6. lecke

Korreláció, rangkorreláció

A korreláció- és regresszió analízis végrehajtására van szükség számtalan, a hétköznapi életben is előforduló probléma kapcsán. Néhány példa regressziós modellek alkalmazására:

- Egy adott jármű féktávolságát a gépkocsi sebessége és a gépkocsivezető reakció ideje hogyan befolyásolja?
- Hogyan befolyásolja a dolgozó életkora a táppénzen töltött napok számát?
- Milyen hatással van az adott bank kihelyezési kölcsön költségére az éves átlagos kölcsön nagysága, az évi összes kölcsönkérők száma, az új kölcsönkérvényezők száma és a bank fizetési skála indexe?
- Hogyan függ az adott idő eltelte után beért paradicsom mennyisége az előre jelzett csapadék mennyiségétől, az előre jelzett napfénytartamtól, a paradicsom fajtájától, valamint a kiültetés időpontjától?
- Milyen kapcsolat van a budapesti lakások kínálati ára és a területe, a szobák száma, valamint a terasz nagysága között?
- Egy autópálya forgalmára miképp hat az autópálya használati díja és az autópályát igénybe vevő gépkocsik száma?
- Egy adott áru keresett mennyisége hogyan változik különböző árak mellett? (Keresleti görbe).
- Egy adott áru kínált mennyisége hogyan alakul különböző árak mellett? (Kínálati görbe).
- Adott tőkeállomány és technológia mellett a felhasznált munka különböző mennyiségeihez mekkora megtermelhető maximális termékmennyiség tartozik? (Rövid távú termelési függvény.)
- Hogyan változik a maximálisan előállítható termékmennyiség, ha nemcsak a felhasznált munka, hanem a tőkeállomány és az üzemméret is változik? (Hosszú távú termelési függvény)
- A munkabér különböző értékei hogyan befolyásolják az egyén által kínált munkamennyiséget? (Egyéni munkakínálati görbe).

1. Korrelációszámítás

Korrelációszámítás esetén az elemzésbe vont metrikus változók közötti kapcsolatot vizsgáljuk.

Két metrikus változó (x,y) közötti kapcsolat vizsgálatának első fázisában pontdiagramot készíthetünk az x-y változópár alapján. A pontdiagram alapján megállapíthatjuk a változópár közötti kapcsolat típusát: lineáris, vagy nem lineáris a kapcsolat. Lineáris kapcsolat esetén a pontok egy képzeletbeli egyenes, nem lineáris kapcsolat esetén egy szabályos görbe körül szóródnak. Mivel a gyakorlatban nagyon gyakran élünk a linearitás feltételezésével, így a továbbiakban erre koncentrálunk. A pontoknak a képzeletbeli egyenes körüli szóródásából következtethetünk arra, hogy milyen szoros kapcsolat van a két változó között. Az egyenes meredekségéből pedig következtethetünk a kapcsolat irányára, ami pozitív, vagy negatív lehet. A pozitív irányú kapcsolat azt jelenti, hogy a két változó azonos irányba változik. Mivel a pontdiagram nem egzakt megoldása a korrelációszámításnak, ezért a kapcsolat erősségének jellemzésére mérőszámokat használunk.

A páronkénti korrelációszámítás alapja a kovariancia, amelynek előjele megmutatja a két metrikus változó közötti **kapcsolat irányát.** Kiszámítása:

$$C_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x} * \bar{y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Lineáris kapcsolat esetén az úgynevezett **lineáris korrelációs együtthatót**, míg nem lineáris kapcsolatok esetén az úgynevezett **korrelációs indexet** használjuk. A lineáris korrelációs együttható kiszámítása:

$$r_{xy} = \frac{C_{xy}}{\sigma_y \sigma_x} = \frac{\sum_{i=1}^{n}(d_x d_y)}{\sqrt{\sum_{i=1}^{n}(d_x^2)}\sqrt{\sum_{i=1}^{n}(d_y^2)}}$$

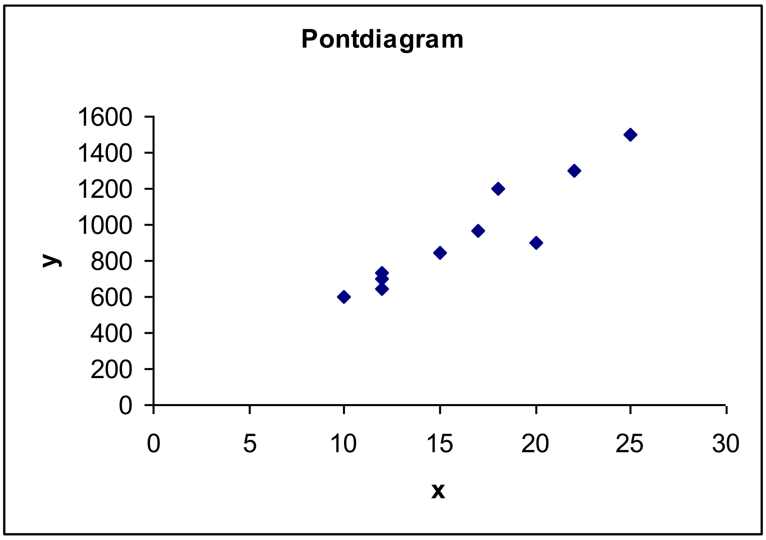
Az *r* lineáris korrelációs együttható értéke [-1;+1] tartományba esik. Előjele megadja a két változó közötti kapcsolat irányát, míg abszolút értéke a kapcsolat erősségét. A nullához közeli érték gyenge, az egyhez közeli érték erős kapcsolatot jelent. Pozitív irányú kapcsolat esetén a két változó ugyanolyan irányba változik. A korrelációs index értéke [0;+1] tartományba esik, és kizárólag a változópár közötti kapcsolat erősségét adja meg.

A lineáris korrelációs együttható alkalmazási feltételei:

- monotonitás,
- linearitás (ha nem teljesül, alá mér a mutató),
- ne legyenek outlierek (kiugró értékek),
- a vizsgált változók normális eloszlásúak legyenek.

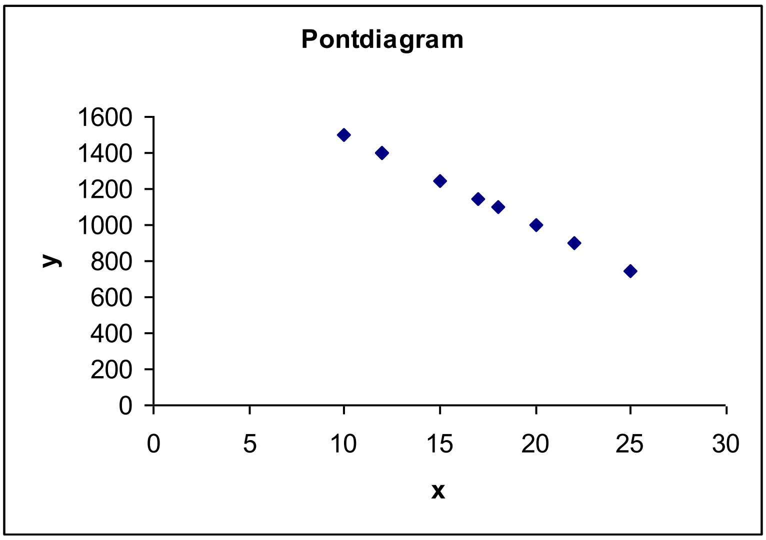
Például, mit állapíthatunk meg az alábbi pontdiagramok alapján?

A)



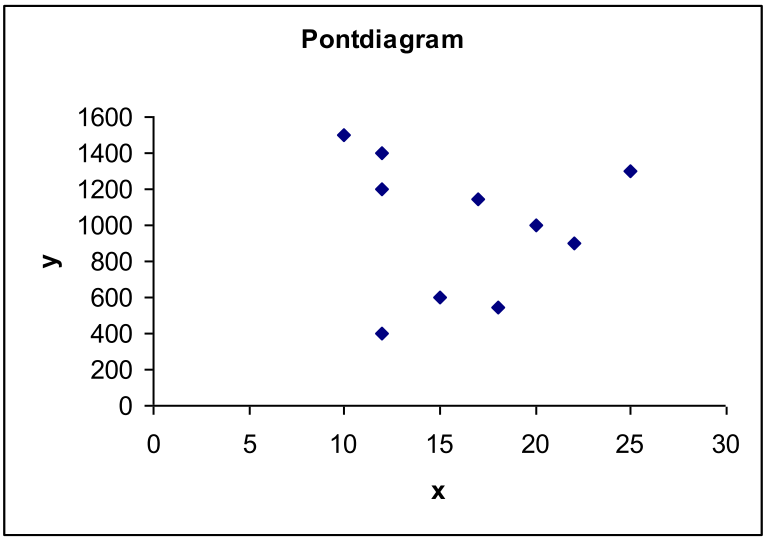
Mivel a pontok nagyon kis mértékben szóródnak egy képzeletbeli, pozitív meredekségű egyenes körül, ezért a két mennyiségi ismérv között pozitív irányú, erős lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke egyhez közeli.

B)



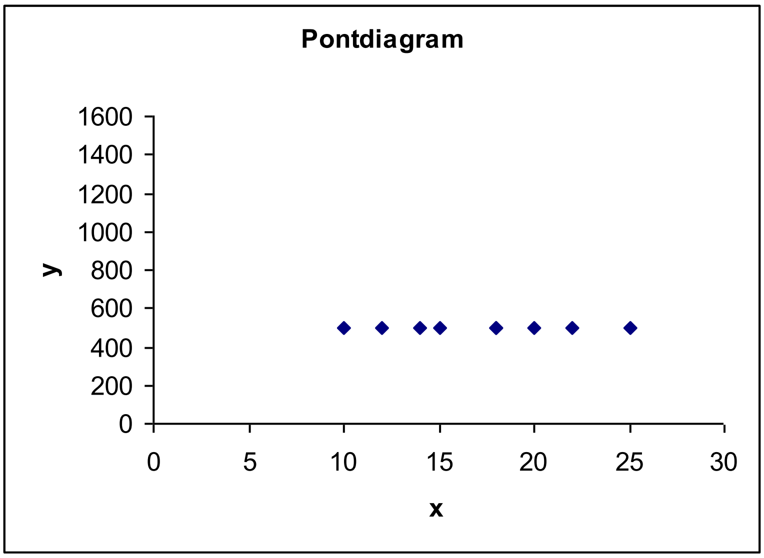
Mivel a pontok kis mértékben szóródnak egy képzeletbeli, negatív meredekségű egyenes körül, ezért a két mennyiségi ismérv között negatív irányú, nagyon erős lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke mínusz egyhez közeli.

C)



Mivel a pontok mindkét dimenzióban nagyon szóródnak, így nem lehetséges egy képzeletbeli egyenes rájuk illesztése, ezért a két mennyiségi ismérv között nagyon gyenge, szinte elhanyagolható lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke nullához közeli.

D)



Mivel a pontok kis mértékben szóródnak egy képzeletbeli, zéró meredekségű egyenes körül, azaz az *x* értékétől függetlenül *y* megközelítőleg konstans, ezért a két mennyiségi ismérv között elhanyagolhatóan gyenge lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke nullához közeli.

Lineáris korreláció esetén érdekes kérdés a kapcsolat szignifikanciájának vizsgálata. A próba nullhipotézise szerint a vizsgált két változó egymástól lineáris független, tehát a korrelációs együttható értéke szignifikánsan nem különbözik nullától. Ehhez az alábbi próbafüggvényt kell alkalmazni:

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

A próbafüggvény mintán felvett értékének kiszámítása után felírhatjuk elfogadási tartományunkat, amely kétoldali próba esetén például: $ET\colon \left(-t_{1-\frac{\alpha}{2}}(v);\ +t_{1-\frac{\alpha}{2}}(v)\right)$, ahol v=n-2. Ha a próbafüggvény mintán felvett értéke beleesik az elfogadási tartományba, akkor elfogadjuk a nullhipotézist, tehát szignifikáns kapcsolat van a két változó között, ha pedig nem esik bele, akkor elvetjük a nullhipotézist, azaz nincs szignifikáns kapcsolat a vizsgált változók között.

Abban az esetben, ha kettőnél több változó közötti kapcsolat vizsgálunk, akkor egyrészt beszélhetünk a változópárok közötti kapcsolatról. Ekkor minden egyes változópárra kiszámíthatjuk a lineáris korrelációs együttható értékét. Ekkor ezeket egy mátrixba rendezve adjuk meg, melyet **korrelációs mátrixnak** nevezünk.

$$\begin{array}{cc} & \begin{array}{cccc} y & x_1 & & x_m \end{array} \\ \mathbf{R} = & \left[\begin{array}{cccc} 1 & r_{yx_1} & \cdots & r_{yx_m} \\ r_{x_1y} & 1 & & r_{x_1x_m} \\ \vdots & & & \\ r_{x_my} & r_{x_mx_1} & & 1 \end{array}\right] \begin{array}{l} y \text{ változó} \\ x_1 \text{ változó} \\ \\ x_m \text{ változó} \end{array} \end{array}$$

Számunkra még a **többszörös korrelációs együttható** lesz fontos, amely a függőváltozónak az együttmozgását méri a magyarázóváltozók együttesével. Ezzel a regresszió témakörében fogunk részletesen foglalkozni.

2. Rangkorreláció

Rangkorreláció nem csak metrikus (skála mérési szintű), hanem ordinális mérési szintű változók között is vizsgálható, amennyiben azok sorrendiségét szeretnénk vizsgálni. Ordinális változók esetén az ismérvváltozatok kategóriáknak is tekinthetők, működhet a kereszttábla-elemzés, de az ismérvek sorrendisége, rangsora is információt hordoz. A rangkorreláció eszközeivel azt vizsgálhatjuk meg, hogy két változó rangsorai között fennáll-e kapcsolat.

A rangkorreláció eljárása első lépésben rangokat társít az egyes ismérvértékekhez, ez történhet csökkenő vagy növekvő sorrendben is, ami fontos, hogy a változók rendezése azonosnak kell, hogy legyen, azaz ha az első változónk ismérvváltozataihoz növekvő sorrendben társítunk rangokat (azaz a legalacsonyabb érték lesz 1), akkor ugyanígy kell eljárnunk a másik vizsgált változó esetén. Ha egy változónak több egyforma értéke is előfordul, akkor a megfelelő sorszámok számtani átlagát rendeljük az azonos értékekhez. Ezeket nevezzük **kapcsolt rang**oknak.

A rangkorreláció kapcsolatmérője a Spearman-féle rangkorrelációs együttható:

$$r_s = 1 - \frac{6\sum_{i=1}^{N}\left(R_{x_i} - R_{y_i}\right)^2}{N(N^2-1)}$$

ahol R_{xi} az első változó rangjait, R_{yi} pedig a második változó rangjait jelöli. A Spearman-féle rangkorrelációs együttható Értéke [-1;+1] közé eshet, előjele és nagysága megadja a kapcsolat irányát és erősségét. A változók (x,y) jelölése a számítás szempontjából irreleváns.

A rangkorreláció szignifikanciájának tesztelését –azaz, hogy szignifikáns kapcsolat van-e a két változó rangsorai között- a hipotézisvizsgálat lépéseit követve tehetjük meg. A vizsgálat nullhipotézise, hogy a két változó rangsora között nincs szignifikáns kapcsolat. A vizsgálat próbafüggvénye:

$$t = \frac{minta - várt}{standard\ hiba} = \frac{r_s - \rho_s}{s_{r_s}} = \frac{r_s - 0}{\sqrt{\frac{1-r_s^2}{n-2}}} = r_s * \sqrt{\frac{n-2}{1-r_s^2}} = \frac{r_s * \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

A próbafüggvény mintán felvett értékének kiszámítása után felírhatjuk elfogadási tartományunkat, amely kétoldali próba esetén például: $ET\colon \left(-t_{1-\frac{\alpha}{2}}(v);\ +t_{1-\frac{\alpha}{2}}(v)\right)$, ahol v=n-2. Ha a próbafüggvény mintán felvett értéke beleesik az elfogadási tartományba, akkor elfogadjuk a nullhipotézist, tehát szignifikáns kapcsolat van a két változó rangsorai között, ha pedig nem esik bele, akkor elvetjük a nullhipotézist, azaz nincs szignifikáns kapcsolat a vizsgált változók rangsorai között.

3. Döntés szoftverrel

3.1. Excel alkalmazása

Az elméleti összefoglalóban szereplő szempontok közül az Excel Adatelemzés bővítménye nem tudja mindegyiket vizsgálni. Ezeknek az alkalmazása előtt érdemes az adatokból pontdiagramot készíteni.

Korrelációanalízis során csak az elemzés alá vont változókat kell megadnunk. Az **Adatelemzés** bővítmény **Korrelációanalízis** menüpontjából érhető el a korrelációelemzés Excelben. A Korrelációanalízis kimenete a korrelációs mátrix. Az *Adatelemzés* bővítmény nem vizsgálja a korrelációs együtthatók szignifikanciáját.

	Változó1	*Változó2*
Változó1	1	
Változó2	–0,980084803	1

A fenti példában a korrelációs mátrixból kiderül, hogy a vizsgált két változó között negatív irányú, erős korrelációs kapcsolat van. A korrelációs mátrix egy szimmetrikus mátrix, a főátlóban levő értékek 1-et vesznek fel, mivel az az adott változó önmagával vett korrelációját mutatja, amely csak és kizárólag 1 lehet, a főátló alatti területen pedig a páronként vett korrelációs együtthatókat láthatjuk, amelyek értéke -1 és 1 közé eshet, előjele a kapcsolat irányát adja meg, abszolútértékét tekintve pedig minél távolabb van a nullától, annál erősebb kapcsolatról beszélhetünk. A fenti példában szereplő -0,98-as érték tehát egy erős, negatív irányú kapcsolatra utal.

3.2. SPSS alkalmazása

Rangkorrelációt és korrelációt is vizsgálhatunk SPSS segítségével. A korrelációs és rangkorrelációs együtthatók kiírása az **Analyze/Correlate/Bivariate** menüpontjában lehetséges. **Korreláció esetén a Pearson, rangkorreláció esetén pedig a Spearman együtthatót kell kiválasztanunk.**

SPSS-ben a két változó együttmozgásáról pontdiagramot a **Graphs/Legacy Dialogs/Scatter/Dot** menüpontban készíthetünk. A korreláció szignifikanciájának vizsgálatához ellenőrizni kell a vizsgálat alkalmazási feltételét, azaz a normális eloszlás teljesülését. A normális eloszlás vizsgálatát az **Analyze/Descriptive Statistics/Explore** menüpontban tudjuk megtenni. Amennyiben teljesül a normális eloszlás, úgy a korrelációelemzés kimenetén található Sig értékeket vizsgálhatjuk. A vizsgálat nullhipotézise ebben az esetben is az, hogy a vizsgált változók között nincs szignifikáns kapcsolat. Ha a Sig. értéke nagyobb, mint 0,05, akkor elfogadjuk a nullhipotézist, tehát nincs szignifikáns kapcsolat a két változó között, ellenkező esetben pedig elvetjük azt, tehát szignifikáns kapcsolat van a két változó között.

JELEN TANANYAG
A SZEGEDI TUDOMÁNYEGYETEMEN KÉSZÜLT
AZ EURÓPAI UNIÓ TÁMOGATÁSÁVAL.
PROJEKT AZONOSÍTÓ: EFOP-3.4.3-16-2016-00014

SZÉCHENYI 2020
Magyarország Kormánya
Európai Unió
Európai Szociális Alap
BEFEKTETÉS A JÖVŐBE