4 TÁRSADALMI JELENSÉGEK TÉRBELI EGYÜTTMOZGÁSA

4.1 Általános szempontok

A területi folyamatok, a tagoltság vizsgálata szinte sohasem szűkül le egy-egy jelenség (mutatószám) térbeli eloszlásának elemzésére (már a fajlagos adatok egyenlőtlenségeinek mérésekor is két jelenséget, jellemzőt kapcsolunk össze), a térbeli együttmozgások elemzésében azonban már kifejezetten a *területi kölcsönhatások,* esetenként az *ok-okozati kapcsolatok* is megjelennek.

A kérdéskörhöz a matematikai-statisztika egyik generális módszercsaládja, *a korreláció- és regressziószámítás* kapcsolódik legszorosabban. E fejezetet erre fűztük fel, területi jelenségek és adatsorok példáit bemutatva. A módszertan általános statisztikai vonatkozásait átfogó statisztikai módszertani munkák részletesen taglalják (*Hunyadi L – Vita L. 2002, Mudruczó Gy. 1981, Moksony F. 1999*).

A *korrelációszámítás önmagában* ritkán szerepel a területi elemzésekben, jelentőségét elsősorban az adja, hogy kapcsolódik az összetettebb módszerekhez (például épp a regressziószámításhoz). Gyakran alkalmazzák a korrelációszámítást azért is, mert számos bonyolultabb matematikai számítás használja az így kapott eredményeket. E módszerek használata során újabban nagyon határozottan *kombinálódik a területi statisztika és a térelemzés* szemlélete, egyre nagyobb hangsúly esik a területi adatok *autokorreláltságára* → **4.3**.

Szignifikancia

Mindkét módszer kapcsán talán a legfontosabb szabály, hogy használatuk, akkor vezet – egyéb feltételek mellett – *megbízható (szignifikáns)* összefüggésekre, ha viszonylag *nagy elemszámú mintából, hosszú adatsorból* számítjuk őket. Az erős szignifikanciájú, megbízható összefüggés – leegyszerűsítve - azt jelenti, hogy a megfigyelési egységek körét véletlenszerűen újabbakkal bővítve, nagy valószínűséggel nem változik az összefüggés iránya és szorossága. A nem-szignifikáns kapcsolat esetében ez már nem áll. Könnyen belátható, hogy mindez szorosan kapcsolódik a megfigyelési egységek számához, hisz például 1000 kísérlet (mérés, adatpár) alapján mért kapcsolat aligha módosul, ha újabb, az 1001. esetet is hozzávesszük, de ugyanez nem áll, ha csak 10 mérésünk van. Kis elemszámok, kevés megfigyelési egység esetében csak a nagy abszolút értékű korrelációs együtthatók szignifikánsak. Formálisan számíthatunk ugyan korrelációs együtthatót a 7 régióra vagy a 20 megyére, de ekkor kevés az alap az összefüggés általánosítására, kiterjesztésére. A hazai területi szintek közül a kistérségek (ma: 168), a városok (ma: 287) vagy az összes település (ma: 3152) adataival számított korrelációk már általában megbízhatóak. (Ne tévesszük tehát össze a statisztika gyakorlatok egyszerű számpéldáit a valóban érdemi következtetéseket lehetővé tevő tudományos elemzéssel!)

A korrelációs kapcsolatok szignifikanciáját azonban nem kizárólagosan az esetszám befolyásolja, hanem a kapcsolat szorossága is. Egy erős (pl. 0,9 körüli) korreláció jóval kisebb valószínűséggel változik egy újabb megfigyelési egységre vonatkozó adatpár hozzávételével, mint egy korrelálatlan (0 körüli együtthatójú) kapcsolat. Végülis a szignifikancia az elemszám és a szorossági szint együttes függvénye, az ún *szignifikancia-tesztekben*[1], táblázatokban ez az összefüggés tükröződik. A szignifikanciát a számítógépes programok többsége (pl. az SPSS is) ellenőrzi.

[1] Ezek koncepciója, leírása az említett matematikai-statisztikai kézikönyvekből megismerhető.

4.2 Korrelációk

A korrelációszámítás valószínűségi változók (jelzőszámok, adatok) közötti kapcsolat *szorosságának* meghatározására szolgáló eljárás. A korrelációszámítás lényege, *döntő lépése* a kapcsolat szorosságának egy mutatószámmal történő tömör jellemzése, azaz a korrelációs együttható értékének kiszámítása. A korrelációnak a regionális elemzésekben három, eltérő jelentést hordozó típusa használatos:

- a közönséges *lineáris korreláció azonos megfigyelési egységekre vonatkozó két adatsor* kapcsolatát méri. Matematikai jelölést használva itt az $r = \mathrm{corr}(x_i\, y_i)$ összefüggés adja a kapcsolat mérőszámát (legismertebb a Pearson-féle korrelációs együttható). A lineáris korrelációs együttható felfogható egyfajta sajátos *egyenlőtlenségi mutatóként* is.

- az *autokorreláció (auto= önmagával vett) egyazon adatsor különböző* (időben eltolt vagy térben szomszédos) *megfigyelési egységekre* vonatkozó értékei közötti kapcsolatot méri.

A kapcsolat idősorokban, azaz az *időbeli* (k-ad rendű) autokorreláció esetében így formalizálható:

$$r = \mathrm{corr}\,(x_i\, x_{i-k}),$$

Itt minden i-edik időponthoz tartozó x_i értékhez ugyanezen x változónak egy k évvel eltolt (k évvel korábbi) adatát rendelve számítjuk ki a korrelációs együtthatót. Ha k = 1, akkor minden év adatát a megelőző év adatával korreláltatjuk. Az eltolás következtében természetesen az adatsor hossza k-val csökken, hisz a kezdőévhez nem rendelhető hozzá korábbi év. A modellben értelemszerűen bármely más időmérték (nap, hét, hónap stb) is szerepelhet.

A területi megfigyelési egységekre vonatkozó adatokból számítható *területi* autokorreláció összefüggése: $r = \mathrm{corr}\,(x_i\, x_{s(i)})$

Ebben a modellben az i-edik megfigyelési (terület) egység x_i adatához a vele szomszédos területegységek (különböző módon számítható és értelmezhető) értékeit, legtöbbször átlagát, $x_{s(i)}$-t rendelve számítjuk ki a korrelációs együtthatót. A területi autokorreláció esetében is beszélhetünk „rendről" (k). Az első rendű autokorreláció az első szomszédok adatait, a másodrendű a második (az első szomszédokkal szomszédos) területegységek adatait rendeli az eredeti értékekhez.

A különböző korrelációk matematikai struktúrája ugyan nagyon hasonló, de tartalmuk lényegesen különbözik. Míg a közönséges lineáris korreláció különböző adatsorok együttmozgását méri, a területi autokorreláció kifejezetten a térbeli elrendeződés mérőszáma egy jelenségen belül[2].

- a *keresztkorreláció két adatsor különböző* (időben eltolt vagy térben szomszédos) *megfigyelési egységekre* vonatkozó értékei közötti kapcsolatot mér, azaz kombinálja a „normális"- és az autokorrelációt (különterű és különidejű események).

Az időbeli keresztkorreláció matematikai sémája (megtartva a fenti jelöléseket):

$$r = \mathrm{corr}\,(x_i\, y_{i-k})$$

A területi keresztkorrelációban a következő összefüggés szerepel:

$$r = \mathrm{corr}\,(x_i\, y_{s(i)})$$

---

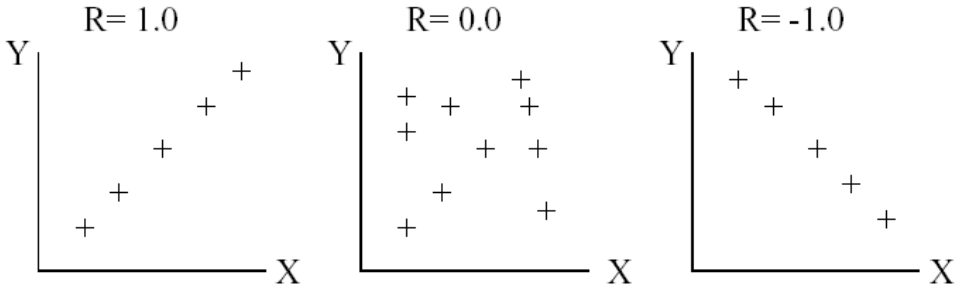
[2] A szakirodalomban újabban vannak olyan kísérletek, amelyek a kétfajta közelítést integrálják. *Lee, S-I. 2001* például egy olyan új indexet javasol, amely két térbeli eloszlást úgy hasonlít össze, hogy az eredményül kapott mérték egyaránt tükrözi az együttmozgást (a lineáris korrelációt) és az alakzatok térbeli jellegét, azaz autokorreláltságát.

Ugyanezen logika alapján beszélhetünk *autoregresszióról és keresztregresszióról* is. Ott a regressziós összefüggésekben szerepelnek a fenti módon definiált adatsorok. Az ún. autoregresszív modellekben a vizsgált változó időbeli lefutásának matematikai-statisztikai magyarázatára ugyanazon változó 1,2,3,...n időegységgel (évvel) eltolt, késleltetett adatainak idősorát használják:

$$y_t = f(y_{t-1}, y_{t-2}, .. y_{t-k}) + e_t$$

4.2.1 Lineáris korrelációs együtthatók

A lineáris korrelációs együtthatók mindegyike -1 és +1 közötti zárt intervallumon vehet fel értékeket, ezeknek a kritikus pontoknak és a korreláltalanságot jelentő 0 értéknek megfelelő összefüggéseket ábrázoltuk a *4.1. ábrán* (a közbülső értékek jelentéséről lásd a *4.1. táblázatot*).



4.1. ábra *A korrelációs együtthatók kulcsértékeinek megfelelő változókapcsolatok*

Pearson-féle lineáris korrelációs együttható (Kiss János Péter)

A korreláció leggyakrabban használt mérőszáma. Két *normális eloszlású; intervallum vagy arány mérési szintű* változó közötti kapcsolat mérésére használható, ha feltételezhető, hogy a két változó között közelítőleg *lineáris összefüggés* van. (Lineáris összefüggés: az egyik változó értékeinek arányában nőnek, vagy csökkennek a másik változó megfelelő értékei). E három feltételnek *egyidejűleg* kell teljesülnie ahhoz, hogy a korrelációszámítást használhassuk. Viszonylag egyszerű számítási módja és tartalmának könnyű felfoghatósága miatt azonban ezek egyikének-másikának hiányában is gyakran alkalmazzák. A módszer legkevésbé a normalitási feltételre érzékeny, úgyhogy az előzetes eloszlásvizsgálatot gyakran mellőzhetjük.

Képlete (a korábbi fejezetek jelölési konvencióit használva):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2 \sum_{i=1}^{n}(y_i - \overline{Y})^2}}$$

A Pearson féle lineáris korrelációs együttható számítási lépései

A mutatószám az x_i és y_i változók értékeinek a számtani átlagaiktól számított eltérésein alapul. (Az eltérést gyakran *d*-vel jelölik, ennek megfelelően egy adatsor 23.-ik helyén álló értéke esetében: $dy_{23} = y_{23} - \overline{y}$.) Ha a változók átlagtól számított eltéréseit páronként összeszorozzuk, majd ezeket összegezzük, akkor megkapjuk a

kapcsolat irányát – hiszen az azonos előjelű eltérések szorzata pozitív, míg a különböző előjelűeké negatív. {Matematikailag kifejezve ezt: $\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})$ vagy egyszerűbben: $\sum_{i=1}^{n}(dx_i \cdot dy_i)$}

E szorzatösszeg nagysága azonban függ a megfigyelések (az összehasonlítandó adatpárok) *n* számától, illetve a két változó mértékegységeitől, valamint szórásától is. Ahhoz, hogy egy általánosítható, további összehasonlításokra alkalmas „dimenziótlan" mérőszámot kapjunk, ki kell szűrnünk ezeknek a tényezőknek a hatását.

A cél érdekében először az eltérések szorzatösszegét elosztjuk a megfigyelések számával (azaz *n*-nel), vagyis a kapott „eltérésszorzatok" számtani átlagát vesszük:

$$C = \frac{\sum_{i=1}^{n}(dx_i \cdot dy_i)}{n}$$

ahol C-t *kovarianciának* nevezzük.

Tulajdonképpen a kovariancia az a mutatószám, amelynek *előjele* megadja a két változó közötti kapcsolat *irányát*, nagysága pedig alapul szolgál a kapcsolat *szorosságának* kifejezéséhez. Bizonyítható, hogy a kovariancia értéke akkor maximális, ha az x_i és y_i változók közötti kapcsolat lineáris és *függvényszerű*. Ilyen esetben a kovariancia az x_i és y_i változók szórásainak szorzatával egyenlő, azaz képlettel kifejezve (és a szórást S-sel jelölve):

$$C = S(x_i)\cdot S(y_i)$$

Ha viszont *nem* függvényszerű a kapcsolat, akkor az eltérő mértékegységekből, illetve az eltérő szórásból adódó különbségek kiszűrését az előbbi összefüggés alapján a

$$\frac{C}{S(x_i)\cdot S(y_i)}$$

képlet segítségével végezhetjük el. És ezzel elértük az eredeti célt, vagyis azt, hogy a korrelációs együttható értéke az alkalmazási feltételeknek megfelelő bármely, és bármilyen egységben mért két változó esetén a {-1; +1} intervallumba essen, azaz: az összefüggések erőssége összevethető legyen. Az eddigi lépéseket egy képletben összegezi a fejezet elején leírt összefüggés.

A *determinációs együttható* a korrelációs együttható négyzete, értéke azt adja meg, hogy a változók egymás szórásának hány százalékát magyarázzák.

Előjelkorreláció, rangkorreláció

Az *előjelkorreláció* gyors, tájékozódó jellegű számításokra használható, valamint bináris (0 és 1, például kicsi-nagy értékekkel azonosítható) változók kapcsolatának mérésére.

$$c = \frac{u - v}{u + v} = \frac{u - v}{n}$$

A képletben *n* a megfigyelési egységek számát, *u* azokat az eseteket jelöli, amikor a két összehasonlított adatsor összetartozó értékei azonos irányban térnek el az adatsor átlagától (vagy mindkettő nagyobb, vagy mindkettő kisebb a saját átlagánál), *v* pedig azon esetek száma, ahol az eltérés ellentétes (az összetartozó adatpárok közül az egyik a saját átlagánál nagyobb, míg a másik kiesebb vagy fordítva). Ha az adatok megegyeznek az átlaggal, akkor mind u, mind v értéke 0,5.

A *Spearman-féle rangkorreláció*

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

A kifejezésben n jelenti a megfigyelési egységek számát, d_i pedig az összetartozó rangszámok különbségeit. A Spearman-féle rangkorrelációval ordinális (sorrendi) adatskálán mért (vagy ilyen adatskálára transzformált) jellemzők együttmozgása mérhető[3].

A rangkorreláció a Pearson-féle lineáris korrelációból vezethető le (*Kovacsics J. 1974,* 198-199. oldal), két nevezetes algebrai összefüggés felhasználásával[4] az átalakításban, ezek:

Az első n természetes szám összege

$$\sum x_i = \frac{n(n+1)}{2}$$

Az első n négyzetszám összege:

$$\sum x_i^2 = \frac{n(n+1)(2n+1)}{6}$$

A rangkorreláció értéke is –1 és 1 közötti lehet, csak 5-nél nagyobb esetszámra alkalmazható.

(Az első publikáció: *C. Spearman 1904* The proof and measurement of association between two things. *American Journal of Psychology*, 15, pp. 72-101.)

A korrelációs együtthatók értékeinek értelmezése

A fenti korrelációs együtthatók hangsúlyozottan a vizsgált jellemzők közötti *lineáris* kapcsolat erősségét mérik. Értékeik (amelyek értelmezésekor gyakorta adnak meg jellemző intervallumokat, egy ilyen osztályozást tartalmaz *a 4.1. táblázat*) nem lineáris kapcsolat esetében megtévesztők lehetnek. Ha két jellemző közötti korreláció alacsony, akkor ez nem azt jelenti, hogy nincs közöttük kapcsolat, csupán azt, hogy nincs lineáris kapcsolat.

r értéke	**A kapcsolat jellege**
r=1	Lineáris függvénykapcsolat, egyenes arányosság van a két jellemző között
0,7 ≤ r < 1	Szoros kapcsolat, egyirányú együttmozgás
0,3 ≤ r < 0,7	Közepes erősségű kapcsolat, egyirányú együttmozgás
0< r ≤ 0,3	Gyenge kapcsolat, egyirányú együttmozgás
0	Nincs lineáris kapcsolat, a két jellemző korrelálatlan
-0,3 ≤ r < 0	Gyenge kapcsolat, ellentétes irányú együttmozgás
-0,7 ≤ r < -0,3	Közepes erősségű kapcsolat, ellentétes irányú együttmozgás
-1 < r ≤ -0,7	Szoros kapcsolat, ellentétes irányú együttmozgás
r = -1	Lineáris függvénykapcsolat, fordított arányosság van a két jellemző között

4.1. táblázat *A korrelációs együtthatók értelmezése*

Az erős korreláció csak két változó együtt járását, összefüggését mutatja, de *nem ad magyarázatot az ok–okozati viszonyokra.* Annak eldöntéséhez, hogy melyik változó mozgatja a másik eloszlását, vagy netán egy harmadik változó hatása bújik meg az összefüggés mögött, logikai magyarázat, illetve egyéb vizsgálatok szükségesek.

[3] Az első publikáció: *C. Spearman 1904* The proof and measurement of association between two things. „*American Journal of Psychology*", 15, pp. 72-101.

[4] A két összefüggés az algebra egyik sajátos módszerével az ún. *teljes indukcióval* bizonyítható.

A „puha” módszerekkel – pl. kérdőíves felméréssel – nyert adatok jelentős része ordinális mérési szintű, ezért ilyenkor csak nagyon indokolt esetben célszerű megkísérelni a korrelációszámítást, és nagyon óvatosan szabad értelmezni az eredményeket. Célszerűbb ilyenkor a már ordinális szinten is értelmezhető rang-, vagy előjel-korreláció használata.

Az ismertetett számítási módból következően az adatsorok nagy és kicsiny értékű elemeinek viselkedése jobban meghatározza a korrelációs együttható értékét, mint az átlag körüliek. Nagyon heterogén – nagyságrendnyivel eltérő értékű – adatokat tartalmazó változók között számított korrelációs együtthatók jelentésének értelmezése ezért bizonytalan. A probléma kiküszöbölése érdekében lehetőség szerint sohasem abszolút, hanem minden esetben fajlagos (pl. lakosságszámhoz viszonyított) adatokból álló változók között kell korrelációt számolnunk. (Ha pl. a magyarországi kistérségek 2000-es adóköteles összjövedelmének, és a diplomás lakosságuk számának összefüggését számoljuk ki, Budapest és a nagyvárosok kiugró értékeinek köszönhetően valószínűtlenül magas, +0.999 értékű korrelációt kapunk. A lakosságszámra vetített, „valódi” korreláció mindössze +0.784, vagyis erős, de azért korántsem determinisztikus. Még jobban megvilágítja a probléma jelentőségét, ha az adóköteles jövedelmet a 2000-ben kiosztott szociális segélyek összegéhez viszonyítjuk. Az abszolút adatokkal számítva +0.553-as együtthatót kapunk, ami arra utalna, hogy a gazdagabb térségekben osztanak több segélyt. A fajlagos adatok esetében viszont az összefüggés iránya megfordul, az együttható -0.519 lesz, vagyis szerencsére megnyugodhatunk: a szociális háló működik, a segélyek a szegényebb térségekbe kerülnek nagyobb arányban.) Ha a fajlagos értékek között is jelentős mértékű a heterogenitás, akkor érdemes a korrelációszámítást a kiugró értékű adat kivételével is elvégezni. (Magyarország esetében ez a helyzet a megyesoros adatok korrelációival: Budapesttel együtt számítva nemritkán jelentősen eltérő a korrelációs együttható értéke, mint csak a 19 megyét alapul véve.)

Térségi szintenként, a területi aggregáció mértékétől függően a korreláció értéke eltérő: általánosságban minél kisebb méretű, és minél több egységből áll egy adott térség, egyre alacsonyabb korrelációs együtthatót valószínűsíthetünk[5]. Az előbbi példáknál maradva: a diplomások aránya és a fajlagos jövedelem közötti kapcsolat a megyék szintjén kiszámítva 0,832-re erősödik, a települések esetében viszont 0,675-re csökken. Hasonló a helyzet a szociális segélyekkel is: a települési szinten csak -0,445, a megyei szinten -0,603 együtthatójú az összefüggés a fajlagos jövedelemmel. A korrelációs együttható értékelésekor ezt a tapasztalati összefüggést is célszerű figyelembe venni. Továbbá, egyebek mellett ezért sem célszerű például kis elemszámú minták esetén – pl. a 7 magyar régióra – korrelációszámítást végezni.

Az eddigiekből is kitűnően *a módszer alkalmazása és a kapott adatok értékelése nagyfokú óvatosságot igényel.* Ezt azért is érdemes külön kiemelni, mert a Pearson-féle korrelációs együttható számítása minden táblázat- és adatbáziskezelő szoftverrel néhány gombnyomással elvégezhető, ami szintén fokozza népszerűségét, s teszi a változók közötti kapcsolatok mérésének talán leggyakoribb módszerévé.

4.2.2 A súlyozás problémája a korrelációszámításban

A területi elemzésekben jellemzően használt adatsorok, indikátorok a területi megfigyelési egységekre *aggregált* értékekből állnak. Ez az adatsorok jó néhány jellegadó értéke (például átlaga) és paramétere (például szórása) esetén felveti a súlyozatlan illetve súlyozott értékek kiszámítását. Ugyanez a probléma megjelenik a korrelációszámításban is, a súlyozás – mivel itt két változó összekapcsolásáról van szó - azonban csak akkor lehetséges, ha a korrelációs együttható esetében az összevetett két adatsorhoz azonos súly rendelhető. Ez számos esetben megoldható: azonos időpontban mért, azonos vetítési alapú fajlagosok (pl. az egy főre jutó GDP és a városi népesség aránya ugyanazon évben megyénként) esetében számítható súlyozott korreláció (a *közös súly* ebben az esetben a megyék népességszáma lesz, s természetesen az átlagok is súlyozottak).

Természetesen a két fajta korreláció értéke különbözni fog, mégpedig annál jobban minél aggregáltabb rendszerrel dolgozunk (ami lényegében azonos azzal, hogy minél egyenlőtlenebbek a súlyok), sok

[5] Itt csak utalunk az ún. *ökológiai tévkövetkeztetés* problémájára (részleteiben lásd *Dusek T. 2004, 6. fejezet*), ami ugyancsak a különböző agggregációs szinteken végzett számítások közötti kapcsolatokat érinti, s óvatosságra int az egy adott aggregációban megállapított összefüggések *átvitelében* más szintekre.

megfigyelési egységből álló, dezaggregált rendszerek esetében a súlyok eloszlása közelít az egyenleteshez, s így a kétfajta korreláció eltérése is csökken.

A közös súly hiánya miatt a legtöbb esetben azonban egyáltalán *nem számítható súlyozott korreláció* (például az egy főre jutó GDP és az útsűrűség közötti korreláció számításakor különböznek a súlyként szóba jövő volumenek, ez esetben a népesség illetve a terület, de ugyanazon mutatószám különböző évre vonatkozó adatsorai esetében is változnak a súlyok).

Mindebből következően a területi adatokból számított korrelációk jellemzően nem az érintett sokaságok egészét jellemző kapcsolatokat mérik, hanem olyan *mintákét*, amelyekben az egyes területegységeket azok *átlagának* megfelelő számértékkel reprezentáljuk.

4.3 Szomszédsági hatások: területi autokorreláció

A korábban bemutatott korrelációs mérőszámok kifejezetten „térmentesek", még akkor is, ha területi adatokból számítjuk őket. Ugyanakkor megfogalmazható az a kérdés, vajon milyen kapcsolatban vannak (korrelálnak-e vagy sem) az egymáshoz közeli területegységek jellemzői. A térbeli egymásrahatások az egymáshoz nagyon közeli, szomszédos helyek között a legvalószínűbbek. Újabban épp e hatásra utalva gyakran idézik *Waldo Tobler* amerikai geográfus-kartográfus professzor[6], „első törvényét". Hogy a „közelhatás" és az együttmozgás milyen jelenségekben van meg, s milyen mértékű, ez az ún. területi autokorrelációs módszerrel számszerűsíthető.

A modern kor gazdasági folyamatainak sajátos ellentmodása (*Quah, D. T. 1996*), hogy miközben a modern gazdaságról elterjedt az a vélemény, hogy mivel „anyagmentes", nincs szüksége térbeli koncentráltságara, agglomerációra, izolált szigetként is létezhet. Ez azonban távolról sincs így. E „térmentes" új időkben különösen felértékelődött a területi autokorreláció problémája (azaz maga a térben való lét szempontja).

A területi autokorrelációs együttható kiszámítása fontos kvantitatív lehetőség arra, hogy *az egymásrahatás (a "közelhatás"),* az együttmozgás szorosságának mértékét különböző jelzőszámok között azonos területi mintában, illetve azonos jelzőszámokét különböző mintákban összehasonlíthassuk. A pozitív területi autokorreláltság regionális tudományi szempontból annak matematikai bizonyítéka, hogy a szomszédság (a közelség) az adott jelenségben egymásrahatással, hasonulással jár. A térbeli autokorreláltság hiánya a társadalmi térben inkább kivétel, mint teljesülő feltétel.

Az autokorreláció fogalma először az idősorok matematikai-statisztikai elemzéséhez kapcsolódva jelent meg. Bevezetéséhez az a felismerés vezetett, hogy számos gazdasági és társadalmi folyamat (pl. a gazdasági növekedés) lefutására, jövőbeni alakulására nemcsak "külső" tényezők hatnak, hanem magának a jelenségnek a vizsgált időpontot megelőző állapotai is. (Ilyen „belső" jellemző a gazdasági növekedés ciklikussága.) Könnyen belátható, hogy az elsőfokú időbeli autokorrelációs együttható értéke lineáris növekedés esetében 1. Éves periódusban ingadozó, oszcilláló, alternáló (pl. 0 és 1 értékeket váltakozva felvevő) idősorokban -1, véletlenszerűen hullámzó idősorokban pedig 0. Autoregresszív modellek természetesen használhatók *területi idősorok* regresszióelemzésére is. Az autokorreláció fogalma és használata a regionális tudományokban azonban nem ezt jelenti (hisz a jelzett esetben továbbra is idősorokról van szó, csak területi megfigyelési egységekben).

A sorbarendezettség szempontját nemcsak időben, hanem más dimenziók mentén is érvényesíthetjük. Ilyen lehet például a *nagyság szerinti* rendezés. E logika szerint vizsgálható például a települések bármely

[6] *Tobler, W. A. 1970* Computer Model Simulating Urban Growth in the Detroit Region. *„Economic Geography",* 2., pp. 234-240. c. tanulmányában a „földrajz első törvényét" így fogalmazta meg: *„everything is related to everything else, but closer things are more closely related"* (nagyjából így: „minden mindennel összefügg, de a közelebbi dolgok erősebben hatnak egymásra".

jellemzőjének autokorreláltsága a lélekszám mentén (az ilyen vizsgálatok vélhetően nagyon erős pozitív autokorreláltságot mutatnak ki).

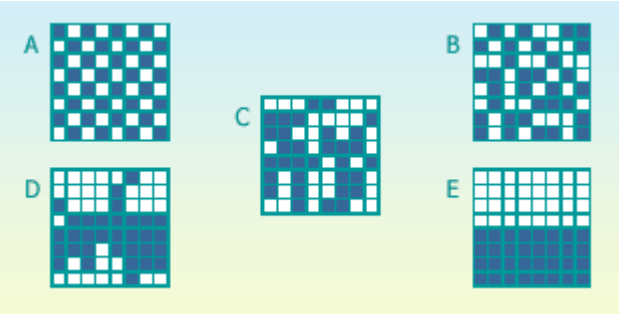
A területi autokorreláció illetve autoregresszió olyan területi modell, amely egy adott társadalmi jelenségnek egy adott helyen, illetve a hellyel szomszédos helyeken mért értékei közötti kapcsolatot, összefüggést mutatja ki. A területi autoregresszív modell – az idősorok analógiájára és a szomszédság fogalmát felhasználva – az adott jelenség térbeli eloszlását, ugyanazon jelenségnek a hely közeli vagy távolabbi környezetében felvett értékeivel írja le, magyarázza → **4.5.1**.

A területi autokorreláció fogalmához az idősor autokorrelációjában megjelenő időkategóriák térbeli megfelelőjének megtalálásán keresztül vezet az út. Ez az időkategória az előtt (illetve után) fogalma, amely az idő linearitása következtében jól meghatározott. A neki megfelelő térkategória a *szomszédság.*

A területi elemzések egyik alapkérdése arra vonatkozik, hogy a vizsgált jelenség területi eloszlásában felfedezhető-e valamilyen szabályszerűség, vagy pedig véletlenszerűnek mondható-e az adatok területi eloszlása. Szabályszerű elrendeződés esetén az egymással szomszédos területegységek adatai egymáshoz hasonlóak lesznek, a nagy érték közelében nagy értékeket találunk (pozitív autokorreláció), vagy épp ellenkezőleg, a szomszédos területek különböznek egymástól, a nagy értékű területek mellett kicsik és a kicsik mellett nagyok helyezkednek el (negatív autokorreláció). A vizsgált térrész egészére jellemző szabályszerűségből az egyedi területegységek szomszédainak értékeire is következtethetünk, illetve a szomszédok ismeretében az adott területegységre vonatkozóan vonhatunk le valószínűségi jellegű állításokat. Autokorrelálatlanság esetén az egyes értékek véletlenszerűen szóródnak a térben, a területi különbségek nem rajzolnak ki szabályos térbeli mintázatot.

A *4.2. ábra* bináris (fekete-fehér) sakktáblamodellje az alábbi öt esetet érzékelteti:

A	extrém negatív autokorreláltság
B	diszperz eloszlás
C	térbeli függetlenség
D	térbeli koncentrálódás (klaszterződés)
E	extrém pozitív területi autokorreláltság



4.2. ábra *A területi autokorreláltság alapesetei egy sakktábla-modellben*

A területi autokorreláltság nemcsak a térkapcsolatok tükrözi, de *mérési gondokat* is okoz a statisztikai vizsgálatokban. A területi autokorreláltság (pl. az egymás melletti területegységek visszatérően hasonló ismérvértékei) olyan torzulásokat okozhat a statisztikai elemzésben, mint amikor egy „ideális" mintavételes eljárásban egy mintául vett adatot *többszörösen szerepeltetnénk* az elemzésben, hatása módosítja, mind az átlagot, mind a szórást.

A kérdésről térképi ábrázolás révén vizuális úton kaphatunk benyomásokat, a területi autokorrelációt mérő különböző indexek segítségével pedig számszerű információ formájában kapunk választ rá.

A területi autokorreláció mérésére többfajta index és eljárás létezik, amelyek egyes lépéseikben (pl. a szomszédság értelmezésében vagy az ezt azonosító szomszédsági mátrixban) különböznek, de a cél mindenütt azonos, a térbeli együttmozgás számszerű mérése. A következőkben a legelterjedtebb területi eutokorrelációs mutatókat ismertetjük.

A Moran-féle I és a Geary féle c (Dusek Tamás)

A területi autokorreláció mérésére szolgáló első, *Moran* által 1950-ben javasolt és azóta leggyakrabban használt mérőszám képlete a következő[7]:

$$I = (N/\Sigma D_{ij})*\Sigma\Sigma(x_i-x)*(x_j-x)*D_{ij}/ \Sigma(x_i-x)^2$$

ahol $(x_i-x)*(x_j-x)$ a területegységek tartozó értékek és az átlagok különbségének a szorzata, D_{ij} a szomszédsági kapcsolatokat leíró mátrix, N a területegységek száma.

A mutató az alábbi tartományokban a következő módon értelmezendő (*Cliff – Ord, 1973, 1981*):
$I>-1/N-1$, pozitív térbeli autokorreláció,
$I=-1/N-1$, nincs térbeli autokorreláció,
$I<-1/N-1$, negatív térbeli autokorreláció.

Gyakorlatilag a nullához közeli értékek az adatok véletlenszerű térbeli eloszlását jelzik, az adatok eloszlása nem terület- és szomszédságfüggő. Ekkor az egyes területegységek szomszédságában ugyanolyan valószínűséggel találunk nagyobb és kisebb értékű területeket is. A szélsőértékek nagysága nem adható meg olyan egyértelműen, mint a korrelációs együtthatónál, mert nagyságuk függ a D_{ij} mátrixban rögzített területi konfigurációtól is. A maximális 1-es értéket végtelen vagy folytonos tér esetén érhetné el, illetve ha a vizsgált terület két, belsőleg homogén, de egymással szomszédsági kapcsolatban nem álló területegységre oszlik. A minimuma szintén a végtelenben közelít a -1-hez.

A regionális elemzésben manapság sokat tesztelt európai regionális fejlettségi tagoltságot vizsgálva 1980-1995 távlatában, az EU régióinak egy lakosra jutó GDP-je tükrében *Le Gallo, J. (2002)*, egy stabilan magas 0,75 és 0,8 közötti pozitív (Moran) területi autokorrelációt mért. Igazolva ezzel a térbeli folytonos átmenet, a fejlettségi zonalitás markáns jelenlétét a térség gazdasági tagoltságában. Az adóköteles jövedelemek hazai területi autokorreláltságát vizsgálja *Dusek T. 2004.*

A *Geary* által 1954-ben javasolt, a Moran-féle I-nél ritkábban használt folytonossági mutató képlete, az előbbi jelölésekkel:

$$c = (N-1/2\Sigma D_{ij})* \Sigma\Sigma D_{ij}(x_i-x_j)^2/\Sigma(x_i-x)^2.$$

A mutató a következő értékeket veheti fel (*Cliff – Ord, 1981*):

$c<1$, pozitív térbeli autokorreláció,
$c=1$, nincs térbeli autokorreláció,
$c>1$, negatív térbeli autokorreláció.

A maximális és minimális értékek nagysága ebben az esetben is a szomszédsági mátrixtól függ.

Mindkét mutató használható ordinális, intervallum és arányskálán rendelkezésre álló adatok elemzésére, a nominális adatokéra pedig alternatív ismérvekké történő átalakításuk után. A mutatók konkrét értékét és eloszlásfüggvényét is két tényező határozza meg:
1. A vizsgált jellemző területi eloszlása.

[7] Az eredeti forrás: *Moran P. A. P. 1950* Notes on continuous stochastic phenomena, „*Biometrica*”, 37., pp. 17-23

2. A szomszédsági kapcsolatok megállapítása és a szomszédsági mátrix súlyai, ami összefüggésben áll a vizsgált terület alakjával és nagyságával.

Ezek közül az első tényező a konkrét érték szempontjából, a második tényező az eloszlásfüggvény alakja miatt fontosabb. Az eloszlásfüggvény vizsgálatát az eredmények megfelelő értékelése követeli meg.

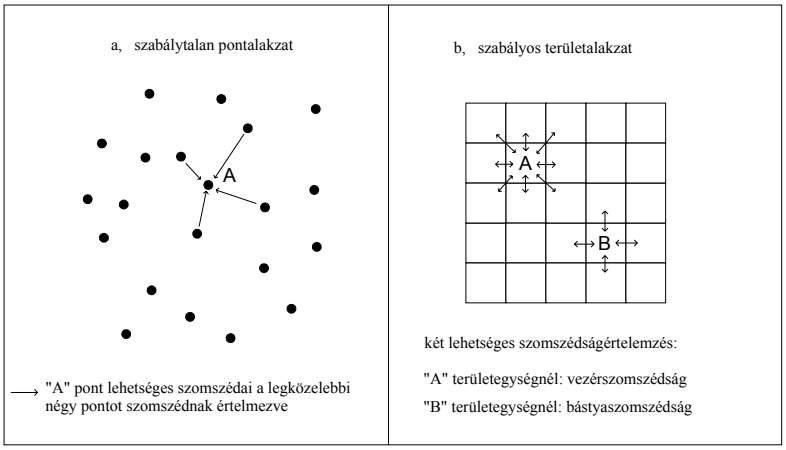
A szomszédsági mátrix

A területi autokorreláció számításának előfeltétele a szomszédsági kapcsolatok megállapítása és a szomszédsági mátrix összeállítása. A szomszédsági mátrix N sorból és N oszlopból áll, i-edik sorának j-edik elemének értéke az i-edik és j-edik területegység szomszédságának hiányában 0, szomszédságuk esetén 0-tól különböző (a *4.2. táblázatban* szereplő szomszédsági mátrixban a szomszédos megyékhez rendelt nem-0 értéket az határozza meg, hogy egy megyének hány szomszédja van. Ha N, akkor egy-egy szomszédhoz a mátrixban 1/N érték – itt két tizedesre kerekítve – tartozik, s így a sorokban szereplő értékek összege 1. Az ilyen mátrixot sor-normalizáltnak hívják.). A megállapodás szerint a területegységek saját maguknak nem szomszédjai, vagyis a mátrix diagonális elemei nullák.

	Bp	Bar	Bacs	Bek	Baz	Cso	Fej	Gyor	Haj	Hev	Kom	Nogr	Pest	Som	Szab	Szo	Tol	Vas	Ves	Zal
Bp	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Bar	0	0	0,33	0	0	0	0	0	0	0	0	0	0	0,33	0	0	0,33	0	0	0
Bacs	0	0,17	0	0	0	0,17	0,17	0	0	0	0	0	0,17	0	0	0,17	0,17	0	0	0
Bek	0	0	0	0	0	0,33	0	0	0,33	0	0	0	0	0	0	0,33	0	0	0	0
Baz	0	0	0	0	0	0	0	0	0,20	0,20	0	0,20	0	0	0,20	0,20	0	0	0	0
Cso	0	0	0,33	0,33	0	0	0	0	0	0	0	0	0	0	0	0,33	0	0	0	0
Fej	0	0	0,17	0	0	0	0	0	0	0	0,17	0	0,17	0,17	0	0	0,17	0	0,17	0
Gyor	0	0	0	0	0	0	0	0	0	0	0,33	0	0	0	0	0	0	0,33	0,33	0
Haj	0	0	0	0,25	0,25	0	0	0	0	0	0	0	0	0	0,25	0,25	0	0	0	0
Hev	0	0	0	0	0,25	0	0	0	0	0	0	0,25	0,25	0	0	0,25	0	0	0	0
Kom	0	0	0	0	0	0	0,25	0,25	0	0	0	0	0,25	0	0	0	0	0	0,25	0
Nog	0	0	0	0	0,33	0	0	0	0	0,33	0	0	0,33	0	0	0	0	0	0	0
Pest	0,14	0	0,14	0	0	0	0,14	0	0	0,14	0,14	0,14	0	0	0	0,14	0	0	0	0
Som	0	0,20	0	0	0	0	0,20	0	0	0	0	0	0	0	0	0	0,20	0	0,20	0,20
Szab	0	0	0	0	0,5	0	0	0	0,5	0	0	0	0	0	0	0	0	0	0	0
Szol	0	0	0,14	0,14	0,14	0,14	0	0	0,14	0,14	0	0	0,14	0	0	0	0	0	0	0
Toln	0	0,25	0,25	0	0	0	0,25	0	0	0	0	0	0	0,25	0	0	0	0	0	0
Vas	0	0	0	0	0	0	0	0,33	0	0	0	0	0	0	0	0	0	0	0,33	0,33
Vesz	0	0	0	0	0	0	0,17	0,17	0	0	0,17	0	0	0,17	0	0	0	0,17	0	0,17
Zala	0	0	0	0	0	0	0	0	0	0	0	0	0	0,33	0	0	0	0,33	0,33	0

4.2. táblázat *A magyar megyék közötti szomszédsági relációk sor-normalizált mátrixa*

Pontalakzatok és területalakzatok esetében sem kerülhető el a szomszédság megállapításánál az önkényes elem, akár szabályos, akár szabálytalan alakzatokról legyen szó. (*4.3. ábra*). Pontalakzatoknál – amelyeknek a gazdasági-társadalmi elemzésben például a települések, lakóházak, kereskedelmi egységek felelhetnek meg – a távolság függvényében lehet kijelölni a szomszédsági relációkat. Például minden pontnál a legközelebb fekvő első kettő-hat pontot lehet szomszédnak tekinteni, vagy bizonyos távolságon belül lévő pontokat szomszédnak venni, vagy a két módszer kombinációjával.



4.3. ábra *Pontalakzat és területalakzat; példák a szomszédságértelmezésekre (Dusek T. 2004, a 25. ábra részlete)*

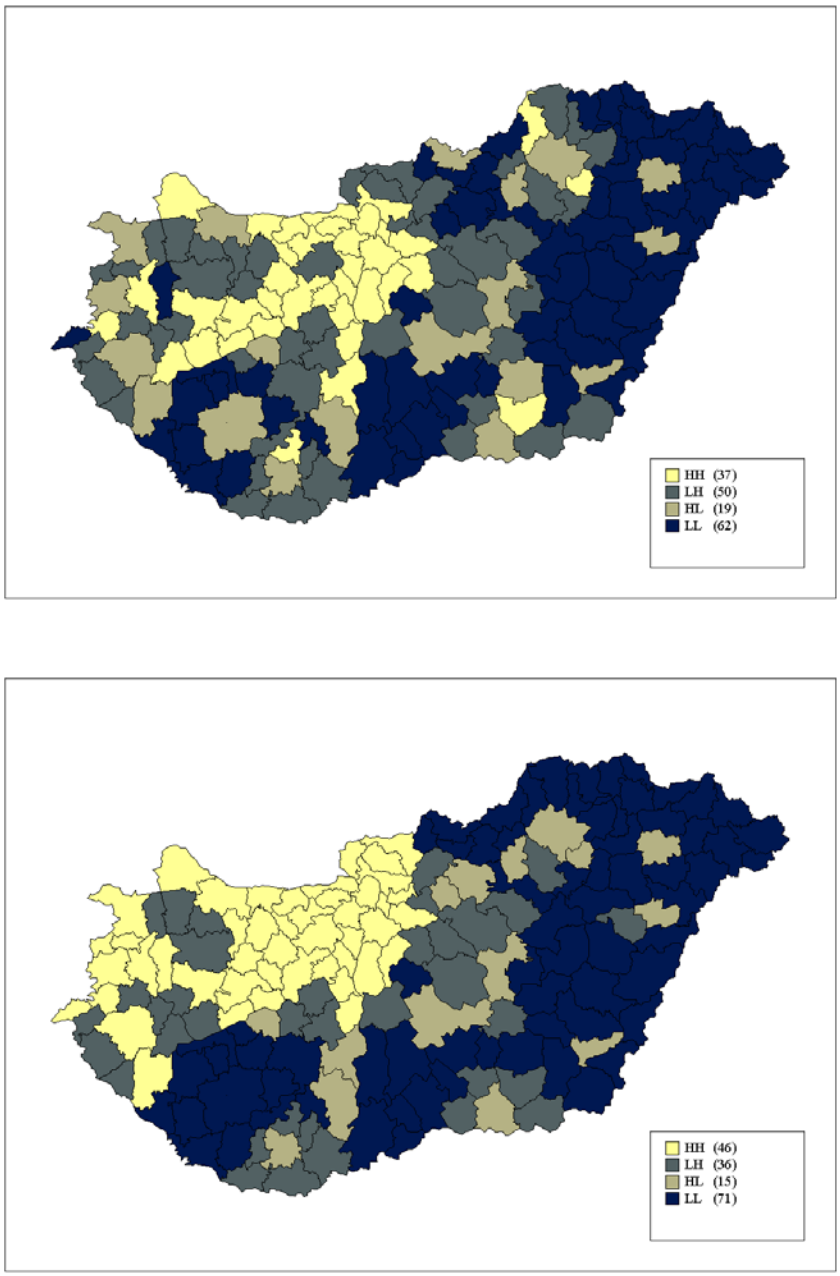
Egy svéd példában (*Lundberg, J. 2002*) a szomszédság többfajta értelmezésével találkozhatunk: A svéd megyék jövedelmét 4 *legközelebbi szomszéd sávot* felállítva vizsgálták (ahol „szomszédnak" 1, 2, 5, 10 legközelebbi megyét tekintettek), de használták osztályozási szempontként *a megyeközpontok közötti utazási időt* (30. 45, 60, 75 perces körzeteket lehatárolva), szerepelt a vizsgálatban az egyszerű *közös határral* való összekapcsoltság is. A szomszédság meghatározásának ugyancsak különböző alternatíváit használja *Aten, B. – Heston, A. (2003),* akik világméretű vizsgálatukban közel 900 területegységet - országot, régiót – elemeztek a jövedelmi autokorreláltság globális illetve lokális jellemzői szerint. Előbbiben a teljes rendszerre számítottak Moran-féle autokorrelációkat, utóbbiban az egyes területegységeket vetették össze „szomszédaikkal". Különösen élesen elüt például fejlettsége tekintetében környezetétől Európában Helsinki, Ázsiában Szingapur városrégiója, az országok közül Izrael és Jordánia, valamint Haiti és a Bahamák esetében párosul a földrajzi közelség nagy jövedelmi lépcsővel.

Területalakzatoknál – amelyeknek megfelelhetnek például országok, megyék, de a települések is – maga a szomszédság léte bizonyos szempontból objektívabb módon dönthető el, amennyiben többnyire egyszerűen a közös határvonallal rendelkező területeket tekintjük szomszédoknak. Az ilyen módon megállapított szomszédság szimmetrikus lesz, „A" területegység szomszédja lesz „B" területegységnek és fordítva. Ezen túlmenően azonban a szomszédság mértékét számos módon lehet súlyozni, például a közös határvonal hossza szerint, a régiókat összekötő hálózatok száma és minősége szerint, ténylegesen megfigyelt áramlások alapján és ezek bármilyen kombinációjával. A módszert tovább lehet bonyolítani a szomszédok szomszédainak, azok szomszédainak stb. számbavételével (másod- , harmad-, n-ed rendű autokorreláció).

A súlyozás az elmondott sokféle szempont érvényesítése miatt nem mindig egyszerű feladat. Minden igényt kielégítő mátrixot azért sem lehet összeállítani, mert a térkapcsolatokra vonatkozó összes információ nem áll a rendelkezésre, mindig csak a statikus állapotokat lehet felmérni. Ez mégsem jelent tényleges problémát, sőt, a többféle súlyozással előállított eredmények összehasonlítása lehetőségeket rejt magában. Ha többféle mátrixszal elvégezzük ugyanazon adatokra vonatkozóan a számításokat, és jelentősen eltérő eredményeket kapunk, akkor azzal a feltételezéssel élhetünk, hogy az alkalmazott súlyozásban megjelenő tényező (távolság, népességszám, határvonal hossza stb.) tényleges befolyást gyakorol a vizsgált jelenség területi kapcsolatainak intenzitására.

A területi autokorreláció bemutatott mérőszámai a teljes adatrendszer térszerkezetéről nagyon összevont információt adnak. Készíthetők azonban olyan (jövedelmi) térképek *(4.4. ábra)*, amelyekben minden területegységet magának illetve a szomszédainak a jövedelszintje szerint típusba sorolunk. Kétszer-kettes osztályozásnál például a magas (átlag feletti) illetve az alacsony (átlag alatti) jövedelem kombinálható a *szomszédok* átlagos értékeinek magas illetve alacsony jövedelmi szintjével. Az ábrán HH jelöli *a vidéki átlagnál* magasabb jövedelmű és szomszédú kistérségeket, LL pedig az átlag alatti jövedelemszint lokális egymásmellettiségét (ez a legszámosabb térségtípus). A HL és LH kategóriák esetében ellentétes a körzet és szomszédeinak jövedelmi (fejlettségi) pozíciója (H = átlag feletti, L= átlag alatti jövedelemeszint). A két térkép határozott regionalizálódást igazol. Jól látszik az,

hogy mára a Dunántúl északnyugati zónája szinte egységes magas jövedelmű térséggé vált – a kilencvenes évek elején még inkább csak a fővárosközeli terekre, s az onnan kiinduló tengelyekre volt jellemző a relatív stabilitás. Érzékelhető a Sajó-völgy vagy a Dél-Dunántúl markáns visszaesésese, s az északi és keleti jövedelemi perifériák széles zónája. Az ország közepén átmeneti lokális viszonyokat tapasztalunk, a problematikus zónákban mindenütt kiugranak a nagyobb városok, megyeszékhelyek térségei[8].



4.4. ábra *Az egy lakosra jutó adóköteles jövedelem lokális hasonlósága 1990-ben (fenn) és 2003-ban (lenn)*

F: *Lőcsei H.* számításai és szerkesztése

[8] Hasonló logikájú térképek a *jövedelem-dinamikát* vizsgálva is készíthetők. Ezzel tesztelhető például az, hogy a gyors vagy lassú *növekedés* is jól körülhatárolható térbeli „klasztereket" rajzol-e ki vagy sem. Az autokorreláltság tetszőleges területi alrendszerekben, például az ország nagyrégióin belül külön-külön is vizsgálható (ezt lásd *Dusek T. 2004. 10.8. fejezet*).

Local Moran I

E térképek tovább finomíthatók a területi autokorreláltság *lokális* mérőszámának felhasználásával. Ilyen index például az, amelyben minden területegység kiválasztott indikátorának (pl. egy lakosra jutó jövedelem) standardizált értékét → **2.5.2** összeszorozzák a szomszédos területegységek együttes átlagos standardizált értékével (3 szomszéd esetén minden szomszéd standardizált jövedelemértéke 1/3-os súlyt kap). A kiválasztott területegység és szomszédai így meghatározott adatának szorzatát nevezi a szakirodalom (*Luc Anselin* amerikai térstatisztikus nyomán) *Local Moran I*-nek. Az index pozitív értéke *hasonló* jellemzők (itt: jellemzően átlag feletti vagy átlag alatti jövedelemszint) *helyi koncentrációját* jelzi, a 0-közeli érték a jövedelemszintek véletlenszerű egymásmellettiségét jelzi, míg a negatív index arra utal, hogy az adott területegység és környéke a vizsgált indikátor (azaz e példában a jövedelemszint) szempontjából erősen különbözik. A standardizálás azzzal az előnnyel jár, hogy a legkülönbözőbb dimenziójú jelzőszámok lokális autokorreláltsági sémája, s az egyes területegységek sajátosságai is összevethetővé válnak, hátránya a mutatónak, hogy nem korlátos. Az I indexek térképezhetők, s ekkor már – eltérően a *4.4. térképtől* – négynél több kategória is érzékeltetheti a lokális együttmozgást. A lokális és globális autokorrrelációs mérőszámok között direkt matematikai kapcsolat van: egy adott térrészben (pl. ország) minden egyes területegységre (pl. N kistérségére) kiszámított lokális autokorrelációs indexek összege egyenlő az előzőekben bemutatott globális Moran index (I) N-szeresével.

Tulajdonképp leginkább a lokális területi autokorreláltság gondolati vonalához kapcsolható *Tiner Tibor területi feszültség*-vizsgálata is (*Tiner T. 2003*), amelyben a szerző a személygépkocsi-ellátottság megyei differenciáltságának példáján azt firtatja, hogy a szomszédos területegységek ellátottsági szintjében mekkorák a különbségek (a „feszültség" nála egy ellátottsági lépcsőt jelent két szomszédos megye között). Az említett vizsgálatban ugyan kérdéses, hogy a statisztikai alapon meghatározott feszültség generál-e kölcsönhatást, más kutatások (például a fejlettségi értelemben nagyon különböző területeket elválasztó határzónák vizsgálata) azonban kétségkívül arra utalnak, hogy a fejlettségi lépcsők – amennyiben erre egyébként van mód – áramlásokat indítanak el (lásd hazánk keleti határait).

4.4 Regressziószámítás a területi elemzésben (Németh Nándor)

A különböző korrelációs együtthatók a változók közötti együttmozgások nagyon tömör mérőszámai. Jelzik ugyan az együttmozgásokat vagy épp a teljesen véletlenszerű értékkombinációkat a megfigyelési egységekben, de a kapcsolat további részleteinek feltárásához újabb eszközökre van szükség. Ezeknek talán legfontosabbika a *regresszió-számítás*, amely a változókapcsolatokat *valószínűségi (sztochasztikus) függvénykapcsolatként* értelmezi és írja le. A valószínű jelleg azt jelenti, hogy a két (vagy több) adatsort a kiválasztott függvény a regressziószámításban csak bizonyos „hibával" illeszti össze. A regressziós kapcsolat a regressziós függvény által magyarázott rész és a hibatag összege[9].

[9] Kötetünk részleteiben nem tárgyalja az idősorok regresszió-elemzését az ún. *trendszámítást*. Ennek kapcsán jegyezzük meg, hogy a legújabb elemző szakirodalomban gyakran találkozhatunk ún. *panel-elemzésekkel is*, amelyekben *területi idősorok együttes regresszióelemzését* végzik. A *The Econometrics of Panel Data: A Handbook of the Theory With Applications (Advanced Studies in Theoretical and Applied Econometrics, Vol 33)* ed. Laszlo Matyas, Patrick Sevestre, Kluwer Print on Demand, 1996. c. kézikönyv áttekinti a legfontosabb panel módszereket.

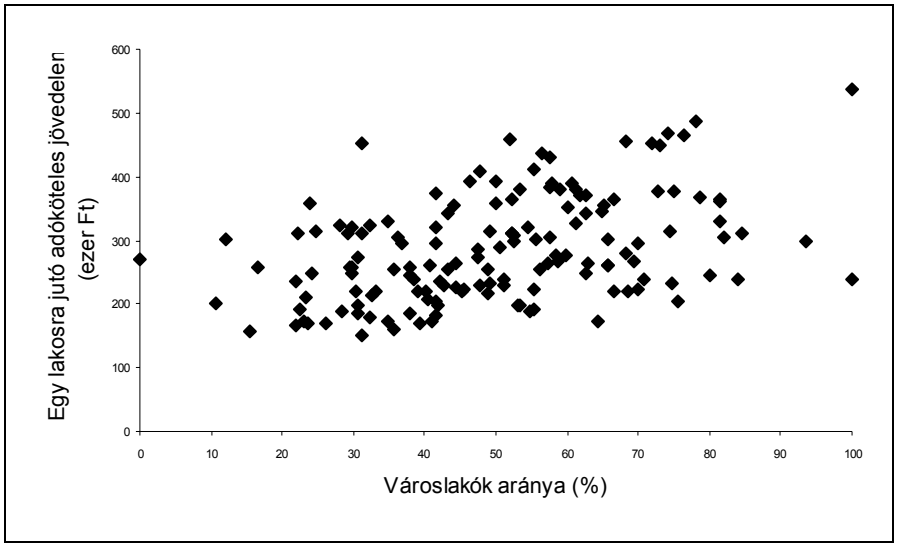
4.4.1 Függő és független változók

Az eljárás bemutatását a két adatsort összekapcsoló, *kétváltozós* regressziószámítással kezdjük.

A kétváltozós regresszió lényege: olyan y = f(x) függvény megtalálása, amely minden megfigyelt x_i-re a megfelelő y_i-hez lehetőleg legközelebbi értéket adja (azaz x értékei alapján kis hibával becsülhetők y értékei). A két változó közötti összefüggést legjobban közelítő függvények matematikailag egyértelműen meghatározhatók, *differenciálszámítási* alapokon kiválasztható az a függvény, amelynek esetében a legkisebb a regressziós becslés hibája.

Bizonyítható, hogy n megfigyelési egység (adatpár) esetében a kapcsolatot egy n-1-ed fokú polinom teljes pontossággal leírja (ennek képe egy olyan vonal, amely áthalad minden x_i y_i ponton). Ebből a függvényből azonban nem olvasható ki a két jellemző közötti kapcsolat tendenciája, interpretációja lényegében lehetetlen. A „teljesen pontos" kapcsolatleírás helyett ezért olyan viszonylag egyszerűbb függvényeket keresnek, amelyek jól interpretálhatók. A leggyakoribb ilyen függvénytípus a lineáris függvény (amelynek képe egy egyenes).

A regresszióelemzést érdemes a két változó értékeinek koordinátarendszerben való ábrázolásával, az ún. *szórásdiagram* felrajzolásával kezdeni (a *4.5. ábra* a 2000. évi városlakók kistérségi aránya – továbbiakban VARPOP – és a kistérségi jövedelemszint – a továbbiakban JOVPOP – közötti összefüggést ábrázolja). Ez nem más, mint egy derékszögű koordináta-rendszer, melynek egyik tengelyén az egyik, a másikon a másik változó értékeit vettük fel. A diagram minden pontja egy-egy kistérségnek felel meg. Azok a kistérségek, ahol viszonylag kicsi a városlakók aránya és emellett még a lakossági jövedelmek szintje is alacsony, a diagram bal alsó sarkának közelében helyezkednek el, míg azok, ahol mindkét vizsgált változónk értéke relatíve magas, a jobb felső sarokban találhatók.



4.5. ábra *A szórásdiagram*

A kapcsolat jellegének feltárása során először azt a nem is könnyű dilemmát kell eldöntenünk, hogy melyik változót kívánjuk magyarázni a másikkal? Estünkben logikusnak az látszik, hogy a városlakók arányának növekedése hozza magával a magasabb kistérségi jövedelemszintet. Ezért itt a városlakók aránya lesz a *magyarázó* (vagy másként *független*), a lakossági jövedelmek szintje pedig az *eredmény* (vagy másként *függő*) változó. A képletekben *y* jelzi a függő, *x* pedig a független változót.

Természetesen azt, hogy egy adott jelenség éppen függő vagy független változóként szerepel-e modelljeinkben, mi magunk döntjük el adott kutatási hipotézisünk és az adott elméleti keretek alapján. *Egyetlen egy társadalmi-gazdasági jelenségre sem mondhatjuk azt, hogy minden körülmények között csak*

valami okaként, vagy éppen okozataként léphet be kutatásainkba. Példánknál maradva: a városlakók népességen belüli aránynövekedésének is vannak okai, tehát más keretek között e mutatónk függő változóként fog szerepelni, míg a jövedelemnövekedés is hatással van egy sor egyéb jelenségre, melyek változásának tehát – részben vagy egészben – okozója lesz.

Mit tudunk megállapítani *4.5. ábránk* segítségével? Sajnos nem sok mindent. Nagy vonalakban látszik ugyan, hogy van valamiféle tendencia két változónk együttes változásában, de ábránk túl kusza, pontjaink elhelyezkedése túl zaklatott ahhoz, hogy határozottan tudjuk állítani: a városlakók arányának növekedése magával vonja a lakossági jövedelmek növekedését is. E kuszaság annak tulajdonítható, hogy a lakossági jövedelmek szintje messze nemcsak a városi lét, a városi körülmények között működő, ily módon sajátos karakterű gazdaság súlyának függvénye egy-egy térségben, hanem számos más tényező is befolyásolja. Így a helyi lakosság kulturális identitása, hagyományai, vagy egyszerűen a települések nagysága, a földrajzi pozíció, a térség elérhetősége, a helyi gazdaság szerkezete, nagyvállalatok jelenléte, és még sorolhatnánk. Hogyan tudjuk mégis e sok-sok tényező ellenére az adatokban alapvetően rejlő tendenciát kimutatni? Lehet-e egyetlen, immár jól magyarázható görbével helyettesíteni e kusza ponthalmazt, kiszűrni az adatok egyediségét, és kiemelni a két változó kapcsolatából azt, ami általános, ami szabályszerű? Az egyik lehetséges megoldás erre az adatok átlagolása: független változónk mentén osztályközöket hozunk létre, majd kiszámítjuk e csoportok átlagait. Ezeket az értékeket *feltételes átlagoknak* hívjuk, míg az őket összekötő görbét *tapasztalati regressziógörbének.* E módszernek azonban számos, itt nem részletezendő gyengéje van, melyek miatt valós empirikus kutatások során már nem szokás alkalmazni. A valódi megoldást az *analitikus regresszió számítás* adja.

4.4.2 A kétváltozós lineáris regresszió számítási menete

E kétváltozós lineáris regresszió használatakor során egyetlen görbébe igyekszünk sűríteni szórásdiagramunk pontjait, és azt a matematikai függvényt keressük, amely e görbét leírja. E *regressziós görbe* tehát mintegy összegzi a két változó kapcsolatát, az abban megnyilvánuló tendenciát. Azt a függvényt, mellyel a két változó közötti általános összefüggést, azaz a regressziós görbét leírjuk, *regressziós egyenletnek* nevezzük. A függő és független változó közötti kapcsolat az őt leíró függvény alakja szerint alapvetően kétfajta lehet: *lineáris* és *nem lineáris.* Mi a továbbiakban főként a *lineáris regresszióval* foglalkozunk, mivel ez a legegyszerűbb, egyúttal legszélesebb körben alkalmazott regressziós függvény-típus. Azt mutatjuk be tehát, hogy – fenti példánknál maradva – miképpen tudjuk megkeresni azt az egyenest, mely legpontosabban megmutatja számunkra, hogy a városlakók arányának növekedésével mennyiben jár együtt a kistérségek lakossági jövedelmeinek növekedése.

Ehhez első lépésként idézzük vissza középiskolai matematika tanulmányainkat, és írjuk fel magunknak az egyenes egyenletét:

$$\hat{Y} = A + BX \qquad (4/1)$$

ahol:

X: a magyarázó változó;

A: regressziós állandó (konstans); értéke megegyezik az egyenes y-tengelyen tapasztalt metszéspontjával, vagyis *A* értéke egyenlő $\hat{Y}$ értékével *X*=0 helyen;

B: regressziós együttható (regressziós koefficiens); ez jelöli az egyenes meredekségét vagy dőlését, vagyis azt mutatja meg, hogy *X* értékének egységnyi növekedése $\hat{Y}$ értékének mekkora mértékű és milyen irányú változását vonja maga után;

$\hat{Y}$: a függő változó regressziós egyenlet alapján becsült értéke; más megfogalmazásban: értékei a függő változónak a független változó adott értékéhez tartozó átlagos szintjét jelzik (feltéve, hogy adott X_i értékekhez több Y_i érték is tartozik, ami területi adatsoroknál ritkán fordul elő). Ezek az átlagértékek fejezik ki azt a tendenciát, szabályszerűséget, amit adatsoraink együttes változásában keresünk. $\hat{Y}$ értékei egytől egyig pontosan a regressziós egyenesen találhatók; ezek építik fel a vizuálisan is megjeleníthető trendvonalat.

Ettől az átlagos szinttől, vagyis $\hat{Y}$-től Y függő változónk tapasztalati, ténylegesen megfigyelt értékei a mérési hibák, illetve a függő változót befolyásoló – részben már említett – egyéb jelenségek miatt kisebb-nagyobb mértékben eltérnek. Ezeket az eltéréseket *reziduumoknak* hívjuk, és általában *„e”* betűvel jelöljük. Függő változónk tapasztalati Y értékei tehát két részből tevődnek össze: egyrészt $\hat{Y}$-ből, mely a szabályszerűséget jeleníti meg, másrészt az egyes *„e”* értékekből, melyek a tendenciától való eltérést, az *egyediséget* adják. Ebből adódóan:

$$Y = \hat{Y} + e \qquad (4/2)$$

Behelyettesítve (4/2)-t (4/1)-be:

$$Y = A + BX + e \qquad (4/3)$$

Míg a (4/1) a két változó közötti együttmozgásra, a szabályszerűségre utal, addig a (4/3) e kapcsolat sztochasztikus jellegére hívja fel a figyelmet: van ugyan tendenciózus együttmozgás, ám az csak átlagosan, *valószínűségi alapon* érvényesül.

Mi a lineáris regressziós egyenlet alkalmazása során azon egyenes egyenletét keressük, ahol az Y és az $\hat{Y}$ közötti eltérések átlagosan a legkisebbek, vagyis ahol a reziduumok összege minimális (4/4).

$$\sum (Y - \hat{Y}) = \text{minimum} \qquad (4/4)$$

Ez a feltétel azonban nem elégséges a legjobban illeszkedő egyenes megleléséhez. Itt ugyanis az $\hat{Y}$ értékei által előálló regressziós egyenesünk egyfajta *átlagértékként* viselkedik, melytől pozitív és negatív irányban is összességében ugyanolyan messze vannak a függő változó egyes Y_i tapasztalati értékei. Reziduumaink egy része pozitív előjelű lesz tehát, ám összegszerűen ugyanolyan mennyiségben fogunk találni negatív eltéréseket is. A különböző előjelű értékek szummázáskor ki fogják oltani egymást, tehát pusztán a reziduumok összege alapján ugyanolyan jónak fogunk ítélni egy rosszul illeszkedő egyenest, mint egy nálánál sokkal jobbat. Éppen ezért egy további feltételt is be kell vezetnünk, melyhez kétféle választási lehetőségünk van: vagy a reziduumok abszolút értékeivel, vagy pedig négyzeteivel dolgozunk a továbbiakban. Mindkét módon megoldjuk az előjelek különbözőségéből adódó problémát. Igazi választási lehetőségünk persze nincs: *a regressziós egyenes egyenletének meghatározásakor mindig a legkisebb négyzetek módszerét használjuk,* azaz a reziduumok négyzetes összegének minimumát keressük (az angol nyelvű szakirodalomban gyakran találkozunk ennek rövidítésével: OLS: Ordinary Least Squeres).

A legkisebb négyzetének követelményét az alábbi formában írhatjuk fel:

$$\sum_i (Y - \hat{Y})^2 \rightarrow \min \qquad (4/5)$$

Gondolatmenetünk továbbvezetéséhez helyettesítsük be a (4/5)-be (4/1)-et:

$$\sum_{i=1}^{n} \left(Y - \hat{Y}\right)^2 = \sum_{i=1}^{n} (Y - A - BX)^2 \qquad (4/6)$$

Mint látható, adott X és Y értékek mellett a reziduumok négyzeteinek összege kizárólag a regressziós állandó (A), valamint a regressziós együttható (B) függvénye. E két paraméter változtatásával érhetjük tehát el, hogy regressziós egyenesünk a lehető legjobban közelítse függő változónk tapasztalati értékeit. Differenciálszámítás segítségével tudjuk egyenletünket megoldani úgy, hogy biztosak lehessünk benne: megtaláltuk A és B azon értékét, mely mellett a reziduumok négyzetösszege minimális lesz. A végeredmény B-re és A-ra a következő lesz:

$$B = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} d_{X_i} d_{Y_i}}{\sum_{i=1}^{n} d_{X_i}{}^2} \qquad (4/7)$$

$$A = \overline{Y} - B\overline{X} \qquad (4/8)$$

(A levezetés részleteit lásd*: Moksony F*. 1999, pp. 59-62.; *Székelyi M. – Barna I. 2002*, pp. 205-209.)

Az alábbiakban egy egyszerű példán keresztül nézzük végig számításaink menetét! A fejezet elején felvázolt problémánál maradva adatbázisunkból kiragadtunk hat kistérséget, majd melléjük rendeltünk két változót: a városlakók arányát, illetve az adóköteles jövedelmek egy lakosra jutó 2000. évi értékét. Ebben a hatelemű sokaságban is arra a kérdésre keressük a választ, hogy vajon a magasabb városodottság magasabb jövedelem szinttel jár-e együtt. Az alábbi egyenletet oldjuk meg tehát:

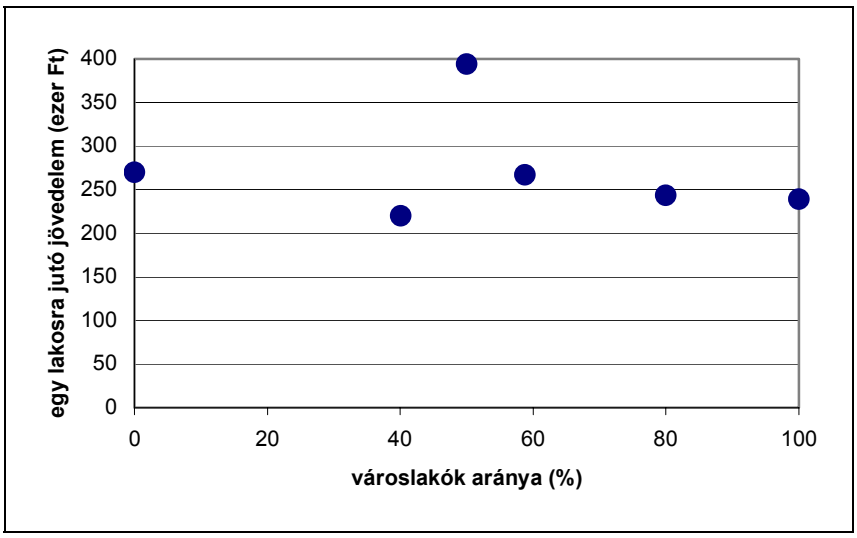
$$\hat{Y} = A + B * URBPOP01 \qquad (4/9))$$

Számításaink kezdő lépéseit a *4.3. táblázat* szemlélteti.

Kistérség	városlakók aránya (%)	jövedelem/fő (ezer Ft)	d_x	d_y	$d_x * d_y$	d_x^2	d_y^2
Dombóvári	60	266	5	-5	-25	25	25
Hajdúböszörményi	100	240	45	-31	-1395	2025	961
Őriszentpéteri	0	270	-55	-1	55	3025	1
Sárvári	50	390	-5	119	-595	25	14161
Sátoraljaújhelyi	40	220	-15	-51	765	225	2601
Szarvasi	80	240	25	-31	-775	625	961
összeg	330	1626			-1970	5950	18710
átlag	55	271					

4.3. táblázat *A regresszió számítás alaplépései*

A fent már bemutatott szórásdiagram segítségével ábrázoljuk a városlakók aránya és a jövedelem együttes eloszlását e hat kistérségben *(4.6. ábra)*!



4.6. ábra *A városlakók aránya és a jövedelem együttes eloszlása hat kistérségben*

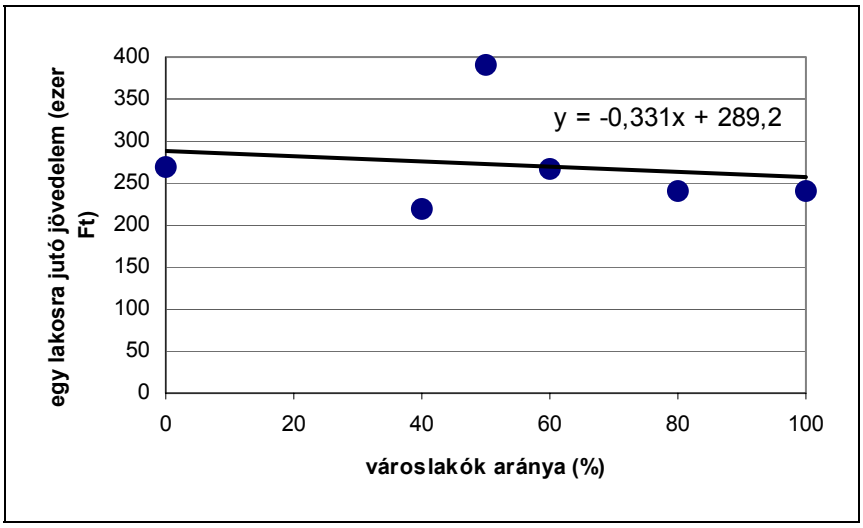
A hat pontra legjobban illeszkedő egyenes meredekségét a (4/7) alapján az alábbiak szerint számítjuk ki:

$$B = \frac{\sum_{i=1}^{n} d_{X_i} d_{Y_i}}{\sum_{i=1}^{n} d_{X_i}^{\ 2}} = \frac{-1970}{5950} = -0{,}331$$

A regressziós egyenes és az y-tengely metszéspontját (4/8) szerint a következő képlet adja meg:

$$A = \overline{Y} - B\overline{X} = 271 - (-0{,}331 * 55) = 289{,}2$$

A *4.6. ábrán* látható szórásdiagram pontjaira legjobban illeszkedő egyenest és annak egyenletét a *4.7. ábra* szemlélteti.



4.7. ábra Két vizsgált változónk kapcsolata a hat kistérségben

Lássuk, mit is kaptunk eredményül! Várakozásunk az volt, hogy a városlakók arányának növekedésével kistérségeink jövedelmi értékei is emelkedni fognak. Ezzel szemben azt látjuk, hogy regressziós egyenesünk balról jobbra lejt, tehát minél inkább növekszik a városlakók aránya, regressziónk szerint annál alacsonyabb lesz a jövedelmek szintje. Megdőlt volna hipotézisünk? E pillanatban úgy tűnik, igen.

A két változó közötti kapcsolat jellegét, összefüggésük irányát - az egyenes szabad szemmel is jól látható balról jobbra való lejtésén túl – a regressziós együttható előjele árulja el számunkra. *Ha B előjele pozitív, két vizsgált változónk között egyenes arányosság, ha pedig negatív, akkor fordított arányosság áll fenn.* Előbbi esetben egyenesünk balról jobbra emelkedik, utóbbi esetben balról jobbra lejt. (Ha regressziós vizsgálatainkat megelőzve két változónk között kiszámítjuk a Pearson-féle lineáris korrelációs együttható értékét, annak előjele már elárulja számunkra a regressziós együttható előjelét is, hiszen e kettő mindig megegyezik egymással.) Mit jelent a -0,331-es érték? Azt, hogy a városlakók arányának minden 1%-os emelkedése átlagosan a lakossági jövedelmek 331 forintos csökkenését vonja maga után.

A regressziós állandó, *A* értéke 289,2, vagyis regressziós egyenesünk ennyi ezer forintnál metszi az y-tengelyt. Ezzel ismerjük egyenesünk magasságát, és még valamit: ez a függő változónak azon értéke, amit regressziónk a független változó nulla értékéhez becsül. Példánknál maradva: regressziónk szerint ennyinek kellene lennie a jövedelemnek abban a kistérségben, ahol nincs egyetlen város sem, így a városlakók aránya nulla.

Mint a *4.3 táblázat* és *ábráink* mutatják, éppen van egy olyan kistérségünk – az Őriszentpéteri – ahol nincs város, így nincsenek városlakók sem. E kistérség lakosainak átlagjövedelme 270 ezer Ft. A regressziós állandó szerint azonban ennek 289,2 ezer Ft-nak kellene lennie. Tévedtünk tehát majdnem húsz ezer forintot. És ha megnézzük a másik öt pontot is, azt látjuk, hogy egyedül a dombóvári kistérség esetében becsültünk a lakosoknak majdnem pontosan annyi jövedelmet, mint amennyi után a valóságban is adóztak. Mennyire vehetjük hát komolyan a regressziós egyenest? Erre a kérdésre felelünk az alábbiakban.

Első lépésként a függő változó, vagyis a jövedelmek heterogenitását írjuk le a variancia, vagyis a szórásnégyzet segítségével. Ehhez szükségünk van az egyes Y_i értékek $\overline{Y}$ -tól való négyzetes eltéréseinek ($d_y^{\,2}$)összegére. Ez az összeg a *4.3 táblázatban* megtalálható, így már csak az elemszámmal kell osztanunk, és előttünk áll a jövedelmek varianciája. Vagyis:

$$\sigma^2(Y) = \frac{\sum_{i=1}^{n} d_y^{\,2}}{n} = \frac{18710}{6} = 3118{,}3$$

Most már csak azt kell tudnunk, hogy ezen belül mekkora az a heterogenitás, amelyet maga a regressziós egyenes fejez ki azzal a tulajdonságával, mely szerint a városlakók arányának térségek közötti változásával egy egyenessel leírható módon változnak a lakossági jövedelmek is. E variancia meghatározásához először is ki kell számítanunk mind a hat kistérség esetében a regresszió által becsült jövedelem értékeket. Ez az egyenes egyenletének ismeretében könnyen megtehető *(4.4. táblázat).*

Kistérség	városlakók aránya (%)	$\hat{Y} = -0{,}331X + 289{,}2$	$\hat{d}_Y = \hat{Y}_i - \bar{\hat{Y}}$	$\hat{d}_Y{}^2$
Dombóvári	60	269,3	-1,7	2,7
Hajdúböszörményi	100	256,1	-14,9	222,0
Őriszentpéteri	0	289,2	18,2	331,6
Sárvári	50	272,7	1,7	2,7
Sátoraljaújhelyi	40	276,0	5,0	25,0
Szarvasi	80	262,7	-8,3	68,5
∑	330	1626,0		652,6
Átlag	55	271,0		

4.4 táblázat *A becsült jövedelmek értékei*

A regressziós egyenes varianciája:

$$\sigma^2(\hat{Y}) = \frac{\sum_{i=1}^{n} \hat{d}_i{}^2}{n} = \frac{652{,}6}{6} = 108{,}8$$

Ez az érték azt mutatja meg, hogy mennyire szóródnának a jövedelmek, ha változásuk csak és kizárólag a városlakók arányának mértékétől függne, vagyis ha a városlakók aránya determinisztikusan határozná meg a lakossági jövedelmek nagyságát. Ez a variancia tehát az az érték, amit a regresszió a függő változó tapasztalati értékeinek heterogenitásából megmagyaráz. Láthatjuk, hogy a két variancia egy nagyságrenddel különbözik egymástól. Regressziónk tehát elég kis részét tudja magyarázni az eredeti jövedelem adatok szórásnégyzetének. De hol keressük a fennmaradt részt? Az éppen a reziduumok varianciájával lesz egyenlő. Ennek számítása a fentiek ismertében már igen egyszerű *(4.5. táblázat).*

Kistérség	jövedelem/fő (ezer Ft) (Y)	$\hat{Y} = -0{,}331X + 289{,}2$	$e = \hat{Y} - Y$	$e = \left(\hat{Y} - Y\right)^2$
Dombóvári	266	269,3	-3,3	11,2
Hajdúböszörményi	240	256,1	-16,1	259,2
Őriszentpéteri	270	289,2	-19,2	369,0
Sárvári	390	272,7	117,3	13769,7
Sátoraljaújhelyi	220	276,0	-56,0	3132,2
Szarvasi	240	262,7	-227	516,3
∑				18057,7

4.5. táblázat *A reziduumok értékének számítása*

A reziduumok varianciája tehát:

$$\sigma^2(e) = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)}{n} = \frac{18057{,}7}{6} = 3009{,}6$$

A teljes variancia a megmagyarázott és a megmagyarázatlan rész összegéből tevődik össze:

$$\sigma^2(Y) = \sigma^2(\hat{Y}) + \sigma^2(e) \qquad (4/10)$$

Itt érkeztünk el feltett kérdésünk megválaszolásához, ugyanis a regresszió által megmagyarázott variancia mutatja meg nekünk, hogy mennyire vehetjük komolyan modellünket, mennyire jó a regressziós egyenes illeszkedése.

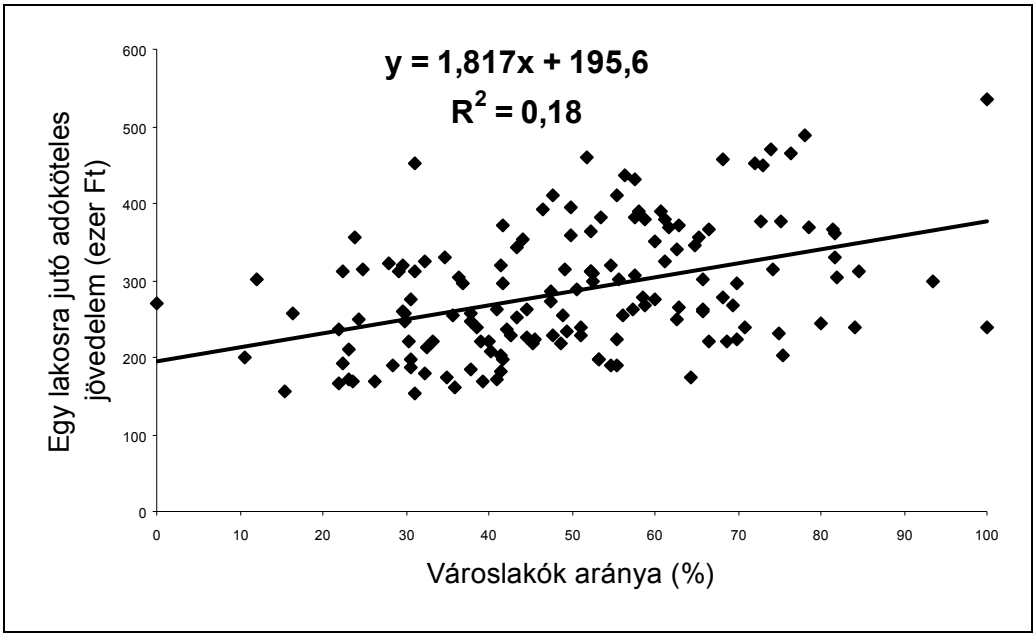
Azt a mérőszámot, ami a megmagyarázott szórásnégyzetet fejezi ki, *determinációs együtthatónak* nevezzük. Jelölése: R^2. Értéke lineáris regresszió esetén mindig egyenlő az ugyanazon két változóra számolt Pearson-féle lineáris korrelációs együttható négyzetével. Értékét általában a (0,1) intervallumban adjuk meg, de szokás százalékos formában is közölni. Ilyenkor értéke azt mutatja meg, hogy a függő változó értékeinek szóródását hány százalékban magyarázza a független változóval fennálló kapcsolata. Kétváltozós regresszió esetén a determinációs együttható egyben a linearitás mérőszáma is: értéke minél közelebb esik 1-hez, megfigyelési egységeink pontjai annál inkább egy egyenes mentén tömörülnek, tehát annál inkább igazoltnak látszik a kiindulási alapnak tekintett linearitási feltétel. A determinációs együttható kiszámításának módja:

$$R^2 = \frac{\sigma^2(\hat{Y})}{\sigma^2(Y)} \qquad (4/11)$$

Példánkban: $R^2 = \frac{\sigma^2(\hat{Y})}{\sigma^2(Y)} = \frac{\frac{652,6}{6}}{\frac{18057,7}{6}} = 0,036$

E hat kistérség esetében tehát a városlakók aránya a lakossági jövedelmek heterogenitásának mindössze 3,6%-át magyarázza meg, ami elenyészően kicsi hányad. Ez alapján akár azt is mondhatnánk, hogy modellünk rossz, hiszen rengeteg egyéb hatást hagytunk megmagyarázatlanul, ami kérdésessé teszi eredményünket is.

Nézzünk most egy valóságos példát! A *4.5. ábrán* már korábban felvázoltuk a 150 kistérség szórásdiagramját, ahol független változóként a városlakók aránya, függő változóként a lakossági jövedelmek szerepelnek. Továbbra is azt keressük, hogy a városodottság növekedése mennyiben jár együtt a jövedelmek növekedésével, vagy éppen csökkenésével. Regressziónk eredményeit a *4.8. ábra* közli.



4.8. ábra *A városodottság és a jövedelmek kapcsolata kistérségeinkben, 2000.*

Mint látjuk, mind a 150 kistérség bevonásával elvégzett regresszió analízisünk gyökeresen más eredményt hozott, mint amit a leegyszerűsített, mindössze hat elemre számított példánk mutatott. A valóságban *magasabb városodottsági szint magasabb lakossági jövedelmekkel jár együtt*. Ezt a regressziós együttható (*B*) értéke bizonyítja: előjele pozitív, tehát független és függő változónk egyenes arányosságban áll egymással. Azt az eredményt kaptuk, hogy a városlakók arányának minden 1%-pontos növekedése általában a lakossági jövedelmek 1817 Ft-os növekedésével jár együtt.

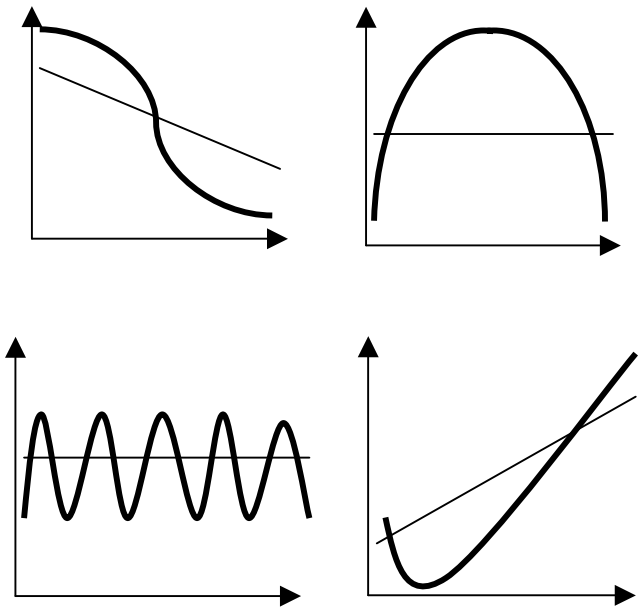
Regressziós állandónk (*A*) értéke 195,6, vagyis az országos trendből következően ennyinek kellene lennie a lakossági jövedelem szintjének abban a kistérségben – itt most az őriszentpéteriben –, ahol nincs város, vagyis a városlakók aránya nulla. Az őriszentpéteri kistérség valós éréke viszont 270 ezer Ft, amiből következően pozitív reziduummal rendelkezik, tehát a lineáris trendvonal „fölött" helyezkedik el. Más szóval: ebben a kistérségben több jövedelem jut egy lakosra, mint ami az ország egészét jellemző trendből következne számára.

A determinációs együttható értéke is jóval magasabb, mint a hat pont esetén: az összes kistérséget figyelembe véve a városlakók arányának szóródása a lakossági jövedelmek heterogenitásának 18%-át magyarázza meg.

A kapott eredmények ismeretében felmerülhet bennünk a kérdés: mi az oka annak, hogy a teljes kistérségi mintát alapvetően más összefüggések jellemzik, mint amiket a hat kiragadott pont esetén tapasztaltunk. Ez alapvetően az alacsony elemszámmal függ össze: hat pontra nem tudunk regressziós modellt építeni. Hogy miért nem? Mert ez esetben szinte teljesen biztosak lehetünk abban, hogy regressziónk nem fog szignifikáns részt magyarázni a függő változó teljes heterogenitásából. E szignifikancia nem függ össze a determinációs együtthatóval: legyen bár az R^2 értéke bármilyen alacsony is, ettől még maga a modell szignifikáns hányadot testesíthet meg eredmény változónk heterogenitásából. (A modell e tulajdonságát F-próbával lehet tesztelni.)

4.4.3 Nem lineáris összefüggések

Az előzőekben már utaltunk rá, hogy a determinációs együttható a linearitás mérőszámaként is funkcionál. E tulajdonságot minden regressziós modell esetében tesztelnünk kell, hiszen empirikus vizsgálatok esetében nem lehetünk biztosak abban, hogy a független és a függő változó közötti kapcsolat valóban lineáris. Különösen akkor kell nagyon ügyelnünk, ha nem készítjük el változóink szórásdiagramját, kizárva ezzel annak lehetőségét, hogy vizuálisan érzékeljük: két változónk *nem lineáris* összefüggést mutat. Az ilyen jellegű kapcsolatok néhány alapesetét mutatják az alábbi ábrák *(4.9. ábra)*.



4.9 ábra *A nem lineáris összefüggések néhány alaptípusa*

A nem lineáris regressziós egyenletek alaptípusai függvényformában a következők:

logaritmikus: $\hat{Y} = A + (B * \ln X)$

polinomiális: $\hat{Y} = A + (B_1 * X) + (B_2 * X^2) + (B_3 * X^3) + ... + (B_n * X^n)$

exponenciális: $\hat{Y} = A * B^X$

hiperbolikus: $\hat{Y} = A + \frac{B}{X}$

hatványkitevős: $\hat{Y} = A * X^B$

Azt, hogy melyik regressziós egyenlet-típus írja le legjobban az adott összefüggést, a determinációs együttható kiszámításával tudjuk eldönteni. A nem lineáris regressziós egyenletek elemzése és értelmezése nehézkesebb, mint a lineáris egyenleteké, használatuk az idősorelemzésben, trendszámításban jóval gyakoribb, mint a területi adataok keresztmetszeti vizsgálatában.

4.4.4 Logaritmikus regressziók (Major Klára)

A regresszió-számításban a lineáris regressziók mellett messze a leggyakoribbak a változók logaritmizált értékeire épülő modellek. Ezért mindenképp indokolt a logaritmizálás tartalmi valamint matematikai és statisztikai okainak felsorakoztató áttekintése.

- Vannak olyan elemzési technikák is, melyek segítségével nem lineáris összefüggést lineárisra tudunk visszavezetni. Ha a függő, illetve a független változók közötti kapcsolat nem lineáris, hanem a függő változót a függetlenek *szorzata* adja (ez a logikája a közgazdaságtan leghíresebb termelési függvényének, az ún. *Cobb-Douglas függvénynek*, amelyben a termelés a munka- és a tőkeráfordítás szorzatának függvénye), a kapcsolat lineáris regresszióval csak logaritmikus átalakítás segítségével vizsgálható:

  az $\hat{Y} = A * X_1 * X_2$ logaritmizálva a következő alakot ölti:

  $$\log \hat{Y} = \log A + \log X_1 + \log X_2$$

  Ez már lineáris összefüggés, s a regresszió-elemzéssel a szokott módon vizsgálható.

- Mivel a legtöbb területi adat nem követi a matematikai-statisztika legalapvetőbb eljárásai során megkövetelt normalitást (azaz nem szimmetrikus, *nem normális eloszlású* → **2.5**), a matematikai kritériumokat közelítendő számos esetben alkalmazzák az alapadatok logaritmikus átalakítását , azaz az eredeti értékek helyett azok logaritmusait használják az elemzésekben. Ez az átalakítás a logaritmusfüggvény monoton növekvő jellege miatt az adatok nagyság szerinti sorrendjét nem változtatja meg (ha *a* kisebb, mint *b*, akkor *loga* is kisebb mint *logb*), az átalakítás viszont közel szimmetrikussá teszi az eloszlást, hisz a kiugróan nagy alapadatokat a logaritmikus transzformáció „összébb húzza". Még inkább előtérbe állítja ezt az átalakítást az, hogy a területi adatoksorok esetén nagyon gyakori a *lognormális* jellegű eloszlás. Ha ekkor a lognormális eloszlású változó helyett annak logaritmusával számolunk, normális eloszlásúra transzformáltuk adatainkat.
- A logaritmizálással jelentősen csökkenthetők a kiugró adatok torzító hatásai a regresszió-számításban. (Egy átlag-közeli magyarázó változó értéknek jóval kisebb hatása van a regressziós összefüggésre, mint a skálák szélein található adatoknak.)
- A logaritmizálás csökkenti az adatok *heteroskedascitásából* eredő torzulást (ez akkor lép fel, ha a függő változó szórása függ a magyarázó változótól, annak növekvő értékei mentén például növekszik)

Logaritmizált adatokon végzett regresszió értelmezése

Hogyan lehet értelmezni az adatok logaritmusán végzett regresszió eredményeit? A választ megmutatjuk mind idősor-, mind keresztmetszeti vizsgálat esetén.

Idősor modellek

Tekintsük két logaritmizált adat különbségét és végezzük el az alábbi átalakítást:

$$y_t - y_{t-1} = \ln\frac{Y_t}{Y_{t-1}} = \ln(1+\gamma) \approx \gamma$$

ahol γ a jövedelem növekedési ütemét jelenti, azaz a példánkban $\gamma=(Y_t-Y_{t-1})/Y_{t-1}$. (A becsléshez azt az ismert összefüggést használtuk fel, hogy ha $x\approx 0$ nullához közeli érték, akkor $\ln(1+x)\approx x$.) Logaritmizált adatok különbsége tehát közel megegyezik az adatok százalékos változásával (növekedési ütem), így a logaritmizált adatokon végzett regresszió esetén az egyenlet együtthatója x és y *százalékos változása* között létesít számszerű kapcsolatot. *A logaritmizált adatokból nyert regressziós egyenlet paraméterének számszerű értéke az y és x változók növekedési ütemének arányával egyezik meg.*

Ehhez a gondolatmenethez tételezzük fel, hogy a regresszióban szereplő y és x változókról nem különböző megfigyelési egységek, hanem egyazon megfigyelési egység különböző időpontbeli adataival rendelkezünk. Jelöljük a különbséget úgy is, hogy nem i-vel, hanem t-vel indexeljük az egyenletet:

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Vonjuk ki az egyenlet t-ik időpontban felírt alakjából a $t-1$ időpontra felírt egyenletet és tekintsünk el átmenetileg a véletlen hibától. A kapott egyenlet mindkét oldalán végezzük el a fent már bemutatott becslést. Eredményül azt kapjuk, hogy $\gamma_Y=\beta\gamma_X$.

Keresztmetszeti vizsgálat

Tekintsük újra a regresszió eredeti formáját, és az adatok i indexe továbbra is jelentse az egyes megfigyelési egységeket. A regressziós egyenlet β paramétere azt mutatja meg, hogy $x_i=lnX_i$-ben bekövetkezett egységnyi változás hogyan hat $y_i=lnY_i$-re. Némi matematikai eszköztár felhasználásával megmutatható, hogy ez az érték az eredeti adatok százalékos változása között létesít kapcsolatot, *azaz azt mutatja meg, hogy X_i 1 százalékos változása hány százalékos változást okoz Y_i-ben.* Az ilyen típusú mutatószámokat *rugalmassági* mutatószámoknak hívjuk.

A logaritmikus adatokon végzett regressziós elemzés β együtthatója tehát rugalmasság, amely a vizsgált adatok százalékos változása közötti kapcsolatot számszerűsíti.

A regressziós egyenlet β paramétere x_i megváltozásának hatását mutatja y_i-re, azaz alkalmazva az analízis elemi eszközeit felírhatjuk, hogy

$$\beta = \frac{dy}{dx} = \frac{d\ln Y}{d\ln X}.$$

β eredeti adatokon való értelmezéséhez tehát az utóbbi derivált értékét kell meghatároznunk. Ehhez az alábbi aprócska „trükköt” használjuk fel. A láncszabály alkalmazásával kapjuk, hogy

$$\frac{d\ln Y}{d\ln X} * \frac{d\ln X}{dX} = \frac{d\ln Y}{dX}$$

Ezt átrendezve és a láncszabály újbóli alkalmazásával kapjuk meg a kívánt összefüggést:

$$\beta = \frac{d\ln Y}{d\ln X} = \frac{d\ln Y}{dX} : \frac{d\ln X}{dX} = \frac{1}{Y}\frac{dY}{dX} : \frac{1}{X} = \frac{X}{Y}\frac{dY}{dX}$$

A kifejezés jobb oldalán szereplő érték már valójában maga a rugalmasság definíciója. Az ennek elméletében kevésbé jártas olvasó számára az értelmezést megkönnyítendő végezzük még el utoljára az alábbi, merőben formális és *heurisztikus* átalakítást:

$$\beta = \frac{X}{Y}\frac{dY}{dX} = \frac{dY}{Y} : \frac{dX}{X}$$

Természetesen ez az átalakítás heurisztikus, azaz a szemléltetést szolgálja, de jól mutatja a rugalmassági együttható értelmezési lehetőségét: *az Y változó relatív (százalékos) változását (dY/Y) viszonyítja az X változó relatív (százalékos) változásához (dX/X).*

Loglineáris regresszió

Semilog (vagy loglineárisnak) nevezzük a regressziót, ha abban a függő változó (y) logaritmizált értékeit nem logaritmizált változókkal magyarázzuk. Például tekintsük az alábbi regressziós egyenletet:

$$\ln y_t = \alpha + \beta x_t + \gamma t + \varepsilon_t$$

Ebben az egyenletben γ jelöli az y változó növekedési ütemét (linearizált trend), a trend körüli ingadozásokat pedig részben a véletlen folyamatok (ε_t), részben az x magyarázó változó ingadozásai okozzák.

Logisztikus modell

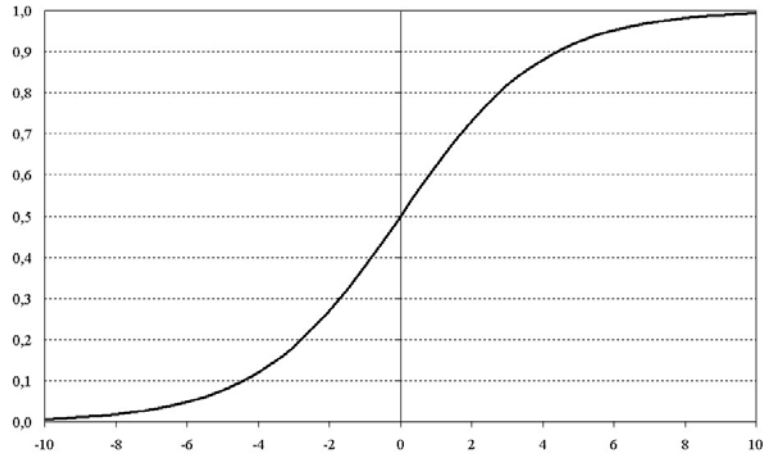
Telítődési, terjedési folyamatok modellezése során lehet szükség a logisztikus függvényre. Ilyenkor olyan folyamatokat vizsgálunk, amelyek esetén meglehetősen speciális időbeni mintát követnek. Például a technológiai fejlődés terjedését először lassú, majd ahogy egyre ismertebbé és kedveltebbé válik, egyre gyorsabb ütem jellemzi. Amennyiben y_t jelöli mondjuk egy technológiai újítás elterjedtségét, úgy a logisztikus összefüggést a következő regresszió írná le:

$$y_t = \frac{1}{e^{-(\alpha+\beta x_t+\gamma t+\varepsilon_t)}+1}$$

Ezen regressziós egyenletben a becsülendő paraméterek nem lineáris függvényei az adatoknak ezért megfelelő adattranszformációra van szükség. A szükséges transzformáció neve: logit transzformáció:

$$\text{logit}(y_t) = \ln\left(\frac{y_t}{1-y_t}\right) = \alpha + \beta x_t + \gamma t + \varepsilon_t$$

Az y adatokon végrehajtott logit transzformáció az eredeti modellt a becsülendő paraméterekben lineárissá alakította át, és így már alkalmazható az OLS technikája (*4.10. ábra).* A függvény értékkészlete a (0, 1) intervallum, így leginkább százalékos adatok magyarázatára (terjedési, telítődési folyamatok) alkalmazható.



4.10. ábra *A logisztikus függvény ($y=1/(e^{-x}+1)$) alakja*

4.4.5 Többváltozós regresszió

Visszatérve a *4.8. ábrán* vázolt példánkra, felmerülhet az a kérdés, hogy a városlakók arányán kívül melyek még azok a tényezők, amik szoros kapcsolatban állnak a lakossági jövedelmekkel úgy, hogy változásuk a jövedelmek átlagos növekedését vagy csökkenését vonja maga után, tehát egy regressziós modellben magyarázó változóként lehet szerepeltetni őket. Ezzel kapcsolatban több társadalmi-gazdasági tényező is eszünkbe juthat. Ez lehet például a lakosság iskolázottsági szintje, ami a munkaerőpiacon való érvényesülésen át hat a jövedelmekre, vagy a cégsűrűség, ami a foglalkoztatás szintjén keresztül befolyásolja a jövedelmeket, vagy éppen a jó elérhetőség, ami elsősorban a nagyvállalatok, külföldi befektetők szemében értékeli fel az egyes térségeket, és segíti elő ezek megtelepedését, ami tömeges jelenségként már határozott jövedelmi előnnyel jár e területek lakói számára. E tényezők jövedelmekre gyakorolt hatását külön-külön jól tudjuk mérni a fent vázolt módon, kétváltozós regresszió analízis segítségével. Mit tegyünk azonban akkor, ha e tényezők jövedelmekre gyakorolt *együttes* hatására vagyunk kíváncsiak? Ekkor kell használnunk a *többváltozós regresszióanalízis* módszerét.

A következőkben is maradjunk fent bemutatott példánknál, tehát továbbra is az a fő kérdés, hogy a városlakók arányának változása mennyiben befolyásolja a kistérségben bevallott lakossági jövedelmek szintjét. Ám ezzel párhuzamosan vizsgáljuk meg azt is, hogy a közúti elérhetőségben tapasztalható differenciák miként módosítják e viszonyrendszert. Ma a közúti közlekedés fő ütőereit az autópályák jelentik. A legtöbb befektetni kívánó cég számára telephely választáskor fontos szempont az adott térség vagy település autópályához való közelsége, hiszen a rövidebb szállítási idő és távolság csökkenti a cég költségeit. Autópályák közelében tehát több befektetőt és nagyobb értékű befektetéseket kell találnunk, mint a közlekedési „árnyékban" lévő területeken. A több befektetés viszont magasabb foglalkoztatottsággal és így több jövedelemmel is együtt jár. Nézzük meg tehát, hogy a városlakók aránya és az egyes kistérségek településeinek legközelebbi autópálya csomóponttól mért átlagos közúti távolsága (APCSPONT) miként befolyásolja a lakossági jövedelmek térségi differenciáit, és e két magyarázó változó közül melyik van erősebb hatással a függő változóra. Ebben az esetben tehát már nem egy egyenes, hanem a két magyarázó változó által kifeszített sík segítségével próbálunk meg becslést adni a függő változóra. Ebben az esetben más egyenlettel kell dolgoznunk, mint amikor csak egy magyarázó változó lépett be a modellbe. A legalább két független változó által kijelölt regressziós síkhoz is egyetlen regressziós konstans tartozik, a regressziós együtthatók száma viszont több lesz: pontosan annyi, ahány független változó a modellt felépíti. Az alábbi egyenletet oldjuk meg tehát:

$$\hat{Y} = A + B_1(URBPOP01) + B_2(APCSPONT)$$

R	R^2 *(R Square)*	Korrigált R^2 *(Adjusted R Square)*	A becslés standard hibája *(Std. Error of the Estimate)*	F	Sig.
0,646	0,417	0,409	62,846	52,56	0,000

	Standardizálatlan együtthatók *(Unstandardized coefficients)*		Standardizált együttható *(Standardized Coefficients)*		
	B	Standard hiba *(Std. Error)*	Béta *(Beta)*	t	Sig.
Konstans *(Constant)*	269,071	17,336		15,521	0,000
URBPOP01	1,527	0,274	0,354	5,563	0,000
APCSPONT	-0,796	0,102	-0,494	-7,772	0,000

4.6. táblázat *Többváltozós regressziónk eredménytáblája*

Számításaink főbb eredményeit a *4.6. táblázat* tartalmazza. (A modellt a területi kutatások talán legkedveltebb szoftverén, az *SPSS* programon futtattuk le → **4.5.7,** így eredményeink szemléltetésénél – a jobb gyakorlati elsajátíthatóság érdekében – nagyrészt e program outputjaihoz igazodunk.)

Többszörös determinációs együttható

A táblázat felső részében szereplő R a többszörös korrelációs, az R^2 pedig a *többszörös determinációs együttható* értéke. Ez utóbbi azt jelzi számunkra, hogy a modellben szereplő két magyarázó változó, vagyis a kistérségek városlakóinak aránya és a kistérségek autópályától való távolsága *együttesen* a jövedelmi differenciák mintegy 42%-át magyarázza meg. Ha visszagondolunk kétváltozós regressziós modellünkre és a *4.8. ábrára,* láthatjuk, hogy a determinációs együttható értéke ez utóbbi vizsgálatnál már lényegesen magasabb, tehát egy jobban illeszkedő regressziós modellt konstruáltunk. Ez azonban törvényszerű volt, hiszen az R^2-nek van egy olyan alapvető tulajdonsága, miszerint ha egy már létező regressziós modell magyarázó változóinak számát egy további változóval bővítjük és úgy futtatjuk le regressziónkat, e második, bővített modellhez tartozó R^2 *mindig nagyobb lesz, mint az eredeti, szűkebb modellhez tartozó* R^2, vagy legalábbis egyenlő lesz azzal. (Ez az egyenlőség azonban elvben létezik ugyan, de *gyakorlati vizsgálatok során szinte soha nem fordul elő.*)

A determinációs együttható éppen e tulajdonságai miatt *nem alkalmas különböző modellek összehasonlítására.* Így nem jó stratégia az R^2 értékeit követnünk akkor, amikor a *formálisan optimális* modell megalkotása a célunk. Itt jegyezzük meg, hogy természetesen *teljes tévút* az, amikor magyarázó változóink összeállításánál egyetlen célunk az R^2 nyakló nélküli növelése, a „tökéletes modell" megalkotása. Egyrészt tökéletes modell nem létezik. Másrészt a vizsgált változók körét mindig a kutatási cél és hipotézis, valamint az alkalmazott elméletek döntik el; a helyes regressziós modell előállítása a maga összes statisztikai törvényszerűségével és lehetőségével együtt is csak eszköz, nem pedig elérendő végcél.

Modellválasztási feladat megoldása során a Theil-féle, szabadságfokkal *korrigált R^2-et* szokás alkalmazni, amely tulajdonságaiban, mint neve is mutatja, alapvetően hasonlít az R^2-hez. (Ennek számítási módját és alkalmazási feltételeit részletesen lásd: *Hunyadi, 2000; Hunyadi – Vita 2002*) A becslés standard hibája – az R^2-hez hasonlóan – a regressziós egyenes, több magyarázó változó esetén a regressziós sík illeszkedésére utaló mérőszám. Értéke a reziduumok szórását mutatja. Minél nagyobb ez a szám, annál inkább számíthatunk kiugró adatokra, melyek igen messze esnek a regressziós egyenes vagy sík által számukra becsült értéktől. Az ilyen extrém jellemzőkkel bíró területegységekre, településekre minden regresszióanalízis során külön figyelmet kell fordítanunk, hiszen jelenlétükkel alapvetően befolyásolhatják a modell paramétereit. De hogy találjuk meg őket? A becslés standard hibája és a reziduumok segítenek nekünk ebben. Általában azokat a térségeket tekintjük a trendtől nagyon távolinak, vagyis extrém, kiugró értéknek, amelyekhez tartozó reziduum nagyobb, mint a becslés standard hibájának háromszorosa. És mit tudunk kezdeni velük? Nem sokat. Ha több ilyen értékünk is van, és reziduumaik még az ésszerű határokon belül vannak, akkor tudomásul kell vennünk jelenlétüket és a teljes mintával dolgoznunk tovább. Gyengíteni fogják ugyan a regresszió erejét, de csak ezért nem hagyhatjuk ki őket vizsgálatainkból. Más helyzet áll elő olyankor, ha alapvetően szinte minden vizsgált területet ugyanaz a trend jellemez, de van egy-két teljesen *extrém régiónk.* Ebben az esetben e kiugró értékeket ajánlatos kihagyni a modellből, és az eredmények interpretálásakor visszatérni rájuk. (Ilyen extrém terület Magyarországon például Budapest. Megyei szintű elemzéseknél rendszeresen ki kell hagynunk a modellekből, mert egészen más folyamatok és tulajdonságok jellemzik, mint a másik 19 megyét. Sajnos még a mai hazai szakirodalomban is találkozunk olyan regressziókkal, melyeknek semmi közük sincs a valós folyamatokhoz egy-két ilyen extrém területegység miatt, ám a nagy, ám teljességgel hamis tudományos következtetéseket mégis papírra veti a szerző.)

Multikollinearitás

Olyan esetben áll fenn e probléma, amikor a regressziós modellben szereplő magyarázó változók nem korrelálatlanok egymással, hanem összefüggő rendszert alkotnak. Ez esetben független változóink nemcsak a függő változót, de jórészt *egymást is magyarázzák*, így hatásuk minden más változóban is megjelenik. Ez vezet végül értelmetlen eredményekhez. A multikollinearitás a regressziós modellek talán legnagyobb ellensége, és a gyakorlatlan elemzők talán e téren tudnak legnagyobbakat, de mindenképpen leglátványosabbakat hibázni. Éppen ezért ezt a jelenséget mindenképpen ki kell szűrnünk modelljeinkből. Ennek legegyszerűbb módja, ha *korrelálatlan magyarázó változókat* építünk

be a regresszióba. Ilyeneket persze a valós életben nagyon ritkán találni, de például főkomponens elemzés segítségével elő lehet állítani faktorok formájában → **5.5**. Erre természetesen nincs mindig lehetőségünk, de nincs is mindig szükség rá.

Egy érdekes összefüggés: a sorrend-nagyság szabály

A regressziószámítás alkalmazásához vezet a területi kutatásokban sokáig determinisztikus modellként kezelt sorrend-nagyság (rang-nagyság) vagy Zipf-szabály (*Auerbach, F. 1913*, *Zipf G. 1949*) is. Ez az összefüggés – egy általános egyensúlyi összefüggést, az ún. *Pareto-szabályt* feltételezve[10] (*Illés I. 1975*) – azt tartalmazza, hogy egy térség (ország) nagyság szerinti sorrendbe rendezett városainak népessége (V_r) és a rangsorban elfoglalt helye (r) között függvényszerű kapcsolat van:, mégpedig a lélekszám a legnépesebb város népességének (P) és a rangszámnak hányadosa:

$V_r = P/r$ vagy átalakítva

$R = P/V_r$

A legkülönböző országokra ellenőrizve ezt az összefüggés (újabban pélául *Ioannides Y. M. – Overman H. G. 2003)* azonban kiderül, hogy a városok eloszlása sokszor nem követi ezt a szabályt. Különösen szembetűnő ez olyan országokban, ahol a vezető városnak kiugróan nagy a lélekszáma („pimate-city dominance”).
Erre hazánk is jó példa. Budapest lélekszámából levezetve a hazai városok népességeloszlását, a Zipf-modell „alatti” eloszlást kapunk, ellentétes logikával, amikor például a 100. város lélekszáma alapján becsüljük meg Budapestét a ténylegesnél jóval kisebb népességet kapunk (*4.11. ábra*).

Megkeresendő a tényleges eloszláshoz jól illeszkedő függvényt, a rangszám r valamely hatványával érdemes számolni (ez elméleti alapesetben k =1):

$V_r = P/r^k$

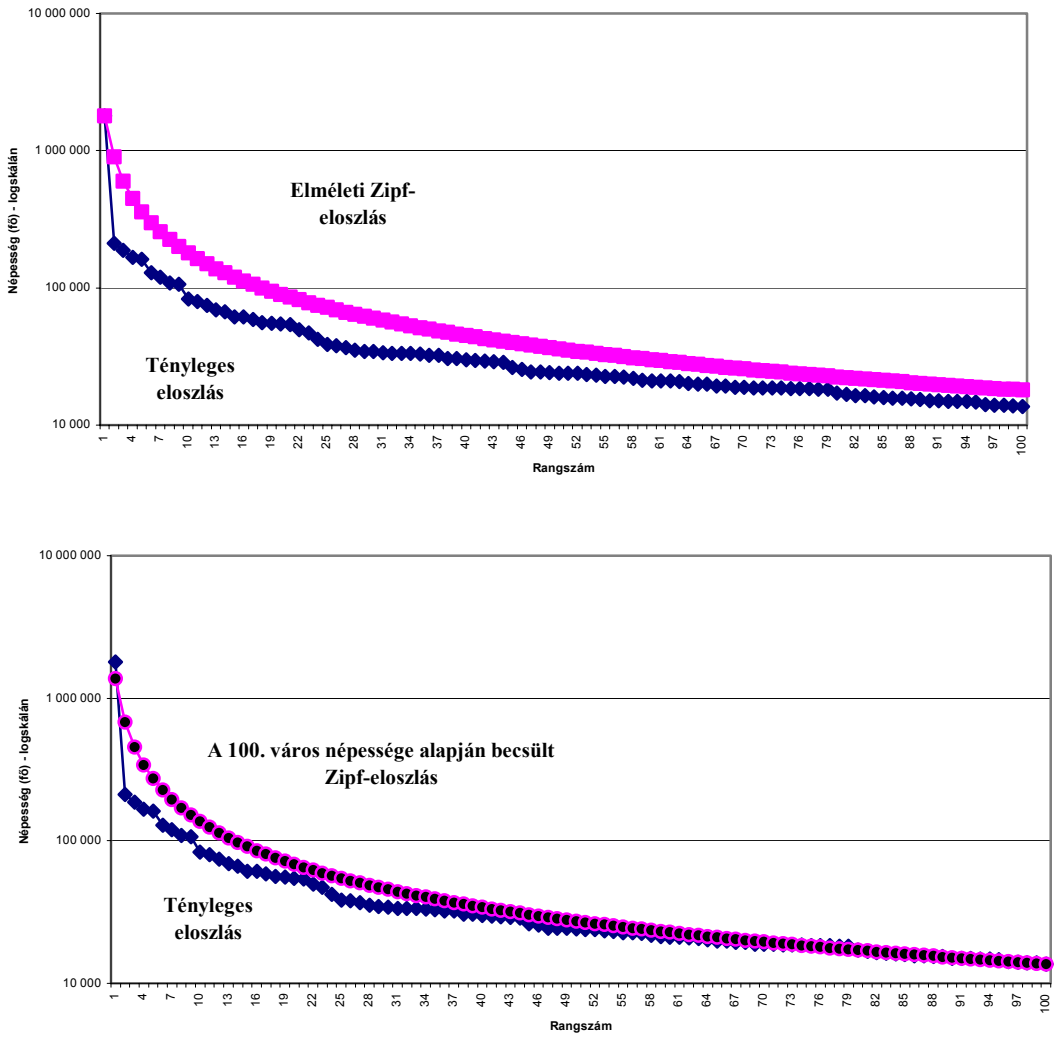
Az egyenletet logaritmizálva:

$\log V_r = \log P + k^* \log r$

Ez pedig egy klasszikus lineáris regressziós egyenlet: bal oldalon log V_r a függő változó, a jobb oldalon log P konstans, *r* 1,2…n értéket felvevő változó, k pedig a becslendő regressziós paraméter.

---

[10] Az eredeti munka*: Pareto, V. Le Cours d' Economie Politique,* Macmillan, London, 1896



4.11. ábra *Az első 100 magyar város lélekszám szerinti eloszlása 2000-ben*

4.4.6 Többváltozós lineáris regresszió az SPSS-ben (Jakobi Ákos)

A közismert statisztikai programcsomag SPSS 8.0 verziójában alkalmazott lépések a következők (csak a legfontosabb beállítási teendőket ismertetjük):

1. Vizsgálati adatbázis összeállítása, illetve megfelelő formátumú behívása
2. *Statistics / Regression / Linear* – belépés a lineáris regresszióanalízis almenübe
3. *Dependent* – a függő változó kiválasztása.
4. *Independent* – a független változó(k) kiválasztása.
5. *Method* – a regressziós eljárás kiválasztása. Az „Enter” metódus az összes változót együttesen alkalmazza a vizsgálatban, a „Forward” eljárás lépésről lépésre vonja be a változókat a vizsgálatba a magyarázó erejük sorrendjében, a „Backward” módszer éppen ellenkezőleg, a magyarázó erő alapján fokozatosan kiveszi a változókat a vizsgálatból.
6. *Statistics* – regressziós alapstatisztikák kiszámítása (pl. kollinearitás-diagnózis, Durbin-Watson próba stb.)
7. *Plots* – az eredmények grafikus megjelenítésének beállításai.

8. *Save* – a mentendő végeredmények kiválasztása. Leggyakrabban a becsült (Predicted Values) és reziduum (Residuals) értékek standardizálatlan (Unstandardized) vagy standardizált (Standardized) változatait mentjük el.
9. *Options* – beállítások, opciók (pl. meghatározható a hiányzó adatok kezelésének módja).

Az eredménytáblák értékelése

A beállítások részletességétől függően számos táblázat, illetve azon belüli részeredmény jeleníthető meg, amelyek közül a legfontosabbakat ismertetjük:

1. *Variables Entered/Removed* – a vizsgálatba bevont vagy eltávolított változók megnevezése lépésenként.
2. *Model Summary* – összegző tábla a regressziós modell eredményességéről. A determinációs együttható (R Square = R^2) és a gyakrabban használt módosított R^2 (Adjusted R Square) értéke megmutatja, hogy a függő változó teljes varianciájából hány százalékot magyaráz regressziós modellünk.
3. *Coefficients* – a regressziós együtthatók táblája. A lineáris regressziós egyenlet megalkotásához szükséges együtthatók standardizálatlan (Unstandardized B) és standardizált formában (Standardized Beta) is leolvashatók, melyek alapján eldönthető, hogy mely független változónk játssza a legmeghatározóbb szerepet egyenletünk szerint a függő változó alakulásában. A tábla közli továbbá a standardizálatlan együtthatók hibaértékeit, valamint az együtthatókhoz tartozó t-értékeket és a kapcsolódó szignifikanciaszintet is.
4. *Residuals Statistics* – az eredményül kapott becsült és reziduum értékek főbb jellemzőit közli a táblázat (minimum, maximum, átlag = Mean, szórás = Std. Deviation)

4.5 A térbeliség regresszió-elemzése

4.5.1. Területi autoregressziós modellek, területi dummyk

Az időbeli autoregressziv modellek mintájára (ahol egy adott y változását önmaga k évvel eltolt értékeivel magyarázzák), felállíthatók olyan regressziós összefüggések, amelyekben a függő változót *szomszédai* hasonló értékeivel magyarázzák (angolul ezeket *spatial lag*-változóknak nevezik). Természetesen többváltozós modellek is szóba jöhetnek, amelyekben a lag-változók mellett a hagyományos közelítésnek megfelelő más, nem a szomszédokra, hanem az eredeti megfigyelési egységekre vonatkozó magyarázó változók is szerepelnek.

A lag-változók mellett a területiség szerepének érzékeltetésére gyakran találkozhatunk a szakirodalomban ún. területi *dummy-változókkal* is. Ezek az adott területi megfigyelési egységek aggregátumait azonosítják. Ha például Európa régió jelentik a kiindulást, dummy váltóként szerepelhetnek egyes országok vagy a kontinens nagyrégiói. Ekkor például a Magyarország-dummyban a 7 hazai régió kap 1-es értéket az összes többi régió pedig 0-át. Ha ez a dummy változó a regresszió-optimalizálás (például a backward eliminació) után is megmarad szignifikáns magyarázó erejűnek, akko az adott jelenségben a „magyarság” önálló faktorként van jelen. Országon belüli vizsgálatoknál dummykkal érzékeltethető egyebek mellet a településnagyság hatása (akkor a különböző nagyságkategóriák adják ezeket a bináris változókat), de sajátos zónák (például a határmentiség) is megjelentethető íly módon a regresszióban. Utóbbi esetben azonban kedvezőbbnek tűnik olyan folytonos távolsági, elérhetőségi változók szerepeltetése, amelyek minden változóhoz a határoktól való távolságot rendeli.

A dummy-változók esetében fontos arra felhívni a figyelmet, hogy a teljes rendszert n aggregátumát leíró változócsoport n-1 dummyból áll, az n.-ik egység ugyanis a többiből kiszámítható. Ha tisztán ilyen magyarázó változókból áll a modell, akkkor az n.-ik megyéhez tartozó érték a modell konstansa

lesz. (Ha a hazai kistérségek elemzéséhez a régiók dummy-jait használjuk, akkkor 6 ilyen változó szerepelhet.)

4.5.2 Térparaméterek

Ha arra a kérdésre keressük a mennyiségi, matematikai választ, hogy a térbeliségnek mekkora szerepe van a társadalom, bizonyos jelenségek tagoltságában, mindenek előtt magát a térbeliséget illetve annak komponenseit kell megragadnunk, számszerűsítenünk. Ilyen lehetőség az, ha a térbeli tagoltság elemeiként három tényezőt:

- a *lokalizálódást* (a földrajzi közelségből eredő egymásrahatást),
- a *regionalizálódást* (a nagytérségi tagoltságot) és
- a földrajzi *centrum-periféria* kettősséget

jelöljük meg sajátos tagozódási tényezőkként.

Ahhoz, hogy számszerű eredményekhez juthassunk, e komponenseket is számszerűsítenünk kell. A lokalitás-tényezőt mérhetjük például a legközelebbi szomszédok adott jellemzőjének hasonlóságával, együttmozgásával, a regionalizmus jelenlétét a földrajzi koordináták segítségével közelíthetjük, a centrum-periféria tagoltságot pedig egy kiválasztott központtól mért távolsággal azonosíthatjuk. Az így számszerűsített térkategóriákat – természetesen más lehetőségek is elképzelhetők, ezek feltárása, azonosítása maga is kutatási feladat lehet – beépítve egy regressziós modellbe, mérhetővé válik a "térhatás".

Ha valamely társadalmi paramétert (P) n pontban (például városban) mérünk, felállítható az alábbi regressziós egyenlet (maga a számítás bármely szokványos számítógépes statisztikai programcsomagban megtalálható regressziószámítási eljárással elvégezhető):

$$P_i = a_o + a_x X_i + a_y Y_i + a_d D_i + a_n N_i + E_i$$

ahol a függő változó: P_i = a vizsgált változó az i-edik pontban

a magyarázó változók:

X_i, Y_i = az i-edik pont földrajzi koordinátái

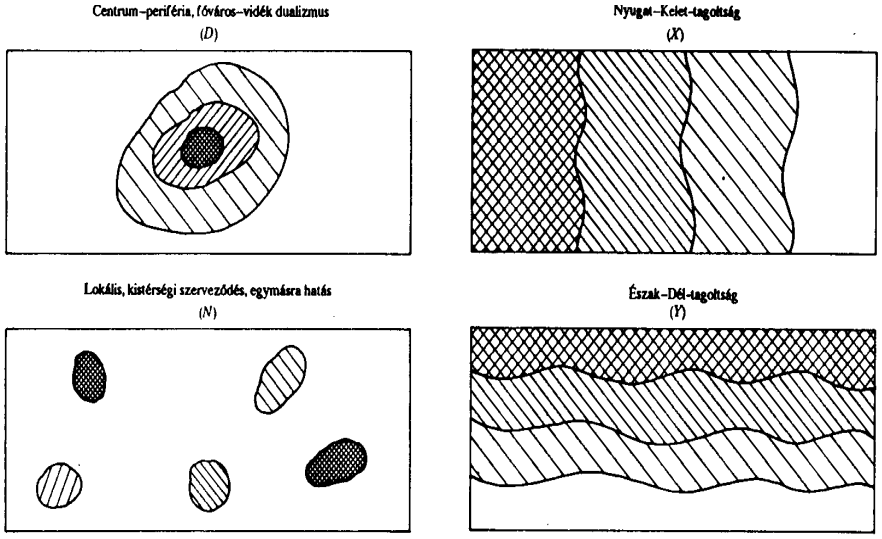
D_i = az i-edik pont távolsága egy adott központtól

N_i = a P_i változó értéke az i-edikhez legközelebbi pontban

E_i = a regresszió hibatagja

a_o, a_x, a_y, a_d, a_n = a regressziós konstansok,

Ha az eljárást Magyarországra konkretizáljuk, akkora a pontok az ország városainak felelhetnek meg, a földrajzi koordináták a városok középpontjának térképi koordinátái lehetnek, legközelebbi szomszédon nyilvánvalóan a legközelebbi várost értjük. A centrum-periféria változóként (D_i) logikus a Budapesttől mért távolságot választani. (A fővároson belüli vizsgálatban a 0 km-kőtől mért távolság lehet ez az érték.) Világosabb és érthetőbb a regressziós egyenlet tartalma, ha vizuálisan is megjelenítjük. A regressziós modellben használt négy sajátos magyarázó változó a *4.12. ábra* szerinti térszerveződési struktúrák jelenlétének vagy hiányának kimutatására alkalmas. Ezek a térmérettől, kiterjedéstől függetlenül felfedezhetők a társadalmi térben, az eljárás alkalmazható egy ország egy térség vagy akár egy város társadalomszerkezeti elemzésében is. Alkalmazási feltétel, hogy a vizsgálatban szereplő mintavételi pontok egyenletesen vagy véletlenszerűen fedjék le a vizsgált területet, ne legyenek üres, illetve túlzsúfolt részek, mert ez a vizsgálat megbízhatóságát rontja – e feltételnek a hazai városhálózat jól megfelel.



4.12.ábra *A regressziós modell magyarázó változói által leírt térszerveződési sémák (Nemes Nagy J. 1993*)

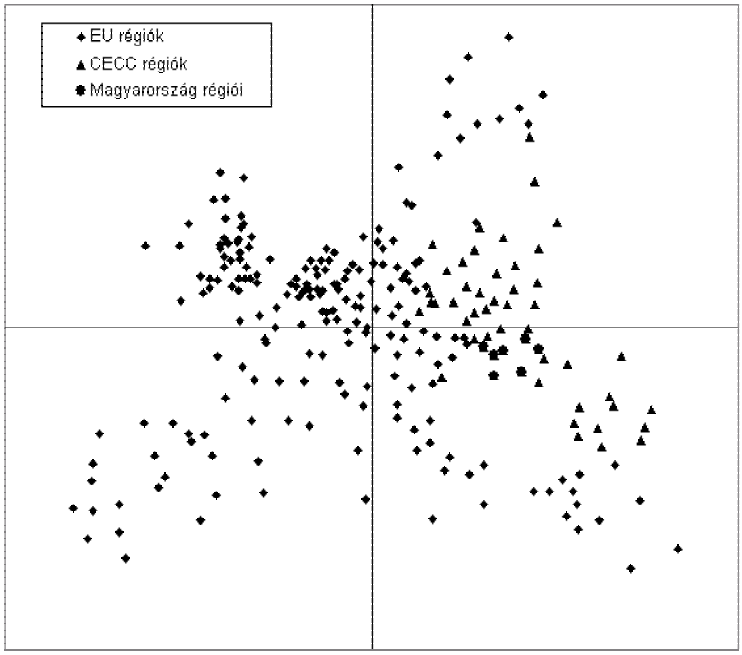
4.5.3 A térbeliség magyarázóereje Európa tagoltságában (Szabó Pál)

A térbeliség magyarázóerejének szerteágazó kutatásából itt csak egy szeletre térünk ki, a földrajzi helyzet kapcsolatára néhány, ún. fő indikátorral (népsűrűség, aktivitási ráta, munkanélküliségi ráta, foglalkozási szerkezet – ágazatokban dolgozók aránya, egy főre jutó GDP).

Térbeli keretként egy „szűkített" Európát választottunk: a tizenöt EU tagállamot és a vizsgálat évében (2000) még csatlakozásra váró, tíz volt szocialista országot (CECC: Bulgária, Csehország, Észtország, Lengyelország, Lettország, Litvánia, Magyarország, Románia, Szlovákia, Szlovénia). A vizsgálatból elhagytuk Ciprust, Máltát, valamint a francia tengerentúli területeket, a Kanári-, az Azori-szigeteket és Madeirát. Regionális szintnek a NUTS 2-et jelöltük meg, így összesen 257 területi egységre tagolt a vizsgálati tér, az EU 204 egysége és a CECC 53 területi egysége. E regionális szint heterogenitását mutatja, hogy a területnagyság szempontjából 119%, a népességszám mentén 80%-os a relatív szórás.

A földrajzi helyzet különböző térparaméterei közül *a fekvést (ÉD, NyK) a határmenti fekvést, a gazdasági centrumtól való távolságot (TÁV)* és *a szomszédságot (SZOM* – a vizsgált indikátorok értéke a szomszédos régiókban*)* választottuk.

A régiókat pontokkal azonosítottuk, úgy, hogy a régió központját (székhelyét vagy ennek hiányában a legnépesebb városát), és annak földrajzi koordinátáit vettük. Ha ezt koordinátarendszerbe rendezzük, akkor a tengelyek irányának észak-délt és kelet-nyugatot, az origónak pedig a pontok koordinátáinak számtani átlagát (vagy a jobb alsó sarkot) választhatjuk (*4.13. ábra*).

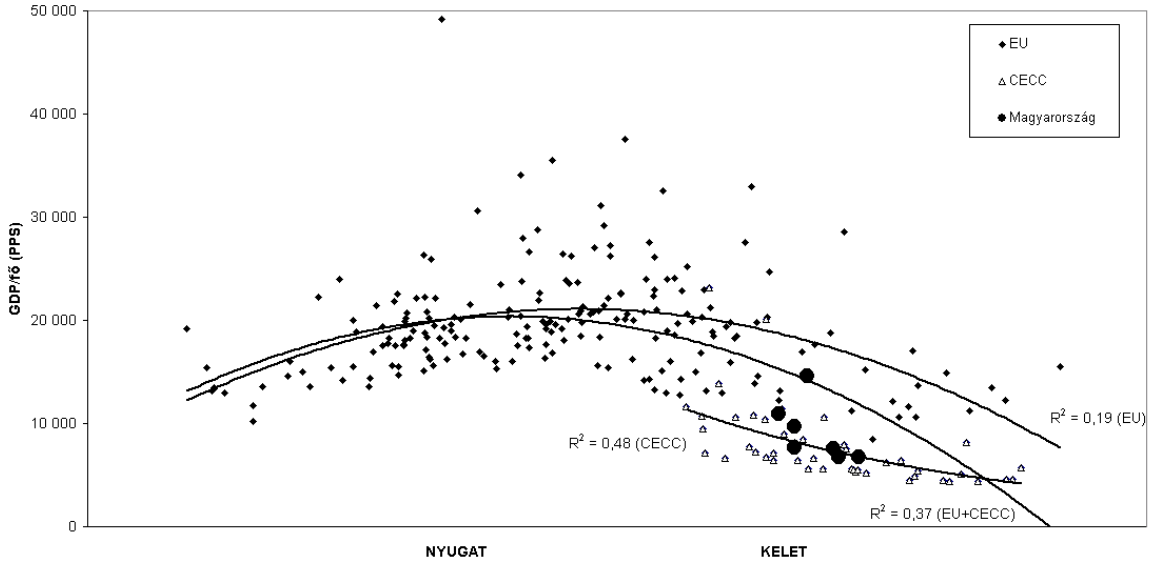


4.13. ábra *„Európa" regionális ponttérképe*

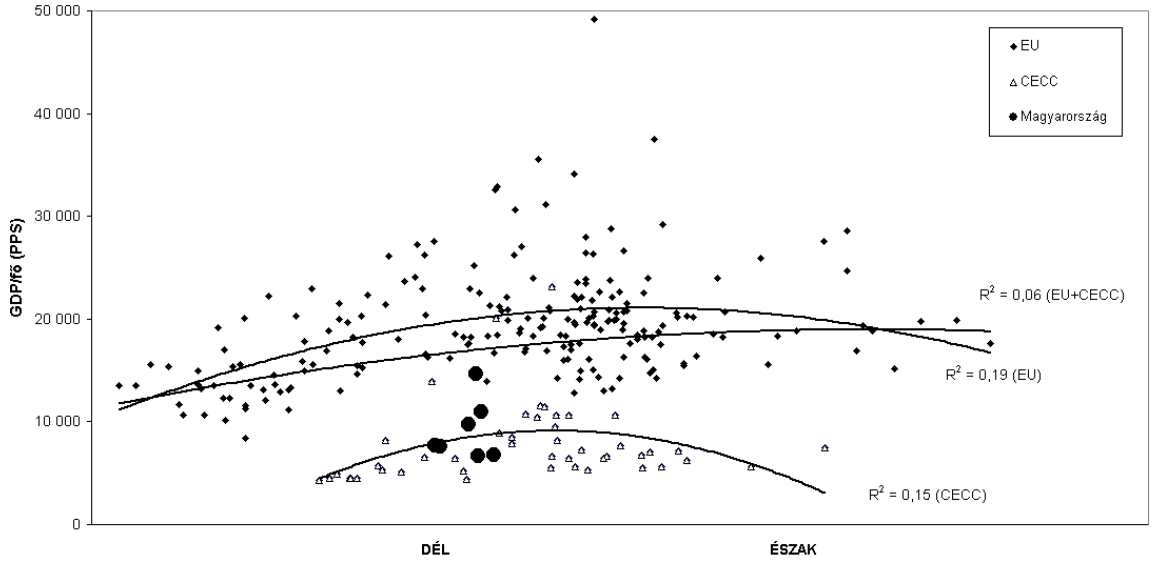
A határmenti fekvés – térképek alapján – könnyen azonosítható. Itt a számszerűsítésre dummy-változót választottnk (tengerparti fekvés=1, nem tengerparti fekvés=0). Gazdasági centrumnak – pontként való értelmezhetősége miatt – a GDP súlypontját választottuk. A gazdasági súlypont Európa esetében valós tartalommal rendelkezik, mivel a fejlett magterület – a Délkelet-Angliától a Benelux államokon, a Rajna-vidéki urbanizációs tengelyén és Svájcon keresztül Észak-Olaszországig húzódó övezet – Európa központi részén fekszik, és a súlypont ennek „közepére" esik. Külön paraméterként teszteltük az Uniós tagságot (ami politikai értelemben „Nyugat-Kelet", de a térben is nagyjából válaszvonalként jelenik meg, dummy változóként: tagság =1, csatlakozási szándék = 0).

Az összetett térszerveződés jellemzőit egy többváltozós lineáris regresszióanalízis segítségével vizsgáltuk, ahol független változóként térparamétereket választottunk (észak-dél, kelet-nyugat, súlyponttól való távolság, szomszédság), függő változóink az említett indikátorok voltak. Az analízis során backward eliminációt alkalmaztunk, így a magyarázóerővel nem vagy csekély mértékben rendelkező térparaméterek kiestek (*4.7. táblázat*).

A hat mutató közül négynél a szomszédsági hatás az egyedüli meghatározó térszerkezeti jegy, kettőnél viszont összetettebb a kép: a szomszédság kiemelkedése mellett más térfaktorok is szignifikánsak. Habár e vizsgálatokból kevésbé domborodott ki, ismert és elismert a centrum-periféria modell léte az Európai Unióban. Több esetben a centrumtól való távolságra, a perifériális helyzetre vezetik vissza például az elmaradottságot, azaz a magterülettől való távolság, mint magyarázótényező szerepel. Azonban ez a térbeli reláció jelentősen megváltozik a bővitéssel, mivel a kelet-közép-európai országok régiói az integráció keleti peremén fekszenek, és döntő többségük fejlettségben jelentősen elmarad az uniós átlagtól. *Új földrajzi dimenzió, a nyugat-kelet ellentét válik az Unió fejlettségi képének egyik fő térbeli jegyévé*. A NyK pozíció magyarázó erejének növekedését jól jelzi az, hogy a determinációs együttható megkétszereződik, az ÉD megosztottság ellenben visszaszorul, a determinációs együttható alig harmadára esik vissza (*4.15.* és *4.16. ábra*), ha csak a „régi" vagy csak az „új" tagokra vagy a két csoportra együtt végezzük el a számításokat (további részletek: *Szabó P. 2003)*.



4.15. ábra *Regionális fejlettség a Nyugat-Kelet pozíció függvényében Európában*



4.16. ábra *Regionális fejlettség az Észak-Dél pozíció függvényében Európában*

	Aktivitási ráta (%)	Munkanélküliségi ráta (%)	Mezőgazdaság (%)	Ipar (%)	Szolgáltatás (%)	GDP/fő (PPS)
R^2	0,79	0,64	0,62	0,52	0,43	0,36
Béta (%)	SZOM (89)	SZOM (80)	SZOM (40) ÉD (-22) TÁV (21) NyK (18)	SZOM (72)	SZOM (65)	SZOM (37) TÁV (-20) ÉD (15)

4.7. táblázat *A regresszióelemzés főbb eredményei*