

## IV. modul: Korreláció- és regresszió-számítás

### 9. lecke Kétváltozós korreláció- és regresszió-számítás: Lineáris

#### 9.1. Kétváltozós korreláció-számítás

A mennyiségi ismérvek között meglévő kapcsolat szorosságát és irányát a korreláció-számítással állapíthatjuk meg.

Ha a mennyiségi ismérvek között nincs kapcsolat, azaz függetlenek egymástól, akkor a szorosság mérőszáma a korrelációs együttható ( $r$ ) nullával egyenlő.

Ha egyértelmű kapcsolat van a két mennyiségi ismerv között, akkor a korrelációs együttható ( $r$ ) értéke  $(-1;1)$  közé esik. Ha a korrelációs együttható pozitív, akkor a két ismerv közötti kapcsolat azt jelenti, hogy az egyik ismerv növekedése maga után vonja a másik ismerv növekedését. Negatív kapcsolat esetén az egyik ismerv növekedése a másik ismerv csökkenését okozza.

A mennyiségi ismérvek eloszlásainak speciális paramétere a kovariancia, amely az átlagtól való eltérések szorzatának számtani átlaga. Az együttes szóródás nagyságrendjét jellemzi, az előjele pedig kifejezi a kapcsolat irányát.

$$C = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n}$$

A kovariancia felhasználásával kiszámítható a lineáris korrelációs együttható:

$$r = \frac{C}{\sigma_x * \sigma_y}$$

ahol

$C$ : a kovariancia,

$\sigma_x$ : az egyik változó szórása

$\sigma_y$ : a másik változó szórása

Ha a változók szórására nincs külön szükségünk, akkor a korrelációs együttható másképp is kiszámolható:

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

Vagy:

$$r = \frac{\sum (x_i * y_i - n * \bar{x} * \bar{y})}{\sqrt{(\sum x_i^2 - n * \bar{x}^2)(\sum y_i^2 - n * \bar{y}^2)}} = \beta_1 * \frac{\sigma_x}{\sigma_y}$$

A korrelációs együttható négyzete a determinációs együttható, amelyet százalékban adunk meg, és azt fejezi ki, hogy az egyik ismerv hány százalékban befolyásolja a másik ismerv változását.

$$R = r^2$$

## 9.2. Kétváltozós regresszió-analízis

Ha két mennyiségi változó közötti függőségi viszonyt valamilyen matematikai képlettel írunk le, akkor regresszió-analízisről beszélünk.

Az ismérvek közötti függőségi viszonyok feltárásával, az összefüggésekben rejlő tendenciák matematikai függvényekkel történő leírásával a **regresszióanalízis** foglalkozik.

A gyakorlati elemző munkában a korreláció- és regresszió-számítást általában együtt, egymást kiegészítve alkalmazzák.

A regresszió-számítás a statisztikai modellezés egyik egyszerű eszköze, ám egyszerűsége ellenére szinte minden, korábban megismert módszertani elemet felhasznál.

A regresszió-számításkor általában meg szoktuk különböztetni a két- a többváltozós eseteket. A kétváltozós regresszió analíziskor két változó kapcsolatát vizsgáljuk, az  $x$  változó az egyik, ez a magyarázó változó, és az  $y$ -lal jelölt (eredményváltozó) változó a másik, amelynek alakulását  $x$  változó befolyásolja. A regresszió-számítás során feltételezzük, hogy az eredményváltozónk ( $y$ ) sztochasztikus kapcsolatban áll a magyarázó változóval  $Y=f(x)$

A regresszió típusának kiválasztásához először ábrázolni kell az adatokat, mivel az ismérvek közötti kapcsolat lényegének megismerésében fontos szerepet játszik a grafikus ábrázolás. Általában pontdiagramot készítünk.

Ezután meg kell határozni a regresszió típusát, ehhez azonban szükséges az adott terület szakmai ismerete is. Lineáris esetben az alábbi függvényt használjuk:

$$y = \beta_0 + \beta_1 \cdot x$$

A függvénytípus kiválasztásával azonban a regressziós függvény meghatározásának problémája még nincs megoldva. A végtelen sok egyenes közül azt az egyet keressük, amely az összefüggést a lehető legjobban leírja. A függvény paramétereit a legkisebb négyzetek módszere segítségével határozzuk meg, vagyis:

$$S = \sum (y_i - y'_i)^2 \Rightarrow \text{minimum}.$$

Azt a becslőfüggvényt keressük tehát, amelyik a mintabeli és a számított értékek közötti különbségek négyzetösszege minimális. Lineáris összefüggés esetén a függvényünk:

$$y = \beta_0 + \beta_1 \cdot x \quad \text{vagy} \quad y = a + b \cdot x$$

Ezt behelyettesítve  $S$ -egyenletébe a következőt kapjuk:

$$S = \sum (y_i - \beta_0 - \beta_1 \cdot x)^2$$

A függvénynek ott van minimuma, ahol a két együttható szerinti parciális differenciáhányadosa egyenlő nullával. Az egyenlet levezetéséből azt kapjuk, hogy:

$$\beta_1 = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i * y_i - n * \bar{x} * \bar{y}}{\sum x_i^2 - n * \bar{x}^2}$$

$$\beta_0 = \bar{y} - \beta_1 * \bar{x}$$

A  $\beta_0$  paraméter azt fejezi ki, hogy az  $x=0$  helyen a függvény éppen ezt az értéket veszi fel, ha a nulla szerepel  $x$  lehetséges értékei között.

A  $\beta_1$  paraméter geometriai értelemben az egyenes meredekségét meghatározó iránytangens, regressziós együtthatóként választ ad arra, hogy az  $x$  változó egységnyi változása átlagosan mekkora változást okoz az  $y$  változóban.

A kétváltozós kapcsolatok esetében megvizsgáljuk, hogy a becslőfüggvény mennyire közelítette meg a mintabeli tapasztalati értéket. Ezt fejezi ki az  $\varepsilon$ -érték, azaz a reziduális szórás, amelyet a regressziós becslés abszolút hibája is egyben:

$$\sigma_e = \sqrt{\frac{\sum (y_i - y'_i)^2}{n - 2}} = \sqrt{\frac{\sum e_i^2}{n - 2}}$$

Ami tulajdonképpen az  $x_i$  pontban vett mintabeli  $y_i$  értékek és az adott  $x_i$  ponthoz tartozó becsült  $y'_i$  értékek közötti eltérések négyzetösszegének a gyöke.

A gyakorlatban kiszámítjuk ennek relatív nagyságát is:

$$CV_{\sigma_e} = \frac{\sigma_e}{\bar{y}} * 100$$

A relatív hiba kifejezi, hogy a regressziós becslések értékei átlagosan hány százalékkal térnek el az eredményváltozó megfigyelt értékeitől. Általában jónak mondjuk a becslőfüggvény illeszkedését, ha:  $CV_{\sigma_e} < 10-15\%$ .

Az eredményváltozó relatív változásának fontos szerepe van a közgazdasági elemzésekben. A relatív változást fejezi ki a rugalmassági együttható:

$$E = \frac{dy}{dx} * \frac{x_0}{y_0}$$

Az  $x$ -magyarázóváltozó adott értékének 1%-os növekedése átlagosan milyen változást eredményez az  $y$ -változó értékében. Ez az érték természetesen minden  $x$ -értékre kiszámítható:

$$E = \beta_1 * \frac{x_i}{y_i}$$

### 9.3. A varianciaanalízis alkalmazása a regressziószámításban

A regressziós együttható tesztelése mellett magának a regressziófüggvénynek a hipotézisellenőrzése is elvégezhető. Ez varianciaanalízissel történik. Ehhez az alábbi számításokat kell elvégezni (az eltérés négyzetösszegek számítását a II. Modulban már tanultuk):

$$SST = \sum (y_i - \bar{y})^2$$

$$SSR = \sum (y'_i - \bar{y})^2$$

$$SSE = \sum (y_i - y'_i)^2$$

$$SST = SSR + SSE$$

Különleges jelentősége van a reziduális négyzetösszegnek (SSE), mivel a megfigyelt értékeknek a regressziós függvény körüli szóródását fejezi ki.

Ha  $SSE=0$ , ez azt jelenti, hogy a függő változó teljes varianciája megmagyarázható a tényezőváltozó segítségével. Minden megfigyelt  $y_i$  érték a regressziós függvényen helyezkedik el. Egyéb tényezőknek nincs hatása az eredményváltozóra, vagyis az ismérvek között függvényyszerű kapcsolat van.

Ha az  $SSE \neq 0$ , akkor a két ismerv között sztochasztikus kapcsolat áll fenn. Minél nagyobb a reziduális négyzetösszeg értéke, annál nagyobb szerepet játszik a függő változó szóródásában.

A varianciatáblázat a következő:

A szórásnégyzet forrása	SS (SQ)	DF(FG)	MS(MQ)
Regresszió	$SSR = \sum (y'_i - \bar{y})^2$	1	$\frac{\sum (y'_i - \bar{y})^2}{1}$
Hibatényező	$SSE = \sum (y_i - y'_i)^2$	n-2	$\frac{\sum (y_i - y'_i)^2}{n-2}$
Teljes	$SST = \sum (y_i - \bar{y})^2$	n-1	-

$$H_0: \beta_1 = 0 \text{ illetve } H_1: \beta_1 \neq 0$$

A nullhipotézist F-próbával ellenőrizzük:

$$F_0 = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE}, \text{ ahol a számláló szabadságfoka } szf_1=1, \text{ a nevezőé pedig } szf_2=n-2. \text{ Ha}$$

számított F-érték kisebb, mint a táblázatbeli, akkor a nullhipotézist elfogadjuk, ellenkező esetben elvetjük.

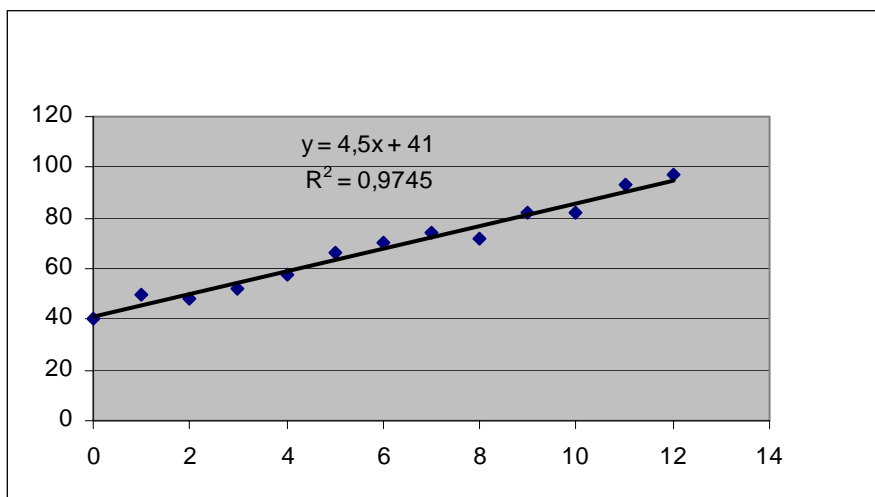
## Bemutató feladat

Egy 60 fős sokaságra a statisztika írásbeli dolgozatok eredményeit és a felkészülésre fordított idő nagyságát vizsgálva 13 elemű mintát vettek, az alábbi eredményekkel:

	<b>Idő (óra) x változó</b>	<b>Dolgozat eredménye (pont) y váltó</b>	$x^2$	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$	$(x - \bar{x}) * (y - \bar{y})$
	0	40	0	-6	36	-28	784	168
	1	50	1	-5	25	-18	324	90
	2	48	4	-4	16	-20	400	80
	3	52	9	-3	9	-16	256	48
	4	58	16	-2	4	-10	100	20
	5	66	25	-1	1	-2	4	2
	6	70	36	0	0	2	4	0
	7	74	49	1	1	6	36	6
	8	72	64	2	4	4	16	8
	9	82	81	3	9	14	196	42
	10	82	100	4	16	14	196	56
	11	93	121	5	25	25	625	125
	12	97	144	6	36	29	841	174
össz.	78	884	650	0	182	0	3782	819
átlag	6	68						

Feladat: számítsuk ki a regressziós függvény paramétereit!

Először ábrázoljuk az adatokat, és illesszünk rá függvényt



$$\beta_1 = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{819}{182} = 4,5$$

$$\beta_0 = \bar{y} - \beta_1 * \bar{x} = 68 - 4,5 * 6 = 41$$

$$y' = 41 + 4,5x$$

$\beta_1$ : azok a hallgatók, akik 1 órával többet tanultak, átlagosan 4,5 ponttal jobb eredményt értek el.

$\beta_0$ : akik nem készültek a dolgozatra, azok átlagosan 41 pontos eredményre számíthatnak.

Ezután összehasonlítjuk a tényleges és a regresszióval becsült adatokat.

Az  $y'$ -t úgy kapom meg, hogy az egyenletbe ( $y'=41+4,5x$ ) rendre behelyettesítem az  $x$ -értékeit. Ezután minden eredeti  $y$ -ból kivonom a kiszámított  $y'$ -értéket, majd az eredményt négyzetre emelem.

Idő (óra) x változó	Dolgozat eredménye (pont) y váltó	$y'$	$y-y'$	$(y-y')^2$
0	40	41	-1	1
1	50	45,5	4,5	20,25
2	48	50	-2	4
3	52	54,5	-2,5	6,25
4	58	59	-1	1
5	66	63,5	2,5	6,25
6	70	68	2	4
7	74	72,5	1,5	2,25
8	72	77	-5	25
9	82	81,5	0,5	0,25
10	82	86	-4	16
11	93	90,5	2,5	6,25
12	97	95	2	4
Összesen				96,5

$$\sigma_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{96,5}{13 - 2}} = 2,96$$

$$CV_{\sigma_e} = \frac{\sigma_e}{\bar{y}} * 100 = \frac{2,96}{68} * 100 = 4,35\%$$

A becslőfüggvény illeszkedése jónak mondható.

$$r = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}} = \frac{819}{\sqrt{182 * 3782}} = 0,987$$

$$R^2 = 97,45 \%$$

A tanulásra fordított idő 97,45%-ban meghatározza az elért pontszámot.

$$SST = 3782$$

$$SSR = 3685,5$$

$$SSE = 96,5$$

A szórásnégyzet forrása	SS (SQ)	DF(FG)	MS(MQ)
Regresszió	3685,5	1	3685,5
Hibatényező	96,5	11	8,77
Teljes	3782	12	-

$$F_0 = \frac{MSR}{MSE} = \frac{3685,5}{8,77} = 420,24$$

A táblázatbeli F-érték

$$F_{11(0,95)}^1 = 4,84$$

A  $H_0$  szerinti feltevést 5%-os szignifikancia szinten elvetjük. Megállapítható, hogy a  $\beta_1$  paraméter értéke szignifikánsan különbözik nullától, vagyis van kapcsolat a két ismerv között.