изразот (10) е равенка на анализа на варијациите, таа покажува дека сумата на квадратите на отстапувања на оригиналните вредности на зависно променливата од нивната аритметичка средина е составена од два дела: од сумата на квадратите на одстапување на вредностите проценети со помошта на регресионата права и аритметичката средина и од сумата на квадратите на одстапување на оригиналните вредности на зависно променливата од аритметичката средина. Ако равенката (10) ја поделиме со бројот N (бројот на оригинални вредности), левата страна од равенката (9) ја претставува вкупната варијација на променливата *(y)*. Првиот член од десната страна претставува збир на квадратите на оној дел на одстапувањето на зависно променливата кој е објаснет со врската помеѓу појавите x и y. Ако тој член го поделиме со N го добиваме објаснетиот дел на варијацијата σ_0. Вториот член од десната страна на изразот (10) претставува збир на квадратите на оној дел на одстапување на зависно променливата кој остана необјаснет. Ако тој израз го поделиме со бројот на набљудувања N се добива необјаснетиот дел на варијацијата σ_N, што всушност е стандарната грешка на проценката:

$$\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N} = \frac{\sum_{i=1}^{N}(y_{Ri} - \bar{y})^2}{N} + \frac{\sum_{i=1}^{N}(y_i - \bar{y}_{Ri})^2}{N}$$

$$\sigma \quad = \quad \sigma_0 \quad + \quad \sigma_N \tag{11}$$

За дефинирање на јакоста, односно корелациониот коефициент, на врската помеѓу истражуваните појави се поаѓа од изразот (10). Коефициентот на корелација се детерминира како :

$$C_C = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{Ri} - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}} \qquad (12)$$

Пеарсновиот коефициент на корелација (12) е дефиниран како квадратен корен од односот помеѓу објаснетиот дел на варијацијата и вкупната варијација. Овој коефициент се движи во интервалот од -1 до 1 што беше порано изложено.

Во наредната табела Таб. 1. нагледно ќе биде даден начинот на пресметнување на резултатите од истражуваните модели.

Табела 1.

N	x_i	y_i	x_iy_i	x_i^2	$(x_i-x^*)^2$	$(y_i-y^*)^2$	$(y_{Ri}-y^*)^2$	$(y_i-y_R^*)^2$
1	x_1	y_1	x_1y_1	x_1^2	$(x_1-x^*)^2$	$(y_1-y^*)^2$	$(y_{R1}-y^*)^2$	$(y_1-y_R^*)^2$
2	x_2	y_2	x_2y_2	x_2^2	$(x_2-x^*)^2$	$(y_2-y^*)^2$	$(y_{R2}-y^*)^2$	$(y_2-y_R^*)^2$
3		y_3	x_3y_3	x_3^2	$(x_3-x^*)^2$	$(y_3-y^*)^2$	$(y_{R3}-y^*)^2$	$(y_3-y_R^*)^2$
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
N	x_N	y_N	x_Ny_N	x_N^2	$(x_N-x^*)^2$	$(y_N-y^*)^2$	$(y_{RN}-y^*)^2$	$(y_N-y_R^*)^2$
	Σx_i	Σx_i	Σx_iy_i	Σx_i^2	$\Sigma(x_i-x^*)^2$	$\Sigma(y_i-y^*)^2$	$\Sigma(y_{Ri}-y^*)^2$	$\Sigma(y_i-y_R^*)^2$

	параметар	проценка	стандардна грешка
$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ $y_R = a + bx$	интерцепт a	$\bar{y} - b\bar{x}$	$\sqrt{\frac{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2 \sum_{i=1}^{N} x_i^2}{N^2 \sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}}$
	косина b	$\frac{\sum_{i=1}^{N} x_iy_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}$	$\sqrt{\frac{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}{N\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}}$

АНАЛИЗА НА ВАРИЈАЦИИТЕ

вид на отстапување	сума на квадрати	сс	проценка на варијациите
модел	$\Sigma(y_{Ri} - y^*)^2$	1	$\Sigma(y_{Ri} - y^*)^2$
грешка	$\Sigma(y_i - y_{Ri})^2$	n-2	$\Sigma(y_i - y_{Ri})^2/(n-2)$
вкупно	$\Sigma(y_i - y^*)^2$	n-1	-

Summary

This paper deals with the application of regression analysis for determination of correlation links between same parameters of the Earth's crust. Liner regression is used because it is proved that it is close to reality. We also present detailed description of the applied mathematical model.

Литература

1. Бикел П, Доксам К. Математичкаѓ статистика I и II, Москва (1983)
2. Боровков А.А. Математичкаѓ статистика-оценка параметров проверка гипотез Москва (1984)
3. Голцман М.Ф. Статистическаѓ интерпретация геофизических данны́х. Ленинград, (1981)
4. Когам Р. И. Статистические рангобы́е критерии в геологии, Москва (1983)
5. Krtolica R. Analiza matematičkih modela stohastičkih sistema sa razdrobljenim parametrima , Beograd, (1979)
6. Milošević V. M. Teorijska statistika i teorija statističkog zaključivanja Beograd, (1983)
7. Stojanović S. Matematička statistika, Beograd, (1980)
8. Цимелъзон И. А. Ткачук И. З. Математические методы́ идентификации модельи в геологии, Москва (1983)