

10. lecke: Két és többváltozós korreláció- és regresszió-számítás: exponenciális és hatványkitevős függvények

Két mennyiségi ismerv közötti kapcsolat leírására sok esetben nem alkalmas a lineáris függvény. Ha az x-változó y-változóra gyakorolt hatásának mértéke függ az x-változó nagyságától, akkor a lineáris regresszió nem alkalmas az adatok közötti kapcsolat elemzésére.

10.1. Hatványkitevős regressziós függvény

Ha a független változó szorzatos növekedésével a függő változó is szorzatosan változik, akkor regressziós függvény az alábbi:

$$y = \beta_0 \cdot x^{\beta_1}$$

Olyan esetekben alkalmazzuk, amikor az x és y változók logaritmusai között van lineáris összefüggés.

β_1 regressziós együttható azt fejezi ki, hogy az x magyarázó változó egységnyi relatív (1%-os) változása mekkora relatív (hány százalékos) változást idéz elő az eseményváltozóban.

Megoldásához linearizálni kell a regressziós függvényt:

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

Látható, hogy az x és az y változók logaritmusai között lineáris a kapcsolat.

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

$$\lg \beta_0 = \overline{\lg y} - \beta_1 * \overline{\lg x}$$

Vezessünk be új ismeretleneket:

$$\lg y = Y; \lg x = X; \lg \beta_0 = B$$

Így a függvényünk az alábbi:

$$Y = B + \beta_1 \cdot x$$

A regressziós együtthatók így már a tanultak szerint számíthatóak:

$$\beta_1 = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$B = \bar{Y} - \beta_1 * \bar{X}$$

$$\beta_0 = 10^B$$

A paraméterek jelentése:

- A β_0 : ha az x=1 része a függvény értelmezési tartományának, akkor van jelentése, azaz az x=1 helyen felvett regressziós érték.
- A β_1 : a magyarázóváltozó 1%-os változása az eredményváltozásban éppen A β_1 %-os változást okoz. A β_1 -együttható elaszticitási (rugalmassági) együttható is egyben, mivel megmutatja hogy az 1%-kal nagyobb x-értékhez hány százalékkal nagyobb vagy kisebb y-érték tartozik.

A kapcsolat szorosságát a korrelációs index fejezi ki:

$$I = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}}$$

$$e_i^2 = \left(y_i - \hat{y}_i \right)^2$$

A korrelációs index értéke:

$$0 \leq I \leq 1$$

A korrelációs index négyzetét százalékban fejezzük ki.

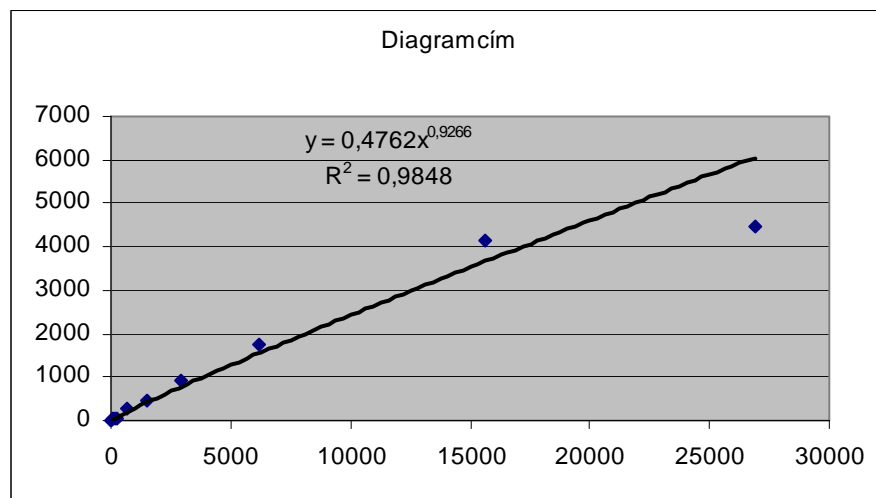
1. Bemutató feladat

Nézzük meg, hogy mennyire van hatással a gyökér súlya a gyökér felületére cukorrépa esetében.

Először az eredeti adatok logaritmusát kell meghatározni. A továbbiakban ezekkel az értékekkel dolgozunk, azaz X és Y értékei kerülnek a képletekbe.

	gyökér súlya x változó	gyökér felülete y váltó	lgx=X	lgy=Y	X ²	X-X _{átlag}	(X-X _{átlag}) ²	(Y-Y _{átlag})	(Y-Y _{átlag}) ²	(X-X _{átlag})*(Y-Y _{átlag})
	26870	4472	4,429268	3,650502	19,61841	1,461268	2,135303	1,222502	1,494511	1,786402
	15660	4152	4,194792	3,618257	17,59628	1,226792	1,505018	1,190257	1,416713	1,460198

	6180	1728	3,790988	3,237544	14,37159	0,822988	0,67731	0,809544	0,655361	0,666245
	2900	904	3,462398	2,956168	11,9882	0,494398	0,244429	0,528168	0,278962	0,261125
	1500	472	3,176091	2,673942	10,08756	0,208091	0,043302	0,245942	0,060487	0,051178
	650	260	2,812913	2,414973	7,912482	-0,15509	0,024052	-0,01303	0,00017	0,00202
	280	48	2,447158	1,681241	5,988582	-0,52084	0,271276	-0,74676	0,557649	0,388943
	130	39	2,113943	1,591065	4,468756	-0,85406	0,729413	-0,83694	0,700461	0,71479
	60	24	1,778151	1,380211	3,161822	-1,18985	1,41574	-1,04779	1,097861	1,24671
	30	12	1,477121	1,079181	2,181887	-1,49088	2,222719	-1,34882	1,819312	2,010925
össz.	54260	12111	29,68282	24,28308	97,37557	0,002824	9,268563	0,003085	8,081486	8,588538
átlag	5426	1211,1	2,968282	2,428308						



$$\beta_1 = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{8,5885}{9,2686} = 0,9266$$

$$B_0 = \bar{Y} - \beta_1 * \bar{X} = 2,4283 - 0,9266 * 2,9683 = -0,3222$$

$$\beta_0 = 10^{-0,3222} = 0,4762$$

$$\hat{y} = 0,4762 * x^{0,9266}$$

10.2. Exponenciális regressziós függvény

Ha az adatok közötti összefüggést a:

$$\hat{y} = \beta_0 * \beta_1^x$$

függvénnyel írható le, akkor exponenciális regresszióról beszélünk. Olyan esetekben alkalmazzuk, ha az y-változó növekedése arányos az adott helyen felvett x-változó értékével. Az exponenciális függvények esetében is igaz, hogy lineáris összefüggés van az eredményváltozó

logaritmus és a magyarázóváltozó között. Hasonlóan a hatványkitevős regresszióhoz, ebben az esetben is visszavezetjük lineáris regresszióra:

$$\lg \hat{y} = \lg \beta_0 + x * \lg \beta_1$$

$$Y = B_0 + B_1 * x$$

$$B_1 = \frac{\sum (x_i - \bar{x}) * (Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_1 = 10^{B_1}$$

$$B_0 = \bar{Y} - B_1 * \bar{x}$$

$$\beta_0 = 10^{B_0}$$

A paraméterek jelentése:

A β_1 regressziós paraméter arra ad választ, hogy az x-változó egységnyi növekedése hányszorosára változtatja az y-változó értékét.

A kapcsolat szorosságát a korrelációs index fejezi ki:

$$I = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}}$$

$$e_i^2 = (y_i - \hat{y}_i)^2$$

A korrelációs index értéke:

$$0 \leq I \leq 1$$

10.3. Választás a különböző regressziós egyenlet-típusok közül

Ugyanarra az adatsorra kiszámolva mindhárom regressziós függvényt, felvetődik a kérdés, hogy melyik jellemzi legjobban a változók kapcsolatát. A függvények kiválasztáshoz az egyenletek illeszkedési módszerét, azaz a legkisebb eltérések-négyzetét használjuk. Az az egyenlet illeszkedik legjobban az adatokra, ahol az $(y_i - \hat{y}_i)^2$ és az $(x_i - \hat{x}_i)^2$ is a legkisebb, illetve ahol a kapcsolat szorosságát kifejező mutató a legnagyobb.

10.4. Többváltozós regresszióanalízis

A társadalmi-gazdasági élet jelenségei összetettebbek, bonyolultabbak annál, mint amit két tényező összefüggése kifejez. Egy-egy jelenség változása általában több tényező változásával van összefüggésben. Az eredményváltozóra ható tényezők körének kibővítésével többszörös vagy többváltozós sztochasztikus kapcsolathoz jutunk.

A többváltozós regresszióanalízis segítségével több ismerv eredményváltozóra gyakorolt hatását vizsgáljuk. A kapcsolat az ismérvek száma szerint 3, 4 stb. változós, a függvény típusa szerint pedig lineáris és nem lineáris lehet.

A többváltozós lineáris regressziós modell az alábbi:

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_m \cdot x_m + \varepsilon$$

Példaként csak 3 változós lineáris kapcsolattal foglalkozunk, de az itt elmondottak akármenynyire változóra általánosíthatóak. Három változó esetén a függvényünk az alábbi:

$$Y' = f(x_1; x_2)$$

Azaz a regressziós függvényünk az alábbi lesz:

$$\hat{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$$

Az egyes paraméterek meghatározásához hasonlóan a kétváltozós regresszióhoz, itt is a legkisebb négyzetek módszerét alkalmazzuk.

Háromváltozós függvény minimumát kell keresni. Ebből meghatározhatóak a regressziós együtthatók.

A statisztikai elemzések során azonban főleg mintával találkozunk és elemzünk, így a minta segítségével becsüljük meg a regressziós függvényt. Ha az Y eredményváltozó értéke valószínűségi változó, de a magyarázóváltozók értékei ismertek, akkor standard lineáris regresszióknak nevezzük.

A legkisebb négyzetek módszerét használjuk a becslésre, és a feladat többváltozós szélsőérték számítással oldható meg. Így β' becslőfüggvénye az alábbi:

$$\beta' = (X^T \cdot X)^{-1} \cdot X^T \cdot y, \text{ feltéve, hogy } X^T \cdot X \text{ inverze létezik.}$$

Konkrét minta esetén a normálegyenletek az alábbiak:

- $\sum y_i = n \cdot b_0 + b_1 \cdot \sum x_{1i} + b_2 \cdot \sum x_{2i}$
- $\sum x_{1i} \cdot y_i = b_0 \cdot \sum x_{1i} + b_1 \cdot \sum x_{1i}^2 + b_2 \cdot \sum x_{1i} \cdot x_{2i}$
- $\sum x_{2i} \cdot y_i = b_0 \cdot \sum x_{2i} + b_1 \cdot \sum x_{1i} \cdot x_{2i} + b_2 \cdot \sum x_{2i}^2$

Vezessünk be új változókat:

- x_{1i} helyett $x_{1i} - \bar{x}_1 = d_{1i}$

- x_{2i} helyett $x_{2i} - \bar{x}_2 = d_{2i}$
- y_i helyett $y_i - \bar{y} = d_y$

A zérussal egyenlő összegek elhagyása után a normálegyenlet maradványaiból a paraméterek könnyen meghatározhatóak.

A 2. és 3. normálegyenletre:

- $\sum d_{1i} \cdot d_y = b_1 \cdot \sum d_{1i}^2 + b_2 \cdot \sum d_{1i} \cdot d_{2i}$
- $\sum d_{2i} \cdot d_y = b_1 \cdot \sum d_{1i} \cdot d_{2i} + b_2 \cdot \sum d_{2i}^2$

Ebből b_1 és b_2 könnyen meghatározható, a középiskolában tanult kétismeretes egyenletek megoldása szerint.

Az első egyenletből pedig meghatározható a b_0 .

$$b_0 = \bar{y} - b_1 \cdot \bar{x}_1 - b_2 \cdot \bar{x}_2$$

A regressziós függvény paramétereinek értelmezése

A becslőfüggvényünk:

$$y' = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$$

- β_1 : Ha x_1 értékét egy egységgel növeljük, miközben x_2 értékeit változatlanul hagyjuk, akkor az eredményváltozó becsült értéke éppen b_1 egységgel változik.
- β_2 : Az x_2 egységnyi növelésével, ha x_1 értékét változatlanul hagyjuk, akkor b_2 az eredményváltozás becsült értékében bekövetkező hatás.

A regressziós együttható tehát kifejezi, hogy egy adott tényezőváltozó egységnyi növekedése mekkora növekedést vagy csökkenést okoz az eredményváltozó becsült értékében, miközben a másik tényezőváltozók értéke változatlan. A regressziós együtthatók tehát 1-1 tényezőváltozó részleges hatását mutatják, ezért ezeket **parciális regressziós együtthatóknak** nevezzük.

A paraméterek értelmezésekor fontos hogy a multikollinearitást figyelembe vegyük. Multikollinearitásnak nevezzük a tényezőváltozók közötti lineáris kapcsolatot. Ez a kapcsolat zavarhatja az eredmények értelmezését.

A parciális regressziós együtthatóhoz hasonlóan a parciális rugalmassági együttható is értelmezhető. Ez a mutató arra ad választ, hogy egy adott tényezőváltozó relatív változása milyen relatív változást eredményez az y eredményváltozóban a másik változók változatlan színvona-
la mellett. Képlete:

$$E_{(y, x_j)} = \frac{dy}{dx_j} \cdot \frac{x_j}{y}$$

Regressziós függvényünkre alkalmazva:

$$E_{(y,x_j)} = \frac{b_j * x_j}{b_0 + b_1 x_1 + b_2 x_2}$$

Mint látható, a parciális rugalmassági együttható nagysága attól függ, hogy azt a tényezőváltozókat milyen színvonal mellett számítjuk ki.

2. Bemutató feladat

10 elemű minta alapján vizsgáljuk meg a szállítási időtartam (y), a szállítási távolság (x_1), és a szállítási tömeg (x_2) közötti összefüggést:

sorszám	száll. Időt. (y)	száll. Táv. (x ₁)	száll. tömeg. (x ₂)	dy	d ₁	d ₂	d ₁ ²	d ₂ ²	d ₁ *d ₂	d ₁ *d _y	d ₂ *d _y	d _y ²
1	10	4	4	-17	-11	-2	121	4	22	187	34	289
2	13	4	5	-14	-11	-1	121	1	11	154	14	196
3	8	2	2	-19	-13	-4	169	16	52	247	76	361
4	20	10	5	-7	-5	-1	25	1	5	35	7	49
5	27	19	5	0	4	-1	16	1	-4	0	0	0
6	35	20	7	8	5	1	25	1	5	40	8	64
7	22	16	6	-5	1	0	1	0	0	-5	0	25
8	40	20	7	13	5	1	25	1	5	65	13	169
9	45	25	9	18	10	3	100	9	30	180	54	324
10	50	30	10	23	15	4	225	16	60	345	92	529
összesen	270	150	60	0	0	0	828	50	186	1248	298	2006
Átlag	27	15	6									

A 2. és 3. normálegyenletre:

- $\sum d_{1i} * d_{yi} = b_1 * \sum d_{1i}^2 + b_2 * \sum d_{1i} * d_{2i}$
- $\sum d_{2i} * d_{yi} = b_1 * \sum d_{1i} * d_{2i} + b_2 * \sum d_{2i}^2$
- $1248 = 828b_1 + 186b_2$
- $298 = 186b_1 + 50b_2$

Megoldás a regressziós együtthatókra:

$$b_1 = 1,025$$

$$b_2 = 2,148$$

$$b_0 = \bar{y} - b_1 * \bar{x}_1 - b_2 * \bar{x}_2$$

$$b_0 = 27 - 1,025 * 15 - 2,148 * 6 = -1,263$$

A háromváltozós regresszió becslése:

$$y' = -1,263 + 1,025x_1 + 2,148x_2$$

A parciális rugalmassági együttható: az x_1 szerinti átlagos rugalmassága:

$$E_{(y,x_1)} = \frac{b_1 \cdot x_1}{b_0 + b_1 x_1 + b_2 x_2} = \frac{1,025 \cdot 15}{-1,259 + 1,025 \cdot 15 + 2,148 \cdot 6} = 0,569$$

Ez azt jelenti, hogy átlagos szállítási távolság és átlagos szállítandó tömeg esetében 1%-os szállítási tömegnövekedés 0,569%-os menetidő növekedést okoz.

az x_2 szerinti átlagos rugalmassága:

$$E_{(y,x_2)} = \frac{b_2 \cdot x_2}{b_0 + b_1 x_1 + b_2 x_2} = \frac{2,148 \cdot 15}{-1,259 + 1,025 \cdot 15 + 2,148 \cdot 6} = 0,477$$

Ez azt jelenti, hogy átlagos szállítási távolság és átlagos szállítandó tömeg esetében 1%-os szállítási tömegnövekedés 0,477%-os menetidő növekedést okoz.

10.5. Többváltozós korrelációs számítás

Célja a többváltozós korreláció szorosságának mérése. Kettőnél több változó esetén vizsgálható:

- páronként,
- továbbá páronként, de a többi változó hatásának kiszűrésével,
- végül pedig az eredményváltozó és az összes tényezőváltozó közötti szorosság is mérhető.

Páronkénti korrelációs együttható:

Két-két változó közötti szorosságot mérjük. A kiszámított korrelációs együtthatókat az R-korrelációs mátrixba rendezzük.

$$R = \begin{bmatrix} r_{yy} & r_{y1} & \dots & r_{ym} \\ r_{1y} & r_{11} & \dots & r_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{my} & r_{m1} & \dots & r_{mm} \end{bmatrix}$$

Ez egy szimmetrikus mátrix. A mátrix fődiagonálisában szereplő korrelációs együtthatók értéke 1.

$$R = \begin{bmatrix} 1 & r_{y1} & \dots & r_{y2} & \dots & r_{ym} \\ \vdots & 1 & \dots & r_{12} & \dots & r_{1m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & 1 & \dots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & 1 \end{bmatrix}$$

$$Y \text{ és } x_1 \text{ között: } r_{y1} = \frac{\Sigma d_1 * d_y}{\sqrt{\Sigma d_1^2 * \Sigma d_y^2}}$$

$$Y \text{ és } x_2 \text{ között: } r_{y2} = \frac{\Sigma d_2 * d_y}{\sqrt{\Sigma d_2^2 * \Sigma d_y^2}}$$

$$x_1 \text{ és } x_2 \text{ között: } r_{12} = \frac{\Sigma d_1 * d_2}{\sqrt{\Sigma d_1^2 * \Sigma d_2^2}}$$

3. Bemutató feladat

$$r_{y1} = \frac{\Sigma d_1 * d_y}{\sqrt{\Sigma d_1^2 * \Sigma d_y^2}} = \frac{1248}{\sqrt{828 * 2006}} = 0,9684$$

$$r_{y2} = \frac{\Sigma d_2 * d_y}{\sqrt{\Sigma d_2^2 * \Sigma d_y^2}} = \frac{298}{\sqrt{50 * 2006}} = 0,9409$$

$$r_{12} = \frac{\Sigma d_1 * d_2}{\sqrt{\Sigma d_1^2 * \Sigma d_2^2}} = \frac{186}{\sqrt{828 * 50}} = 0,9141$$

$$R = \begin{bmatrix} 1 \dots 0,9684 \dots 0,9409 \\ \dots 1 \dots 0,9141 \\ \dots 1 \end{bmatrix}$$

Szoros, pozitív irányú kapcsolat van a menetidő és a távolság, illetve a menetidő és a rakomány súlya között. A távolság és a rakomány súlya között is erős a sztochasztikus kapcsolat.

Parciális korrelációs együttható:

Megmutatja, hogy milyen szoros a kapcsolat valamelyik kiválasztott tényező és a függő változó között, ha a többi tényezőváltozó hatását mind a vizsgált tényezőváltozóból, mind az eredményváltozóból kiszűrjük.

$$Y \text{ és } x_1 \text{ között, ha } x_2 \text{ hatását kiszűrjük: } r_{y1 \bullet 2} = \frac{r_{y1} - r_{y2} * r_{12}}{\sqrt{(1 - r_{y2}^2) * (1 - r_{12}^2)}}$$

$$Y \text{ és } x_2 \text{ között, ha } x_1 \text{ hatását kiszűrjük: } r_{y2 \bullet 1} = \frac{r_{y2} - r_{y1} * r_{12}}{\sqrt{(1 - r_{y1}^2) * (1 - r_{12}^2)}}$$

$$x_1 \text{ és } x_2 \text{ között, ha } y \text{ hatását kiszűrjük: } r_{12 \bullet y} = \frac{r_{12} - r_{y1} * r_{y2}}{\sqrt{(1 - r_{y1}^2) * (1 - r_{y2}^2)}}$$

4. Bemutató feladat

$$r_{y1 \bullet 2} = \frac{r_{y1} - r_{y2} * r_{12}}{\sqrt{(1 - r_{y2}^2) * (1 - r_{12}^2)}} = \frac{0,9684 - 0,9409 * 0,9141}{\sqrt{(1 - 0,9409^2) * (1 - 0,9141^2)}} = 0,7888$$

$$r_{y2 \bullet 1} = \frac{r_{y2} - r_{y1} * r_{12}}{\sqrt{(1 - r_{y1}^2) * (1 - r_{12}^2)}} = \frac{0,9409 - 0,9684 * 0,9141}{\sqrt{(1 - 0,9684^2) * (1 - 0,9141^2)}} = 0,5508$$

$$r_{12 \bullet y} = \frac{r_{12} - r_{y1} * r_{y2}}{\sqrt{(1 - r_{y1}^2) * (1 - r_{y2}^2)}} = \frac{0,9141 - 0,9684 * 0,9409}{\sqrt{(1 - 0,9684^2) * (1 - 0,9409^2)}} = 0,0343$$

Látható, hogy lazább a kapcsolat, ha kiszűrjük a másik tényező hatását, a páronkénti kapcsolatot tehát a harmadik változó hatása mindegyik esetben felerősítette.

Többszörös korrelációs együttható

Egy speciális korrelációs együttható, amely az y eredményváltozó és a magyarázóváltozók alapján becslt regressziós értékek kapcsolatának szorosságát méri.

$$R_{y \bullet 1,2,...m} = \frac{\Sigma d_y * \Sigma d_{y'}}{\sqrt{\Sigma d_y^2 * \Sigma d_{y'}^2}}$$

A páronkénti korrelációs együtthatókból is kiszámolható, 3 változós esetben az alábbi:

$$R_{y \bullet 1,2,...m} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2 * r_{y1} * r_{y2} * r_{12}}{1 - r_{12}^2}}$$

A többszörös korrelációs együttható négyzete a többszörös determinációs együttható, amely megmutatja, hogy az eredményváltozó teljes szórásnégyzetéből mekkora a regresszióknak tulajdonítható, tehát a magyarázóváltozókkal magyarázható hányad.

5. Bemutató feladat

$$\begin{aligned} R_{y \bullet 1,2,...m} &= \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2 * r_{y1} * r_{y2} * r_{12}}{1 - r_{12}^2}} = \\ &= \sqrt{\frac{0,9784^2 + 0,9409^2 - 2 * 0,9684 * 0,9409 * 0,9141}{1 - 0,9141^2}} = \\ &= 0,978 \end{aligned}$$

$$R^2 = 0,9568^2 = 95,6\%$$