Linearna korelacija

Korelacija je mjera linearne zavisnosti dviju serija podataka $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$. Drugim riječima, ako su točke $(x_1,y_1), (x_2,y_2), \ldots (x_n,y_n)$ grupirane oko regresijskog pravca, onda govorimo da su podatci **korelirani** (**linearno korelirani**). Na osnovi toga govori se da su pripadne veličine x,y korelirane. Razina koreliranosti mjeri se **koeficijentom korelacije**

$$r := \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Vrijedi:
(1) $-1 \le r \le 1$

(2) Ako je r pozitivan onda je i koeficijent smjera regresijskog pravca pozitivan i obratno (ako je r>0 onda je a>0, a ako je r<0 onda je a<0)

(3) Što je r bliži 1 ili -1 to su veličine značajnije linearno korelirane, tj. podatci su bliže regresijskom pravcu, a što je r bliži 0, podatci su razbacaniji.

(4) Ako je r = 1 ili r = -1 onda su sve točke (x_i,y_i) na regresijskom pravcu, tj. podatci su potpuno linearno zavisni.

(5) Ako je r=0 onda nema nikakve linearne zavisnosti među veličinama.

Primjer 1. Odredimo koeficijent korelacije za podatke

x_i |	0	1	2	3	4
y_i |	6	3	1	-2	-3

iz Primjera 2. u lekciji Metoda najmanjih kvadrata.

Napravimo tablicu poput one za metodu najmanjih kvadrata, samo dodajmo redak s y_i^2.

x_i |	0	1	2	3	4 ||	10
y_i |	6	3	1	-2	-3 ||	5
x_i^2 |	0	1	4	9	16 ||	30
x_iy_i |	0	3	2	-6	-12 ||	-13
y_i^2 |	36	9	1	4	9 ||	59

$$r = \frac{5\cdot(-13) - 5\cdot 10}{\sqrt{5\cdot 30 - 10^2}\cdot\sqrt{5\cdot 59 - 5^2}} = \frac{-115}{30\sqrt{15}} \approx -0.98976$$ (na pet decimala)

što je visoka razina linearne zavisnosti. Uočite da je r<0, što je u skladu s tim da je a<0 (sjetite se da je jednadžba regresijskog pravca y = -2.3x+5.6).

Napomena. Naredba LinReg uz podatke o parametrima također bi nam izbacila i vrijednost koeficijenta *r*.

Podatci u sljedećem primjeru su vrlo nisko linearno korelirani.

Primjer 2. Odredimo koeficijent korelacije za podatke

x_i |	-3	-2	-1	0	1	2	3
y_i |	0	2	-2	3	2	1	-1

Unesimo podatke u kvadratnu 7x7 mrežu kao na slici. Vidimo da podatci ne prate ni jedan pravac, pa procjenjujemo da je koeficijent korelacije blizu nule.

								||
x_i	-3	-2	-1	0	1	2	3	|| 0
y_i	0	2	-2	3	2	1	-1	|| 5
x_i^2	9	4	1	0	1	4	9	|| 28
x_iy_i	0	-4	2	0	2	2	-3	|| -1
y_i^2	0	4	4	9	4	1	1	|| 23

Sad je $r = \frac{7 \cdot (-1) - 0 \cdot 0}{\sqrt{7 \cdot 28 - 0^2} \cdot \sqrt{7 \cdot 23 - 5^2}} \approx$ -0.042875 (na šest decimala)

što je praktički jednako nuli. Također, može se provjeriti da je pripadna suma kvadrata odstupanja za linearnu regresiju jednaka 19. 392 857 (na šest decimala), što također upućuje na vrlo slabu linearnu vezu.

Linearnu korelaciju ne treba shvatiti kao jedini oblik zavisnosti dviju veličina (serija podataka). Dvije veličine mogu biti vrlo jasno zavisne, a da im je koeficijent (linearne) korelacije jednak nuli; to samo znači da su one linearno nekorelirane. To pokazuje sljedeći primjer.

Primjer 3. Odredimo koeficijent korelacije za podatke

x_i |	-3	-2	-1	0	1	2	3
y_i |	9	4	1	0	1	4	9

Ucrtavanjem podataka vidimo da oni ne prate ni jedan pravac. Kako je $\sum x_i = 0$ i $\sum x_i y_i = 0$, vidimo da je r=0. Dakle, podaci su linearno nekorelirani. S druge strane, oni su zavisni. Naime, povezani su relacijom y = x^2 (točke su na paraboli).

Često se postavlja pitanje koji koeficijenti znače visoku, koji nisku, a koji srednju linearnu koreliranost. Na to pitanje nema jasnog odgovora. On ovisi i o znanstvenom području na koje se primjenjuje, a unutar znanstvenog područja na konkretan problem koji se razmatra. Na

primjer, u psihologijskim istraživanjima, u pravilu, čim je r>0.5 smatra se da je koreliranost značajna, a ako je r>0.8 vrlo značajna, dok u preciznim fizikalnim ili kemijskim istraživanjem često niti r=0.9 ne upućuje na značajnu koreliranost.

Primjer 4. U sljedećoj tablici su u prvom redku bodovi prvih 9 najboljih rezultata postignutih iz kolokvija na Matematici 1, a u drugoj su odgovarajući bodovi iz Matematike 2.

x_i |	103	93	84	81	81	80	79	79	78
y_i |	99	73	82	85	77	79	73	55	83

Odredimo regresijski pravac i koeficijent korelacije. Komentirajmo rezultate.

Da dobijemo predodžbu, podatke predočavamo u koordinatnom sustavu.

Predviđamo pozitivan koeficijent korelacije jer podatci prate glavnu dijagonalu (ali ne visok, jer podatci variraju). Procijenjujemo da je koeficijent regresijskog pravca nešto manji od 1. Zadatak se može izraditi prema uzoru na prijašnje primjere. Mi ćemo se poslužiti grafičkim kalkulatorom, koji ima gotov program za metodu najmanjih kvadrata i linearnu korelaciju. Dobijemo, zaokružujući na dvije decimale,
$y= 0.79x+12.29$ i $r = 0.56$.
Komentar. Dobili smo $a=0.79<1$, što je u skladu s činjenicom da su rezultati Matematike 2, nešto niži od rezultata Matematike 1. Koeficijent korelacije nije blizu 1, ali je veći od 0.5, što, pri ovakvoj problematici upućuje na nezanemarivu korelaciju.
Suma kvadrata odstupanja jednaka je, na četiri decimale, 763.7222 što izgleda veliko, ali taj rezultat treba tumačiti tako da je prosječno odstupanje oko 9 bodova, što i nije tako veliko.

Zašto nas je u ovom primjeru zanimala linearna korelacija među rezultatima?
Zato što intuitivno prihvaćamo da će rezultati iz Matematike 2 biti približno proporcionalni onima iz Matematike 1, tj. da odprilike jednake razine usvajanja nekog znanja uvjetuju odprilike jednake razine usvajanja novog znanja koje počiva na starom. Naravno da to ne vrijedi za svakog konkretnog pojedinca, već u prosjeku.

Obrazloženje formule za koeficijent korelacije.

Seriju od n podataka možemo shvatiti kao vektor u n-dimenzionalnom prostoru. Sjetimo se što za vektore znači da su linearno zavisni (linearno korelirani).

Slike upućuju na to da su dva vektora linearno zavisna (kolinearna, proporcionalna) ako i samo ako je kut među njima nul-kut ili ispruženi kut (od 180 stupnjeva). Što je kut bliže 0 ili 180, to vektore možemo smatrati više linearno koreliranim, a što je bliže pravom kut, tj. 90 stupnjeva, manje koreliranim. Za kut φ među vektorima $\mathbf{u}=(u_1,u_2,...,u_n)$ i $\mathbf{v}= (v_1,v_2,...,v_n)$ vrijedi

$$\cos\varphi = \frac{u_1v_1 + u_2v_2 + \ldots + u_nv_n}{\sqrt{\sum u_i^2} \cdot \sqrt{\sum v_i^2}}$$

(izraz u brojniku je **skalarni produkt** vektora, a u nazivniku su **norme-duljine** vektora).

Vrijedi:

(1) $-1 \le \cos\varphi \le 1$

(2) Ako je $\cos\varphi > 0$ onda je kut među pravcima šiljast, a ako je $\cos\varphi < 0$ onda je taj kut tup

(3) Što je $\cos\varphi$ bliži 1 ili -1 vektori su sve bliže tome da budu proporcionalni (linearno zavisni), a što je $\cos\varphi$ bliži 0, vektori su to manje kolinearni.

(4) Ako je $\cos\varphi = 1$ ili $\cos\varphi = -1$ onda su vektori linearno zavisni; tada je kut od 0 stupnjeva i $\mathbf{v} = c\cdot\mathbf{u}$, za $c>0$, ili je kut od 180 stupnjeva i $\mathbf{v} = c\cdot\mathbf{u}$, za $c<0$.

(5) Ako je $\cos\varphi = 0$ onda su vektori okomiti pa su najudaljeniji od kolinearnosti.

Dakle, $\cos\varphi$ ima svojstva analogna onima koje ima *r*. To znači da je $\cos\varphi$ koeficijent kolinernosti dvaju vektora (slično kako je r koeficijent linearne korelacije dviju serija podataka).

Da bismo razmatranje s vektorima primijenili na razmatranje serija podataka, treba umjesto serija $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$ gledati pomaknute serije

$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$ i $y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y}$

gdje su $\bar{x}$ i $\bar{y}$ aritmetičke sredine podataka.

Naime, ako od tražene linearne veze

y=ax+b

oduzmemo istinitu relaciju

$\bar{y} = a\bar{x} + b$,

dobit ćemo proporcionalnost

$y - \bar{y} = a(x - \bar{x})$.

Ako bi zadane točke zaista zadovoljavale tu jednadžbu, bilo bi

$y_1 - \bar{y} = a(x_1 - \bar{x})$

$y_2 - \bar{y} = a(x_2 - \bar{x})$

.

$y_n - \bar{y} = a(x_n - \bar{x})$.

Međutim, to vrijedi samo približno, a za mjeru te približnosti razumno je uzeti kosinus kuta među vektorima $\mathbf{u}:=(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})$ i $\mathbf{v}:=(y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$.

Zato je prirodno koeficijent korelacije definirati kao

$$r := \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}}$$

Na prvi pogled, to nije ona formula koju smo napisali na početku, međutim, dobije se:

$$r := \frac{(\sum x_i y_i) - \bar{x}(\sum y_i) - \bar{y}(\sum x_i) + n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2) - 2\bar{x}(\sum x_i) + n\bar{x}^2} \cdot \sqrt{(\sum y_i^2) - 2\bar{y}(\sum y_i) + n\bar{y}^2}}$$

$$= \frac{(n\sum x_i y_i) - (\sum x_i)(\sum y_i) - (\sum y_i)(\sum x_i) + (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - 2(\sum x_i)(\sum x_i) + (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - 2(\sum y_i)(\sum y_i) + (\sum y_i)^2}}$$

$$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

kako smo i na početku imali.