TÖBBVÁLTOZÓS ADATELEMZÉS

Jegyzetek és példatárak a matematika egyetemi oktatásához sorozat

Algoritmuselmélet
Algoritmusok bonyolultsága
Analitikus módszerek a pénzügyben és a közgazdaságtanban
Analízis feladatgyűjtemény I
Analízis feladatgyűjtemény II
Bevezetés az analízisbe
Complexity of Algorithms
Differential Geometry
Diszkrét matematikai feladatok
Diszkrét optimalizálás
Geometria
Igazságos elosztások
Introductory Course in Analysis
Mathematical Analysis – Exercises I
Mathematical Analysis – Problems and Exercises II
Mértékelmélet és dinamikus programozás
Numerikus funkcionálanalízis
Operációkutatás
Operációkutatási példatár
Parciális differenciálegyenletek
Példatár az analízishez
Pénzügyi matematika
Szimmetrikus struktúrák
Többváltozós adatelemzés
Variációszámítás és optimális irányítás

Kovács Erzsébet

TÖBBVÁLTOZÓS ADATELEMZÉS

Budapesti Corvinus Egyetem

Typotex

2014

Lektorálta: Ágoston Andrea

ISBN 978 963 279 243 9
Készült a Typotex Kiadó (`http://www.typotex.hu`) gondozásában
Felelős vezető: Votisky Zsuzsa
Műszaki szerkesztő: Hajabács Enikő

Készült a TÁMOP-4.1.2-08/2/A/KMR-2009-0045 számú,
„Jegyzetek és példatárak a matematika egyetemi oktatásához” című projekt keretében.

KULCSSZAVAK: Adatelemzés, többváltozós matematikai statisztika, társadalmi és gazdasági adatok elemzése, SPSS alkalmazások, elemi statisztikák, statisztikai táblák, kereszttábla, tanuló algoritmusok, klaszterelemzés, regressziószámítás, logisztikus regresszió, főkomponens elemzés, faktoranalízis, diszkriminanciaanalízis, többdimenziós skálázás, sajátérték-sajátvektor feladatok megoldása.

ÖSSZEFOGLALÁS: A közgazdasági képzésben a Többváltozós adatelemzés és a Többváltozós statisztikai modellezés c. tárgyak hallgatóinak készült jegyzet az elemzési módszerek matematikai háttérének és az alkalmazás előfeltételeinek bemutatása után az SPSS-ben elvégezhető elemzés technikáját és a mintapéldák eredményeinek értelmezését tárgyalja. Az alapok ismertetése során kitérünk az adatok „előkészítésére” is. Valós gazdasági, pénzügyi és demográfiai adatok elemzése mellett egyszerű számpéldákkal is illusztráljuk az elemzési munka buktatóit. Az elemi statisztikai módszereket követően ismertetjük a statisztikai táblázás lehetőségeit, majd sorba vesszük a pénzügyi területen használt legfontosabb többváltozós adatelemző módszereket: a klaszterezést, a lineáris és logisztikus regresszió elemzést, a diszkriminancia-analízist, a faktorok keresését és a többdimenziós skálázást lehetőségeit.

A tananyaghoz kapcsolódó adattáblák letölthetők innen:
`https://www.typotex.hu/index.php?page=ELTE%20TTK`

Tartalom

Bevezetés

A jegyzet a Többváltozós adatelemzés és a Többváltozós statisztikai modellezés című tárgyak hallgatói számára készült, és a féléves kurzus során tárgyalt főbb módszereket ismerteti.

Adatokkal minden szakember találkozik, és az adatokból kinyerhető információ értéke felbecsülhetetlen. A személyi számítógépek elterjedésével népszerűvé váltak a többváltozós statisztikai módszerek, közülük is elsősorban a feltáró elemzések. A statisztikai szoftverek könnyen és gyorsan végzik el a kért elemzést, a megfelelő adatok kiválasztása, a korrekt alkalmazás, valamint az eredmények értelmezése, a következtetések levonása időt és odafigyelést igényel. Nem haszontalan Winston Churchill egy mondását idézni:

„The only statistics you can trust are those you falsified yourself."

A jegyzet nyolc fejezete hármas tagolású:

- a matematikai háttér bemutatása, az alkalmazás előfeltételei,
- az SPSS-ben elvégezhető elemzés technikája és
- a mintapélda eredményeinek értelmezése követik egymást.

A matematikai alapok ismertetése során kitérünk az adatok „előkészítésére" is. Az SPSS 20.0 változatán alapul az elemzési lehetőségek bemutatása, és a futtatás beállítása mellett egy-egy mintapélda eredménytábláit is megadjuk. A jegyzetben valós gazdasági, pénzügyi és demográfiai adatok elemzése mellett egyszerű számpéldák is szerepelnek, amelyek az elemzési buktatókra hívják fel a figyelmet. Az elemzési láncok lehetősége, a módszerek kombinált alkalmazása terjedelmi okokból nem került be az írott anyagba.

Az előző félévekben sok hallgatóval dolgoztam együtt a tárgyak keretében. Érdeklődésük, összegyűjtött adataik és elemzéseik sokat segítettek abban, hogy elkészüljön a jegyzet. Név szerint is köszönöm Ágoston Kolosnak, Csicsman Józsefnek és Kovács Eszternek, hogy figyelmesen elolvasták, javító ötleteikkel gazdagították az anyagot. Minden, a szövegben maradt esetleges hiba és pontatlanság arra vár, hogy a kurzus hallgatói jelezzék nekem!

A lektor munkáját és a TÁMOP által nyújtott támogatást külön is köszönöm.

Budapest, 2013. szeptember

Kovács Erzsébet

1. Leíró és feltáró adatelemzés

A többváltozós adatelemzés alapja az „adat”, ami a számítógépes elemzés érdekében mátrixba rendezett. Szokásos elrendezése szerint soraiban találjuk a megfigyeléseket, és az oszlopok tartalmazzák a megfigyeléseken mért változókat. Ezért a többváltozós adatelemzés módszerei közötti választás előtt célszerű az adattábla tartalmát, kitöltöttségét áttekinteni.

Kezdő lépésként a bevont változókat egyenként vizsgáljuk meg. Szükség lehet a mérési skálák beállítására, sőt néha a skálák transzformációjára, az eloszlásokra vonatkozó előfeltevések ellenőrzésére.

A változók jellemzőinek feltárása mellett a megfigyelt értékekre is fordítsunk figyelmet. A hiányzó adatok pótlása, a kilógó egyedek feltárása, esetleg kiszűrése is az elemzés előkészítő szakaszában történik. A megfigyelt értékek csoportokra bontása, valamely kategória szerinti alminták vizsgálata is ebben a szakaszban végezhető el. Az alapos, körültekintő leíró és feltáró elemzéssel a többváltozós adatelemző munkánk sikerét alapozzuk meg.

1.1. A változók mérési skálája

Az adatok szerzése, gyűjtése több módon történhet, ezért nem mindig mi határozzuk meg a változók mérési skáláját. De az elemzések megkezdése előtt át kell tekinteni, hogy melyik változó milyen skálán van mérve, hiszen statisztikai mutatószámokat is a mérési szint szerint kell választani.

Elméleti megfontolások alapján négyféle mérési szintet[1] különböztetünk meg, amelyeket az egyszerűbbtől a bonyolultabbak felé haladva ismertetünk. Kvalitatív (minőségi) skálának nevezzük összefoglalóan a nominális és az ordinális skálákat. Kvantitatív (mennyiségi) skála az intervallum és az arányskála.

- **Nominális skálán** mérünk, ha csak megkülönböztetést jeleznek a számok vagy a betűk. Ilyenkor általában nem is egyértelmű, hogy egy-egy kategóriát mivel jelölünk. A nominális skálán belül megkülönböztetünk kétértékű (dichotom) és több kategóriából álló változókat.
    - A férfi-nő megkülönböztetésre a 0-1, az 1-2, de az F-N is teljesen megfelel.
    - Ugyanígy például a budapesti kerületeket is azonosíthatjuk arab vagy római számokkal is. Ilyenkor az egymás utáni számok nem adnak információt arról, hogy melyik kerület jobb vagy rosszabb, sőt a szomszédos számok sem jelentenek hasonlóságot.

[1] További példák találhatók itt: http://en.wikipedia.org/wiki/Level_of_measurement

- Az irányítószámok, a telefonszámok, rendszámok stb. mind nominális szinten mért adatok.

- **Ordinális skálán** mért adat már preferenciát is jelez. Két megfigyelés esetén az egyenlő, (leg)nagyobb vagy (leg)kisebb információt is látjuk a változókhoz rendelt számokból. A számok közötti különbség azonban nem értelmezhető. Itt is használhatunk kétértékű (dichotom) és több kategóriából álló változókat. Kétértékű ordinális változó mutatja pl. a megfelelt-nem felelt meg, az igaz-hamis, egészséges-beteg kategóriákat. Több kategóriára számos példa adható.
  - Az életkorokat gyakran ötéves korcsoportokban használjuk, ha a tényleges kor ismerete nem ad több információt, vagy túl kevés megfigyelésünk van egyedi adatok elemzéséhez.
  - A településeket megadhatjuk úgy, hogy 1=500 fő alatti falu, 2=500-1000 fő közötti falu, 3=1000-2000 közötti település, és így tovább. A lakónépesség létszáma szerinti kategóriákat használjuk a tényleges létszám megadása/ismerete nélkül.
  - A jövedelemsávok, a gépjárművek teljesítmény kategoriák is ordinális adatot jelentenek, hiszen a számok között aritmetikai művelet nem értelmezhető.
  - Betűkkel megadott ordinális skálát is ismerünk, pl. külföldi egyetemeken A-F között osztályoznak, vagy az országkockázatra, tőzsdei cégek minősítésére is gondolhatunk.
  - A kérdőíves vizsgálatokban leggyakrabban páratlan (5,7,..) fokú ordinális skálán lehet a válaszokat megadni. Ilyenkor a számok mellett szövegesen is szerepel a válasz: 1: teljesen nem ért egyet, 2: nem ért egyet, 3: nincs véleménye, 4: egyetért, 5: teljesen egyetért.

- **Intervallum skálán** mért adatok között már eltérést is számolunk és értelmezünk. Az intervallum hossza a két megfigyelés közötti eltérést tükrözi.
  - Ha az időjárást Celsiusban mérjük, akkor az átlaghőmérséklet változását jellemezni tudjuk.
  - A fizetések vagy a hitelösszegek ismeretében az átlagos értékek és az átlagtól való eltérések kiszámítása mellett akár a két változó közötti kapcsolatot is jellemezni tudjuk.
  - Az egyetemi vizsgadolgozatok pontozása is intervallum szintű adatot jelent. Ebből kategória határokat kijelölve ordinális szinten mért osztályzatot képezünk.
  - Több minősítő cég 0-100 közötti pontszámmal, azaz intervallum skálán értékeli az országkockázatot.

- **Az arányskála** speciális intervallumskála, amelyen mért adatok között kitüntetett nulla pont is van, és két megfigyelés aránya is értelmezhető, nemcsak a különbségük.
    - A testmagasság és a testsúly egyaránt arányskálán mért változók.
    - Az életkor is arányskálán mérhető, hiszen a születés pillanatához nulla életév tartozik.
    - A Kelvin fokban mért hőmérsékletnek is van abszolút nulla foka, ez a -273.15° Celsius.
    - Napokban, hónapokban, években mért tartamokat (befektetés, hitel, életbiztosítás jellemzésére) is arányskálán mérünk.

Ha csak egy-egy változót elemzünk, akkor is fontos a mérési szint pontos ismerete. A mérési szintnek megfelelő leíró statisztikai mutatók kiválasztásához az 1.2. alfejezet ad útmutatást.

A többváltozós elemzések többségükben azonos mérési skálát igényelnek. Ennek érdekében gyakran skála-transzformációt hajtunk végre, ami fel- és leértékelés is lehet. Magasabb szintű skálára áttérni csak többlet információ birtokában lehet. A skála leértékelése, a különbségek helyett kategóriák kialakítása sokszor hasznosan tömöríti az információt. A kategória képzés hatékony módját a 2 fejezet ismerteti. A könyv további fejezeteiben bemutatunk majd más skála-transzformációs lehetőségeket is.

1.2. Leíró statisztikák kiválasztása az adatok mérési skálája alapján

Leíró statisztikát készítünk, ha nem állítunk fel és tesztelünk hipotézis(eke)t, csak a változók és a megfigyelések jellemzése a célunk. Leggyakrabban központi értéket vagy szóródási jellemzőt számítunk, az eloszlás alakját mutatjuk be numerikus és/vagy grafikus eszközökkel. Vizsgálhatjuk a teljes adatállományt együtt, vagy részekre tagolva is.

Az SPSS-ben az **Analyze/Descriptive Statistics** menűpont alatt találunk három eljárást, amelyek több mutató:

- A „**Frequencies" funkció** választásával a nominális és ordinális változók kategóriáihoz tartozó gyakoriságok listázása válik lehetővé. Továbbá gyakoriságokat és relatív gyakoriságokat is megadó ábrákat is készíthetünk itt. Emellett tetszőleges skálán mért adatokat is elemezhetünk, mert minden statisztikai mutatót felajánl ez a menüpont is választási lehetőségként.
- A **„Descriptive" funkció** az intervallum vagy arány skálájú változók leírására, jellemzésére csak numerikus statisztikákat számol. Itt kérhetjük és menthetjük el a változók sztenderdizált értékeit.

- Az **Explore**[2] **funkciót** választjuk, ha almintákat is feltételezünk, vagy egy kategóriaképző – nominális/ordinális – változó szerint tagoljuk a megfigyeléseket, és intervallum vagy arányskálán mért változó(k)ra leíró statisztikát készítünk. A „feltárás" elnevezés arra utal, hogy ez az elemzés megelőzi pl. a két minta átlagának egyezésére vonatkozó hipotézis megfogalmazását, a normalitási teszt elvégzését, stb.

Mindegyik eljárás megengedi, hogy egyszerre több változót válasszunk ki, és ezek mindegyikére elvégzi az összes általunk kért műveletet. Ezért célszerű egyszerre csak azonos mérési szintű változókat felsorolni, így csak a szakmailag korrekt eredményeket állítjuk elő.

Az 1.1. táblázatban összefoglaljuk azt, hogy melyik SPSS menűpontban találhatók meg a leíró statisztika eszközei a mérési skálák szerinti bontásban. A magasabb szintű mérési skálákon az előző skálákhoz rendelt eljárások mindig alkalmazhatók. **D** jelöli a Descriptive, **F** a Frequency és **E** az Explore funkciót.

1.1. táblázat: Elemzési célokat megvalósító funkciók

Cél / Skála	Nominális	Ordinális	Intervallum/arány
Központi tendencia	Módusz **F, E**	Módusz **F,E** Medián **F, E** Minimum, Maximum **F,D,E**	Átlag **F,D,E**
Szóródás	Gyakoriság, relatív gyakoriság **F**	Terjedelem **F,D,E** Interkvartilis terjedelem **E**	Szórás, variancia, sztenderd hiba **F,D,E**
Eloszlás - numerikus	-	-	Ferdeség, csúcsosság **F,D,E** Normalitási teszt **E**
Eloszlás - grafikus	Gyakoriságra oszlop- és kördiagram **F**	Stem&leaf **E**	Hisztogram **F, E** boxplot **E**

A legfontosabb leíró statisztikai mutatókat röviden áttekintjük, és a képleteket is megadjuk.

[2] Az Explore nemcsak alminták összehasonlítására alkalmas. Egyetlen homogén minta esetében a Descriptive-vel azonos eredményeket ad, továbbá nyesett átlagot is számol.

- **Mean**: számtani átlag, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, ahol n a megfigyelések száma (1.1)

  Az elméleti várható érték (m) általában nem ismert. Értékét az (1.1) szerint számított mintabeli átlaggal ($\bar{x}$) helyettesítjük.

- **Range**: terjedelem= maximum-minimum

- **Variance**: szórásnégyzet, a sokaságban: σ^2 , ennek mintabeli becslése s^2 és gyöke a szórás, s. A szórás angol neve standard deviation, röviden: Std. dev.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{1.2}$$

- **Std.Error**: az átlag sztenderd hibája: $\frac{\sigma}{\sqrt{n}}$ vagy becslése $\frac{s}{\sqrt{n}}$ (1.3)

- **Skewness**: ferdeségi mérték, képlete: $\gamma_1 = \frac{\frac{1}{n}\sum (x_i - m)^3}{\sigma^3}$

  A ferdeség negatív értéke balra hosszan elnyúló eloszlást, a pozitív értéke pedig jobbra elnyúló eloszlást jelez. Ha nulla közeli a mutató, akkor szimmetrikus az eloszlás. (De itt ne csak a normális eloszlásra gondoljunk, mert az U alakú eloszlás is szimmetrikus.)

  A ferdeség varianciája = $\frac{6n(n-1)}{(n-2)(n+1)(n+3)}$. E variancia gyöke: $SE(\gamma_1)$ szerepel „standard error" elnevezéssel az eredményeket bemutató 1.2. táblában.

  A ferdeség torzítatlan becslése $\hat{\gamma}_1 = \frac{n\sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$ (1.4)

  A nullhipotézis szerint a ferdeség=0. A ferdeségi mutató és a sztenderd hiba hányadosát hasonlítjuk az (n-1) szabadsági fokú Student eloszlás kritikus értékéhez.

  A ferdeséghez tartozó t-teszt képlete: $t = \gamma_1 \,/\, SE(\gamma_1)$ (1.5)

- **Kurtosis**: csúcsosság, mérőszáma: $\gamma_2 = \frac{\frac{1}{n}\sum (x_i - m)^4}{\sigma^4}$, értéke sztenderd normális eloszlás esetében = 3. Ezt levonva közvetlenül *(γ_2 - 3)* alakban kapjuk a mutatót az SPSS-ben. Más gépi programok ezt „kurtosis excess" néven adják meg.

  A csúcsosság varianciája = $\frac{4(n^2-1)\left[SE(\gamma_1)\right]^2}{(n-3)(n+5)}$. E variancia gyöke szerepel „standard error" elnevezéssel az 1.2. táblázatban.

  A csúcsossági mutató torzítatlan becslése:

$$\hat{\gamma}_2 = \frac{n(n+1)\sum (x_i - \bar{x})^4 - 3(n-1)\left[\sum (x_i - \bar{x})^2\right]^2}{(n-1)(n-2)(n-3)s^4} \qquad (1.6)$$

  A csúcsossági mutató és a sztenderd hiba $(SE(\gamma_2))$ hányadosát hasonlítjuk az (n-1) szabadsági fokú Student eloszlás kritikus értékéhez. A csúcsossági mutatóhoz tartozó t-próba képlete: $t = \gamma_2 \,/\, SE(\gamma_2)$ (1.7)

A pozitív csúcsosság a normális eloszlás sűrűségfüggvényénél hosszabb, vastagabb farok részt, a központi érték körüli tömörülést vagy mindkettőt jelezheti. A negatív érték lapult eloszlásra utal, amelynek a haranggörbénél rövidebb, vékonyabb farok része van, és középen sem sűrűsödnek a megfigyelések.

A lapultság minimális értéke –2, mert a ferdeség és a csúcsosság mértéke között fennáll a következő egyenlőtlenség: csúcsosság ≥ (ferdeség^2 – 2)

A ferdeség csak az egyik oldalon, a csúcsosság a mindkét oldalon előforduló extrém értékek előfordulását jelezheti. Az **extrém, outlier megfigyelések** nagy hatással lehetnek az átlagra és a szórásra, ezért érdemes grafikusan (például hisztogramon) is megnézni a változók alakját.

- A **mintaátlag ferdesége:** $\gamma_1 / \sqrt{n}$ és **csúcsossága:** γ_2 / n. A mintanagyság növelésével csökken a ferdeség, és még gyorsabban csökken a csúcsosság.

Van néhány egyszerű, de hasznos nagyságrendi összefüggés a leíró statisztikák között, amire itt felhívjuk a figyelmet.

- Szimmetrikus eloszlás esetén az átlag=medián=módusz, míg eltérésük ferde eloszlásra utal.
- Pozitív ferdeségű az eloszlás, ha módusz<medián<átlag, és negatív ferdeségű, ha átlag<medián<módusz áll fenn.
- A medián kevésbé érzékeny az adathiányra és a szélső értékekre, mint az átlag.
- A terjedelem közelítőleg a szórás négyszerese.

Az SPSS nem számol **relatív szórást**, amely a szórás és az átlag hányadosa. A Csebisev egyenlőtlenségen alapuló hüvelykujj szabály alapján magas a szórás, ha ez az arány meghaladja a kettőt. Ez arra utal, hogy az adatrendszerben több alminta lehet, ezek feltárását grafikus módszerekkel érdemes elvégezni.

A pénzügyi adatokban általában a szórás a kockázat mértéke, a biztosításban pedig a relatív szórás méri a kockázatot. A relatív szórás alkalmazását indokolja az is, hogy így a különböző mértékegységet kiküszöböljük, tehát pl. különböző valutanemben kifejezett változók szórása is így vethető össze.

Ha egy változónak nagy a szórása, akkor ez a változó mentén megvalósítható nagyobb szeparációs képességet jelzi. Az alacsony szórás az átlag körül koncentrálódó (általában csúcsos eloszlású) megfigyelésekre utal.

A „Descriptive" **a sztenderdizált „z-score" változók elmentését** is lehetővé teszi. A zérus átlagú és egységnyi szórású új változó ferdesége és csúcsossága nem változik meg.

$$z_x = \frac{x - \bar{x}}{s} \tag{1.8}$$

Normális eloszlás (és/vagy nagy minta) esetén a központi határeloszlás tétel alapján a sztenderdizált változó $z_{\bar{x}} = \frac{\bar{x} - m}{s / \sqrt{n}}$ standard normális eloszlású lesz, kis mintára pedig (n-1) szabadságfokú Student t-eloszlást követ.

Több érv szól a változók sztenderdizálása mellett. A mértékegység kiküszöbölése, az ismert átlag és szórás különösen akkor hasznos, ha többváltozós elemzést végzünk, azaz egyszerre több változót használunk.

A fejezet végén óvjuk az olvasót attól, hogy bármely programcsomagot mechanikusan alkalmazzon. A szórás mintából történő becslésekor az SPSS-ben (n-1) szerepel a nevezőben, akár kicsi a minta, akár nagy. A csúcsossági mutatóból – előzetes figyelmeztetés nélkül – levonja az SPSS a sztenderd normális eloszlásra jellemző hármat. Az R-ben pedig a >range(x) menüpont nem a terjedelmet adja meg, hanem a minimum és a maximum értékeket írja ki egymás mellé.

1.3. Leíró statisztikák kiszámítása és értelmezése

A számítási eredményeket a megismételhetőség érdekében az SPSS mintapéldák között található World95.sav adathalmazon mutatjuk be, amely 109 ország adatait tartalmazza. Az első lépésben a férfiak és nők várható élettartamára készültek számítások. Ezek az információk a befektetési döntések, pl. az életjáradék és különösen a nyugdíj számításához fontosak. Bár nem szerepel az adat nevében, ezek a születéskor várható élettartamok, és a két nemre számolt átlagok között a világ

minden országában eltérés van. Az 1.2. táblázatban a **Frequency**-ben készített részeredmények láthatók.

Hiányzó adat nincs erre a két változóra, a medián természetesen megegyezik az 50%-os percentilissel, és figyelmeztetést kapunk, hogy több móduszú a nők várható élettartamát mérő változó. A negatív ferdeség a hisztogramon (1.1. ábra) is látható, tehát a magasabb várható élettartam értékek a gyakoribbak. Az (1.4) szerinti ferdeségre számolt (1.5)-beli t-teszt értéke -5 körüli, azaz minden szokásos szignifikancia szint mellett elvethető, hogy szimmetrikus az eloszlás, hisz értéke nem nulla. A csúcsosság/lapultság értéke nem tér el szignifikánsan a zérustól, mindkét nemre a t-teszt kisebb, mint egy. Nem koncentrálódnak tehát túlzottan a várható élettartamok az átlag körül. Az élettartamok összege (Sum) nem hordoz lényegi információt.

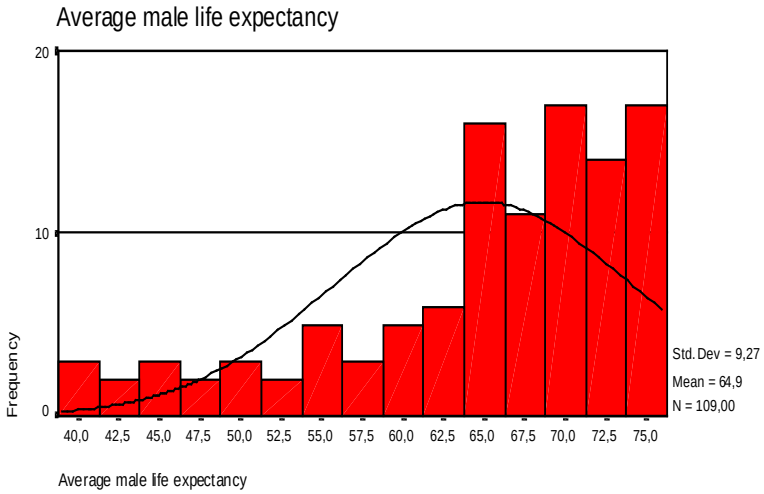
A percentilisek és a kvartilisek alapján megállapítható az élettartam eloszlások több jellemzője. Érdekes az, hogy a legalacsonyabb életkilátású 10 százaléknyi népességnél 2 évnyi élettartam eltérést kaptunk, míg a legfelső 10 %-ban már 6 év a nők javára a különbség.

1.2. táblázat: Frequency-ben előállított eredmények

Statistics

		Average female life expectancy	Average male life expectancy
N	Valid	109	109
	Missing	0	0
Mean		70,16	64,92
Std. Error of Mean		1,01	,89
Median		74,00	67,00
Mode		75[a]	73
Std. Deviation		10,57	9,27
Variance		111,76	85,98
Skewness		-1,109	-1,080
Std. Error of Skewness		,231	,231
Kurtosis		,213	,336
Std. Error of Kurtosis		,459	,459
Range		39	35
Minimum		43	41
Maximum		82	76
Sum		7647	7076
Percentiles	10	52,00	50,00
	20	59,00	57,00
	25	66,50	61,00
	30	68,00	63,00
	40	70,00	65,00
	50	74,00	67,00
	60	76,00	69,00
	70	78,00	71,00
	75	78,00	72,50
	80	79,00	73,00
	90	80,00	74,00

a. Multiple modes exist. The smallest value is shown



1.1. ábra: Hisztogram és a normális eloszlás sűrűségfüggvénye

Az 1.3. táblázatban a **Descriptive**-ben előállított valamennyi részeredményt bemutatjuk. Értékeik természetesen megegyeznek azokkal, amiket a Frequency-ben kaptunk, csak elrendezésük más. Itt is több változó kérhető egyszerre, de statisztikai összehasonlítást most sem végzünk.

Azt a szembetűnő különbséget, ami a férfiak és a nők várható élettartama között látható, a konfidencia intervallumok összevetésével vagy t-próbával lehet tesztelni.

1.3. táblázat: Leíró statisztikák

Descriptive Statistics

		Average female life expectancy	Average male life expectancy	Valid N (listwise)
N	Statistic	109	109	109
Range	Statistic	39	35	
Minimum	Statistic	43	41	
Maximum	Statistic	82	76	
Sum	Statistic	7647	7076	
Mean	Statistic	70,16	64,92	
	Std. Error	1,01	,89	
Std. Deviation	Statistic	10,57	9,27	
Variance	Statistic	111,762	85,984	
Skewness	Statistic	-1,109	-1,080	
	Std. Error	,231	,231	
Kurtosis	Statistic	,213	,336	
	Std. Error	,459	,459	

Az (1.8) szerinti sztenderdizálás nem csak a mértékegység kiszűrése miatt hasznos, hanem az összehasonlítást is segíti. A pozitív értékek átlag feletti, a negatívok pedig átlag alatti eredeti értéket jeleznek. Ezeket két vagy több változó mentén egyszerre is láthatóvá tudjuk tenni egy pontdiagramon (Scatter plot), ahogy ezt az 1.2. ábra mutatja. Mivel behúztuk az átlagokat jelző koordináta tengelyeket, a négy sík negyedben jól tudjuk jellemezni az országokat. Az első sík negyedben a mindkét változó szerint átlag feletti értékkel rendelkező országokat látjuk. Magyarország és a szomszédos országok a harmadik negyedben helyezkednek el, azaz az egy főre jutó GDP és a népesség növekedése szerint is átlag alatti értékek jellemezték térségünket 1995-ben.

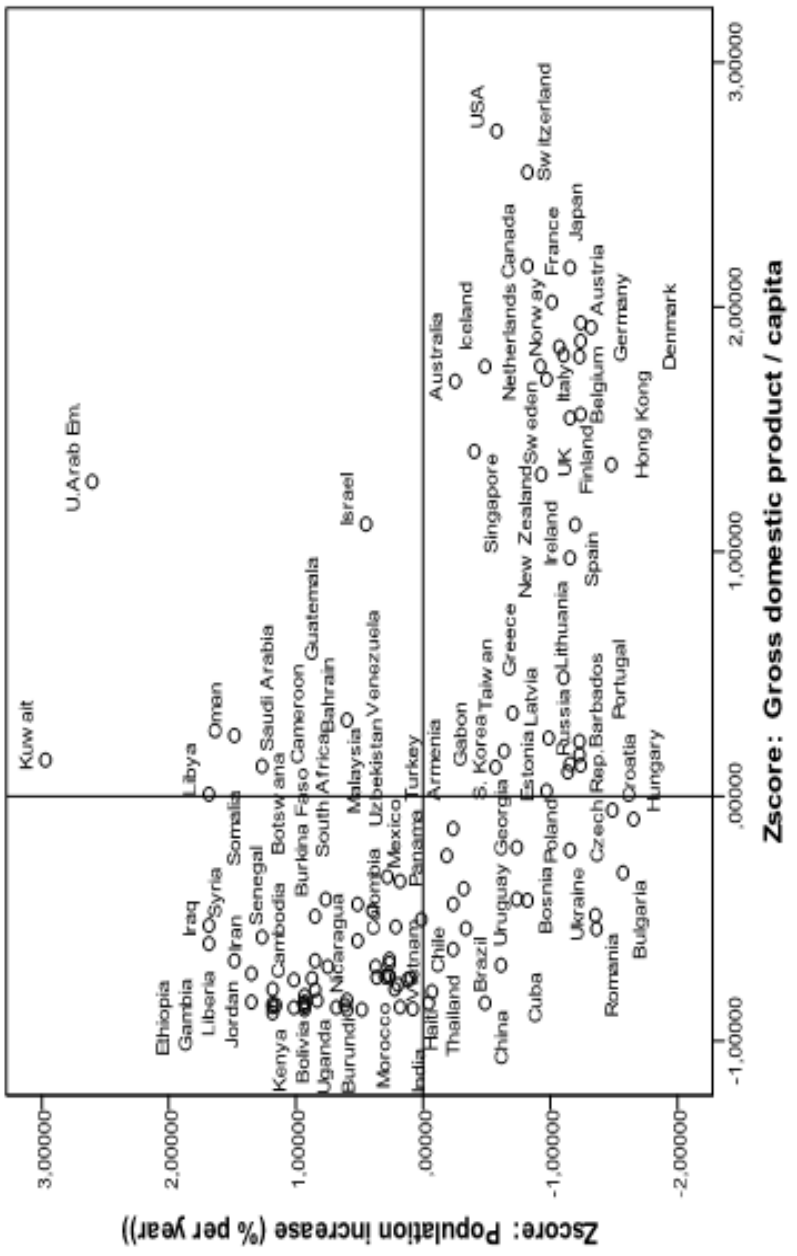
Az is szembetűnő az 1.2. ábrán, hogy negatív előjelű, bár nem teljesen lineáris a két változó kapcsolata, és kevés olyan ország van, ahol mindkét változó az átlag felett van.

Érdemes figyelni arra is, hogy az eredeti adatokban a GDP/fő változó terjedelme és szórása jóval nagyobb, mint a népesség növekedés százalékos adatának terjedelme. A sztenderdizált változók terében a terjedelem éppen fordított nagyságot mutat, miközben mindkét átlag 0 és a szórások egységnyiek, ahogy ez az 1.4. táblázatban látható.

1.4. táblázat: Az eredeti és a sztenderdizált változók jellemzői

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Population increase (% per year))	109	-,3	5,2	1,682	1,1976
Zscore: Population increase (% per year))	109	-1,65535	2,97072	,000	1,000
Gross domestic product / capita	109	122	23474	5859,98	6479,836
Zscore: Gross domestic product / capita	109	-,88551	2,71828	,000	1,000
Valid N (listwise)	109				



1.2.ábra: Országok a sztenderdizált változók terében

Házi feladat: Bizonyítandó

a) Az eredeti és a sztenderdizált változók ferdesége és csúcsossága megegyezik.

b) Normális eloszlású alapsokaság esetében az *s* és a $\sqrt{n}\left(\bar{x}-m\right)$ függetlenek, ezért korrelációjuk zérus.

c) Tetszőleges eloszlás esetén az *s* és a $\sqrt{n}\left(\bar{x}-m\right)$ két tag közötti korreláció= $\gamma_1 \Big/ \sqrt{\gamma_2+2}$, ez a normalitástól való eltérést is jelzi.

1.4. Az extrém pontok és az alminták statisztikai elemzése

Két változó statisztikai jellemzőinek összevetése, az egyedi, extrém értékek azonosítása és az adatállományban levő alminták, kategóriaváltozók (factor) mentén képzett csoportok vizsgálata az Explore menűpontban végezhető el. Az itt előállított (az 1.2. és 1.3. táblázattal megegyező) eredményeket nem mutatjuk be ismét, csak azokat, amiket többletként kapunk.

a) **Konfidencia intervallum** (1-α) megbízhatósági szinten: $\bar{x} \pm t_{\alpha/2,(n-1)} \cdot \frac{s}{\sqrt{n}}$ képlettel számolható. A megbízhatósági intervallum szélességét a sztenderd hiba mellett a t-statisztika is befolyásolja. A megfigyelésszám növekedésével csökken mind a sztenderd hiba, mind a t-érték, tehát nagyobb mintában szűkebb intervallumot kaphatunk.

A nők várható élettartamára az alsó és felső határ: 68,15-72,16 év, a férfiak adataira 63,16-66,68 év adódik. A két intervallum nem fedi át egymást, ezért a megfelelő tesztek elvégzése nélkül[3] is mondhatjuk, hogy jelentős, statisztikailag szignifikáns az eltérés.

b) **Trimmed mean, azaz nyesett átlag:** a nagyság szerint sorba rendezett megfigyelések középső 90 százalékára számított átlag. A rendezett minta két végén 5-5%-ot elhagyunk. Szimmetrikus eloszlás esetén a közönséges és a nyesett átlag megegyezik. Nem normális eloszlás és extrém értékek előfordulása esetén az így számított átlag értelmezése javasolt. A várható élettartam adatokra a férfiak esetében 65,59, a nőknél 70,96 a nyesett átlag. Mindkét eloszlás erősen balra ferde, ezért a nyesett átlag nagyobb, mint a közönséges számtani átlag.

A nyesett átlag számításának két változata van:

[3] Így a tesztelés előfeltételeit sem kell ellenőrizni. A normális eloszlás például a ferdeség miatt nem áll fenn.

- Ha a nyesés során (*0,05n*) egész, akkor ennyi megfigyelést hagyunk el, és a fennmaradó értékek egyszerű összege a nyesett átlag számlálója. A nevezőben pedig (*0,9n*) áll.
- Ha (*0,05n*) nem egész szám, akkor *k* és *(k+1)* egészek közé esik. Az első *k* és az utolsó *k* darab megfigyelést elhagyja a gép, a *(k+1)-edik* elem és az *(n-k)-adik* elem súlya pedig a zárójelben álló két tag minimuma lesz: *min(k+1-0,05n; 0,05n-k)* a számtani átlag számításakor. A köztes megfigyelések súlya egy.

c) A centrumtól távoli megfigyelések súlyozása **M-esztimátorok** alkalmazásával is történhet. (Nem elhagyjuk a távoli értékeket, hanem csökkenő súlyt adunk nekik.) Az M-esztimátorok révén becsült „korrigált átlagok"általában az átlag és a medián közé esnek, nem rangsorolhatók, nem mondható meg, hogy melyik a jobb.

Az esztimátorok képzése a helyzeti közép (*T*) becslése után következik. A helyzeti közepet az alábbi egyenlet megoldásával kapjuk:

$\sum_{i=1}^{k} f_i \Psi(\frac{x_i - T}{s}) = 0$, ahol f_i a gyakoriság, s „szórás" és ψ páratlan függvény.

Az egyenlet másik alakja:

$$\sum_{i=1}^{k} f_i \left(\frac{x_i - T}{s} \right) \omega \left(\frac{x_i - T}{s} \right) = 0, ahol \omega(u) = \frac{\Psi(u)}{u}$$

A gyakoriságokkal szorzunk, hogy *T* kifejezhető legyen:

$$\sum_{i=1}^{k} f_i \left(\frac{x_i}{s} \right) \omega \left(\frac{x_i - T}{s} \right) - T \frac{\sum f_i}{s} \omega \left(\frac{x_i - T}{s} \right) = 0$$

Átrendezve *T* az x adatok súlyozott átlaga:

$$T_{k+1} = \frac{\sum f_i x_i \omega(\frac{x_i - T_k}{s})}{\sum f_i \omega(\frac{x_i - T_k}{s})}$$

Látjuk, hogy *T* csak iterációval adható meg, a T_{k+1} kifejezhető a T_k –ból. T_0-t nem adja meg az SPSS leírása, de ez az érték általában a medián.

Az iteráció leáll, ha

i) $\left| T_{k+1} - T_k \right| \le 0{,}005 \cdot \frac{T_{k+1} + T_k}{2}$ vagy

ii) *k>30*.

A helyzeti középtől való eltérésből reziduálist kapunk. A reziduális számlálója a mediántól való eltérés, míg a nevezője a minta mediánjától való abszolút értékes eltérések mediánja.

$$u_i = \frac{x_i - T}{s} = \frac{x_i - Medián(x)}{Medián|x_i - Medián(x)|}$$

Az ω(u) függvény - mint súly - a reziduális nagyságához kapcsolódik. Az SPSS-ben a súly megválasztására elérhető c1)-c4) eljárás a kidolgozóiról kapta a nevét.

c1) Huber esztimátorában:

$$\omega(u_i) = \begin{cases} 1, & ha |u_i| \le 1{,}339 \\ (1{,}339/u_i)\,\mathrm{sgn}(u_i), & ha |u_i| > 1{,}339 \end{cases}$$

Itt 1,339-től változó előjellel csökkenő, előtte pedig 1 a súly.

c2) Tukey két súlyt használ. A 4,685-nél nagyobb abszolút értékű, sztenderdizált reziduálisra 0 súlyt ad, a kisebbekre pedig a centrumtól való távolsággal fordított arányos a súly.

$$\omega(u_i) = 1 - \left(\frac{u_i}{4{,}685}\right)^2, ha \ \ |u_i| \le 4{,}685, \qquad és \qquad 0 \ \ különben$$

c3) Hampel súlyfüggvénye 4 szakaszból áll:

a) A súly $\omega(u_i)$ = 1, ha az $|u_i| \le 1{,}7$

b) $\omega(u_i) = \frac{1{,}7}{u_i} \cdot \mathrm{sgn}(u_i)$, ha a $1{,}7 < |u_i| \le 3{,}4$

c) $\omega(u_i) = \frac{1{,}7}{u_i} \cdot \frac{8{,}5 - |u_i|}{8{,}5 - 3{,}4} \mathrm{sgn}(u_i)$, ha a $3{,}4 < |u_i| \le 8{,}5$

d) Ha pedig az $|u_i| > 8{,}5$ akkor a súly = 0.

c4) Andrews szinusz függvényt javasolt, ebben nincs törés.

A súly $\omega(u_i) = \frac{1{,}34\pi}{\pi \cdot u_i} \cdot \sin\left(\frac{\pi \cdot u_i}{1{,}34\pi}\right)$, ha $|u_i| \le 1{,}34 * \pi$ (~4,2).

1.5. táblázat: A „korrigált" átlagok számítása

M-Estimators

	Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
Average female life expectancy	73,06	74,51	73,09	74,55
Average male life expectancy	66,85	67,30	66,44	67,33

a. The weighting constant is 1,339.

b. The weighting constant is 4,685.

c. The weighting constants are 1,700, 3,400, and 8,500

d. The weighting constant is 1,340*pi.

A negatív ferdeség miatt mindkét változóra mind a négyféle korrigált átlag meghaladja a számtani átlagot, sőt a nyesett átlagot is. A nők várható élettartamának minden M-esztimátora magasabb a 95%-os konfidencia intervallum felső határánál, míg a férfiakra számolt Hampel-féle érték beleesik a konfidencia intervallumba.

Az élettartambecslés pontossága azért kiemelten fontos, mert a fejlett országokban ez a mutató folyamatosan emelkedik. Két megállapítást tehetünk ebben a szakaszban:

- Érdemes évről évre friss adatokat gyűjtve megismételni a számításokat.
- Célszerű a fejlett és a fejlődő országokat külön csoportban vizsgálni, hogy homogénebb almintáink legyenek.

d) **Interquartile range**: **interkvartilis (belső) terjedelem**, a felső kvartilis (75%) és az alsó kvartilis (25%) közti különbség: $IQR=Q_3 - Q_1$, és ez a doboz diagram (box-plot) dobozának magasságát adja meg.

A várható élettartamokra 1.3. ábrán látható a közös doboz-diagram, eredeti nevén **Box-plot.** A doboz közepén levő vonal a medián, a dobozban a megfigyelések 50%-a található. A doboz alja: az első kvartilis: Q_1 , teteje a felső kvartilis: Q_3.

Felfelé és lefelé addig húzzuk a vonalat, amíg az alábbi kettő közül az első bekövetkezik:

- elérjük a tényleges maximumot vagy minimumot,
- fel/lemérjük az interkvartilis terjedelem 1,5-szeresét.

A fenti tartományon kívül eső megfigyelés outlier (jele: o).

A kilógó (Outlier) pontok tartománya:

alul: $Q_1 - 3IQR$; $Q_1 - 1{,}5IQR$

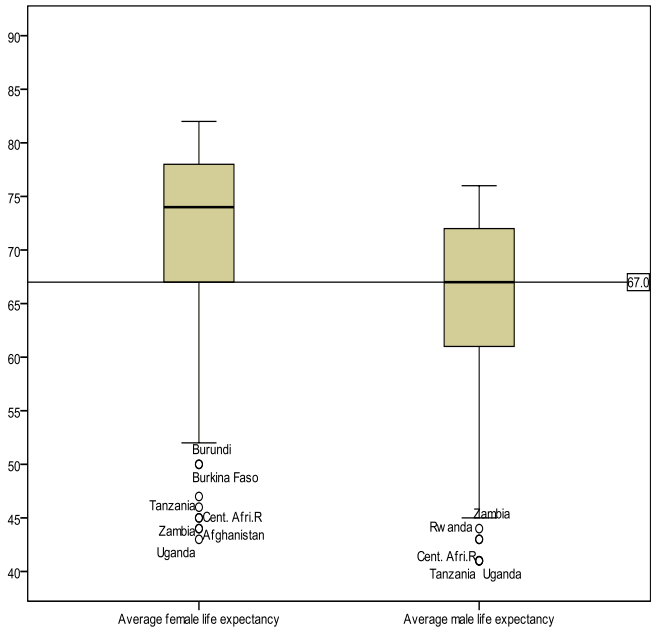
felül: $Q_3 + 1{,}5IQR$; $Q_3 + 3IQR$

A háromszoros interkvartilis terjedelemnél távolabbi megfigyelések az extrém pontok (jelük:*):

alul: $x \leq Q_1 - 3IQR$

felül: $x \geq Q_3 + 3IQR$

Bár az élettartam kvartilisek eltérőek, különbségünk mindkét nemre 12 év, ezért a dobozok magassága azonos. Az eloszlások ferdék, ezért a vonalkák hossza felfelé és lefelé eltérő. Az outlier országok számmal vagy névvel írathatók ki. Itt csak lefelé vannak kilógó – nagyon alacsony várható élettartamú országok – melyeket az országnév-címkék azonosítanak. Az 1.3. ábrába behúztuk a férfi medián életkort (67 év). Szembetűnő, hogy a nők alsó kvartilise is a férfi-medián vonal felett van. Azaz az országok 75%-ában tovább élnek a nők 67 évnél, míg a férfiaknál csak 50% ez az arány.



1.3. ábra: Doboz diagram 2 változóra

e) Az **extrém értékek** listája minden változóra az 5 legnagyobb és az 5 legkisebb megfigyelést sorolja fel akkor is, ha ezek nem valóban kilógó pontok. Az „extrém" listát össze kell vetni a box-plottal vagy a stem&leaf ábrával, hogy a tényleges belső távolságokról meggyőződhessünk.

f) A **Stem&leaf** ábra a gyakoriságokat adja meg, és felsorolja az egyes osztályokban[4] előforduló értékeket. A megfigyelt érték utolsó számjegye a levél (leaf). Erről az ábráról például azonnal megállapítható, hogy a 75 éves kor mellett a nők másik módusza a 78, mert mindkettő 9-9 országban fordul elő. (1.4. ábra)

Nagyobb minta esetében egy-egy levélke több (egymáshoz közeli) esetet jelképez. A minimum vagy maximum előtti szakadást, és a terjedelmen belüli üres kategóriákat is láthatjuk egy ilyen ábrán. is láthatjuk egy ilyen ábrán.

```
Average female life expectancy Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     9 Extremes    (=<50)

     3       5 .  223

     3       5 .  455

     2       5 .  77

     5       5 .  88889

     1       6 .  3

     3       6 .  455

     6       6 .  677777

     7       6 .  8888899

     6       7 .  000001

     6       7 .  222333

    14       7 .  444445555555555

    11       7 .  66666777777

    16       7 .  888888888999999

    14       8 .  00000001111111

     3       8 .  222

 Stem width:    10

 Each leaf:       1 case(s)
```

1.4. ábra: Stem-and-leaf gyakorisági ábra

[4] Ordinális skálán mért adatok is megjeleníthetők így.

Házi feladat: Bizonyítandóak az alábbi állítások:

- A nyesés hatására a változó szórása biztosan csökken.
- A nyesés után az átlag lehet azonos, kisebb, sőt nagyobb is, mint az eredeti adatok átlaga.

1.5. A normalitásvizsgálat numerikus és grafikus módszerei

A normalitás vizsgálatának két mutatószámát, a ferdeség és a csúcsosság mérőszámait már ismertettük az 1.2. alfejezetben. Mindkettőre nullhipotézist állítottunk fel, és t-teszttel vizsgáltuk a normális eloszlástól való eltérés mértékét.

Bár az SPSS nem számolja, a ferdeség és csúcsosság részeredményeinek ismeretében könnyen meghatározható **Jarque-Bera – normalitás tesztje**[5], ha a mintából becsült ferdeség (4) és csúcsosság (6) négyzeteit összegezzük az alábbiak szerint, ahol *n* a minta mérete:

$$JB = \frac{n}{6}\left(\gamma_1^{\,2} + \frac{1}{4}\gamma_2^{\,2}\right)$$

A JB teszt használata csak nagy minta[6] esetén ajánlott, és a JB értékét a khi-négyzet eloszlással vetjük egybe. A teszt szabadsági foka kettő, hisz két négyzetszámot adunk össze.

Eredményeink alapján (JB_férfi= 21,702 és JB_nő=22,549) mindkét változóra el kell vetni a normalitási feltevést, hiszen a khi-négyzet kritikus értéke 5,99 (ha a szabadsági fok=2 és p=0,05)

Ha a minta elég nagy, akkor χ^2 próbát végezhetünk annak a hipotézisnek a tesztelésére, hogy a változó normális eloszlást követ. Az SPSS két normalitás tesztet számol a leíró statisztikák között. A Shapiro-Wilks tesztet értékeljük n<50-re, nagyobb mintára a Kolmogorov-Szmirnow teszt számított értéke alapján következtetünk.

1.5.1. Kolmogorov-Szmirnov próba

Itt az empirikus eloszlás függvény és a normális eloszlás összevetését úgy végezzük, hogy a sokasági várható értéket és a szórást is a mintából becsüljük. Ezt a változatot Lilliefors 1967-ben javasolta.

Az adatokat nagyság szerint sorba rendezzük, majd standardizáljuk: $z_{(i)} = \left(x_{(i)} - \overline{x}\right)\!/\,s$. Ehhez a z-hez tartozó sztenderd normális

[5] Ökonometriából is ismert lehet a JB teszt: Jarque, Carlos M. és Bera, Anil K. (1980). "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". *Economics Letters* **6** (3): 255–259.

[6] Mivel 109 adatból dolgozunk, alkalmazható a J-B teszt.

eloszlás függvényértéke: $\Phi(z_{(i)})$. Az empirikus eloszlásfüggvény lépcsős függvény, 0 és 1 között i/n értéket vesz fel.

Így $D_i = |\, i/n - \Phi(z_{(i)}) \,|$ eltérések maximuma, $\max_i D_i$ lesz a teszt függvény értéke. Szabadsági foka n, azaz a megfigyelések száma.

A nem-parametrikus[7] próbák blokkjában is készíthető egymintás K-S teszt, de ott a $\max_i D_i$ helyett $\sqrt{n}\max_i D_i$ adódik.

1.5.2. Shapiro-Wilk W mutató

Az SPSS által közölt másik tesztet Shapiro és Wilk publikálta[8] 1965-ben. Itt is a növekvő sorba rendezett $x_{(i)}$ adatokból indulunk ki. A W mutató számlálójában levő súlyokat ($\underline{a}$ vektor) a sorba rendezett adatok átlaga ($\underline{m}$ vektor) és kovariancia mátrixa (V) alapján határozzuk meg. A teszt szabadsági foka a megfigyelések száma.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$(a_1, \ldots, a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}$$

1.6. táblázat: Normalitás próbák

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Average female life expectancy	,174	109	,000	,860	109	,000
Average male life expectancy	,164	109	,000	,882	109	,000

a. Lilliefors Significance Correction

[7] A nem-parametrikus próbák nem valamely eloszlást jellemző paraméter becsült értékét tesztelik.

[8] Shapiro, S. S.- Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". *Biometrika* **52** (3-4): 591–611. A Biometrika folyóirat nagyon sok, statisztikai szempontból jelentős írást jelentetett meg. Az ELTE Könyvtárában olvashatók is a régi újságok.

Az 1.6. táblázat alapján mindkét változóra elvetjük a normalitási feltevést[9], mert a K-S teszt empirikus szignifikancia szintje mindkét változóra kisebb, mint 0,05.

1.5.3. Grafikus normalitás vizsgálat

Grafikus normalitás vizsgálatot[10] is kapunk az Explore-ból Q-Q plot néven. Ez a kvantilisek[11] ábrája, innen kapta nevét, azaz a Q-Q-t. Ha a vízszintes tengelyen az életkort, a függőlegesen pedig a sztenderd normális eloszlás u változóját ábrázoljuk, akkor az $u = \Phi^{-1}\left[\Phi(\frac{x-\overline{x}}{s})\right] = \frac{x}{s} - \frac{\overline{x}}{s}$ transzformáció után a normális eloszlású változó értékei a 45 fokos egyenes mentén helyezkednek el, vagy az átló körül véletlenszerűen szóródnak.

Ha a normalitási feltevés helyes, csak a paraméterekben tévedtünk, akkor az egyenes helyzete más lesz.

Ha a normalitás nem teljesül, amint ez az 1.5. ábrán is látható, akkor a pontok szisztematikusan térnek el az egyenestől.

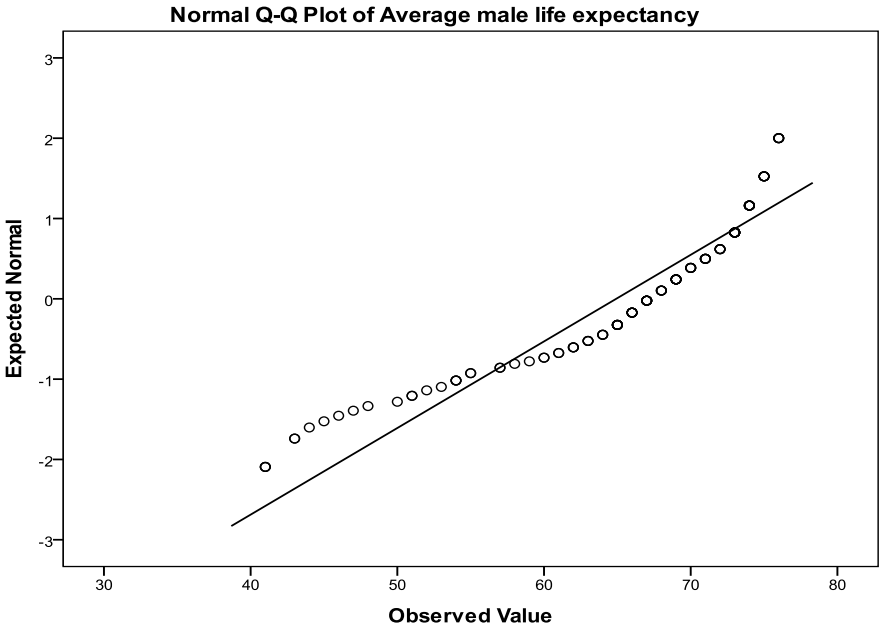
A férfiak várható élettartama a tesztek alapján sem követett normális eloszlást. Nagyon alacsony átlagéletkorban jóval több országban halnak meg, mint ami a normális eloszlás alapján várható lenne. 60 körüli várható élettartamot kevesebb országban látunk, és 75 fölött ismét magasabb a megfigyelt, mint a várt gyakoriság. A Q-Q ábrához megkapjuk a feltételezett és a megfigyelt eloszlás eltérését mutató változatot is, melynek neve: Detrended Q-Q, és a 1.6. ábrán látható.

---

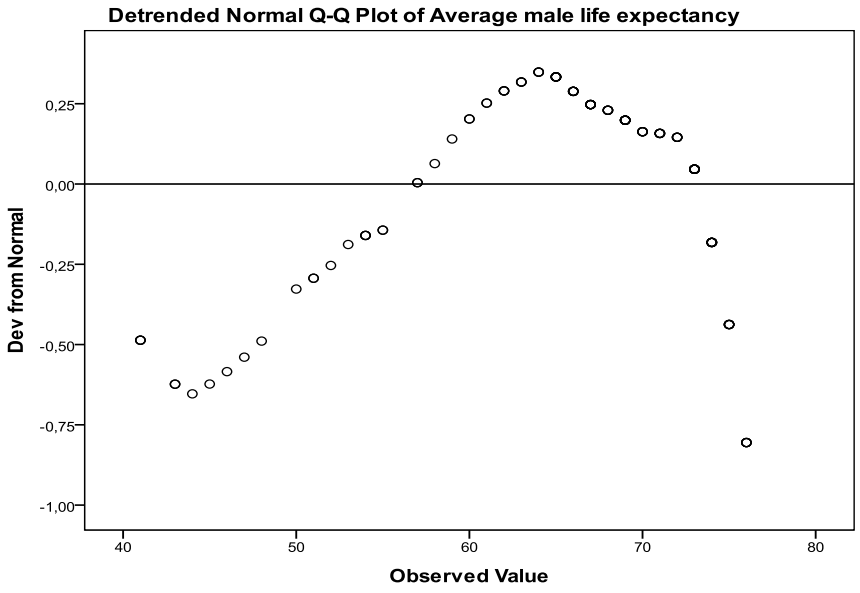
[9] Az 1.1.ábrán a hisztogramot látva biztosak lehettünk a döntésben, szinte felesleges volt a teszt.

[10] Ajánlott olvasmány a témához Hunyadi László cikke a 2002. januári Statisztikai Szemlében.

[11] A kvantilisek között a legismertebbek a másodrendű kvantilis= medián, a negyedrendű=kvartilisek, a tized-rendűek, azaz a decilisek, és a századrendűek, a percentilisek.



1.5. ábra: Grafikus normalitás vizsgálat Q-Q ábrán



1.

6. ábra: A normális eloszlástól való eltérés ábrája

Ha az a célunk, hogy normális eloszlásúvá transzformáljunk egy ferde eloszlású változót, akkor több lehetőség közül választhatunk.

- Szóba jöhet a szélső, extrém értékek elhagyása. Ez akkor igazán hasznos, ha kevés ilyen adatunk van, és ezek távol vannak a megfigyelések többségétől.
- A pozitív ferdeségű mutatók logaritmálása vagy az adatokból való gyökvonás ajánlott, ez legtöbbször hatékonyan orvosolja a problémát.

A pénzügyi mutatók, a biztosítási összegek és más jövedelem-adatok eredendően pozitív ferdeségűek, mert a kisebb értékek előfordulása gyakoribb. A szélső értékek elhagyása alapos megfontolást igényel a pénzügyi elemzésekben. Egy különösen nagy összegű hitelt felvevő adós vagy egy hatalmas kárt bejelentő biztosított adatainak elhagyása az egész számítás értelmét megkérdőjelezheti!

A Transform / Compute Variable menűpontban megtaláljuk az aritmetikai függvények között mind a tízes alapú, mind a természetes alapú logaritmust.

A WORLD95.sav-ban szereplő mutatók közül egy főre jutó GDP pozitív ferdeségű (1,146, és st. hibája 0,231) ezért transzformáljuk. A GDP/fő tízes-alapú logaritmusát tartalmazza az adatállomány, ezért most az e-alapú logaritmust, az ln(gdp)-t készítjük el. Ha összevetjük a két transzformált változót, akkor mindkettő a szimmetrikushoz közelebbi eloszlást követ, ferdeségük azonosan -0,243 és a sztenderd hiba 0,231.

A K-S teszt alapján már nincs elegendő bizonyítékunk arra, hogy a normalitást 5%-os valószínűségi szinten elvessük a 1.7. táblázat szerint, míg a kismintás W mutató továbbra is elvetné a normalitási feltevést.

1.7. táblázat: A logaritmálás hatása a tesztekre

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Gross domestic product / capita	,204	109	,000	,800	109	,000
Log (base 10) of GDP_CAP	,085	109	,053	,950	109	,000
Lngdp (base e)	,085	109	,053	,950	109	,000

a. Lilliefors Significance Correction

Házi feladat: Bizonyítandó, hogy az x adatsorra készített $\log_{10}(x)$ és az ln x átlaga és szórása eltér, de a két adatsor ferdesége és csúcsossága megegyező lesz.

1.6. Idősoros adatok statisztikai elemzése

Az adatelőkészítéshez tartozó lépés az idősoros adatok differenciájának képzése is. A pénzügyi életben számos idősor, pl. hozam, árfolyam adat gyűlik, de az időbeli egymásutániság miatt nem tekinthetők független megfigyeléseknek, és nem stacionáriusak. A differencia képzésével kiküszöböljük ezeket, és így leíró statisztikai elemzéseket végezhetünk, korrelációt számolhatunk, és a páronkénti lineáris korreláción alapuló további modelleket illeszthetünk.

Az adatokat az importálás után SPSS állományként[12] elmenthetjük. A változók mérési skáláját érdemes ellenőrizni, mert nem mindig sikerül tökéletesen az átvitel.

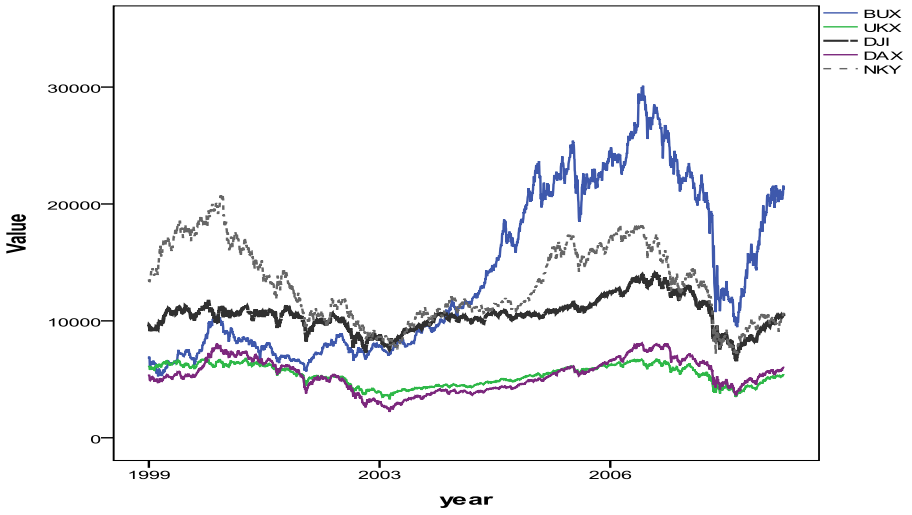
A számításokat az Indexek.xls adatállomány megnyitásával és importálásával végezhetjük el. Ebben 1999.01.07. és 2009.12.31. között hétköznapokon öt tőzsdei index értékeit látjuk. A megfigyelések száma 2753, de mivel ezek egymást követő napok mért adatai, ezért nem véletlenszerű és egymástól nem független megfigyeléseink vannak.

Az adatsorok egymástól eltérő alakulását jól mutatja a Multiple Line Chart, ahol az egyedi értékeket választva (Values of individual cases) kaphatjuk meg a 1.7. ábrát.

A legnagyobb hullámzást a BUX mutatja, míg az angol (UKX) és a német (DAX) indexek első látásra is együttmozognak, azaz kointegráltak[13].

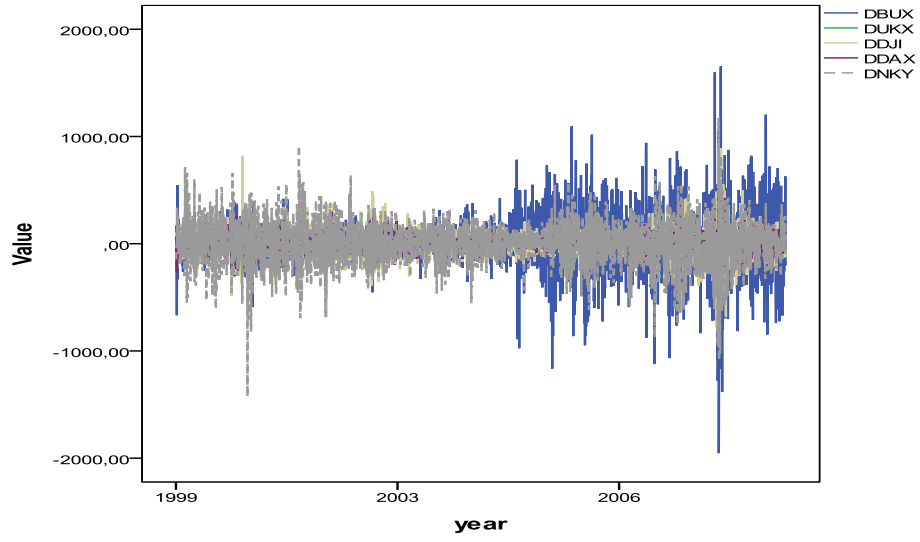
[12] Az SPSS egy munkalapos Excel állományt tud közvetlenül beolvasni, ha az első sorban a változók rövid neve áll. (A név legyen maximum 8 alfanumerikus karakter hosszú, célszerű ékezet nélküli, angol betűket használni, speciális karakterek nélkül.)

[13] Két idősort kointegráltnak nevezünk, ha együtt mozognak az időben, de ok-okozati kapcsolatot nem tételezünk fel közöttük. Ökonometria könyvek részletesen foglalkoznak ezzel a módszerrel.



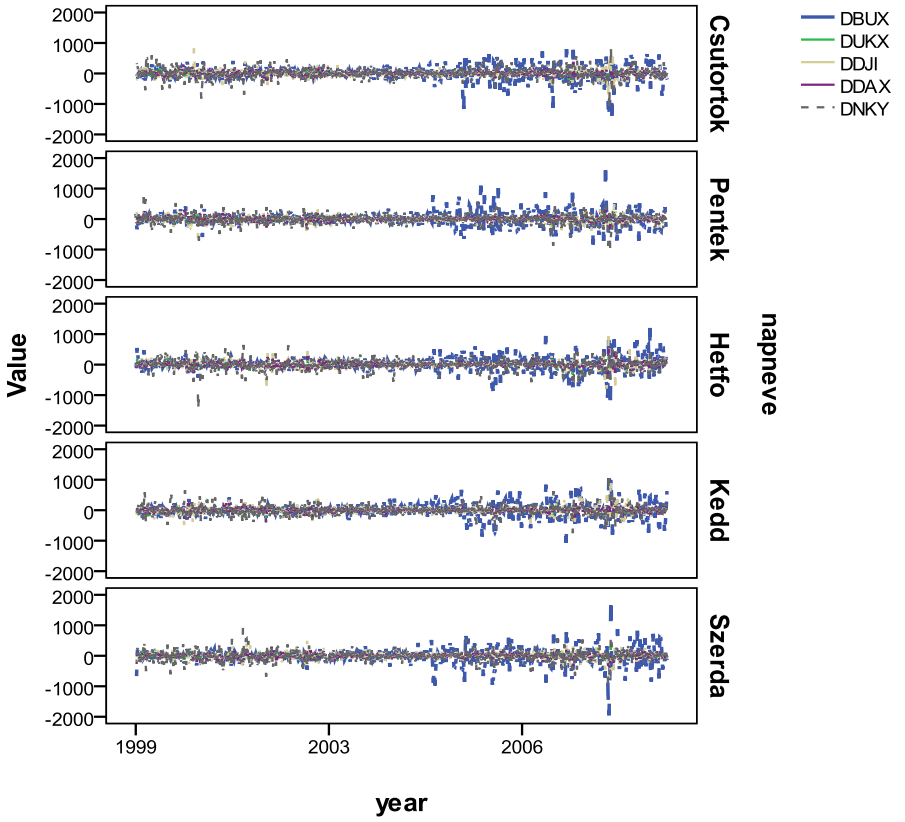
1.7. ábra: Az eredeti 5 tőzsdeindex 11 éves adatsorai

De most nem közvetlenül az idősorok viselkedését elemezzük. Célunk az egymást követő napokra képzett különbségek elemzése. Ezek már stacionáriusok, ahogy az 1.8. ábra mutatja.



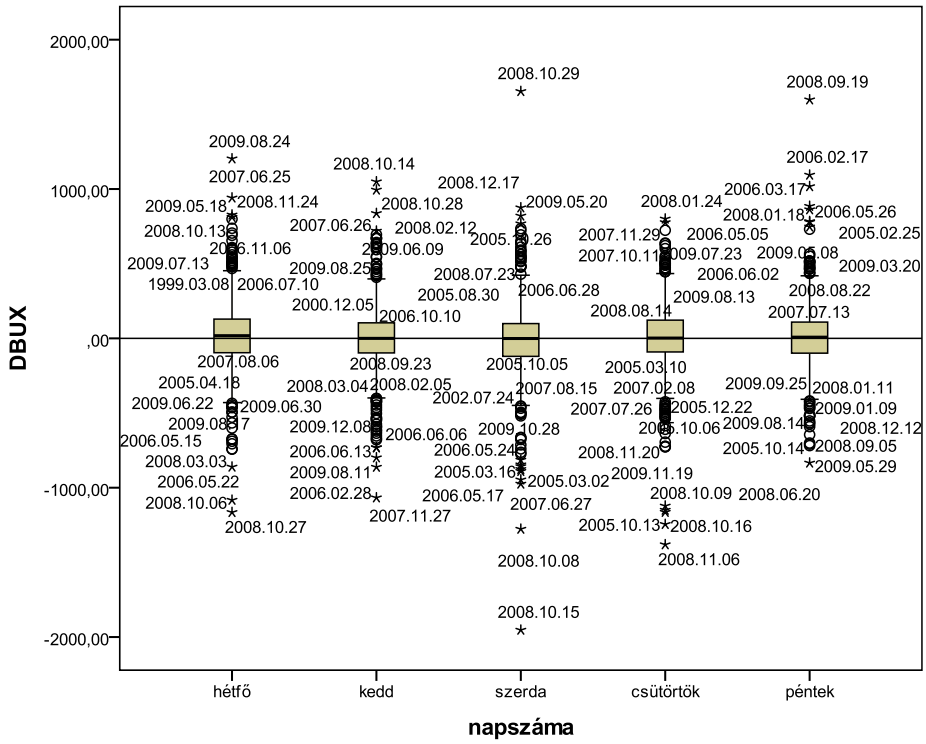
1.8. ábra: Az 5 tőzsdeindex első differenciáinak idősora

Érdekes kérdés, hogy az egyes napok szerint különböznek-e a differenciák. Ezt részben a panel ábrákon tekinthetjük meg (1.9. ábra), részben az Explore-ban factor=napok beállítással számolhatjuk ki, és dobozdiagramon ábrázolhatjuk. (1.10. ábra) Az adott nap differenciája az jelenti, hogy az előző napról erre átlépve hogyan változtak az indexek. Tehát a hétfői differencia a hétfő-péntek különbséget méri.



1.9. ábra: A differenciák napok szerint bontott idősorai

Az 1.10. ábrán a dobozdiagramok egymás mellett mutatják a napokra vonatkozó magyar adatokat. Az öt doboz közepén a medián vonalat látjuk, ami általában nem zérus. Látható, hogy a dobozok magassága kicsi, azaz a változások 50%-a nem volt jelentős.



1.10. ábra: A magyar differenciák dobozdiagramjai naponként

A magyar és a német adatokból képzett differenciákra számolt eredmények egy részét a „Report” beállítással tömörebb formában tartalmazza az 1.8. és az 1.9. táblázat. A napok közötti átlagok eltérése mellett a relatív szórások hatalmas értékei érdemelnek figyelmet. A szórás/átlag értékek a százat is meghaladják a magyar keddi adatokra! A magyar adatok nagyobb terjedelméhez nagyobb szórás is tartozik

A változások átlaga szerdánként a magyar és a német adatokra negatív, tehát keddről szerdára inkább volt csökkenés, mint növekedés. Ez a „fekete” szerda[14] megállapítás mind az öt országra érvényes. A japán és az amerikai átlagos differencia emellett még pénteken, az angol átlag pedig kedden negatív.

[14] 2008. október 15-ére volt minden országban nagy esés, kivéve Japánt. Ott másnap, október 16-án érték el a változások mélypontját.

1.8. táblázat: BUX index első differenciának statisztikai mutatói napok szerint

Case Summaries

DBUX

napszáma	N	Mean	Minimum	Maximum	Std. Deviation
hétfő	525	21,8571	-1165,00	1203,00	250,27327
kedd	559	2,3971	-1067,00	1049,00	241,33509
szerda	559	-13,1878	-1953,00	1654,00	275,93169
csütörtök	557	3,4147	-1381,00	800,00	250,26170
péntek	552	12,8786	-834,00	1598,00	240,67750
Total	2752	5,2522	-1953,00	1654,00	252,15855

1.9. táblázat: DAX index első differenciának statisztikai mutatói napok szerint

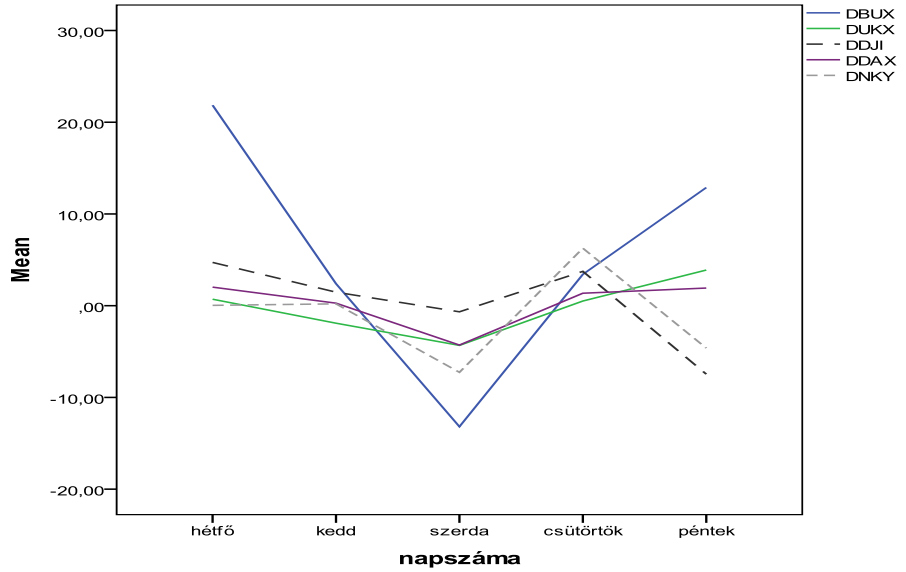
Case Summaries

DDAX

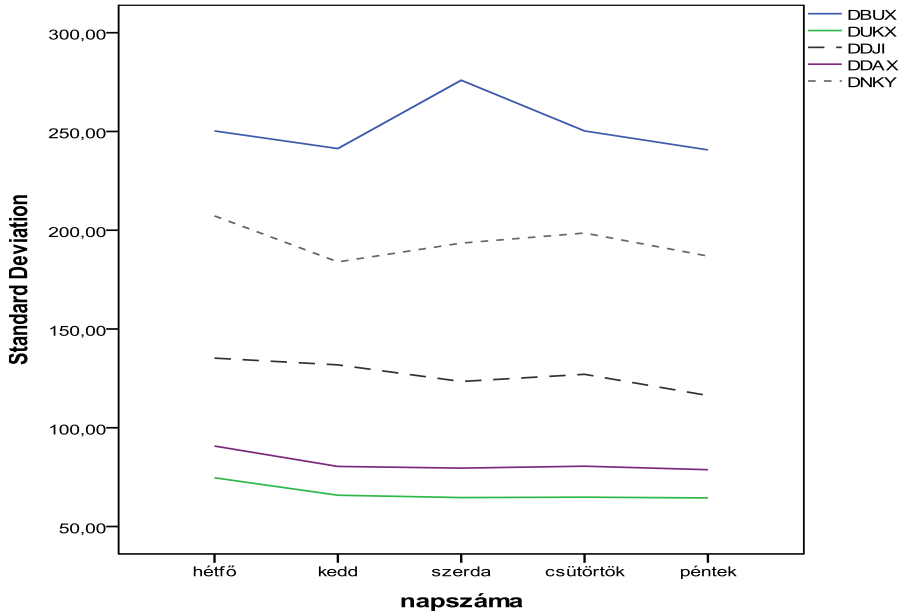
napszáma	N	Mean	Minimum	Maximum	Std. Deviation
hétfő	525	2,0229	-524,00	518,00	90,73243
kedd	559	,2755	-396,00	488,00	80,41003
szerda	559	-4,2934	-337,00	298,00	79,56389
csütörtök	557	1,3591	-353,00	382,00	80,53497
péntek	552	1,9221	-343,00	327,00	78,76485
Total	2752	,2304	-524,00	518,00	81,99164

Az 1.11. ábra a napokra számított átlagokat és az 1.12. ábra a napokra képzett szórásokat mutatja országonként. Ezek az ábrák „Multiple line, Summaries of separate variables” beállítással készültek, ahol a kategória tengelyt a napok jelentik.

Az angol és a német tőzsdei adatok nullához közeli átlagos változása és legkisebb szórása a legszembetűnőbb a két ábrán.



1.11. ábra: Az öt index változásainak átlaga a 11 év során



1.12. ábra: Az öt index változásainak szórása a 11 év adataiból

Házi feladat:

A 1.8. és a 1.9. táblázat eredményeit érdemes előállítani és áttekinteni az amerikai, az angol és a japán adatokra is

2. Kategóriák és kereszttáblák elemzése

Ha vizsgált adathalmazunkban több változó van, feltételezhető, hogy vannak közöttük független változó-párok, és vannak olyanok is, amelyek hatnak egymásra vagy kölcsönös kapcsolatban állnak egymással. A kapcsolat létének és erősségének feltárására több módszer áll rendelkezésünkre, melyek közül a mérési skálák ismeretében választhatunk. A legegyszerűbb eljárások a következők:

- Két nominális, két ordinális vagy vegyes (nominális és ordinális) skálán mért változókra vonatkozó megfigyeléseket kereszttáblába rendezzük, és függetlenségi hipotézist fogalmazunk meg.
- Ordinális skálájú változókra (Spearman) rangkorrelációt számolunk.
- Intervallum (vagy arány) skálán mért változók közötti lineáris kapcsolatot korrelációval mérjük.

Kettőnél több változó kapcsolatrendszerének vizsgálatára a későbbi fejezetekben szereplő módszerek alkalmazhatók.

2.1. Kategóriák előállítása

A gazdasági-pénzügyi elemzések többségében sok adatból kiindulva képezhető kategória vagy index, aminek az értelmezése könnyebb, mint az eredeti adatok minősítése. Ilyen például az ország-kockázati besorolás, ahol a besorolási kategória változása, például egy „leminősítés" bejelentése a részletek közlése és ismerete nélkül is információt ad egy országról.

A banki és biztosítói gyakorlatban is sok olyan adat áll az elemző rendelkezésére, amelyet csoportosítva, kategorizálva érdemes felhasználni. Példaként a következők említhetők:

- A hiteltörlesztésben késedelmes ügyfelek besorolása a legalább 30, 60 és 90 napos késedelmi kategóriába.
- A biztosításban a kockázatelbírálás folyamata, melynek bináris kimenetele az ügyfél kockázatának vállalása vagy elmenűponta, vállaláskor pedig esetleg magasabb díjosztásba sorolás.
- A gépjármű felelősségbiztosításban a bónusz-málusz rendszer fokozatai.

- A testtömeg index (BMI) arány skálán számítható, hisz képlete = testsúly (kg)/ magasság (méter)^2, mégis értékelése 4 kategóriába[15] sorolva történik:

  Sovány, ha BMI < 18,5

  Normál testalkatú 18,5 - 24,9 között

  Túlsúlyos 25 - 29,9 között

  Erősen testes, túlsúlyos, ha BMI > 30.

Az ügyfelek további ismert tulajdonságai kapcsolatban állhatnak a kategória-besorolással. Elemezni érdemes például azt, hogy az egyén neme, életkora, családi állapota, jövedelme, a gépjármű típusa közül melyik és milyen hatású. Itt azonban felmerül az eltérő mérési skálák problémája, továbbá az, hogy elegendő megfigyelésünk van-e.
Az életkor vagy a jövedelem mérése intervallum skálán történik, de egy-egy életkorhoz vagy jövedelem szinthez nem feltétlenül tartozik sok egyén. Ezért statisztikailag indokolt a skálákat transzformálni, és ordinális mérési szintű kategóriákba sorolni az ilyen változókat. A továbbiakban a kategóriákat használva a kereszttáblákat lehet elemezni.

A skála-transzformáció ebben az esetben a skála leértékelését jelenti, azaz információt veszítünk.

Eredeti és új skála neve	Nominális	Ordinális
Ordinális	Szélső értékek összevonása, középső megtartása	Kevesebb kategória képzése
Intervallum vagy arány	Az átlagos és az átlagtól eltérő értékek kategorizálása	Az átlagos és az átlagtól felfelé valamint lefelé eltérő megfigyelések osztályba sorolása

Az értékek és a kategóriák összevonására nemcsak a skála változtatása miatt kerül sor. Szükség lehet erre, akkor is, ha egy-egy osztályba kevés megfigyelés került. Erre az SPSS/Transform/Recode into Different Variables használata ajánlható, hogy az eredeti adatok is megmaradjanak.

A kategorizálás/diszkretizálás számos módon elvégezhető. Szakmai megfontolások alapján és az eloszlást megvizsgálva érdemes választani az alábbiak közül.

- Kerekítést alkalmazunk, amikor a legközelebbi egész számot tartjuk meg: az életkort is csak években mérjük, a jövedelmet 1000-re, százezerre kerekítve adjuk meg.

[15] Sportolók, idősebbek értékelésére más határok alkalmazhatóak.

- Egyenlő hosszú kategóriákat képzünk, pl. 5 éves életkor tartományokba soroljuk az embereket, vállalkozásokat.
- Egyenlő gyakoriságú csoportokat hozunk létre, pl. kettéosztjuk a mediánnál, 10 csoportot képzünk a decilisek mentén vagy 4 csoportot a kvartilisek szerint.
- Osztályozással, amikor a kategóriahatárokat előre kijelöljük. (Ilyen a dolgozatok pontozását követően megállapított érdemjegy is.)
- Előzetes kategória határok kijelölése nélkül, a több dimenzióban leghasonlóbb megfigyelések csoportba sorolásával, amit klaszterelemzéssel[16] készíthetünk el.

Mielőtt az eljárásról döntünk, érdemes megvizsgálni az adatok lehetséges tagolását. Ehhez felhasználhatjuk az SPSS/ Transform/Visual binning menűpontját, amely grafikus és numerikus megközelítést is alkalmazva többféle felosztást tud megjeleníteni.

a) Egyenlő hosszú intervallumokat kérve az alábbiak közül 2 értéket kell beírni:

- Első metszéspont
- Metszéspontok száma
- Intervallum hossza

b) Egyenlő percentilisekre bontást kérve az egyik értéket kell megadni:

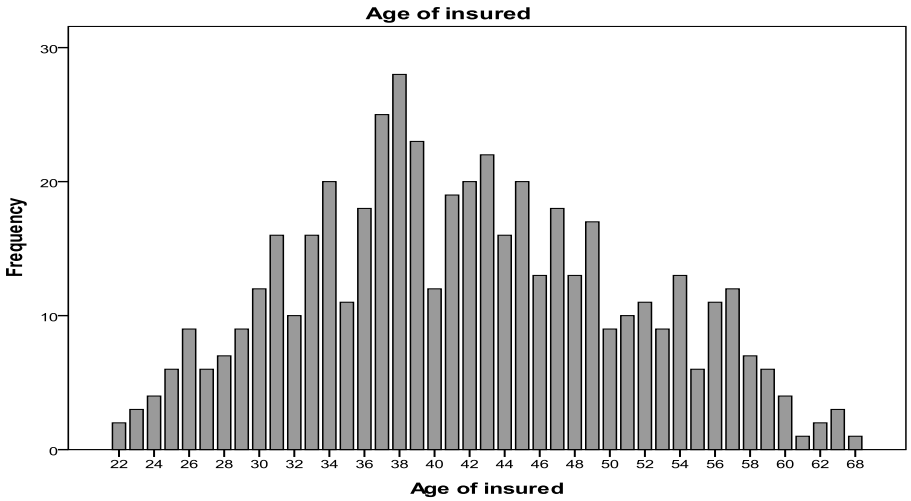
- Metszéspontok száma (3 metszéspontra 25%-os felosztás adódik)
- Intervallum hossza (20% megadása 4 metszéspontot ad!)

c) Az átlag és a szórás alapján az átlag körül 1, 2 vagy 3-szoros szórásnyi intervallumokat választhatunk, ha az előzetesen ábrázolt adatok hisztogramja normális eloszláshoz hasonló képet mutat.

Ha megnyitjuk a Program Files\SPSS\tutorial\samplefiles\autoaccidents.sav adatokat, és az 500 ügyfél életkor megoszlását oszlopdiagramon[17] ábrázoljuk, akkor a 2.1. ábrán látható, hogy érdemes a 22-68 év közötti vezetőket kevesebb kor-kategóriába sorolni, mert egy-egy életkorhoz – statisztikai szempontból – kevés ember tartozik.

[16] A klaszterelemzés módszercsaládot a 3. fejezetben mutatjuk be.

[17] Példánkban az oszlopdiagram nem egyezik meg a hisztogrammal. A hisztogram nulla előfordulást jelezne 63 évnél és 65-67 év között, mivel nincs ezekhez az életévekhez tartozó ügyfél. Az oszlopdiagram csak a megfigyelt értékeket tükrözi.



2.1. ábra: A vezetők életkorának oszlopdiagramja

Arra érdemes figyelni, hogy ha egyenlő hosszú intervallumokat készítünk, akkor a „középső" kategóriában nagyon sok egyén lesz, a szélsőkben pedig nagyon kevés.

Minél csúcsosabb az eloszlás, annál erőteljesebben jelentkezik ez a probléma.

A statisztikai megfontolások (legalább 5-10 megfigyelés essen egy intervallumba) mellé értelmezési szempontokat is érdemes figyelembe venni. Ha általában 10 éves intervallumokban közölnek adatokat, akkor készítsünk mi is ilyen felosztást. A kezdő értéket megadva és 4 kategóriát kérve a Paste gombbal az alábbi Syntax utasítást állítjuk elő:

```
* Visual Binning.

*age.

RECODE  age (MISSING=COPY) (LO THRU 28.0=1) (LO THRU 38.0=2)
(LO THRU 48.0=3) (LO THRU 58.0=4) (LO

   THRU HI=5) (ELSE=SYSMIS) INTO age10.

VARIABLE LABELS  age10 'Age of insured (Binned)'.

FORMATS  age10 (F5.0).

VALUE LABELS  age10 1 '<= 28' 2 '29 - 38' 3 '39 - 48' 4 '49 - 58' 5 '59+'.

VARIABLE LEVEL  age10 (ORDINAL).

EXECUTE.
```

Érdemes bináris kategorizálást alkalmazni a balesetek számára, így a balesetmentesen vezetőket elválasztjuk a balesetet szenvedőktől. Ezt az SPSS/Transform/Recode into Different Variables funkciójával kapjuk: a nullák megmaradnak, a többi érték 1 lesz. (A címkébe beírhatjuk, hogy „egy vagy több".) Végül pedig kereszttáblában ellenőrizzük, hogy mind az 500 megfigyelés átkódolása megtörtént, és nem veszítettünk adatot.

```
RECODE accident (0=0) (ELSE=1) INTO accid.
EXECUTE.
```

		accid		
		zero accident	one or more accident	Total
Number of accidents past 5 years	0	122	0	122
	1	0	139	139
	2	0	107	107
	3	0	63	63
	4	0	39	39
	5	0	19	19
	6	0	9	9
	7	0	2	2
Total		122	378	500

A csoportok kialakítása után kereszttáblában vizsgáljuk a balesetek száma és a vezető életkor-csoportja közötti kapcsolatot. Az előkészítő lépések után tekintsük át a kereszttábla elemzés módszertanát.

2.2. Keresztábla készítése és elemzése

Ebben a fejezetben a nominális és/vagy ordinális skálán mért változókra[18] felírható kombinációs táblákkal foglalkozunk, és a változók közötti kapcsolatot mérjük.

2.2.1. Matematikai-statisztikai háttér

A kereszttábla elemzésekor a két változó közötti függetlenség hipotézisét vizsgáljuk, és a függetlenség elvetésekor az asszociációs kapcsolat erősségét mérjük. A változók közötti kapcsolatrendszerre azonban számos más hipotézis is felírható.

a) Két nominális vagy ordinális mérési szintű változó esetén kereszttáblába rendezzük az együttes előfordulásuk gyakoriságait:

Változók	B_1	B_2	…..	B_c	Összesen
A_1	f_{11}	f_{12}		f_{1c}	m_1
A_2	f_{21}	f_{22}			m_2
…			f_{ij}		m_i
A_r	f_{r1}			f_{rc}	m_r
Összesen	n_1	n_2	n_j	n_c	n

Kétdimenziós táblára öt modell illeszthető.

b) A táblában a várt gyakoriságok (F) alakulására felírható modellek közül a legegyszerűbb a minimális vagy null-modell. Ekkor a tábla minden cellájában egyenlő gyakoriságot tételezünk fel, az összes megfigyelést szétosztjuk az összes cella (rc) között: $F_{ij} = n / rc$ (2.1)

1.Példa: Várt gyakoriságok a null-modellben

A táblában a megfigyelt peremgyakoriságok szerepelnek, amelyek nem feltétlenül egyeznek meg a várt gyakoriságok sor- és oszlopösszegeivel.

Változók	B_1	B_2	B_3	Összes
A_1	100/6	100/6	100/6	60
A_2	100/6	100/6	100/6	40
Összes	10	50	40	100

[18] Nominális és intervallum változók közötti kapcsolat vizsgálatára például a szóráselemzés alkalmazható.

c) Feltételezhetjük, hogy a várt gyakoriságokra csak az egyik változó hat. Az elsőrendű hatás egyik modelljében csak a sorváltozó hat, az adott kategória összes gyakoriságát egyenletesen szétosztjuk az oszlopok között, mert az oszlopvalószínűség konstans. Ekkor

$$F_{ij} = m_i / c \qquad (2.2)$$

2. Példa: Sorhatás modelljében várt gyakoriságok

Változók	B_1	B_2	B_3	Összes
A_1	60/3	60/3	60/3	60
A_2	40/3	40/3	40/3	40
Összes	10	50	40	100

d) Elsőrendű modellt az oszlopváltozó hatására is felírhatunk, az oszlop összes gyakoriságát egyenlően elosztjuk a sorok között. Ekkor a sorvalószínűség konstans, és a várt gyakoriság: $F_{ij} = n_j / r$ (2.3)

3. Példa: Oszlophatás modelljében várt gyakoriságok

Változók	B_1	B_2	B_3	Összes
A_1	10/2	50/2	40/2	60
A_2	10/2	50/2	40/2	40
Összes	10	50	40	100

e) Elsőrendű modellt illesztünk akkor is, ha sor- és oszlopváltozók egymástól független hatását tételezzünk fel. Ekkor a függetlenség modelljét írjuk fel, amelyben a sor és az oszlop összegeket is figyelembe vesszük a várt gyakoriság becslésekor:

$$F_{ij} = m_i \, n_j / n \qquad (2.4)$$

4. Példa: Függetlenségi modell várt gyakoriságai

Változók	B_1	B_2	B_3	Összes
A_1	60*10/100	60*50/100	60*40/100	60
A_2	40*10/100	40*50/100	40*40/100	40
Összes	10	50	40	100

f) Az egyes változók egyedi hatása mellett kölcsönhatásuk, azaz másodrendű hatás is szerepel a telített modellben. Ez a modell teljesen a megfigyelt gyakoriságok alapján becsli a várt előfordulásokat: $F_{ij} = f_{ij}$ (2.5)

Ez utóbbi esetben tökéletes az illeszkedés, az előbbiekben viszont mérni kell a megfigyelt és a várt gyakoriságok eltérését. Az öt modell tovább vizsgálható

loglineáris modellezéssel. Ez az eljárás terjedelmi korlátok miatt nem szerepel a jegyzetben.

A **függetlenség** feltételezése mellett előforduló eltérések mértékét a Pearson által javasolt khi-négyzet próbával (2.6), likelihood arány teszttel (2.7) vagy lineáris asszociációs teszttel (2.8) mérjük.

- $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(f_{ij} - m_i n_j / n)^2}{m_i n_j / n}$, szabadságfok: (r-1)(c-1) (2.6)

- Likelihood arány teszt: $L(f) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} \ln \frac{f_{ij}}{m_i n_j / n}$, sz.fok: (r-1)(c-1)

(2.7)

- Lineáris asszociáció tesztje[19] $\chi^2 = (n-1)r^2$, (2.8)

ahol r a lineáris korreláció. A teszt szabadsági foka 1.

Ha a függetlenség hipotézisét elvetjük, akkor mérni kell az asszociáció szorosságát. Erre számos mutatószám létezik, közülük a szakmai feltételezések és a mérési skála alapján választunk. Az asszociációs mérőszámok ismertetését az SPSS-ben elérhető csoportosításban mutatjuk be.

2.2.2. Keresztábla elemzés megvalósítása az SPSS-ben:

A leíró statisztikák között találjuk a kereszttábla elemzést annak ellenére, hogy itt már hipotézisvizsgálatot[20] végzünk.

Analyze/Descriptive/Crosstabs választás után a következő beállításokat tehetjük**:**

1. **Sor- és oszlopváltozó kijelölése**[21] az elemzés célja szerint.
2. **Layer:** rétegekre, alcsoportokra bontható a kereszttábla, így vizsgáljuk a 2 változó függését, az eredményeket is így bontva kapjuk.

[19] Ez a lineáris asszociáció Mantel-Haenszel-féle tesztje**.**

[20] Az eloszlásmentes vagy más néven nem paraméteres tesztek családjába tartozik a khi-négyzet próba.

[21] Csak nomiális és/vagy ordinális változókat választunk. Intervallum/arány skálájú változók előzetesen kategóriákra bontandók a Recode menűponttal.

A Crosstab menü STATISTICS opció használata

A) Nominális változókra számítható három khi-négyzet alapú asszociációs mérőszám (2.9)-(2.11), melyek szimmetrikusak és 0-1 között mérnek:

$$\text{Phi} = (\chi^2/n)^{1/2} \tag{2.9}$$

A (2.9) mutató értelmezését nehezíti, hogy a khi-négyzet várható értéke a szabadságfok (varianciája pedig annak kétszerese), ezért kevés megfigyelés esetén Phi >1 is előfordulhat.

$$\text{Cramer-V} = \left(\frac{\chi^2}{n(q-1)} \right)^{1/2} \tag{2.10}$$

Ahol a (2.10) nevezője az aszimptotikus sztenderd hiba: ASE(V)= $(n(q-1))^{-1/2}$ és V/ASE(V)~N(0,1). A (2.10)-ben q=min(r,c).

$$\text{Kontingencia együttható CC} = (\chi^2/(n+\chi^2))^{1/2} \tag{2.11}$$

Nominális változókra PRE[22]-alapú nem-szimmetrikus mérőszámokat is választhatunk:

A Guttman által javasolt Lambda mutatónak három változata van:

1. ha B oszlopkategória ismert és az A változó i. sorába esést becsüljük, akkor

$$\lambda_{a|b} = \frac{\sum_j \max f_{ij} - \max m_i}{n - \max m_i} \tag{2.12}$$

2 ha a sor szerinti besorolás ismert, akkor

$$\lambda_{b|a} = \frac{\sum_i \max f_{ij} - \max n_j}{n - \max n_j} \tag{2.13}$$

3. szimmetrikus mutató:

$$\lambda = \frac{\sum_j \max f_{ij} - \max m_i + \sum_i \max f_{ij} - \max n_j}{2n - \max m_i - \max n_j} \tag{2.14}$$

Goodman-Kruskal tau mértékének is 3 változata van, itt csak egyet írunk fel, amely azt méri, hogy a hibavalószínűség relatív csökkenése mekkora, ha a sorváltozó szerinti kategória ismert.

[22] PRE: Proportional Reduction of Errors= relatív hibacsökkenés= $(\text{hiba}_1 - \text{hiba}_2)/\text{hiba}_1$.

$$\tau_{b|a} = \frac{n\sum_i \sum_j f_{ij}^2 / m_i - \sum_j n_j^2}{n^2 - \sum_j n_j^2} \tag{2.15}$$

Bizonytalansági (Uncertainty) együttható (Likelihood-arány teszten alapuló) sor/oszlop mutató, PRE elven mér:

$$UC = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} \log(m_i\, n_j / n f_{ij})}{\sum_{i=1}^{r} m_i \log(m_i / n)} \tag{2.16}$$

A két utóbbi mutatószám a G-K tau (2.15) és az UC (2.16) értéke aszimptotikusan konvergál az (r-1)(c-1) szabadsági fokú khi-négyzet eloszláshoz. Szélsőértékük:

- 0, ha az oszlop szerinti kategória ismeretében nem csökken a sor-variancia
- 1, ha az oszlop szerinti kategória ismeretében teljesen lecsökken a sor-variancia

B) Az ordinális változókra alkalmas mértékek nemcsak szorosságot, hanem irányt is mérnek, ezért értékük -1 és 1 között lehet.

Gamma (Goodman-Kruskal) $\gamma = (P-Q)/(P+Q)$ (2.17)

ahol $P = \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} S_{ij}$ és $Q = \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} D_{ij}$, továbbá

S az egyezően rendezett megfigyelések száma, azaz vagy i>k és j>l, vagy i<k és j<l teljesül egyszerre. Az f_{12} –höz képest (+) jelöli az ilyen cellákat az alábbi kis táblában.

D az eltérően rendezett párok száma, vagy i>k és j<*l*, vagy i<k és j>*l*, ezeket f_{12} –höz képest (-) jelöli az alábbi táblában:

	f_{12}		
-		+	+
-		+	+
-		+	+

A Somers-féle d mutatónak 3 változata[23] van, ezek az i=k és a j=*l* „egyezéseket” is figyelembe veszik.

[23] A Goodman-Kkruskal tau és a Somers d mutatók nevezői megegyeznek.

Ha az oszlopban van a függő változó: $d_{B/A}$= (P-Q)/D_r , ahol $D_r = n^2 - \sum_i m_i^2$

Ha a sorban van a függő változó: $d_{A/B}$= (P-Q)/D_c , ahol $D_c = n^2 - \sum_j n_j^2$

Ha szimmetrikus a két változó: $$d = \frac{P-Q}{1/2(D_r + D_c)} \tag{2.18}$$

A Kendall-féle tau-b a mértani átlaggal osztja az eltérést:

$$\tau_b = \frac{P-Q}{\sqrt{D_r D_c}} \tag{2.19}$$

Sztenderd hibája: ASE(τ_b)={(4n+10)/9(n^2 –n)}$^{1/2}$.

Kendall tau-c $\tau_c = \frac{q(P-Q)}{n^2(q-1)}$, ahol q=min (r,c) (2.20)

C) További mutatók:

Kappa: (Cohen mutatója) négyzetes táblára, csak a diagonális elemeket használja, pozitív értéke két döntéshozó véleménye közötti egyezést méri.

$$K = \frac{n\sum_i f_{ii} - \sum_i m_i n_i}{n^2 - \sum_i m_i n_i} \tag{2.21}$$

Kockázat (Risk): 2x2 táblára számolható, ha nincs üres cella. Az első oszlopba sorolás relatív kockázata ($f_{11}(f_{21}+f_{22})$)/(f_{21} (f_{11} +f_{12})) mellett a második oszlopba sorolás relatív kockázata is számolható, és a kettő hányadosaként az esélyhányadost R=(f_{11} f_{22} /f_{12} f_{21}) is becsli. Konfidencia-intervallumot is kapunk mindháromra. Az esélyhányadosra az alsó és felső határ:

$$R \cdot \exp(-z_{1-\alpha/2}\nu); R \cdot \exp(+z_{1-\alpha/2}\nu) \quad ahol \quad \nu = \left(\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}}\right)^{1/2}$$

McNemar teszt: csak négyzetes táblára alkalmazható. Ismételt mérésre a változást teszteli (before-after, initial-final hatások), a diagonálison kívüli elemekre épül: MC=f_{12}-f_{21} (2.22)

Nagy mintára $\chi^2 = \frac{(|f_{12} - f_{21}| - 1)^2}{f_{12} + f_{21}}$ és df=1

Cohran és Mantel-Haenszel statisztika: csak bináris változókra alkalmazható (dichotom factor, dichotom response) egy vagy több kontrolváltozó esetén. Ha

logisztikus regresszióban alkalmazzuk, akkor azt teszteli, hogy az oszlopváltozónak (kezelésnek) nincs hatása:

$$\log \frac{p_{ij}}{1-p_{ij}} = \mu + \beta_i + \tau_j \ ,$$

és a nullhipotézis szerint a j indexű τ oszlopváltozók megegyeznek.

- Korrelációs együtthatót is számolhatunk a kereszttábla elemzése során, amit kiválasztva egyúttal a Spearman-féle rangkorrelációt (és mindkettő t-tesztjét) is megkapjuk.
- Az eta mutató is kérhető, ha a nominális változónak, mint szempontnak a hatását mérjük az intervallum szinten mért változóra.

A kereszttábla elemzésben a tesztek nagy mintára alkalmazhatók, aszimptotikusan követik a feltételezett eloszlást. **Exact teszt** számolható az SPSS-ben binomiális, Poisson vagy hipergeometriai eloszlás feltételezése mellett, ha a megfigyelések száma nem több mint 20-30, és a változóknak háromnál nincs több kategóriájuk.

Végül grafikus ábrázolást is választhatunk a kereszttáblában vizsgált összefüggés szemléltetésére.

Kombinált oszlopdiagramot kérhetünk „Clustered bar chart" néven. A sorok számával megegyező beosztást látunk a vízszintes tengelyen, és mindegyiknél annyi oszlop szerepel, ahány kategóriája van az oszlopváltozónak. Az oszlopok magassága az együttes gyakoriság, ami a függőleges tengelyen jelenik meg.

2.2.3. 1. mintapélda

Az USA 242 felsőoktatási intézményét az iskola jellege valamint a tulajdonos alapján rendeztük, és a két ismérv közötti függetlenség hipotézisét teszteljük.

A kereszttáblában nincs üres cella, és teljesül az, hogy cellánként minimum 5 megfigyelést várunk. A cellákban a megfigyelt gyakoriságok mellett kérhetjük a várt gyakoriságok, a százalékok (sor-, oszlop-, teljes) és a reziduálisok (közönséges és sztenderdizált eltérések) feltüntetését.

Milyen tulajdonú? * iskola típusa Crosstabulation

Count

		iskola típusa		
		főiskola	egyetem	Total
Milyen tulajdonú?	állami	6	86	92
	magán	33	37	70
	egyházi	53	27	80
Total		92	150	242

A függetlenség hipotézisét minden valószínűségi szint mellett elvethetjük, hiszen a khi-négyzet tesztnél p<0,05 teljesül:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	68,264[a]	2	,000
Likelihood Ratio	77,976	2	,000
Linear-by-Linear Association	65,552	1	,000
N of Valid Cases	242		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 26,61.

Az eredmények között szereplő **lineáris asszociációs mérték** (linear-by-linear) akkor értelmezhető, ha a sor- és oszlopváltozók természetes módon rendezettek. Ekkor a sorokhoz u_i és az oszlopokhoz v_j tetszőleges számokat rendelve, és a gyakoriságokkal súlyozva: $LL = \sum\sum u_i v_j f_{ij}$ adódik. Az összeget sztenderdizálva khi-négyzet eloszlású statisztikát kapunk. A nullhipotézis azt mondja ki, hogy nincs sor-oszlop interakció. Példánkban a kategóriák rendezettsége nem teljesül, ezért nem értelmezzük.

A „tulajdonos" változó nominális, az „iskola típusa" ordinális. **Vegyes kapcsolatra** az SPSS-ben nincs külön mérőszám, ezért a nominális változókra javasolt mértéket választjuk. Egyes szakmákban kialakult hagyománya van annak, hogy melyik mérőszámot használják.

Ha azt gondoljuk, hogy a két változó között kölcsönös kapcsolat van, akkor a szimmetrikus mutatók közül kell választanuk. Összehasonlítani két kereszttáblát csak azonos asszociációs mérték alapján lehet. **A mérőszámok értéke általában különböző**. Példánkban a három szimmetrikus kapcsolat-mérték közül kettő egybeesik, mert az iskolatípus változónak két kategóriája van, és ezért a Cramer V-ben q-1=min(r,c)-1=1 kerül a nevezőbe. A (10) szerint számolva a szignifikancia szint lényegében nulla, közepesen szoros a kapcsolatot a két változó között.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,531	,000
	Cramer's V	,531	,000
	Contingency Coefficient	,469	,000
N of Valid Cases		242	

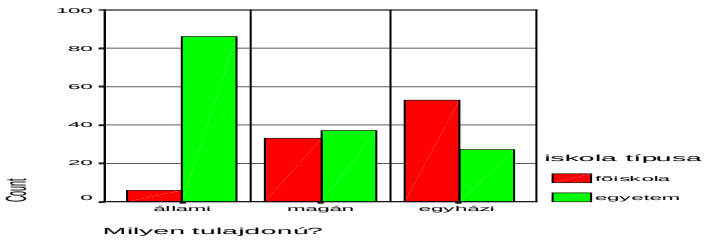
A kétféle oksági irányt feltételező mértékek közrefogják a szimmetrikus mértéket. Mindig szakmai megfontolás alapján választunk, nem a nagyobb számot értelmezzük! Ha nem szimmetrikus kapcsolatot tételezünk fel, akkor feltevéssel kell élnünk arra, hogy melyik a függő változó, és azt a sort kell értékelnünk az output táblában.

Gondolhatjuk azt, hogy a tulajdonos dönti el, hogy egyetemet vagy főiskolát alapít, tehát a típus a függő változó. De az az érvelés is helyes lehet, hogy a már működő iskolát veszi/kapja meg a tulajdonos, tehát fordított is lehet az oksági kapcsolat.

Directional Measures

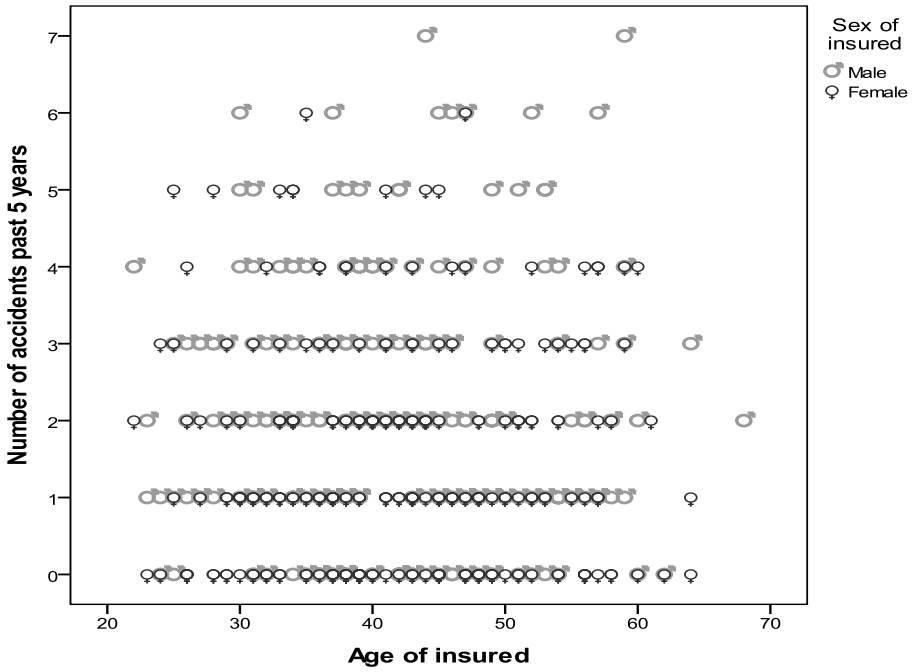
			Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,302	,055	4,889	,000
		Milyen tulajdonú? Dependent	,313	,042	6,655	,000
		iskola típusa Dependent	,283	,082	2,959	,003
	Goodman and Kruskal tau	Milyen tulajdonú? Dependent	,152	,029		,000
		iskola típusa Dependent	,282	,048		,000
	Uncertainty Coefficient	Symmetric	,183	,035	5,225	,000
		Milyen tulajdonú? Dependent	,147	,028	5,225	,000
		iskola típusa Dependent	,243	,046	5,225	,000

A kombinált oszlopdiagram szemlélteti, hogy az állam döntően egyetemeket finanszíroz, míg az egyházak inkább főiskolákat működtetnek.



2.2.4. 2. mintapélda

Ha van egy feltevésünk, például az, hogy a fiatalabb férfiak és a középkorú nők okoznak autóvezetés közben több balesetet (lásd a Pontdiagramot a 2. ábrán), akkor ennek teszteléséhez a kategorizált életkor változót és a nemet is figyelembe vesszük. Ismét a Program Files\SPSS\tutorial\sample files\autoaccidents.sav adatokat használjuk.



2. ábra: Az életkor, a nem és a balesetek száma

Többféle hipotézist fogalmazhatunk meg és tesztelhetünk, ha az autoaccident.sav állományhoz megnyitjuk az Analyze/Descriptive Statistics/Crosstabs –ot.

a) A balesetek száma és a nemek közötti függetlenségét vizsgáljuk először. A nominális változókra elérhető asszociációs mutatókat kérjük, hisz az ügyfél neme nominális változó.

Az első Pearson-féle khi-négyzet teszt értéke 16,584 (az empirikus szignifikancia p=0,02), tehát elvethetjük a függetlenséget, de a táblázat alján figyelmeztetést találunk: 4 cellában a várt gyakoriságok nem érik el az ötöt. Ez a 6 és 7 balesetet okozók alacsony száma miatt következett be. Ilyenkor az 5 vagy több baleset összevonása, az 5+ kategória kialakítása segít. A többi értéket változtatás nélkül átmásoljuk. Az új változó neve acc6, hogy emlékezzünk a kategóriák számára.

Sex of insured * acc6 Crosstabulation

Count

		acc6						
		0	1	2	3	4	5-6-7	Total
Sex of insured	Male	46	69	54	38	23	20	250
	Female	76	70	53	25	16	10	250
Total		122	139	107	63	39	30	500

A várt gyakoriságok már minden cellában kellő számban vannak, és a függetlenséget a szokásos 5%-os valószínűségi szinten elvethetjük, hisz $p=0,012<0,05$.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	14,666[a]	5	,012
Likelihood Ratio	14,833	5	,011
Linear-by-Linear Association	12,990	1	,000
N of Valid Cases	500		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 15,00.

Ha azt gondoljuk, hogy a vezető neme befolyásolja a balesetek számát, akkor az acc6 Dependent sorokat olvassuk. A Lambda mutató nem támasztja alá állításunkat, mert értéke statisztikailag nullának tekinthető. A vezető nemének ismeretéből alig 2%-nyi információt szerzünk a balesetek számára.

.**Directional Measures**

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,061	,032	1,814	,070
		Sex of insured Dependent	,124	,060	1,926	,054
		acc6 Dependent	,017	,033	,497	,619
	Goodman and Kruskal tau	Sex of insured Dependent	,029	,015		,012[c]
		acc6 Dependent	,006	,003		,009[c]
	Uncertainty Coefficient	Symmetric	,013	,006	1,950	,011[d]
		Sex of insured Dependent	,021	,011	1,950	,011[d]
		acc6 Dependent	,009	,005	1,950	,011[d]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

d. Likelihood ratio chi-square probability.

A vezetők nemét a balesetmentes-balesetes kettősséggel is összevethetjük. A függetlenséget elvetjük, mertekkor a khi-négyzet 9,758 (p=0,002), és a relatív kockázatot is mérjük.

Sex of insured * accid Crosstabulation

		accident		
		zero accident	one or more	Total
Sex of insured	Male	46	204	250
	Female	76	174	250
Total		122	378	500

Annak relatív kockázata, hogy egy ügyfelet balesetmentesnek minősítünk, 0,605. A balesetet okozó kategóriába sorolás relatív kockázata 1,172. Ezek hányadosa megadja az esélyhányadost (odds ratio), a 0,516-t, ami a gyakoriságokból közvetlenül is számolható: (46*174)/(76*204). Erre kapunk egy ½ körüli konfidencia intervallumot. Tehát a vezető neme a károkozásra nincs érdemi hatással.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Sex of insured (Male / Female)	,516	,340	,784
For cohort accid = zero accident	,605	,439	,835
For cohort accid = one or more accident	1,172	1,060	1,297
N of Valid Cases	500		

b) A balesetek száma és az életkor-kategóriák kapcsolatára készített kereszttáblában nincs elegendő bizonyíték a függetlenség hipotézisének elvetésére. Ezt állapítjuk meg akkor is, ha a bináris balesetváltozóra és a 10 évnyi hosszú életkor-kategóriákra számolunk. A khi-négyzet értéke 1,58 (p=0,812). A függetlenség elvetésekor nem értelmezzük az asszociációs mérőszámokat, hiszen azok értéke nem különbözik szignifikánsan a nullától.

c) A vezető neme változó rétegképző (Layer) lehet, amit beírva a két nemre és a teljes mintára is kereszttáblát számol a program. Külön tudunk tehát dönteni a férfiak és a nők csoportjában arról, hogy az életkor és a baleset okozása[24] között van-e kapcsolat.

Így a három táblára egyszerre látjuk, hogy a balesetmentes-balesetet okozó és az 5 életkor kategória közötti függetlenség hipotézisét egyik esetben sem vethetjük el. A szabadsági fok mindhárom esetben (5-1)(2-1)=4. Az 59 év feletti vezetők száma

[24] A biztosítók egy időszakban meglepve tapasztalták, hogy a 45-50 éves nők nevén levő autókra milyen sok kárbejelentés érkezik. Az ok természetesen nem a nők romló vezetési rutinja, hanem az, hogy éppen felnőtt, jogosítványt szerzett a fiú, aki az anyja kocsiját kéri kölcsön. (Azóta a biztosítás megkötésekor jelezni kell, ha több személy vezeti az autót.) A példa tanulsága, hogy nagyon óvatosan kell a kereszttáblában a kategória változókat megválasztani. Nem a tulajdonos, hanem a használó neme és életkora a fontos, ha ezt is rögzíti a biztosító adatbázisa.

kicsi, ezért a táblázat alján üzenet figyelmeztet, hogy a várt gyakoriság 5 alatt maradt.

Chi-Square Tests

Sex of insured		Value	df	Asymp. Sig. (2-sided)
Male	Pearson Chi-Square	2,880[a]	4	,578
	Likelihood Ratio	3,040	4	,551
	Linear-by-Linear Association	,000	1	,992
	N of Valid Cases	250		
Female	Pearson Chi-Square	2,606[b]	4	,626
	Likelihood Ratio	2,511	4	,643
	Linear-by-Linear Association	,000	1	,998
	N of Valid Cases	250		
Total	Pearson Chi-Square	1,580[c]	4	,812
	Likelihood Ratio	1,573	4	,814
	Linear-by-Linear Association	,002	1	,966
	N of Valid Cases	500		

a. 2 cells (20,0%) have expected count less than 5. The minimum expected count is 1,66.

b. 1 cells (10,0%) have expected count less than 5. The minimum expected count is 2,43.

c. 1 cells (10,0%) have expected count less than 5. The minimum expected count is 4,15.

3. Klaszterelemzés

A klaszterező eljárások csoportosítása

Az osztályozó eljárások családjának egyik ágába sorolható a klaszterelemzés, amely többféle módszer és konkrét eljárás összefoglaló neve. Alapgondolata az, hogy előre nem ismert besorolás esetében is feltárható a halmazon belül egymáshoz leginkább hasonló (közeli) „egyedek" csoportja. Egyed alatt érthetjük a megfigyelési egységet vagy a változót, mindkettőre végezhető osztályozás.

A klaszterező módszerek két fő csoportja:

- a hierarchikus osztályozás és
- a nemhierarchikus osztályozás.

A hierarchikus osztályozás két megközelítéssel végezhető.

Az **összevonó** (agglomeratív) hierarchikus eljárás kezdetben mind az *n* elemet külön osztálynak tekinti, majd lépésenként egy-egy összekapcsolást végez. Összesen (n-1) lépésben[25] elvégzi azt az összevonás-sorozatot, amely végül egyesít minden egyedet. Ez a folyamat grafikusan – két dimenzióban – megjeleníthető. Ha az adott lépésben már k csoport van, akkor a következő összekapcsolást maximum k(k-1)/2 távolság összehasonlításával lehet kiválasztani. A konkrét összevonás 7 eljárásváltozattal valósítható meg az SPSS-ben.

A **felosztó** (divizív) hierarchikus eljárás minden egyes lépésben – valamilyen döntési kritérium alapján – kettéosztja a megfigyeléseket, így az eljárás (2^{n-1}-1) felosztás megvizsgálása után fejeződik be. A magas lépésszám miatt ezt az eljárást a gyakorlatban nem alkalmazzák.

A nemhierarchikus osztályozás a témakör szakmai ismerete alapján előre adott k számú osztályra bontja a mintát. Az n számú elem k nem üres csoportba $\frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^n$ féleképpen sorolható be. A képlet alapján n=8 megfigyelést k=2 csoportba (1/2)(-2+28) = 127 változatban lehet besorolni.

Ha a struktúra feltárásának kezdetén a csoportok számát nem ismerjük, akkor minden $1 \le k \le n$ számra el kellene végezni a felosztást, hogy a k elfogadható értékét megtaláljuk. Nagyméretű feladatok esetében ez az út járhatatlan, ezért ilyenkor a $k \le \sqrt{n/2}$ hüvelykujj szabályt követjük. Hasznos lehet a hierarchikus klaszterezés

[25] Ha p számú változóra végzünk összevonást, akkor (p-1) lesz a lépések száma.

összevonó változatát elvégezve, struktúrafeltáró elemzést készítve „tájékozódnunk” a klaszterszámról, bár nagy elemszám esetében nem kapunk áttekinthető képet.

A következőkben a legismertebb, számítógépes algoritmussal is rendelkező klaszterező eljárásokat mutatjuk be. A témakör áttekintését segíti az elemzés döntési pontjainak előzetes áttekintése:

Ha az adatok előzetes csoportosítása nem ismert, akkor 3.1. fejezet szerint járhatunk el.

- A távolsági vagy hasonlósági mérőszámok közötti tájékozódást segíti a 3.1.1. alfejezet.
- Az összevonó eljárás kiválasztásakor a 3.1.2. alfejezet ad útmutatást.
- Ha a minta szerkezetét tanulmányozzuk, akkor 3.1.3. alfejezet segít.
- A számítógépes futtatás lépéseit a 3.4.1. alfejezet mutatja be.

Ha a megfigyelésekből képezhető klaszterek számára feltevéssel élünk, akkor a 3.2. fejezetet követhetjük.

- A számítógépes megvalósítás lépéseit a 3.4.2. alfejezet mutatja be.

3.1. Hierarchikus klaszterezés

A hierarchikus módszerek legfőbb sajátossága az, hogy a csoportosításhoz nem kell megadni a mintában létező (vagy feltételezett) csoportok számát.

Általában 3 lépést[26] hajtunk végre:

- Az induló adatokból[27] hasonlósági vagy távolság-mátrixot készíthetünk.
- Értelmezzük az egyedek és a csoportok egymáshoz való közelségét.
- Ábrázoljuk az összevonási folyamatot.

E három lépés során számos részdöntést hozzunk, amelyek következtében eltérő eredményeket kaphatunk. Az egyedek közti távolságot számos mérőszámmal mérhetjük, közülük például a mérési skála alapján választhatunk. A már egy klaszterbe sorolt egyedek távolságát a többi egyedtől (vagy klasztertől) származtatott távolsággal mérjük, amely szintén többféleképpen értelmezhető. Ezért fontos, hogy a lehetőségeket áttekintsük, és az adatrendszer sajátosságainak leginkább megfelelő távolságmértéket és összevonó eljárást megtaláljuk.

---

[26] A lépések megegyeznek akár eseteket, akár változókat osztályozunk. Ezért ezt a szempontot csak akkor említjük, ha szükséges.

[27] Az is előfordulhat, hogy ez a lépés kimarad, mert inputként már a távolsági vagy a hasonlósági mátrixot ismerjük.

3.1.1. Távolsági és hasonlósági mértékek

Az elemzés célja alapján választunk, hogy távolságot vagy hasonlóságot számolunk. De azt, hogy a két fő csoporton belül melyik mérőszámmal dolgozunk, az adatok mérési skálája alapján kell eldönteni. A részletes ismertetés előtt az 3.1. táblázatban összefoglaljuk az egyes mérési szintekre alkalmazható mutatók nevét vagy képletszámát.

3.1. táblázat: Mérési szintek szerinti mutatószámok képletei

Mérési szint / Mutató	Távolsági mutató képlete	Hasonlósági mutató képlete
Nominális vagy ordinális skálán mért változók	–	Khi-négyzet és Phi mutató (2. fejezet)
Intervallum vagy arány skálán mért változók	(3.1) – (3.4)	Pearson-korreláció, bezárt szög koszinusza (4. fejezet)
Bináris skálán mért változók	(3.5) – (3.10)	(3.11) – (3.14)

- **Intervallum skálán mért adatok között mért távolság**

Az SPSS alapértelmezésben a négyzetes euklideszi távolságot javasolja, amely az i és a k egyedek között (3.1) szerint számolható, ahol j index jelzi az egyedeket vagy a változókat:

$$d_{ik}^{\ 2} = \sum_j (x_{ij} - x_{kj})^2 \qquad (3.1)$$

A Csebisev metrika csak a legnagyobb eltérést méri: $d_{ik} = \max_j \left| x_{ij} - x_{kj} \right|$ (3.2)

Míg a city-block (vagy Manhattan) metrika összegzi az eltéréseket[28]:

$$d_{ik} = \sum_j \left| x_{ij} - x_{kj} \right| \qquad (3.3)$$

„Négyszer-négy”[29] távolság néven eltérő hatványkitevőt és gyököt választhatunk:

[28] A változók előzetes sztenderdizálása nagyon fontos azért, hogy ne különböző mértékegységben mért eltéréseket adjunk össze.

[29] A „customized” lefordítva „felöltöztetett” lenne. Mivel a p és az r 1-4 között változhat, ezért 4*4 mutatóként említjük.

$$d_{ik} = \left(\sum_j \left| x_{ij} - x_{kj} \right|^p \right)^{1/r}, \qquad (3.4)$$

ami p = r esetén megegyezik a Minkowski metrikával.

A hasonlóság mérésére a két vektor által bezárt szög koszinuszát és a Pearson-féle korrelációs együtthatót választhatjuk.

- **Nominális vagy ordinális skálán mért adatok (Counts)**

Alapértelmezés szerint a kereszttábláknál szokásos khi-négyzetet vagy a Phi-négyzetet kapjuk, amelyek esetekre is és változókra is számolhatók, és hasonlóságot

$$\chi^2(x,y) = \sum_i \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum_i \frac{(y_i - E(y_i))^2}{E(y_i)} \qquad \textit{és} \qquad PHI^2 = \chi^2 \Big/ \sqrt{n}$$

mérnek. A khi-négyzet mutató nagyon érzékeny a minta nagyságára, *n*-re.

A khi-négyzet képletében a várható érték a függetlenség feltételezése melletti gyakoriságként határozható meg x-re és y-ra.

- **Bináris változók**[30]

Mesterségesen is képezhetünk bináris változókat, ha csak a tulajdonsággal rendelkezés vagy nem rendelkezés a fontos. Azt, hogy két egyed (X és Y) mennyire hasonlít, a tulajdonságok együttes előfordulásának gyakoriságát tartalmazó kereszttáblából olvassuk ki. (3.2. táblázat)

3.2. táblázat: Együttes gyakoriságok

X / Y	(1)	(0)	Összesen
(1)	a	b	a+b
(0)	c	d	c+d
Összesen	a+c	b+d	a+b+c+d

Ha összesen p (=a+b+c+d) tulajdonság alapján hasonlítjuk össze X és Y egyedet, akkor b esetben csak X-re, és c esetben csak Y-ra voltak jellemzők a vizsgált ismérvek. Ezek felhasználásával számos távolságmérőszám képezhető, itt az SPSS

[30] Nincs általánosan ismert magyar neve egyik mértéknek sem, ezért itt is az angol elnevezés szerepel. Az SPSS 27 távolsági és hasonlósági mérőszámot kínál fel bináris változókra, ugyan mindre nem térünk ki, de a felsoroltakat klaszterezzük is.

által felajánlottak közül hatot mutatunk be. Egymással nem összehasonlíthatóak, mert a felső határuk különböző, bár mindegyiknek zérus[31] a minimuma.

Euklideszi: $d = \sqrt{b+c}$ (négyzete az alapértelmezés) (max: √p) (3.5)

Size difference: $d = \dfrac{(b-c)^2}{(a+b+c+d)^2}$ (max:1) (3.6)

Pattern difference $d = \dfrac{(bc)}{(a+b+c+d)^2}$, (max: ¼) (3.7)

Variance: $d = \dfrac{(b+c)}{4(a+b+c+d)}$, (max: ¼) (3.8)

Shape: $d = \dfrac{(a+b+c+d)(b+c)-(b-c)^2}{(a+b+c+d)^2}$, (max: 1) (3.9)

Lance-Williams: $d = \dfrac{(b+c)}{(2a+b+c)}$, (max: 1) (3.10)

A hasonlóság mérése sok bináris asszociációs mutatóval valósítható meg. Ezek csoportosíthatók aszerint, hogy a 0-0 értékpár (d gyakoriságú) előfordulását szerepeltetik-e a számlálóban és/vagy a nevezőben. A súlyozás szerint is vannak különböző mértékek: egyenlő súlyt vagy dupla súlyt kaphatnak a párok. A mutatók egy része 0 és 1 között mér, itt az 1 jelzi a maximális hasonlóságot. De vannak olyanok is, amelyek felső határa a végtelen.

Simple matching: $\dfrac{a+d}{a+b+c+d}$ (max:1) (3.11)

Jaccard: $\dfrac{a}{a+b+c}$ (max:1) (3.12)

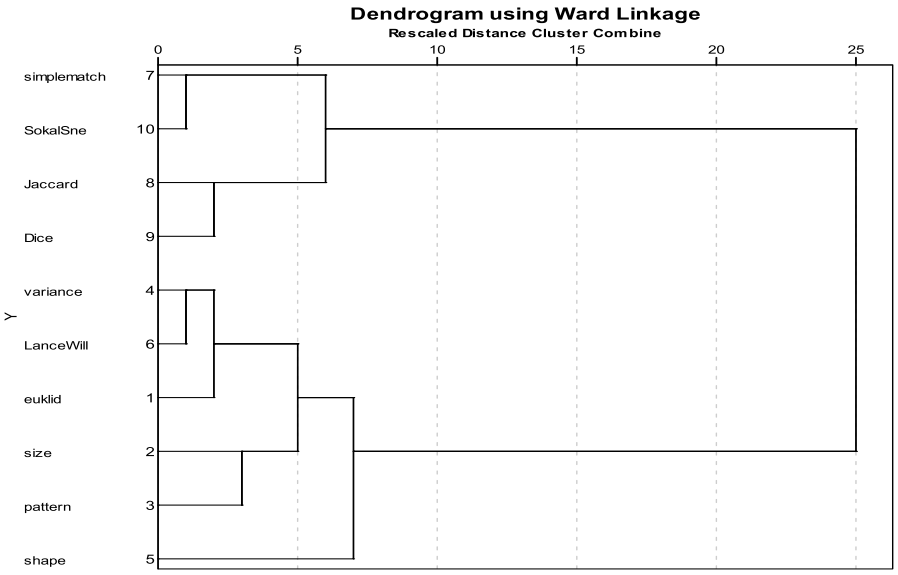
Dice: $\dfrac{2a}{2a+b+c}$ (max:1) (3.13)

[31] Zérus adódhat akkor is, ha b=c=0, vagyis tényleg nem különböznek, de úgy is, pl. (3.6)-ban, ha b=c, és egyik sem 0. Külön probléma az, hogy a d szám mit jelent. Attól, hogy egyformán nem rendelkeznek a vizsgált tulajdonságokkal, még nem biztos, hogy hasonlóak.

Sokal-Sneath 3. mutatója: $\frac{a+d}{b+c}$ (max: ∞) (3.14)

A bináris mutatók eltéréseit és egymáshoz viszonyított helyzetét a fejezet témaköréhez igazodva a hierarchikus klaszterezés Ward[32] elvű összevonó eljárásával készült ún. dendrogramon[33] szemléltetjük a 3.1. ábrán.

Az adattáblát a könyvhöz is csatoljuk, így a számításokat ellenőrizni lehet. Világosan elválik az első blokkban a négy hasonlósági mutató, majd egy klasztert alkot a hat távolsági mérőszám. Az egyes mutatócsoportokon belül is láthatunk tagozódást. A (3.9) képlettel megadható Shape mutató összetettsége miatt csak az utolsó előtti lépésben csatlakozik a távolság-mérőszámok csoportjához. Természetesen ez az eredmény is függ attól, hogy milyen adatok alapján és milyen eljárással hasonlítjuk össze a mutatókat. Itt bináris változókkal jellemeztük az egyes mérőszámok tartalmát, felépítését.



3.1. ábra: Bináris mutatók klaszterezése hasonlóságuk alapján

[32] A Ward elv lényegét a következő alfejezet ismerteti.
[33] Az ábra tulajdonságait a 3.1.3. alfejezet ismerteti.

3.1.2. Összevonó eljárások

Az SPSS-ben hét agglomeratív eljárás található, melyek lényegében hat megfontolás szerint mérik a csoportok közötti távolságot. Lance és Williams (1966) megmutatta, hogy e különbözőségek ellenére a klaszterek távolsága a (3.15) közös képlettel írható fel. A képletben szereplő:

$$D(IJ,K)=\alpha_I D(I,K)+\alpha_J D(J,K)+\beta D(I,J)+\gamma \left| D(I,K)-D(J,K) \right| \quad (3.15)$$

Az összevonás kezdetén *D(I,J)* két eredeti megfigyelés közötti minimális távolság. Az I és a J egyének vagy klaszterek összevonása már megtörtént, most a K (egyén vagy csoport) hozzákapcsolását vizsgáljuk. A további lépésekben az α, β, γ paraméterek, mint súlyok megválasztásával bármelyik összevonó eljárás elvégezhető. A 3.3. táblázatban az egyes hierarchikus összevonó eljárások és a távolság-paraméterek megfeleltetése látható.

3.3. táblázat: Távolságok súlyozása[34] Lance-Williams együtthatókkal

Eljárás	α_I	α_J	β	γ
1.Egyszerű lánc	1/2	1/2	0	-1/2
2.Teljes lánc	1/2	1/2	0	1/2
3.Átlagos lánc	$n_I/(n_I+n_J)$	$n_J/(n_I+n_J)$	0	0
4. Centroid	$n_I/(n_I+n_J)$	$n_J/(n_I+n_J)$	$-\alpha_I\alpha_J$	0
5. Medián	1/2	1/2	-1/4	0
6. Ward	$(n_I+n_K)/(n_I+n_J+n_K)$	$(n_J+n_K)/(n_I+n_J+n_K)$	$-n_K/(n_I+n_J+n_K)$	0

Ez a „közös gyökér” a hierarchikus eljárások egyik szép tulajdonsága, de ez okozza az alkalmazások során a legnagyobb nehézséget, mert az eltérő eljárások[35] eltérő felosztást, és így eltérő dendrogramot eredményeznek. Ezért több változatban célszerű elvégezni a klaszterezést. Így, ha a különböző eljárásokból egymással összhangban levő felosztások adódnak, akkor stabilabb a kapott felosztás. Mivel a hierarchikus módszereknél a korábban besorolt elemek áthelyezése nem valósítható meg, a kezdeti lépések döntő jelentőségűek.

Más szerzők (pl. Krzanowski (2000)) amellett érvelnek, hogy a csoportosítandó elemek természetét tanulmányozva előre kell módszert választani. Ezzel elkerülhető a sok fölösleges futtatás, valamint az, hogy az előzetes elvárásainknak legjobban megfelelő eredményt választjuk. Mindkét megközelítés megfontolandó, ezért a

[34] A súlyok az átlagos lánc, a centroid és a Ward eljárásánál a klaszterek tagszámától függnek

[35] Emlékeztetünk arra, hogy a sokféle hasonlósági és távolságmérték közötti választás lehetősége még további klaszter-kombinációkat eredményezhet.

módszerválasztás megkönnyítése érdekében tekintsük át részletesebben a klaszterező eljárások főbb jellemzőit.

Ha a klaszthereljárások matematikai tulajdonságait tekintjük, akkor fontos megjegyezni, hogy az egyedek közötti távolságok monoton transzformációjára csak az egyszerű lánc és a teljes lánc módszerek invariánsak[36].

A klaszterek geometriai alakja eltérő az egyes eljárásoknál. Az egyszerű lánc módszer jellemzője a lánchatás, vagyis bizonyos elemeket közbeeső elemek láncolata révén kapcsol össze. A közös klaszterbe kerüléshez elegendő az is, ha a csoport egyetlen tagjához hasonlít a vizsgált egyed, így az eljárás térösszehúzó hatású. A lánchatás érvényes a medián módszernél is, ahol az utoljára kapcsolódó pontnak döntő hatása lehet a klaszterezés további menetére.

Viszonylag zárt, „gömbölyű" klasztereket kapunk, ha a teljes lánc, az átlagos lánc vagy a centroid módszerekkel végezzük az osztályozást. Ekkor egy-egy klaszter elemei egymáshoz nagyon közeliek. A legtávolabbi szomszéd elv alapján inkább új klaszterek képződnek egy-egy következő lépésben, nem a meglevők csoportokhoz kapcsolódnak az újabb egyedek. Ezt tértágító hatásnak nevezi a szakirodalom, míg az átlagos lánc elv térkonzerváló hatásúnak tekinthető. A teljes lánc módszer egyenlő átmérőjű, a Ward módszer pedig egyenlő elemszámú klaszterek kialakítására törekszik.

Ha az adatok klasztereződése nem egyértelmű, akkor a centroid és a medián módszer alkalmazása során problémát okozhat az inverzió előfordulása. Ekkor az összevonás későbbi lépésében megtörik a monoton növekedés, és kisebb távolság adódik, mint a korábbi szintek klaszterei között mért legkisebb távolság.

További – bár a klaszterezésben nem lényegi – problémát okoz az, ha a távolsági vagy a hasonlósági mátrixban megegyező elemek vannak. Ekkor – különösen az összevonás elején– többféle felosztás adódhat, és ez az értelmezést nehezíti.

3.1.3. Dendrogramok értékelése, összehasonlítása

A hierarchikus összevonó eljárások közös tulajdonsága, hogy az n számú egyedet (n-1) lépésben összevonják egyetlen egy csoportba. Az összevonási folyamat ábrázolása dendrogramon történik. Ez egy kétdimenziós ábra, melynek speciális szerkezete van. Az egyik tengelyen az összevont elemeket látjuk, a másikon pedig azt a távolságértéket, amelynél az összevonás megtörtént. Kezdetben (0 távolsági szinten) minden megfigyelés egyedül van, a végén (általában 25 maximális távolságértékre átskálázva) már minden pont egyetlen csoportban van. Ha többféle távolságmértékkel és/vagy eltérő eljárásokkal is elvégezzük a klaszterezést, akkor nagy valószínűséggel különböző dendrogramokat kapunk, amelyek hasonlóságát meg kell vizsgálni.

---

[36] Például a távolságok logaritmusát véve eltérő felosztás és eltérő dendrogram adódik, ha nem a legközelebbi vagy a legtávolabbi szomszéd elvet követjük.

Az összevonási folyamatot tükrözi maga a dendrogram, de további elemzést igényel a megfelelő klaszterszám leolvasása. Ehhez az összevonás rendjét és távolságszintjeit mutató táblázat ad információt.

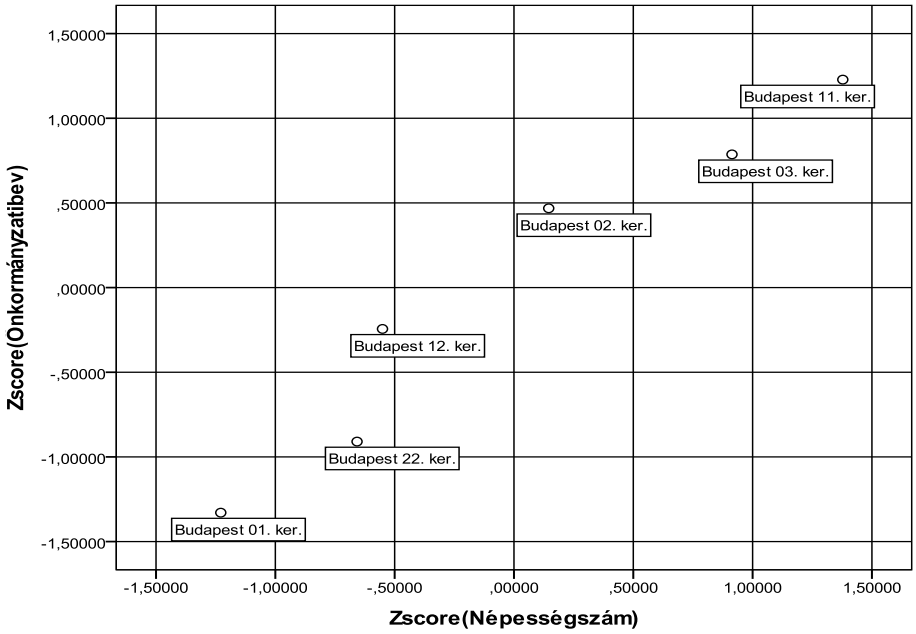
- Két dendrogramot összehasonlíthatunk úgy is, hogy az összekapcsolódásokat vetjük egybe. Az *n(n-1)/2* pontpárra meghatározzuk, hogy az egyes dendrogramokban hányadik összekapcsolódás után kerültek egy csoportba, és a két összevonási adatsorra korrelációt számítunk.
- Elemzői szokás a 40%-os távolságszint (10-es rescaled distance) alatti csoportok számát leolvasni, és ezt elmenteni. Így két összevonó eljárás eredménye kereszttáblázatban is összevethető. Mivel a klaszter-azonosítók nominális változók, a 2. fejezetben bemutatott asszociációs mérőszámokkal mérhetjük a felosztások hasonlóságát.

Fontos azonban megjegyezni, hogy bármilyen gondosan választottunk távolságmértéket és klaszterező eljárást, bárhogyan hasonlítottuk össze a dendrogramokat, nem kapunk végleges választ arra a kérdésre, hogy hány csoportba sorolható a vizsgált adathalmaz. A struktúrafeltárás ezen eljárása csak exploratív célra alkalmas, az ábra alapján hipotézis fogalmazható meg a mintabeli csoportok számára. Továbbá hatékonyan segíti a dendrogram az extrém értékek feltárását, hiszen a magas távolság szinten és/vagy az összekapcsolódás későbbi szakaszában látható megfigyelések egyedi jellege szembetűnő. Ismét emlékeztetjük az olvasót arra, hogy a változókat is lehet klaszterezni, és az összekapcsolódásukat dendrogramon ábrázolni. Ekkor a változó-fürtökből a dimenziócsökkentés lehetséges mértékéről kapunk statisztikai képet.

Ha szakmai ismeretek alapján előre tudjuk, hogy hány csoport van a vizsgált mintában, akkor ne alkalmazzuk az agglomeratív eljárásokat, mert azok nem alkalmasak egy várt felosztás reprodukálására. Ilyen feladatok megoldására választhatjuk a nem-hierarchikus klaszterezést, vagy a konkrét céltól függően számos más sokváltozós statisztikai eljárást.

3.1.4. Az összevonó algoritmus lépéseinek követése egy mintapéldán

Hat budai kerületet mutatunk be két változó terében (3.2. ábra), hogy egyszerűen, akár kézi számolással is ellenőrizni tudjuk a klaszterezés folyamatát. Az ábráról leolvasható, hogy három kerület (II., III. és XI.) mindkét változó szerint átlag feletti értékekkel rendelkezik, míg a másik három átlag alatti értékeket ér el.



3.2. ábra: Hat budai kerület két – sztenderdizált – változó terében

Mivel a számítások csak az egyszerű lánc és a teljes lánc esetén követhetők szemmel is, ez utóbbi eljárást mutatjuk be.

Euklideszi távolságok négyzeit számolva a hat kerület között, a távolságmátrixban kiemelve láthatók az összevonáskor számolt távolságok. (3.4. táblázat)

Az első lépés minden eljárásnál azonos, a minimális távolságú két pont kerül összevonásra. Ezt példánkban a III. és a XI. kerület között látjuk: 0, 411.

A második lépés során azt nézzük, hogy III. és XI. kerület együtt milyen távol van a többiektől. Most a legnagyobb távolságok - hiszen ez a legtávolabbi szomszéd elv néven is ismert - legkisebbikét keressük: ez a II. kerület lenne: 2,096 távolságra a XI.-től (mivel 0,691 távolságot ért el a III. –tól). De mégsem itt történik összevonás, hiszen a XII. és a XXII. kerület közötti távolság kisebb: 0,454.

3.4. táblázat: A hat kerületre páronként mért euklideszi távolságok négyzete

Proximity Matrix

Case	Squared Euclidean Distance					
	Budapest 01. ker.	Budapest 22. ker.	Budapest 12. ker.	Budapest 02. ker.	Budapest 03. ker.	Budapest 11. ker.
Budapest 01. ker.	,000	,502	1,637	5,119	9,066	13,335
Budapest 22. ker.	,502	,000	,454	2,543	5,345	8,713
Budapest 12. ker.	**1,637**	**,454**	,000	,993	3,207	5,888
Budapest 02. ker.	5,119	2,543	,993	,000	,691	2,096
Budapest 03. ker.	9,066	5,345	3,207	,691	,000	,411
Budapest 11. ker.	**13,335**	8,713	5,888	**2,096**	**,411**	,000

This is a dissimilarity matrix

A harmadik lépésben arról kell döntenünk, hogy a már meglévő két klaszterünk (2-2 elemmel) milyen távol van egymástól és a további két egyedüli kerülettől. Itt a következő számok legkisebbikét választjuk:

- (III+XI) – II: 2,096
- (III+XI) – I: 13,335
- (III+XI) – (XII+XXII): 8,713
- (XII+XXII) –II: 2,543
- (XII+XXII) –I: 1,637

A negyedik lépésben ismét a két klaszterünk és a még egyedül álló II. kerület közötti maximális távolságokat vesszük szemügyre, de a legkisebb távolságot választjuk:

- (III+XI) – II: 2,096
- (XII+XXII+I) –II: 2,543

Az ötödik lépés az utolsó, mivel hat kerület van a példában. Az eddigi lépések miatt itt már csak a két klaszter közötti távolság meghatározása maradt hátra. Nem volt extrém helyzetű kerület, amelyik eddig nem kapcsolódott sehová.

- (III+XI+II) –(XII+XXII+I): 13,335

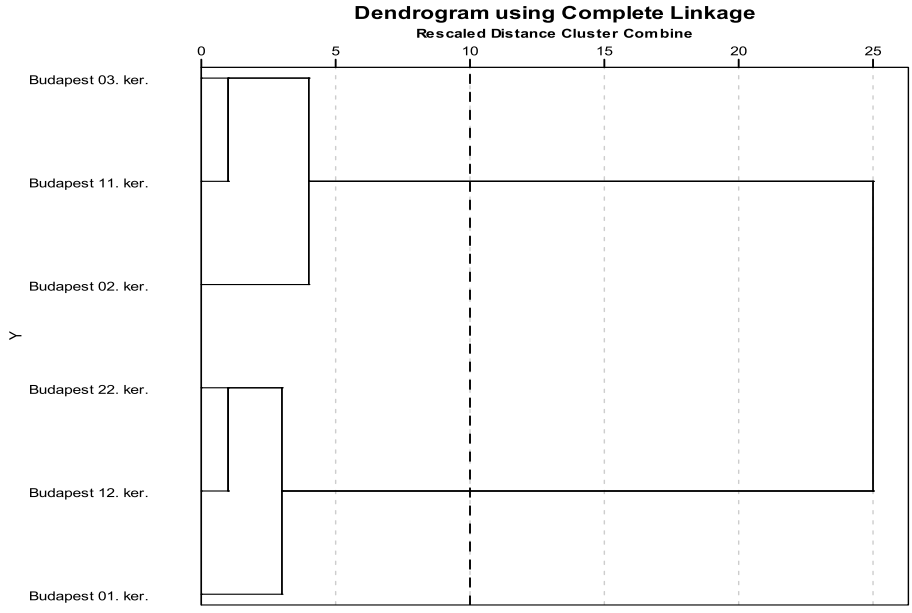
Az összevonás menetét a 3.5. táblázat és a 3.3. ábra is mutatja. Mivel az utolsó lépésben nagyot nő a klaszterek közötti belső távolság, érdemes két klasztert megkülönböztetni.

3.5. táblázat: A hat kerület összevonása 5 lépésben

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	11	,411	0	0	4
2	22	12	,454	0	0	3
3	1	12	1,637	0	2	5
4	3	2	2,096	0	1	5
5	1	2	13,335	3	4	0

3.3. ábra: A hat kerület kapcsolódása alapján két klaszterbe sorolható



3.2. Nem-hierarchikus klaszterezés

A nem-hierarchikus módszerek közül a leggyakrabban alkalmazott – és a hierarchikus klaszterezéshez a leghasonlóbbak – a diszjunkt klasztereket előállító **partícionáló módszerek**. A különböző eljárások általános menete a következő:

- a kezdeti klaszterek kialakítása, és az egyedek[37] szétosztása az euklideszi távolság[38] szerinti legközelebbi kezdő klaszterbe,
- új klaszterközéppontok számítása,
- az egyedek átsorolása a legközelebbi középponthoz.

Az iteráció, a klaszterek közötti mozgás addig folytatódik, amíg változnak a középpontok.

Az első és a második lépés végrehajtása többféleképpen történhet, ezért több eljárásváltozat ismert.

A kezdeti klaszterek kialakítását a csoportok k számának a megadásával kezdjük. A megfelelő k megválasztása szakmai tapasztalaton vagy korábbi statisztikai elemzésen (pl. hierarchikus klaszterezésen) alapulhat. Az SPSS-ben a MacQueen féle **k-középpontú klaszterezés**[39] végezhető.

A k-középpontú klaszterezés értelmezése két fő kérdést vet fel.

1. A csoportszám megfelelő-e? Az egyedek arányos szétosztása a klaszterek között nem követelmény, de a nagy aránytalanság fontos információt hordoz. Az egyelemű klaszterek a kilógó, a többiektől nagyon eltérő tulajdonságú egyedek létére vagy túl magas csoportszámra figyelmeztetnek. A nagy elemszám pedig azt jelzi, hogy érdemes a csoportszám növelésével megismételni a klaszterezést.

   A klaszterközéppontok és a köztük levő euklideszi távolságok előállítása is segíti az értelmezést és a klaszterek megkülönböztetését. Ezt kiegészíthetjük azzal, hogy az egyes egyedeknek a saját klaszterük középpontjától mért távolságát is meghatározzuk. A távolságok alapján dönthetünk az egyes csoportok szétvágásáról vagy összevonásáról, azaz a k növeléséről vagy csökkentéséről.

2. A változók szignifikáns szerepet játszanak-e az osztályozásban? Az egyedek osztályozásán túl vizsgálható az is, hogy a figyelembe vett p változó mindegyike jelentős szerepet játszott-e a klaszterek

---

[37] Itt csak a megfigyelések klaszterekbe sorolása lehetséges. A változók csoportosítása nem választható.
[38] Ebben a „Quick-cluster"-nek is nevezett eljárásban távolságmérték sem választható.
[39] A középpontok változása itt követhető:
http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

megkülönböztetésében. Az egyes klaszterek varianciáit kiszámolva a csoportok alakját hasonlíthatjuk össze, mivel az azonos variancia-kovariancia mátrix azonos alakot jelez. A szóráselemzés (F-próba)[40] segítségével kiválaszthatjuk a csoportokat elkülönítő változókat, és így akár dimenziócsökkentést is végrehajthatunk a következő lépésben.

Ezekre az értelmezési kérdésekre részben választ kaphatunk, ha elkészítjük a klaszterkönyök ábrázoláshoz szükséges számításokat:

a) Először k=2 beállítással klaszterelemzést készítünk, és a szórásfelbontó (ANOVA) táblázatban ellenőrizzük a változók megkülönböztető erejét.

i. Ha a változóra vonatkozó parciális F statisztika „alacsony", azaz az empirikus szignifikancia szint meghaladja a 0,05-t, akkor a változó elhagyásával megismételjük a futtatást.

ii. Ha minden változó megkülönböztető ereje elégséges, azaz az empirikus szignifikancia szintek kisebbek, mint 0,05-t, akkor elmentjük a klaszterazonosítókat.

iii. Az elmentett klaszterazonosítókat kategóriaképző változóként használva a szignifikáns változókra szórásfelbontást végzünk. Az ANOVA táblázatból rendre összegezzük a változókra számolt külső eltérések négyzetösszegét, majd a teljes eltérések négyzetösszegét, és a két összeg hányadosát képezzük. Így megkapjuk a klaszterezéssel megmagyarázható eltérések hányadát.

b) Elvégezzük k=3,4,5…-re az előző lépéssorozatot. A maximális lépésszám/klaszterszám egy hüvelykujj szabály[41] szerint a mintaméret (n) felének a gyöke, azaz $k \leq \sqrt{n/2}$.

c) Az ANOVA táblázatból k=2,3,4….re képzett hányadosokat ábrázoljuk, és megállapítjuk, hogy a k szám mentén meddig emelkednek[42] „határozottan" a megmagyarázott eltérések.

3.3. A klaszterelemzés eredményének értékelése

A klaszterező eljárások nagyon népszerűek, sokféle területen alkalmazzák az ismeretlen adatstruktúrák feltárására. Ennek részben az az oka, hogy sem a hierarchikus, sem a nemhierarchikus klaszterezéshez nem tartoznak matematikai előfeltételek. Ugyanakkor nem rendelhető hozzá célfüggvény sem, amivel az

[40] Csak leíró és nem tesztként való alkalmazásról van szó, mert a matematikai előfeltételek (normális eloszlás és azonos csoport-varianciák) teljesülését nem vizsgáljuk.

[41] Ezt a gyakorlati szabályt felül kell bírálnunk akkor, ha sok egyedi megfigyelésünk van, amelyek 1-1 elemű klasztereket alkotnak.

[42] Ha a monoton növekedő értékek sorában a növekedés lelassul, akkor nem érdemes több klasztert képezni.

osztályozás jósága mérhető lenne. Ezért, mielőtt a számítógépes megvalósításra térünk, összefoglaljuk a klaszterezéssel kapcsolatos legfontosabb megállapításokat és követelményeket, amelyek támpontot jelenthetnek a kapott eredmények értékelésében.

- Nyilvánvaló kívánalom, hogy a klaszterezés eredménye független legyen a megfigyelések sorrendjétől.

Ezt a követelményt nem teljesíti az SPSS Quick-cluster eljárása. A Kerületek2010.sav megadott adatállományon elvégezhető az ellenőrzés. Ha k=3 beállítással sztenderdizált változók terében klaszterezünk, akkor más és más eredményt kapunk, ha az abc-ben felsorolt megfigyeléseket klaszterezzük, vagy ha az adatállomány változói közül bármelyik szerint növekvő sorrendbe rendezzük a klaszterezés előtt az adatokat. Eltéréseket tapasztalunk a három kezdőpontban, a magpont megválasztása tehát érzékeny az adatok sorrendjére. De eltérő a végső felosztás és a klaszterek elemszáma is!

- Jól definiáltak legyenek a klaszterek abban az értelemben, hogy azonos megfigyelt adatokból azonos felosztást kapjunk. Ha vannak egyenlő távolság illetve hasonlósági értékek, akkor az eljárás önkényesen választ közülük, és emiatt ez a tulajdonság több eljárásnál nem teljesül.

- A folytonosság követelménye az, hogy az adatokban bekövetkező kis változások kis változást eredményezzenek a felosztásban.

- A stabilitás követelménye azt jelenti, hogy ha egy egyedet elveszünk vagy hozzáadunk a megfigyelésekhez, akkor az osztályozásban nagyon kis változás következzen be. Ez egy láncban összekötő kapcsot képező pont esetében nem teljesül. A stabilitási követelmény részének tekinthető az az elvárás is, hogy ha egy klaszter minden egyedét (hierarchikus esetben a dendrogram egy ágát) kihagyjuk, akkor a többi elem tagozódása invariáns legyen erre a változtatásra.

- Gyakori követelmény, hogy az osztályozás eredménye invariáns legyen a különbözőségek monoton transzformációjára. Itt említjük meg az adatok lineáris transzformációjára való invariancia követelményét is, amely például a sztenderdizált adatok használatát teszi lehetővé. Ha a vektorok hajlásszögének koszinuszából számítunk távolságot, akkor a pontok közötti távolság nem arányosan változik.

- A klaszterek érvényessége (validitása) négy kritérium alapján vizsgálható. *Külső* követelményként értelmezhető az, ha ismert csoportokba tartozó egyedekből veszünk mintát, és arra végezzük el a klaszterezést. *Belső* követelménynek tekinthetők azok a mutatók, amelyekkel az eredeti és a származtatott távolságok illeszkedését mérjük. Harmadik megközelítést jelent a *megismételhetőség* kritériuma, amelynek lényege a kettéosztott megfigyelések klaszterezése és a felosztások összevetése. A klaszterek érvényességének *relatív* kritériuma az adatmátrix több eljárás szerinti

klaszterezését, és a felosztások közötti egyezés mérését fogalmazza meg, de csak jól elkülönülő és gömb alakú struktúrák esetében tekinthetjük az egyező felosztásokat úgy, mint amelyek a természetes csoportok létét igazolják.

- A robusztusság követelménye a kilógó pontok hatásának csökkentését jelenti. Ha több nem tipikus, „távoli" pont van a mintában, akkor ezek jelentősen befolyásolhatják a felosztást olyan eljárások esetében, amelyek a belső eltérés-négyzetösszeget minimalizálják. Ilyenkor a csoportokon belüli azonos kovariancia-struktúra feltevése téves lehet, pedig az optimalizáló eljárások csak azonos alakú csoportok feltárására alkalmasak.

A klaszterelemző módszerek és a számítógépes eljárásváltozataik alkalmazásával kapott csoportosítások értelmezése és értékelése nagy szakmai felkészültséget és körültekintést igényel. Érdemes más sokváltozós módszereket, például sokdimenziós skálázást (8. fejezet) és diszkriminancia analízist (7. fejezet) is végezni, hogy a minta szerkezetéről megbízható megállapításokat fogalmazhassunk meg.

3.4. A megvalósítás lépései az SPSS-ben

Az **ANALYZE/CLASSIFY** úton elindulva a hierarchikus és a nem-hierarchikus módszerek közül kell először választanunk. A struktúrafeltárás logikája miatt a hierarchikus klaszterezés eljárásaival kezdjük a futtatást.

3.4.1. Hierarchikus klaszterezés

Először azokat a változókat kell kiválasztani, amelyeket csoportosítunk, vagy amelyek terében csoportosítjuk a megfigyeléseket. A LABEL-ben cimkét, azonosítót rendelhetünk a megfigyelésekhez.

Ezt követően 4 parancsgomb alatt tárulnak fel a választási lehetőségek.

1) STATISTICS/Statisztikák:

i. Az összevonás menetét mutatja az „Agglomeration schedule". Ha kérjük, akkor látható, hogy az összekapcsoláskor mennyi volt az egyedek közötti távolság. Ebből észrevehető az inverzió fellépése.

ii. Az induló távolsági vagy hasonlósági mátrixot „Proximity matrix" néven láthatjuk.

iii. Ha van elképzelésünk a belső tagozódásról, akkor a „Solution"-ben adhatjuk meg a konkrét számot. Beírható egyetlen szám: „Single"(=k), vagy egy tartomány: „Range" (2 és n-1 között), de üresen hagyva is elkészül a klaszterezés.

2) PLOTS/Ábrák:

A kapcsolódás szintjét és menetét mutató dendrogram kérhető[43] ábraként. Az ábra csak kisebb feladatokra látványos, 50-nél több megfigyelésre egy képernyőn nem tekinthető át.

3) METHODS/Módszerek: Itt 7 eljárásból választhatunk, és további fontos beállításokat tehetünk meg.

3a) Az eljárások

- Átlagos lánc a csoportok között[44] (ez az alapértelmezés az SPSS-ben)
- Átlagos lánc a csoportokon belül[45]
- Legközelebbi szomszéd vagy egyszerű lánc
- Legtávolabbi szomszéd vagy teljes lánc
- Centroid eljárás
- Medián eljárás
- Ward eljárása

3b) Távolsági vagy hasonlósági mérték választása

Itt nyílik mód a mérési skála figyelembe vételével a távolsági vagy a hasonlósági mértéket megjelölésére, és a különböző mértékegységek miatt indokolt sztenderdizálásra:

- Measure: Interval, Counts, Binary
- Standardize: 7 féleképpen szűrhető ki a mértékegység.

4) SAVE/Mentés: Elmenthetjük azt az egy vagy többféle felosztást, amit az induláskor az 1) lépés szerint iii.-ben megadtunk.

3.4.2. Nem-hierarchikus klaszterezés, k-középpontú eljárás

Ekkor a klaszterek számát (k) szakmai ismeretek vagy a hierarchikus klaszterek ábrája alapján előre meg kell adni.

A futtatás beállítása:

1. Változók kiválasztása
2. Label: megnevezések feltüntetése
3. Number of clusters: klaszterek száma (default=2)

---

[43] Icicle nevű diagramot is kaphatunk, de a képernyőn és nyomtatásban is áttekinthetőbb a dendrogram.
[44] Az összevonandó n és m elemű csoportokra nxm távolság átlagát számolja.
[45] Az n és az m elemű csoportok távolságainak átlagát az elemek egyesítése után (n+m)(n+m) elemre számolja.

4. Method/Módszer kétféle lehet:

    a. „Iterate and classify”= iteráció során a besorolt elemekre új klaszterközéppontot számol, újra besorolja a mintaelemeket

    b. „Classify only”: a kezdeti középpontokhoz való közelség szerint szétosztja a mintát, nem keres új magpontokat.

5. Iterate/Iteráció: Ha kérünk iterációt, azaz a 4.a. szerint haladunk, akkor még további két lehetőséget kínál fel az SPSS. Itt választható a folyton változó átlag: „Use running mean”

    a. Default = nem kérjük. Ekkor az összes elem szétosztása után számol klaszter középpontokat.

    b. Ha kérjük, akkor minden egyes elem besorolása után kiszámolja a klaszterek centrumait, mielőtt a további elemek osztályozására sor kerül.

6. Save/Mentés: „Cluster membership” = a klaszter azonosító számokat és „Distance from cluster center” = a klaszterközépponttól mért távolságokat hozzárendeli minden egyes megfigyeléshez.

7. Options/Lehetőségek: Itt további fontos statisztikákat kapunk.

    a. A kezdeti (Initial) klaszter-középpontokat felsorolja.

    b. Kérésre megkapjuk változónként a klaszterek közötti és a klaszteren belüli eltérésnégyzetösszegek hányadosát is tartalmazó ANOVA táblát az F-teszttel. A magas F érték (alacsony szignifikancia szint mellett) parciálisan jelzi az egyes változók megkülönböztető erejét. Itt az F-próbát nem egy nullhipotézis ellenőrzésére használjuk. (Nem úgy értelmezzük, mint a szóráselemzésnél, ahol a nullhipotézis az lenne, hogy a csoportátlagok között nincs különbség.)

    c. Minden elemre kiírathatjuk a képernyőre annak a klaszternek a számát, ahová besorolást nyert.

    d. Megkapjuk a monitoron az euklideszi távolságot minden megfigyelés és a saját középpontja között, továbbá a középpontok között is.

3.5. Települések klaszterezése

E fejezet célja az elméleti tudás elmélyítése és tapasztalatszerzés a gyakorlati megvalósításban. Ezért a könyvhöz tartozó adatállományok közül az 50 település (23 budapesti kerület és 27 környező település) 2010-es adatait használva a számítások elvégzése után válaszoljon a következő kérdésekre. A saját eredményeit vesse össze a közölt megoldással.

1) **kérdés**: Hány csoportot/dimenziót alkotnak a változók?

A megoldás lépései:

A – sztenderdizált – változókat hierarchikus klaszterezéssel vizsgáljuk, több dendrogramot készítünk. Több – intervallum skálára alkalmas távolságmérőszám kiválasztása is indokolt. A bináris változókat külön kell elemezni, hiszen egyidejűleg nem lehet kétféle távolságmértéket választani.

2) **kérdés**: Hogyan tagolódnak a települések? Valóban elválik egymástól a 23 kerület és a többi Budapest környéki település? Milyen klaszterszámot érdemes feltételezni?

A megoldás lépései:

A – sztenderdizált – változók terében hierarchikus klaszterezéssel vizsgáljuk a településeket, ismét több eljárást alkalmazunk, több dendrogramot készítünk. Több – intervallum skálára alkalmas – távolságmérőszám kiválasztása is indokolt.

3) kérdés: Ha k=2 beállítással készít k-közép klaszterezést, akkor a település típussal azonosnak tekinthető felosztás adódik?

A megoldás lépései:

Az előzetesen – sztenderdizált – változók terében k=2 klaszterezéssel besoroljuk a településeket. Megvizsgáljuk az ANOVA táblázatbeli F-teszt és p szignifikancia szint alapján, hogy minden változónak van-e megkülönböztető ereje. A nem-szignifikáns változókat elhagyva megismételjük a klaszterezést, és elmentjük a klaszter-azonosítókat. Végül kereszttáblázatban összevetjük a település jellege és a klaszterazonosítók alapján kapott besorolást.

4) kérdés: Hány klasztert érdemes megkülönböztetni?

A megoldás lépései:

Klaszterkönyök keresése a 3.2. alfejezetben leírtak szerint.

<u>Az eredmények részletes bemutatása</u>

1) kérdés eredményei: Hány csoportot/dimenziót alkot a 16 változó?

A leíró statisztikák 3.6. táblázatából[46] látható, hogy a relatív szórás (Szórás/átlag) sehol sem éri el a kettőt, tehát kilógó, nagyon extrém értéket mutató települések nincsenek. A változók nagy része pozitív ferdeségű, csak 5 változó tekinthető szimmetrikus eloszlásúnak. Három változó erősen csúcsos, a többi alakja nem szignifikánsan tér el a haranggörbétől.

[46] Helytakarékosság miatt töröltük a 3. táblából, hogy összesen 50 megfigyelésből számoltuk minden változó statisztikai mutatóit, egyiknél sincs hiányzó érték.

3.6. táblázat: Leíró statisztikai mutatók értékei

	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
					Statistic	Std. Error	Statistic	Std. Error
Terület	209,00	6192,00	2773,2400	1626,80908	,265	,337	-,611	,662
Népességszám	1830,00	144992,00	43987,3800	37985,55365	,851	,337	-,165	,662
Odavándorlás	85,00	9866,00	1661,7600	1669,25394	2,804	,337	11,379	,662
Elvándorlás	73,00	4301,00	1269,5000	979,83421	1,018	,337	,885	,662
Állandóodavándorlás	38,00	3338,00	846,0200	689,12321	1,590	,337	3,064	,662
Állandóelvándorlás	39,00	2370,00	729,7200	545,24479	,927	,337	,620	,662
Önkormányzatibev	454454,0	29214720,0	10362302,30	8173829,71762	,417	,337	-1,002	,662
	0	0	00					
Vendéglátóhely	4,00	838,00	249,9200	245,67606	1,093	,337	,012	,662
Lakásállomány	582,00	75846,00	21293,0200	20665,54774	1,082	,337	,443	,662
Épitettlakások	,00	1178,00	170,3000	245,57581	2,542	,337	6,765	,662
Álláskeresők	40,00	3380,00	1130,1800	1013,86467	,787	,337	-,685	,662
Odavanperfo	,0212	,0705	,043197	,0152623	,296	,337	-1,259	,662
Elvanperfo	,0179	,0596	,034468	,0109296	,415	,337	-1,015	,662
ÁllElvanperfo	,0103	,0373	,020327	,0074053	,862	,337	-,220	,662
Állodavanperfo	,01	,06	,0254	,01245	,809	,337	-,160	,662
Álláskeresőkaránya	,01	,04	,0252	,00645	,391	,337	,333	,662

Az előkészítő lépés, a változók sztenderdizálása után is több döntési pontunk van.

a) A változókat a számítások elvégzése előtt és a hierarchikus klaszterezésen belül is sztenderdizálhatjuk.

> Ez csak akkor változtatja meg az eredményeket, ha vannak hiányzó adatok. Az előzetes sztenderdizálásban minden változóra felhasználjuk az összes elérhető adatot, azaz különböző megfigyelésszám lehetséges. Míg a „belső" sztenderdizálás során a „közös", hiánytalan adatállomány kerül felhasználásra.

b) Az elemzésben szereplő változók között távolságot és hasonlóságot is mérhetünk. Ettől függően eltérő összevonási adatokat kapunk. Az Agglomeration Schedule a 3.7. táblázatban azonos sorrendben és 15 lépésben kapcsolódik össze a 16 változó a négyzetes euklideszi (növekvő) távolság és a csökkenő hasonlóságot jelző korrelációs együttható alapján.

Az 1. számú változó, a Terület mérőszám elkülönül a többi változótól, csak az utolsó három lépésben kapcsolódik a többiekhez.

c) Az összevonási struktúrát mutató dendrogramon mindig 25 a maximális távolság, bármilyen mutatót és eljárást választunk. Itt az átlagos lánc elvű klaszterezés ábráját[47] mutatjuk be, behúzva a 40%-os távolsági szintvonalat. A 3.4. ábra azt jelzi, hogy két nagyobb változócsoportunk van, és két változó (*Terület és Álláskeresők aránya*) távol van / nem korrelál a többiekkel és egymással sem.
Az első nagy klaszterben 10 változót találunk, amelyek a településeken mért *létszámot, méretet* mutatnak. Míg a második klaszterben négy olyan változó van, amelyek *létszámarányos* mutatók.
A 16 változó tehát nem képezhető le 2 dimenzióba a két „kilógó", magasabb távolságnál kapcsolódó változó miatt, de a többi 14 változó határozottan két csoportba különíthető el[48].

[47] Ezen az adatállományon azonos az ábra, ha a távolság- és a hasonlóságmértéket változtatjuk, vagy ha előzetesen sztenderdizáljuk az adatokat. A második esetben minden változó z-score-ja szerepel feliratként.
[48] Ez hasznos információt jelent a későbbi faktorelemzéshez (6. fejezet).

3.7. táblázat: Az összevonás lépései

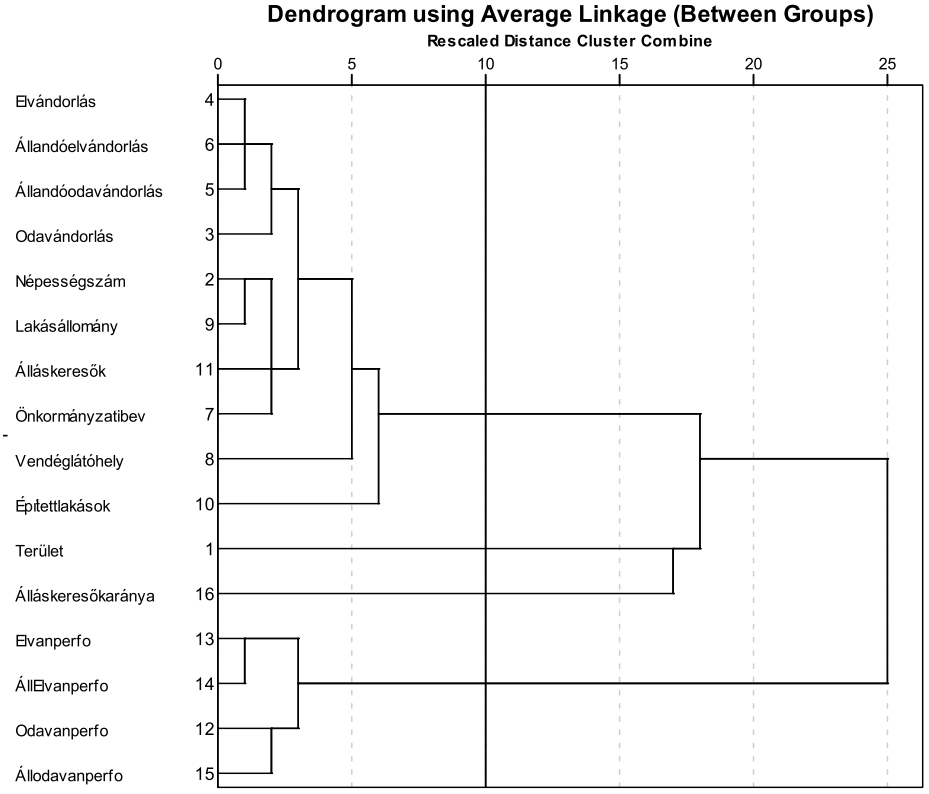
Agglomeration Schedule

Stage	Cluster Combined		Sq. Euc. Distance Coefficients	Stage Cluster First Appears		Next Stage	Correlation Coefficients
	Cluster 1	Cluster 2		Cluster 1	Cluster 2		
1	4	6	,947	0	0	3	,990
2	2	9	1,970	0	0	7	,980
3	4	5	4,078	1	0	5	,958
4	13	14	5,874	0	0	9	,940
5	3	4	6,697	0	3	10	,932
6	12	15	8,304	0	0	9	,915
7	2	11	9,116	2	0	8	,907
8	2	7	11,276	7	0	10	,885
9	12	13	12,077	6	4	15	,877
10	2	3	14,156	8	5	11	,856
11	2	8	22,933	10	0	12	,766
12	2	10	31,518	11	0	14	,678
13	1	16	87,459	0	0	14	,108
14	1	2	94,310	13	12	15	,038
15	1	12	132,195	14	9	0	-,349

c) Az összevonási struktúrát mutató dendrogramon mindig 25 a maximális távolság, bármilyen mutatót és eljárást választunk. Itt az átlagos lánc elvű klaszterezés ábráját[49] mutatjuk be, behúzva a 40%-os távolsági szintvonalat. A 3.4. ábra azt jelzi, hogy két nagyobb változócsoportunk van, és két változó (*Terület és Álláskeresők aránya*) távol van / nem korrelál a többiekkel és egymással sem.

[49] Ezen az adatállományon azonos az ábra, ha a távolság- és a hasonlóságmértéket változtatjuk, vagy ha előzetesen sztenderdizáljuk az adatokat. A második esetben minden változó z-score-ja szerepel feliratként.

Az első nagy klaszterben 10 változót találunk, amelyek a településeken mért *létszámot, méretet* mutatnak. Míg a második klaszterben négy olyan változó van, amelyek *létszámarányos* mutatók.
A 16 változó tehát nem képezhető le 2 dimenzióba a két „kilógó", magasabb távolságnál kapcsolódó változó miatt, de a többi 14 változó határozottan két csoportba különíthető el[50].



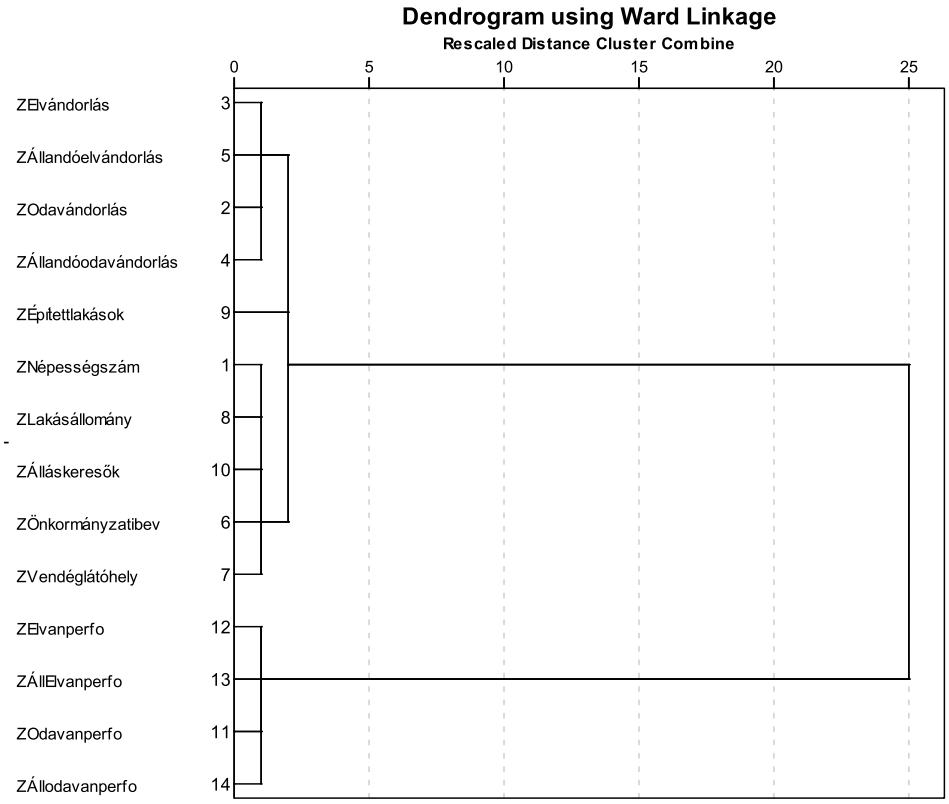
3.4. ábra: A változók összekapcsolódása az átlagos távolságok alapján

Gondoljunk arra is, hogy ha kihagyjuk a *Terület és Álláskeresők aránya* változókat, akkor a többi 14 változó közötti távolság fogja hasonlóan kitölteni a dendrogramon a helyet, mert a maximális távolság e két csoport között látható.

Második dendrogramként a sztenderdizált változókra Ward eljárással képzett klasztereket mutatjuk be. A változók klaszterződése hasonló, tehát stabilan elválnak az eredeti és az egy főre vetített mutatók. Mivel a Ward eljárás a belső

[50] Ez hasznos információt jelent a későbbi faktorelemzéshez (6. fejezet).

eltérések négyzetösszegét minimalizálja, itt a maximális értéke 451,676, ez tartozik a 25 távolságszinthez a 3.5. ábrán.



3.5. ábra: A változók összekapcsolódása az eltérés-négyzetösszegek alapján

2) kérdés eredményei: Hogyan tagolódnak a települések? Valóban elválik egymástól a 23 kerület és a többi 27 Budapest környéki település? Milyen klaszterszámot érdemes feltételezni?

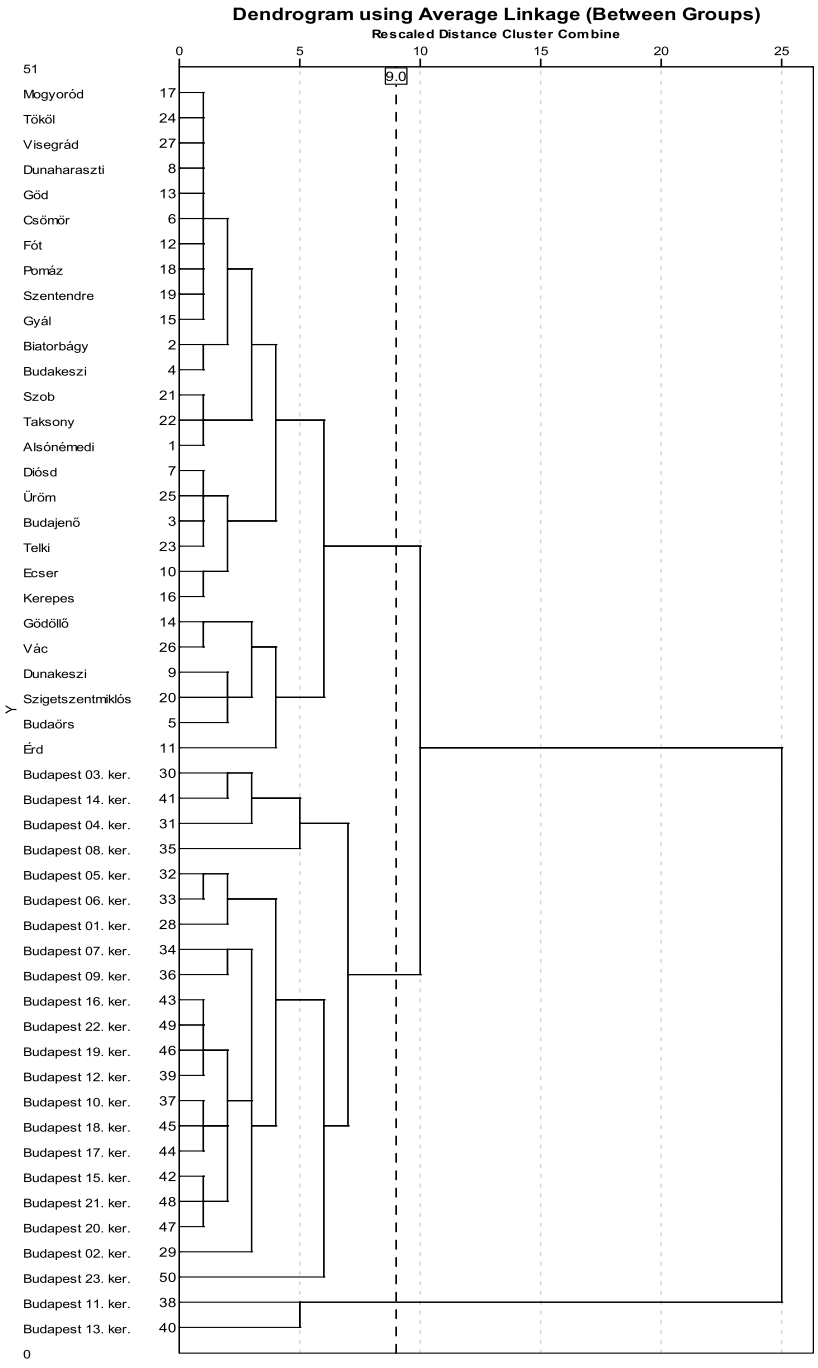
A klaszterezéshez nem tartozik hüvelykujj szabály, hogy hány változót és hány megfigyelést célszerű használni, ezért elkészíthetjük a teljes 16 dimenziós változótérben képzett település-dendrogramot. Az euklideszi távolság négyzetére az átlagos lánc elvű összekapcsolás (3.6. ábra) inkább **3 klasztert** mutat, mint kettőt. Egyrészt határozottan elkülönül a főváros XI. és XIII. kerülete, másrészt a fővároson kívüli településeket és a többi kerületet is érdemes megbontani. A 10, mint vágási szint nem előírás, most túlságosan nagy és heterogén klasztert jelentene, ha együtt vizsgálnánk a 48 települést. Ezért a 9-es szint alatt olvassuk le a klaszterszámot, példánkban a hármat.

Ellenőrzést jelent a másik távolsági mutató vagy a másik klaszterező eljárás alkalmazása. A Ward elv mentén képzett település-klaszterek (3.7. ábra) egyértelműen 2 csoportot mutatnak, és itt már éles a budapesti kerület – nem főváros kettéválás. Ha azonban kisebb belső eltéréseket engedünk meg, azaz homogénebb klaszterekt keresünk, akkor (8-as vágási szinten) három klaszter különböztethető meg. Így négy (nagy) budapesti kerület elkülönül a főváros többi részétől.

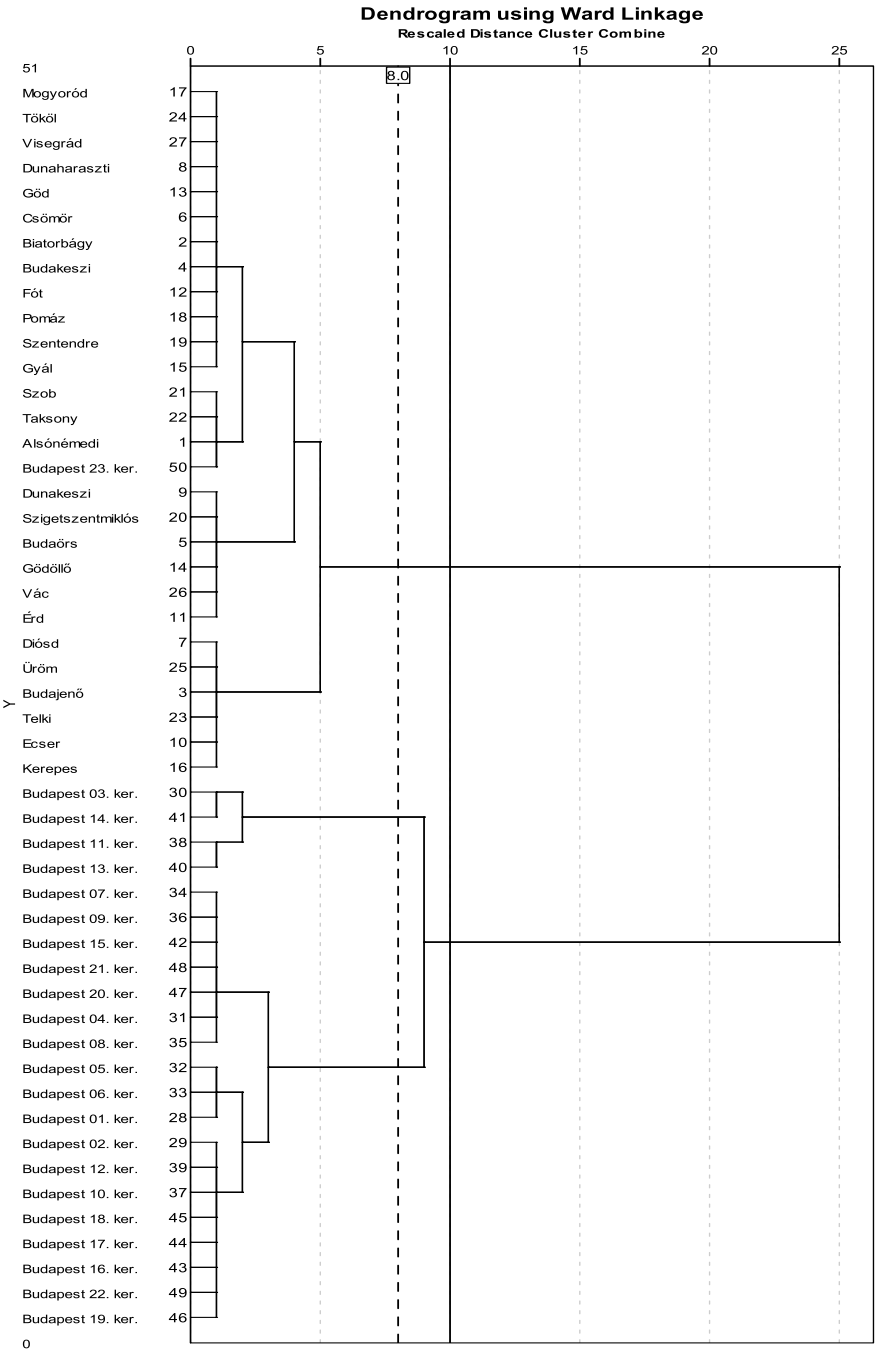
Ha a legtávolabbi szomszéd elvet választjuk, akkor is 3 klaszter látható a dendrogramon, de a XI. és XIII. kerület elvállása mellett nem a főváros – többi település a metszés alapja, hanem további 2-5 kisebb, de vegyes klasztert kapunk.

Nem rejtjük véka alá azt, hogy a választási döntések eredményre gyakorolt hatása óriási. Ha például az euklideszi távolság (négyzete) helyett csak abszolút értékes eltérést választunk, azaz a nagy eltéréseket nem súlyozottan vesszük figyelembe, akkor teljesen megváltozik az ábra.

Az elemző felelőssége tehát óriási, hogy hányféle számítást készít, és végül melyik megoldást tekinti a további elemzéshez jó alapnak. A hierarchikus klaszterezéssel tehát csak egy feltevést kapunk a klaszterszámra, amit elmenthetünk, és ez alapján tovább vizsgálódunk. Most az átlagos lánc elv 3 klaszterét és a Ward módszer 2-3 klaszterét is elmentjük.



3.6. ábra: A települések összevonása átlagos lánc elven



3.7. ábra: A települések összevonása Ward módszerével

3) kérdés eredményei: Ha k=2 beállítással készít k-közép klaszterezést, akkor a település típussal azonosnak tekinthető felosztás adódik?

Az előzetesen – sztenderdizált – változók terében k=2 klaszterezéssel besoroljuk a településeket. A magpontokhoz való besorolás 9 iterációs lépésben lezárul, és az ANOVA táblázatbeli F-teszt és p szignifikancia szint alapján két változónak nem szignifikáns a megkülönböztető ereje. Ezért a *Terület* (p=0,233) és az *Álláskeresők aránya* (p=0,555) elhagyásával 14 változó terében megismételjük a 8 lépéses klaszterezést, és mivel minden változó megkülönböztető erővel rendelkezik, elmentjük a klaszter-azonosítókat.

A szórásfelbontást mutató ANOVA táblázat (3.8. táblázat) megadása csak leíró célokat szolgál, a klaszterképzésben nem kerül sor hipotézisvizsgálatra. Mivel nem tételezzük fel, hogy a csoportátlagok megegyeznek, nem is vizsgáljuk az F-teszt előfeltételeinek[51] teljesülését. Az azonban kiolvasható a 3.6. táblázatból, hogy a legerősebben megkülönböztető változók a *Népességszám* (F=117,476), majd a *Lakásállomány* (F=110,563), továbbá hasonló erőt képvisel az *Önkormányzati bevétel* (F=96,613) és az *Álláskeresők száma* (F=95,990).

[51] Tehát nem kell ellenőrizni a változók szerinti normális eloszlást és a csoportonkénti azonos varianciát.

3.8. táblázat: A változók klaszterek közötti és klaszteren belüli eltérésnégyzet-összegei

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(Népességszám)	34,786	1	,296	48	117,476	,000
Zscore(Odavándorlás)	16,659	1	,674	48	24,726	,000
Zscore(Elvándorlás)	25,157	1	,497	48	50,646	,000
Zscore(Állandóodavándorlás)	17,241	1	,662	48	26,059	,000
Zscore(Állandóelvándorlás)	26,362	1	,472	48	55,895	,000
Zscore(Önkormányzatibev)	32,736	1	,339	48	96,613	,000
Zscore(Vendéglátóhely)	28,776	1	,421	48	68,297	,000
Zscore(Lakásállomány)	34,167	1	,309	48	110,563	,000
Zscore(Építettlakások)	8,563	1	,842	48	10,165	,003
Zscore(Álláskeresők)	32,666	1	,340	48	95,990	,000
Zscore(Odavanperfo)	18,797	1	,629	48	29,873	,000
Zscore(Elvanperfo)	25,881	1	,482	48	53,733	,000
Zscore(ÁllElvanperfo)	22,046	1	,562	48	39,260	,000
Zscore(Állodavanperfo)	22,298	1	,556	48	40,084	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

A felosztás szerint az 1. klaszterbe tartozó 28 település kisebb népességű, mint az átlag[52], kevesebb ott a lakás, a bevétel, a vendéglő, továbbá abszolút számban az átlagnál kisebb ott a vándorlás, de a kisebb létszámra vetítve arányaiban átlag feletti az oda- és elvándorlás. (3.9. táblázat)

A 2. klaszterbe a többi 22 település került, amelyek az első 10 változó szerint az átlagnál nagyobbak, míg az utolsó 4 változó szerint az átlagnál kisebb értékekkel bírnak.

[52] A sztenderdizált változók használata azért is előnyös, mert így az előjel mutatja, hogy a zérus átlaghoz képest milyen tulajdonsággal rendelkeznek a klaszterek.

3.9. táblázat: A klaszterközéppontok változónként számított értékei

Final Cluster Centers

	Cluster	
	1	2
Zscore(Népességszám)	-,73935	,94100
Zscore(Odavándorlás)	-,51165	,65120
Zscore(Elvándorlás)	-,62875	,80023
Zscore(Állandóodavándorlás)	-,52051	,66247
Zscore(Állandóelvándorlás)	-,64363	,81916
Zscore(Önkormányzatibev)	-,71723	,91284
Zscore(Vendéglátóhely)	-,67245	,85585
Zscore(Lakásállomány)	-,73274	,93258
Zscore(Építettlakások)	-,36683	,46688
Zscore(Álláskeresők)	-,71646	,91186
Zscore(Odavanperfo)	,54349	-,69172
Zscore(Elvanperfo)	,63773	-,81165
Zscore(ÁllElvanperfo)	,58859	-,74912
Zscore(Állodavanperfo)	,59195	-,75339

A kérdésre válaszolni tudunk, ha kereszttáblázatban összevetjük a település jellege és a klaszterazonosítók alapján kapott besorolást. Az agglomerációból a 2. klaszterbe, a „nagyok" közé sorolt település Érd, míg a fővárosi kerületek közül kettő került az 1. klaszterbe: az I. és a XXIII. kerület, amelyek valóban mind a 10 méretmutató szerint kisebbek, mint a Budapest többi kerülete. (3.10. táblázat)

A kétféle felosztásra a függetlenségi hipotézist elvetjük (khi-négyzet teszt értéke 38,681, p=0,000) és az asszociáció a Phi és a Cramer V mutatóra azonosan[53] nagyon szoros: 0,880 (p=0,000)

[53] A 2x2 táblázat szabadsági foka 1, ezért egyezik meg itt a két mutató.

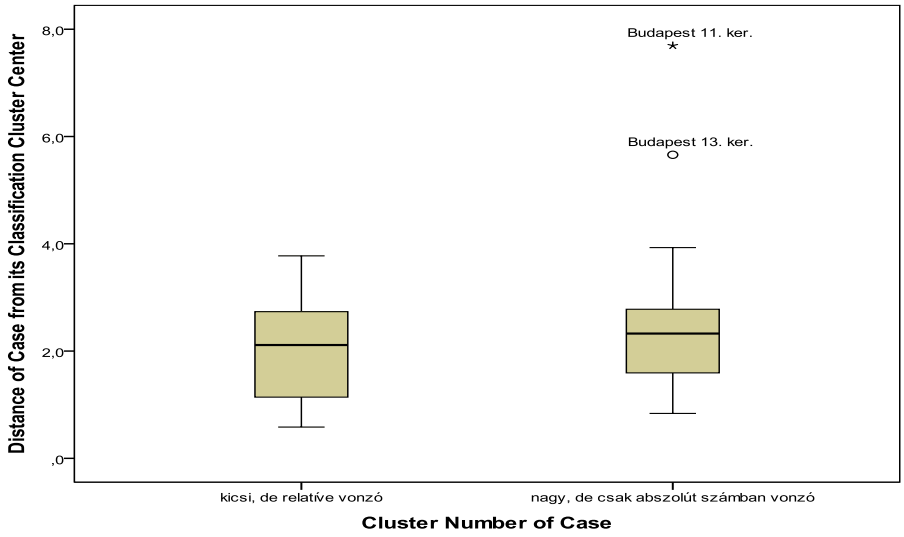
3.10. táblázat: A települések és a klasztertagok kereszttáblája

Kerület * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case		
		1	2	Total
Kerület	Agglomeráció	26	1	27
	Kerület	2	21	23
Total		28	22	50

Még egy ellenőrzési lehetőséget érdemes használni arra, hogy valóban stabil-e a két klaszteres felosztás. A településekre elmenthető, hogy mekkora a saját klaszterközéppontjuktól mért távolságuk. Ezeket pedig dobozdiagramon (3.8. ábra) ábrázolva látjuk, hogy a két klaszter közel azonos belső homogenitással bír, hiszen azonos méretűek a dobozok és közel azonos a távolságok medián vonala. Az eltérés csak annyi, hogy a 2. klaszterbe tartozó XI. és XIII. kerületek távolabb vannak a középponttól. Ha kettőről háromra, négyre vagy ötre emeljük a klaszterszámot, akkor is e kerületek alkotnak önálló klasztert. (Három klaszter esetén még a XIV. kerület csatlakozik hozzájuk.) Ilyen dobozdiagramot érdemes a klaszterek szerinti bontásban az eredeti változókra is készíteni. Akkor világosan látható, hogy az ANOVA táblázat szerint szignifikáns változók dobozai eltérő magasságban vannak.



3.8. ábra: A két klaszterben mért belső távolságok

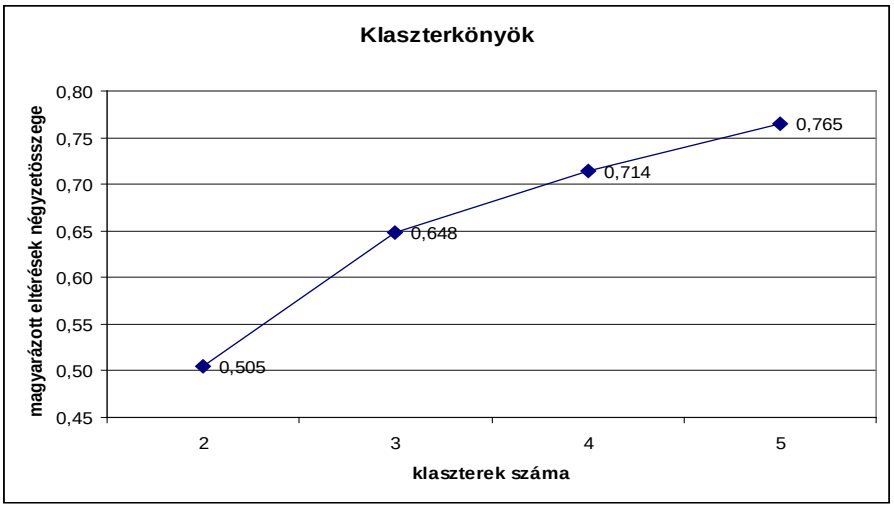
4) kérdés eredményei: Hány klasztert érdemes megkülönböztetni?

Az n=50 elemszám miatt maximum 5 klasztert érdemes előállítani. A k=3,4,5 futtatásokat a 3) lépés szerinti beállításokkal megismételjük, és az eredményeket elmentjük.

Ezt követi a csoportátlagok összehasonlítása az egy-utas ANOVA táblák alapján. Azért nem többváltozós (MANOVA) eljárást alkalmazunk, mert érdemes megnézni minden változó parciális hozzájárulását a csoportok közötti különbséghez.

A külső – klaszterek közötti – eltérések négyzetösszege és a teljes – a minta egészére mért – eltérések négyzetösszege a klaszterek által megmagyarázott eltérések hányadát adja meg. Ez a mérőszám csak külön számítással, például EXCEL-ben képezhető, ezért az SPSS output táblákra kattintva EXPORT menűpontsal kimásoljuk az ANOVA táblákat. Az összegzések után elkészíthető a klaszterkönyök ábra, amely mutatja, hogy további klaszterek előállításával mennyivel növelhető a magyarázott eltérések hányada. (3.9. ábra)

Két klaszter képzésével a különbségek 50%-át, három klaszterrel pedig 64%-át tudjuk megmagyarázni. A további klaszterek előállítása már kisebb arányú és mértékű növekedést eredményez, ezért a háromklaszteres megoldást fogadjuk el.



3.9. ábra: A klaszterszámok és a magyarázott eltérések kapcsolata

Összegzésül érdemes figyelni arra, hogy a klaszterek általában nem rangsorolhatók. A 3.11. táblázatban megmutatjuk a három klaszterre kapott középpontokat. Látható, hogy a 10 méret-mutató alapján 2-1-3 a sorrend, azaz 2. klaszter (XI, XIII, XIV. kerület) a legnagyobb, ezt követi az 1. klaszter (21

tag, benne Érd és húsz fővárosi kerület), végül a 3. klaszterben (26 település) vannak a legkisebb átlagok. Míg a négy létszámarányos mutatóra a 3-2-1 sorrend adódik, bár a rendezettség értelme kérdéses, hiszen az oda- és elvándorlás hasonló értékpárokat jelez.

3.11. táblázat: A háromklaszteres felbontás középpontjai

Final Cluster Centers

	Cluster		
	1	2	3
Zscore(Népességszám)	,61404	**2,22201**	-,75234
Zscore(Odavándorlás)	,19085	**3,01267**	-,50176
Zscore(Elvándorlás)	,41749	**2,44990**	-,61989
Zscore(Állandóodavándorlás)	,23395	**2,73049**	-,50401
Zscore(Állandóelvándorlás)	,47605	**2,17202**	-,63512
Zscore(Önkormányzatibev)	,63625	**1,99471**	-,74405
Zscore(Vendéglátóhely)	,59034	**1,87814**	-,69352
Zscore(Lakásállomány)	,57831	**2,51209**	-,75695
Zscore(Épített lakások)	,02224	**2,89130**	-,35158
Zscore(Álláskeresők)	,66578	**1,68841**	-,73256
Zscore(Odavanperfo)	-,90322	,51612	**,66998**
Zscore(Elvanperfo)	-,86323	-,54159	**,75971**
Zscore(ÁllElvanperfo)	-,75900	-,74313	**,69879**
Zscore(Állodavanperfo)	-,82118	-,33622	**,70206**

4.Többváltozós regressziószámítás[54]

Az eljárás alapgondolata ismerős mindenkinek, aki már tanult statisztikát. Mégis érdemes egy fejezetnyit foglalkozni a regressziószámítással, mert a cikkek, tanulmányok használják az eljárást, és a könyv további fejezeteiben is többször visszautalunk erre a megközelítésre.

Többváltozós lineáris regressziós modellt írunk fel akkor, ha több független magyarázó változó lineáris kombinációjával becsüljük a magyarázni kívánt y változót. A regressziós becslés elvégzése és az eredmények értékelése számos döntést igényel. Tekintsük át először ezeket a főbb döntési pontokat.

a) Az adatok közvetlenül alkalmasak regressziós modell illesztésére vagy adatelőkészítést kell végeznünk? A 4.1. alfejezet és a 4.2.1. alfejezet ad betekintést a részletekbe.
    - A magyarázó változó normális eloszlású-e, és ha nem, akkor milyen (például logaritmus) adat-transzformáció indokolt?
    - A független változók relatív szórásaira teljesül-e a kisebb, mint 2 feltétel? Ha nem, akkor vannak-e extrém értékű, kihagyható megfigyelések?
    - A pontdiagram alapján az y-x párok lineáris kapcsolata fennáll-e? Ha nem, akkor linearizáló transzformáció végezhető-e?
    - A független változók közötti páronkénti korrelációk gyengék-e? Ha nem akkor szakmai vagy statisztikai szempontok alapján válogatjuk ki a modell magyarázó változóit?
b) Az illesztés menete, a változók közötti szelekció végrehajtása. A 4.2.2., a 4.2.4. és a 4.2.5. alfejezetek mutatják az eljárás lépéseit.
    - Melyek a statisztikai értelemben legerősebb magyarázó erővel bíró változók? Mely tesztek támasztják alá a változószelekciót?
    - Létezik-e lineáris modell, vagy minden becsült együttható nullának tekinthető?
    - Milyen tesztekkel és hogyan minősíthető a modell egésze?

[54] A regressziószámítás alapmodelljét és az együtthatók becslését szolgáltató legkisebb négyzetek módszerét ismertnek tételezzük fel.

c) A magyarázó változók közötti kapcsolatrendszer megfelelő-e? A 4.2.3. és a 4.2.6. alfejezeteket tartalmaznak útmutatást erre a kérdésre.

- Milyen mutatókra támaszkodhatunk annak mérésekor, hogy túlzott multikollinearitás fellépett-e?
- Mely változók elhagyásával küszöbölhető ki a multikollinearitás?

d) Modell diagnosztika, hibatagok viselkedése, kiugró pontok kezelése. A 4.2.7. alfejezet hasznos az alábbi kérdések megválaszolásakor.

- Megfelelő magyarázó erejű modellt kaptunk-e?
- A hibatagok normális eloszlásúak-e?
- A hibatagok szórása azonos-e, nem lépett fel heteroszkedasztikitás?
- Vannak-e nagyon erős hatást gyakorló megfigyelések a mintában? Ezek elhagyása indokolt-e?

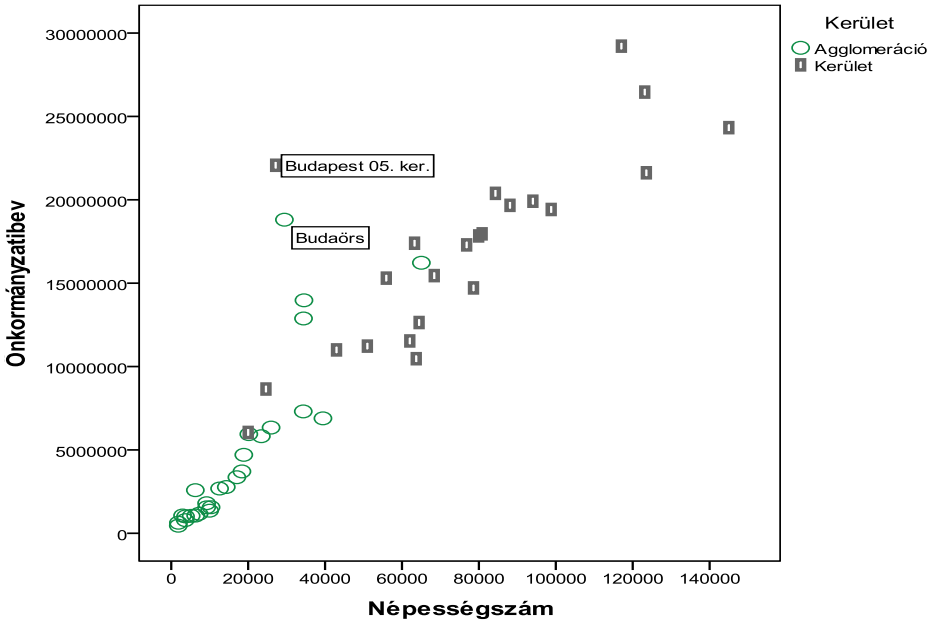
4.1. Az adatok áttekintése, előzetes megfontolások

Az induló adatok között szereplő változókat intervallum vagy arány skálán mérjük, és feltételezzük, hogy az n számú megfigyelés homogén sokaságból származik. Az *y* függő változó normális elosztást követő n elemű oszlopvektor. A p darab magyarázó változót és a konstanshoz tartozó egyeseket az n(p+1) méretű *X* mátrix tartalmazza. A magyarázó változók között kétértékű, dummy változók is szerepelhetnek.

A regressziószámítás két legfőbb lépése az együtthatók becslése és a regressziós modell tesztelése. De sok egyszerű numerikus és grafikus vizsgálati lépést megtehetünk a becslés és a tesztelés előtt. A ferdeség és csúcsosság mellett a relatív szórás kiszámítása képet ad az y változóról. Grafikus módszerekkel, például 2-3 dimenziós pontdiagram készítésével már a regressziós modell felállítása előtt meggyőződhetünk arról, hogy közelítően teljesülnek-e az előfeltételek, használható lesz-e a regressziós modell. Mivel grafikus ábra magasabb dimenzióban nem készíthető, ezek a lépések nem helyettesítik a modell jóságát vizsgáló teszteket, de a teljesen hasznavehetetlen számítások megelőzésére alkalmasak.

Az y és egy-egy x változó pontdiagramján láthatóvá tehetünk sok fontos részletet. Ebben az alfejezetben[55] a Kerületek2010.sav adatállományt használjuk. A 4.1. ábra Budapest 23 kerületének és az agglomeráció további 27 településének népességszámát és az önkormányzati bevétel nagyságát mutatja. Ez az ábra alkalmas arra, hogy ellenőrizzük a 4.1. táblázatban szereplő követelményeket. Érdemes további lehetséges magyarázó változókra is ábrát készíteni a modell illesztése előtt.

[55] A kerületek adatainak további elemzése a 4.2.10-ben szerepel.



4.1. ábra: Kilógó pontok hatása a regressziós egyenesre

A 4.1. ábrán látható kilógó pontok szerepeltetése az adatok között meredekebb regressziós egyenest eredményezne. Ha mindkét változó mentén kilógó megfigyelést találunk, annak kettős hatása lehet:

- Ha a megfigyelt lineáris tendencia mentén – de a többiektől távolabb – van egy pont, akkor szerepeltetése a mintában felerősíti a modell jóságát.
- Ha nem a megfigyelt lineáris tendencia mentén találunk távolabbi pontot, akkor a pont elhagyása javítja az illeszkedést, figyelembe vétele pedig nem lineáris modellt igényel.

4.1. táblázat: Mikor alkalmasak az adatok lineáris regressziós modell illesztésére?

Elméleti követelmények	*Következtetés a pontdiagram alapján*	*Döntés*
Lineáris-e a kapcsolat, jogos-e a lineáris modell illesztése, vagy más függvénytípust célszerű feltételezni?	*A népességszám és az önkormányzati bevétel együttes növekedése, lineáris kapcsolata fennáll.*	+
Az x növekedésével az y adatok szórása változatlan marad-e, a hibatag konstans szórása feltételezhető-e?	*A népesség növekedésével az önkormányzati bevételek szórása enyhén növekedik, bár a kisebb lakosság mellett is van két helyen jelentősebb eltérés az általános tendenciától.*	*?*
Vannak-e kilógó pontok, és milyen az elhelyezkedésük? Egy vagy mindkét dimenzióban kilógnak-e?	*Budapest V. kerülete és Budaörs népességszáma alapján inkább kicsik, míg a bevételük jóval magasabb, tehát az egyik dimenzióban kilógó megfigyelések.*	-
Homogén-e a minta, vagy alminták láthatók, amelyekben más-más tendencia érvényesül a változók között?	*Az adatok homogenitása megfelelő, nem mutatnak a fővárosi kerületek más tendenciát, mint a környékbeli települések.*	+
Az egyes x pontokhoz tartozó y értékek normális eloszlást[56] követnek-e, a tesztek elvégezhetőek lesznek-e?	*Ez csak hisztogramon látható, vagy a ferdeség és csúcsosság mutatókkal írható le. Statisztikailag elfogadható a feltevés.*	+

Ha összegezzük döntéseinket – amiket természetesen a további magyarázó változókra is elvégeztünk –, akkor már csak a magyarázó változók egymás közötti korreláltságát kell megvizsgálnunk, hogy választani tudjunk a modellépítés két útja között:

I) Megerősítő szemlélet: A szakmai tudásunk alapján előre rögzített magyarázó változók körét egyszerre, egy lépésben vonjuk be a modellbe. Így bekerülhet a modellbe statisztikai értelemben nem szignifikáns magyarázó változó is. Ekkor a modell utólagos értékelésével győződünk meg arról, hogy elfogadható-e a modell egésze, és minden változó szignifikáns szerepet játszik-e a becslésben.

II) Feltáró szemlélet: A lehetséges magyarázó változók halmazát megadva lépésenkénti regressziós eljárással minden lépésben egy-egy változót vonunk

[56] A normalitás a regressziós együtthatók becsléséhez nem szükséges, csak akkor kell feltételeznünk, ha t-próbát végzünk, és konfidencia intervallumot írunk fel.

be[57] a modellbe, és a bevont változók elhagyhatóságát is lépésenként ellenőrizzük. Így olyan modell adódik, ami statisztikai értelemben a „lehető legjobb", de előfordulhat, hogy szakmailag nehezebben értelmezhető.

A kétféle megvalósítás számítási lépései nem térnek el érdemben. Minden illesztett modell jóságát négy fő lépésben értékelhetjük:

a) Parciálisan vizsgáljuk egy-egy magyarázó változó hatását/erejét t-próbával.

b) Vizsgáljuk azt, hogy az összes magyarázó változó együttesen szignifikáns kapcsolatban van-e az eredményváltozóval, ezért mérjük az R^2 és a korrigált R^2 értékét, valamint elvégezzük az F-próbát.

c) A hibatagok megfelelő viselkedését ellenőrizzük.

d) A megfigyeléseknek a becslésre gyakorolt egyedi hatását vizsgálni kell.

Az eddig ismertetett döntési pontokat a 4.2. táblázatban foglaljuk össze.

4.2. táblázat: A regressziós modellek és tesztek áttekintése

Regressziós modell	I) megerősítő	II) feltáró
Változók bevonása	egyszerre, egy lépésben	szelektálva
a) Változók ereje	minden változót tesztelni kell (t-próba), és a változók között lehet multikollinearitás	minden bevont változó szignifikáns (de a konstans nem mindig!)
b) Modell egésze	az R^2 és a korrigált R^2, valamint az F-próba alapján minősítjük	az adott változókörből ez a legjobb lineáris modell, de ez elég jó-e?
c) Hibatagok	normális eloszlását és homo-szkedasztikus jellegét ellenőrizni kell	azonos az I) modellel
d) Egyedi megfigye-lések hatása	a túlzott áttétel-hatást mérni, vizsgálni kell, és a zavaró pontokat elhagyni	azonos az I) modellel

[57] A változók lépésenkénti bevonása mellett van a teljes modellből induló, a változókat lépésenként kihagyó változat is, ezekkel majd a megvalósítási részben foglalkozunk.

4.2. A regresszió matematikai háttere

A többváltozós lineáris modell mátrix-egyenlete:

$$y = X\beta + \varepsilon \text{ , ahol} \tag{4.1}$$

az y n elemű vektor, X mátrixnak n sora és (p+1) oszlopa van, az ismeretlen együtthatók β vektora (p+1) elemű, az ε hibatag n elemű.

A modell alkalmazásának feltételei:

- A hibatag normális eloszlású, várható értéke zérus, varianciája konstans, és a hibatagok nem autokorreláltak.
- A magyarázó változók lineárisan függetlenek, értékük mérési hibát nem tartalmaz.
- A megfigyelések száma és a magyarázó változók száma között fennáll, hogy n>5p.

E feltételek teljesülése esetén a (p+1) regressziós együttható legkisebb négyzetes becslése:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{4.2}$$

A megoldás előállítható, ha az inverz létezik, azaz ha X rangja[58] (p+1).

A reziduumok varianciája:

$$\sigma^2 = \frac{\varepsilon^T \varepsilon}{n - p - 1} \text{, ahol} \qquad \varepsilon = y - \hat{y} = y - X\hat{\beta} \tag{4.3}$$

A modellben levő szórásnégyzet felbontása

Az együtthatók becsült értékét a továbbiakban b-vel, és a becsléshez tartozó reziduumokat e-vel jelöljük:

$$e = y - Xb \text{ .}$$

A teljes eltérések négyzetösszege (SST: Sum of Square of Total) az egyváltozós modellhez hasonló alakú, ez az y változó szórásnégyzetének n-szerese:

[58] Az inverz létezik, ha X oszlopvektorai lineárisan függetlenek. A gyakorlatban előfordul, hogy valamelyik változó kifejezhető a többi lineáris kombinációjaként, vagy erősen korrelálnak egymással. Ebben az esetben multikollinearitás lép fel, és ekkor lépésenkénti regressziót célszerű végezni.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = y^T y + n\bar{y}^2 \quad (4.4)$$

Az SST az y változó megfigyelt értékeiből kiszámítható, de most az a célunk, hogy két részre bontsuk[59]: *SST=SSR+SSE*

- Az x magyarázó változók által a regressziós modellben megmagyarázott hányad (*SSR*: Sum of Square of Regression) a lehető legnagyobb legyen.
- A meg nem magyarázott rész, az ún. hibahatás (*SSE*: Sum of Square of Error) pedig minél kisebb legyen.

A hiba-variancia (s^2) az *SSE* jelölésű eltérés-négyzetösszegből osztással kapható meg:

$$SSE=e^Te= \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \left(y - \hat{y}\right)^T\left(y - \hat{y}\right) = (y - Xb)^T(y - Xb)$$, és

$$s^2 = (e^T e)/(n - p - 1) \quad (4.5)$$

A regressziós együtthatók szórásnégyzete a hibavariancia (4.5) segítségével határozható meg. Egy *b* regressziós együttható varianciája az $(X^TX)^{-1}$ megfelelő diagonális eleméből adódik:

$$Var(b_j)=s^2\, diag_j\{(X^TX)^{-1}\} \quad (4.6)$$

A regressziós eltérés-négyzetösszeg nagysága különbségként is megkapható:

$$SSR = SST\text{-}SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = y^T Xb - n\bar{y}^2 \quad (4.7)$$

Az egyenletek felírása után következik a tesztelés, azaz annak eldöntése, hogy eredményes volt-e a modell illesztése. Ennek eldöntéséhez azt a nullhipotézist teszteljük, hogy a b_j meredekségek mind zérussal egyenlők, azaz nincs érdemi magyarázó ereje a modellnek. A teszteléshez felírt szórásfelbontó (ANOVA) táblázat (4.3. táblázat) tartalmazza az eddig ismertetett eltérés-négyzetösszeg tagokon túl az átlagos négyzetösszegeket (*MS*), valamint az F-próba értékét.

Azzal, hogy az együtthatók legkisebb négyzetes becslése során az *SSE*-t minimalizáljuk, egyúttal az *SSR*-t maximalizáljuk. Az átlagos négyzetösszegek aránya – az F-hányados –is „nagy" lesz, ha van lineáris regressziós összefüggés a

[59] Az itt alkalmazott jelölés - bár igen elterjedt - csak az egyik lehetőség. Lehet a Sum of Square két része „Explained" és „Residual", akkor épp fordítva van a tartalmuk, mint ahogy itt szerepel.

magyarázó változók és az eredményváltozó között. Ezt a próbafüggvényhez tartozó szignifikancia szint jelzi.

4.3. táblázat: Szórásnégyzet felbontása és tesztelése

A variancia forrása	Eltérés négyzetösszeg	Szabadság fok	Átlagos négyzetösszeg	F-hányados
Regresszió Hibatag	*SSR* *SSE*	p n-p-1	*MSR=SSR/p* *MSE=SSE/(n-p-1)*	*F=MSR/MSE*
Teljes	*SST=SSR+SSE*	n-1	-	-

4.3. A változók közötti korreláció mérése és szerepe a regressziós modellben

A megfigyelések halmazát és a változók körét is szakmai megfontolások alapján választjuk ki, mégis előfordulhat, hogy

- túl sok magyarázó változónk van,
- a magyarázó változók nem függetlenek,
- a változók nem lineárisan kapcsolódnak a függő változóhoz.

A korrelációs együttható (4.8) szerinti képlete centírozott adatokra egyszerűbb alakot ölt, és így közvetlenül látható, hogy a két változó között az n-dimenziós térben bezárt szög koszinuszával azonos értéket ad:

$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum x_i^2\cdot\sum y_i^2}}=\frac{x^T\cdot y}{|x|\cdot|y|}=\cos\alpha \tag{4.8}$$

A függő és a p számú magyarázó változó közötti páronkénti korrelációt tartalmazó (p+1)x(p+1) méretű *R* korrelációs mátrixból a szignifikancia szintek alapján képet kapunk a multikollinearitás mértékéről. A korrelációs mátrix szimmetrikus, a

főátlójában egyesek állnak. A mátrixban található bármely *r* korrelációs együtthatóhoz tartozó szignifikanica szint a t-próba alapján állapítható meg, ahol

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{4.9}$$

Ez a t-teszt (n-2) szabadságfokú Student-eloszlást követ.

4.4. Érdemes-e több változót egyidejűleg bevonni a regressziós modellbe?

Többváltozós modellt csak akkor érdemes becsülni, ha ez érdemben javítja az illeszkedést az egy magyarázó változóhoz képest. Döntésünkhöz globális mutatókat és parciális teszteket használhatunk.
Először a modell egészét minősítő három globális mutatót tekintjük át:

a) determinációs együttható és korrigált változata
b) a modell sztenderd hibája
c) a lineáris modell létét ellenőrző F-teszt

a) Legelterjedtebb a determinációs együttható (a többszörös korreláció négyzete[60]) mellett ennek korrigált (adjusztált) változata az illeszkedés jóságának mérőszámaként:

$R^2 = SSR/SST = 1 - SSE/SST$ ezért $0 \le R^2 \le 1$

$$R_{adj}{}^2 = R^2 - \frac{p(1-R^2)}{n-p-1} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} \tag{4.10}$$

ahol p a modellben szereplő független változók száma.

A korrekció azért szükséges, mert több változó bevonásával R^2 nő, és túl optimista képet mutat a modell illeszkedéséről. Az R^2 és a korrigált változata is százalékosan értelmezhető. Mindkettő azt méri, hogy a modellbe bevont magyarázó változók az eredményváltozó varianciájának hány százalékát magyarázzák meg. E mutatószámokhoz teszt nem kapcsolódik.

b) A regressziós modell sztenderd hibája a (4.3) négyzetgyökének mintabeli becslése.

$$s = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.11}$$

---

[60] Csak kétváltozós modellben érvényes az, hogy a determinációs együttható a közönséges korrelációs együttható négyzete. Érdemes elolvasni Hunyadi László: „A determinációs együtthatóról” című cikkét, mely a Statisztikai Szemle 2000. szeptemberi számában jelent meg. (78. évf. 9. sz. 753-765. oldal)

Hüvelykujj szabályként érdemes megnézni, hogy *s* kisebb-e, mint egy-egy magyarázó változó szórása. Ha nem kisebb, akkor a modell illesztése nem ér annyit sem, mintha a függő változó átlagát tekintenénk becslésnek.

c) A variancia-analízis azt a nullhipotézist teszteli, hogy a b_j meredekségek mind zérussal egyenlők (csak a konstans különbözik szignifikánsan nullától), míg az alternatív hipotézis szerint van zérustól különböző b_j. A nullhipotézis elfogadása azt jelenti, hogy az adott változókkal felírt regressziós modell nem alkalmas y becslésére. Ha elvetjük a nullhipotézist, abból még nem következtethetünk arra, hogy jó becslést tudunk adni a függő változóra, mert lehetnek a modellben nem szignifikáns paraméterű magyarázó változók. Az ANOVA táblából számolt próbafüggvényt globális F-próbának nevezzük.

A modell parciális vizsgálata – a regressziós együtthatók egyenkénti tesztelése – t-próbával történik. A nullhipotézis szerint $H_0 : \beta_j = 0$ és a kétoldali alternatív hipotézis: $H_1 : \beta_j \neq 0$.

A tesztfüggvény Student-eloszlást követ, képlete $t = b_j \Big/ s_{b_j}$, (4.12)

ahol s_{bj} az (4.5) szerinti becsült variancia gyöke. A t-próba szabadságfoka n-j-1, ahol j azt jelzi, hogy a j-edik változót vontuk be a modellbe. A t-eloszlás segítségével (1-α) valószínűségi szintű konfidencia intervallum is felírható az elméleti β_j paraméterre:

$$b_j \pm t_{\alpha/2,(n-j-1)} \cdot s_{b_j} \tag{4.13}$$

A sztenderdizált regressziós együtthatók számítása a (4.14) képlettel[61] történik, ezekre külön tesztet nem kell végezni.

$$beta_j = b_j \cdot s_{x_j} \Big/ s_y \tag{4.14}$$

A sztenderdizált béta nem azonos az elméleti modell β együtthatójával. Értéke a szórások arányától függően kisebb vagy nagyobb is lehet, mint a becsült *b* együttható. Az abszolút értékben legnagyobb értékű változót tekinthetjük a modell legfontosabb magyarázó változójának.

Közvetlen, közvetett és teljes hatás (kitekintés)

A regressziós együtthatók értelmezésekor fontos hangsúlyozni, hogy a magyarázó változók függetlenségét feltételeztük a becslés során. A modellben a b_0 konstans azt

[61] Ha a modellben egyetlen x magyarázó változó van, akkor beta= r, ahol r a közönséges korrelációs együttható.

az alapértéket adja meg, amit y akkor vesz fel, ha minden x_j értéke nulla. A b_j együttható pedig azt a közvetlen hatást méri, hogy mennyivel változik y, ha x_j egy egységgel nő, miközben a többi magyarázó változó értéke változatlan.

Ha a magyarázó változók lineáris függetlensége nem teljesül, akkor y és x_j között a teljes hatást (b_{yj}) a közvetlen hatás (b_j) és az x_j-vel korreláló (pl. x_k) magyarázó változó(ko)n keresztül megvalósuló közvetett hatások együtt adják.

Így $b_{yj} = b_j + b_k * b_{jk}$, ahol b_{jk} az x_k–nak mint magyarázó változónak az x_j–re, mint függő változóra felírt regressziós együtthatója. A direkt és az indirekt hatások feltárása **út-elemzéssel**[62] valósítható meg.

4.5. A változó szelekciót megvalósító lépésenkénti regresszió

A lépésenkénti regresszió 4 eljárással végezhető el, de háromnak közös jellemzője az, hogy egy lépésben egyetlen változó bevonásáról vagy elhagyásáról döntünk. A döntés alapja a parciális F-próba:

$$F_p = \frac{R^2 - R_O^2}{1 - R^2} \cdot \frac{n - p - 1}{q} \tag{4.15}$$

ahol R^2 az aktuális, p magyarázó változós becslés, R_O^2 pedig az előző modell determinációs együtthatója, q pedig az adott lépésben bevont változók száma (általában q=1).

Az F-hányados szabadságfoka a számlálóban q és a nevezőben (n-p-1).

A t-próba négyzete megegyezik ezzel a parciális F-teszttel, amelyet azért számítunk, hogy mérjük az éppen bevont x_j változó magyarázóerejének szignifikanciáját.

Az újabb változók bevonásával R^2 monoton nő a differencia csökkenése mellett. Így eldöntendő kérdés, hogy szignifikánsan nő-e a determinációs hányados az adott változó(k) bevonásával. A beléptetés és kihagyás kritériuma F rögzített nagysága, vagy az F-hez kapcsolódó szignifikancia szint megválasztása lehet.

Ha újabb magyarázó változókat vonunk be a modellbe, akkor az ANOVA táblázatban SSE csökken és SSR nő. Az átlagos négyzetösszegek (MS) változásának iránya már nem egyértelmű, mert a nevezők is változnak, ezért F értékének alakulásáról biztosat nem állíthatunk.

Ha rögzített α valószínűségi szinthez tartozó F-érték mellett (4.15)-ből kifejezzük az R^2 változását, akkor a (4.16) döntési kritériumhoz jutunk. Bevonásra érdemes a változó, ha

---

[62] Angol neve Path analysis, az SPSS-ben nem szerepel.

$$R^2 - R_O^2 > \frac{q}{n-p-1}(1-R^2)F_{\dot{\alpha},q,(n-p-1)} \tag{4.16}$$

A lépések során meghatározásra kerülnek itt előjel nélkül a parciális korrelációk is:

$$R_{parc} = \sqrt{\frac{R^2 - R_O^2}{1-R_O^2}} \tag{4.17}$$

A számláló gyökét részkorrelációnak nevezzük (Part correlation). Ha az újonnan belépő változó valóban korrelálatlan a modellbe már bevont változókkal, akkor a részkorreláció jelentősen nő a vizsgált lépésben.

A lépésenkénti modellezés változatai:

- Forward szelekció: minden lépésben azt a magyarázó változót vonjuk be, amelyiknek a parciális F-tesztjéhez a legkisebb p valószínűség tartozik. A bevonási folyamat addig folytatódik, amíg ez a p az előre rögzített maximum (PIN) alatt marad, vagy minden változó bevonásra került.
- Backward elimináció: az induló lépésben az összes változó a modellben van, és lépésenként azt az egyet hagyjuk ki, amelyiknek a legkisebb a parciális korrelációja. Ekkor a parciális F-teszthez a maximális p valószínűség tartozik. Leáll a kiküszöbölés, ha p kisebb, mint a küszöb (POUT), vagy nincs már változó a modellben.
- Stepwise módszer: a forward eljárást úgy módosítjuk, hogy minden lépésben ellenőrizzük a modellbe korábban bevont változók p valószínűségét, és ha p>POUT, akkor a változót kihagyjuk a modellből. Nem kerülünk végtelen ciklusba, ha PIN≤POUT. Szokásos beállítás: PIN=0,05 és POUT=0,10.
- Remove eljárás: belépteti az összes változót (mint az „Enter” módszer), majd elhagyja egyszerre az összes változót, és összehasonlításként csak a konstans tagot tartalmazó modell eredményeit közli.

4.6. A magyarázó változók közötti korreláció, a multikollinearitás

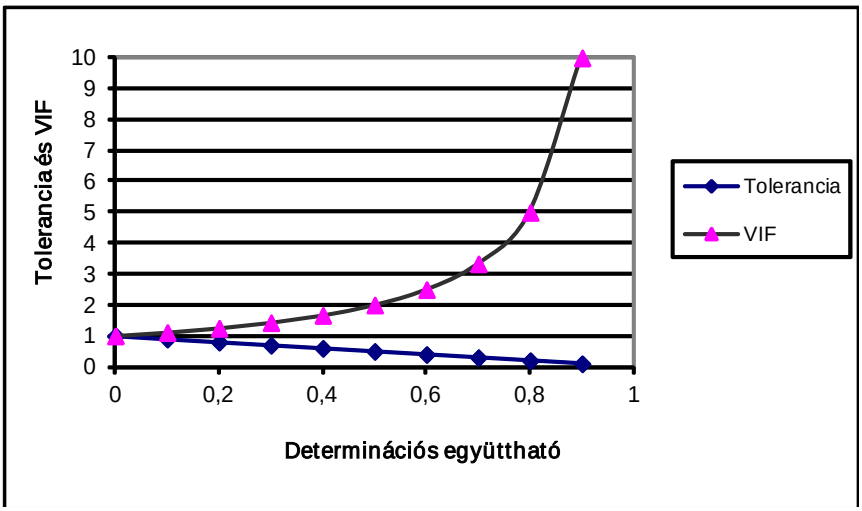
A magyarázó változók függetlenségére vonatkozó elvárást akkor is megsérthetjük, ha lépésenkénti szelekciót végzünk, mert a bevonásnál a modell magyarázó erejének javulásán van a hangsúly. Ezt a közvetett hatások még fokozzák is. Ezért a lépésenkénti regressziós modellezésnél különösen indokolt a modellbe bevont magyarázó változók közötti korreláció, a multikollinearitás mérése, melyre négy mérőszámot ismertetünk.

a) A **tolerancia** mérték annak a többszörös determinációs együtthatónak a komplementere, amely azt méri, hogy az i-edik magyarázó változót az

összes többi *x* milyen szorosan határozza meg: $Tol = 1 - R_i^2$. A kicsi (nullához közeli) tolerancia jelenti azt, hogy közel függvényszerű a kapcsolat a magyarázó változók között.

b) A **variancia infláló faktor** (VIF) a tolerancia reciproka: $VIF_i = 1/(1 - R_i^2)$. Ezért ha a magyarázó változók között szoros kapcsolat van, a VIF végtelen nagy lehet. Ha a változók ortogonálisak, akkor a VIF egységnyi. A VIF_i egyúttal a sztenderdizált magyarázó változókból képzett $(X^T X)^{-1}$ mátrix i-edik diagonális eleme. Ez a képlet szerepel (4.5)-ben a regressziós együtthatók szórásnégyzetének becslésekor. Ezért multikollinearitás fellépésekor nő a VIF, és emiatt nagy lesz a *Var(b)*, továbbá széles lesz az együttható konfidencia intervalluma. A VIF-hez kritikus küszöb nem adható, de hüvelykujj szabály szerint 2-ig elfogadható, 5-ig „tűrhető", öt felett pedig veszélyes.

A két mutató ellentétes alakulását mutatja a 4.2. ábra.



4.2. ábra: A multikollinearitás két mérőszámának alakulása

c) Az $(X^T X)$ centírozatlan, de a szórással leosztott[63] adatokból képzett szorzatmátrix sajátértékeit (λ_i) előállítva és nagyság szerint rendezve **kondíciós index** (CI) képezhető:

$$CI_i = \sqrt{\lambda_{max} / \lambda_i}\ ,\ \text{ahol i=1,…,(p+1)}$$

[63] Ilyen mátrix főátlójában egyesek állnak.

Ha a magyarázó változók között szoros korreláció van, akkor a maximális sajátérték nagy, a többi lambda gyorsan csökken, ezért a kondíciós index is nagy. Hüvelykujj szabály, hogy 1-5 között gyenge, 5-10 között zavaró a multikollinearitás. Ha 10 feletti az index, akkor komoly kollinearitás áll fenn. Ha sok sajátérték közel nulla, akkor az adatokban bekövetkező kis változások nagy változást idéznek elő a becsült együtthatókban.

A nagy sajátértékek száma azt jelzi, hogy hány dimenziós térben jeleníthetők meg a „független" változók. A főkomponens elemzés, amelyet egy későbbi fejezet ismertet, ilyen adathalmazok elemzésére alkalmas.

d) **Variancia hányadot** számíthatunk minden regressziós együtthatóra (a konstansot is beleértve), hogy a regressziós együtthatók varianciáit a sajátértékek (és az általuk jelzett merőleges tengelyek) között szétosszuk. Egy-egy együttható oszlopának összege tehát egységnyi.

Soronként vizsgálva a variancia hányadot, multikollinearitási problémára utal, ha egy-egy nagy kondíciós index sorában több regressziós együtthatónak magas a variancia hányada.

4.7. Az egyedi megfigyelések hatása a becslésre

Eddig az *X* mátrix oszlopaira, a változók szerepére koncentráltunk. Most a sorokat vizsgáljuk, az egyes megfigyelések fontosságát, befolyását mérjük. Az angolul „leverage"-ként megjelenő fogalom áttételhatást jelent. Ezzel a mérőszámmal azonosíthatók az extrém helyzetű megfigyelések is. A hibatagokat is megfigyelésenként vizsgáljuk, valamint távolságot is mérhetünk, mielőtt extrém helyzetűnek minősítünk egy megfigyelést.

4.7.1. A becslést befolyásoló pontok feltárása

A becslést befolyásoló pontok feltárásához a (4.2)-ben felírt becslőegyenlet

$$\hat{B} = (X^T X)^{-1} X^T y$$

mindkét oldalát szorozzuk balról *X* mátrixszal. Ekkor azonosságot kapunk, ahol *H* (nxn)-es mátrix a leképezés[64] mátrixa.

$$X\hat{B} = \hat{y} = X(X^T X)^{-1} X^T y = Hy \tag{4.18}$$

[64] *H* mátrix angol neve „hat matrix".

A (4.18)-ból látható, hogy *H* közvetlen kapcsolatot teremt a függő változó megfigyelt értékei (*y*) és becsült értékei $(\hat{y}_i)$ között.

A *H* mátrix segítségével a hibatagok vektora

$$e = y - \hat{y} = y - Hy = (E - H)y \,, \tag{4.19}$$

ahol *E* az egységmátrix, és így az eltérés-négyzetösszegek is felírhatók:

$$SSE = y^T (E - H)y \quad \text{és} \quad SSR = y^T Hy - n\bar{y}^2 \,.$$

H mátrix szimmetrikus, diagonális elemei (jelölje h_{ii}) azt a hatást fejezik ki, amit az i-edik megfigyelés (X mátrix i-edik sora) gyakorol az összes magyarázó változón keresztül a regressziós becslésre.

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \tag{4.20}$$

Megmutatható, hogy $\sum_{i=1}^{n} h_{ii} = p + 1$, hiszen ennyi az *X* oszlopainak száma, és $1 \geq h_{ii} \geq 1/n$. Egy megfigyelés „áttétel" hatása átlagos, ha értéke (p+1)/n, és befolyásoló, jelentős megfigyelést jelez, ha az átlag kétszeresét meghaladja, azaz $h_{ii} \geq 2(p+1)/n$.

Könnyebb az értelmezés, ha a h-ból a minimális 1/n értéket levonjuk, és az origóhoz tolt hatást (centered leverage) vizsgáljuk:

$$h_{ii} - \frac{1}{n} \tag{4.21}$$

Mivel így 0 és (n-1)/n közötti értéket kaphatunk, gyakorlati szabály adható a *(h –1/n)* eltolással kapott mértékre:

- 0,2 alatti érték mellett a megfigyelések bevonhatók a becslésbe
- 0,2 és 0,5 között kockázatos a becslés elvégzése
- 0,5 felett kerülendő a megfigyelések bevonása a regressziós becslésbe.

Az SPSS kézikönyv által javasolt másik szabály szerint p>6 és (n-p)>12 esetén 3p/n a bevonási küszöb. Ha a megfigyelések száma és a magyarázó változók száma közötti n>5p ajánlást is figyelembe vesszük, akkor 3/5=0,6 feletti értéket elérő megfigyelést semmiképpen nem veszünk figyelembe a regressziós modell becslésekor.

Minden megfigyelt érték *h* súllyal befolyásolja a becslést:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} = \sum_{j=1}^{n} h_{ij} y_j \quad \text{, ahol } h_{ij} = x_i^T (X^T X)^{-1} x_j .$$

A legkisebb négyzetes becslés nagyon érzékeny az extrém (x_i, y_i) megfigyelés-párokra. Ha a megfigyelt *y* érték extrém, és/vagy az *x* értékektől függő *h* súly nagy, akkor erős hatást gyakorolnak a becslésre. Egyszerűbb a hatások értelmezése, ha az *X* mátrixban a független változók átlagtól vett eltérései, a centírozott adatok vannak. Ekkor egy magyarázó változó esetén *h* azt fejezik ki, hogy az *x* változó egy-egy megfigyelt értéke milyen távol van az átlagtól: $h_{ii} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}$.

4.7.2. Hibatagok előállítása és elemzése

A hibatagok, vagy elterjedt szóval reziduálisok vizsgálata nagyon szerteágazó terület. Az elvárások között szerepel, hogy normális eloszlást követnek, függetlenek és állandó a szórásuk.

- Hisztogramot érdemes készíteni, melyről a reziduálisok eloszlása látható, és a normális eloszlástól való eltérés grafikusan megjeleníthető. A reziduálisok ábráit az eredmények bemutatásánál tekintjük át.
- A QQ plot is a hibatagok normális eloszlástól való eltérését mutatja.
- Ha a hibatagokat az y adatok mentén ábrázoljuk, akkor a szórások homoszkedasztikus jellege is szemléltethető.
- Az egyik legismertebb teszt a Durbin-Watson statisztika, amely a hibatagok autokorrelálatlanságát teszteli, ezért idősoros adatok esetében célszerű értelmezni, keresztmetszeti elemzésben nincs létjogosultsága.

Mivel a megfigyelt és a becsült értékek eltérése többféleképpen mérhető, többféle reziduális számítható és elemezhető. A közönséges reziduálisok ($e_i = y_i - \hat{y}_i$) mellett számolható úgy is hibatag, ha egy-egy megfigyelést kihagyunk. Az i-edik megfigyelés (x,y) koordinátáinak elhagyásával nyert becslés és az így számított reziduális[65] indexében szerepel a kihagyott elem: $e_{(i)i} = y_i - \hat{y}_{(i)i}$.

Ha az i-edik megfigyelés erősen befolyásolja a becslést, akkor a két hibatag nagyon eltérő. A két reziduális között a hatás (h_{ii}) értéke teremt kapcsolatot:

---

[65] Az egy megfigyelés törlése, kihagyása után számított hibatag angol neve „deleted” residual. Hunyadi-Mundruczó-Vita: Statisztika c. könyve sorelhagyásos módszert említ.

$$e_{(i)i} = e_i /(1 - h_{ii}) \tag{4.22}$$

Mivel h nem-negatív, $\left|e_{(i)i}\right| \geq \left|e_i\right|$, de nagyméretű, homogén mintában egy-egy megfigyelés kihagyása miatt a kétféle reziduális nem térhet el jelentősen egymástól. Míg a reziduálisok négyzetösszege=SSE, addig a törlések után becsült reziduálisok négyzetösszege[66] PRESS= $\sum_{i=1}^{n} e_{(i)i}^2$. A két összeg hányadosa (PRESS/SSE) jelzi, hogy mennyire érzékeny a regressziós becslés a kihagyott megfigyelésekre. Ha sok és/vagy nagyon távoli (outlier) pont volt a mintában, akkor a PRESS/SSE arány jóval nagyobb, mint egy.

A reziduálisok „nagyságának" megítélését segíti a sztenderdizálás. A közönséges reziduálisokat osztva a (4.4) gyökével, az s szórással, sztenderdizált hibatagokat kapunk:

$$z_i = e_i / s \tag{4.23}$$

Mivel a regressziós becslésből származó hibatagok varianciája torzított, $Var(e_i) = \sigma^2(1 - h_{ii})$, a z_i szórásnégyzete nem egységnyi. Az egységnyi varianciát biztosítja, ha a (4.24) szerint sztenderdizáljuk a hibatagokat. Az így kapott reziduálisok abszolút értékben nagyobbak lesznek (4.23)-beli párjaiknál:

$$r_i = e_i / s\sqrt{1 - h_{ii}} \tag{4.24}$$

A (4.24)-ben a sztenderdizáláshoz használt *s* szórás nem független az e_i hibatagtól, ezért ezt szokták belsőleg studentizált reziduálisnak is nevezni, megkülönböztetve a kihagyással számolt, külsőleg studentizált reziduálistól, t_i–től, amelynek eloszlása Student eloszlást követ:

$$t_i = e_i / s_{(i)}\sqrt{1 - h_{ii}} \tag{4.25}$$

Ez a (4.25)-ben számolt t-statisztika méri az e_i –ben azt, hogy *y* mennyire tér el a becsléstől, és h_{ii} –ben pedig azt, hogy az *x*-ek hatása milyen jelentős. Ha gyanítjuk, hogy valamelyik megfigyelés nagyon rendhagyó, akkor az erre kiszámolt t-értéket összevethetjük a Student eloszlás kritikus értékével. A Student-eloszlás szabadságfoka (n-p-2). Nagy megfigyelésszám mellett normális eloszlás alkalmazható.

Az áttekinthetőség érdekében a 4.4. táblázatban foglaljuk össze a reziduálisok tartalmát, képletét és az SPSS-ben szereplő rövid elnevezést.

---

[66] A sorkihagyásokkal számolt eltérés-négyzetösszegek angol neve: Predicted Residual Sum of Squares= PRESS.

4.4. táblázat: Hibatagok változatai

A reziduális tartalma, (angol neve), betűjele	**Képletének száma**	**SPSS-neve**
Közönséges reziduális (unstandardized): e	(4.19)	res
Az i-edik megfigyelés kihagyásával számított reziduális (deleted): $e_{(i)}$	(4.22)	dre
Sztenderdizált közönséges reziduális: z	(4.23)	zre
Studentizált reziduális, megfigyelés kihagyva, szórás a teljes mintából (studentized): r	(4.24)	sre
Studentizált reziduális, a szórás is kihagyással számolva (studentized deleted): t	(4.25)	sdr

4.7.3. A becslést befolyásoló távoli pontok feltárása, kihagyási döntés

Mahalanobis távolság alapján kiválaszthatjuk azokat a potenciális megfigyeléseket, amelyek kilógónak (outliernek) tekinthetők. A Mahalanobis távolság d_M kétféleképpen is kiszámítható.

a) $d_M = (n-1)(h_{ii} - 1/n)$, (4.26)

b) $d_M^2 = (\hat{y}_{(i)} - \hat{y})^T S^{-1} (\hat{y}_{(i)} - \hat{y})$, ahol S a változók kovariancia[67] mátrixa.

Cook javasolta a D-statisztika számítását, amelyben az i-edik megfigyeléssel és e pont kihagyásával készített lineáris regressziós becsléseket vetjük össze az i-edik pontban: $D_i = \sum_{i=1}^{n} (\hat{y}_{(i)i} - \hat{y}_i)^2 / (p+1)s^2$

A Cook-féle D egyszerűbben kiszámítható a (studentizált) reziduális és a hatásmérték felhasználásával:

$$D_i = \frac{e_i^2 \cdot h_{ii}}{(p+1)s^2(1-h_{ii})^2} = r_i^2 \frac{h_{ii}}{(p+1)(1-h_{ii})} \quad (4.27)$$

Hüvelykujj-szabály alapján az egynél nagyobb D_i –t adó megfigyelésekre kell odafigyelni.

[67] Ha a változók korrelálatlanok, akkor megegyezik az euklideszi távolsággal.

A diagnosztikát segítő további mértékek a regressziós együtthatókat és a becsült értékeket vetik össze, mérve azok változását, ha egy-egy megfigyelést kihagyunk.

DfBeta[68] mutatóval a j-edik regressziós együttható[69] érzékenységét mérjük, ha az i-edik megfigyelést elhagyjuk:

$$DfBeta_{ji} = (b_j - b_{(i)j}) / c_{jj} s_{(i)} \tag{4.28}$$

ahol c_{jj} az együttható szórásától függő korrekciós tényező, négyzete az $(X^TX)^{-1}$ diagonálisában található. Figyelmet érdemel az i-edik megfigyelés, ha (4.28) abszolút értéke meghaladja a $2/\sqrt{n}$ küszöbszámot.

A sztenderdizált változatot a regressziós együttható sztenderd hibájával történő osztás után kapjuk, és az előjelet is figyelembe vesszük: $StDfBeta_i = DfBeta_i / s_b$.

Cook D mutatójához hasonlóan a becsült értékeket hasonlítja össze a DfFits mérték, amelyben a (24)-beli *r* helyett (4.25) szerinti *t* szerepel. A DfFits egyesítve mutatja azt a hatást, amit az i-edik megfigyelés kihagyása gyakorol az egyes regressziós együtthatókra, b_0–ra, b_1-re, stb.: $DfFits_i = \left(\hat{y}_i - \hat{y}_{(i)i} \right)$

Mivel az eltérést itt sem emeljük négyzetre, DfFits előjelét is vizsgálhatjuk. Az összehasonlíthatóság érdekében (4.29) szerint sztenderdizáljuk az eltéréseket, és az abszolút értékben $2\sqrt{p/n}$ -nél nagyobbakat kiemelten kezeljük:

$$StDfFits_i = \left(\hat{y}_i - \hat{y}_{(i)i} \right) / s_{(i)} \sqrt{h_{ii}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} \tag{4.29}$$

Végül a kovariancia-hányados mutatóval zárjuk a megfigyelések hatásának vizsgálatát. Az adatokból becsült kovariancia mátrixot (*S*) is képezhetjük az i-edik megfigyelés kihagyása után ($S_{(i)}$). Ezek determinánsainak hányadosa:

$$CovRatio = \frac{\left| S_{(i)} \right|}{\left| S \right|} = \frac{(n-p)^p}{\left[(n-p-1) + t_i \right]^p (1 - h_{ii})} \tag{4.30}$$

Ha a kovariancia-hányados értéke ~1, akkor nem jelentős az i-edik eset hatása.

Az összetevőket vizsgálva megállapítható[70], hogy $\left| CovRatio - 1 \right| \le 3p/n.$

[68] A Df rövidítés a differenciára utal.
[69] A konstans tag, b_0 is vizsgálható így.
[70] Belsey, Kuh és Welsch 1980-ban adták meg a felső határt.

Egyszerűbb alakot kapunk, ha egyetlen magyarázó változónk van. Ekkor azokra a megfigyelésekre kell különösen figyelnünk, amelyek kovariancia-hányadosa nagyobb, mint (1+3/n) vagy kisebb, mint (1-3/n).

4.8. A megvalósítás lépései az SPSS-ben

Az ANALYZE/REGRESSION/LINEAR utat követve a nyitó oldalon először

- a függő (dependent) változót és
- a független (independents) változókat kell megadni.

A **módszer** alapértelmezés szerint Enter, vagyis minden független változót bevon az eljárás. Mintapéldánkban lépésenként felépített (stepwise) modellt ismertetünk. A népességnövekedés becsléséhez 6 magyarázó változót jelöltünk ki.

- Megadható még „selection" változó, amellyel almintát képzünk, ezzel most nem élünk.
- Címkézzük az országok nevével az eseteket a „case label"-ben.

Az outputok listája a következő 4 gomb mögött tárul fel: **Statistics, Plots, Save, Options.**

A beállítás menete és az eredmények sorrendje jelentősen eltér. Először azt tekintjük át, hogy mit érdemes kérni, majd azt, hogy mit hogyan értelmezünk.

I. Statistics

- A regressziós együtthatók becslése mellett konfidencia intervallumot és kovariancia mátrixot kérhetünk.
- A modell illeszkedését, az R^2 változását, leíró statisztikát (átlag, szórás, megfigyelések száma), parciális korrelációt és multikollinearitási mértékeket választhatunk.
- A reziduális a Durbin-Watson tesztet és esetenkénti diagnosztikát kérhetünk. Ha az n nagy, érdemes csak az outlier eseteket kiíratni, amelyek az átlagtól 2-3 szórásnyi távolságra vannak.

II. Plots

A regressziós becslés összevethető a reziduálisok különböző fajtáival. A reziduálisok normális eloszlásáról a hisztogram és a normális eloszlástól való eltérés ad képet.

III. Save

Ez a gomb öt csoportba sorolva ajánlja fel az elmenthető eredményeket.

1. Becsült értékek (közönséges, sztenderdizált és korrigált becslés, valamint a becslés sztenderd hibája minden egyes megfigyelésre külön-külön)
2. Reziduálisok (közönséges, sztenderdizált, studentizált, kihagyott és kihagyva studentizált)
3. Távolságok egyenként mérve: Mahalanobis, Cook-D és az áttétel-hatás értékek
4. A befolyást mérő statisztikák (DfBeta és DfFit sztenderdizálva is, kovariancia hányados)
5. Konfidencia intervallum a regressziós becslés minden pontjára az átlaghoz és egy egyedi ponthoz képest, választható megbízhatósági szinten.

IV. Options

- A beléptetés az F-hez tartozó valószínűség (alapérték: Entry: 0,05, Removal: 0,10) vagy az F teszt értékének kiválasztásával szabályozható.
- Alapértelmezés szerint van konstans tag a modellben, de itt kihagyható.
- A hiányzó értékek páronkénti vagy soronkénti kihagyását, estleg az átlaggal való helyettesítését kérhetjük.

4.9. A számítási eredmények bemutatása

A népesség növekedési ütemét (y) becsüljük az SPSS-ben elérhető World95.sav adatállomány alapján. Az egyes táblák angol és magyar nevének megadása után röviden értékeljük a részeredményeket.

Descriptive statistics – a leíró statisztikák közül a változók átlagát és szórását, valamint a megfigyelések számát kapjuk meg. 109 ország adatai között sokszor hiányzik a napi kalória-bevitelt mérő változó. Ilyen esetben az alapértelmezés szerint a regressziós becslés az egész sort kihagyja („listwise"), ezért 75 adatból számolt statisztikákat kapunk. (4.5. táblázat) Az eredmények közül AIDS-esek számának relatív szórása[71] több mint 4, ez túlzott mértékű heterogenitást[72] jelent, a modellbe bevonni nem célszerű.

[71] Szórás/átlag= relatív szórás, a kettőnél nem nagyobb érték a kedvező. Az átlag előjelétől eltekintünk.
[72] Nincs népességre vetítve az adat, és az USA kiugróan magas betegszáma megnöveli a szórást.

4.5. táblázat: Leíró statisztikák

Descriptive Statistics

	Mean	Std. Deviation	N
Population increase (% per year))	1,821	1,143	75
Average female life expectancy	68,81	11,41	75
Average male life expectancy	63,88	10,11	75
Infant mortality (deaths per 1000 live births)	47,021	38,731	75
Gross domestic product / capita	5853,16	7149,52	75
Daily calorie intake	2753,83	567,83	75
Aids cases	**11067,40**	**48111,34**	75

Correlations: a függő és a magyarázó változókra páronkénti korrelációk, szignifikancia szintek és a minta mérete szerepel a táblázatban. A multikollinearitás már itt észlelhető, egyes magyarázó változók között szinte függvényszerű kapcsolat van. Az AIDS változó nem korrelál szignifikánsan a népességnövekedéssel, bevonásra nem kerülhet. (4.6. táblázat)

4. 6. táblázat: Korrelációs mátrix

Correlations

	Population increase (% per year))	Average female life expectancy	Average male life expectancy	Infant mortality (deaths per 1000 live births)	Gross domestic product / capita	Daily calorie intake
Population increase (% per year))	1,000	-,582	-,529	,617	-,665	-,609
Average female life expectancy	-,582	1,000	,989	-,962	,675	,775
Average male life expectancy	-,529	**,989**	1,000	-,946	,657	,765
Infant mortality (deaths per 1000 live births)	,617	**-,962**	**-,946**	1,000	-,690	-,777
Gross domestic product / capita	-,665	,675	,657	-,690	1,000	,751
Daily calorie intake	-,609	,775	,765	-,777	,751	1,000
Aids cases	-,094	,044	,032	-,075	,285	,167

Bevont és kihagyott változók lépésenkénti felsorolása: a 2. lépésben bevont csecsemőhalandóságot az 5. lépésben eltávolítja a stepwise eljárás.

A Model Summary táblázatban (4.7. táblázat) a többszörös korreláció és determinációs együttható, a korrigált R^2, a regressziós modell standard hibája szerepel lépésenként. Mivel az ötödik lépésben redukáltuk a modellt, az összes mutató csökkent. A Durbin-Watson tesztet nem értelmezzük.

4.7. táblázat: A változások követése

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin - Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,665	,443	,435	,859	,443	58,006	1	73	,000	
2	,700	,491	,476	,827	,048	6,751	1	72	,011	
3	,722	,521	,501	,808	,031	4,558	1	71	,036	
4	,752	,565	,540	,775	,044	7,015	1	70	,010	
5	,745	,555	,536	,779	**-,010**	1,640	1	72	,204	1,887

Az R^2 változását az előző és az adott lépésbeli mérték különbsége adja, a változás jelentőségét az F-teszt alapján ítélhetjük meg. Az F-próba változásának szignifikanciáját is F-teszt méri.

Az ANOVA táblázat is lépésenként készül. Az MSR, az MSE és az F-hányados az első négy lépesben fokozatosan csökken, majd az ötödik lépésben a redundáns változó elhagyása után mindhárom magasabb lesz. (4.8. táblázat)

4.8. táblázat: Szórásnégyzet felbontása lépésenként

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42,832	1	42,832	58,006	,000
	Residual	53,904	73	,738		
	Total	96,737	74			
2	Regression	47,453	2	23,727	34,663	,000
	Residual	49,283	72	,684		
	Total	96,737	74			
3	Regression	50,426	3	16,809	25,770	,000
	Residual	46,310	71	,652		
	Total	96,737	74			
4	Regression	54,644	4	13,661	22,719	,000
	Residual	42,092	70	,601		
	Total	96,737	74			
5	Regression	53,658	3	17,886	29,479	,000
	Residual	43,079	71	,607		
	Total	96,737	74			

A regressziós együtthatók becslése az elemzés célja.
Az együtthatókat sztenderd hibáikkal osztva a t-teszt értékét kapjuk. A lépésenkénti eljárás hatására csak a nullától szignifikánsan különböző együtthatójú változók maradnak a modellben. Ha az induló adatokat sztenderdizáljuk, akkor egyből sztenderdizált együtthatókat, bétákat kapunk, amelyek az x 1%-os változásának y-ra gyakorolt hatását fejezik ki.

A táblázatban szereplő zero-order korrelációk az adott x és az y közötti közönséges Pearson korrelációk. A parciális korrelációk (4.19) a már bevont magyarázó változók hatását szűrik ki, ezért alacsony értékük (például a 4. lépésben a női várható élettartam bevonása után a csecsemőhalandóság) multikollinearitásra utal. A rész-korreláció a parciális korreláció számlálója.

A kollinearitási statisztika két mutatót ad. A tolerancia=1- R_i^2, azaz az i-edik változónak az összes többi magyarázó változóval való determinációs együtthatójának komplementere. Értéke 1, ha egy magyarázó változó van, utána egyre csökken. Már a 3. lépésben erős multikollinearitás van, amint azt a korrelációs mátrixnál is észleltük.

A VIF a tolerancia reciproka. A 4. lépéstől az egymással szorosan korreláló férfi és női várható élettartam együtt szerepel a végső modellben, ezért a VIF túl magas, két változóra is öt felett van. (4.9. táblázat) Ezek alapján a modell alkalmazása megkérdőjelezhető.

4.9. táblázat: A regressziós modell együtthatói

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2,444	,129		19,007	,000					
	Gross domestic product / capita	,000	,000	-,665	-7,616	,000	-,665	-,665	-,665	1,000	1,000
2	(Constant)	1,830	,267		6,861	,000					
	Gross domestic product / capita	,000	,000	-,457	-3,938	,000	-,665	-,421	-,331	,524	1,907
	Infant mortality (deaths per 1000 live births)	,009	,003	,302	2,598	,011	,817	,293	,219	,524	1,907
3	(Constant)	-2,772	2,171		-1,277	,206					
	Gross domestic product / capita	,000	,000	-,462	-4,072	,000	-,665	-,435	-,334	,524	1,908
	Infant mortality (deaths per 1000 live births)	,024	,008	,809	3,073	,003	,817	,343	,252	,097	10,292
	Average male life expectancy	,061	,029	,540	2,135	,036	-,529	,246	,175	,105	9,491
4	(Constant)	,389	2,402		,162	,872					
	Gross domestic product / capita	,000	,000	-,435	-3,976	,000	-,665	-,429	-,313	,520	1,925
	Infant mortality (deaths per 1000 live births)	,011	,009	,384	1,281	,204	,817	,151	,101	,069	14,447
	Average male life expectancy	,209	,062	1,850	3,357	,001	-,529	,372	,265	,020	48,859
	Average female life expectancy	-,175	,066	-1,749	-2,649	,010	-,582	-,302	-,209	,014	70,160
5	(Constant)	3,331	,705		4,722	,000					
	Gross domestic product / capita	,000	,000	-,461	-4,268	,000	-,665	-,452	-,338	,538	1,859
	Average male life expectancy	,221	,062	1,953	3,566	,001	-,529	,390	,282	,021	47,823
	Average female life expectancy	-,221	,066	-2,203	-3,934	,000	-,582	-,423	-,312	,020	49,982

a. Dependent Variable: Population increase (% per year))

Az éves népesség növekedést becslő egyenletben a konstans (3,331) mellett a GDP/fő és a férfi valamint a női várható élettartam szerepel. Ez a három magyarázó változó egymással is szorosan korrelál – a tolerancia alacsony, a VIF pedig túl magas – ezért a modellben gondok lesznek. A sztenderdizált regressziós együtthatók alapján a női várható élettartam hatása a legerősebb, mivel a -2,203 abszolút értékben meghaladja a másik két bétát.

A modellben nem szereplő változók listájából a következő lépést lehet megállapítani. A (k+1) lépésben az a változó kerül bevonásra, amelynek a legnagyobb (és még szignifikáns) a t-tesztje. (4.10. táblázat)

A sajátértékek és a kondíciós indexek a 4.11. táblázatban találhatók. Látható, hogy minden lépésben egy nagy[73] sajátérték van, ami arra utal, hogy maximum két független dimenzió van, amibe a magyarázó változók tömöríthetők. A kondíciós index már a 3. lépésben meghaladja a veszélyes szintet, a 30-t. A regressziós együtthatók varianciáinak szétosztása nem sikerült, már a második lépés magas variancia hányadot jelez. (A számok százalékosan értelmezhetők.) A magyarázó változók mögött azonos sajátérték húzódik meg, ezért tömöríthetők, egymástól nem függetlenek. Ilyen esetben érdemes főkomponens vagy faktor előállítására gondolni.

[73] Az egységnyinél nagyobb sajátérték számít „nagynak”. Erről részletes magyarázatot a főkomponensek ismertetésekor adunk.

4.10. táblázat: A még be nem vont változók statisztikái

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2,444	,129		19,007	,000					
	Gross domestic product / capita	,000	,000	-,665	-7,616	,000	-,665	-,665	-,665	1,000	1,000
2	(Constant)	1,830	,267		6,861	,000					
	Gross domestic product / capita	,000	,000	-,457	-3,936	,000	-,665	-,421	-,331	,524	1,907
	Infant mortality (deaths per 1000 live births)	,009	,003	,302	2,598	,011	,617	,293	,219	,524	1,907
3	(Constant)	-2,772	2,171		-1,277	,206					
	Gross domestic product / capita	,000	,000	-,462	-4,072	,000	-,665	-,435	-,334	,524	1,908
	Infant mortality (deaths per 1000 live births)	,024	,008	,809	3,073	,003	,617	,343	,252	,097	10,292
	Average male life expectancy	,061	,029	,540	2,135	,036	-,529	,246	,175	,105	9,491
4	(Constant)	,389	2,402		,162	,872					
	Gross domestic product / capita	,000	,000	-,435	-3,976	,000	-,665	-,429	-,313	,520	1,925
	Infant mortality (deaths per 1000 live births)	,011	,009	,384	1,281	,204	,617	,151	,101	,069	14,447
	Average male life expectancy	,209	,062	1,850	3,357	,001	-,529	,372	,265	,020	48,859
	Average female life expectancy	-,175	,066	-1,749	-2,649	,010	-,582	-,302	-,209	,014	70,160
5	(Constant)	3,331	,705		4,722	,000					
	Gross domestic product / capita	,000	,000	-,461	-4,268	,000	-,665	-,452	-,338	,538	1,859
	Average male life expectancy	,221	,062	1,953	3,566	,001	-,529	,390	,282	,021	47,823
	Average female life expectancy	-,221	,056	-2,203	-3,934	,000	-,582	-,423	-,312	,020	49,982

a. Dependent Variable: Population increase (% per year))

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Gross domestic product / capita	Infant mortality (deaths per 1000 live births)	Average male life expectancy	Average female life expectancy
1	1	1,636	1,000	,18	,18			
	2	,364	2,120	,82	,82			
2	1	2,081	1,000	,03	,03	,03		
	2	,848	1,567	,00	,22	,10		
	3	7,095E-02	5,416	**,97**	**,74**	**,87**		
3	1	3,027	1,000	,00	,02	,00	,00	
	2	,854	1,883	,00	,20	,02	,00	
	3	,117	5,077	,00	,78	,11	,01	
	4	1,048E-03	**53,754**	**1,00**	,00	**,87**	**,99**	
4	1	4,000	1,000	,00	,01	,00	,00	,00
	2	,858	2,160	,00	,18	,02	,00	,00
	3	,141	5,318	,00	,80	,07	,00	,00
	4	1,121E-03	59,727	,86	,00	,73	,10	,02
	5	2,202E-04	134,760	,14	,01	,19	**,90**	**,98**
5	1	3,538	1,000	,00	,02		,00	,00
	2	,451	2,800	,00	,57		,00	,00
	3	1,038E-02	18,464	,97	,38		,01	,01
	4	2,604E-04	116,571	,03	,03		,99	,99

a. Dependent Variable: Population increase (% per year))

4.11. táblázat: Sajátértékek és variancia hányadok

A reziduálisok statisztikái

Először két országot látunk a 4.12. táblázatban, amelyek sztenderdizált reziduálisa kívül esnek a (-3;+3) intervallumon. Mindkettőnek pozitív előjele van, azaz a modell alulbecsli a megfigyelt értéket. Felülbecslés negatív reziduális esetén fordul elő.

4.12. táblázat: Kilógó megfigyelések

Casewise Diagnostics[a]

Case Number	COUNTRY	Std. Residual	Population increase (% per year))	Predicted Value	Residual
80	Kuwait	4,497	5,2	1,737	3,503
87	U.Arab Em.	4,348	4,8	1,413	3,387

a. Dependent Variable: Population increase (% per year))

A további (4.22)-(4.25) képletek szerint számított reziduálisokat megfigyelésenként az adatállományhoz csatolja az SPSS, míg a főbb statisztikai jellemzőket összefoglaló táblába rendezve kapjuk meg. (4.13. táblázat)

4.13. táblázat: A reziduálisok statisztikái

Residuals Statistics[a]

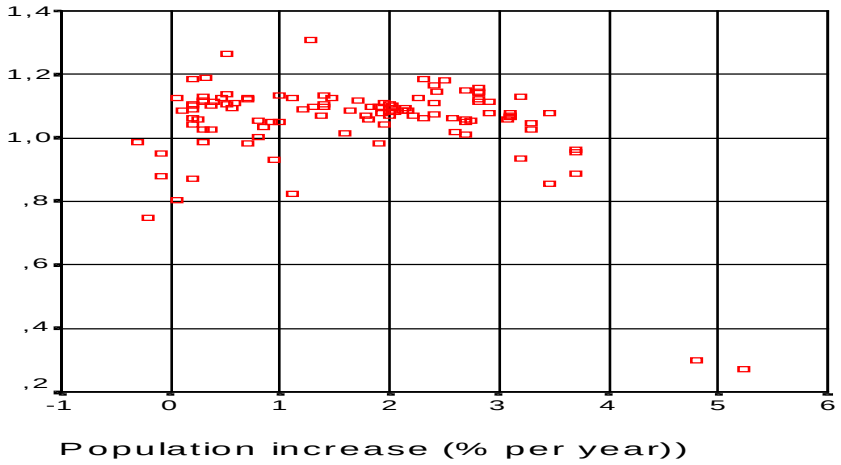
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	,149	3,544	1,753	,812	109
Std. Predicted Value	-1,964	2,023	-,079	,954	109
Standard Error of Predicted Value	,106	,348	,178	5,001E-02	109
Adjusted Predicted Value	9,415E-02	3,544	1,756	,818	109
Residual	-1,936	3,503	-7,10E-02	,850	109
Std. Residual	-2,485	4,497	-,091	1,091	109
Stud. Residual	-2,450	4,572	-,089	1,095	109
Deleted Residual	-1,936	3,620	-7,33E-02	,873	109
Stud. Deleted Residual	-2,452	5,404	-,083	1,134	109
Mahal. Distance	,375	13,787	3,095	2,367	109
Cook's Distance	,000	,196	,014	,029	109
Centered Leverage Value	,005	,186	,042	,032	109

a. Dependent Variable: Population increase (% per year))

Itt megállapítható, hogy a különböző reziduálisok mindegyike inkább jobbra ferde, mint szimmetrikus, hisz a reziduálisok átlagai a minimum és a maximum között nem középen vannak. A reziduálisok az adatállományban egyenként is megőrzésre kerülnek, és részletesen értékelhetők a különböző hibatagok. Minden hiba-számítási mód mellett a 4.12. táblázatban látott két megfigyelés, a 80. Kuwait és a 87. Egyesült Arab Emirátusok lóg ki a megfigyelések közül. Ezen országok illeszkedése

is gyenge. A (4.29) képlet szerinti StDfFits értékeket úgy kapjuk meg, ha az origóhoz igazított hatás-értékekhez hozzáadjuk az 1/n=1/75 számot.
Az origóhoz tolt hatás (leverage) maximális mértéke alatta marad az óvatosságra intő 0,2 küszöbnek. A maximális értéket Brazília éri el, ezért a (26) összefüggés alapján a Mahalanobis távolság maximuma (13,787) is Brazíliához tartozik. Lettország (11,5) és Ukrajna (9,5) távolságai szintén nagyok. Ugyanakkor a Cook-féle távolság sehol sem haladja meg az egyet, ezért igazi outliereket nem tudunk azonosítani.

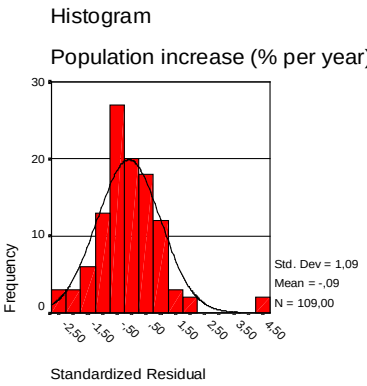
A kovariancia hányados erősen ingadozik az egy körül, többször kilép abból a sávból, amit az 1±3p/n képlet megad. (4.3. ábra) Nagyobb a kovariancia mátrix determinánsa, ha Brazíliát vagy Lettországot hagyjuk ki (1,2 feletti hányadosok). Csökken a determináns, ha Kuvait vagy az Egyesült Arab Emirátusok marad ki (0,4 alatti CR).



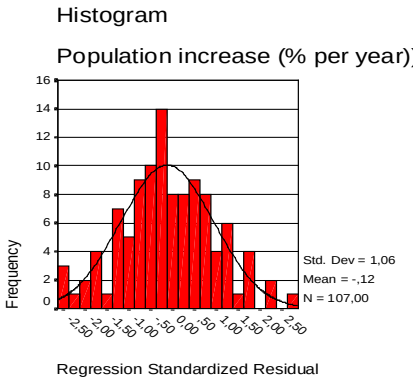
4.3. ábra: Kovariancia hányados

Reziduálisok ábrái

a) A reziduálisok statisztikáiból láttuk, hogy a 80. és 87. országok rontják az illeszkedést. A 4.4/a. ábrán még e két ország reziduálisai is szerepelnek, míg a 4.4/b hisztogram a kihagyásukkal készült regressziós modell sztenderdizált hibatagjait mutatja.



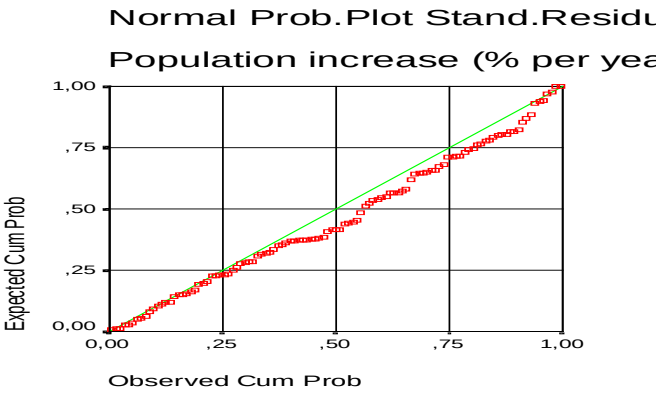
4.4/a. ábra 109 ország hibatagjai



4.4/b. ábra: 107 ország hibatagjai

b) Normális valószínűség ábrája: Ha a reziduálisok normális eloszlást követnek, a pontok a 45 fokos egyenes mentén helyezkednek el. A sztenderdizált reziduálisokat és a normális eloszlás feltételezésével várt hibatagokat jelző pontok a 4.5. ábrán nem esnek az egyenesre, de nincs is markáns eltérés köztük.

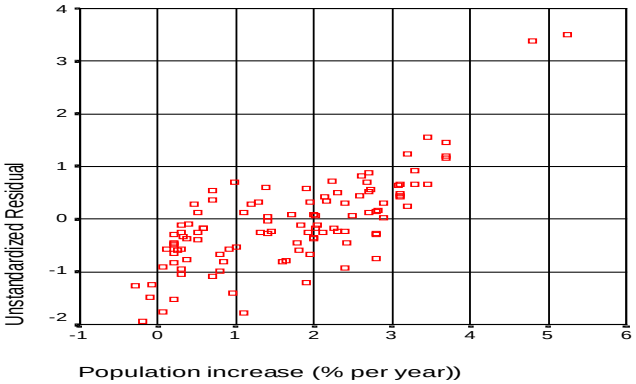
Általában elmondható, hogy az egyenes alatti vagy feletti pontok a szimmetria hiányát jelzik. Az egyenes elejénél vagy végénél lévő néhány távoli pont kilógó megfigyelésekre utalna. Ha a pontsorozat távolodik, akkor lapult vagy csúcsos az eloszlás.



4.5. ábra: A sztenderd reziduálisok normális eloszlásának grafikus vizsgálata

c) Szokás az is, hogy a vízszintes tengelyen y-t vagy valamelyik x változót, a függőleges tengelyen a reziduálist tüntetjük fel. A nulla körüli, nem növekvő, függvénykapcsolatot nem mutató reziduálisok a lineáris modell megbízhatóságát

támasztják alá. A 4.6. ábrán a 80. és 87. országok a nagy reziduálisok miatt külön állnak, és a hibatagok növekednek[74].

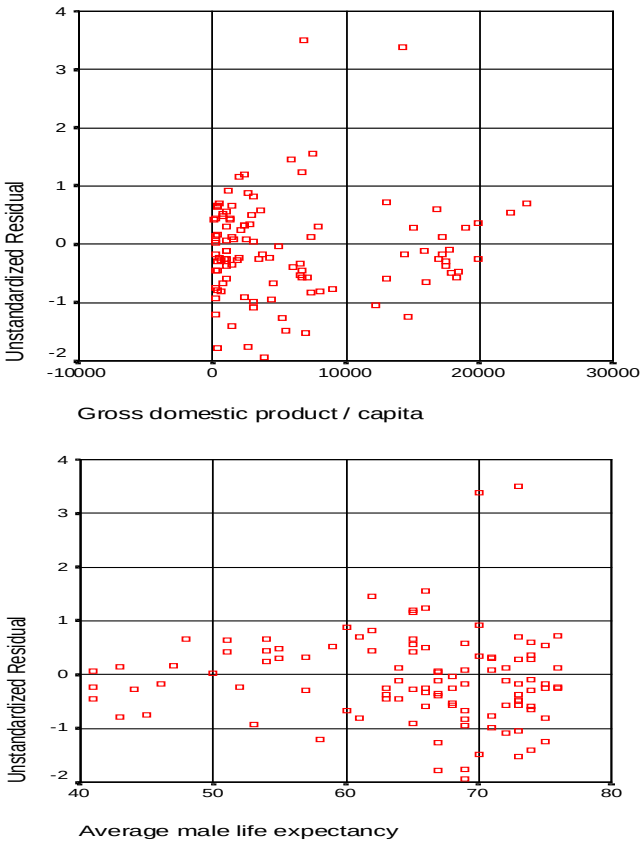


4.6. ábra: A függő változó mentén növekvő reziduálisok

Az átlag körüli és az egyedi megfigyelésekhez tartozó 95%-os megbízhatósági szintű konfidencia sávok is ábrázolhatók a Graphs/Line/Multiple beállítással. Nagyon sok ország megfigyelt népességnövekedése esik kívül az alsó és a felső becsült értéken.

A független változók közül kettőt kiválasztva mutatjuk be a hibatagok viselkedését. A 4.7. ábrán a GDP/fő változóra csökkenő, a férfiak várható élettartamára vetítve növekvő reziduálisokat látunk.

[74] Ilyenkor adat-transzformációt ajánlott alkalmazni, pl. y vagy x, esetleg mindkettő logaritmusát célszerű venni.



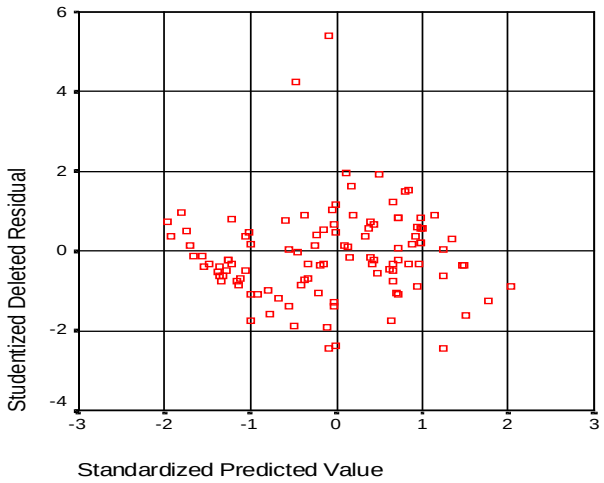
4.7. ábra: A magyarázó változók és a reziduálisok

d) A Studentizált – sorkihagyással számolt – reziduálisokat az y tengelyen, a standardizált becsült értékeket az x tengelyen ábrázolva a modell érvényessége jól szemléltethető. A normalitás teljesülése esetén a reziduálisok 95%-a (-2;+2) közé esik. A 4.8. ábrán felfelé nagyon kilógó (80, 87) országokat már azonosítottuk. Lefelé haladva kicsivel (–2) alatt találjuk balról jobbra haladva Bulgáriát, Romániát és Kínát, ahol jóval kevesebb gyerek születik, mint amennyit a modell alapján várunk. Éppen 5 kilógó ország fér bele száz körüli minta esetén a 95%-os tartományba.

Itt a minta mérete és a hiányzó adatok kezelése kapcsán fontos technikai megjegyzést kell tennünk:

- 109 ország van a World95.sav-ban. De csak 75 országnak van teljes adatsora a regressziós modellben felsorolt függő és magyarázó (1+6) változóra. Ezért a táblák egy részében, például a 7. és 8. táblázatban n=75-ből számolt szabadságfok szerepel.

- A változószelekciót követően azonban kimarad az a három magyarázó változó, amelyeknek 34 országra hiányzik értéke. Így a felépített regressziós modellt már 109 ország adataiból becsülte az SPSS. Reziduálist is 109 országra számol és ábrázol a számítógép.



4.8. ábra: Melyik országok nélkül lenne nagyon más a regressziós egyenes?

4.10. Összefoglalás: A bemutatott modell illeszkedésének minősítése

Az adathalmaz kiválasztott változóin szinte a regressziószámítás összes gyengéjét sikerült bemutatni, miközben a 4 magyarázó változóval készített lépésenkénti modell minden teszten „átment". Mégis felmerültek az alábbi problémák:

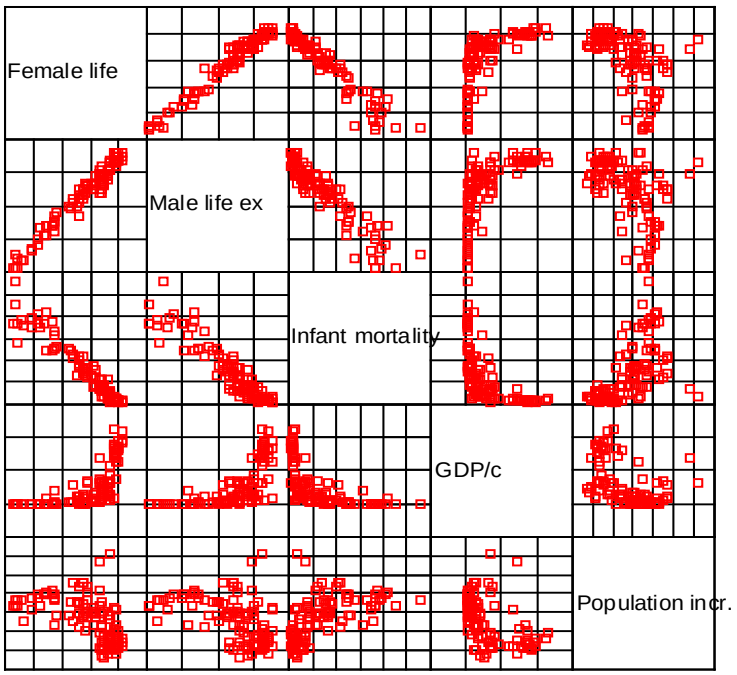
- a megfigyelések halmaza nem homogén,
- a magyarázó változók nem függetlenek,
- a determinációs együttható nem elég magas,
- a reziduálisok kívül esnek a kedvező tartományon, és szórásuk nem konstans.

Mit lehet tenni a modell javítása érdekében?

Ezek a problémák nem egymástól függetlenül jelentkeznek. Ha kihagyjuk például a két outlier országot (80 és 87), akkor az R^2 0,54-ről 0,64-re nő.

De a gyenge modell legfőbb oka az, hogy a lineáris modell feltételezése nem állja meg a helyét. A függő változó és a magyarázó változók kapcsolata nem írható le lineáris függvénnyel, amint ezt a 4.9. ábra mutatja. Az első három magyarázó változó szoros lineáris kapcsolatban van, ami erős

multikollinearitást okoz, a GDP hatása viszont nem lineáris. A lépésenkénti regresszió a megadott változók közül készítette el a lehető legjobb becslést, ami szakmai értelemben nem jó, további elemzésekre nem alkalmas.



4.9. ábra: Változó-párok pontdiagramja

A változó-transzformációkra és a nemlineáris regresszióra itt nem térünk ki, mivel a jegyzetben ismertetésre kerülő többi sokváltozós eljárás megalapozásához a lineáris regressziós modell szükséges.

4.11. Önálló elemzési feladatok

Válaszoljon az alábbi kérdésekre és a Kerületek2010.sav adatállományból számolva ellenőrizze az elgondolásait.

1. feladat
Legyen a függőváltozó az Önkormányzati bevétel.
Kívánalom, hogy normális eloszlású legyen? igen/nem
Milyen módon ellenőrizhető, hogy teljesül-e a normalitás?
a) Grafikusan:
b) Numerikusan:

2. feladat
A magyarázó változók közé választandó a következő 7 változó:
Népességszám
Odavándorlás
Elvándorlás
Vendéglátóhely
Lakásállomány
Épített lakások
Álláskeresők
a) A relatív szórások kettő alatt vannak?
b)A magyarázó változók közötti korrelációk szignifikánsak?
c) A STEPWISE eljárás fontos? Igen/nem

3. feladat
Elemezze **együtt,** egy regressziós modellben az 50 települését az 1. és a 2. feladat változói alapján.
a) Hány magyarázó változó került bevonásra?
b) Milyen a modell illeszkedése?
c) A reziduálisok viselkedése megfelelő-e?
d) Vannak-e kilógó kerületek/települések az adatok között?
e) A „kerület" státusz változó dummy-ként szerepelhet-e a modellben? Igen/nem Bevonásra kerül? Igen/nem
HOMOGÉN az adathalmaz, közös tendencia jellemző a kétféle településre? Igen/nem

4. feladat
Külön illesztendő lineáris regressziós modell a 23 kerületre és a többi 27 falura/városra.
a) Más magyarázó változók kerülnek be a két modellbe?
b) Melyik modell illeszkedik jobban?
c) Melyek a kilógó kerületek/települések az adatok között?

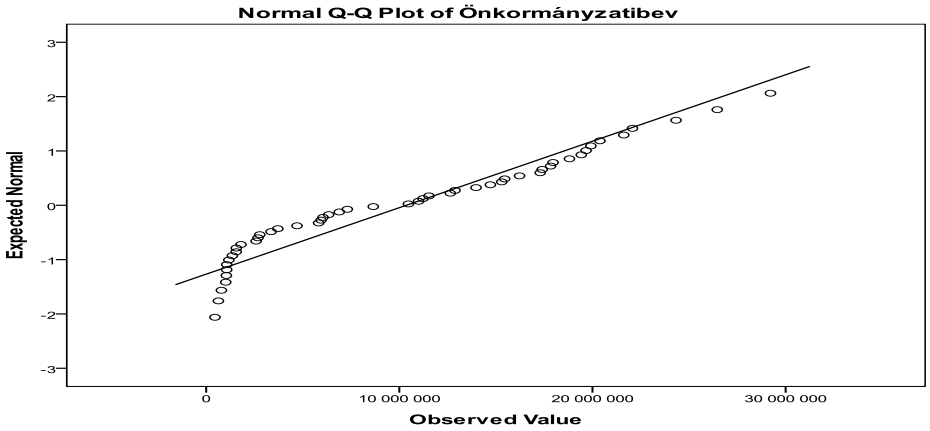
5. feladat
a) Hogyan lehet csak konstansban eltérő modellt illeszteni két almintára?
b) Ha magas a kondíciós index, akkor a regressziós modell helyett melyik módszer alkalmazása lehet indokolt?

4.12. Megoldások

1. feladat

A függővváltozó (Önkormányzati bevétel) normális eloszlása elvárás. A normalitás ellenőrizhető grafikusan és numerikusan is.
a) Grafikusan két lehetőség is adódik:
i) Hisztogram

ii) QQ plot: a 45 fokos egyenestől a kisebb értékeknél tapasztalunk eltérést, azaz a kis önkormányzati bevétellel rendelkező települések gyakoribbak, mint a normális eloszlás szerint várt előfordulás.



b) Numerikusan több adatot nézhetünk:

i) ferdeség 0,417 ±2*0,337 és csúcsosság -1,002±2*0,662 mérőszámok konfidencia intervallumai tartalmazzák a nullát, az eltérés nem szignifikáns

ii) Kolmogorov-Szmirnov vagy Shapiro-Wilk teszt (éppen n=50 a megfigyelések száma)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Önkormányzatibev	,132	50	,029	,916	50	,002

a. Lilliefors Significance Correction

Mindkét tesztre 5%-os szignifikancia szinten elvethető a normális eloszlás.
Teljesül a normalitás? Nem egyértelmű a válasz! De a ferdeség és csúcsosság értékekre tekintettel elkészítjük a regressziós becslést.

Figyelem! Érdemes kipróbálni a következőket, bár egyik révén sem kapunk a tesztek szerint normális eloszlást:

- az önkormányzati bevétel logaritmusa normális eloszlású-e

- az egy főre jutó önkormányzati bevétel eloszlása milyen?

- az egy főre jutó önkormányzati bevétel logaritmusa milyen alakú?

2. feladat

a) A relatív szórások (szórás/átlag hányadosok) közül néhány meghaladja az egyet, de a kettőt egyik sem közelíti meg. Tehát a változók mentén a minta nem heterogén.

b) A 7 magyarázó változók közötti páronkénti korreláció mind szignifikáns és pozitív. A legkisebb r=0,629 (épített lakások és vendéglátóhely), a legnagyobb r=0,980 (népesség szám és lakásállomány)

c) A Stepwise eljárás fontos, mert nagyon jelentős multikollinearitás áll fenn.

3. feladat

Lineáris regressziós modellben az 50 település adatai alapján az önkormányzati bevétel becslésére

a) 4 lépésben 3 magyarázó változót von be, de csak kettőt tart bent.

1. lépés: az önkormányzati bevétellel legerősebben korreláló lakásállomány bevonása

2. lépés: a vendéglátóhely változó bevonása

3. lépés: népességszám bevonása

4. lépés: a népesség és a lakás változók erős korrelációja miatt lakásállomány változó kihagyása

Itt fontos figyelni arra, hogy ez a „legjobb" regressziós modell, ami a korlátozó feltételeket figyelembe véve felépíthető. De vajon a kiválasztott két változó helyett mind a hét magyarázó változó főkomponensbe tömörítve, egyetlen faktorként nem ad-e jó, használható becslést az önkormányzati bevételre? Az önkormányzati bevétel és a 7 változóból (83%-ot megőrző) faktor közötti korreláció= 0,899.

b) A modell illeszkedése nagyon jó, a korrigált R-négyzet 0,858.

- Az F-tesztek minden lépésben alátámasztják a lineáris modell létét.

- A két változó tolerancia értéke 0,379, és a variancia infláló faktor 2,641, ami nem túl magas. (Két magyarázó változó esetén indokolt, hogy közös a Tol és a VIF érték, hisz egymást magyarázzák.)

- A kondíciós index 5,222 értéke sem jelez a két változó és a konstans között túlzott erejű kapcsolatot.

c) A reziduálisok eloszlása a hisztogramon normálishoz közeli alakú.

Correlations

		Önkormányzatibev	Népességszám	Odavándorlás	Elvándorlás	Vendéglátóhely	Lakásállomány	Építettlakások	Álláskeresők
Pearson Correlation	Önkormányzatibev	1,000	,904	,764	,867	,845	,907	,581	,844
	Népességszám	,904	1,000	,859	,945	,788	,980	,650	,927
	Odavándorlás	,764	,859	1,000	,933	,715	,851	,789	,743
	Elvándorlás	,867	,945	,933	1,000	,758	,919	,721	,872
	Vendéglátóhely	,845	,788	,715	,758	1,000	,845	,629	,716
	Lakásállomány	**,907**	**,980**	,851	,919	,845	1,000	,691	,887
	Építettlakások	,581	,650	,789	,721	,629	,691	1,000	,602
	Álláskeresők	,844	,927	,743	,872	,716	,887	,602	1,000

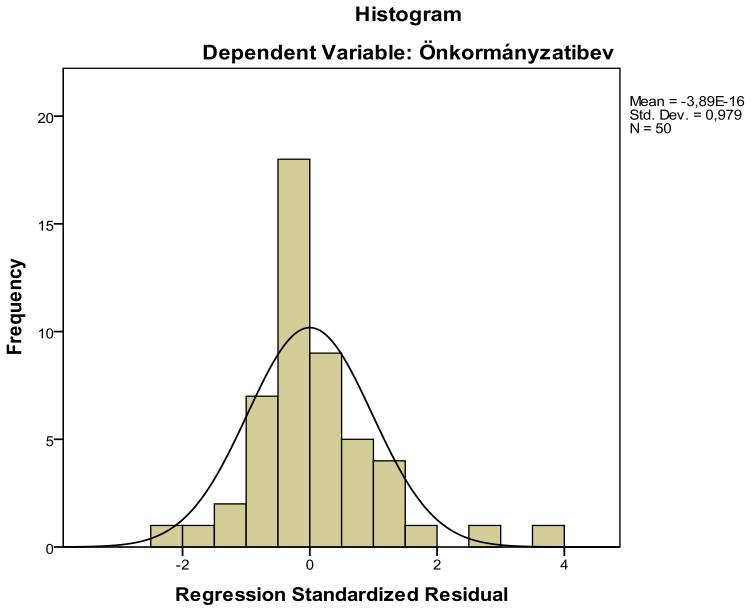
2. feladat táblázata

Model Summary[e]

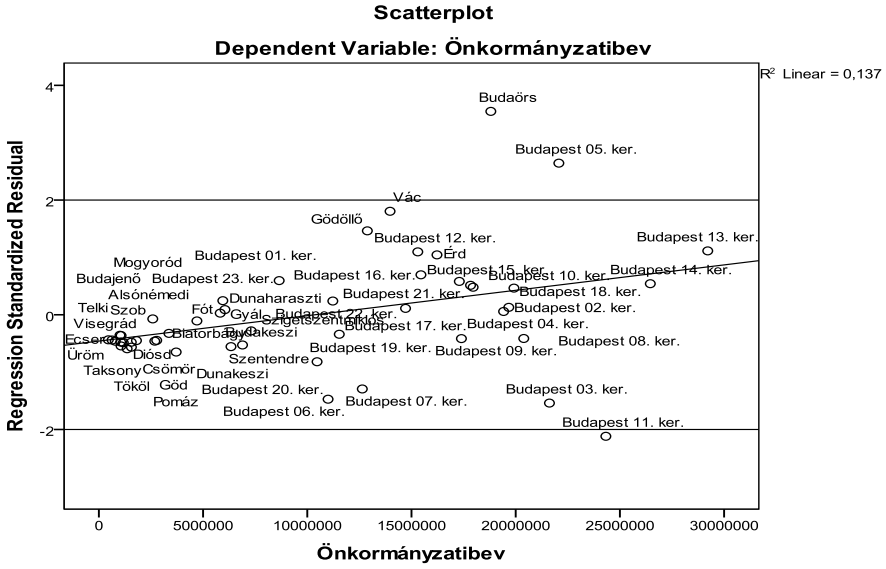
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,907[a]	,822	,818	3483896,503	,822	221,722	1	48	,000
2	,918[b]	,844	,837	3300174,144	,022	6,493	1	47	,014
3	,930[c]	,864	,855	3107961,410	,021	6,993	1	46	,011
4	,929[d]	,863	,858	3084032,434	-,001	,279	1	46	,600

a. Predictors: (Constant), Lakásállomány
b. Predictors: (Constant), Lakásállomány, Vendéglátóhely
c. Predictors: (Constant), Lakásállomány, Vendéglátóhely, Népességszám
d. Predictors: (Constant), Vendéglátóhely, Népességszám
e. Dependent Variable: Önkormányzatibev

3.feladat táblázata



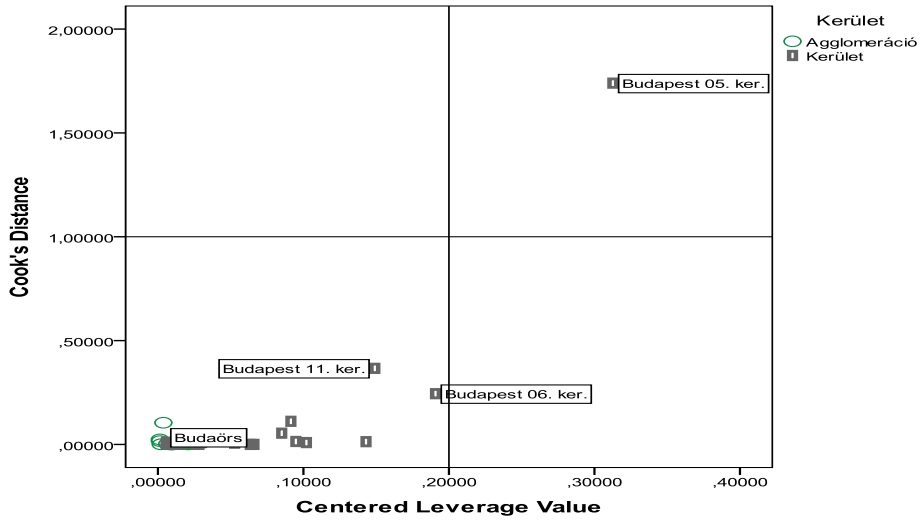
A pontok szórása enyhe növekedést mutat, a nagyobb önkormányzati bevételhez kicsit magasabb sztenderdizált rezidálisok tartoznak (R-négyzet=0,137). Csak Budaörs (3,546) és az V. kerület esik kívül a [-2;+2] intervallumon, míg a XI. kerület a határ közelében van.



d) Vannak-e kilógó kerületek/települések az adatok között?

- Itt a sztenderd reziduálisok ábrája alapján Budaörs és az V. kerület említhető. Mindkettőnek alulbecsli az önkormányzati bevételét a modell.

- Az egyedi áttétel hatások és a Cook-távolság terében vizsgálva egyedül az V. kerület kerül a kritikus értékeken kívülre.



Érdemes átgondolni, hogy Budaörs és Budapest V. kerület miben térnek el és miben hasonlítanak:

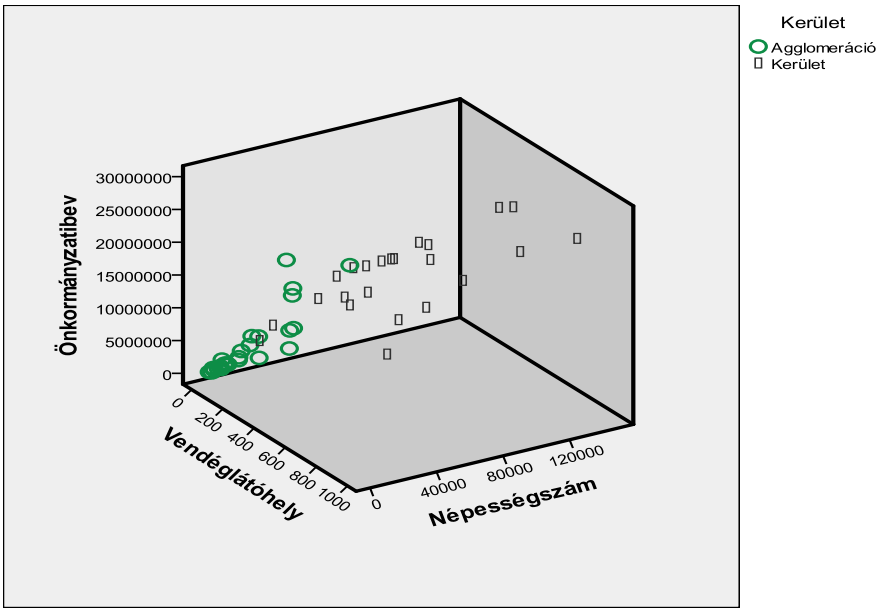
- A reziduálisaik nagyok, 2-4 közötti értékük azt jelzi, hogy jelentősen alulbecsülte a modell az ott mért önkormányzati bevételeket. Itt más magyarázó változók figyelembe vétele is indokolt lenne.

- Az áttétel hatása egyiknek sem éri el a 0,5-öt, tehát egyik elhagyása sem indokolt. De a 0,2 és 0,5 közötti érték arra utal, hogy a V. kerület a becslésre erősen hat.

- a Cook-távolság csak az V. kerületre magas. Ha elhagynánk az V. kerületet a regressziós becslés során, akkor a 49 pontból készített regressziós becslés jelentősen eltérne az 50 pontból számolt modelltől.

e) A „kerület" státusz változó dummy-ként szerepelhetne a modellben, de nincs szignifikáns szerepe, ezért nem került bevonásra. Ez azt is jelenti, hogy a kerületekre és az agglomeráció településeire nem egymással párhuzamos modell illeszkedik.

Az adatállomány két része homogén, közös – lineáris – tendencia jellemzi a három változó kapcsolatát, amint ezt a 3D-s pontdiagram is mutatja.



4. feladat

Ha külön illesztünk lineáris regressziós modellt a 23 kerületre és a többi 27 falura/városra, nagyon eltérő regressziós modelleket kapunk. A számításokat két úron végezhetjük el:

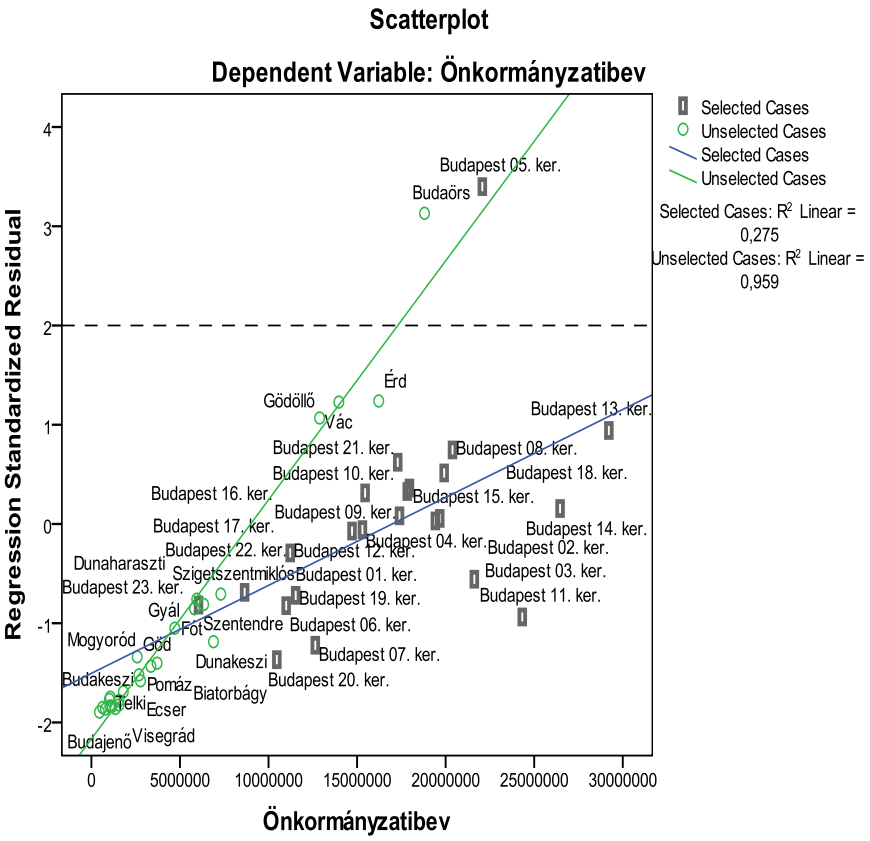
i) A regressziós modellezésen belül Selection: Kerület=1 (majd 0) beállítással futtatva mind a kiválasztott, mind a másik almintára elkészül a becslés, és mindkét részre megkapjuk a főbb statisztikai jellemzőket.

ii) Ha előre leszűrjük az egyik almintát, és csak ezt használjuk a regressziós blokkban, akkor a másik almintára nem kapunk semmilyen eredményt.

Most az i) szerint jártunk el, és két részmodell eredményeit vetjük össze az a) –b) –c) kérdések mentén.

a) –c) kérdések	Kerületi adatok saját modellje	*Kerületi adatok agglomerációra*	Agglomerációs adatok saját modellje	*Agglomerációs adatok kerületre*
magyarázó változó(k)	lakásállomány	*lakásállomány*	Odavándorlás Építettlakások Elvándorlás	*Odavándorlás* *Építettlakások* *Elvándorlás*
modell illeszkedése	Többszörös R=0,851	*Többszörös* *R=0,854 (!)*	Többszörös R=0,939	*Többszörös* *R=0,522*
kilógó települések	V. kerület	*Budaörs*	nincs	*13 kerület*

A kilógó kerületek/települések az adatok között nemcsak a felsorolásból, hanem a sztenderd reziduálisok ábrájáról is látható. Itt csak a kerületi adatok modelljéből számolt reziduálisokat mutatjuk be, de mindkét almintára. Látható, hogy az agglomeráció településeire határozottan növekednek a reziduálisok, tehát ott további magyarázó változók bevonása indokolt. Ez teljesül is, hisz az agglomerációra illesztett modellben 3 magyarázó változó szerepel. Ugyanakkor 3 három változós modellben a vándorlási mutatók VIF-értéke 40 feletti és a kondíciós index 26,687, a multikollinearitás tehát túl erősen van jelen. Mindent összevetve a két alminta együttes kezelésével statisztikai értelemben jobb modellt kaptunk.



5. feladat

a) Csak konstansban eltérő modellt illeszteni két almintára úgy lehet, hogy az almintát azonosító dummy (d= 0 vagy 1) változót a modellbe bevonjuk. Így $y = b_0 + \underline{b}\underline{x} + b_2 d$ az alapmodell lesz, ha d=0. Míg d=1-re b_2–vel magasabb vagy alacsonyabb értéket becslünk b_2 előjelétől függően.

b) Ha magas a kondíciós index, akkor a regressziós modell helyett faktor (vagy főkomponens) elemzés alkalmazása indokolt. De legyünk tudatában annak, hogy ez is a változók szoros lineáris kapcsolatára épít. Nem lineáris kapcsolat esetén előzetes linearizáló transzformáció indokolt.

5. Logisztikus regresszió

A lineáris regresszió tárgyalása során éppen csak utaltunk a nemlineáris regresszióra. Mi ennek az oka? Az, hogy a nemlineáris jelleg számtalan függvényformát takar. További módszertani elágazást jelent az, amikor az y függő változó nem folytonos, hanem két vagy több kategóriával rendelkező változó. Ha ilyen elemzési feladat adódik, akkor használhatjuk a kereszttáblát, vagy a kereszttáblára illeszthető loglineáris modellt[75]. Ez – éppúgy, mint a lineáris regressziószámítás – is az általánosított lineáris modell család (GLM) speciális esete.

Ebben a fejezetben egy további GLM modellt, a logisztikus regressziós modellcsalád legegyszerűbb modelljét, a bináris logisztikus regressziót, az un. logit modellt tárgyaljuk. A módszer fontosságát, alkalmazhatóságát az utóbbi években megjelent számos cikk[76] is bizonyítja.

A logisztikus regresszió alkalmazási célját tekintve az osztályozó eljárások[77] közé sorolható,

mert akkor használhatjuk, ha előre definiált, egymást kölcsönösen kizáró csoportok egyikébe soroljuk be a megfigyeléseket a magyarázó változókból nyert információ alapján. Ha az eredményváltozónak több lehetséges kimenete van, akkor multinomiális logisztikus regresszióról beszélünk. A logit modell akkor

---

[75] Ezt részletesen tárgyalja: Füstös-Kovács-Meszéna-Simonné (2004): Alakfelismerés c. könyve.

[76] Hunyadi László: A logisztikus függvény és a logisztikus eloszlás, Statisztikai Szemle 2004.10-11.

Hajdu Ottó: A csődesemények logit-regressziójának kismintás problémái Statisztikai Szemle, 2004. 4. .

Fülöp Péter: A bináris logit modellek használatának és tesztelésének eszközei, Statisztikai Szemle 2002. 3.

Bartus Tamás: Logisztikus regressziós eredmények, Statisztikai Szemle 2003. 4.

Gray R.-Kovács E.: Az általánosított lineáris modell és biztosítási alkalmazásai, Statisztikai Szemle, 2001. 8.sz.

[77] A klasszifikációs módszerek közül foglalkozik ez a könyv a Klaszterelemzéssel (3. fejezet) és a Diszkriminancia elemzéssel (7. fejezet). Klaszterelemzést akkor végzünk, ha ismeretlen kategória határok mellett tárjuk fel a belső struktúrát. A diszkriminancia elemzés a logisztikus regresszióhoz hasonló feladatok megoldására – számos előfeltevés teljesülése esetén – alkalmazható. A logisztikus regresszióval végzett diszkriminálást akkor érdemes választani, ha a változók együttes eloszlása nem tekinthető normálisnak, és a variancia-kovariancia mátrixok nem egyenlők.

alkalmazható, ha az eredményváltozónak csak két, egymást kölcsönösen kizáró kategóriája van.

5.1. A logit modell és az induló adatok

Az eredményváltozó, Y (response, függő változó) 0-1 értékű bináris változó, amely többek között azt fejezheti ki, hogy

- a hitelt felvevő ügyfél csődbe jutott vagy törleszt,
- az ügyfél felmondta-e a szerződését, azaz lemorzsolódott vagy előfizető maradt,
- a páciens felgyógyult vagy nem élte túl a balesetet,
- egy játékterembe belépő személy kockáztatott vagy nem játszott stb.

A magyarázó változók között lehetnek nominális, ordinális vagy magasabb (intervallum és arány) skálán mért változ ók is. A nominális vagy ordinális szinten mért x változók lehetséges értékei (szintjei) közül egyet (általában az elsőt vagy az utolsót) rögzítjük, ezekhez viszonyítva becsüljük a függő változóra gyakorolt hatást. A magyarázó változók szintjeinek kombinációt is rögzíthetjük (Pl. 1500 cm3 alatti autót vezető férfi), ezek a kovariánsok.

Az y eredményváltozó kategóriáinak bekövetkezése (pl. y=1, a csőd előfordulása) az x magyarázó változókból (jövedelem, életkor, eladósodottság) nem becsülhető a hagyományos legkisebb négyzetek módszerével az $y = \beta_0 + \underline{\beta x}$ lineáris regressziós modellel az alábbi okok miatt:

A dichotom y nem normális eloszlású, hanem Bernoulli B(1,p) eloszlást követ. Az y=1 bekövetkezésének a valószínűsége *p*. Várható értéke: $E(y)=P(y=1)=p$ és varianciája: $Var(y)=p(1-p)$. Így a variancia a *p* valószínűségtől függ, nem konstans.

A magyarázó x változó egy egységnyi változása nem a teljes tartományon eredményez azonos változást y értékében.

A lineáris regresszióval becsült érték nem feltétlenül esik a [0;1] intervallumba, pedig az y=1 bekövetkezésének valószínűséget becsüljük.

Az említett problémák megoldása érdekében a Cox[78] (1970) által javasolt logit transzformációt alkalmazunk, hogy a becsült *p* érték a [0;1] tartományban maradjon, és ne növekedjen/csökkenjen a „széleken" túl gyorsan, úgy, mint ahogy ez a lineáris regresszióval történő becslésnél előfordul.

A logit transzformáció azt jelenti, hogy a függő változó helyett a hitel vissza nem fizetés valószínűségének (*p*) és a törlesztés valószínűségének (*1-p*) hányadosát logaritmáljuk, és erre illesztünk (5.1) szerint (itt egyváltozós) lineáris modellt:

---

[78] Cox D.R. 1966-ban írt először a logisztikus kvalitatív függő változók elemzéséről. 1970-ben pedig „Analysis of binary data" címen könyvet is publikált a témában.

$$\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \mathrm{logit}(\mathrm{p}) = \beta_0 + \beta_1 \mathrm{x} \quad , \tag{5.1}$$

ahol *p/(1-p)* az odds[79], és ennek logaritmusa, azaz az esély logaritmusa a logit.

5.2. A logit modell paramétereinek becslése

Az (5.1) egyenletben három ismeretlen van: p, β_0 és β_1 .

Hogyan becsüljük annak valószínűségét, hogy az ügyfél hitelképes, és a modell alapján inkább a hitelképesek csoportjába soroljuk-e? Általánosan megfogalmazva az Y kimenet előrejelzése, azaz az ügyfél klasszifikációja hogyan végezhető el?

Mivel az y eloszlása ismert, esetünkben Bernoulli eloszlású, a mintából a legvalószerűbb – Maximum Likelihood (ML) – becslést készítjük el.

Első lépésben tekintsünk el az x adatoktól, még csak az y=1 és az y=0 bekövetkezések gyakoriságát ismerjük. Likelihood függvényt írunk fel (5.2) szerint a B(1,p) eloszlású változóra:

$$L(p) = \prod_{i=1}^{n} (p)^{y_i} \cdot (1-p)^{(1-y_i)} \tag{5.2}$$

Ennek logaritmusát deriváljuk p szerint:

$$\ln L = \sum_{i=1}^{n} y_i \ln p + \sum_{i=1}^{n} (1-y_i) \ln(1-p)$$

$$\frac{d \ln L}{dp} = \frac{\sum y_i}{p} - \frac{\sum (1-y_i)}{1-p} = 0$$

Mivel az n számú megfigyelésből k esetben y=1 és (n-k) esetben y=0 következett be, az összegzésben $\Sigma y=k$ és $\Sigma(1-y)=n-k$ írható. Ekkor $k/p=(n-k)/(1-p)$, amit rendezve $k=np$ adódik, azaz a relatív gyakorisággal történő becslés formuláját kaptuk:

$$\hat{p} = \frac{k}{n} \tag{5.3}$$

Ha tehát x magyarázó változót nem vonunk be a modellbe, a kockázat (csőd) becsült valószínűsége például n=25 és k=15 esetén P(y=1)=15/25=0,6 lesz. A klasszifikációt úgy végezzük, hogy akire ennél nagyobb valószínűséget becslünk, azt a

[79] Az „odds” a szótár szerint „valószínűség”, de ez a fordítás nem helyes, mert a két valószínűség hányadosa egynél nagyobb is lehet. A továbbiakban az „odds” szót használjuk, vagy esélynek fordítjuk.

„kockázatosak" közé soroljuk, míg a 0,6 alatti értékűek a másik kategóriába[80] kerülnek.

Ezt az eredményt úgy is értelmezhetjük, hogy minden egyes x értékhez (pl. életkorhoz, jövedelmi kategóriához, eladósodottsági rátához) egyetlen közös $p_i = \pi$ valószínűség tartozik.

Ez a feltevés a gyakorlatban általában nem igaz. A p_i valószínűség változik, ha az x_i magyarázó változók értékeit figyelembe vesszük. Tipikus példaként említhető a halálozási (q_x) vagy az életben maradási (p_x) valószínűség. Mindkettő függ az életkortól, életmódtól, vagyoni helyzettől stb.

Ha a bekövetkezési valószínűség becsléséhez a magyarázó változókat is bevonjuk a logit modellbe, az ML becslés jóval komplikáltabbá válik.

Az esélyek logaritmusa, a log-odds lesz az x magyarázó változók lineáris függvénye:

$$\ln(odds) = \ln\left(\frac{p}{1-p}\right) = \log it(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{5.4}$$

vagy

$$odds = \left(\frac{p}{1-p}\right) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) = e^{\beta^T x} \tag{5.5}$$

Ebből kapjuk a becsült feltételes valószínűséget:

$$\hat{p} = \frac{p}{1-p+p} = \frac{p/(1-p)}{1+p/(1-p)} = \frac{e^{\beta^T x}}{1+e^{\beta^T x}} \tag{5.6}$$

A regressziós paraméterek becsléséhez az (5.7) szerinti likelihood függvényt írjuk fel, és az (5.6) szerinti becslést behelyettesítve kapjuk (5.8)-at:

$$L(b_0, b_1, \ldots, b_p) = \prod_{i=1}^{n} (p_{ib})^{y_i} \cdot (1-p_{ib})^{(1-y_i)} \tag{5.7}$$

---

[80] Ez a k/n érték lehet beállítva „*cut-value*", azaz döntési küszöbértéknek a futtatásban. A számítógép alapbeállításában ez ½.

$$L(\underline{b}) = \prod \left[\frac{\exp(\sum_j b_j x_{ij})}{1+\exp(\sum_j b_j x_{ij})} \right]^{y_i} \cdot \left[\frac{1}{1+\exp(\sum_j b_j x_{ij})} \right]^{1-y_i} \quad (5.8)$$

Ha csak egyetlen x változónk van, akkor két paramétert (b_0 és b_1) becslünk. Mivel a *b* becslésekre nincsen explicit formula, a számítógép számos b_0 és b_1 értékpárt behelyettesít, hogy megtalálja azt az értékpárt, amelyre az *L(b)* a maximumát felveszi. Ez az iteratív Newton-Raphson eljárás.

A becsült b paraméterek felhasználásával (5.9) egyenletből (5.10) szerint kapunk becslést *p*-re:

$$\log it(p_i) = \hat{b}_0 + \hat{b}_1 x_i \quad (5.9)$$

$$\hat{p} = \frac{1}{1+e^{-(b_0+b_1 x)}} \quad (5.10)$$

Ha x=0, akkor (5.11)-ből belátható, hogy a becsült érték 0 és 1 között van:

$$\hat{p} = \frac{1}{1+e^{-b_0}} \leq 1 \quad (5.11)$$

A logisztikus görbe nyújtott S-alakját a b_1 előjele határozza meg. Ha $b_1>0$, akkor emelkedő az S-görbe, és a b_1 a növekedés sebességét fejezi ki. Ez a hatás parciális és additív.

Értelmezni az $exp(b_1)$ kifejezést szoktuk, ami azt mutatja meg, hogy az x egy egységnyi növekedése hányszorosára változtatja meg az esélyt, az odds-t. Ez a hatás parciális és multiplikatív, amint ezt (5.12) mutatja.

$$odds = \left(\frac{p}{1-p} \right) = \exp(\beta_0 + \beta_1(x_1+1) + \ldots + \beta_p x_p) = e^{\beta x} \cdot e^{\beta_1} \quad (5.12)$$

Ha $b_1>0$, akkor $exp(b_1)>1$, az esély növekedik, míg $b_1<0$ esetében $exp(b_1)<1$, ami csökkenti az esélyt. Ha $b_1=0$, akkor az esélyhányados értéke 1, vagyis x változásával arányosan változik az odds.

A b_1 közvetlen értelme az esélyhányados logaritmusához kapcsolható:

$$\log \left[\frac{p(x+1) \,/\, (1-p(x+1))}{p(x) \,/\, (1-p(x))} \right] = \log \frac{p(x+1)}{1-p(x+1)} - \log \frac{p(x)}{1-p(x)} = \left(b_0 + b_1(x+1)\right) - (b_0 + b_1 x) = b_1$$

Ha $b_1>0$, akkor a hányados is nagyobb egynél, az x növekedésénél jobban nő az esély. Míg ha $b_1<0$, akkor az esélyhányados kisebb egynél, az x növekedéséhez csökkenő esély tartozik.

További érdekes kérdés, hogy milyen x érték mellett adódik ½ valószínűség, azaz mikor lesz teljesen bizonytalan a helyzet (és használhatatlan a modell)?

$$\hat{p} = \frac{1}{1+e^{-(b_0+b_1x)}} = 1/2, \text{ ha } e^{-(b_0+b_1x)} = 1$$

Az egyenlőség akkor teljesül, ha a kitevő nulla. Ez két esetben állhat fenn, ha x= $-b_0/b_1$, vagy ha $b_0=b_1=0$.

A statisztikai becslést általában követi a konfidencia intervallumok felírása, nullhipotézisek felállítása és tesztelése.

Az (1-α) megbízhatósági szinthez tartozó konfidencia intervallumokat több magyarázó változót tartalmazó logit modell együtthatóira írjuk fel. Az x_j egységnyi változásának hatása két alakban is vizsgálható:

a) a logit változására: $b_j \pm z_{\alpha/2} se(b_j)$,

b) az odds-ra pedig: $e^{b_j \pm z_{\alpha/2} se(b_j)}$, (5.13)

Mivel az értelmezésben is kitüntetett szerepe van *exp(b)*-nek, a konfidencia intervallumot[81] is (5.13) szerint érdemes vizsgálni. Ha az intervallum tartalmazza az egyet, akkor az x változó hatása nem szignifikáns.

A logit modellben az együtthatókra felírt nullhipotézist parciálisan teszteljük. A regressziós modellhez hasonlóan H_0: $\beta_j = 0$ hipotézist vizsgáljuk. Nagy mintára a $z=b_j/se(b_j)$ hányados sztenderd normális eloszlást követ. Itt egy- és kétoldali alternatív hipotézist is vizsgálhatunk.

Csak kétoldali alternatív hipotézist (H_{alt}: $\beta_j \neq 0$) tesztelhetünk a Wald-statisztikával *(W)*, ahol: $W=z^2$, és ez 1 szabadsági fokú khi-négyzet eloszlást követ.

Ha *z* és *W* „nagy” és mellette az empirikus szignifikancia szint $p < 0,05$, akkor x_j hatása szignifikáns, H_0-t elvetjük.

5.3. A logit modell illeszkedésének jósága

A modell jósága több tényező együttes elemzése alapján állapítható meg. Először parciálisan vizsgáljuk a modellt. A téves besorolásnál megkapjuk a reziduálisokat. A reziduális az eredeti y=1 esemény p valószínűsége és a becsült p_b eltérése: $e_x = p-p_b$.

[81] Az SPSS outputjában ezt külön kell kérni.

Az (5.14) szerinti sztenderdizált reziduálisok

$$e_z = \frac{p - p_b}{\sqrt{p_b(1-p_b)/n}} \tag{5.14}$$

nagy megfigyelésszám (n>30) mellett sztenderd normális eloszlást követnek, négyzetösszegük khi-négyzet eloszlású lesz.

A modell egészét több mérőszámmal is tudjuk minősíteni. A globális minősítéshez a klasszikus illeszkedésvizsgálatot a Pearson-féle khi-négyzet teszt-függvénnyel végezhetjük el.

Az illeszkedés vizsgálat további mérőszámai közül a Lagrange-multiplikátor (score) a Pearson–féle khi-négyzet elv alapján számolható, a megfigyelt (*f*) és a várt (*np*) gyakoriságok sztenderdizált eltérés-négyzetösszege:

$$\chi^2 = \sum_x \frac{(f_x - n_x p_{xb})^2}{n_x p_{xb}(1-p_{xb})} \tag{5.15}$$

Ha egy kovariáns változó kategóriáira nem teljesül az, hogy a becsült gyakoriságok nagysága legalább öt, akkor Hosmer-Lemeshow tesztet kell alkalmazni, hogy megállapítsuk, szignifikáns-e a megfigyelt és a várt gyakoriságok eltérése. A számítógép akkor is elvégzi ezt a homogenitásvizsgálatot, ha kellő számú megfigyelés esik egy-egy kategóriába, ezért röviden áttekintjük a Hosmer-Lemeshow teszt lépéseit.

A bináris (y) változóra és a becsült (*p*) valószínűségekre 2xg méretű kereszttáblát készítünk. Általában g=10 sort, azaz deciliseket határozunk meg.

A becsült valószínűségeket növekvő sorrendbe rendezzük és decilisekre bontjuk.

Összegüket decilisenként osztjuk a decilis elemszámával (s~n/10).

A második tag komplementerét vesszük minden decilisre: 1-Σp/s.

Megfigyelt (*M*) és várt (*V*) gyakoriságok eltérését négyzetre emeljük, és a nevezőben a második tag komplementere is szerepel:

$$\chi^2 = \sum (M-V)^2 / \left(V(1-\sum p/s\right)$$

A fenti összeg khi-négyzet eloszlást követ. A teszt kritikus értéke g-2 szabadsági fok mellett adódik. A számítógép az empirikus szignifikancia szint megadásával segíti a döntést. Ha ez kisebb, mint 0,05, akkor elvetjük a homogenitást.

A likelihood arány elven történő tesztelés azért fontos, mert hierarchikus modellekre is jól használható. Az (5.16) szerint ez azt fejezi ki, hogy egy x magyarázó változó bevonása javít-e az illeszkedésen ahhoz képest, ha csak a konstans szerepel a modellben:

$$LR = -2\ln\left(\frac{L_{b_0}}{L_{b_1}}\right) \tag{5.16}$$

A számlálóban szerepelhet az induló modell, a nevezőben pedig az újabb x változók bevonásával készült – bővített – modell.

Ebből számolható többféle R^2 mutatószám is, az egyiket McFadden javasolta:

$$R^2 = 1 - \frac{\ln L(\hat{b}) - (k+1)}{\ln L(0)}, \tag{5.17}$$

ahol (k+1) az összes becsült b paraméterek száma, L(0) pedig a null-modell. Ez a mérőszám a likelihood függvényben bekövetkezett változást méri, ezért közvetlen – a lineáris regressziós szórásnégyzet felbontáshoz hasonló – értelme nincsen.

5.4. A logit modell illesztése az SPSS-ben

Az általánosított lineáris modellek többsége, köztük a logit modell is több útvonalon érhető el az SPSS-ben. A logit modell becslésének most azt a változatát ismertetjük, amelyet a regressziós modellezésen belül található.

Regression /Binary Logistic választást követően először a függő és a magyarázó változókat jelöljük ki.

Dependents: y változó megadása (a 0-1 értékpár hozzárendelését az elemző dönti el, a becslés az y=1-re készül)

Covariates: x-ek listája, itt a változók közötti interakció is megadható

Method:

- Enter eljárás: a felsorolt x változók mindegyikét egyszerre lépteti be a logit modellbe,
- Forward (Conditional, LR és Wald változatok): lépésről lépesre szignifikáns változókkal bővíti a modellt
- Backward (Conditional, LR és Wald változatok): lépésről lépesre szűkíti a modellt, ha nem szignifikáns minden megadott x változó.

A lépésenként választó eljárásokon belüli további három lehetőség közül választhatunk:

- A Wald teszt értéke szerinti szignifikáns változó beléptetése (vagy a nem szignifikáns x kihagyása).
- A likelihood arány (LR) legnagyobb változását eredményező változó bevonása/kihagyása, ahol a maximum likelihood elven becsült paraméterekkel számolt *LR=-2*[*lnL(redukált)-lnL(teli)*] khi-négyzet eloszlást követ, és a szabadsági foka a két modell változószáma között mért

különbség.

- A feltételes (Conditional) statisztika alapján történő választás is LR alapon történik. De itt a redukált modellben az együtthatók közötti kovarianciákat is felhasználó feltételes becsléssel számolódnak az együtthatók.

Három további beállítási lehetőség kínálkozik még:

a) A „Categorical" gomb alatt a magyarázó változók, a kovariánsok szintjei közül választhatunk referencia kategóriát: az első vagy az utolsó kategóriához viszonyíthatjuk a többi kategóriának a bekövetkezési valószínűségre gyakorolt hatását.

b) A „Save" gomb a Regresszió elemzés (4) fejezetében tárgyalt opciókhoz nagyon hasonló mentéseket tesz lehetővé:

Elmenthetjük a becsült valószínűséget, és a javasolt csoportba sorolást (Predicted probability, Group membership)

Az egyes változóknak a modellre gyakorolt hatását (Influence) a Cook mérték, a Leverage values és a DfBeta(s) adja meg, mindhárom elmenthető.

A reziduálisok vizsgálatára pedig öt változatban kerülhet sor, mert a sztenderdizált és nem-sztenderdizált reziduálisok mellett menthető a Studentizált reziduális, a logit reziduális és a deviancia mértéke is.

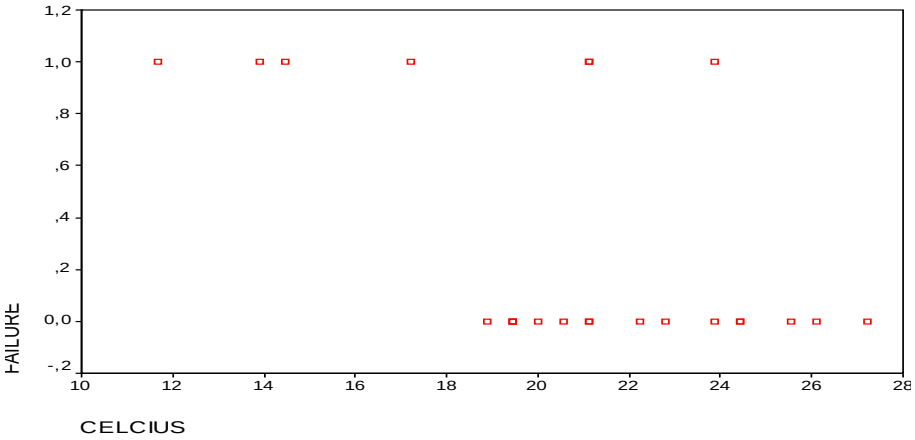
c) Az „Option" megnyomásával számos további részeredmény állítható elő. A klasszifikációt mutató ábra, az illeszkedés jóságának mutatói, azok a kilógó értékek, amelyek reziduálisai 2 szórásnyinál nagyobbak, a modellbeli változók közötti korrelációk kérhetők. Az iteráció beállított maximális lépésszáma 20, de ez változtatható. Az exp(β)-ra becsült konfidencia intervallum megbízhatósági szintje is eltérhet az alapértelmezésben választható 95%-tól. A b_0 konstans is választható vagy kihagyható a modellből. Továbbá itt található a döntés kritikus értéke, a klasszifikációs pont („cutoff")=0,5-re, mint alapértékre beállítva. Ezt akkor használjuk, ha a véletlenre bízzuk a besorolást, nincs előztes ismeretünk a csoportba tartozásról, vagy egyforma valószínűséggel eshetnek a megfigyelések az egyik vagy a másik kategóriába. Az értéket a relatív gyakoriságok ismeretében megváltoztathatjuk, és ezzel az osztályozást befolyásolni tudjuk. A logit modell alkalmazásakor visszatérünk a klasszifikációs pont értékének megadására.

5.5. LOGIT modell illesztése

Célunk a sikeres repülés[82] valószínűségének becslése a külső hőmérséklet ismeretében. 23 adatpár áll rendelkezésünkre: a külső hőmérséklet Celciusban és a sikeres visszatérés vagy a kudarc ténye. A 23 repülésből 7 végződött kudarccal,

[82] Az elemzés a repülési kudarc egyik okaként az alacsony hőmérsékletet tárta fel. De természetesen a vizsgálat célja lehet az is, hogy mekkora hőmérséklet mellett lehet kellően magas valószínűséggel számítani a sikeres visszatérésre.

ebből a becsült valószínűség: 7/23= 0,304. Az 5.1. ábra alapján ez azonban nem konstans valószínűség, mert a hőmérséklet emelkedésével csökkenni látszik a kudarc. A logit modell illesztésével a becslés során felhasználjuk a hőmérsékleti adatokat, és teszteljük a modell erejét.



5.1. ábra: A sikeres és kudarcos felszállások a hőmérséklet függvényében

A 0. lépésnek nevezi a program azt, amikor még csak – az 5.1. táblázatban látható – becsült konstans van a modellben, ekkor a valószínűséget (5.11) szerint kapjuk meg:

$$\hat{p} = \frac{1}{1+e^{0.827}} = 0{,}3043$$

, ami éppen megegyezik a *k/n=7/23* relatív gyakorisággal.

5.1. táblázat: A logit modellbeli konstans és a Wald teszt

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-,827	,453	3,328	1	,068	,438

A konstans szerepe a logit modellben a Wald teszt alapján 5%-os valószínűségi szinten nem szignifikáns.

Ebben a lépésben az esély, azaz a p/(1-p) hányados éppen *exp(-0.827)=0,438*, ami természetesen megegyezik 7/16-dal. A likelihood függvény (5.8) szerint a konstanssal is felírható: $L(0)=\left[\frac{0{,}438}{1+0{,}438}\right]^{7} \cdot \left[\frac{1}{1+0{,}438}\right]^{16} = 7{,}268*10^{-7}$.

Ennek logaritmusát (lnL= -14,134), majd (-2)-szeresét vesszük, mert ez követ khi-négyzet eloszlást.

Az 5.2. táblázatban háromlépéses iteráció után -2lnL=28,267 található. Ehhez az értékhez viszonyítjuk a logit modell illeszkedésének javulását a további lépésekben.

5.2. táblázat: 3 lépéses iteráció a konstans becslésére

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	28,277	-,783
	2	28,267	-,826
	3	28,267	-,827

a. Constant is included in the model.
b. Initial -2 Log Likelihood: 28,267
c. Estimation terminated at iteration num parameter estimates changed by less

Az 5.3. táblázat szerinti „Score" is khi-négyzet eloszlást követ és szignifikáns nagyságú, ez jelzi számunkra, hogy található még a logit modellbe be nem vont, de bevonható (szignifikáns hatású) változó, ezért folytatjuk az eljárást.

5.3. táblázat: A következő lépésben bevonható változó

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	CELCIUS	7,231	1	,007
	Overall Statistics		7,231	1	,007

Az 5.4. táblázatban a Newton-Raphson iteráció 4 lépése során becsült b_0 és b_1 együtthatók láthatók. Megállapíthatjuk azt is, hogy x bevonásával nőtt a likelihood függvény értéke, mert itt a -2loglikelihood= 20,315, és ez az induló 28,267-hez képest 7,952-vel kisebb.

5.4. táblázat: Az illeszkedés javulása

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients Constant	Coefficients CELCIUS
Step 1	1	21,185	4,834	-,269
	2	20,359	6,896	-,380
	3	20,315	7,559	-,415
	4	20,315	7,613	-,418

a. Method: Enter
b. Constant is included in the model.
c. Initial -2 Log Likelihood: 28,267
d. Estimation terminated at iteration number 4 bec log-likelihood decreased by less than ,010 perce

Az 5.5. táblázatban az iteráció negyedik lépésének loglikelihoodja (LL) mellett két további mutatót találunk. Ezek a regresszió számításból ismert determinációs együtthatóhoz hasonló tartalmúak. A szakkönyvekben szereplő McFadden-féle R^2 mutatót (5.17) az output nem tartalmazza. A szakirodalomban[83] számos szerző óv a pszeudo-mutatók direkt értelmezésétől, főleg több modell közötti választásra lehet ezeket használni.

5.5.táblázat: A modell „determinációs" együtthatói

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	20,315	,292	,413

Cox és Snell (5.18) mutatója közvetlenül a likelihoodból számolható, és eszerint mintegy 30%-ban határozza meg a hőmérséklet a repülési kudarc esélyét:

$$R_{Cox}{}^2 = 1 - \left(\frac{L(0)}{L(1)} \right)^{2/n} \tag{5.18}$$

Cox-Snell mutatóját a maximális értékkel leosztja Nagelkerke. Az így számolt (5.19) együttható mindig magasabb értéket ad. Itt 41,3%-os determináltságot jelez:

$$R_N^2 = R_{Cox}^2 / \left(1 - L(0)^{2/n}\right) \tag{5.19}$$

A becsült együtthatók outputja előtt kapjuk meg az osztályozás jóságát, vagyis azt, hogy a hőmérsékletet figyelembe véve a repülések 87%-át helyesen osztályozza a modell, amint ezt az 5.6. táblázat mutatja. Az összesített százalékot is befolyásolja, de különösen az egyes kategóriákhoz helyesen besorolt megfigyelések aránya érzékeny a küszöbszám (cut value) beállítására.

[83]Számos fórumon vitatják, hogy pszeudo-mutatók egyáltalában értelmezhetők-e, nem jobb-e a megfigyelt és a várt gyakoriságokat összevető Hosmer-Lemeshow teszt alkalmazása. http://stats.stackexchange.com/questions/3559/which-pseudo-r2-measure-is-the-one-to-report-for-logistic-regression-cox-s

5.6. táblázat: Klasszifikációs táblázat

Classification Table[a]

Observed			Predicted		
			FAILURE		Percentage Correct
			success	failure	
Step 1	FAILURE	success	16	0	100,0
		failure	3	4	57,1
	Overall Percentage				87,0

a. The cut value is ,500

A logit modell együtthatói és a tesztek az 5.7. táblázatban találhatók. Az additív hatást kifejező b_1 = -0,481 negatív, tehát a hőmérséklet növekedésével csökken a kudarc logitja. A multiplikatív hatást kifejező $\exp(b_1)$=0,658 pedig azt jelzi, hogy 1 Celsius foknyi hőmérséklet-emelkedés 0,658 szorosára változtatja a kudarc esélyét. 95%-os szignifikancia szinten 0,449 és 0,965 közötti ez a hatás, tehát biztosan csökken a kudarc esélye. A hőmérsékletet mérő változó tehát a modellben szignifikáns.

5.7. táblázat: A logit modell együtthatói

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	CELCIUS	-,418	,195	4,601	1	,032	,658	,449	,965
	Constant	7,613	3,933	3,747	1	,053	2025,098		

a. Variable(s) entered on step 1: CELCIUS.

A modell alapján a becsült valószínűség: P(y=1)= 1/(1+exp(-7,613+0,418x)

Ha x= 20, akkor p=0,3221-t kapunk. Ezek a becsült valószínűségek elmenthetők, és a reziduálisok is kiszámíthatók. Példánkban a 18. megfigyelés sztenderd reziduálisa kívül esik a (-2;+2) tartományon (5.8. táblázat), mert a magas hőmérséklet miatt alacsony valószínűséget (0,086) becsült a modell, de ez kudarcos repülés volt.

5.8. táblázat: Az outlier megfigyelések listája

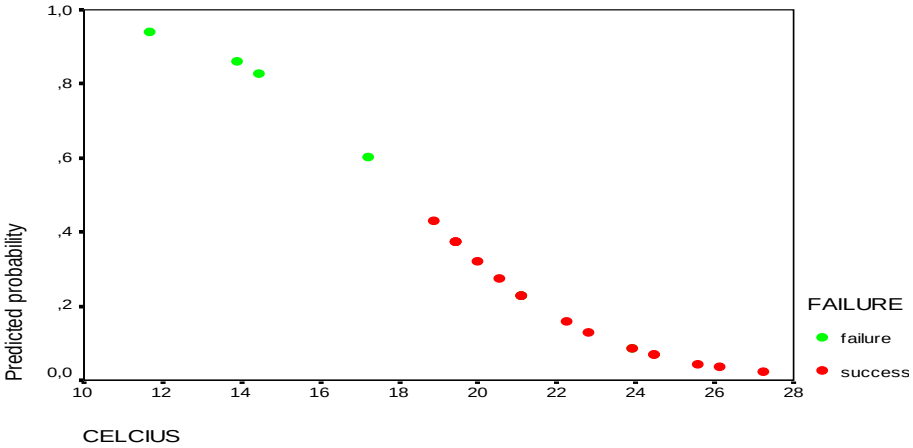
Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable	
		FAILURE			Resid	ZResid
18	S	f**	,086	s	,914	3,269

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Végül a becsült valószínűségeket pontdiagramon (5.2. ábra) ábrázolva mutatjuk be a logit modell egyik eredményét: 0,6 és 1 közötti valószínűséggel kudarcra számíthatunk, ha 18 Celsius fok alatti a hőmérséklet, míg melegebb időben a kudarc valószínűsége gyorsan – de nem lineárisan – csökken.



5.2. ábra: A hőmérséklet és a becsült valószínűségek

5.6. Mintamodell a lemorzsolódásra

A Telco.sav adatállomány lemorzsolódási (churn) adatait Logit modellel vizsgáljuk. Először a múlt havi adatokból (Frequency funkcióval) a lemorzsolódás gyakoriságát állapítjuk meg, amit az 5.9. táblázat mutat.

5.9. táblázat: Lemorzsolódott ügyfelek száma és gyakorisága

Churn within last month

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	726	72,6	72,6	72,6
	Yes	274	27,4	27,4	100,0
	Total	1000	100,0	100,0	

A bináris logisztikus regresszióban függő változó a „churn”, kovariánsok pedig az ügyfelek „személyi” adatai. Az alábbi beállítás (PASTE menűpont-sorozat) mellett illesztjük a LOGIT modellt:

```
LOGISTIC REGRESSION VARIABLES churn
/METHOD=FSTEP(WALD) tenure marital income gender longmon age address employ
/CONTRAST (marital)=Indicator
/CONTRAST (gender)=Indicator
/SAVE=PRED PGROUP COOK LEVER DEV
/PRINT=GOODFIT CI(95)
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.28).
```

- A módszer kiválasztásánál feltételezzük, hogy az ügyfelek adatai korrelálnak egymással, ezért a 8 változó között szelekciót kérünk, a beléptetés a Wald teszt alapján történik.
- Az ügyfél családi állapota és a neme kategória változók, ezeket beállítjuk, és az utolsó kategóriát, mint referenciát adjuk meg. Így az 5.10. táblázat szerint a „férfi" és a „nem házas" szerepelhetne – ha szignifikáns hatása lenne – a bevont változók között.

5.10. táblázat: Kategória változók kódolása és modellbeli szerepe

Categorical Variables Codings

		Frequency	Parameter coding
			(1)
Gender	Male	483	1,000
	Female	517	,000
Marital status	Unmarried	505	1,000
	Married	495	,000

- Az elmentési lehetőségek közül többet is kiválasztunk: (5.6) alapján a becsült valószínűség mellé a törlési kategóriát, egyedi megfigyelések hatását (leverage és Cook távolság), majd a devianciát is kérjük.
- Az együtthatók becslése mellé az (5.13) szerint a 95%-os konfidencia intervallumot is kérjük.
- A besorolási szintet (cut-value) 0,5 helyett 0,28-ra állítjuk.

Az eredménytáblák egy részét rövid értékeléssel együtt mutatjuk be.

a) A 8 változó közül három került bevonásra, és így a modell magyarázó ereje az (5.19) képlet szerint 23%-os, gyenge-közepes.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1020,513[a]	,143	,206
2	1004,542[a]	,156	,226
3	1000,758[a]	,159	,231

b) A klasszifikációs tábla a harmadik lépésben 66%-os arányban ismeri fel a maradó ügyfeleket, és 73,4%-ban a lemorzsolódókat. Összességében 68%-os az eredetivel megegyező, sikeres besorolás.

Classification Table[a]

	Observed		Predicted		
			Churn within last month		Percentage Correct
			No	Yes	
Step 1	Churn within last month	No	478	248	65,8
		Yes	74	200	73,0
	Overall Percentage				67,8
Step 2	Churn within last month	No	475	251	65,4
		Yes	72	202	73,7
	Overall Percentage				67,7
Step 3	Churn within last month	No	479	247	**66,0**
		Yes	73	201	73,4
	Overall Percentage				68,0

a. The cut value is ,280

c) A három lépésben bevont változókhoz a becsült együtthatók, azok sztenderd hibái és a Wald tesztek (szabadsági fokkal és szignifikancia szinttel együtt) követik egymást a „Variables int he Equation" táblázatban. Az Exp(B) oszlopra irányítsuk figyelmünket, hogy a hatások irányát és mértékét is értékelni tudjuk.

A magasabb jövedelem valamelyest emeli a törlés esélyét (1,002). míg a szerződés tartama (tenure) 0,962-szeresére, a munkahelyen ledolgozott idő hossza pedig 0,949-szeresére csökkentik a lemorzsolódást.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	tenure	-,046	,004	123,346	1	,000	,955	,947	,962
	Constant	,462	,136	11,574	1	,001	1,587		
Step 2[b]	tenure	-,039	,005	71,389	1	,000	,962	,953	,971
	employ	-,042	,011	14,720	1	,000	,959	,939	,980
	Constant	,599	,142	17,679	1	,000	1,820		
Step 3[c]	tenure	-,039	,005	70,898	1	,000	,962	,954	**,971**
	income	,002	,001	3,974	1	,046	1,002	1,00003	**1,00368**
	employ	-,053	,012	18,087	1	,000	,949	,926	**,972**
	Constant	,564	,143	15,538	1	,000	1,757		

a. Variable(s) entered on step 1: tenure.

b. Variable(s) entered on step 2: employ.

c. Variable(s) entered on step 3: income.

Önálló munkára javasolt feladatok:

Az életkor, a lakóhelyen töltött idő és a munkahelyen töltött idő főkomponensét előállítva és elmentve kapott PCA-Score szerepeltethető a LOGIT modellben az eredeti három változó helyett.

1/a) Vesse össze ennek a Logit modellnek az eredményeit a fentebb bemutatott részeredményekkel.

1/b) ROC görbe segítségével mutassa meg, hogy a besorolás pontossága mennyire tér el.

Megoldás:

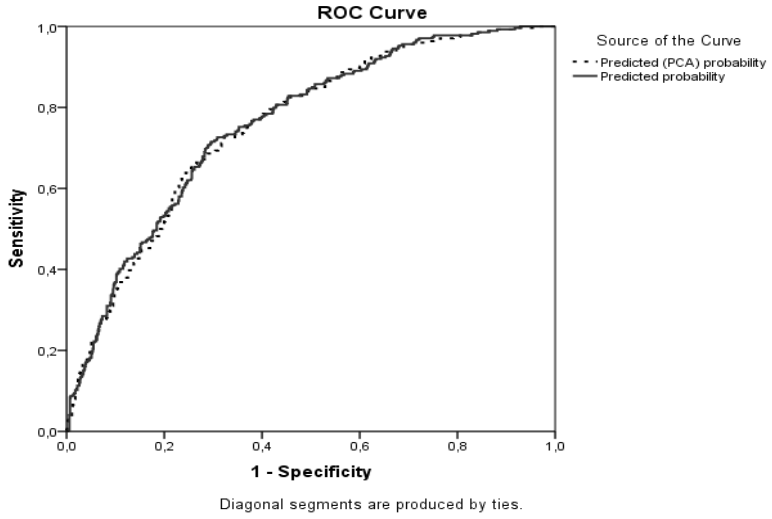
1/a) A főkomponens jól értelmezhető, 70 % feletti információsűrítést jelez. Magasabb score 0,532 és 0,814 közötti mértékben csökkenti a törlés esélyét. A jövedelem adat így nem került be a modellb, ami a következő oldalon látható.

1/b) A két modell AUC értéke 3 ezreléknyi eltérést mutat, a ROC görbék szinte egybeesnek.

Area Under the Curve

Test Result Variable(s)	Area
Predicted (PCA) probability	,755
Predicted probability	,758

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	tenure	-,046	,004	123,346	1	,000	,955	,947	,962
	Constant	,462	,136	11,574	1	,001	1,587		
Step 2[b]	tenure	-,037	,005	60,535	1	,000	,964	,955	,973
	FAC1_1	-,419	,109	14,829	1	,000	,658	,532	,814
	Constant	,087	,167	,274	1	,601	1,091		

a. Variable(s) entered on step 1: tenure.

b. Variable(s) entered on step 2: FAC1_1.

További feladat:

Más változók bevonásával keressen magasabb R-négyzetet elérő, és pontosabb besorolást adó modellt.

5.7. A modellválasztás grafikus eszköze

Mivel a számítógépes statisztikában is az angol nyelv dominál, számos olyan módszer és teszt van, ami eredeti angol nevén vált ismertté. Ilyen a kezdetben jelek felismerésére alkalmazott ROC-görbe[84] (Receiver Operating Curve) és AUC mérték (AUC=Area Under the Curve) is, amelyek alkalmasak arra, hogy több logit modell közül a legjobb felismerő képességűt ki tudjuk választani.

Kezdetben egy 2x2-es kereszttáblába rendezzük adatainkat. Így összevethető a kezdeti és a modell által adott besorolás. A jelölést nehezíti, hogy nem egyértelmű, mi számít jó vagy rossz megfigyelésnek. Ha a szerződés elmenűpontát, törlését vagy ügyfél lemorzsolódást elemezzük, akkor ez üzleti szempontból nem kedvező, de ennek becslésére irányul a modell. Ezért a táblázatban az „1” és a „0” kódokat is feltüntettük aláhúzandó, hogy az „1” jelűek helyes besorolása, azonosítása a logit modell célja. A döntések mellett zárójelben az előfordulások számát is megadjuk. Összesen a+b+c+d=n megfigyelést sorolunk be.

Tényleges/Döntés	Jó, befogadott (1)	Rossz, elutasított (0)
Jó, kedvező (1)	Helyes döntés (a)	Téves döntés (b)
Rossz (0)	Téves döntés (c)	Helyes döntés (d)

A ROC-görbe két tengelyén a fenti négy cellából két arányszámot készítünk és vetünk össze.

- Az y tengelyen d/(c+d) arány jelenik meg, ami a teszt érzékenységét méri. Itt az elutasított d számú rossz/csődös ügyfelek aránya az összes rossz/csődös arányában látható.
- Az x tengelyen b/(a+b) arány látható. Ez az elutasított b számú jókat az összes jóügyfél arányában méri. Ezt téves riasztásnak is nevezzük.

A döntési táblát a logit modell alapján kapjuk meg, ami a döntési érték (cut-value) beállításától függően más és más lesz. A ROC-görbe egy-egy pontja azt mutatja

[84] Történelmi érdekesség, hogy a jelfelismerés a II. világháború idején Pearl Harbor 1941-es megtámadását követően vált szükségessé. A radarok használatának célja az ellenséges repülők és a saját repülőgépek által adott jelek megkülönböztetése volt. A ROC-görbe szélesebb körű alkalmazása az 1970-es évek óta jellemző: kockázatcsökkentésre, orvosi tesztek értelmezésére is használni kezdték.

meg, hogy bizonyos döntési értékhez milyen x=b/(a+b) és y=d/(c+d) számpárok tartoznak.

Mivel a logit modellben nemcsak folytonos, hanem kategória-változók is szerepelhetnek, a ROC-görbe emelkedése sem folytonos, szakadások is lehetnek benne.

A 45 fokos egyenesen az elutasított rosszak aránya (y) épp megegyezik az elutasított jók arányával (x), ez a modell használhatatlanságát fejezi ki.

A ROC-görbe annál jobb modellt jelez, minél gyorsabban és minél magasabbra emelkedik a 45 fokos egyenes felé. A görbe alatti terület nagyságát a trapezoidokból számolt AUC-mérték adja meg. Ennek maximális értéke=1.

Több modell közötti választásra kiválóan alkalmas az AUC mérték. Hüvelykujj-szabály szerint az alábbi kategóriákkal jellemezhetők a logit modellek:

- 0,90-1 = kiváló
- 0,80-0,90 = jó
- 0,70-0,80 = közepes
- 0,60-0,70 = gyenge
- 0,50-0,60 = nem alkalmas a modell a megkülönböztetésre.

A görbét és a görbe alatti területet az 5.8. alfejezet példáján mutatjuk be.

A statisztikában használt első és másodfajú hiba tartalmilag kapcsolódik a ROC-görbéhez, de a ROC-görbe és az AUC mérték összetettebb információt adnak, bár valószínűségi szint nem tartozik hozzájuk. Emlékeztetőül az elsőfajú hiba alfa=c/n, rosszat befogadunk, míg a másodfajú hiba béta=b/n, jót elutasítunk (~ x tengely).

5.8. További logisztikus modellek

Ha a függő változónak kettőnél több kategóriája van, akkor két utat követhetünk:

a) Visszavezetjük a feladatot kétkategóriásra úgy, hogy

i) Egy kategóriát megtartunk, a többieket összevonjuk.

ii) A k számú kategória miatt (k-1) dummy változót vezetünk be, és k-1 logit modellt illesztünk

b) Multinomiális modellt illesztünk úgy, hogy az egyik kategóriát referencia kategóriának választjuk, és a többi (k-1) kategóriával minden egyes független változóra összehasonlítjuk. Egy-egy megfigyelést a legnagyobb valószínűségű kategóriába sorol az eljárás.

Alkalmazási előfeltevése a multinomiális logisztikus regressziónak sincs, se a független változók normális eloszlása, se a szórásnégyzetek egyezése nem szükséges.

Gyakorlati feltétel az, hogy a megfigyelések száma tízszerese legyen a változók számának, azaz n >10 p.

A Probit modellt is megemlítjük ebben a részben, bár ez továbbra is kétértékű függő változót becsül. A nevét a probability+unit szavak összekapcsolásából kapta, és az y=1 érték valószínűsége normális eloszlást feltételezve határozható meg.

$$P(Y = 1|X) = \Phi(X'\beta)$$

A modellben az x változók hatását kifejező β együtthatók maximum likelihood elven becsülhetők.

6. Faktorelemzés

A faktorelemzés három esetben kiemelten hasznos módszer. Ezek rövid bemutatása mellett példákkal is igyekszünk az olvasó figyelmét megragadni.

a) Látens változó előállítása

Komplex problémák elemzése a célunk, amikor a vizsgálni kívánt jelenség(ek) közvetlenül nem is mérhető(k). A megfigyelt, mérhető változókból állítjuk elő a látens (nem megfigyelhető) változókat, amelyeket faktoroknak nevezünk. Ilyen faktor lehet például a gazdasági vagy társadalmi fejlettség, a jólét, a települések vagy a piacok fejlettsége, egy „méret" vagy egy indexszám, ami több mutatószámból „keverhető ki". Ha egy faktor az eredmény, akkor rangsorolhatjuk is a megfigyeléseinket.

b) Dimenziócsökkentés

Az összes információ lehető legnagyobb hányadának megőrzése mellett keressük a minimális dimenziószámot, és azokat a faktorokat, amelyek már egymásra merőleges tengelyeket adnak meg. Így akár grafikusan is láthatóvá tehetjük a homogén adathalmazt alkotó megfigyeléseink szerkezetét ebben a redukált dimenziójú térben.

c) Független komponensek előállítása

Mivel a gazdasági és társadalmi változók többsége erősen korrelált, több – egymással kölcsönös kapcsolatban álló – változó egyidejű figyelembevétele nem lehetséges olyan módszerek alkalmazásakor, amikor a változók függetlensége alapfeltétel. A változók közül néhánynak a kiválasztása helyett képezzük az egymásra merőleges helyzetű faktorokat, amelyek független változókként használhatók például egy regressziós modellben.

A faktorelemzés több módszer összefoglaló[85] neve. Közülük a két legismertebbet tárgyaljuk részletesebben:

- Főkomponens-elemzés (Principal Component Analysis=PCA)
- Faktorelemzés (Principal Axis Factoring=PAF)

Egy-egy változó szórásnégyzetének felbontásakor három összetevőt különböztetünk meg: Teljes variancia = Közös variancia + Egyedi variancia + Hiba variancia

[85] A faktorelemzést összefoglalóan használjuk, ahogy a regressziószámítást is említjük, de mindig pontosítani kell, hogy milyen modellről van szó.

A két módszer döntően ebben a felbontásban különbözik, mert

- Főkomponenseket készítünk, ha a közös és egyedi varianciát együtt magyarázzuk, és csak a hibatagtól vonatkoztatunk el. Ekkor a p számú egymással korreláló változó közötti kapcsolatrendszert vizsgáljuk feltáró szemléletben, és egymással korrelálatlan változókká transzformáljuk az eredeti változókat, de a változók között ok-okozati kapcsolatot nem tételezünk fel. A változók lineáris kapcsolataira építve keressük az előre általában meg nem határozott számú ortogonális tengelyt.
- Faktorelemzést végzünk, ha csak a közös varianciát modellezzük. Ilyen alkalmazások során statisztikai modell húzódik meg a változók kapcsolatrendszere mögött, tehát megerősítő elemzést végzünk. A háttérben meghúzódó faktor hatásaként alakul a megfigyelt változók értéke úgy, ahogy az adatállományban látható.

A módszercsalád további eljárásairól is részletesen ír Füstös-Kovács-Meszéna-Simonné (2004): Alakfelismerés című könyve.

Bevezető példaként a főváros kerületeit és a környező településeket[86] kívánjuk összehasonlítani az életminőség szempontjából Az 50 megfigyeléshez rendelkezésünkre áll számos változó, amelyek egymással korrelálnak. Az adatokban „mérethatás" van: ahol több a népesség, ott több a lakás, de ahová többen vándorolnak, ott több az újonnan épített lakás is. Ezek a hatások kölcsönösek, tehát az ok-okozati irány nem mindig nyilvánvaló. Az elemzés célja most nem egy kiemelt változó megmagyarázása a többivel, mint a regressziós modellben, hanem azt keressük, hogy hány dimenzióban lehet leírni az életminőséget, mint látens változót. [87]

6.1. A főkomponenselemzés

Az eljárás alapgondolata az, hogy az egymással páronként lineárisan korreláló változók együtteséből ortogonális transzformáció révén előállítjuk a korrelálatlan főkomponenseket úgy, hogy az első néhány komponens leírja a változók összes szórásnégyzetének elég nagy hányadát, és így alacsonyabb dimenzióba képezhetjük le megfigyeléseinket. Ha az induló változók közötti korrelációk gyengék, akkor az eredeti változókkal többé-kevésbé megegyező számú és tartalmú komponenseket kapunk.

[86] A Kerületek.sav adatbázist használjuk ebben a fejezetben.
[87] A példa eredményeit a 6.1.3. alfejezetben követheti az olvasó.

6.1.1. A főkomponens elemzés matematikai háttere

Induló adatainkat az *X* mátrixba rendezzük, ahol a sorokban n megfigyelés, az oszlopokban p változó található. Hüvelykujj-szabályként javasolható, hogy n≥ 5p teljesüljön.

A főkomponensek négy tulajdonsággal írhatók le. Egyszerűbb a felírás, ha feltesszük, hogy a p db változó centírozott, az eredeti adatok helyett az átlagtól való eltérést használjuk.

1) Az y főkomponensek a mért x változók lineáris kombinációi, így az n-elemű főkomponensek felírhatók:

$\underline{y}_1 = \underline{X}\,\underline{a}_1$, $\underline{y}_2 = \underline{X}\,\underline{a}_2$,..., $\underline{y}_p = \underline{X}\,\underline{a}_p$, vagy mátrix alakban:

$\underline{Y} = \underline{X} \cdot \underline{A}$, ahol az $\underline{A}$ (pxp)-s.

2) A lineáris kombináció együtthatóinak négyzetösszege minden főkomponensre egy legyen, az elsőre így írható fel: $\underline{a}_1^{\,T} \cdot \underline{a}_1 = 1$

3) A főkomponensek varianciája monoton csökken: Var $(y_1)\geq \mathrm{Var}(y_2)\geq ...\geq \mathrm{Var}(y_p)\geq 0$ és a variancia:

$$Var(\underline{y}_1) = \frac{1}{n}\underline{y}_1^T\underline{y}_1 = \frac{1}{n}(\underline{X}\,\underline{a}_1)^T(\underline{X}\,\underline{a}_1) = \underline{a}_1^T\frac{\underline{X}^T\underline{X}}{n}\underline{a}_1 = \underline{a}_1^T\underline{S}\,\underline{a}_1 \rightarrow \max,$$

ahol

$\underline{S}$: a megfigyelt változók pxp-s méretű kovariancia mátrixa. Ha feltesszük azt is, hogy a változók standardizáltak, akkor $\underline{S}$ helyett $\underline{R}$ korrelációs mátrix szerepel.

4) A főkomponensek páronként korrelálatlanok: $r(y_1, y_2)=0$

A továbbiakban az $\underline{R}$ korrelációs mátrixból indulunk ki.

A 2) és a 3) tulajdonság együtt feltételes szélsőérték feladatot ad, ennek megoldását a Lagrange multiplikátorok módszerével végezzük.

$$L = \underline{a}_1^T\underline{R}\underline{a}_1 - \lambda_1(\underline{a}_1^T\underline{a}_1 - 1) \rightarrow \max \qquad (6.1)$$

A parciális deriváltat egyenlővé tesszük nullával:

$$\frac{\partial L}{\partial \underline{a}_1^T} = 2\underline{R}\underline{a}_1 - 2\lambda_1\underline{a}_1 = \underline{0}$$

Egyszerűsítve és rendezve λ_1 sajátértékű és $\underline{a}_1$ sajátvektorú egyenletrendszerhez jutunk:

$$\underline{R}\underline{a}_1 = \lambda_1\underline{a}_1 \qquad (6.2)$$

és $(\underline{R} - \lambda_1\underline{E})\underline{a}_1 = \underline{0}\ \lambda$

A homogén egyenletrendszernek csak a nem-triviális ($\underline{a} \neq \underline{0}$) megoldását keressük. Ekkor a mátrix determinánsa zérus:

$$\left|\underline{R} - \lambda_{\underline{1}} \underline{E}\right| = 0 \tag{6.3}$$

A pxp méretű mátrix determinánsának kifejtésével megkapjuk a p-ed fokú polinom gyökeit, a sajátértékeket, amelyek monoton csökkenő sorrendbe rakhatók. Mivel $\underline{R}$ mátrix szimmetrikus és pozitív definit mátrix[88], a legkisebb sajátérték is nemnegatív:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

A sajátértékek szorzata a mátrix determinánsát adja. Minél közelebb vannak a legkisebb sajátértékek a nullához, annál közelebb van a determináns értéke is a nullához.

A sajátértékek összege a mátrix nyoma, ezért a korrelációs mátrix felbontásakor

$$\sum_{i=1}^{p} \lambda_i = p \tag{6.4}$$

A kovariancia mátrixra $\sum_{i=1}^{p} \lambda_i = \sigma_1^2 + \ldots + \sigma_p^2$ teljesül. Ha a változók különböző mértékegységűek voltak, akkor nincs értelme a varianciákat összeadni. Ilyenkor fontos, hogy az adatokat előzetesen sztenderdizáljuk, vagy a korrelációs mátrix felbontását végezzük el. Ha korrelációs mátrix dekompozícióját végezzük, akkor a sajátértékek és a sajátvektorok eltérnek a kovariancia mátrix felbontásával kapott eredményektől. A két változat eredményei egymásból közvetlenül nem állíthatók elő. Ha mégis kovariancia mátrixból dolgozunk, akkor az alábbiakat tartsuk szem előtt:

Jól értelmezhetők a komponensek, ha	Miért fontos ez?
Minden változó azonos mértékegységű.	A skála változásával változik a főkomponens.
A változók varianciája közel azonos.	A nagy szórású változó dominálja a főkomponenst.

Mivel $\underline{R}$ (és $\underline{S}$) szimmetrikus, pozitív definit mátrixok, a sajátértékeik nemnegatívok.

A különböző sajátértékekhez tartozó $\underline{a}_1, \ldots, \underline{a}_p$ sajátvektorok pedig ortogonálisak, és a 2) feltétel miatt egységnyi hosszúak[89].

---

[88] Az $\underline{S}$ kovariancia mátrix is szimmetrikus és pozitív definit, ennek $\underline{S}$ sajátértékei is nemnegatívok.

[89] A normáltság miatt csak egy elemzésen belül hasonlíthatóak össze a sajátvektorok elemei.

Ha balról szorozzuk az $\underline{a}$ vektorral a (6.2) egyenletrendszert, akkor látható, hogy a 3) tulajdonság alapján a főkomponens szórásnégyzete a sajátérték:

$$\underline{R}\underline{a}_1 = \lambda \underline{a}_1 \quad /*\underline{a}_1^T$$

$$\underline{a}_1^T \underline{R}\underline{a}_1 = \lambda_1 \left(\underline{a}_1^T \underline{a}_1\right) = \lambda_1 \tag{6.5}$$

Egy főkomponens relatív fontosságát a $\lambda_j / \sum_{k=1}^{p} \lambda_k$ hányados mutatja, százzal szorozva százalékos formában adható meg a főkomponens által hordozott össz-információ.

A j-edik sajátértékhez a homogén egyenletrendszer megoldása[90] adja a j-edik sajátvektort, és ezzel előállítható a j-edik főkomponens. A főkomponensek korrelálatlanságát a sajátvektorok ortogonalitása biztosítja.

A sajátvektorok $\underline{A}$ mátrixával felírható az összes megfigyelés származtatott koordinátája*:* $\underline{Y}=\underline{XA}$

A főkomponens értéke (score) az i-edik megfigyelés „elhelyezkedését" mutatja a j-edik főkomponens tengelyen:

$$y_{ij} = \underline{a}_j^T \underline{x}_i \tag{6.6}$$

Összehasonlítható sajátvektorokat ($\underline{c}$ =component loading-ot, súlyt) kapunk, ha az $\underline{R}$ (vagy $\underline{S}$) mátrix nem egységnyi hosszú sajátvektorait előállítjuk:

$$\underline{c}_j = \sqrt{\lambda_j}\,\underline{a}_j \text{, amelyre } \left|\underline{c}_j\right| = \sqrt{\underline{c}_j^T \underline{c}_j} = \sqrt{\lambda_j \underline{a}_j^T \underline{a}_j} = \sqrt{\lambda_j}$$

vagy másképpen $\sum_{i=1}^{n} c_{ij}^2 = \lambda_j$ (6.7)

A c_{ij} jelentése: az i-edik változó és a j-edik komponens[91] közötti korreláció, amelyben a (6.2) mátrix-alakját használjuk fel:

---

[90] A sajátvektorok előjele tetszőleges, mert a homogén lineáris egyenletrendszer megoldásakor van szabad ismeretlen.

[91] A korreláció számításakor osztunk az Y komponensek szórásával, azaz a sajátértékek gyökével.

A $\underline{\Lambda}$ diagonális mátrix, főátlójában a sajátértékek szerepelnek. Az X-beli változók sztenderdizáltak, szórásuk egységnyi.

$$corr(\underline{X},\underline{Y}) = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \mathrm{cov}(\underline{X},\underline{Y})\underline{\Lambda}^{-1/2} = \frac{1}{n}\underline{X}^T(\underline{X}\,\underline{A}\underline{\Lambda}^{-1/2}) = \underline{R}\underline{A}\underline{\Lambda}^{-1/2} = \underline{A}\underline{\Lambda}^{1/2} = \underline{C}$$

A $\underline{C}$ mátrix minden eleme korrelációs együttható, de a $\underline{C}$ nem korrelációs mátrix, mivel a főátlójában az egyesek helyett az azonos indexű változó és komponens közti korrelációs együttható szerepel, és a mátrix nem szimmetrikus. (6.7) szerint az oszlopelemek négyzetösszege a sajátértéket adja. Egy-egy sor elemeinek négyzetösszege a változónak a főkomponensek által megmagyarázott varianciája, azaz a kommunalitás:

$$\sum_{j=1}^{p} c_{ij}^2 = h_i^2 = 1 \tag{6.8}$$

Fontos kapcsolat van $\underline{R}$ és $\underline{C}$ között:

$$\underline{R} = \underline{C}\underline{C}^T = \underline{A}\Lambda\underline{A}^T, \tag{6.9}$$

azaz a változók páronkénti korrelációit tökéletesen reprodukálják a változók és a főkomponensek korrelációinak szorzatai, valamint a sajátvektorok és sajátértékek mátrixai. A (6.9)-et úgy is megkapjuk, ha (6.2)-t mátrix alakban felírjuk, és jobbról szorozzuk:

$$\underline{R}\underline{A} = \underline{A}\underline{\Lambda} \qquad /*\underline{A}^T$$

Mivel az ortogonális mátrix transzponáltja megegyezik az inverzével, a szorzás után

$$\underline{R} = \underline{A}\underline{\Lambda}\underline{A}^T = \sum_{i=1}^{p} \lambda_i \underline{a}_i \underline{a}_i{}^T \tag{6.10}$$

teljes reprodukciót kapunk, ha az összes változó mentén p-ig összegzünk.

A kétféle input mátrixot és a sajátvektorok hosszát tekintve a $\underline{C}$ mátrix elemei négyfélék:

Input mátrix / Sajátvektor hossza:	$\underline{a}^T\underline{a}=1$	$\underline{a}^T\underline{a}=\lambda$
R korrelációs mátrix	$c_{ij}=a_{ij}\sqrt{\lambda_j}$	$c_{ij}=a_{ij}$
S kovariancia mátrix	$c_{ij}=a_{ij}\sqrt{\lambda_j}/\sigma_i$	$c_{ij}=a_{ij}/\sigma_i$

A korrelálatlan komponenseket tehát az eljárás végén megkapjuk, de hogyan valósulhat meg másik célunk, a dimenziócsökkentés?

Ha a legkisebb sajátérték(ek) nagysága zérus, akkor a hozzá(juk) tartozó sajátvektort, és így a főkomponenst sem állítjuk elő. Általában azonban csak

közelítik a λ-k a nullát, és ilyenkor felvetődik a kérdés, hogy hány főkomponens kell?

Mivel a varianciák monoton csökkenőek, az első k darab komponens nagyobb hányadot képvisel az összvarianciából, mint bármely másik k darab komponens. Ezért az utolsó (p-k) komponens figyelmen kívül hagyásáról dönthetünk úgy, hogy

- megadjuk előre a k számot,
- az egynél nagyobb sajátértékűeket vesszük,
- meghatározzuk azt a százalékot, amennyi információt meg akarunk őrizni.

Döntésünknek természetesen következményei lesznek. A változók és főkomponensek korrelációit tartalmazó $\underline{C}$ mátrix mérete nem p*p, hanem p*k lesz, a (6.8) szerinti kommunalitások kisebbek lesznek, mint egy, illetve a (6.9) és a (6.10) szerinti tökéletes reprodukálás sem valósul meg.

Ha az egynél kisebb sajátértéket elhagyjuk, az $\underline{A}$ mátrixnak is p-nél kevesebb oszlopa van. Az összegzés i=1-től k-ig (k≤p) megy, ami nem reprodukálja teljesen a korrelációs mátrixot. A redukált korrelációs mátrix:

$$\hat{R} = \sum_{i=1}^{k} \lambda_i \underline{a}_i \underline{a}_i^T \tag{6.11}$$

6.1.2. A megvalósítás lépései az SPSS-ben

Az **Analyze/Dimension Reduction/Factor** lépésekkel lehet a módszerek közül választani és főkomponens-elemzést végezni.[92]

A **változók kiválasztásával** kezdjük úgy, hogy törekedjünk az n>5p szabály betartására.

A **Selection**> menűpontsal egy kategóriaváltozó kijelölésével almintát adhatunk meg. Ez akkor hasznos, ha azt feltételezzük, hogy az almintákban más faktorstruktúra jellemző. Az SPSS ilyenkor az alminta adatait használva készíti el a becslést a teljes mintára.

A) Descriptives, azaz leíró statisztikák

E funkció alatt számos fontos előkészítő eredmény szerepel. A 6.1. táblázatban összefoglaljuk, hogy mit és miért kérünk, majd az egyes eredmények előállításához szükséges képleteket (zárójelben a sorszámuk) ismertetjük.

[92] A beállításokat az output táblák sorrendjében ismertetjük.

6.1. táblázat: PCA leíró statisztikák

Választható részeredmények	**Értelmezésük**
Egyváltozós leíró statisztikák	A változók eredeti átlaga és szórása. A magas relatív szórásra figyelni kell, hiszen homogén adathalmazból dolgozunk.
Korrelációs mátrix, szignifikancia szintek és a mátrix determinánsa	Változók közötti *lineáris* kapcsolatok szignifikánsak-e? Egyhez közeli determináns gyenge páronkénti korrelációkat jelez. $\|\underline{R}\|\approx 0$ esetén szorosak a korrelációk.
Korrelációs mátrix inverze	Parciális[93] és többszörös[94] korreláció mérése
Kaiser-Meyer-Olkin mérték (12)	Ha kisebb, mint 1/2, a minta nem alkalmas főkomponens-elemzésre. 0,5-0,7 között gyenge, 0,7-0,8 között közepes, 0,8 felett jó a PCA
Anti-Image korrelációs mátrix főátlója (13)	MSA[95] mértékek változónként, az 1-hez közeli érték a kedvező
Anti-Image korr. mátrix többi eleme	A parciális korrelációk (-1)-szeresei
Bartlett-teszt (gömbölyűségi) khi^2 – próba (14)	H_0 :$\underline{R}=\underline{E}$, a változók függetlensége elvethető-e (a többdimenziós normalitást feltételezi)

A Kaiser-Meyer-Olkin mérték számításakor az egész **minta megfelelőségét** (MSA: Measure of Sampling Adequacy) vizsgáljuk. A számlálóban a közönséges korrelációk négyzeteinek összege szerepel, kivéve a főátlóbeli egyeseket. A nevezőben pedig ehhez még hozzáadódnak a parciális korrelációk négyzetei. (A számlálóban p(p-1)/2 tag, a nevezőben p(p-1) tag szerepel.)

$$\text{KMO}= \sum_{i\neq j}\sum r_{ij}^{\ 2} \Big/ \left\langle \sum_{i\neq j}\sum r_{ij}^{\ 2} + \sum_{i\neq j}\sum p_{ij}^{\ 2} \right\rangle \qquad (6.12)$$

A KMO mérték 0 és 1 között lehet. Ha a KMO=1, akkor a parciális korrelációk nullák.

---

[93] A parciális korreláció az inverz mátrix főátlóbeli elemeiből is meghatározható. Ha az első két változó kapcsolatából p-2 változó hatását kiszűrjük: $r_{12\bullet 34\ldots p} = -q_{12} \Big/ \sqrt{q_{11}q_{22}}$, ahol q az inverz mátrix megfelelő eleme. Ha nem zavaró, akkor a részletes kiírás helyett p_{ij} szerepel.

[94] Egy többszörös korreláció értéke az inverz mátrix azonos indexű eleméből meghatározható: $R_{1\bullet 23\ldots p} = \sqrt{1-1/q_{11}}$, és a mutató mindig pozitív.

[95] MSA: Measure of Sampling Adequacy.

Az Anti-Image korrelációs mátrix (AIC) főátlójában a változónként kiszámolt KMO értékek állnak. A mutató az i-edik változóra:

$$MSA_i = \sum_{i \neq j} r_{ij}^2 \Big/ \left\langle \sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} p_{ij}^2 \right\rangle \qquad (6.3)$$

A mutató nagy értéke fontos változót és közös faktor létét jelzi. Ha kicsi (0,5 alatti) valamely MSA, akkor a változó kihagyásával javítható a modell.

Az AIC főátlón kívüli elemei a parciális korrelációk (-1)-szeresei. Jó a faktormodell, erősek a közös faktorok, ha a parciális korrelációk nullához közeliek. Ez azt jelenti, hogy az egyedi faktorok közötti korreláció is közel nulla.

Hüvelykujj szabály szerint minősíthetjük az eredményt, ahogy a 6.2. táblázat jelzi.

6.2. táblázat: A minta megfelelőségének értékelése KMO és MSA mértékek alapján

KMO és MSA értéke	**Minősítés (és teendő)**
0,9 felett	Kiváló, mert kicsik a parciális korrelációk
0,8-0,9	jó
0,7-0,8	közepes
0,5 felett	megfelelő
=0,5	Ha a korrelációs mátrix elemeinek négyzetösszege egyenlő a parciális korrelációk négyzetösszegével. Az alkalmazás kérdéses.
0,5 alatt	Elfogadhatatlan a módszer alkalmazása, mert • nem elég szorosak a lineáris korrelációk • túl magasak a parciális korrelációk (MSA 0,5 alatt: az adott változót ki kell hagyni.)

A KMO=0,5 adódhat úgy, hogy megkérdőjelezhető az alkalmazás:

- Ha összesen két változóra próbálunk főkomponens illeszteni. Ekkor a parciális korrelációban nincs kiszűrhető változó.
- Gépi beállítás miatt (hogy elkerüljük a nullával való osztást) is kaphatunk ilyen értéket, ha a korrelációs mátrix egységmátrix.

A **Bartlett-teszt** alapfeltevése az, hogy többváltozós normális eloszlású sokaságból[96] vettük a mintát, és az eredeti változók függetlenek, azaz az $\underline{R}=\underline{E}$. Ezt likelihood-arány teszttel vizsgáljuk, ahol $|\underline{R}|=\Pi\lambda_i$, és H_0: $\underline{R}=\underline{E}$.

$$\chi^2 = -a\log|\underline{R}| \text{ , ahol a = n-1-(2p+5)/6 és a szabadsági fok= p(p-1)/2} \qquad (6.14)$$

Főkomponens-elemzés csak akkor végezhető, ha elvetjük a nullhipotézist, azaz nem tekinthetők függetlennek a változók.

Itt kapjuk meg a kezdeti megoldást. Az eredeti változók egységnyi szórásnégyzete mellett a főkomponens-elemzéssel kapott (6.8) szerinti *h* kommunalitások állnak. Az i-edik változó varianciájának a közös faktorok együtt ekkora hányadát magyarázzák. Felső határát csak akkor éri el, ha mind a p db komponenst előállítjuk:

$$h_i^2 = \sum_j c_{ij}^{\ 2} \le 1$$

Az outputok között kapjuk meg a (6.9) szerint számolt **reprodukált korrelációs mátrixot**. Ennek főátlójában a kommunalitások (a közös faktorok által magyarázott variancia) találhatók.

B) Az „Extraction" blokkban választunk faktorelemző eljárást.

A főkomponens elemzés (PCA) az alapmódszer, és az egynél nagyobb sajátértékekhez (Kaiser kritérium) tartozó sajátvektorokat állítja elő, ha nem kérünk „k" számú faktort. Itt kérhető a Scree plot[97] ábra is. Ez megmutatja, hogy a sajátértékek nagysága hogyan csökken. A hirtelen csökkenés után megállunk, a további komponensek elhanyagolható mértékben javítják a modell illeszkedését. A kis sajátérték a véletlen hibát méri, nem egy látens közös komponens varianciája. Ha a változók gyengén korrelálnak, akkor nem csökken meredeken a Scree plot, nem csökken a dimenzió.

C) A „Rotation" blokkban rotált megoldást[98] állíthatunk elő, ha egynél több faktorunk[99] van.

A faktorok elforgatása történhet úgy, hogy a forgatás után is merőlegesek maradnak, és úgy is, hogy a faktorok korreláltak lesznek. Az ortogonális forgatás biztosítja azt, hogy a faktorok által nyújtott információ nem redundáns, de a vizsgált jelenségek faktorai lehetnek egymással összefüggőek is.

---

[96] Mivel többdimenziós normalitási teszt nincs, legalább nagy minta álljon rendelkezésünkre!

[97] A Scree plot vízszintes tengelyén a faktorok száma, függőleges tengelyén pedig a sajátértékek láthatók.

[98] A rotáció jelentőségét mutatja be Hajdu Ottó cikke a Statisztikai Szemle 2004. X-XI. dupla számában.

[99] A rotálás a PAF eljárás közös faktorainak értelmezésekor nagyon fontos.

Az ortogonális forgatás egyik változata a Kaiser által javasolt **Varimax** eljárás. A kommunalitások és a magyarázott összvariancia nem változik, de a sajátértékek igen. A „nagy" loadingok négyzetei egyhez, a kicsik nullához közeliek lesznek a forgatás után. Ha $\underline{B}=\underline{AT}$, ahol $\underline{T}$ a transzformáció ortogonális mátrixa, a Varimax kritérium felírható:

$$V=\sum_{q=1}^{k}\left\langle \frac{\left[\sum_{j=1}^{p} b_{jq}^{4}-\left(\sum_{j=1}^{p} b_{jq}^{2}\right)^{2}/p\right]}{p}\right\rangle \rightarrow \max$$

, és k a faktorok száma, k≤p. (6.15)

A ferdeszögű (Oblique) forgatást a Direct **Oblimin** eljárás végzi. Ekkor a főkomponensek közötti korrelációk mátrixa nem lesz egységmátrix, és nem adható meg az, hogy egyes változók szórásnégyzetének mekkora hányadát képviseli egy-egy faktor.

Ebben a részben kérhető a „**Loading plot"**, amely a változókat ábrázolja a faktorok terében.

D) További eredményeket kapunk a **Factor Scores** blokkban.

A score együttható-vektor p elemű, a sajátérték gyökéből és a hozzátartozó sajátvektorból számolható, minden változóhoz kiíratható:

$$\underline{a}/\sqrt{\lambda} \tag{6.16}$$

Az adatállományban jelenik meg a faktor score együttható mátrixa, amely mentése során három eljárás[100] közül választhatunk. Ha regressziós becsléssel készül, értelmezése is a standardizált regressziós együtthatókéhoz hasonló. Ezek adják a redukált dimenziójú térben az eredeti megfigyelések sztenderdizált koordinátáit, azaz minden oszlop átlaga 0 és szórása egységnyi. A regressziós becslés:$\underline{R}^{-1}\underline{C}$, akkor készíthető el, ha létezik a korrelációs mátrix inverze. A (6.9) és (6.10) egyenletek alapján belátható, hogy $\underline{R}^{-1}\underline{C}=\underline{A\Lambda}^{-1/2}$

A faktor score mátrix (n*k) méretű, és elemei: $\underline{Y_z}=\underline{X A\Lambda}^{-1/2}$, azaz $\underline{Y}$ főkomponensek sztenderdizált értékeit tartalmazzák.

[100]Bartlett eljárást és Anderson-Rubin becslést is választhatunk, amelyek a sajátértékek és a sajátvektorok felhasználásával adják meg az eredményt.

E) Az Options-ban a hiányzó adatok kezelését**,** adott szint alatti kis korrelációk kihagyását, és a többiek nagyság szerinti rendezését választhatjuk.

6.1.3. A PCA eredmények bemutatása és értelmezése

Budapest 23 kerülete és a fővárost körülvevő 27 település 2010-es adataira végzünk főkomponens elemzést. (Kerületek2010.sav)

Az első szakaszban csak négy változót használunk. Azt vizsgáljuk, hogy a lakónépességre vetített oda- és elvándorlást mérő négy változó milyen hatékonysággal sűríthető-e egyetlen v**ándorlás komponensbe**?

Kérdés: Javul vagy romlik a modell illeszkedése, ha nem létszámra vetített mutatókat használunk, hanem a vándorlást leíró eredeti abszolút számokat?

Válasz: A mérethatás miatt erősebbek a korrelációk, így az eredeti változók jobban sűríthetők egy főkomponensbe. De ne áldozzuk fel a korrekt alkalmazást ennek érdekében.

Mivel a kerületek és az agglomeráció települései eltérő változó-struktúrát is mutathatnak, érdemes a relatív szórást ellenőrizni a 6.3. táblázatban. Egyik szórás/átlag hányados sem közelíti meg a kettőt, mint kritikus értéket[101].

6.3. táblázat: A négy változó statisztikai jellemzői

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Odavanperfo	,043197	,0152623	50
Elvanperfo	,034468	,0109296	50
ÁllElvanperfo	,020327	,0074053	50
Állodavanperfo	,025357	,0124537	50

A változók mértékegységei nem különböznek, de nagyságrendi eltérések lehetnek, ezért a 6.4. táblázatban megadott korrelációs mátrixból indulunk. Minden korrelációs együttható szignifikáns, nem látunk blokkokat a változók között. Ebből feltételezhető, hogy a négy változóból egy főkomponens fog képződni. A mátrix nullához közeli (0,002) determinánsából sejthető, hogy a sajátértékek határozottan csökkenő sorozatot alkotnak.

[101] Lehet szigorúbb (pl. 0,7) kritikus értéket is választani, itt ez is teljesül.

6.4. táblázat: Az eredeti változók korrelációs mátrixa

Correlation Matrix[a]

		Odavanperfo	Elvanperfo	ÁllElvanperfo	Állodavanperfo
Correlation	Odavanperfo	1,000	,877	,838	,915
	Elvanperfo	,877	1,000	,940	,884
	ÁllElvanperfo	,838	,940	1,000	,908
	Állodavanperfo	,915	,884	,908	1,000
Sig. (1-tailed)	Odavanperfo		,000	,000	,000
	Elvanperfo	,000		,000	,000
	ÁllElvanperfo	,000	,000		,000
	Állodavanperfo	,000	,000	,000	

a. Determinant = ,002

A Kaiser-Meyer-Olkin (KMO) teszt 0,746-os értéke alapján adataink alkalmasak főkomponens elemzésre, és a Barlett-féle khi-négyzet teszt alapján minden szokásos szignifikancia szinten elvetjük a változók függetlenségének hipotézisét. (6.5/a. táblázat)

6.5/a. táblázat: PCA alkalmazhatósági tesztek

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,746
Bartlett's Test of Sphericity	Approx. Chi-Square	280,951
	df	6
	Sig.	,000

A változók egyedi alkalmasságát a 6.5/b. táblázat alsó mátrix főátlója adja meg. Az egyedi MSA értékek a KMO körül ingadoznak, egyik változó kihagyása sem indokolt, mindegyik meghaladja a 0,5 küszöböt. A főátlón kívül a parciális korrelációk (-1)-szeresei kaptak helyet.

6.5/b. táblázat: A változók egyedi alkalmasságának mérése

Anti-image Matrices

		Odavanperfo	Elvanperfo	ÁllElvanperfo	Állodavanperfo
Anti-image Covariance	Odavanperfo	,127	-,049	,032	-,073
	Elvanperfo	-,049	,087	-,062	,015
	ÁllElvanperfo	,032	-,062	,080	-,048
	Állodavanperfo	-,073	,015	-,048	,093
Anti-image Correlation	Odavanperfo	,752[a]	-,466	,316	-,667
	Elvanperfo	-,466	,755[a]	-,737	,170
	ÁllElvanperfo	,316	-,737	,718[a]	-,551
	Állodavanperfo	-,667	,170	-,551	,758[a]

a. Measures of Sampling Adequacy(MSA)

Ha sztenderdizált adatokkal dolgozunk, akkor kezdetben minden változó szórásnégyzete egységnyi (Initial), és ebből az egynél nagyobb varianciájú, „fontos" főkomponens(ek) bizonyos hányadot magyaráz(nak) (Extraction), amint ezt a 6.6. táblázat mutatja. Ha a magyarázott hányad túlságosan alacsony lenne[102], akkor a változót célszerű lenne kihagyni a futtatásból. Példánkban mind a négy változó esetében 90% közeli vagy ezt meghaladó a megőrzött információ. A négy kommunalitás összege pedig 3,6 felett van, ami előre jelzi, hogy a teljes megőrzött információ is 90% felett lesz.

6.6. táblázat: A teljes variancia megőrzött hányada

Communalities

	Initial	Extraction
Odavanperfo	1,000	,894
Elvanperfo	1,000	,931
ÁllElvanperfo	1,000	,923
Állodavanperfo	1,000	,933

[102] Ha a kommunalitás kisebb, mint 0,25, akkor a változó egyetlen faktorral sem korrelál közepesen, mert $0{,}5^2 = 0{,}25$. A kommunalitás többszörös determinációs együtthatóként értelmezhető.

A megmagyarázott variancia hányada 3,861/4= 92%, így a négydimenziós térből képzett egyetlen komponenssel csak 8%-át veszítjük el az eredeti információból. (6.7. táblázat) A második komponens jóval kevesebb információt hordoz, mint egy eredeti változó, mivel varianciája (0,183) kisebb, mint egy. Ha ilyen erős az egyetlen komponens, amit előállítunk, akkor főfaktornak is szokás nevezni az eredményt.

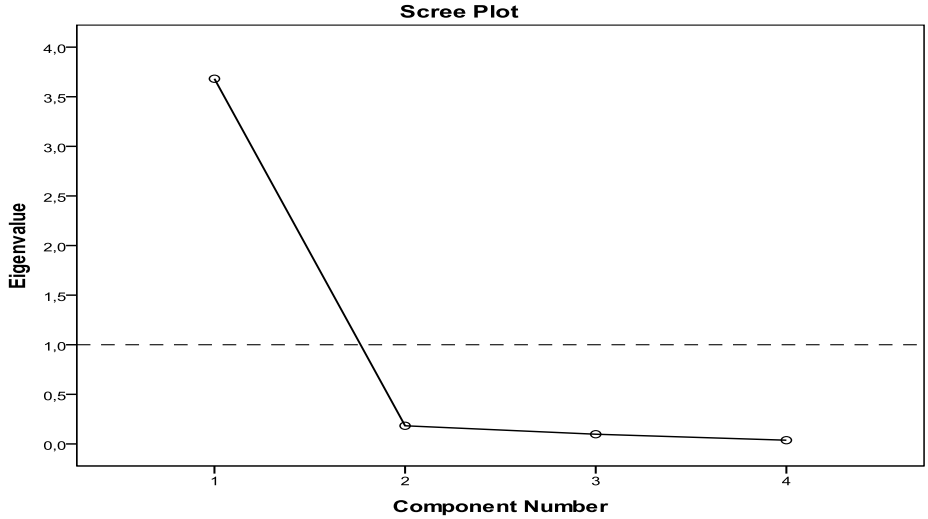
6.7. táblázat: A főkomponensek sajátértékei és relatív fontosságuk

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,681	92,036	92,036	3,681	92,036	92,036
2	,183	4,576	96,612			
3	,098	2,448	99,060			
4	,038	,940	100,000			

Extraction Method: Principal Component Analysis.

A sajátértékek monoton csökkenő sorozatát mutatja a 6.1. ábra. Ha a második, és a további komponensek csökkenése nem elég határozott, akkor az SPSS-ben a főkomponensek kívánt számát beállítva megismételjük a futtatást.



6.1. ábra: A sajátértékek sorozata

Az értelmezés szempontjából a komponens mátrix (6.8. táblázat) az egyik legfontosabb eredmény. Ez tartalmazza a változók és a főkomponens közötti korrelációkat, azaz a $\underline{C}$ mátrix első oszlopát. Minden változó szorosan és pozitív előjellel korrelál a komponenssel. Ez azt jelenti, hogy a komponens alapján a lakónépességre vetített magasabb oda- és elvándorlási adatokkal rendelkező kerületek és agglomerációs települések magasabb koordinátával rendelkeznek. (Nehezebb lenne értelmezni a kétpólusú, pozitív és negatív korrelációkat is tartalmazó komponens jelentését.)

6.8. táblázat: A változók és a főkomponens közötti korrelációk

Component Matrix[a]

	Component
	1
Odavanperfo	,946
Elvanperfo	,965
ÁllElvanperfo	,961
Állodavanperfo	,966

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

A PCA célja az, hogy az eredeti változók közötti korrelációkat jól megőrző, de kevesebb számú komponenst állítson elő. Ezért nemcsak a főkomponens(ek) nagyságát figyeljük, hanem az $\underline{R}$ reprodukálásának mértékét is. A 6.9. táblázat főátlójában a 6.6. táblázatban szereplő kommunalitásokat látjuk, a főátlón kívül pedig a (6.11) szerint számolt reprodukált korrelációk találhatók. A 6.4. táblázatbeli eredeti korrelációk és a 6.9. táblázat felső fele közötti eltéréseket reziduálisként adja meg a 6.9. táblázat alsó része.

A reziduálisok között abszolút értékben a legnagyobb a -0,070, amely arra utal, hogy az odavándorlás/fő és az állandó elvándorlás/fő között mért (0,838) korrelációt a főkomponens alapján némileg felülbecsüljük (0,909). Ez az egyetlen korreláció, ahol a becslési hiba meghaladja a 0,05-t. (Ezt a b. jelű megjegyzés is rögzíti.)

6.9. táblázat: A korrelációk becsült értékei és a hibatagok

Reproduced Correlations

		Odavanperfo	Elvanperfo	ÁllElvanperfo	Állodavanperfo
Reproduced Correlation	Odavanperfo	,894[a]	,912	,909	,914
	Elvanperfo	,912	,931[a]	,927	,932
	ÁllElvanperfo	,909	,927	,923[a]	,928
	Állodavanperfo	,914	,932	,928	,933[a]
Residual[b]	Odavanperfo		-,035	-,070	,002
	Elvanperfo	-,035		,013	-,048
	ÁllElvanperfo	-,070	,013		-,021
	Állodavanperfo	,002	-,048	-,021	

Extraction Method: Principal Component Analysis.
a. Reproduced communalities
b. Residuals are computed between observed and reproduced correlations. There are 1 (16,0%) nonredundant residuals with absolute values greater than 0.05.

A faktortérbeli ábrához ismernünk kell a település-score-okat. Ezeket a főkomponens(ek)re, mint tengely(ek)re vonatkozó koordinátákat a (6.16) szerint számolt sztenderdizált regressziós együtthatókat (6.10. táblázat) használva állítjuk elő. Ha egy-egy település négy változóra megfigyelt értékeit behelyettesítjük az első oszlop alapján felírható regressziós egyenletbe, akkor megkapjuk az adott kerület vagy település koordinátáját az első főkomponens terében.

6.10. táblázat: A főkomponens együtthatók regressziós becslése

Component Score Coefficient Matrix

	Component
	1
Odavanperfo	,257
Elvanperfo	,262
ÁllElvanperfo	,261
Állodavanperfo	,262

Extraction Method: Principal Component Analysis.
Component Scores.

Mivel a főkomponens átlaga zérus, a pozitív koordináták „nyitott” települést jeleznek, ahol oda- és elvándorlás is jellemző, míg a negatív értékek a lakónépesség arányában „zártabb” településekhez tartoznak.
Összegezve a számításokat egy nagyon erős vándorlási komponenst kaptunk, amely az információ 92%-át megőrzi. A fővárosi kerületek és a Budapest közeli települések részletes vándorlási adatai helyett ez az egyetlen adatsor is használható a továbbiakban.

6.2. A faktorelemző módszercsalád további eljárásai

Ha az Analyze/Dimension Reduction/Factor úton elindulunk, az **„Extraction**” részben választhatunk másik eljárást.

Eddig az alapváltozatot, a **főkomponens elemzést** (PCA) ismertük meg. Ekkor azt tételezzük fel, hogy a korrelációs mátrixot tökéletesen reprodukálni tudjuk az $\underline{R}=\underline{A}\Lambda\underline{A}^T=\underline{C}\underline{C}^T$ szorzattal, ha a változókkal megegyező számú főkomponenst állítunk elő, azaz $\underline{Y}=\underline{X}\underline{A}$, ahol $\underline{Y}$ és $\underline{X}$ (nxp)-s mátrixok, $\underline{A}$, Λ és $\underline{C}$ pedig (pxp) méretűek.

A tökéletes reprodukció nem kizárólagos cél, és nem is mindig reális elvárás. Ha csak néhány közös faktort tételezünk fel, amelyekkel leírhatók a változók, akkor más eljárást választunk.

Legkisebb négyzetek módszerének (LKNM) súlyozatlan és súlyozott változatát használhatjuk, ha a faktorok száma adott, és keressük azt a faktorstruktúrát, amely minimalizálja a megfigyelt és a reprodukált korrelációs mátrixok közti p(p-1) eltérés négyzetösszegét. Csak a diagonális elemeken kívüli eltéréseket mérjük. A súlyozott LKNM-ben a korrelációkat a változók egyediségének[103] reciprokával súlyozzuk.

Maximum Likelihood (ML) faktoreljárást választhatunk, ha a változók többdimenziós normális eloszlást követnek, és a megfigyelt korrelációs mátrix a populáció korrelációs mátrixának „legvalószerűbb” becslése. Itt is az egyediség reciprokával súlyozunk, és iterációval kapjuk a megoldást. Adott k faktorszám mellett tesztelni kell az illeszkedés jóságát. A k-faktoros modell jóságát mérő statisztika (képlete: $n \cdot \ln\left|\hat{R}\right| / \left|R\right|$) nagy minta esetében khi-négyzet eloszlást követ.

Jó az illeszkedés, ha a próbafüggvény szignifikancia szintje magas. A 0,05 alatti alacsony szignifikancia szint esetén (k+1) faktorra megismételjük a futtatást. A faktorok száma nem haladhatja meg azt a legnagyobb egész számot, amire teljesül a következő egyenlőtlenség: $k<1/2(2p+1-(8p+1)^{1/2}$

Principal-axis factoring (PAF): Főfaktor módszer a főkomponens elemzéshez hasonló elvet követ, de az induló korrelációs mátrix diagonálisaiban álló egyeseket a becsült kommunalitásokkal cseréli ki. Ezt a redukált korrelációs mátrixot veti alá sajátérték-sajátvektor felbontásnak. A kívánt számú faktor előállítása után becsli a

[103] Egyediség=1-kommunalitás

faktormátrixban a „loading” súlyokat, ebből újrabecsli a kommunalitásokat, és az iteráció addig folytatódik, míg két egymást követő eredmény már csak minimálisan tér el. Ezen eljárás során több matematikai probléma vetődik fel, melyeket a modell ismertetése során tárgyalunk.

6.2.1. A faktorelemzés modellje

A centírozott (átlagtól való eltéréssel megadott) megfigyelések mátrixa felírható a közös faktorok lineáris kombinációja és az egyedi faktorok összegeként:

$\underline{X}=\underline{F}\underline{L}^T+\underline{H}$, melyben (6.17)

- $\underline{X}$ mérete (nxp), ahol n a megfigyelési egységek és p a változók száma
- $\underline{F}$ (nxk)-s, ahol k a közös faktorok száma (k< p)
- $\underline{L}$ (pxk)-s, a faktorsúlyok mátrixa (loading)
- $\underline{H}$ (nxp)-s egyedi faktor, hibatag mátrix.

Feltevések:

- A faktorok lineárisan függetlenek:

$\underline{F}^T\underline{F}/n=\underline{E}$, ahol $\underline{E}$ egy (kxk)-s egységmátrix (6.18)

- A közös faktor és a hibatag korrelálatlan: $\underline{F}^T\underline{H}=\underline{H}^T\underline{F}=\underline{0}$ (6.19)
- A hibatagok függetlenek, azaz variancia-kovariancia mátrixuk (pxp)-s diagonális mátrix: $\underline{H}^T\underline{H}/n=\underline{U}^2$ (6.20)

A megfigyelt változók korrelációs mátrixát (6.17) alapján felbontjuk, és a (6.18)-(6.20) feltevéseket felhasználva a faktorelemzés alapegyenletét kapjuk:

$$\underline{R}=\underline{X}^T\underline{X}/n=1/n(\underline{F}\underline{L}^T+\underline{H})^T(\underline{F}\underline{L}^T+\underline{H})=\underline{L}\underline{L}^T+\underline{U}^2 \quad (6.21)$$

Ha a korrelációs mátrix diagonális elemeiből levonjuk a hibatagok varianciáit, a változóknak a közös faktorok által magyarázott részét, a kommunalitásokat kapjuk.

Az $\underline{U}^2$ ismeretében az $\underline{R}$-$\underline{U}^2$ redukált korrelációs mátrix sajátérték-sajátvektor felbontását kell elvégezni:

$$\underline{R}_{red}=\underline{L}\underline{L}^T \quad (6.22)$$

A hibatagok varianciája ($\underline{U}^2$ főátlója) általában nem ismert, értékét a többszörös korrelációs együttható komplementereként becsüljük, vagy a kommunalitásból számoljuk: $u_i^2=1-h_i^2$ (6.23)

Mivel általában a kommunalitásokat sem ismerjük, alapértelmezés szerint a többszörös korrelációs együttható négyzete adja a kommunalitás becslését. Használható a PCA futtatásával kapott kommunalitás is, vagy a korrelációs mátrixban szereplő maximális páronkénti korrelációs együttható abszolút értéke.

A (6.22)-ben felírt redukált korrelációs mátrix sajátérték-sajátvektor felbontásakor:

$$\underline{R}_{red} = \underline{L}\underline{L}^T = \underline{V}\underline{\Lambda}\underline{V}^T \tag{6.24}$$

írható fel, melyben a $\underline{V}$ a sajátvektorok mátrixa, $\underline{\Lambda}$ pedig a sajátértékek diagonális mátrixa, és így $\underline{L}=\underline{V\Lambda}^{1/2}$ áll fenn.

A faktorok forgatása (rotációja)

Legyen $\underline{T}$ az ortogonális transzformáció mátrixa, melyre $\underline{T}^T\underline{T}=\underline{T}\underline{T}^T=\underline{E}$.

Az L faktorsúly mátrixot bármelyik eljárással (PCA, PAF, ML,…) állítottuk elő, a rotálás hatására: L*=LT lesz. De Rred=LLT= LTTTLT =L*L*T fennáll, azaz a redukált korrelációs mátrix és főátlójában a kommunalitások változatlanok maradnak.

Kétdimenziós térben az óra járásával egyező forgatást eredményez az alábbi transzformációs mátrix:

$$\underline{T} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

A főkomponens elemzéstől eltérően a faktorelemzésnek nem mindig van ***megfelelő*** megoldása, mert a redukált korrelációs mátrix nem pozitív definit.

- Csak a pozitív definit mátrixra teljesül az, hogy minden sajátérték nem-negatív. Ezért a faktorelemzésben a sajátértékek között negatívok is lehetnek, ezek pedig nem megfelelő megoldások, mert a sajátértékek a faktorok varianciáit fejezik ki, amelyek biztosan nem-negatív értékek.
- További problémát okoz az, hogy ha vannak negatív sajátértékek is, akkor az első néhány nagy pozitív sajátérték összege nagyobb lehet, mint a redukált mátrix nyoma, azaz a diagonális elemek összege. Ilyen esetben úgy tűnhet, hogy a dimenziócsökkentés után megőrzött információ meghaladja a 100%-ot.
- Problémát okozhat az is, ha a (6.21) alapegyenlet megoldása során kapott eredmény nem teljesíti a változó és a faktor közti kapcsolat szorosságát mérő korrelációs együtthatókkal szembeni elvárásokat, és/vagy a hibatag varianciájára negatív érték adódik.

Az említett problémák előfordulását kis mintapéldán mutatjuk be.

Három változónk korrelációs mátrixa legyen a következő:

$$\underline{R} = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix}$$

és k=1 faktort tételezünk fel, azaz az $\underline{F}$ mátrix (nx1)-s vektor, az $\underline{L}$ pedig 3 elemű vektor.

A megfigyelések (nx3)-s méretű $\underline{X}$ mátrixa tehát oszloponként így írható fel:

$X_{i1} = l_1 F_i + h_{i1}$, ahol i=1,…,n

$X_{i2} = l_2 F_i + h_{i2}$

$X_{i3} = l_3 F_i + h_{i3}$

A loadingok és a hibatagok a (6.21) alapegyenlet értelmében egyenlők a korrelációs mátrix elemeivel az alábbiak szerint:

$$\begin{array}{lll} 1 = l_1^2 + u_1^2 & 0{,}9 = l_1 l_2 & 0{,}7 = l_1 l_3 \\ & 1 = l_2^2 + u^2{}_2 & 0{,}4 = l_2 l_3 \\ & & 1 = l_3^2 + u^2{}_3 \end{array}$$

Ha a 0,7 és 0,4 korrelációs együtthatókra felírt egyenleteket elosztjuk egymással, akkor l_3 kiesik, és például l_2 kifejezhető: $l_2 = 4/7 l_1$

Ezt behelyettesítve

$0{,}9 = l_1 l_2 = 4/7 l_1^2$ és innen $l_1^2 = 1{,}575$

Gyökvonás után $l_1 = \pm 1{,}255$

Egyik érték sem megfelelő, mivel l_1 az (egységnyi szórású) változó és a (szintén egységnyi szórású) faktor közötti korrelációt méri, és a korreláció maximuma 1.

A főátlóban pedig az első hibatag szórásnégyzetére negatív szám (1-1,575 = -0,575) adódik, és ez sem megfelelő érték. Létezik tehát megoldás, de a kapott eredmény nem fogadható el. Valós méretű feladatok esetében halmozottan jelentkezhetnek a problémák, ezért csak stabil, jól felépített modell birtokában javasolható a főfaktorok előállítása.

6.2.2. A PAF eredmények bemutatása és értelmezése

Az országok politikai, gazdasági és pénzügyi kockázatát több szakértő különböző módon és eltérő gyakorisággal méri, de feltételezhetjük, hogy létezik a háttérben egy közös ország-kockázat faktor, és a publikált kockázati mértékek ennek a hatását tükrözik. Ezt az elméleti megfontolást szem előtt tartva végzünk főfaktor elemzést a Világbank által közzétett három kockázati mérőszámra. Mindhárom kockázati mérték 0 és 100 között mér, a nagyobb érték jelenti a kisebb kockázatot.

Az „Investmentclimate.sav” adatok három változójára Dimension Reduction/ Extraction/ Principal axis factoring választással faktort állítunk elő. A többi beállítás a PCA futtatással megegyezik, egy faktor esetében rotálás nem végezhető.

A leíró statisztikák (6.11. táblázat) szerint a sok tényezőből súlyozottan készített (kompozit) mutató átlaga magasabb, szórása kisebb, mint a nemzetközi bankok és a gazdasági elemzők szakértői véleményét tükröző két mérőszám.

6.11. táblázat: Átlagok és szórások

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Composite ICRG risk rating	74,365	11,355	31
Institutional Investor credit rating	64,610	26,538	31
Country credit worthiness rating (Euromoney)	68,597	24,109	31

A mutatók között nagyon szoros, 0,9 feletti a páronkénti korreláció, ezért megalapozottnak tűnik feltételezésünk, hogy közös faktor létezik. (6.12. táblázat)

6.12. táblázat: Korrelációs mátrix

Correlation Matrix[a]

		Composite ICRG risk rating	Institutional Investor credit rating	Country credit worthiness rating (Euromoney)
Correlation	Composite ICRG risk rating	1,000	,921	,925
	Institutional Investor credit rating	,921	1,000	,992
	Country credit worthiness rating (Euromoney)	,925	,992	1,000
Sig. (1-tailed)	Composite ICRG risk rating		,000	,000
	Institutional Investor credit rating	,000		,000
	Country credit worthiness rating (Euromoney)	,000	,000	

a. Determinant = 2,398E-03

A tesztek (6.13. táblázat) is azt bizonyítják, hogy adataink alkalmasak látens változó előállítására. Ez a rész megegyezik a PCA és a PAF eljárásoknál.

6.13. táblázat: Alkalmassági tesztek

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,739
Bartlett's Test of Sphericity	Approx. Chi-Square	169,936
	df	3
	Sig.	,000

A közös faktor által magyarázott variancia hányadát mutató kommunalitások (6.14. táblázat) első oszlopa a PCA eredményt mutatja, második oszlopa pedig a főfaktorhoz tartozó kommunalitást.

6.14. táblázat: PCA és PAF kommunalitások

Communalities

	Initial	Extraction
Composite ICRG risk rating	,856	,860
Institutional Investor credit rating	,983	,988
Country credit worthiness rating (Euromoney)	,984	,994

Extraction Method: Principal Axis Factoring.

A 6.15. táblázat alapján a redukált korrelációs mátrix sajátértéke és relatív fontossága (2,842 és 94,7%) valamivel kisebb, mint az eredeti korrelációs mátrix dekompozíciójából kapott sajátérték (2,892).

6.15. táblázat: PCA és PAF sajátértékek

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,892	96,403	96,403	2,842	94,734	94,734
2	9,959E-02	3,320	99,723			
3	8,325E-03	,277	100,000			

Extraction Method: Principal Axis Factoring.

A PCA komponens mátrixa (*C*) helyett itt *L* faktormátrixot (6.16. táblázat) ad az eljárás, amelyből látható, hogy a főfaktor és mindhárom változó között nagyon szoros pozitív korreláció van.

6.16. táblázat: Főfaktor súlyok

Factor Matrix[a]

	Factor
	1
Composite ICRG risk rating	,927
Institutional Investor cre rating	,994
Country credit worthines rating (Euromoney)	,997

Extraction Method: Principal Axis Factc
a. 1 factors extracted. 4 iterations re

A faktor score-ok regressziós becslésében (6.17/a. és 6.17/ b. táblázat) viszont jelentősen eltérnek az együtthatók, bár mindkét számítás a sztenderdizált regressziós együttható (béta) értékeket adja.

*6.17./a táblázat: **PAF** eljárással számolt sztenderdizált regressziós együtthatók*

Factor Score Coefficient Matrix

	Factor
	1
Composite ICRG risk rating	,026
Institutional Investor credi rating	,308
Country credit worthiness rating (Euromoney)	,668

Extraction Method: Principal Axis Facto

6.17./b táblázat: PCA eljárással számolt sztenderdizált regressziós együtthatók

Component Score Coefficient Matrix

	Compone nt
	1
Composite ICRG risk rating	,334
Institutional Investor credit rating	,342
Country credit worthiness rating (Euromoney)	,342

Extraction Method: Principal Component Analysis.

Az eredeti korrelációk előállítása a főfaktorral nagyon jól sikerült, a főátlón kívüli reziduálisok zérusnak tekinthetők a 6.18. táblázat alapján. Meggyőződtünk tehát arról, hogy egy főfaktort feltételező modellünk jól illeszkedik a mért változókhoz, tehát a kockázati faktor alkalmas arra, hogy az országokat kockázat szerint rangsoroljuk, csoportosítsuk.

Felvetődik azonban a kérdés, hogy mennyire más a PAF és a PCA eredménye? Mivel a változók közötti korrelációk nagyon szorosak voltak, és a 6.14. valamint a 6.6. táblázat alapján a két eljárás eredményei nem térnek el jelentősen, nem meglepő, hogy a PCA és a PAF koordináták közötti determinációs együttható 0,9861. A koordináták egyezését a 6.2. ábra mutatja. Tökéletes egybeesést a 45^{o} egyenes pontjai mutatnak. Nagyobb eltérést csak Törökország score-jai között találunk, mivel a PAF (-0,69) jelentősen felülbecsli a főkomponens (-1,01) koordinátát.

6.18. táblázat: Az eredeti korrelációk előállítása a főfaktorral

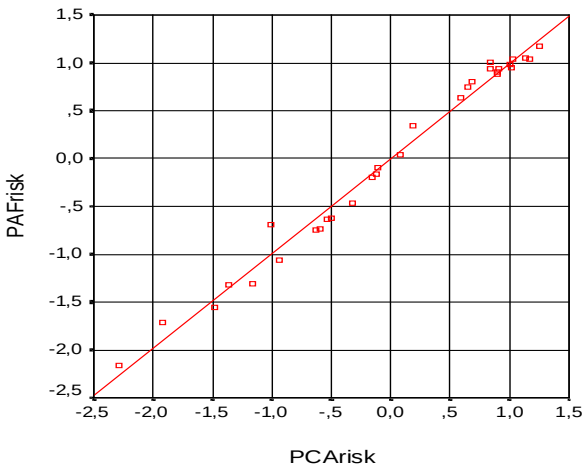
Reproduced Correlations

		Composite ICRG risk rating	Institutional Investor credit rating	Country credit worthiness rating (Euromoney)
Reproduced Correlation	Composite ICRG risk rating	,860[b]	,922	,925
	Institutional Investor credit rating	,922	,988[b]	,991
	Country credit worthiness rating (Euromoney)	,925	,991	,994[b]
Residual[a]	Composite ICRG risk rating		,000	,000
	Institutional Investor credit rating	,000		,000
	Country credit worthiness rating (Euromoney)	,000	,000	

Extraction Method: Principal Axis Factoring.

a. Residuals are computed between observed and reproduced correlations. There are 0 (,0%) nonredundant residuals with absolute values > 0.05.

b. Reproduced communalities



6.2. ábra: PCA és PAF koordináták pontdiagramja

6.3. A faktorelemzés további kihívásai

Nem célunk a tisztelt olvasó megtévesztése. Nem kapunk mindig egyetlen és főleg jól értelmezhető faktort/főkomponenst a futtatás végén. Most a gyakorlatban előforduló nehézségekre is mutatunk példát úgy, hogy a fejezet elején feltett kérdésre keressük a választ, azaz a települések életminőségét mérjük.

6.3.1. Abszolút és relatív mutatók elemzése

A Kerületek2010.sav adatállományban a tényleges vándorlási adatok, mint abszolút számok mellett a lakónépességre vetített – relatív – mutatók is szerepelnek. Melyiket érdemes az elemzésbe bevonni? Ezen szakmai kérdés mellé további statisztikai részkérdések is feltehetők:

a) Melyik változókörre kapunk jobban illeszkedő faktormodellt?

b) Mely részeredmények változnak, ha egyik vagy másik változócsoportot vonjuk be?

c) Egy közös modellben elemezzük a változókat, vagy két faktor-futtatást készítsünk?

Készítsük el és ellenőrizzük eredményeinket négy változatban: csak az abszolút (A10) változókra, az abszolút mellett relatív (AR10) mutatók felhasználásával, valamint külön változókörre (K6, K4) futtatás esetén. A változók listája a 6.19. táblázatban szerepel.

A továbbiakban csak néhány részeredményt emelünk ki. Érdemes a négy változatot önállóan elkészíteni és tanulmányozni.

6.19. táblázat: A változók szerepe a négy különböző modellben

Változók és modellek	A modell illeszkedése, főbb következtetések
(A10) n=50 és p=10 Népességszám Odavándorlás Elvándorlás Állandóodavándorlás Állandóelvándorlás Önkormányzatibev Vendéglátóhely Lakásállomány Építettlakások Álláskeresők	**KMO mérték: 0,850** Legkisebb kommunalitás: Épített lakások (0,587) Egy feletti sajátérték és %: 8,454 (84,54%) Az 1. komponens tartalma: eleve nagy méretű és vándorlásban is kiemelkedő település A 2. komponens: nincs Összesített minősítés: jól értelmezhető modell

(AR10) n=50 és p=10 Népességszám Önkormányzatibev Vendéglátóhely Lakásállomány Építettlakások Álláskeresők Odavanperfo Elvanperfo ÁllElvanperfo Állodavanperfo	KMO mérték: 0,828 Legkisebb kommunalitás: Épített lakások (0,756) Egy feletti sajátérték és %: 6,8 (68%) és 2,045 (20,45%) Az 1. komponens tartalma: méret és életfeltételek A 2. komponens tartalma: vándorlás Összesített minősítés: rotálás után jól értelmezhető modell *(ezt részletesen is bemutatjuk az 5.3.2.-ben)*
(K6) n=50 és p=6 Népességszám Önkormányzatibev Vendéglátóhely Lakásállomány Építettlakások Álláskeresők	KMO mérték: 0,822 Legkisebb kommunalitás: Épített lakások (0,558) Egy feletti sajátérték és %: 4,965 (82,75%) Az 1. komponens tartalma: méret és életfeltételek Összesített minősítés: jól értelmezhető modell
(K4) n=50 és p=4 Odavanperfo Elvanperfo ÁllElvanperfo Állodavanperfo	KMO mérték:0,746 Legkisebb kommunalitás: odavándorlás/fő (0,894) Egy feletti sajátérték és %: 3,681 (92%) Az 1. komponens tartalma: vándorlás Összesített minősítés: jól értelmezhető modell

Vajon miért van az, hogy háromszor egy faktoros, és egyszer két faktoros eredmény adódott? Miből ered ez a különbség?

Ismét a mérethatásra emlékeztetünk. A mutatók többsége egymással együttmozog, erős a multikollinearitás, ezért az (A10) modell KMO-ja a legmagasabb. Ha az egyik mutató nagyobb értéket ér el, akkor a másik is magasabb. De a második modellben, az (AR10)-ben relatív mutatók is szerepelnek, és ezek különülnek el a többi változótól. Ez azzal magyarázható, hogy a létszámhoz viszonyított vándorlás másként alakulhat, mint a vándorlás önmagában.

Hasonlót lehet tapasztalni vállalati adatok elemzése esetén is. Más lesz a komponensek tartalma és értelme, ha az árbevétel, az eredmény, stb. mutatókat összesen értékben használjuk, vagy ezeket egy főre vetítjük.

6.3.2. Kétdimenziós megoldás értelmezése, ábrázolása

Az elemzések során az a gyakoribb, hogy nem sikerül egyetlen faktorba tömöríteni az összes változót, hanem több, egynél nagyobb sajátérték adódik. Ez nem von le semmit az eredmények erejéből, sőt lehetőséget teremt két vagy háromdimenziós ábrák készítésére, a megfigyelések szerkezetének feltárására.

A PCA/PAF futtatások első néhány táblázata (leíró statisztika, korreláció, KMO, Bartlett teszt, anti-image korrelációk, kommunalitások) nem tér el az eddig bemutatott output tábláktól, ezért ezeket itt nem közöljük. Csak az újabb eredmények értékelő bemutatására törekszünk. Két tengelyre már rotálás is kérhető, és ez a 6.20. táblázatban látható újabb eredményeket ad.

6.20. táblázat: Eredeti sajátértékek és rotált megoldás

	Initial Eigenvalues			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6,800	67,999	67,999	4,774	47,738	47,738
2	2,045	20,453	**88,452**	4,071	40,714	**88,452**
3	,391	3,915	92,367			
4	,313	3,126	95,492			
5	,140	1,401	96,894			
6	,101	1,009	97,902			
7	,092	,918	98,820			
8	,079	,788	99,608			
9	,029	,295	99,903			
10	,010	,097	100,000			

A 10 változóból kinyert 88%-nyi összes információ nem nőhet meg a rotálás során, de a tengelyek közötti szétosztás 68+20 százalékról indulva 48+40%-ra, azaz jelentősen megváltozik. (Kivételes esetekben a második komponens sajátértéke rotálás után meghaladhatja az elsőt!)

A rotálás a faktorok értelmezésében, a változók tengelyekhez rendelésében, a tiszta struktúra kialakításában segít. A komponens mátrix C rotálás előtti (6.21. táblázat) és utáni (6.23. táblázat) elemeit, valamint a két ábrát (6.3/a. és 6.3/b.) is bemutatjuk, hogy e művelet hatását érzékeltetni tudjuk.
Az első pillantásra értelmezhetetlen komponens mátrixot látunk a 6.21 táblázatban. Szinte minden változó közepes vagy erős korrelációt mutat mindkét faktorral, az épített lakások és az odavándorlás/fő mutatók közel azonosan korrelálnak mindkét tengellyel, tehát mintha középen, a 45 és a 135 fokos egyenes mentén lennének. (Ezt megerősíti a 6.3/a. ábra)

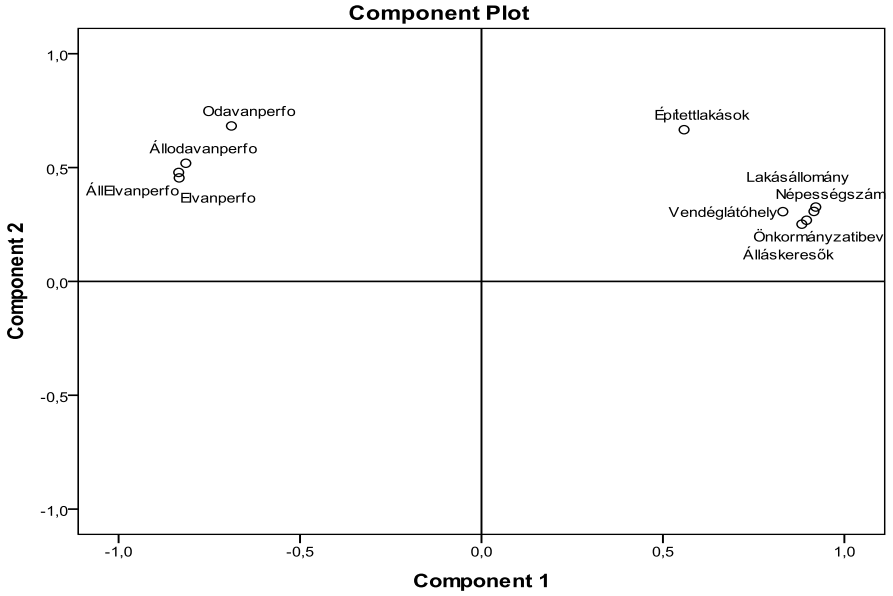
6.21. táblázat: Eredeti komponens mátrix

Component Matrix[a]

	Component	
	1	2
Népességszám	,916	,307
Önkormányzatibev	,896	,269
Vendéglátóhely	,830	,306
Lakásállomány	,921	,327
Építettlakások	,558	,666
Álláskeresők	,883	,251
Odavanperfo	-,689	,683
Elvanperfo	-,834	,479
ÁllElvanperfo	-,833	,455
Állodavanperfo	-,815	,519

Extraction Method: Principal Component Analysis.

a. 2 components extracted.



6.3/a. ábra: 10 változó leképezése két dimenzióba

6.22. táblázat: A forgatás mértéke

Component Transformation Matrix

Component	1	2
1	,758	-,653
2	,653	,758

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

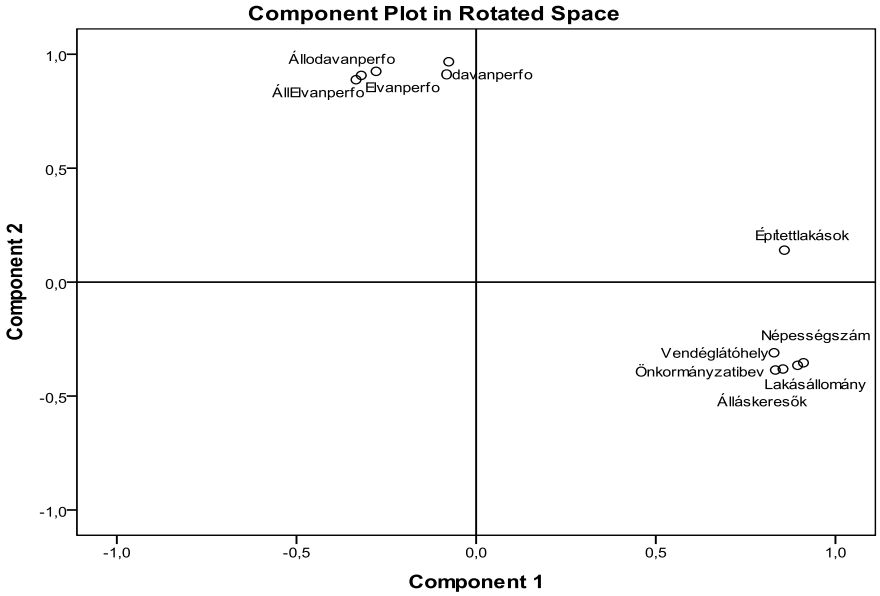
A variancia maximálizáló rotáció (6.22. táblázat) megtalálja azt a 40 fok[104] közeli szöget, amivel a kis súlyok még kisebbek, a nagyok pedig még nagyobbak lesznek, és kialakul egy értelmezhetőbb struktúra a 6.21/b táblázatban és a 6.3/b. ábrán.

6.23. táblázat: Rotált komponens mátrix

Rotated Component Matrix[a]

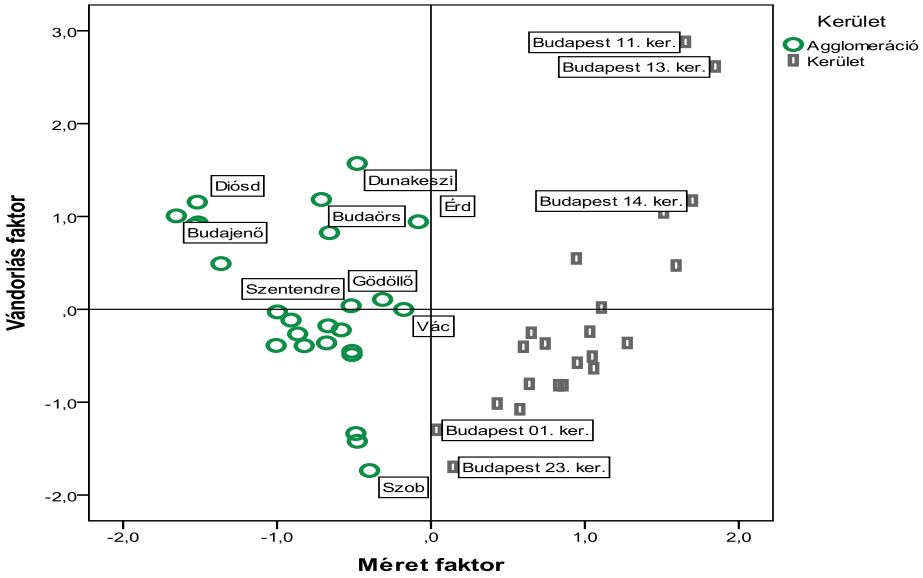
	Component	
	1	2
Népességszám	**,895**	-,365
Önkormányzatibev	**,854**	-,381
Vendéglátóhely	**,829**	-,310
Lakásállomány	**,911**	-,354
Építettlakások	**,858**	,140
Álláskeresők	**,833**	-,386
Odavanperfo	-,076	**,967**
Elvanperfo	-,320	**,907**
ÁllElvanperfo	-,334	**,888**
Állodavanperfo	-,278	**,925**

[104] Mivel cosα=0,758, a szög 40-41 fok között van.



6.3/b. ábra: 10 változó leképezése rotált tengelyekre

A változók elhelyezkedése alapján a síknegyedeket is jellemezni tudjuk a 6.4. ábrán, ahol a települések szerkezete látható. Emlékezzünk rá, hogy 10 változóból kiindulva, 88%-os információsűrítés után kaptuk a kétdimenziós vetületet!



6.4. ábra: 50 település 2 dimenziós faktortérben

Az első tengely szétválasztja a fővárost (átlag felettiek) és az agglomerációt (átlag alattiak). Ez felveti azt a kérdést, hogy a két almintára vajon külön elemzést kell-e végezni? A választ az alfejezet végén adjuk meg.

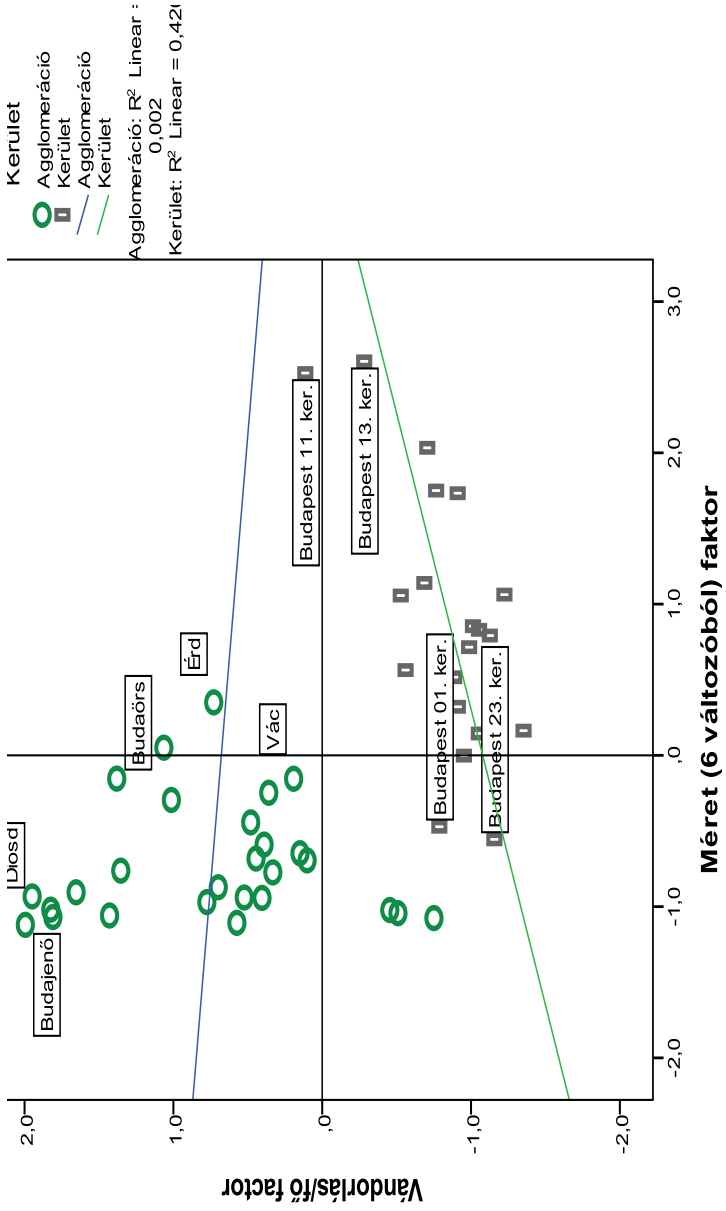
Az első síknegyedben csak fővárosi kerületek vannak, ezek az átlagnál nagyobb méretűek (létszám, lakás) és jobb életfeltételt jelentenek, hisz több a vendéglő és magasabb az önkormányzati bevétel. Ezek vándorlási mutatók szerint is vonzó célpontok. Balra fent a XI. és XIII. kerületet látjuk. (A III., IX. és XIV. kerületek találhatók még itt.)
Alattuk, a negyedik síknegyedben vannak Budapest további kerületei. Ezek kisebb méretűek, és nem jellemző rájuk nagy vándorlás. Legalul van az I. és a XXIII. kerület.
A második síknegyedben a kisebb, de vonzó célpontok között Budajenő és Diósd, mellettük vannak átlag közeli mérettel és jelentős vándorlással: Érd, Budaörs és Dunakeszi.
Az origóhoz legközelebbi pontunk, amely mindkét faktor szerint átlagos értékű: Vác.
A harmadik síknegyed a kisebb és zártabb településeket, falvakat foglalja magában. Ide tartozó pontként Szob említhető.
Közös modell tehát a megoldás vagy a két változóhalmaz külön sűrítését célszerű megpróbálni? Ezzel válaszolunk az 6.3.1. b) és c) kérdésekre is.

Először tekintsünk rá ismét a 6.3/a és a 6.3/b ábrákra. A rotálással nem sikerült teljesen tiszta struktúrát kapni, hiszen a 6.23. táblázat komponens mátrixában még több közepes korreláció látható. Nem teljesül az az elvárás, hogy egy-egy változó csak egy komponenssel korrelál.

Ha két számítássorozatot végzünk, és a 6.3.1-ben vázolt (K6) valamint (K4) elemzéseket egymástól elhatárolva végezzük el, akkor az előállított főkomponensek merőlegessége nem lesz elvárt. A K6=méret és a K4=vándorlás faktorok közötti korreláció -0,552 lesz, tehát valóban nem merőlegesek egymásra. A 6.5. ábra mutatja a külön becsült score-ok terében a megfigyelt kerületeket és településeket.

Három fontos megjegyzést érdemes átgondolni:

- A fővárosi kerületek értékei pozitív korrelációt mutatnak: a nagyobb méretű kerületekben nagyobb vándorlást jeleznek az adatok. (R-négyzet=0,420)
- Az agglomerációban viszont nem korrelál a két komponens egymással. (R-négyzet=0,002)
- Az 50 megfigyelésre tehát úgy adódik negatív korreláció, hogy a két almintában pozitív korreláció, valamint korrelálatlanság tapasztalható.



6.5. ábra: 50 település 2 külön becsült faktor terében

Ha ilyen eredményeket tapasztalunk, akkor nem érdemes erőltetni az összes változó egy modellben való sűrítését. Sőt azt is meg kell fontolni, hogy a két almintára jellemző komponenseket külön állítsuk elő.

Amikor arról döntünk, hogy a teljes mintára vagy külön fővárosra és külön agglomerációra készüljön a modell, akkor újabb korlátba ütközhetünk. Az alminták használata kisebb elemszámokat eredményez. Ha 23 és 27 a megfigyelések száma, akkor az n>5p hüvelykujj szabály miatt csak 4-5 változó egyidejű használata célszerű.

A számítások két úton végezhetők el:

1) Előre leszűrjük az adatokat a SELECT menűpontban, és csak az egyik felét használjuk. Ilyenkor csak a vizsgált almintára kapjuk meg a faktor-score-okat.

2) A faktor-futtatáson belül használjuk szelekciós változónak a „kerület” nevű dummy változót, ami a kerületekre=1, különben=0. Így a teljes adatállományra elkészül a faktor-score-ok becslése. Végül a két futtatás eredménye numerikusan és grafikusan vethető össze.

6.4. Idősorok faktorelemzése

Az öt tőzsdeindex elemzését már az 1. fejezetben megkezdtük, most folytatjuk. Nem a valóságtól elrugaszkodott az a feltételezés, hogy ezek viselkedése az időben együttmozog, még akkor is, ha nem tudjuk, hogy melyik okozza a másik változását. Inkább az a jogos feltevés, hogy a hátterében egy meg nem figyelhető faktor – nevezhetjük világ-kockázatnak, tőzsdei bizonytalanságnak – húzódik meg. Ennek a látens tényezőnek a feltárása elvégezhető faktorelemzéssel. A fejezetben ismertetett lépések a közönséges, és nem a dinamikus faktorelemzést[105] követik.

6.4.1. Differenciák faktorelemzése

Az Indexek.sav adatállományban a tőzsdeindexekből képzett differenciák már stacionáris viselkedésűek, ezért alkalmasak lehetnek főkomponens(ek) előállítására. Ugyanakkor a differenciák relatív szórása túl magas, a lineáris korrelációk (6.24. táblázat) pedig nem elég szorosak, ami megkérdőjelezi a homogén adatállomány mögött meghúzódó közös faktor feltevésünk teljesülését. Érdemes észrevenni, hogy New York differencia-adatai kevésbé korrelálnak a többi tőzsdével. Ebből számítani lehet arra, hogy gyengébb lesz az információ-sűrítés.

[105] A dinamikus faktorelemzés eljárást Bánkövi György – Veliczky József – Ziermann Margit dolgozták ki 40 évvel ezelőtt, és mutatták be számos írásukban. Számítógépes változata nem része a statisztikai programcsomagoknak.

6.24. táblázat: Korrelációs együtthatók

Correlation Matrix[a]

		DBUX	DUKX	DDJI	DDAX	DNKY
Correlation	DBUX	1,000	,486	,280	,468	,282
	DUKX	,486	1,000	,477	,796	,284
	DDJI	,280	,477	1,000	,542	,102
	DDAX	,468	,796	,542	1,000	,270
	DNKY	,282	,284	,102	,270	1,000
Sig. (1-tailed)	DBUX		,000	,000	,000	,000
	DUKX	,000		,000	,000	,000
	DDJI	,000	,000		,000	,000
	DDAX	,000	,000	,000		,000
	DNKY	,000	,000	,000	,000	

A KMO mutató értéke az outputban:0,751, ami közepes modellt jelez, de a DNKY (New Yorki tőzsde) kommunalitása a 6.25. táblázat szerint nagyon alacsony, a változó elhagyása megfontolandó. A gyenge korreláció és az alacsony kommunalitás a nem lineáris kapcsolatból adódhat. Ha jelentősége miatt nem az elhagyás mellett döntünk, akkor a második faktort érdemes előállítani, amiben különválik New York, hiszen a 6.26. táblázat sajátértékei közül a második nagyon közel van egyhez, és közel 19 százalékkal emeli az összesen megőrzött információt.

6.25. táblázat: A differencia-változókból megőrzött információ

Communalities

	Initial	Extraction
DBUX	1,000	,470
DUKX	1,000	,776
DDJI	1,000	,454
DDAX	1,000	,794
DNKY	1,000	,198

Extraction Method: Principal Component Analysis.

6.26. táblázat: 5 indexből 1 vagy 2 komponens képezhető

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,691	53,830	53,830	2,691	53,830	53,830
2	,944	18,881	72,711			
3	,651	13,029	85,740			
4	,514	10,280	96,020			
5	,199	3,980	100,000			

Extraction Method: Principal Component Analysis.

6.4.2. Tőzsdehányadosok faktorelemzése

Az Indexek.sav adatállományban a tőzsdeindexekből képzett hányadosak is szerepelnek, ezek is stacionárius viselkedésűek, ezért alkalmasak lehetnek főkomponens(ek) előállítására.

A hányadosok (ráták) relatív szórásai nagyon kicsik, mind az öt 0,1 alatti (6.27. táblázat), a lineáris korrelációk (6.28. táblázat) pedig kicsit változtak: néhol nőttek, néhol csökkentek. A KMO=0,754 hajszányit javult, és ha két komponenst kérünk (6.29. táblázat), akkor minden kommunalitás megfelelő (6.30. táblázat)

6.27. táblázat: A relatív szórások ellenőrzése

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
RBUX	1,0006	,01703	2753
RUKX	1,0000	,01330	2753
RDJI	1,0001	,01299	2753
RDAX	1,0002	,01669	2753
RNKY	1,0000	,01594	2753

6.28. táblázat: A tőzsdehányadosok közötti korrelációk

Correlation Matrix[a]

		RBUX	RUKX	RDJI	RDAX	RNKY
Correlation	RBUX	1,000	,506	,299	,468	,301
	RUKX	,506	1,000	,488	,790	,295
	RDJI	,299	,488	1,000	,573	,119
	RDAX	,468	,790	,573	1,000	,260
	RNKY	,301	,295	,119	,260	1,000
Sig. (1-tailed)	RBUX		,000	,000	,000	,000
	RUKX	,000		,000	,000	,000
	RDJI	,000	,000		,000	,000
	RDAX	,000	,000	,000		,000
	RNKY	,000	,000	,000	,000	

a. Determinant = ,161

6.29. táblázat: A második komponens előállítása megfontolandó

	Initial Eigenvalues			Original and Rotation Sums of Squared				
	Total	% of Variance	Cumulative %	Total	% of Variance	Total	% of Variance	Cumulative %
1	2,732	54,639	54,639	2,732	54,639	2,358	47,156	47,156
2	,939	18,790	**73,429**	,939	18,790	1,314	26,272	**73,429**
3	,634	12,680	86,109					
4	,494	9,882	95,991					
5	,200	4,009	100,000					

6.30. táblázat: Két komponens mellett a kommunalitások megfelelőek

Communalities

	Initial	Extraction
RBUX	1,000	,545
RUKX	1,000	,782
RDJI	1,000	,672
RDAX	1,000	,821
RNKY	1,000	,851

Extraction Method: Principal Component Analysis.

A két komponens tartalmát a rotálás után a 6.31. táblázatban és a 6.6. ábrán megvizsgálva észrevehetjük a budapesti tőzsde épp „középen” van, egyrészt

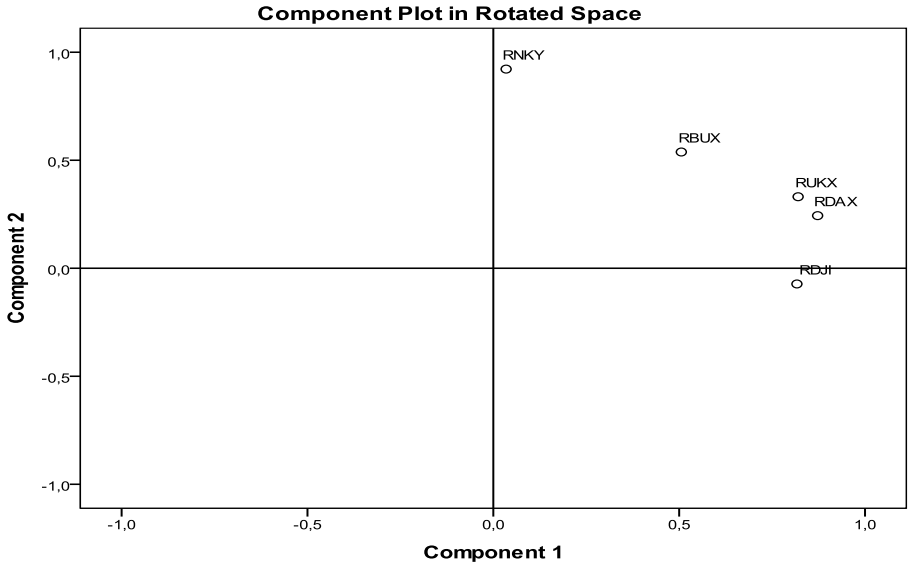
együttmozog az angol-német-japán tőzsdékkel az 1. komponens pozitív korrelációi alapján, másrészt erősebben együttmozog az amerikai adatokkal, mint bármelyik másik nagy tőzsde.

6.31. táblázat: Rotált tőzsdehányados komponensek

Rotated Component Matrix[a]

	Component	
	1	2
RBUX	**,506**	**,538**
RUKX	**,820**	,331
RDJI	**,817**	-,073
RDAX	**,872**	,244
RNKY	,034	**,922**

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.



6.6. ábra: Tőzsdeindex hányadosok faktortérben

Utolsó megfontolásként gondoljuk át a közös faktor feltevést és a PAF faktorbeállítás alkalmazását. A két faktor által megőrzött összes információ 53%-ra esik vissza, míg a PCA 73,4%-ot jelzett a 6.27. táblázatban.
Mivel a kezdeti kommunalitások (6.32. táblázat) az alacsony eredeti korrelációk miatt kicsik, összességében is gyenge eredményeket kapunk.

6.32. táblázat: A faktormodell kommunalitásai

Communalities

	Initial	Extraction
RBUX	,293	,412
RUKX	,654	,737
RDJI	,334	,378
RDAX	,675	,877
RNKY	,121	,252

Extraction Method: Principal Axis Factoring.

A New Yorki tőzsde ráta elhagyása az elemzésből statisztikailag határozottan javasolható, de emellett a német és a magyar kommunalitás is alacsony. A rátaváltozók mögött a vizsgált 11 évben nem húzódott meg 1-2 közös faktor.

7. Diszkriminancia elemzés

7.1. A diszkriminanciaelemző eljárás alapgondolata

Megfigyeléseink sok esetben nem homogének, és már előzetesen csoportokba sorolva állnak rendelkezésünkre a változók mentén mért értékek. A csoportosítás szempontjai lehetnek a jövedelmi viszonyok vagy a fizetőképesség éppúgy, mint az iskolai végzettség, földrajzi, területi elv vagy más szakmai megfontolások. Statisztikai megfontolásokból a mintán belüli eltérések csökkentése érdekében statisztika eljárások alkalmazásával (pl. klaszterelemzéssel) is csoportosíthatjuk az egyedeket.

Most azt vizsgáljuk, hogy melyik változó milyen szerepet játszik az adott, ismert csoportosításban. Célunk az, hogy a megfigyelt p számú változó olyan lineáris kombinációit állítsuk elő, amelyek a lehető legjobban elkülönítik a g számú osztályba tagolt mintát. Ha ez(ek) a diszkrimináló függvény(ek) nem képes(ek) az előre megadott felosztás teljes reprodukálására, akkor az eljárás megadja a függvény(ek) alapján javasolt csoportosítást.

7.2. A diszkriminancia elemzés alkalmazásának feltételei

A **lineáris** döntési függvényt két előfeltevés mellett keressük:

1. a változók többváltozós normális eloszlást követnek, és
2. minden csoportnak azonos a kovariancia mátrixa.

   Mivel a számítási lépések sorába többváltozós normalitási teszt[106] nincs beépítve, e feltétel teljesüléséről csak „hozzávetőlegesen" győződhetünk meg. A változókra külön-külön grafikus vagy numerikus normalitásvizsgálatot végezve feltárhatjuk azokat a változókat, amelyek eloszlása erősen eltér a normálistól. Ha változó-transzformációval sem tudjuk biztosítani a normális eloszlást, akkor biztosan el kell vetnünk az együttes normális eloszlás feltevését. E mögött az a valószínűségszámítási tétel húzódik meg, hogy a többváltozós normális eloszlás peremeloszlásai biztosan normális eloszlást követnek, de a tétel nem megfordítható.

A csoport kovarianciákat a Box-féle M és ennek F-eloszlású transzformáltja teszteli. Ez a teszt érzékeny a normalitástól való eltérésre, ezért egyenlőtlennek ítélhetünk kicsit eltérő kovariancia mátrixokat akkor, ha a normalitási feltevés nem helytálló. Mivel az M kiszámításában a kovarianciák eltérését a csoportok méretével

[106] Az SPSS-ben nem szerepel olyan statisztikai próba, amellyel a többváltozós normalitás tesztelhető.

súlyozzuk, kis eltérések is szignifikánsnak tűnnek, ha nagy a csoport mérete[107]. Kis méretű csoportokra a lineáris diszkrimináló függvény alkalmazható akkor is, ha a kovariancia mátrixok kissé eltérőek. Ha a kovariancia mátrixok nem egyenlők – de a minta elég nagy – akkor kvadratikus diszkriminancia függvény alkalmazása ajánlható. Ilyen választást az SPSS nem tesz lehetővé.

Ha csak két osztályunk van, azaz dichotom változóval írható le a csoportosítás, akkor a logisztikus regresszió alkalmazása célravezető. E módszernél ugyanis kevesebb előfeltevést kell figyelembe vennünk. Ezt a módszert az 5. fejezet ismerteti.

Vegyes mérési skálájú adatok elemzésére számos nemparametrikus módszer áll rendelkezésre, ilyenkor nem célszerű diszkriminancia elemzést végezni. Problémát okoz az, hogy diszkrét változókra normális eloszlást tételezünk fel, vagy az, hogy ordinális skálán mért változókra kovariancia nem számítható.

Az induló adatok:

Ismerjük p számú változó terében a legalább intervallum szinten mért adatokat, és egy további oszlopban szerepel a csoportosítást megadó nominális változó. A csoportok elemszáma eltérő lehet.

A matematikai háttér:

Az ismert csoportosításból kiindulva a többváltozós szóráselemzés alapgondolatát követjük. Előfeltevéseink:

- A csoportbeli megfigyelések függetlenek és véletlen mintából származnak.
- A független változók többdimenziós normális eloszlást követnek minden csoportban.
- A variancia-kovariancia mátrixok azonosak minden csoportban.

A főátlagtól mért teljes eltérések négyzetösszege két részre bontható: a csoportok közötti és a csoporton belüli eltérések négyzetösszegére[108].

$$T = K + B, \text{ ahol } T = X^T X, \tag{7.1}$$

ha centírozott adataink vannak, azaz X elemei már a főátlagtól való eltéréseket tartalmazzák.

X mátrix (nxp) méretű, ahol a g csoport elemszámai eltérőek lehetnek: $\sum_{i=1}^{g} n_i = n$.

[107] Ha minden csoport elemszáma közel azonosan nagy, akkor ennek nincs torzító hatása. A súly szerepe akkor fontos, ha vegyesen vannak nagyon nagy és nagyon kisméretű csoportjaink.

[108] Ha többváltozós elemzést végzünk, akkor átlagvektorok és eltérés négyzetösszeg mátrixok írhatók fel, méretük (pxp).

A *B* mátrixban az összes megfigyelésre összegezzük a csoportátlagoktól való négyzetes eltéréseket. Alternatív számítása a csoport-kovariancia mátrixok[109] (*S*) súlyozott összege:

$$B = \sum_{i=1}^{g} (n_i - 1) S_i \tag{7.2}$$

A megfigyelt változók lineáris kombinációjaként állítjuk elő a diszkrimináló függvényt, ahol a *c* együtthatók a főkomponens elemzéshez hasonlóan normalizáltak[110]:

$$y = Xc \text{ és } c^T c = 1 \tag{7.3}$$

Különböző c együttható vektorokhoz tehát különböző diszkrimináló függvények tartoznak. Az y vektor értékei nem megfigyeltek, de a centírozás miatt az átlaga zérus, varianciája[111] pedig (7.3) és 7.1) felhasználásával a külső és a belső eltérés négyzetösszeg mátrixokból állítható elő:

$$y^T y = (Xc)^T (Xc) = c^T X^T Xc = c^T Tc = c^T (K + B)c = c^T Kc + c^T Bc \tag{7.4}$$

Most nem egyszerűen az y variancia maximalizálása a célunk. Feladatunk olyan c együttható becslése, amely mellett a csoportok a lehető legjobban különböznek egymástól, és a belső eltérések kicsik, azaz a külső eltérések maximumát és a belső eltérések minimumát egyszerre keressük, a hányadosukat maximalizáljuk:

$$\lambda = \frac{c^T Kc}{c^T Bc} \to \max \tag{7.5.a}$$

Mindkét oldal logaritmusát vesszük, és *c* szerint deriváljuk, a derivált zérus helyét keressük:

$$\ln \lambda = \ln(c^T Kc) - \ln(c^T Bc)$$

$$\frac{\partial \ln \lambda}{\partial c^T} = \frac{2Kc}{c^T Kc} - \frac{2Bc}{c^T Bc} = 0$$

---

[109] A többváltozós variancia-elemzésben a csoportok variancia-kovariancia mátrixának egyezését tételezzük fel. Ezek összege is invertálható, ha egy csoport S mátrixa invertálható. Probléma csak akkor lép fel, ha az elemzésbe bevont változók között nagyon szoros a korreláció.

[110] A gyakorlatban a csoport kovarianciák súlyozott átlagát is figyelembe vesszük: $c^T S_p c = 1$ pótlólagos feltételt alkalmazunk. Ha a változók minden csoportban korrelálatlanok és egységnyi szórásúak, akkor (7.3) szerint számolunk, mert *S*=*E*.

[111] Itt még csak a számlálót írjuk fel, nem osztjuk (n-1)-vel.

Az egyenletet $c^T Kc$-vel végig szorozzuk, és (7.5.a) alapján λ-t behelyettesítjük, *c*-t kiemeljük, így sajátérték-sajátvektor egyenletrendszert kapunk:

$$\begin{aligned} &Kc - \lambda Bc = 0 \\ &(B^{-1}K - \lambda E)c = 0 \end{aligned} \qquad (7.5.b)$$

A megoldást megkapjuk, ha létezik a B^{-1}, azaz a *B* rangja p. A *K* mátrix rangja = min (g-1;p), ezért a szorzatuké sem lehet ennél több. Ha (g-1) kisebb, mint p, akkor (g-1) különböző sajátértéket kapunk. Ha p a kisebb, akkor p számú eltérő sajátérték és hozzátartozó sajátvektor határozható meg. Tehát a diszkrimináló függvények számának felső korlátja a (g–1) és a p közül a kisebb érték.

A j-edik diszkrimináló függvény a λ_j sajátértékhez[112] tartozó sajátvektorral írható fel: $y_j = Xc_j$. Ezeket a sztenderdizálatlan együtthatókat használva a származtatott, (itt használt elnevezéssel) kanonikus térbe képezzük le az eredetileg p dimenzióban megfigyelt pontokat.

A j-edik függvény együtthatóit általában sztenderdizáljuk, azaz szórásával osztjuk. Így a változók hatásának erőssége összehasonlíthatóvá válik. (Hasonló okból számítjuk ki a regressziós modellnél a b mellett a béta együtthatókat is.)

Az egyes diszkrimináló függvények erejét a λ_j sajátértékek fejezik ki. Ha a sajátértékek összegével osztjuk a λ_j-t, akkor az adott függvény szétválasztó erejét százalékban fejezzük ki. Bármely másik *c* együttható vektor kevésbé különíti el a csoportokat, mint a maximális (első) sajátértékhez tartozó c_1.

A diszkrimináló függvények együttes szétválasztó erejét a sajátértékekből (7.6) szerint számított – Wilks lambdának nevezett – Λ mutató méri, amely megegyezik a belső és teljes eltérés négyzetösszeg mátrixok determinánsainak arányával. Mivel a nagy λ_j sajátértékek jelzik az erős diszkrimináló függvényt, a Wilks-lambda kicsi értéke utal szignifikáns függvény(ek)re:

$$\Lambda = \prod_{j=1}^{k} \frac{1}{1+\lambda_j} = \frac{|B|}{|T|} \qquad (7.6)$$

Azt, hogy hány függvény mentén van szignifikáns különbség a csoportok között, szükséges-e mind a k kiszámítható függvény az elkülönítéshez, Bartlett nyomán khi-négyzet próbával teszteljük. Wilks lambdáját (7.7) szerint khi-négyzet eloszlásúvá transzformáljuk. A nullhipotézis szerint a diszkrimináló függvény(ek) hatása nem szignifikáns.

$$\chi^2 = -(n-1-\frac{g+p}{2})\ln\Lambda \qquad (7.7)$$

[112] Itt nem jelent kiválasztási szabályt az, hogy a sajátértékek egynél nagyobbak-e.

a szabadságfoka: (p-r)(g-r-1), ahol r a kihagyott függvények száma.

Az y értékek alapján távolságot számíthatunk egy új, korábban nem osztályozott pont és a csoport átlagok között, hogy az új megfigyelést a hozzá leghasonlóbbakkal egy osztályba soroljuk.

7.3. A diszkriminancia elemzés számítási lépései

A diszkriminancia elemzést előzetesen már csoportokba sorolt adatokra végezzük, mégis a csoportosító eljárások blokkjában található ez az eljárás.

ANALYSE/CLASSIFY/DISCRIMINANT lépéseket követve a nyitó oldalon a következőket találjuk:

Grouping Variable: kategória változó megadása

Define Range: a legkisebb és legnagyobb vizsgálandó kategóriát jelezzük.

Pl. 5 fokú osztályozás esetén min=3 és max=5 kijelölésével csak a közepes vagy annál jobb érdemjegyű diákokat csoportosítjuk.

Independents: azok a változók kerülnek ide, amelyek kombinációja előállíthatja a döntési függvényt.

- Enter: ha minden változót bevonunk a döntési függvénybe
- Stepwise, ha csak a szignifikáns változókat kívánjuk szerepeltetni. (Ha a változók korrelálnak egymással, ezt érdemes választani.)

Statistics gombra kattintva a leíró statisztikák közül választhatunk:

- ❑ Means (a változók átlagai)
- ❑ Anova (egy-egy változó F-tesztje)
- ❑ Box M mutató (a csoportok kovariancia-mátrixainak egyezését méri)

A függvényegyüttható:

- ❑ Fisher félék (közvetlenül az osztályozást segítik), vagy
- ❑ Standardizálatlanok (a döntési függvényeknek az eredeti térben való ábrázolásához és a csoportok középpontjainak meghatározásához használhatók)

A mátrixok között pedig

- ❑ Csoporton belüli korrelációk
- ❑ Csoporton belüli kovarianciák
- ❑ Csoportok közti kovarianciák

- Teljes kovariancia megvizsgálására van lehetőség.

„**Enter independents together**” választása esetén módszert nem választhatunk, a **Method** gomb nem aktív. Ha a változókat lépésenként vonjuk be a döntési függvénybe, amint ezt a következő alfejezet ismerteti, akkor a belépési kritérium kiválasztásával módszert is választunk.

A **Select>>** gomb segítségével egy újabb változó kijelölésével almintát választhatunk ki, és csak erre készül a diszkriminancia elemzés.

Classify gombra kattintva

- a priorok értékéről dönthetünk. Alapértelmezés szerint a csoportok mérete egyenlő, de választhatjuk azt is, hogy a tényleges mintanagyság alapján becsüljük a csoportok valószínűségét.
- Kovariancia mátrix: alapértelmezés szerint a változók kovariancia mátrixait a csoportokon belül számoljuk (Within-groups). A másik lehetőség (Separate-groups) nem a változó, hanem a diszkrimináló függvények kovariancia mátrixait számolja. Ha a függvények száma kisebb, mint a változóké, akkor eltér a két eredmény.
- Display: itt adjuk meg azt, hogy mit kérünk outputként. Az összegző eredmények mellett – ha nem túl nagy a minta –, érdemes esetenként vizsgálni a besorolást. Egy-egy elem kihagyásával (n-1) megfigyelésre elvégezve az osztályozást észrevehetjük az eredményre jelentős befolyást gyakoroló megfigyeléseket.
  - Casewise result
  - Summary Table
  - Leave-one-out-classification
- Plots:
  - Combined groups: egy ábrán mutatja az összes csoport középpontjait és elemeit. (neve: All-groups scatterplot) 1 függvény esetén hisztogramot rajzol.
  - Separate groups: ahány csoport, annyi külön ábra készül. 1 függvény esetén változónként hisztogramot rajzol.
  - Territorial map: a származtatott térbeli térképen szerepelnek a csoportátlagok, a csoportokat jelző számokból képzett „vonalak” pedig elhatárolják a térrészeket egymástól. Csak két vagy több függvény esetén készíthető.

A Save utasítás zárja a sort.

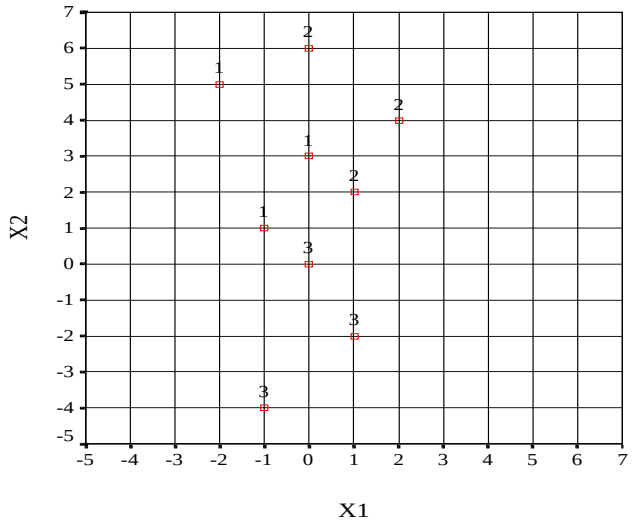
- Predicted group membership választással az új besorolást mentjük el.
- Discriminant scores: Ez adja meg a becsült értéket a döntési függvények terében (ha alacsonyabb dimenzióba jutottunk, akkor ez nagyon hasznos, például ábrázolhatóvá válnak a megfigyelések).
- Probability of group membership: a posteriorokat mutatja.

7.4. Az eredmények részletezése, értelmezése

A grafikus szemléltetést is lehetővé tevő kis példával kezdjük ezt az alfejezetet.

A három csoportba sorolt, csoportonként 3-3 megfigyelésünket kívánjuk két dimenzióban szétválasztani, ezért két diszkrimináló függvényt keresünk.

Induló adataink ábráján (7.1. ábra) látható, hogy a második változó mentén jóval nagyobb az adatok ingadozása (a terjedelem 10 egység), míg az elsőn az átlagok egymáshoz közelibbek (itt 4 egység a terjedelem).



***7.1. ábra**: Három csoport, kilenc pont*

A pontok koordinátái:

Csoport	1		1	1	2	2	2	3	3	3
X_1	**-2**		0	-1	0	**2**	1	1	0	-1
X_2	5		3	1	**6**	4	2	-2	0	**-4**

Az SPSS eredménylistájának rendjét követve haladunk. A 7.1. ábra pontjaira együttesen (Total) készített alapstatisztikákat, valamint a csoportonként és változónként számított átlagokat és szórásokat mutatja a 7.1. táblázat.

7.1. táblázat: Változónkénti átlagok és szórások

Group Statistics

CSOPORT		Mean	Std. Deviation	Valid N (listwise) Unweighted
1	X1	-1,00	1,00	3
	X2	3,00	2,00	3
2	X1	1,00	1,00	3
	X2	4,00	2,00	3
3	X1	,00	1,00	3
	X2	-2,00	2,00	3
Total	X1	,00	1,22	9
	X2	1,67	3,28	9

A csoportátlagok változónkénti egyezésének tesztjét bemutató 7.2. táblázatban Wilks-lambda elnevezés szerepel. Ez nem azonos sem a (7.5)-ben, sem a (7.6)-ban szereplő lambda mértékkel.

7.2. táblázat: Wilks 1. lambda mutatója

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
X1	,500	3,000	2	6	,125
X2	,279	7,750	2	6	,022

Itt az egyes változókra külön-külön számoljuk ki klasszikus, egyváltozós statisztikai értelemben azt, hogy a csoporton belüli eltérések négyzetösszege (SSB) hogyan aránylik a teljes eltérés négyzetösszeghez (SST), az arány SSB/SST=lambda. Az eltérések nagyságát az egyváltozós F-teszttel vizsgáljuk:

$$F(x_i) = \frac{1 - lambda}{lambda} \cdot \frac{n-g}{g-1} = \frac{SSK/(g-1)}{SSB/(n-g)},$$

ahol a számláló szabadságfoka (g-1), a nevezőé pedig (n-g).

Példánkban csak a második változó szerint különböznek szignifikánsan a csoportok[113], az első mentén a csoportátlagok nem különülnek el statisztikai értelemben ($F(x_1)= 3$ és $p_1 =0,125 >0,05$).
A 7.3. táblázatban szereplő egyesített (pooled) kovariancia mátrixot (7.2) szerint szorozva a *B* belső eltérések négyzetösszeg-mátrixát kapjuk, és ez a 7.4. táblázat csoportonként adott kovariancia mátrixaiból kiszámítható. Az egyesített korreláció a csoportonként számított korrelációk elemszámmal súlyozott átlaga. Általában nem egyezik[114] meg a teljes korrelációs mátrix elemeivel, amelyet úgy számítunk, hogy az n elemet egyetlen homogén mintának tekintjük.

7.3. táblázat: A belső kovariancia mátrix elemei

Pooled Within-Groups Matrices[a]

		X1	X2
Covariance	X1	1,000	-,333
	X2	-,333	4,000
Correlation	X1	1,000	-,167
	X2	-,167	1,000

a. The covariance matrix has 6 degrees of freedor

A 7.4. táblázatban látható, hogy az 1. és 2. csoport kovariancia mátrixbeli elemei, azaz a kovariancia-struktúrájuk teljesen megegyező, míg a 3. csoporté eltérő.

7.4. táblázat: A csoportok kovariancia mátrixai és a teljes kovariancia mátrix

Covariance Matrices[a]

CSOPORT		X1	X2
1	X1	1,000	-1,000
	X2	-1,000	4,000
2	X1	1,000	-1,000
	X2	-1,000	4,000
3	X1	1,000	1,000
	X2	1,000	4,000
Total	X1	1,500	,125
	X2	,125	10,750

a. The total covariance matrix has 8 degre

[113] Erre utalt x_2 jóval nagyobb terjedelme is.
[114] Képzeljünk el két változó mentén 3 csoportot úgy, hogy a csoportok elemei kis köröket formáznak, a csoporton belül szinte nincs korreláció. A 3 csoport értékei viszont mindkét változó szerint növekednek, ezért a 3 csoport a 45 fokos egyenes mentén helyezkedik el. Ekkor a teljes mintára számított korreláció egyhez közeli lesz.

A 7.5. táblázatban a szóráselemzés gondolatmenetét követve a csoport kovarianciák azonosságát teszteljük, amihez először a csoport kovariancia mátrixok determinánsának logaritmusát vesszük. Példánkban az első csoportban $|S_1| = \begin{vmatrix} 1 & -1 \\ -1 & 4 \end{vmatrix} = 3$, ebből ln3=1,0986, az egyesített (poolozott) kovarianciára pedig:

$|S_p| = \begin{vmatrix} 1 & -1/3 \\ -1/3 & 4 \end{vmatrix} = 3\frac{8}{9}$, ennek természetes alapú logaritmusa 1,358.

7.5. táblázat: Csoport kovarianciák determinánsainak logaritmusa

Log Determinants

CSOPORT	Rank	Log Determinant
1	2	1,099
2	2	1,099
3	2	1,099
Pooled within-groups	2	1,358

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

A 7.6. táblázatban Box M-mutatóját számítjuk. M kis értéke jelenti a kovariancia mátrixok jó egyezést, ezt F-teszttel ellenőrizzük.

$$M = [\prod_{k=1}^{g} |S_k|^{(n_k-1)/2}] / |\overline{S}|^{(n-g)/2}, \text{ahol}$$

$$\overline{S} = \sum_{k=1}^{g} (n_k - 1) S_k / (n - g) \qquad \textit{és} \qquad n = \sum_{k=1}^{g} n_k$$

$F = -2b \ln M$, ahol b az adott feladatra jellemző szorzószám[115].

[115] A b értéke megtalálható pl Jobson: Applied Multivariate Data Analysis c. könyvének 221. oldalán. A változók és a csoportok száma, az egyes csoportokban található elemek súlyozottan, különböző hatványokon figyelembe véve biztosítják azt, hogy M transzformált értéke F-eloszlást kövessen. Ezek a képletek adják a szabadságfokokat is.

7.6. táblázat: Box-M és F-teszt a csoport kovarianciák egyezésére

Test Results

Box's M		1,557
F	Approx.	,133
	df1	6
	df2	897,231
	Sig.	,992

Tests null hypothesis of equal population covariance matrices.

Mivel az F=0,133 és a szignifikancia szint 0,992, a minta nem mond ellent a nullhipotézisnek, a csoport kovarianciák nem térnek el jelentősen.

A 7.1.-7.6. táblázatokból a diszkriminancia elemzés korrekt végrehajtásához szükséges előkészítő lépéseket és teszteket ismertük meg. Ezek alapján mintafeladatunk alkalmas a diszkrimináló függvény(ek) előállítására.
Először a (7.5.b)-ben szereplő ($\mathbf{B}^{-1}\ \mathbf{K}$) mátrix λ_j sajátértékeit és azok relatív fontosságát kapjuk meg a 7.7. táblázatban. Az első függvényhez tartozik a legnagyobb csoportok közötti változékonyság, ezért szétválasztó ereje mindig magasabb, mint a további függvényeké. Mivel (g-1)=2 és p=2, két sajátérték van, 2 diszkrimináló függvény állítható elő, és az első függvény 76%-át magyarázza a külső eltéréseknek (2,867/(2,867+0,904)=0,76).

7.7. táblázat: A diszkrimináló függvény jellemzői
Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2,867[a]	76,0	76,0	,861
2	,904[a]	24,0	100,0	,689

a. First 2 canonical discriminant functions were used in the analysis.

A 7.7. táblázat utolsó oszlopában a kanonikus korreláció azt méri, hogy milyen szoros az asszociáció a kapott diszkriminancia értékek (mint függő változók) és a csoportok között. Kiszámítása és értelmezése megegyezik az ANOVA-ból ismert eta-négyzet mutató gyökével, ahol eta-négyzet a csoportok közötti és a teljes eltérés négyzetösszegek hányadosa. Itt azt méri, hogy a diszkrimináló „score”-ok változékonyságát milyen arányban magyarázza a csoportbesorolás. Közvetlen

összefüggés áll fenn eta-négyzet és a döntési függvény λ_j sajátértéke között: $\eta_j^2 = \frac{\lambda_j}{1+\lambda_j}$,

példánkban $(0{,}861)^2 = 0{,}74 = 2{,}867/3{,}867$ és $(0{,}689)^2 = 0{,}47 = 0{,}904/1{,}904$.

A 7.8. táblázatban másodszor találkozunk az outputban Wilks lambdával. Ezzel itt a függvények (és nem az eredeti változók) hatását mérjük (7.6) szerint. Lambda (Λ) értéke alacsony, ha a 7.7. táblázatban van nagy sajátérték, ami azonos azzal, hogy a belső eltérések kicsik a teljes eltérésekhez képest. Ha az elhagyott függvények száma, r=0, akkor a min(p, g-1) korlát által meghatározott összes függvényt felhasználjuk a csoportok szétválasztásához.

Az első két függvény által meg nem magyarázott heterogenitás 0,136, mert

$$\Lambda_{2\ függvény} = \frac{1}{1+2{,}867} \cdot \frac{1}{1+0{,}904} = 0{,}136$$

$$\Lambda_{1.\ függvénynélkül} = \frac{1}{1+0{,}904} = 0{,}525$$

7.8. táblázat: Szignifikáns függvények kiválasztása

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,136	10,982	4	,027
2	,525	3,543	1	,060

Példánkban a (7.7) szerint felírt első khi-négyzet értéke magas (valószínűsége kisebb, mint 0,05), arra utal, hogy szükséges k-r=2 függvényt használni a csoportok elkülönítéséhez. Az első diszkrimináló függvény elhagyása után a többi (esetünkben a második) függvény nem szignifikáns részét magyarázza a csoportok közti eltérésnek.

$\chi^2 = -(9-1-\frac{2+3}{2})\ln 0{,}136 = 10{,}98$ szabadságfoka: (2-0)(3-0-1)=4

$\chi^2 = -(9-1-\frac{2+3}{2})\ln 0{,}525 = 3{,}543$ szabadságfoka: (2-1)(3-1-1)=1

A döntési függvény értelmezése szempontjából az egyik legfontosabb eredményt a 7.9. táblázatban találjuk. Mivel a sajátvektorok nagysága függ az eredeti változók szórásától, a teljes mintában mért szórással sztenderdizált változókból (is) számítunk diszkrimináló együtthatókat. Ezeket a regressziós bétához hasonlóan értelmezzük,

ezért mondhatjuk, hogy az első függvényben a második változó hatása erősebb, mint az első változóé, míg a második függvényben fordított a helyzet.

7.9. táblázat: Sztenderdizált diszkriminancia együtthatók

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
X1	,386	,938
X2	,989	-,224

$$y_1 = 0{,}386\left(\frac{x_1}{s_1}\right) + 0{,}989\left(\frac{x_2}{s_2}\right) \text{ és}$$

$$y_2 = 0{,}938\left(\frac{x_1}{s_1}\right) - 0{,}224\left(\frac{x_2}{s_2}\right)$$

Példánkban $s_1 = \sqrt{1{,}5}$ *=1,2247 és* $s_2 = \sqrt{10{,}75} = 3{,}2404$.

A változóknak a diszkrimináló függvényhez való hozzájárulását a sztenderdizált együtthatók mellett korrelációval is kifejezhetjük. A 7.10. táblázat elemei a főkomponens elemzésnél megismert struktúra mátrixhoz hasonlóan a változók és a döntési függvények közötti korrelációs együtthatók.

7.10. táblázat: Változók és függvények korrelációi

Structure Matrix

	Function	
	1	2
X2	,925*	-,380
X1	,221	,975*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functio
Variables ordered by absolute size of correlation within fu

*. Largest absolute correlation between each variable a any discriminant function

A struktúra mátrixból leolvashatjuk, hogy az első függvény mentén növekvő értékek tartoznak azokhoz a megfigyelésekhez, amelyeknek mindkét koordinátája növekszik, és x_2–vel a kapcsolat nagyon szoros. A második tengely mentén elért értéket viszont csökkenti az, ha x_2 magas, de x_1 hatása erős, pozitív.

A 7.11/a. táblázatban szereplő sztenderdizálatlan együtthatókból írjuk fel a döntési függvényt, és a konstans segítségével ábrázolhatjuk is a diszkrimináló függvényeket az eredeti térben.

$$0{,}386x_1 + 0{,}495x_2 - 0{,}824 = 0$$

$$0{,}938x_1 - 0{,}112x_2 + 0{,}187 = 0$$

Az ábrázolás természetesen csak azért lehetséges, mert az eredeti feladat kétdimenziós.

7.11/a. táblázat: Nem sztenderdizált diszkriminancia együtthatók

Canonical Discriminant Function Coeffic

	Function	
	1	2
X1	,386	,938
X2	,495	-,112
(Constant)	-,824	,187

Unstandardized coefficients

A 7.11.a táblázat eredményei különböznek, ha az induláskor sztenderdizáljuk a változókat (7.11/b. táblázat), de 7.11/a és 7.11/b elemei a teljes szórások segítségével egymásból származtathatók. Az első oszlopban például: 0,472= $(0{,}386)(1{,}5)^{1/2}$ és 1,622=$(0{,}495)(10{,}75)^{1/2}$, ahol 1,5 és 10,75 a változók varianciái.

7.11/b. táblázat: Sztenderdizált változókból számolt nem sztenderdizált együtthatók

Canonical Discriminant Function Coefficien

	Function	
	1	2
Zscore(X1)	,472	1,149
Zscore(X2)	1,622	-,367
(Constant)	,000	,000

Unstandardized coefficients

Ha a kanonikus térben ábrázolni kívánjuk megfigyeléseinket, akkor a sztenderdizálatlan sajátvektorokra van szükségünk. A sajátvektorok fontos tulajdonsága, hogy előjelük önkényes. Erre a tényre az értelmezéskor kell különösen figyelni.

A sztenderdizálatlan együtthatókkal számítjuk ki a csoportok centroidjainak (vagy bármely más egyednek) a koordinátáit a származtatott, kanonikus térben (7.12. táblázat).

Példánkban az első csoport átlagpontja (-1,+3), ezt mindkét diszkrimináló függvénybe behelyettesítve kapjuk a centrum új koordinátáit:

0,386(-1) +0,495(3) –0,824 = 0,274

0,938(-1) –0,112(3) +0,187 = -1,087

7.12. táblázat: Csoportközéppontok a kanonikus térben

Functions at Group Centroids

	Function	
CSOPORT	1	2
1	,274	-1,087
2	1,540	,677
3	-1,813	,410

Unstandardized canonical discriminant functions evaluated at group means

A csoportátlagok átlaga zérus a diszkrimináló térben. A tengelyek mentén mért szórás pedig a megfelelő sajátértékek gyöke, ezért az első tengely mentén jobban szóródnak a pontok, mint a függőleges tengely mentén.

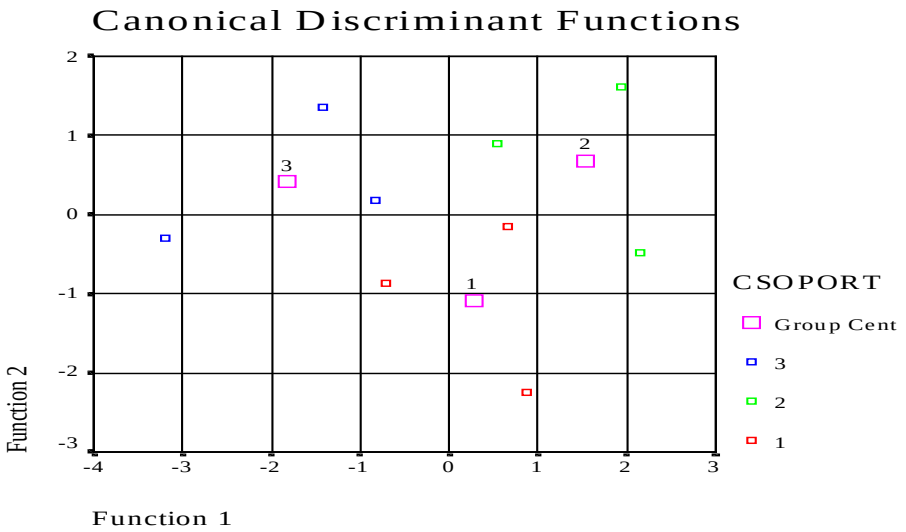
Fontos hangsúlyozni, hogy általában dimenziócsökkentést is végrehajtunk a diszkriminancia elemzéssel ha p>(g-1), mivel az eredeti p dimenziós adathalmazt k (ahol k≤ min(p, g-1)) dimenziós térbe képezzük le. A sajátvektorokkal előállított diszkrimináló tengelyek ortogonálisak.

Egy új megfigyelés csoportba sorolásához kiszámítjuk a diszkrimináló score-okat (y_{ij}) a 7.11/a. táblázat együtthatóiból, és a 7.12. táblázatbeli csoportátlag score-któl (y_{0j}) mért négyzetes euklideszi távolságok legkisebbike határozza meg a besorolást:

$$\min_i \left\{ \sum_{j=1}^{k} (y_{0j} - y_{ij})^2 \right\}, \text{ ahol i=1,…,g.}$$

Az output részeként megkapjuk a kanonikus térbeli ábrát (territorial map), ahol az átlagok körül a csoportok elemei is láthatók. (7.2. ábra)

Mivel kétdimenziós volt az eredeti feladat, a 7.1. és a 7.2. ábra összevetéséből látható, hogy a csoportok más-más sík negyedben vannak, mint az eredeti ábrán, ami a lineáris kombinációban szereplő együtthatók nagyságának és előjelének a következménye.



7.2. ábra: Pontok a kanonikus térben

Az osztályozás jóságának megítélésében több részeredmény segít.

Először a megfigyelések eredeti, a csoportosítással megadott, a priori eloszlását közli a 7.13. táblázat. Mivel a három csoport azonos méretű volt, minden csoport priorja $P(G_i)= n_i /n= 3/9$. A futtatás során a prior a minta empirikus eloszlását követi, vagy a csoportok egyenlő valószínűségét $P(G_i)= (1/g)$ tételezzük fel.

7.13. táblázat: Klasszifikációs statisztika

Prior Probabilities for Groups

CSOPORT	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1	,333	3	3,000
2	,333	3	3,000
3	,333	3	3,000
Total	1,000	9	9,000

A korábban megismert sztenderdizált és sztenderdizálatlan kanonikus együtthatók mellett a Fisher, R.A. által javasolt lineáris diszkrimináló függvények szerepelnek a 7.14. táblázatban. Ezek a csoportonként meghatározott együtthatók alkalmasak arra, hogy közvetlenül az eredeti térben elvégezzük az osztályozást. Abba a csoportra soroljuk a vizsgált egyedet, amelyikre a legnagyobb diszkrimináló érték adódik. Ez

a döntési szabály nem csak a számításokban figyelembe vett pontokra működik, hanem új, eddig nem ismert megfigyelés utólagos osztályozására is alkalmas. A gyakorlatban pl. banki ügyfelek hitelminősítésére használható a lineáris diszkrimináló függvény. Előnye, hogy minden eredeti változót figyelembe vesz, nem redukálja a dimenziót, és nem eredményez nehezen értelmezhető redukált térbeli tengelyeket.

A Fisher-féle *a* együtthatóvektorok számításához a csoport átlagvektorok közötti eltéréseket és a csoportokon belüli kovariancia mátrixokat használjuk. Ezt a függvényt akkor alkalmazhatjuk, ha teljesül a normalitási feltevés. Két csoport esetén: $a = S_p^{-1}\left(\overline{x_1} - \overline{x_2}\right)$

7.14. táblázat: Fisher döntési függvénye

Classification Function Coefficients

	CSOPORT		
	1	2	3
X1	-,771	1,371	-,171
X2	,686	1,114	-,514
(Constant)	-2,513	-4,013	-1,613

Fisher's linear discriminant functions

Ha a harmadik csoportba sorolt (0,0) pontot vesszük, akkor éppen a konstansok adják a Fisher-függvény értékét, és valóban a harmadik csoportban kapjuk a legnagyobb értéket, a (-1,613)-t.

Ha egy új pontot vizsgálunk, amelynek koordinátái (2,3), akkor az 1. csoportra –1,997, a másodikra 2,071, és a harmadikra –3,497 adódik. A függvény alapján a (2,3) pontot a 2. csoportba soroljuk.

A kanonikus függvény és a lineáris diszkrimináló függvény alapján készített osztályozás eredménye megegyezik, ha az összes kanonikus függvényt előállítjuk és felhasználjuk.

A 7.15. táblázat minden megfigyelésre közli az előzetes és a javasolt besorolást, feltételes valószínűséget és posteriort ad. Az eljárás a Bayes-tételen alapul, ahol annak valószínűsége, hogy a D diszkriminancia score-ral rendelkező egyed az i-edik csoportba tartozik:

$$P(G_i|D) = \frac{P(D|G_i) \cdot P(G_i)}{\sum_{i=1}^{g} P(D|G_i) \cdot P(G_i)}$$

Minden egyed abba a csoportba sorolódik át, ahol a legnagyobb a posterior valószínűség.

Van a táblázatban egy „négyzetes Mahalanobis távolság” oszlop is, amely a csoportközépponttól mért négyzetes eltérés a belső kovarianciák kiszűrése után,

valamint olvashatók a kanonikus diszkrimináló függvény(ek) mentén mért score értékek. Ez utóbbiak a származtatott térbeli koordináták, amiket a 7.2. ábrán láttunk.

A 7.15. táblázat alsó fele azt az osztályozást mutatja, ahol az adott egyed kihagyásával (n-1) elemre készült a diszkriminancia függvény. Így két pont besorolásának megváltoztatására tesz javaslatot az eljárás. Az 1. csoport 2. pontjának eredeti koordinátái (0;3), és ez tényleg közelebb van a 2. csoport (1;2) pontjához (d^2=2), mint az 1. csoportbeli másik két ponthoz. Hasonlóan ellenőrizhető a (0;6) pont 1. csoportba való átsorolására tett javaslat.

15. táblázat: Megfigyelésenkénti eredmények

Casewise Statistics

	Case Number	Actual Group	Highest Group					Second Highest Group			Discriminant Scores	
			Predicted Group	P(D>d | G=g)		P(G=g | D=d)	Squared Mahalanob is Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanob is Distance to Centroid	Function 1	Function 2
				p	df							
Original	1	1	1	,424	2	,973	1,714	2	,025	9,000	,877	-2,249
	2	1	1	,598	2	,534	1,029	2	,431	1,457	,659	-,149
	3	1	1	,598	2	,691	1,029	3	,281	2,829	-,715	-,863
	4	2	2	,424	2	,745	1,714	1	,255	3,857	2,143	-,485
	5	2	2	,598	2	,988	1,029	1	,011	10,029	1,925	1,615
	6	2	2	,598	2	,761	1,029	1	,170	4,029	,550	,901
	7	3	3	,598	2	,965	1,029	1	,020	8,829	-1,428	1,348
	8	3	3	,598	2	,668	1,029	1	,272	2,829	-,824	,187
	9	3	3	,301	2	,994	2,400	1	,006	12,600	-3,188	-,304
Cross-validated[a]	1	1	1	,060	2	,929	5,625	2	,071	10,781		
	2	1	2**	,458	2	,574	1,563	1	,342	2,596		
	3	1	1	,273	2	,509	2,596	3	,441	2,885		
	4	2	1**	,194	2	,763	3,281	2	,236	5,625		
	5	2	2	,273	2	,984	2,596	1	,013	11,178		
	6	2	2	,273	2	,543	2,596	1	,370	3,365		
	7	3	3	,273	2	,878	2,596	2	,065	7,788		
	8	3	3	,273	2	,493	2,596	1	,427	2,885		
	9	3	3	,004	2	,985	11,250	1	,015	19,688		

For the original data, squared Mahalanobis distance is based on canonical functions.
For the cross-validated data, squared Mahalanobis distance is based on observations.
**. Misclassified case
a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

Az osztályozás jóságát összefoglalóan a 7.16. táblázat minősíti. Az eredeti és a javasolt besorolás szerint egyező elemek száma és aránya szerepel csoportonként a táblázatban, majd ezek átlagaként az egész osztályozást minősítő egyetlen százalék szerepel a táblázat alatt. A táblázat alsó fele az egy-egy elem kihagyásával készült (cross-validated) osztályozás jóságát mutatja.

7.16. táblázat: Az osztályozás eredménye

Classification Results[b,c]

			Predicted Group Membership			
		CSOPORT	1	2	3	Total
Original	Count	1	3	0	0	3
		2	0	3	0	3
		3	0	0	3	3
	%	1	100,0	,0	,0	100,0
		2	,0	100,0	,0	100,0
		3	,0	,0	100,0	100,0
Cross-validated[a]	Count	1	2	1	0	3
		2	1	2	0	3
		3	0	0	3	3
	%	1	66,7	33,3	,0	100,0
		2	33,3	66,7	,0	100,0
		3	,0	,0	100,0	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 100,0% of original grouped cases correctly classified.

c. 77,8% of cross-validated grouped cases correctly classified.

Eddig csak azzal foglalkoztunk, hogy az összes megfigyelt változó egyidejű bevonásával készítsünk döntési függvényt. Az elemzések során gyakran előfordul az, hogy több változót tartunk érdemesnek arra, hogy a diszkrimináló függvényben szerepeljen, mint ahánynak szignifikáns szerepe van a csoportok elválasztásában. A többváltozós regresszió-számításhoz hasonlóan itt is a lépésenkénti változó bevonás elvét követhetjük, ha a Stepwise módszert választjuk.

7.5. A változók lépésenkénti bevonásával végzett diszkriminancia elemzés

Az SPSS 5 kritériumot kínál fel, ha a változókat lépésenként (stepwise) kívánjuk bevonni a diszkrimináló függvény előállításába. Ezek a kritériumok nem rangsorolhatók, nincsen közöttük egy, amelyik minden adathalmaz esetén megadja a legjobb szétválasztó függvényt. Mind az öt eljárás abból indul ki, hogy először azt a változót kell bevonni, amelyik mentén a csoportátlagok a leginkább különböznek.

Ezt követően lépésenként egy további változó bevonására vagy elhagyására kerül sor, amelyek kiválasztása az alábbi elvek szerint történik.

1. **Wilks lambda elve**: A (7.6) szerint a változókra kiszámított lambda és transzformáltja, (1-lambda)/lambda alkalmas arra is, hogy egy további változó bevonása utáni változás jelentőségét mérje. Mivel a kis lambda és a nagy F érték arra utal, hogy a változó mentén jelentősen különböznek az átlagok, most a p változós modell után a (p+1) változós döntési függvény diszkrimináló erejét mérjük:

$$F_{change} = \frac{n-g-p}{g-1} \cdot \frac{1-\lambda_{p+1}/\lambda_p}{\lambda_{p+1}/\lambda_p}$$

Ha F nagy (a szignifikancia szintje <0,05), akkor a bővítést érdemes végrehajtani, mert a belső, nem magyarázott eltérések jelentősen csökkennek az új változó bevonásával. A modellben szereplő változót kihagyjuk, ha az adott lépésben az F a kihagyási küszöb alá esik. A szelekció szabályozható, mert alapértelmezés szerint az F belépési és kihagyási küszöbértéke rögzített[116]. Ettől eltérhetünk, és választhatjuk bevonási szignifikancia szintnek a 0,05-t, kihagyási küszöbnek pedig a 0,10-t.

A Mahalanobis-féle általánosított távolság központi szerepet játszik a további négy kritériumban.

2. A **Mahalanobis távolságot** maximalizáló változót vonjuk be minden lépésben a döntési függvénybe. Azt a változót keressük, amely mentén a két legközelebbi csoport (A és B) középpontjának távolsága a legnagyobb:

$D_{AB}^2 = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}(\bar{x}_{iA} - \bar{x}_{iB})(\bar{x}_{jA} - \bar{x}_{jB})$, ahol a képletben szereplő w a csoportokon belüli kovariancia mátrix inverzének megfelelő eleme, p a modellbeli változók száma.

A Mahalanobis távolság, mint változó szelekciós kritérium alkalmazása a következő lépéseket jelenti:

a) Mind a g(g-1)/2 csoport-párra p-dimenzióban Mahalanobis távolságot számolunk.

b) Kiválasztjuk a két legközelebbi csoportot[117], azaz a minimális D^2 értéket.

---

[116] Az F-eloszlás kritikus értékét a számláló (g-1) és a nevező (n-g) szabadsági foka is meghatározza, ezért a táblázatban több helyen található 5%- mellett 3,8 körüli érték, pl. (g-1)=4 és (n-g)=8, vagy g-1=2 és n-g=13. Nagyobb megfigyelésszám mellett csökken a kritikus F-érték.

[117] Két csoport esetében ez a lépés kimarad.

c) A D^2–ben szereplő összeadandó négyzetösszegek (i=j) közül kiválasztjuk a maximálisat. Ez lesz a következő lépésben bevonandó változó indexe.

3. Ha a **legkisebb F arány elv** alapján választjuk ki a döntési függvény következő változóját, akkor a Mahalanobis távolságot a csoportok elemszámával súlyozzuk:

$$F = \frac{(n-1-p)n_A n_B}{p(n-2)(n_A+n_B)} D^2_{AB}$$

Az a változó kerül bevonásra, amelyik a legnagyobb - csoportok közti - F értéket adja. Mivel itt az A és B csoport méretét[118] is figyelembe vesszük, a 2. és a 3. kritérium alapján eltérő változót vonhatunk be egy adott lépésben a diszkrimináló függvénybe.

4. A **Rao-féle V mutató**[119] is a Mahalanobis távolságból indul ki, de itt egy-egy csoport átlagát viszonyítjuk a főátlaghoz minden egyes modellbeli változó mentén. Minél inkább eltérnek csoportátlagok és a főátlag, annál nagyobb Rao V-je.

$$V = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij} \sum_{k=1}^{g} (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j)$$

A maximális V-t kiválasztva azonosítjuk a legerősebben megkülönböztető változót. Mivel Rao V-mutatója közelítőleg p(g-1) szabadságfokú khi-négyzet eloszlást követ, egy változó bevonása után a V változása is khi-négyzet eloszlású. Így tesztelhetjük, hogy a modell bővítése szignifikáns változást okozott-e. Egy változó bevonása révén csökkenhet is Rao V-je. Ezt megakadályozandó megadhatunk egy minimális V-t (VIN), aminek az alapértéke 0.

5. A **meg nem magyarázott variancia összege** (Sum of unexplained variance, minimális variancia), mint szelekciós elv közvetlen kapcsolatban áll a Mahalanobis távolsággal.

---

[118] Az $(n_A n_B)/(n_A + n_B)$ maximumát akkor veszi fel, ha $n_A = n_B$. A súlyozás miatt más (AB) csoportra kapjuk a legkisebb F értéket, mintha a mérettől függetlenül választjuk ki a legközelebbinek ítélt két csoportot. Az első változó kiválasztásakor p=1, ezért (n-1-1)/(n-2) ki is esik a képletből.

[119] Más néven is említi a szakirodalom: „Lawley-Hotelling trace", azaz L-H nyoma.

Két csoport szétválasztása úgy is felfogható, hogy 0 és 1 értékkel kódolt dummy változóra, mint függő változóra illesztett többváltozós regresszió. A meg nem magyarázott varianciát minimalizáló változót keressük, amit a többváltozós regressszós modellben ($1-R^2$) mér.

Belátható, hogy a Mahalanobis távolság és a determinációs együttható arányos egymással, $R^2 = cD^2$, ahol c konstans.

7.6. Példa a szelekciós kritériumok alkalmazására

Válasszuk ki a Kényszerértékesítés.sav adatállományt, amely 5 negyedévre (2011. IV. és 2012. I.-IV. negyedév között) Budapest és a megyék bontásában részletezi az adatokat. Keressük meg azokat a diszkrimináló függvényeket, amelyek a negyedévek mentén a lehető legjobban elkülönítik a megyéket. (Itt most minden csoportban, azaz negyedévente azonos számú megfigyelésünk van, de az azonos csoportméret nem elvárás a diszkriminancia elemzés alkalmazása során.)

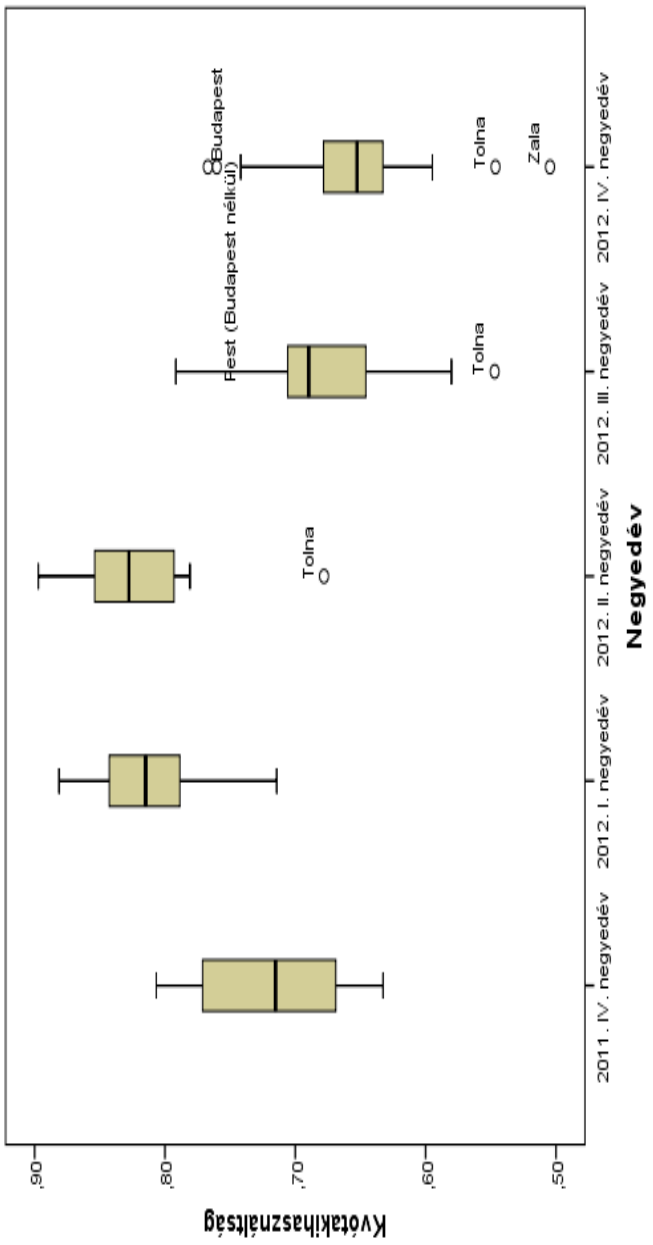
A futtatás beállítása:

- Csoportosító változó: negyedev (1;5)
- Független változók: x1: Kvóta alapja (db), x2: Kvóta alapján kijelölhető maximum (db), x3: Kényszerértékesítésre kijelölt (db), x4: Kvótakihasználtság (%)
- Stepwise módszer, az 5 elv egymás utáni alkalmazása

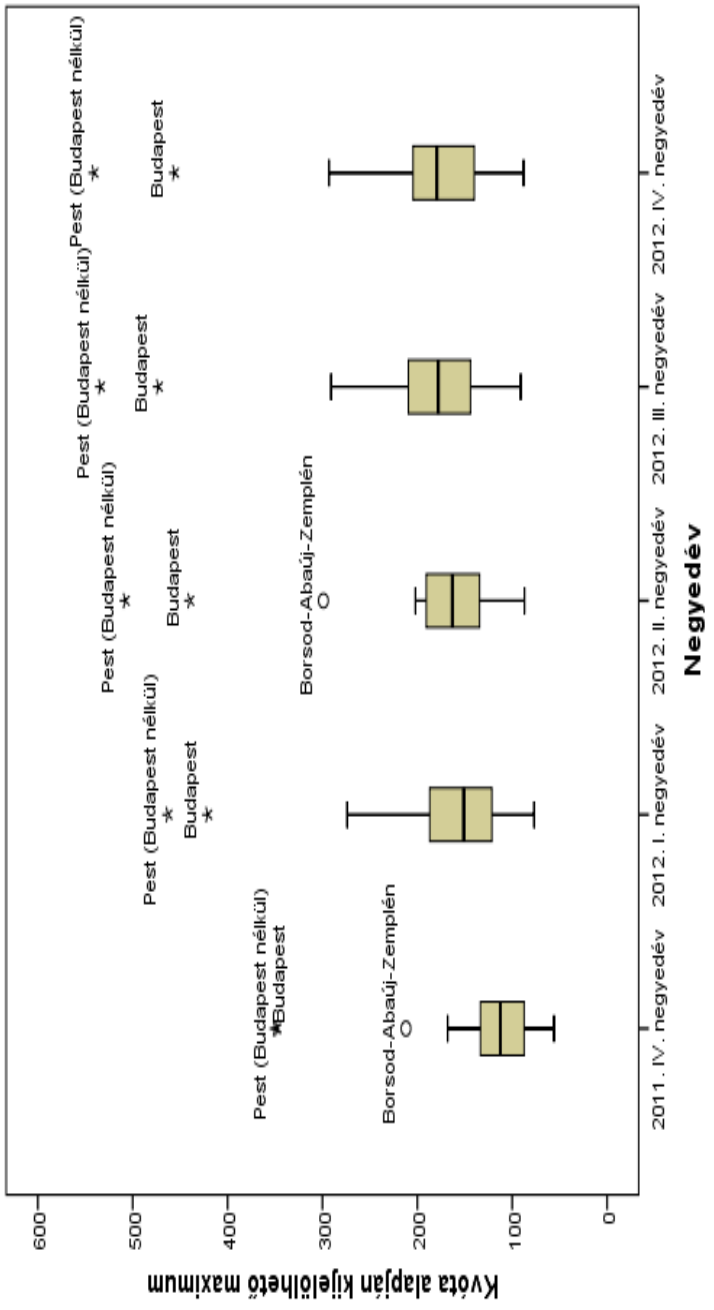
Az eredmények részletezése előtt tekintsük át a leíró statisztikák közül az Explore–ban előállított Boxplot ábrákat két változóra.

A 7.3. ábrán látható, hogy a kezdeti időszaknál jóval magasabb volt 2012. első felében a kihasználtság, míg az év második felében alacsonyabb százalékok jellemzőek. Az eltérések miatt ez a változó megkülönböztető erőt mutat.

A 7.4. ábrán a maximális lakásszámok dobozdiagramjai láthatóak. A negyedévek eltérése csekély, ezért ez a változó várhatóan nem kerül bevonásra, nem fog szerepelni a diszkrimináló függvényben.



7.3.ábra: A kvótakihasználtság alakulása az öt negyedévben



7.4. ábra: A kvóta alapján kijelölhető maximumok az öt negyedévben

A változók egyedi megkülönböztető szerepéről a 7.17. táblázat statisztikái alapján döntünk. A kvóta kihasználtság változóra az átlagok egyezését elvetjük az F-próba alapján. (p=0,000).

7.17. táblázat: Csoportátlagok egyezésének tesztjei 5 negyedévre

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Kvóta alapja	,992	,180	4	95	,948
Kvóta alapján kijelölhető maximum	,942	1,461	4	95	,220
Kényszerértékesítésre kijelölt	,942	1,457	4	95	,221
Kvótakihasználtság	,388	37,478	4	95	,000

Ezen a ponton számos elemzői kérdés fogalmazódik meg.

- Mivel öt csoportunk és 4 változónk van, a (g-1)=4 lesz a döntési függvény számát meghatározó felső korlát.
- Mely változókat és milyen súllyal vonjuk be a diszkriminálásba?
- Ténylegesen hány döntési függvény képezhető?
- Milyen sikeres lesz a negyedévek elkülönítése?

A lépésenkénti beválogatás több szelekciós elv szerint készíthető el. Az első sikeres, a feltételeknek eleget tevő és statisztikailag jól értelmezhető megoldás megtalálása azonban több előkészítő lépést igényel. A lépések megadása mellett kitérünk arra, hogy milyen feltételek nem teljesülése tette szükségessé az újabb lépéseket. (Ez természetesen nem jelenti azt, hogy mindig ilyen – és ilyen sorrendben végrehajtott - korrekciókra van szükség.)

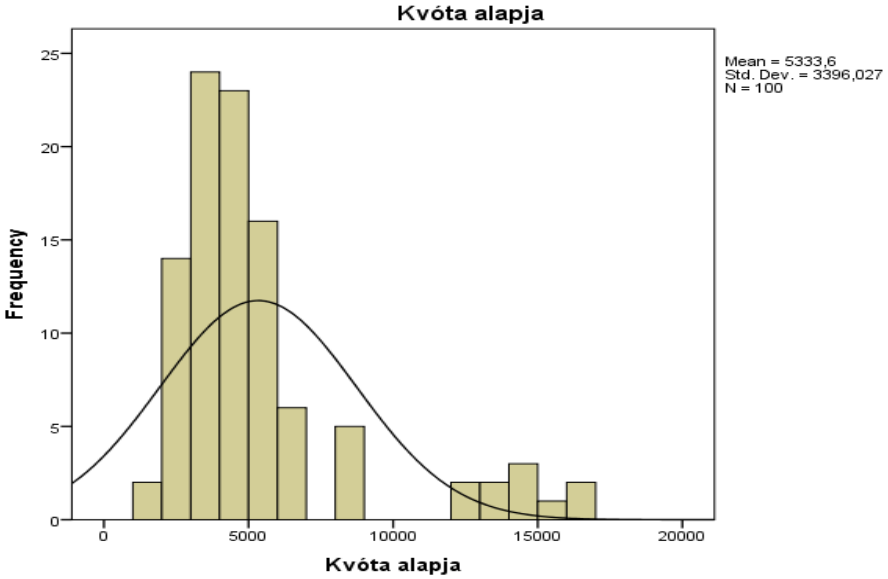
1) Az eredeti változókat és öt negyedévet használva keressük a diszkrimináló függvényt. Ekkor a magas M érték és az alacsony szignifikancia szint (0,000) miatt a kovariancia mátrixok egyezésének hipotézisét el kell vetnünk.

Test Results

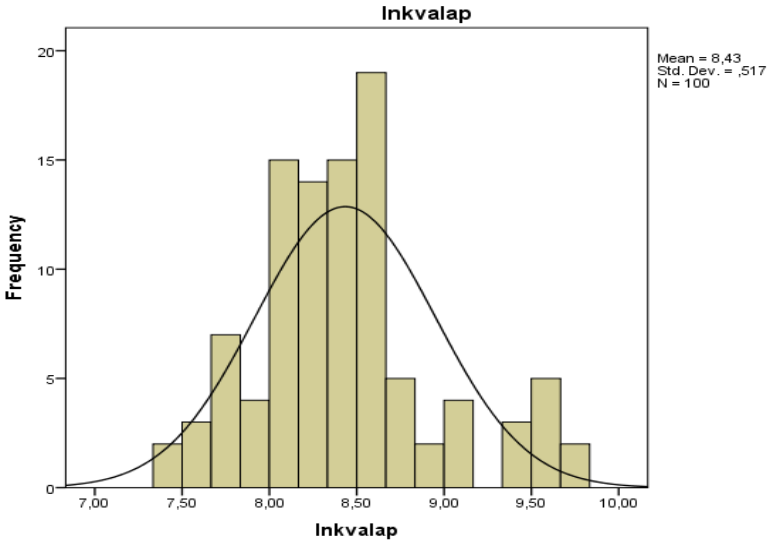
Box's M		324,382
F	Approx.	7,359
	df1	40
	df2	19908,088
	Sig.	,000

.

2) Az első három eredeti változó logaritmusát és a kihasználtsági rátát, valamint öt negyedévet használva keressük a diszkrimináló függvényt. A változók logaritmusát véve a pozitív ferdeségű változók jobban közelítik a normális eloszlást. Példaként a pozitív ferdeségű kvóta alapot és természetes alapú logaritmált értékeit mutatjuk be a 7.5/a és 7.5/b. ábrán.



7.5/a. ábra: A kvóta alapja változó gyakorisága



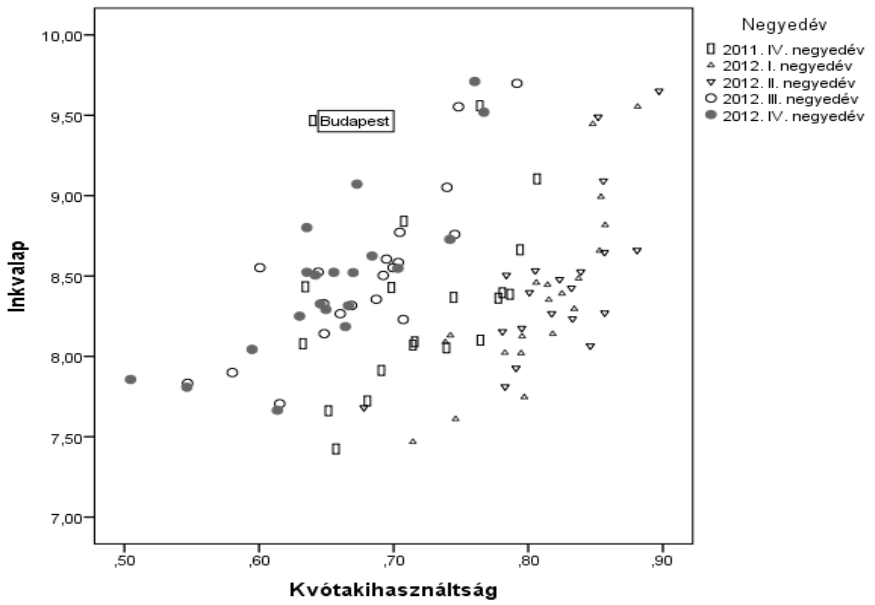
7.5/b. ábra: A kvóta alapja változólogaritmusának gyakorisági ábrája

Ezzel a változókörrel a kovariancia mátrixok eltérése kisebb mértékű, a Box-féle M értéke hatodára csökkent, de még elvetjük az egyezésüket (szignifikancia szint 0,002<0,05).

Test Results

Box's M		53,537
F	Approx.	2,076
	df1	24
	df2	24918,584
	Sig.	,002

3) Az első három eredeti változó logaritmusát és a kihasználtsági rátát, valamint a 2012. év négy negyedévét használva keressük a diszkrimináló függvényt. Ezt a lépést az indokolja, hogy a 2011. év negyedik negyedévére számolt kovariancia mátrix tért el leginkább a többitől, mert Budapest 2011. IV. negyedévei adata a 7.6. ábra szerint távol van a többi ponttól.



7.6. ábra: Öt negyedév adatai két változó terében

A 2011. IV. negyedévi adatok nélkül az F teszt 0,819 értéke és a hozzá tartozó 0,598-as szignifikancia szint alapján a kovariancia mátrixok egyezésének hipotézise nem vethető el.

Test Results

Box's M		7,743
F	Approx.	,819
	df1	9
	df2	66191,846
	Sig.	,598

A Wilks lambda érték alapján 2 változó került be a diszkrimináló függvénybe. Az eredmények bemutatása és értelmezése az SPSS-ben közölt sorrendet követi.

A 7.18. táblázatban a 2012. évi négy negyedéves csoportosítás mellett látható a változókra külön-külön számolt átlagok F tesztje. A csoportátlagok egyezését csak a kvótakihasználtság változó esetében vethetjük el.

7.18. táblázat: Csoportátlagok egyezésének tesztjei 4 negyedévre

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Kvótakihasználtság	,335	50,215	3	76	**,000**
lnkvalap	,987	,330	3	76	,803
lnkvmax	,986	,365	3	76	,778
lnkenyszer	,982	,459	3	76	,711

A kovariancia mátrixok egyezésének F tesztjét ellenőrizve és a nullhipotézist elfogadva a diszkrimináló függvénybe bevont változókat adja meg a 7.19. táblázat.

7.19. táblázat: A Wilks lambda elv alapján beválasztott két változó

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	Kvótakihasználtság	,335	1	3	76,000	50,215	3	76,000	,000
2	lnkvalap	,146	2	3	76,000	40,383	6	150,000	,000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

Ebből a két változóból képezhető két diszkrimináló függvény, amelyek közül az elsőnek nagyon magas (0,924) a kanonikus korrelációja, azaz a negyedévek által alkotott csoportok és a döntési függvény mentén felvett értékek között erős asszociációs kapcsolat van a 7.20. táblázat alapján.

7.20. táblázat: A két függvény és a 4 negyedév közötti kanonikus korreláció

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	5,808[a]	99,9	99,9	,924
2	,005[a]	,1	100,0	,068

a. First 2 canonical discriminant functions were used in the analysis.

A két diszkrimináló függvény **együttesen** szignifikánsan (khi-négyzet teszt szignifikancia szintje=0,000) megkülönbözteti a négy negyedévre megfigyelt adatokat, de a második függvény önmagában nem szignifikáns (szig=0,839) részét magyarázza a csoportok közötti eltéréseknek.

7.21. táblázat: Függvények szignifikáns szerepének tesztelése

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,146	146,130	6	,000
2	,995	,351	2	,839

A függvények számát megismerve a tartalmát is megkapjuk, ha a 7.22., 7.23. és 7.24. táblázatokat áttekintjük.

A 7.22. táblázat Struktúra mátrix nevet viseli, mert az összes változó és a két függvény közötti korrelációs együtthatókat tartalmazza. Az első függvénnyel pozitívan korrelál a kvóta kihasználása, míg a második függvényt döntően a kvótalap logaritmusa határozza meg. A lépésenkénti kiválasztás nem engedi a nem szignifikáns, a bevont változókkal is korreláló változók (a kényszer és a maximum) szerepeltetését a döntési függvényben.

7.22. táblázat: A változók és a függvények közötti korrelációs együtthatók

Structure Matrix

	Function	
	1	2
lnkvalap	-,038	,999*
lnkenyszer[b]	,048	,997*
lnkvmax[b]	-,053	,996*
Kvótakihasználtság	,584	,812*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.
*. Largest absolute correlation between each variable and any discriminant function
b. This variable not used in the analysis.

A diszkrimináló függvényt két alakban: sztenderdizált és sztenderdizálatlan együtthatókkal felírva is megkapjuk. A regressziós bétákhoz hasonló tartalmú a sztenderdizált együttható (7.23. táblázat) azt jelzi, hogy az első függvény mentén növekvő értékű koordináták tartoznak a magas kvótakihasználtsághoz és az alacsonyabb kvótaalaphoz. A második függvény pedig a magas kvótaalapra ad magas koordinátát.

7.23. táblázat: A sztenderdizált együtthatók értékei

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Kvótakihasználtság	1,627	,062
lnkvalap	-1,322	,950

A 7.24. táblázatban az eredeti változók terében is ábrázolható – sztenderdizálatlan - döntési függvény együtthatói kaptak helyet. Ezekbe a függvényekbe behelyettesítve a negyedéveket jellemző átlagokat kapjuk a 7.25. táblázatban látható „centrum”, azaz átlagpontokat.

7.24. táblázat: A két döntési függvény együtthatói

Canonical Discriminant Function Coefficients

	Function	
	1	2
Kvótakihasználtság	29,786	1,135
lnkvalap	-2,570	1,847
(Constant)	-,309	-16,455

Unstandardized coefficients

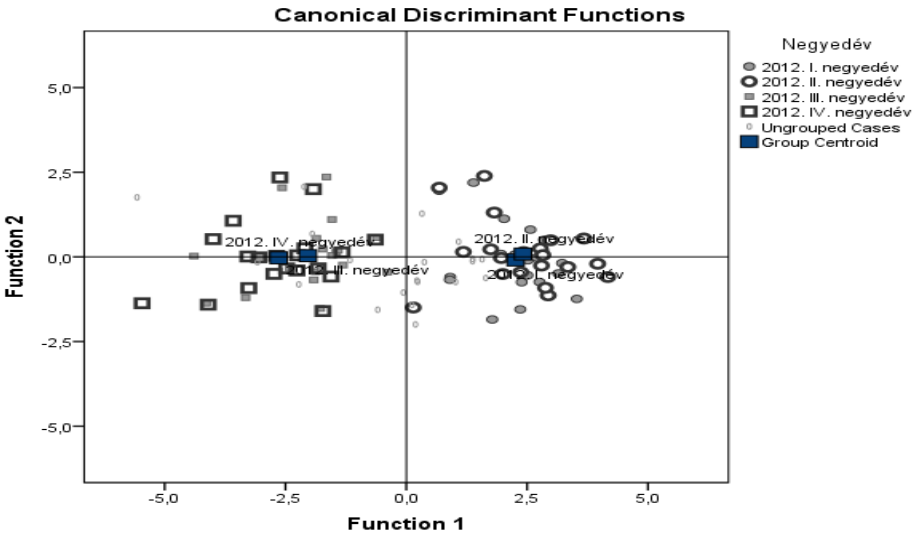
7.25. táblázat: A négy negyedév középpontjai a kanonikus döntési térben

Functions at Group Centroids

Negyedév	Function	
	1	2
2012. I. negyedév	2,263	-,092
2012. II. negyedév	2,414	,085
2012. III. negyedév	-2,035	,035
2012. IV. negyedév	-2,642	-,028

Unstandardized canonical discriminant functions evaluated at group means

A 7.7. ábrán látható, hogy az első diszkrimináló függvény mentén jelentősebb a megyék szóródása, mint a függőleges tengelyen. Azt is leolvashatjuk a 7.25. táblázat és a 7.7. ábra alapján, hogy az átlagpontok nem különülnek el markánsan a négy negyedévre. Ezért érdemes az osztályozó mátrix alapján (7.26. táblázat) az elkülönítés sikerét ellenőrizni, amely nem éri el a 60 százalékot. Az első és a második negyedév, valamint a harmadik és a negyedik negyedév nem különíthető el markánsan, hiszen ezeken belül a nagyobb lakásállománnyal rendelkező főváros és Pest megye másként viselkedik, mint a kisebb megyék.



7.7. ábra: A negyedévek elkülönülése a kétdimenziós kanonikus térben

7.26. táblázat: Az eredeti és a döntési függvény szerinti besorolás osztályozó mátrixa

Classification Results[a,c]

Negyed év		Predicted Group Membership				Total
		2012. I. negyedév	2012. II. negyedév	2012. III. negyedév	2012. IV. negyedév	
Megyék db		**10**	**10**	**0**	**0**	**20**
		8	**12**	**0**	**0**	**20**
		0	**0**	**14**	**6**	**20**
		0	**0**	**9**	**11**	**20**
Százalék		50,0	50,0	,0	,0	100,0
		40,0	60,0	,0	,0	100,0
		,0	,0	70,0	30,0	100,0
		,0	,0	45,0	55,0	100,0

a. 58,8% of original grouped cases correctly classified.

c. 51,3% of cross-validated grouped cases correctly classified.

Az öt lehetséges kritériumot egymás után lefuttatva nem egybehangzó[120] eredményt kapunk. Mind az öt esetben két változó kerül be a függvénybe, de nem ugyanaz a két változó!

Vessük össze a 7.27. táblázatban azt, hogy az egyes lépésekben melyek a kiválasztott változók és mennyire sikeres a döntési függvénnyel az osztályozás.

7.27. táblázat: A szelekciós kritériumok hatása az eredményekre

	Wilks lambda (min)	Minimális Variancia	Mahalanobis távolság(max)	F hányados (max)	Rao - V
1. lépés	kvóta-kihasználtság	kvóta-kihasználtság	kvóta-kihasználtság	kvóta-kihasználtság	kvóta-kihasználtság
2. lépés	lnkvótaalap	lnkvótaalap	*lnkényszer*	*lnkényszer*	lnkvótaalap
azonosan besorolt	58,8%	58,8%	*60%*	*60%*	58,8%

7.7. Egyéni munkára javasolt további feladatok

1) A Kényszerértékesítés.sav adatokra lefuttatva a lépésenkénti diszkriminancia elemzés 5 változatát, mely – további - részeredmények különböznek, melyek egyeznek meg?

   Megoldás:

   2 féle eredmény adódik, melyek a 7.27. táblázat szerint különböznek:

   - Box-M és F teszt
   - függvények együtthatói
   - centrumpontok

2) Készítse el a döntési függvénybe bevont változók terében a pontdiagramot, és szerkessze bele a nem sztenderdizált együtthatókkal a döntési egyeneseket.

[120] Egyes adatállományokra az öt változószelekciós elv azonos eredményt ad. Most tapasztaltunk némi eltérést.

8. Sokdimenziós skálázás

8.1. Az eljárás alapgondolata

A sokdimenziós skálázás (Multidimensional Scaling=MDS) a feltáró módszerek családjába tartozik. Geometriai hátterében az a feltevés áll, hogy a térben minden megfigyelésnek megfelel egy pont, és a hasonlóbb pontok közelebb vannak egymáshoz. Az MDS alkalmazásakor nem fogalmazunk meg sztochasztikus modellt, nem tételezünk fel oksági kapcsolatot, nem állítunk fel tesztelendő hipotézist. A skálázással az adatok között mért különbözőségekből nyerünk információt, származtatunk koordinátákat a skálatérképen. Majd a származtatott koordináták közötti távolságokat összevetjük az eredetileg ismert különbözőségekkel, és törekszünk az eltérések minimalizálására. Az MDS elemzés célja hasonló ahhoz, amit a főkomponens elemzésnél tűzünk ki: az objektumok közötti eltéréseket megőrizve csökkentjük a tér dimenzióját, objektív skálát hozunk létre egy redukált dimenziójú térben.

Az induló adatok

A mátrixok száma és a mérési skála szerint több modell létezik.

- Az (nxp) méretű mátrixba rendezett adatok mérési skálája lehet intervallum szintű, ismerhetjük a kategória gyakoriságokat, és bináris változóval mérhetjük a tulajdonsággal rendelkezést vagy nem rendelkezést. Ekkor az adatok mérési skálájának megfelelő hasonlósági vagy távolság mérőszámot választva hasonlítjuk össze páronként az n számú megfigyelést vagy a p darab változót.
- Az eredeti adatok ismerete nélkül is rendelkezésünkre állhat egy (nxn) vagy egy (pxp) méretű hasonlósági vagy távolságmátrix[121]. A hasonlósági és távolság mérőszámokat részletesen a 3. klaszter-fejezet ismerteti.
- Különböző időpontokban, eltérő körülmények között vagy más személyek, csoportok által mért hasonlóságok, távolságok mátrixaiból is végezhetünk skálázást. Ekkor az egyéni különbségek feltárását végezzük el.

A matematikai háttér

A megfigyelt különbözőségekből MDS térbeli koordinátákat származtatunk, és a koordináták között euklideszi távolságot számítunk. Ismert, hogy n pont közötti eltéréseket (n-1) dimenzióban tökéletesen tudunk ábrázolni. A skálázás célja az,

[121] Ha nem fontos a hasonlóság és a távolság megkülönböztetése, akkor általánosan különbözőségi mátrixot említünk.

hogy alacsonyabb dimenziójú térben jelenítse meg a pontokat, és feltárja a természetes csoportokat, mintabeli struktúrákat[122].

A skálázó módszerek két fő típusát különböztetjük meg.

- Klasszikus (vagy metrikus) skálázásról beszélünk akkor, ha a fő koordinátákat[123] keressük, és az induló különbözőségeket euklideszi távolsággal mérjük. A metrikus modellben lineáris függvénykapcsolat van a különbözőségek (δ) és a skálatérképen mért távolságok (d) között, és a modell intervallum szintű: d=a+bδ vagy arány skálájú, ha a=0 a lineáris függvényben.
- A modell lehet nem-metrikus[124], ha a skálatérképen a távolságok (d) ordinálisan (pl. monoton függvénnyel) kapcsolódnak az eredeti különbözőségekhez (δ). Nem-metrikus modellt célszerű használni, ha az eredeti adatok ordinálisak, pl. rangszámok.

8.2. Koordináták meghatározása klasszikus skálázással

Induljunk ki az alapesetből, $\underline{X}$ mátrix tartalmazza az n pont koordinátáit a p dimenziós térben. A levezetést egyszerűsíti, ha bevezetjük az (nxn)-s méretű[125] $\underline{B}$ mátrixot, amelynek elemei a pontok közti szorzatok:

$$b_{rs} = \sum_{j=1}^{p} x_{rj} x_{sj} \qquad \text{ahol r, s = 1,…,n} \qquad (8.1)$$

A négyzetes euklideszi távolságok $\underline{D}^2$ mátrixának általános eleme felírható (8.1) felhasználásával:

$$d_{rs}^2 = \sum_{j=1}^{p} \left(x_{rj} - x_{sj} \right)^2 = b_{rr} + b_{ss} - 2b_{rs} \qquad (8.2)$$

Miután $\underline{X}$-ből könnyen felírható $\underline{D}$, vizsgáljuk meg a fordított problémát. Tegyük fel, hogy ismerjük a távolságok négyzeteit, de nem ismertek a koordináták. Két lépésben oldjuk meg a feladatot, először $\underline{B}$-t becsüljük, majd $\underline{B}=\underline{X}\underline{X}^T$ szorzattá bontjuk.

---

[122] Hasonló a célja a klaszterelemzésnek is.

[123] A metrikus skálázás atyja Torgerson (1952, 1958). Gower a „principal coordinates analysis" elnevezést javasolta erre a modellre, de rövidítése, a PCA nem különbözik a főkomponens elemzéstől, ezért inkább a metrikus skálázás terjedt el.

[124] Kruskal (1964) dolgozta ki a nem-metrikus eljárást, amit ordinális skálázás néven is említ a szakirodalom.

[125] Az eljárás matematikai lépéseinek ismertetése során az n megfigyelést jelenítjük meg általában p-nél alacsonyabb dimenzióban. A p változó skálázása hasonló lépések alkalmazásával végezhető el.

Ahhoz, hogy egyértelmű megoldást kapjunk, fel kell tételeznünk, hogy a koordináták átlaga 0, azaz $\sum_{r=1}^{n} x_{rj} = 0$ minden j-re. Ez az egyszerűsítés azt eredményezi, hogy a (8.1)-ben megadott b_{rs} sor- és oszlopösszegei is nullák lesznek. Ezt felhasználva, és (8.2)-t összegezve a sorindex, az oszlopindex, majd mindkettő szerint kifejezhetjük b_{rs}–t a távolságmátrix elemeiből az alábbiak szerint:

$$\sum_{r=1}^{n} d_{rs}^2 = tr(B) + nb_{ss} \text{ ,ebből} \qquad b_{ss} = \sum d_{rs}^2 / n - tr(B)/n = d_{r\bullet}^2 - tr(B)/n \tag{8.3}$$

$$\sum_{s=1}^{n} d_{rs}^2 = nb_{rr} + tr(B) \text{ ,és } b_{rr} = \sum d_{rs}^2 / n - tr(B)/n = d_{\bullet s}^2 - tr(B)/n \tag{8.4}$$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = 2ntr(B) \tag{8.5}$$

ahol tr(B) a B mátrix főátlóbeli elemeinek összege, azaz a mátrix nyoma, az indexben szereplő pontok pedig a sor- és oszloptávolságok átlagára utalnak.

Ha (2)-ből kifejezzük b_{rs} –t: $b_{rs} = \frac{1}{2}(b_{rr} + b_{ss} - d_{rs}^2)$

és behelyettesítjük (8.3)-(8.5) átalakított alakjait:

$$b_{rs} = \frac{1}{2}(d_{r\bullet}^2 + d_{\bullet s}^2 - d_{\bullet\bullet}^2 - d_{rs}^2) = \frac{-1}{2}(d_{rs}^2 - d_{r\bullet}^2 - d_{\bullet s}^2 + d_{\bullet\bullet}^2) \tag{8.6}$$

A koordináták származtatásának első lépésében (8.6) szerint kettős centírozást végeztünk. Most a $\underline{B}$ mátrix sajátérték-sajátvektor dekompozíciójával folytatjuk az eljárást.

Ha (8.6)-ban négyzetes euklideszi távolságok vannak, akkor belátható, hogy $\underline{B}$ mátrix szimmetrikus, pozitív definit mátrix, amelynek a rangja k. Így $\underline{B}$-nek van k darab pozitív sajátértéke, melyek nagyság szerint sorba rendezhetőek ($\lambda_1 \geq \lambda_2 \geq ... \lambda_k > 0$). Diagonális mátrixuk jele $\underline{\Lambda}$. A hozzájuk tartozó egységnyi hosszú sajátvektorok ($\underline{v}_1, \ldots \underline{v}_k$) is kiszámíthatók, és (nxk)-s mátrixuk $\underline{V}$. A további (n-k) sajátérték zérus, ezért k dimenziós térben kapjuk meg a megoldást. Tehát B mátrix felbontásával megkapjuk a keresett koordinátákat:

$$\underline{B} = \underline{V\Lambda V}^T = \underline{XX}^T \text{ , ahol } \underline{X} = \underline{V\Lambda}^{1/2}. \tag{8.7}$$

Megjegyzések a klasszikus skálázás eredményeinek értelmezéséhez

- Ha k<p, akkor az eredeti térnél alacsonyabb dimenziójú térben jelenítjük meg a megfigyelt pontokat.
- Mivel a sajátvektorok előjele tetszőleges, a származtatott koordináták értelmezése nem mindig esik egybe az eredeti változók terének irányaival. (Például kétdimenziós térben nem várjuk el, hogy az első sík negyedben legyenek a mindkét tulajdonság szerint „jobb" megfigyelések.)
- A koordináta tengelyek nem is azonosíthatók közvetlenül az eredeti változókkal. Többváltozós regresszió-számítás végezhető annak megállapítására, hogy melyik változó milyen erős hatást gyakorol egy-egy tengelyen mért koordinátákra.
- Ha a $\underline{B}$ mátrix (8.6) szerinti előállításkor nem az euklideszi távolságok négyzeteit ismerjük, akkor $\underline{B}$ nem pozitív szemidefinit, és nem k, hanem n darab sajátértéke lesz, melyek között lesz legalább egy zérus[126], és lehetnek negatívok is. Így nem egyértelmű, hogy hány nagy sajátérték van, és hány dimenzióban kell kiszámítani a koordinátákat. Ilyenkor az javasolható, hogy annyi kis pozitív sajátértéket hagyjunk el, hogy összegük megegyezzen a negatív sajátértékek összegével. Így a megmaradó „nagy" sajátértékek összege egyenlő lesz a mátrix nyomával.
- Bár a klasszikus skálázás robusztus az euklideszi távolságtól való eltérésre, nagy eltérő távolság mértékek használata nem ajánlott. Ilyen esetekre nagy negatív sajátérték, vagy sok közepes méretű pozitív sajátérték figyelmezteti az alkalmazót.

A metrikus skálázás és a főkomponens elemzés eredményei között közvetlen kapcsolat van, ha a korrelációs mátrix felbontását és az egységnyi varianciát eredményező sztenderdizált euklideszi távolságok skálázását vetjük össze. Ha az (nxp)-s $\underline{X}$ mátrix elemei az átlagtól való eltérések, és X rangja k<min(n;p), akkor az $\underline{X}^T\underline{X}$ és az $\underline{X}\underline{X}^T$ szorzatmátrixok sajátértékei megegyeznek, sajátvektoraik viszont különböző elemszámúak. Ha a normalizált sajátvektorokat[127] hasonlítjuk össze, akkor egymásból közvetlenül előállítható eredményeket kapunk. Az i-edik megfigyelésre vonatkozó főkomponensek score-ok ($\underline{X}\underline{a}_i$) négyzetösszege éppúgy λ_i , mint a skálázással kapott koordináták négyzeteinek összege. A (8.8)-ban felírt egyenlőségben a sajátvektorok önkényes előjelétől eltekintünk:

$$\sqrt{\lambda_i}\,\underline{v}_i = \underline{X}\,\underline{a}_i \tag{8.8}$$

Ha az eredmények azonosak, akkor mikor alkalmazzuk a főkomponens elemzést, és mikor a skálázást? Főkomponens elemzést célszerű végezni, ha az induló

[126] Lesz zérus sajátérték, mivel B minden sorában az elemek összege nulla.
[127] A komponensek négyzetösszege =1.

adatmátrixban n>5p, mert ekkor a (p*p) méretű $\underline{X}^T\underline{X}$ dekompozíciója jelent kisebb feladatot.

8.3. Ordinális skálázás

Egyes tudományterületeken, különösen a pszichológiában előfordul az, hogy a különbözőségek számszerű értéke kevésbé fontos, mint a különbözőségek sorrendje. Ilyenkor az eredeti adatok helyett csak a rangszámokat használjuk, és arra törekszünk, hogy az n pont között származtatott távolságok (közelségek, angolul proximities=p*) 2-3 dimenzióban[128] jó egyezést mutassanak a különbözőségekkel. Ez a követelmény nem elégséges ahhoz, hogy egyértelmű megoldást kapjunk, ezért feltesszük, hogy pontjaink az origó körül helyezkednek el, és az origótól mért távolságok négyzetgyöke egységnyi.

A nem-metrikus skálázás iterációval végezhető. Feltételezünk egy kezdeti konfigurációt a p* dimenziós térben, e koordinátákból a pont-párokra származtatott euklideszi távolságot (d_{rs}) számolunk, és ezeket összevetjük a megfigyelt különbözőségekkel (δ_{rs}). Ha a távolságok sorrendje megegyezik a különbözőségek sorrendjével, akkor megfelelő kezdeti konfigurációt találtunk. A tökéletes egyezés ritkán érhető el, csak gyenge monotonitást követelünk meg, azaz a különbözőségek azonosságát nem, csak a távolságok egyezését engedjük meg:

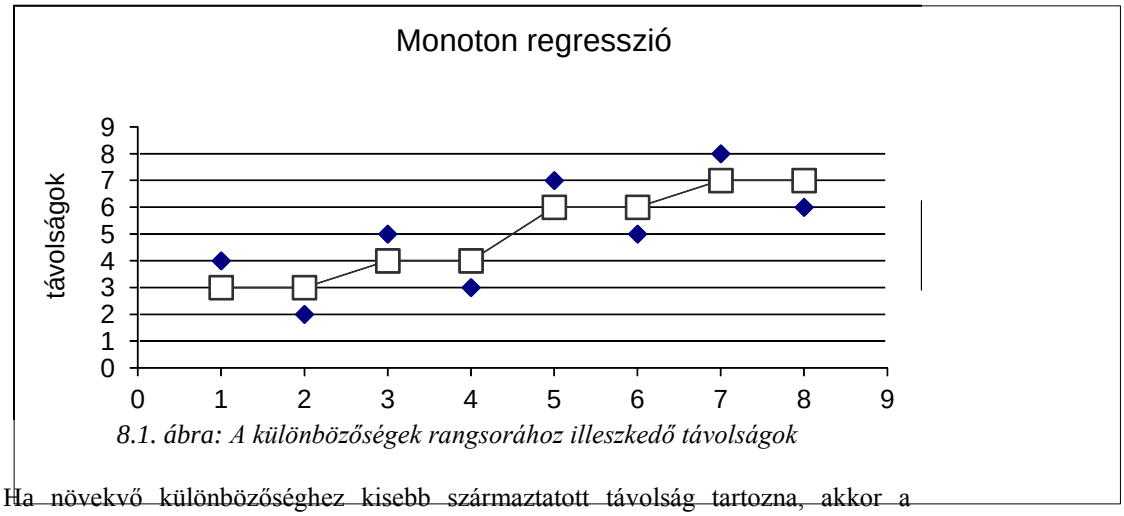
ha $\delta_{rs}<\delta_{tu}$, akkor $\hat{d}_{rs} \leq \hat{d}_{tu}$ álljon fenn.

A d becsült értékét monoton regresszióval állítjuk elő. Ennek során az egymással megegyező különbözőségekre általában nem teszünk külön kikötést, mert az egyező különbözőségekhez egyező távolságok megkövetelése konvergencia problémát okozhat. Monoton regresszió alkalmazását mutatja a 8.1. táblázat és a 8.1. ábra.

8.1. táblázat: A különbözőségek rangsorához illeszkedő távolságok becslése monoton regresszióval

Különbözőség	1	2	3	4	5	6	7	8
Távolság	4	2	5	3	7	5	8	6
Becsült táv.	3	3	4	4	6	6	7	7

[128] Itt a tényleges dimenziószám nem ismert. A keresett dimenziószámot az illeszkedés alapján próbálgatással állapítjuk meg. Egyes szakterületeken, pl. az archeológiában egy dimenziós eredményt, azaz időbeli sorrendet határoznak meg skálázással.



8.1. ábra: A különbözőségek rangsorához illeszkedő távolságok

Ha növekvő különbözőséghez kisebb származtatott távolság tartozna, akkor a monoton regresszió vízszintes görbe lesz, mert a becsléshez a távolságok átlagát vesszük. A gyakorlatban előfordul, hogy csak több lépéssel biztosítható a gyenge monotonitás.

A kezdeti konfiguráció megfelelő, ha az abból számított és a becsült távolságok eltérése kicsi. Az illeszkedésének jóságát a Kruskal által javasolt célfüggvénnyel, a **Stress[129]-függvénnyel** mérjük:

$$S = \left[\sum_{r<s} (d_{rs} - \hat{d}_{rs})^2 / \sum_{r<s} d_{rs}^2 \right]^{1/2} \qquad (8.9)$$

Az S a [0;1] tartományon vesz fel értékeket, és Kruskal véleménye szerint S<0,05 jó illeszkedést, S>0,20 gyenge illeszkedést jelent. Az illeszkedés jóságának megítélésekor ne felejtsük el azt, hogy n és p* is befolyásolja az S értékét. Több pont vagy kevesebb dimenzió esetén nyilván magasabb normalizált reziduális eltérés négyzetösszeg adódik.

A nem-metrikus skálázással elért megoldás általában csak lokális minimumot szolgáltat, és nem mindig konvergál. Több kezdeti konfigurációt[130] érdemes kipróbálni a kiválasztott p* dimenzióban, és a dimenziószám változtatása mellett érdemes figyelni az S változását. Ha a dimenziószám függvényében felrajzoljuk az S alakulását, akkor látjuk, hogy milyen jelentős a Stress csökkenése a magasabb dimenzióban.

---

[129] STRESS= Standardized Residual Sum of Squares

[130] Ilyen kezdeti konfigurációnak választhatjuk a metrikus skálázással kapott koordinátákat is.

Összefoglalva megállapíthatjuk, hogy a metrikus és az ordinális skálázás hasonló eredményre vezet, ha euklideszi távolságokból indulunk ki, de nem euklideszi távolságnál csak a nem-metrikus skálázás alkalmazása javasolható.

8.4. A megvalósítás lépései az SPSS[131]-ben

Az MDS térbeli koordináták kiszámítása és az ábrázolás az

ANALYSE/SCALE/MULTIDIMENSIONAL SCALING lépéseket követve végezhető el.

A nyitó oldalon először azt kell megadni, hogy 1) az input távolságmátrix, vagy 2) az (nxp)-s ***X*** megfigyelési mátrixból számítjuk a távolságot:

1) Data are distances

Ha távolságmátrixból indulunk, akkor a mátrix alakjáról is információt kell adnunk, mert a távolságmátrix lehet

- Négyzetes, szimmetrikus. Ekkor a sorokban és az oszlopokban ugyanazok vannak felsorolva, és különbözőségük az összevetés sorrendjétől függetlenül azonos. Ez a leggyakoribb távolságértelmezés.
- Négyzetes, aszimmetrikus. A sorokban és az oszlopokban most is ugyanazok vannak felsorolva, de különbözőségük mértéke más az alsó és a felső háromszögben (pl. kilométerben és mérföldben is megadjuk két-két város távolságát).
- Háromszög (Rectangular) alakú. Ilyen mátrixunk van, ha az egyik csoport minden eleme azonos távolságra van a másik csoport elemeitől, és a csoporton belüli távolságokról nincs információnk. Formailag az *X* (nxp) adatmátrix is ilyennek tekinthető, mivel n általában nem egyezik meg p-vel.

2) **Create distances from data**

Ebben az esetben a listából kiválasztjuk a változókat.

a) Először arról kell döntenünk, hogy a **megfigyelések** (n darab) vagy a **változók** (p darab) közötti különbséget mérjük, mert az első esetben (nxn), a másodikban (pxp, ahol p>3) lesz a távolságmátrix mérete.

b) A változók **mérési skáláját** is meg kell adni, vegyes skála választása nem lehetséges.

- Intervallum skálán hat távolságmérték[132] választható, alapértelmezés az euklideszi távolság. Választható négyzetes euklideszi, Csebisev, city-blokk, Minkowski vagy „customized” tavolság.

---

[131] Az SPSS későbbi változatai általában kényelmesebbek, több lehetőséget ajánlanak fel. Úgy tapasztaltam, hogy az MDS-ben ez nem sikerült.

- Gyakoriságokra két mérőszámot találunk. A függetlenség feltételezése melletti khi-négyzet és a phi-négyzet számítható.
- Bináris skálán hat mértéket kínál a program. Ezek részhalmazát képezik a klaszterezésnél megismert mértékeknek.

c) **Sztenderdizálhatjuk** az adatokat a változók szerint (alapértelmezés) vagy az egyes eseteken belül hatféle értelemben.

A sztenderdizálással kaphatunk

- 0 várható értékű és 1 szórású z változót,
- (-1,+1) tartományon mozgó értékeket, ha a terjedelemmel osztunk,
- (0,1) között változó értéket, ha a minimumot vonjuk le minden értékből, és a terjedelemmel osztunk,
- egységnyi kiterjedésű relatív értéket, ha a maximális értékkel osztunk,
- egységnyi várható értékű változót, ha az átlaggal osztunk (Ha az átlag zérus, minden megfigyeléshez egyet hozzáadunk.),
- egységnyi szórású változót, ha a szórással osztunk.

A Model menűpont vezet el a modellválasztáshoz, ahol először a modell mérési szintjét adjuk meg.

a) Level of Measurement

- Ordinális szinten mért adatokra a Kruskal-féle **nem-metrikus skálázást** hajtjuk végre monoton transzformációval.
- Intervallum vagy arányskálát választva **metrikus skálázást** végzünk.

b) A skálázó modellek másik lehetséges csoportosítása attól függ, hogy hány mátrixunk van.

- Euklideszi távolság modellt választunk, ha egyetlen mátrixunk van. Ekkor **klasszikus skálázást** (KMDS) hajtunk végre, amely lehet metrikus és nem-metrikus is.
- Ha több - azonos méretű - mátrixunk van, amelyek az egyéni különbségeket[133] írják le, akkor **INDSCAL** eljárást végzünk.

---

[132] A távolságmértékeket a klaszterelemzésnél részletesen tárgyaltuk. Emlékeztetőül: a customized távolság a koordináta eltéréseket p-edik hatványra emeli, majd ezek összegéből r-edik gyököt von. A p és r megfelelő megválasztásával a többi távolságot megkaphatjuk, kivéve a Csebisev mértéket, amely a maximális koordináta-eltéréssel egyenlő.

[133] Az egyéni különbségek eredhetnek abból, hogy különböző időpontokban, különböző feltételek között mérünk valamit, vagy különböző végzettségű emberek véleményét

c) A távolságmátrix egyes elemeinek értelme függhet attól, hogy a mátrix mely részében található. Erről adunk információt, ha a „**Conditionality"** 3 lehetősége közül választunk.

- Matrix: szimmetrikus távolságmátrix, ez az alapértelmezés. Az eltérések azonos mérési skálán kerültek számszerűsítésre.
- Row: a sorokban például különböző szakértőket sorolunk fel, akiknek a szubjektív ítéletei alapján mérjük egyes termékek hasonlóságát, és feltételezzük, hogy a szakértők eltérő skálát használnak. (Aszimmetrikus és háromszög mátrixokra használható.)
- Unconditional: akkor használjuk, ha több azonos méretű mátrixunk van. Így például három-utas faktorelemzést is végrehajthatunk, ha intervallum vagy arány skálán mért adatok távolságát számítjuk.

d) A modellspecifikáció negyedik fontos lépése a dimenziószám meghatározása. Minimum (1 az alapérték) és maximum (6) adható meg. E két értékre és köztük minden egész számra megkapjuk az eredménytáblákat.

Opciók a skálázásban

Az opciók között ábrákat választhatunk, és konvergencia kritériumot állíthatunk be.

a) Ábrák:

- Group plots: egy közös térben ábrázolja a pontokat a kiszámított koordináták alapján. Annyi ábra készül, amennyi a tér dimenziójának mértéke a megadott minimum és maximum között. Egyúttal kapunk egy pontdiagramot is, amely az eredeti távolságok (x tengely) és az MDS térbeli távolságok (y tengely) illeszkedését mutatja.
- Individual subject plot, szimmetrikus távolságmátrixra kérhető.
- Adatmátrix megjelölése esetén az induló és a skálázással kapott távolságmátrixot látjuk kinyomtatva. Ezek illeszkedését mutatja a pontdiagram.
- Modell és összegzés: az eredményt befolyásoló beállításokról ad összefoglalót. Akkor célszerű használni, ha több futtatás készül, és így látjuk, hogy miben különböznek egymástól.

b) Három **kritérium** beállítását változtathatjuk meg. Az a követelmény állítja le az iterációt, amelyik először teljesül.

- S-stress konvergencia: Leáll az iterációs eljárás, ha a célfüggvény (S-stress) változása kisebb, mint 0,001. Kisebb számmal pontosabb megoldást

---

kérdezzük, stb. Az Individual Differences Scaling rövidítéséből ered az eljárás INDSCAL elnevezése.

kapunk, nagyobb érték megadásával rövidebb a számítási idő. Zérus megadásával 30 iterációs lépést hajt végre az SPSS.

- Minimum S-stress: leáll a program, ha (az alapértelmezés szerint) 0,005 alatti S célfüggvény-értéket kapunk. Gyakorlati szabály, hogy kiváló az illeszkedés, ha S kisebb, mint 0,05. Ez vagy egy nagyobb érték kevesebb iterációt igényel. Bármely 0 és 1 közti szám megadható.
- Maximális iteráció szám: 30 az alapérték, de növelhető.

Alapbeállítás szerint a nullánál kisebb távolságokat hiányzó adatként kezeli az SPSS.

8.5. Az eredmények részletezése, értelmezése

Budapest 23 kerületének vizét jellemeztük 4 változó mentén, és euklideszi távolságot számítottunk a sztenderdizált változókra. 2 és 3 dimenziós megoldást is kértünk az összehasonlítás érdekében.

Mivel magasabb dimenzióban tökéletesebb az illeszkedés, mindig a maximális dimenziószámhoz tartozó megoldást kapjuk meg először. Mivel az output nem tagolt, számokkal tördelve, szakaszosan fűzünk megjegyzéseket az eredményekhez.

1) **A háromdimenziós megoldás**

Az iteráció a 3. lépésben leáll, mert a célfüggvény csökkenése kisebb, mint egy ezred.

Iteration history for the 3 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	**Improvement**
1	,04234	
2	,03342	,00892
3	,03308	**,00034**

Iterations stopped because
S-stress improvement is less than ,001000
Stress and squared correlation (RSQ) in distances
RSQ values are the proportion of variance of the scaled data (disparities)
in the partition (row, matrix, or entire data) which
is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.
For matrix
Stress = ,02618 RSQ = ,99630

Az illeszkedés három dimenzióban kiváló, S=0,02618 kisebb, mint 0,05. Az adatok és a távolságok megfelelését mérő R^2 nagyon magas: 0,9963

2) A koordináták
A vetületeket megkapjuk három dimenzióban, de sajnos közvetlenül a „mentés” nem lehetséges.

Configuration derived in 3 dimensions				
Stimulus Coordinates				
		Dimension		
Stimulus Number	Stimulus Name	1	2	3
1		1,2151	1,1148	,2766
2		,3576	,8341	1,2442
3		1,7544	-,3214	-,1385
4		,9237	-1,0688	-,2229
5		,6276	2,9403	,0390
6		1,0598	-,6098	-,7026
7		,6630	-,8249	-,2454
8		-,3203	-1,5557	1,2401
9		-,8091	-,9016	1,2216
10		1,1973	,0522	-1,0767
11		-,4442	1,1190	-,6994
12		,6089	1,2245	,8052
13		,6436	-,1807	,0765
14		,5090	-,3929	,6542
15		,6622	-,6255	-,1427
16		,9285	-,8751	-,1241
17		,5104	-,0429	,0703
18		,0396	-,0856	-,2694
19		-,8464	-,7769	-1,1336
20		-1,4968	1,1220	-,4716
21		-1,9283	,0073	-,5381
22		-2,8726	,1205	,6089
23		-2,9831	-,2731	-,4715

3) Az iteráció lépései

A kétdimenziós iteráció is a harmadik lépésben áll meg.

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	**Improvement**
1	,16331	
2	,14217	,02114
3	,14173	,**00044**

Iterations stopped because
S-stress improvement is less than ,001000
Stress and squared correlation (RSQ) in distances
RSQ values are the proportion of variance of the scaled data (disparities)
in the partition (row, matrix, or entire data) which
is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.
For matrix **Stress = ,12402 RSQ = ,93216**

Az illeszkedés a dimenziócsökkenés miatt romlott, S=0,124 értéke 0,10 és 0,15 közé esik, itt közepes illeszkedésről beszélünk. A távolságok determináltsága 93,2%.

4) A kétdimenziós koordináták

Ezek természetesen nem egyeznek meg a háromdimenziós megoldás első két tengelyére vonatkozó koordinátákkal.

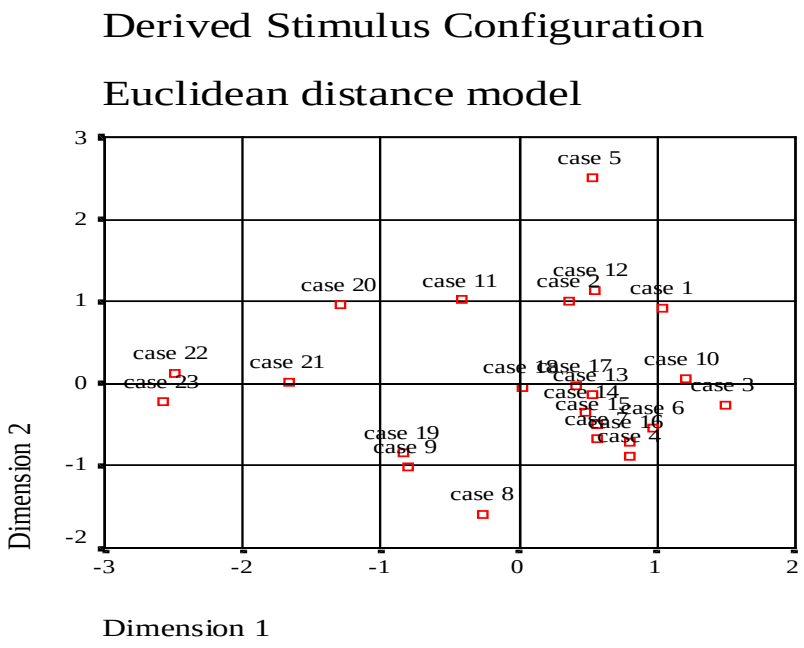
Stimulus Coordinates

	Dimension	1	2
1	VAR1	1,0377	,9246
2	VAR2	,3620	1,0082
3	VAR3	1,4920	-,2629
4	VAR4	,8007	-,8864
5	VAR5	,5252	2,5081
6	VAR6	,9701	-,5448
7	VAR7	,5713	-,6737
8	VAR8	-,2607	-1,5925
9	VAR9	-,8046	-1,0084
10	VAR10	1,2050	,0636
11	VAR11	-,4095	1,0159
12	VAR12	,5454	1,1262
13	VAR13	,5318	-,1242

14	VAR14	,4737	-,3509
15	VAR15	,5610	-,4893
16	VAR16	,7986	-,7019
17	VAR17	,4160	-,0153
18	VAR18	,0301	-,0453
19	VAR19	-,8319	-,8313
20	VAR20	-1,2869	,9575
21	VAR21	-1,6589	,0195
22	VAR22	-2,4946	,1221
23	VAR23	-2,5737	-,2187

5) Csoporttérbeli ábra

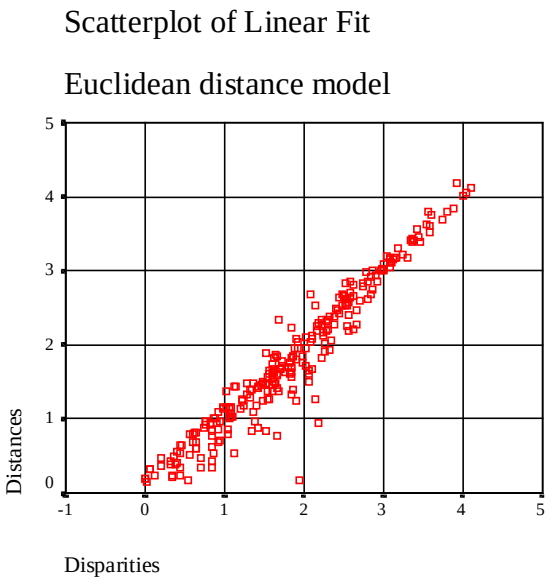
Dimenziónként kapjuk a csoporttérbeli ábrákat. Itt csak a kétdimenziós térképet mutatjuk be. Feliratozást nem lehet választani, a megfigyeléseket mindig sorszámokkal azonosítjuk (8.2. ábra).



8.2. ábra: MDS térkép két dimenzióban

6) Az eredeti és a számított távolságok egyezésének pontdiagramja

Az ábrán is látható, hogy nem tökéletes az illeszkedés, mert eredetileg relatíve távol levő pont-pár (disparitás=2) nagyon közel került a skálatérképen (distance= 0,1). A háromdimenziós megoldás pontdiagramján a távolság-párok szinte tökéletesen a 45 fokos egyenesen fekszenek. (8.3. ábra)



8.3. ábra: Az eredeti és a számított távolságok egyezése

8.6. Az egyéni különbségek skálázása (INDSCAL)

Az MDS alkalmazásának különösen fontos esete az, amikor több időpontra vonatkozó megfigyelésünk van, vagy különböző körülmények[134] között gyűjtöttünk adatokat, vagy több egyén véleményét ismerjük.

Ha az n számú megfigyelést a p változó terében több időpontban mértünk, akkor 3 dimenziós adattömbünk van, amelyben az általános elem x_{ivt} , ahol i=1,…,n a megfigyelések indexe, v=1,…,p a változók azonosítója, és t=1,…,T az időpontokat jelzi. Ha nem az időbeni különbségek a döntőek, hanem a megfigyelés körülményei, vagy az egyéni vélekedések, akkor ezt a k index jelzi az x_{ivk} jelölésben, ahol k=1,…,K.

Most is adódhat olyan feladat, amelyben a megfigyelések, vagy a p számú változó kapcsolatrendszerét, a köztük levő távolság vagy hasonlóság alapján vizsgáljuk,

[134] Fizikai kísérleteknél ilyen pl. a hőmérséklet változtatása, egy kezelés vagy beavatkozás előtt és után való mérés, a biztosításmatematikában a technikai kamatláb különböző mértéke mellett elvégzett számítások.

tehát (nxn) vagy (pxp) méretű különbözőségi mátrixból áll rendelkezésünkre több, amelyeket különböző időpontokban, különböző feltételek teljesülése mellett gyűjtöttünk. Input mátrixunk tehát háromdimenziós. Általános eleme δ_{ijk}, ahol i és j az összehasonlított eseteket vagy változókat, k pedig a mátrix harmadik dimenzióját, az egyént, az időt vagy a körülményt jelöli.

Az időpontok vagy a környezet változása általában befolyásolja a változók vagy megfigyelések kapcsolatrendszerét, és ez a hatás úgy jelenik meg, mintha az egyes időpontokban más és más súlyt rendelnénk a közös MDS térkép koordinátáihoz. Ezt a súlyozott euklideszi modellt nevezzük az egyéni különbségek skálázásának, ahol a különbözőségek stabilitását vizsgálhatjuk úgy, hogy az ismétlődően megfigyelt mátrixokra az egyéni különbségeket feltáró INDSCAL eljárást alkalmazzuk.

A számítások során előállítjuk a közös **dimenziós térben** az MDS koordinátákat, amelyek azt a helyzetet tükrözik, amikor az ismétlődően rendelkezésre álló mátrixok szisztematikusan nem különböznek. Az egyedi y koordináták között közönséges euklideszi távolságot számítunk, és ezen távolságok (monoton vagy lineáris) függvényei az eredeti különbözőségek:

$$\delta_{ijk} = f(d_{ijk}), \quad ahol \quad d_{ijk} = \left(\sum_{s=1}^{r} \left(y_{iks} - y_{jks} \right)^2 \right)^{1/2}$$

A közös tér feltételezésére tett hipotézist ellenőrizzük azzal, hogy az egyes időpontok vagy körülmények között mért adatokban rejlő egyediséget kifejezzük, és mint az MDS tengelyekre vonatkozó súlyokat számszerűsítjük.

Az egyedi terek (y) és a csoport tér (x) között az egyedi súlyok teremtenek kapcsolatot:

$$y_{iks} = \sqrt{w_{ks}} \cdot x_{ik} \quad és \quad y_{jks} = \sqrt{w_{ks}} \cdot x_{jk},$$

ezért a közös térben mért távolság a súlyozott közös koordinátákból is előállítható:

$$d_{ijk} = \left(\sum_{s=1}^{r} w_{ks} \left(x_{is} - x_{js} \right)^2 \right)^{1/2}$$

A w súly tehát a k-adik egyénre (időpontra vagy körülményre) és az MDS koordinátára vonatkozó, 0 és 1 közötti szám. A súly négyzete az s-edik dimenzió fontosságát fejezi ki. A súlyok sor-négyzetösszege determinációs együtthatóként értelmezhető, és a k-adik „egyén” távolságai és különbözőségei közti megfelelés mértékét fejezi ki.

$$\sum_{s=1}^{r} w_{ks}^2 = R_k^2$$

Minden egyén súlyai egy (rxr) méretű diagonális W_k mátrixba rendezhetők.

8.7. Az INDSCAL megvalósítása az SPSS-ben

A futtatás beállítása megegyezik az alapbeállítással, két kiegészítéssel:

- a „Modell” gomb alatt kell jelezni, hogy több azonos méretű mátrixunk van, ezért egyéni különbségeket skálázunk,
- továbbá az „Opciók” részben az ábráknál kérjük az „Individual subject plot” ábrát[135].

Az eredmények áttekintése közben részletezzük az illeszkedés jóságának mutatóit.

A WORLD95 adatokat futtatjuk, 4 változó hasonlóságát tárjuk fel INDSCAL-lal, úgy, hogy a régió változó 6 kategóriáját használjuk.

Változóink: írástudás, városi népesség aránya, férfi és női várható élettartam. A változókat sztenderdizáljuk, euklideszi távolságot számolunk, és 2 dimenziós megoldást kérünk.

Az eredmények részletezése, értelmezése

Az eredményeket a klasszikus MDS-hez hasonló szerkezetben kapjuk, ezért most is tagoljuk.

1) A célfüggvény változása az iteráció során

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
0	,17198	
1	,15957	
2	,15683	,00274
3	,15654	,00029

Iterations stopped because S-stress improvement is less than ,001000

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula 1.

Matrix	Stress	RSQ	Matrix	Stress	RSQ
1	,214	,803	2	,146	,935
3	,044	,988	4	,134	,934
5	,076	,970	6	,234	,699

Averaged (rms) over matrices

[135] Az SPSS az egyedi tereket nem rajzolja le.

Stress = ,15664 RSQ = ,88810

Az illeszkedés jóságára adott korábbi minősítést itt nem alkalmazzuk, mert a közös térben nem várunk el az egyes régióktól jó illeszkedést. Három lépés után már nem javul jelentősen az illeszkedés. Régiónként nézve a 3. térségben kiváló, az 5. térségben jó az illeszkedés.

A végső Stress (0,15664) nem a régiós célfüggvények átlaga, az R-négyzet (0,8881) viszont az egyes csoportok mérőszámainak egyszerű számtani átlaga, tehát a 89% azt jelenti, hogy átlagosan jó az illeszkedés.

2) Koordináták a közös térben

Configuration derived in 2 dimensions
Stimulus Coordinates

		Dimension	
	Stimulus	1	2
1	URBAN	-1,0130	-1,4641
2	LIFEEXPF	,9989	,2403
3	LIFEEXPM	1,0010	-,1126
4	LITERACY	-,9869	1,3364

Ezek alapján készül el a közös térben az ábra, amelyből az egyedi súlyok gyökével szorozva az egyedi terekben a változók ábrázolhatók.

3) Az egyedi súlyok és a „weirdness" (W) index

Subject Weights

Subject	Weirdness	Dim 1	Dim 2
1	,6807	,8741	,1980
2	,9029	,9649	,0652
3	,7561	,2130	,9709
4	,0843	,7641	,5914
5	,7032	,2554	,9510
6	,0709	,6557	,5184
Overall importance of each dimension:		,4699	,4182

Az egyedi súlyok négyzetgyökével szorozzuk a közös koordinátákat az egyes dimenziókban. A számokból látható, hogy a 2. régió (Közép-Kelet Európa) adja az első tengelynek a maximális súlyt, az 5. régió (Közel-Kelet) pedig a legkisebbet. A második tengely fontosságát a 3. és az 5. régió hangsúlyozza magas súllyal.

Az egyes dimenziók általános fontossága megegyezik a dimenzió súlyok négyzetösszegének egy csoportra eső átlagával: $\sum_{k=1}^{6} w_{k1}^2 / 6 \geq \sum_{k=1}^{6} w_{k2}^2 / 6$

A számítások természetéből adódik, hogy az első dimenzió fontosabb (0,4699), mint a második (0,4182).

A dimenzió-súlyok előtt álló W-indexek 0 és 1 között vehetnek fel értéket. Értelmezésükhöz rövid útmutatást is ad az output. A minimumot akkor kapja az „egyén" (esetünkben egy régió), ha a súlyai az átlagos súlyokkal arányosak. Most a 6. régióé a legkisebb index (0,07), ami arra utal, hogy itt szokásos, átlagos a változók kapcsolatrendszere. (A 45^o egyeneshez közel fekszik a súlyt jelző pont.)

A maximumhoz közeli index azt jelzi, hogy az adott régió súlyaránya nagyon szokatlan, az átlagtól erősen eltérő. Egy az index, ha csak egyetlen tengelyre vonatkozik nagy súly, a többi tengelyhez kicsi súlyt rendel az egyén. Példaként a 2. régió említhető.

A súlyok terének értelmezése figyelmet igényel. Itt nem a súlyok közti távolság, hanem az origóból a súlyt jelölő ponthoz húzott vektorok között bezárt szögeket értelmezzük. Ha kicsi a bezárt szög két súly-vektor között, akkor mondhatjuk, hogy a két egyén hasonlóan súlyozza a dimenziókat. A 45^oegyeneshez közeli vektor tipikus, az attól távoli vektor sajátos súlyt jelez.

A W-index kiszámításához a súly-vektort normalizáljuk[136]:

$w_{ks}^n = w_{ks} / \left(\sum_{k=1}^{K} w_{ks} \right)$ és egységnyi hosszú, vele lineárisan összefüggő vektort állítunk elő:

$$v_{ks} = w_{ks}^n / \left(\sum_{s=1}^{r} (w_{ks}^n)^2 \right)^{1/2}$$

Az egyéni súly-vektor és a 45^o egyenes által bezárt szög radiánja kiszámítható, ha figyelembe vesszük, hogy a maximális szög radiánja a dimenziószámból határozható meg: $\cos^{-1}(r^{-1/2})$.

A W-index (WI) képlete: $$WI = (\cos^{-1}\left[r^{-1/2}\right] \left(\sum_{s=1}^{r} v_{ks} \right)) / (\cos^{-1}(r^{-1/2}))$$

4) Az egyedi hatások lineáris mértéke: Flattened Weights

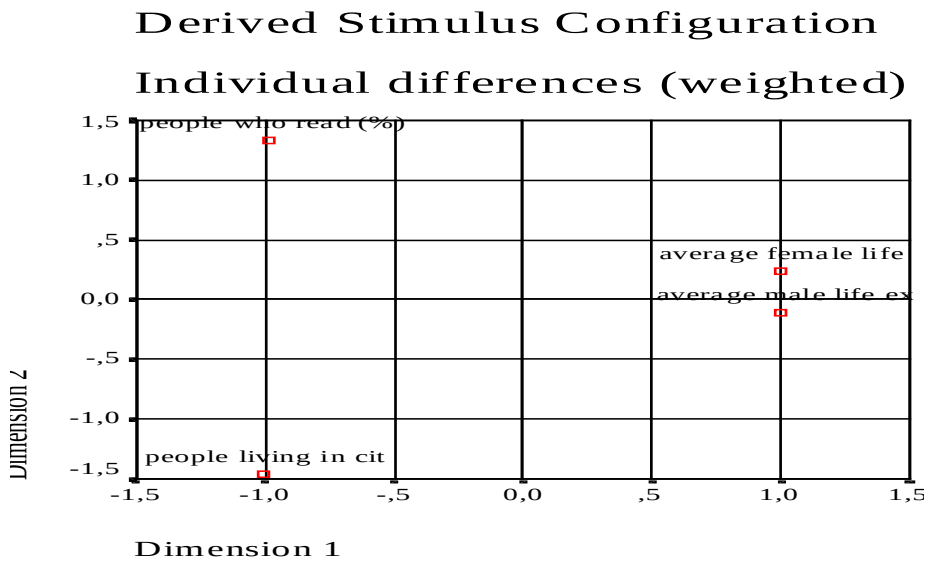
Mivel az egyedi súlyok közötti szögek értelmezhetők, nem a súlyok koordinátái, ezért a szögekből újra pontokat származtatunk, hogy a köztük látható távolságokat értelmezni tudjuk. Ezeket a „lapított" súlyokat (r-1) dimenzióba való vetítéssel kapjuk, és az egyénekre is (r-1) dimenzióban jelennek meg. Az új súlyok lineárisan értelmezhetők, és összegük minden tengelyre zérus. Példánkban a két dimenziós súly-térben mindkét tengellyel 45^o szöget bezáró egyenest húzunk, és erre vetítjük a

[136] A normalizált súlyokat nem kapjuk meg, de az index kiszámításához elvégzi az SPSS a számítást.

régiók súlyait. Az átlagos súlyú régió most nulla-közeli F-súlyt kap, az első tengelyt preferálókhoz nagy pozitív, a második tengelyt kiemelőkhöz pedig nagy negatív súlyt rendel az eljárás.

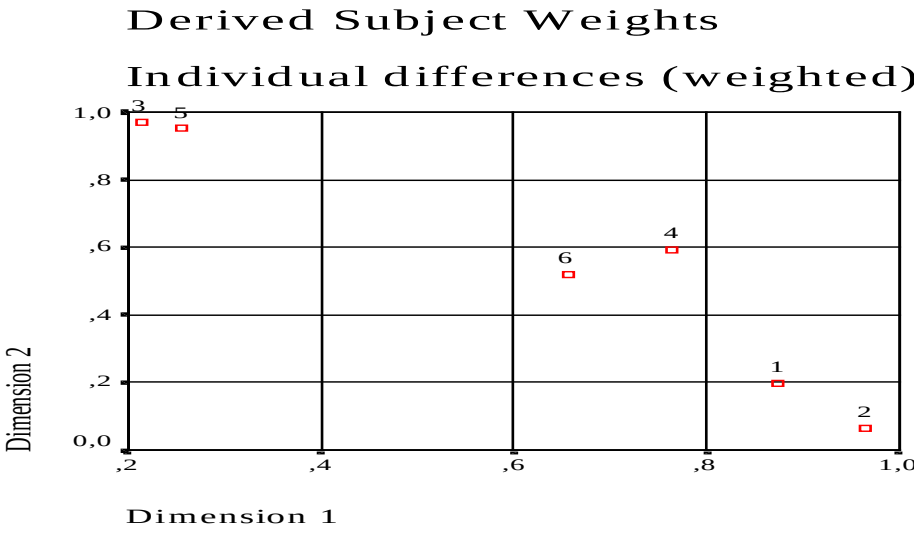
5) Ábrák az INDSCAL-ban

a) Csoport térben láthatók a változók (8.4. ábra) vagy a megfigyelések.



8.4. ábra: A változók közelsége

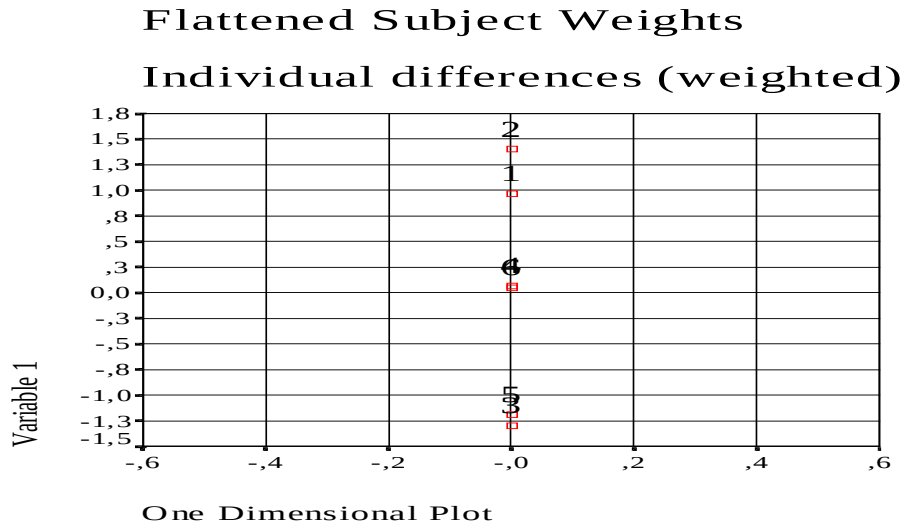
b) A régiók eltérően súlyozzák az egyes tengelyeket (8.5. ábra).



8.5. ábra: A régiók tengely-súlyai

c) A különbözőségek és a távolságok lineáris illeszkedését mutató ábra megegyezik a klasszikus MDS ábrával, ezért külön nem közöljük.

d) A lineáris súlyok ábrája – egy dimenzióban a 8.6. ábrán látható.



8.6. ábra: A tengely-súlyok egy dimenziós vetületei

A 4. és a 6. régió lineáris (Flattened) jelzőszámai az origó közelében egymásra esnek, mert súlyaik nagyon közel kerületek az egy dimenzióba történő vetítés során egymáshoz.

8.8 Önálló elemzési feladatok

A Kényszerértékesítés.sav adattáblázat alkalmas az egyéni különbségek megjelenítésére, akár a negyedévek, akár a területi különbségek szerint bontjuk meg a mintát.

1) Mutassa meg, hogy időben – azaz az öt negyedév szerinti bontásban vizsgálva az egyéni különbséget, eltérő-e a négy változó
    a. x1: Kvóta alapja (db),
    b. x2: Kvóta alapján kijelölhető maximum (db),
    c. x3: Kényszerértékesítésre kijelölt (db),
    d. x4: Kvótakihasználtság (%) közötti kapcsolatrendszer.
2) A regionális különbségek statisztikai jelentőségét is feltárhatja az MDS eljárással, ha a megyék szerint méri a négy változó
    a. x1: Kvóta alapja (db),
    b. x2: Kvóta alapján kijelölhető maximum (db),
    c. x3: Kényszerértékesítésre kijelölt (db),
    d. x4: Kvótakihasználtság (%) terében az egyéni különbségeket.

Források

Carol Alexander (2007): Market Models, A Guide to Financial Data Analysis, John Wiley&Sons, Ltd

Chatfield, C. And Collins, A. J. (2000): Introduction to Multivariate Analysis, Chapman & Hall/CRC, Boca Raton st al., (Reprint, First edition 1980)

Csendes Tibor (2001): Bevezetés a számítógépes statisztikába, Novadat, Szeged

Füstös László –Meszéna György – Simonné Mosolygó Nóra (1997): Térstatisztika, Aula Kiadó, Budapest

Füstös László – Kovács Erzsébet – Meszéna György – Simonné Mosolygó Nóra (2004, 2007): Alakfelismerés. Sokváltozós statisztikai modellezés a társadalomtudományokban ÚjMandátum Kiadó, Budapest

Green, Samuel B. – Salkind, Neil J. – Akey Theresa M. (2000): Using SPSS for WINDOWS. Analyzing and Understanding Data, Prentice Hall International (UK) Ltd, London (Second Edition)

Hajdu Ottó (2003): Többváltozós statisztikai számítások, KSH, Budapest

Horvai György (2001): Sokváltozós adatelemzés (Kemometria), Nemzeti Tankönyvkiadó, Bp.

Hunyadi László (2001): Statisztikai következtetéselmélet közgazdászoknak, KSH, Budapest

Hunyadi László – Mundruczó György – Vita László (1997): Statisztika, AULA Kiadó, Budapest (II. kiadás)

Jobson, J. D. (1992): Applied Multivariate Data Analysis, Volume I & II, Springer-Verlag, New York et al. (Second Edition)

Johnson, Dallas E. (1998): Applied Multivariate Methods for Data Analysts, Duxury Press, Pacific Grow (California)

Ketskeméty László – Izsó Lajos – Könyves Tóth Előd (2011): Bevezetés az IBM SPSS Statistics programrendszerbe, 3. kiadás, Artéria Stódió Kft, Budapest

Krzanowski, W. J. (2000): Principles of Multivariate Analysis. A User's Perspective, Oxford University Press, Oxford (Revised Edition)

Maindonald, J.-Braun, W. J. (2008): Data Analysis and Graphics. Using R- an Example-Based Approach, 2nd Edition, Cambridge Press

Norusis Maria, J. [SPSS Inc.] (1994): SPSS Professional Statistics 6.1., SPSS Inc., Chicago

SPSS Inc. (1998): SPSS Base 8.0. Applications Guide, SPPS Inc., Chicago

Székelyi Mária – Barna Ildikó (2002): Túlélőkészlet az SPSS-hez. Többváltozós elemzési technikákról társadalomkutatók számára, Typotex Kiadó, Bp.