Regresszióanalízis

Ha valamely kisérlet kapcsán két véletlen mennyiséget: X és Y-t is vizsgálunk, akkor elsősorban az érdekel bennünket, hogy függetlenek-e egymástól a kérdéses v.v-k vagy pedig értékeik között valamilyen kapcsolat áll fenn.
Gyakori eset, hogy X és Y értékei között nincs determinisztikus kapcsolat, egyikből a másik nem mondható meg egyértelműen, de felvett értékeik között bizonyos tendencia mutatkozik, pl. hogy X nagy értékei Y nagy értékeivel és X kis értekei Y kis értékeivel jelentkezik együtt.
Ha pl. X a Duna vízállása egy adott pontban mérve, és Y pedig a talajvízszint egy Dunához közeli pontban, akkor ha X és Y közt nincs is szigorú függvénykapcsolat, de közös tendencia tapasztalható közöttük.
Ezt a kapcsolatot fejezzük ki a kovarianciával ill. a korrelációs együttható segítségével.

Legyen X és Y valószínűségi változók. Ekkor X, Y <u>kovarianciája</u>

$$E(XY) - E(X)E(Y)$$

jelölés: cov(X,Y)

Megjegyzések:

1. Tudjuk, hogy ha X és Y független v.v-k, akkor $E(XY) = E(X)E(Y)$, ezért ekkor $cov(X, Y) = 0$.
2. Megadható olyan X, Y, hogy $cov(X, Y) = 0$, de X, Y v.v-k nem függetlenek.

Példa: Legyen X és Y olyan v.v-k, hogy

$$P((X,Y)=(0,0))=0,5$$

$$P((X,Y)=(2,1))=0,3$$

$$P((X,Y)=(3,2))=0,2.$$

Ekkor

$$P(X=0)=0,5, \quad P(X=2)=0,3, \quad P(X=3)=0,2 \Rightarrow$$

$$E(X)=1,2$$

$$P(Y=0)=0,5, \quad P(Y=1)=0,3, \quad P(x=2)=0,2 \Rightarrow E(Y)=0,7$$

és

$$P(XY=0)=0,5, \quad P(XY=2)=0,3, \quad P(XY=6)=0,2 \Rightarrow$$

$$E(XY)=1,8$$

így

$$cov(X,Y)=1,8-1,2*0,7=0,96$$

Legyen X és Y valószínűségi változók. Ekkor X, Y korrelációs együtthatója

$$\frac{cov(X, Y)}{D(X)D(Y)}$$

jelölés: r(X,Y)

Megjegyzések:

1. Ha X és Y független v.v-k, akkor $r(X, Y) = 0$
2. Vannak nem független X, Y v.v-k, melyekre $r(X, Y) = 0$.
3. Ha X, Y v.v-k esetén $r(X, Y) = 0$, akkor X és Y-t korrelálatlannak mondjuk.
4. Megmutatható, hogy minden X, Y esetén

$$-1 \leq r(X, Y) \leq 1$$

5. Az $|r(X, Y)| = 1$ pontosan akkor teljesül, ha $Y = aX + b$, tehát a korrelációs együttható azt mutatja meg, hogy mennyire lineáris a kapcsolat az X és Y v.v-k között.

Az előző példában

$$E(X^2) = 0^2 * 0,5 + 2^2 * 0,3 + 3^2 * 0,2 = 3$$

$$D^2(X) = E(X^2) - E(X)^2 = 3 - 1,2^2 = 1,56 \Rightarrow D(X) = 1,25$$

Hasonlóan

$$E(Y^2) = 0^2 * 0,5 + 1^2 * 0,3 + 2^2 * 0,2 = 1,1$$

$$D^2(Y) = E(Y^2) - E(Y)^2 = 1,1 - 0,7^2 = 0,61 \Rightarrow D(Y) = 0,78,$$

ezért

$$r(X, Y) = \frac{cov(X, Y)}{D(X)D(Y)} = \frac{0,96}{1,25 * 0,78} = 0,96$$

A regresszióelmélet legfontosabb feladata prognózis felállítása az Y v.v értékére, ha X v.v-ról tudjuk, hogy milyen értéket vesz fel ($X = x$). Erre az optimális választ az ún. feltételes várható érték adja meg:

$$E(Y|X = x)$$

Minden szóba jöhet x-re ha megadjuk a feltételes várható értéket, akkor ezzel az ún. <u>regressziós görbét</u> határozzuk meg.

Példa: Jelölje X a Duuna vízgyüjtő területén bizonyos időtartam alatt lehulott csapadék mennyiségét, Y pedig a Duna vízállását adott szelvényen. Előjelzést szeretnénk adni a vízállás alakulására. Ha tudjuk, hogy $X = x$, akkor Y v.v-ra vonatkozó legjobb jóslatunk az Y v,v feltételes várható értéke, az $E(Y|X = x)$ számérték lesz.

Sajnos a regressziós görbe kiszámítása általában reménytelenül nehéz feladat.
Fontos speciális eset az, amikor az X, Y együttesen ún. kétdimenziós normális eloszlást követ. Ekkor a regressziós görbe a következő egyenes:

$$E(Y|X = x) = E(Y) + r(X, Y)\frac{D(Y)}{D(X)}(x - E(X)),$$

tehát egy olyan egyenes, mely átmegy az $(E(X), E(Y))$ ponton és meredeksége $r(X, Y)\frac{D(Y)}{D(X)}$.

Általában is egy fontos kérdés annak megválaszolása, hogy ha adott $X = x$ esetén milyen a, b valós számokkal érdemes Y-ra

$$ax + b$$

értékkel tippelni.

Feladat: Határozzuk meg azokat az a, b számokat, hogy az

$$E((Y - (aX + b))^2)$$

várható érték minimális legyen.

Megmutatható, hogy ennek a minimuma az

$$a = r(X, Y)\frac{D(Y)}{D(X)}$$

$$b = E(Y) - E(X)r(X, Y)\frac{D(Y)}{D(X)}$$

ami másképp

$$\frac{y - E(Y)}{D(Y)} = r(X, Y)\frac{x - E(X)}{D(X)}$$

alakba is írható. Ezt az egyenest regressziós egyenesnek hívjuk.

Ha az X, Y v.v-kra n megfigyelést végzünk, akkor a regressziós egyenest az

$$\frac{y-\overline{y}}{S_y^*} = \widehat{r}\frac{x-\overline{x}}{S_x^*},$$

ahol $\overline{x}$ az X_i-k átlaga, S_x^* a korrigált empirikus szórás és az empirikus korrelációs együttható:

$$\widehat{r} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \overline{xy}}{S_X^* S_Y^*}$$

azaz, ha (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ mintához írjuk fel az

$$y = ax + b$$

regressziós egyenes, akkor

$$a = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$b = \frac{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

Példa: Tegyük fel, hogy az (X, Y)-ra a következő 10 elemű mintánk van:

$$(3,10), \quad (2,7), \quad (6,16), \quad (4,10), \quad (3,12)$$

$$(1,4), \quad (6,17), \quad (5,13), \quad (4,14), \quad (2,9)$$

Határozza meg a regressziós egyenest!

Megoldás:

$$\sum_{i=1}^{10} x_i = 36, \quad \sum_{i=1}^{10} x_i^2 = 156$$

$$\sum_{i=1}^{10} y_i = 112, \quad \sum_{i=1}^{10} x_i y_i = 461$$

ezért

$$a = \frac{10 * 461 - 36 * 112}{10 * 156 - 36^2} = 2,19$$

$$b = \frac{156 * 112 - 36 * 461}{10 * 156 - 36^2} = 3,32$$