

**Szegedi Tudományegyetem Gazdaságtudományi Kar
Közgazdaságtudományi Doktori Iskola**

**A multikollinearitás vizsgálata lineáris regressziós modellekben
A PETRES-féle *Red*-mutató vizsgálata**

Doktori értekezés

Készítette:

Kovács Péter

Témavezető:

Dr. Petres Tibor
egyetemi docens

Szeged
2008

Tartalomjegyzék

ELŐSZÓ	2
I. AZ ÉRTEKEZÉS PROBLÉMÁJÁNAK DEFINIÁLÁSA, CÉLJAI, HIPOTÉZISEI	3
I. 1. A PROBLÉMA MEGFOGALMAZÁSA	3
I. 2. REGRESSZIÓSZÁMÍTÁS	3
I. 3. A MULTIKOLLINEARITÁS FOGALMA	5
I. 4. A PETRES-FÉLE <i>RED</i> -MUTATÓ	8
I. 5. A DISSZERTÁCIÓ CÉLJA ÉS A KUTATÁSI HIPOTÉZISEK	9
II. A MULTIKOLLINEARITÁS SZAKIRODALMÁNAK ÁTTEKINTÉSE	13
II. 1. A MULTIKOLLINEARITÁS KÖVETKEZMÉNYEI	13
II. 2. A MULTIKOLLINEARITÁS FELISMERÉSE, MÉRÉSE, ILLETVE MÉRŐSZÁMAINAK BÍRÁLATA	15
II. 3. A MULTIKOLLINEARITÁS NEGATÍV KÖVETKEZMÉNYEINEK CSÖKKENTÉSE	33
II. 4. AZ ISMERTETETT MUTATÓK ÉS ELJÁRÁSOK EMPIRIKUS SZEMLÉLTETÉSE	40
III. A PETRES-FÉLE <i>RED</i>-MUTATÓ VIZSGÁLATA	54
III. 1. A <i>RED</i> -MUTATÓ SZÁMSZERŰSÍTÉSE A SAJÁTÉRTÉKEK ISMERETE NÉLKÜL	54
III. 2. A MULTIKOLLINEARITÁS ELLIPTIKUS MODELLJE	58
III. 3. A <i>RED</i> -MUTATÓ ÉS A REGRESSZIÓS PARAMÉTEREK VARIÁNCIÁI	67
III. 4. A <i>RED</i> -MUTATÓ ELOSZLÁSÁNAK VIZSGÁLATA	74
III. 5. A <i>RED</i> -MUTATÓ ÉS A <i>KMO</i> -MUTATÓ KAPCSOLATA	78
III. 6. A MULTIKOLLINEARITÁS EGY MÁSIK MÉRÉSI LEHETŐSÉGE	81
IV. A KUTATÁS EREDMÉNYEINEK ÉRTÉKELÉSE, TOVÁBBI KUTATÁSI IRÁNYOK	86
IV. 1. ÉRTÉKELÉS, TÉZISEK	86
IV. 2. KUTATÁSI IRÁNYOK A JÖVŐRE VONATKOZÓAN	89
V. IRODALOMJEGYZÉK	91
VI. MELLÉKLETEK	96
VI. 1. ÁBRAJEGYZÉK	96
VI. 2. TÁBLAJEGYZÉK	97
VI. 3. PÉLDA A <i>DEF</i> -MUTATÓ HIÁNYOSSÁGÁRA	98
VI. 4. A II. 4. ALFEJEZET EMPIRIKUS PÉLDÁJÁNAK DEFLÁLT ADATAI 2004-ES ÁRSZÍNVONALON, ILLETVE A SZÁMÍTÓGÉPES ELEMZÉS NÉHÁNY KIMENETE	100
VI. 5. A MULTIKOLLINEARITÁS ELLIPTIKUS MODELLJE HÁROM TÉNYEZŐVÁLTOZÓ ESETÉN	112
VI. 6. A TÉNYEZŐVÁLTOZÓK KORRELÁCIÓS MÁTRIXA SAJÁTÉRTÉKEINEK ÉS RECIPROKÖSSZEGÜKNEK A VISELKEDÉSE HÁROM TÉNYEZŐVÁLTOZÓ ESETÉN, RÖGZÍTETT <i>RED</i> ÉRTÉK MELLETT	113
VI. 7. A <i>RED</i> -MUTATÓ EMPIRIKUS ELOSZLÁSFÜGGVÉNYÉNEK MEGHATÁROZÁSÁHOZ HASZNÁLT ALGORITMUS	118
VI. 8. PUBLIKÁCIÓIM, KONFERENCIA-ELŐADÁSAIM	119

Előszó

Jelen disszertáció a Szegedi Tudományegyetem Statisztikai és Demográfiai Tanszékén készült. A Tanszék oktatói-kutatói munkájába először 2000-ben kapcsolódtam be, akkor még demonstrátorként. A Tanszéken folyó szakmai munka jelentősen bővítette látókörömet.

Az értekezés elkészítése során nagyon sok segítséget és támogatást kaptam.

Ezért, elsőként köszönettel tartozom témavezetőmnek, dr. Petres Tibornak a Statisztikai és Demográfiai Tanszék egyetemi docensének. Az ő által vezetett kutatások, „együttgondolkodások” és a közös szakmai munka jelentősen hozzájárult az értekezés témaválasztásához, illetve elkészítéséhez, valamint a szakmai fejlődésemhez.

Köszönet illeti dr. Lukovics Miklóst, aki többször is gondosan átolvasta munkám és hasznos tanácsaival, észrevételeivel sokat segített.

Köszönettel tartozom az elő-opponenseimnek is, dr. Rappai Gábornak és dr. Krámlí Andrásnak, akik egyrészt nagyon jó és hasznos ötleteikkel elősegítették az értekezés végső formátumának elérését, másrészt ötleteik rendkívül gondolatébresztők voltak a későbbi kutatásaimra nézve.

Továbbá hálás vagyok minden kollegámnak, akik ötleteikkel, tanácsaikkal segítettek.

Nem utolsó sorban szeretném megköszönni családomnak a kitartást és a támogatást.

I. Az értekezés problémájának definiálása, céljai, hipotézisei

I. 1. A probléma megfogalmazása

A mai világban nagymértékben növekszik a döntéshozók információigénye. Az adatok mennyiségének nagymértékű növekedése nem jár együtt automatikusan a megfelelő mértékű információnövekedéssel. Igazából a döntéshozóknak ma már nem az adatok hiányával, hanem azok bőségével kell szembenézniük. Éppen ezért, empirikus elemzéseknél *lényeges kérdés a metrikus adatok információtartalma*, mivel a nagyon nagy mennyiségű adat gyakran kevés információt hordoz, azaz nagymértékű a *redundancia*. *Ez utóbbi alatt a vizsgálat szempontjából újabb információt, érdemleges közlést már nem tartalmazó, „felesleges” adatokat értjük* (PETRES – TÓTH, [2006]). Különösen igaz ez a lineáris regressziós modellek alkalmazásakor. Ahhoz, hogy ezekben a modellekben meg tudjuk ragadni a redundancia megjelenési formáját, a következő alfejezetben röviden összefoglalom a regressziószámítás – a továbbiak megértéséhez feltétlenül szükséges – alapismereteit.

I. 2. Regressziószámítás

Többváltozós empirikus elemzéseknél a statisztikai módszerek közül – talán – leggyakrabban a regressziós modell kerül alkalmazásra, melynek legismertebb típusa a standard lineáris regressziós modell. Ez mátrixalgebrai jelöléssel az

$$(1) \quad \tilde{\mathbf{y}} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$$

formában is felírható, ahol

$\tilde{\mathbf{y}}$: az eredményváltozó n komponensű oszlopvektora;

$\tilde{\mathbf{X}}$: a tényezőváltozók n sorból és $(m+1)$ oszlopból álló mátrixa, ahol az első oszlop mindig egy $\tilde{\mathbf{x}}_0$ összegező vektor;

$\tilde{\boldsymbol{\beta}}$: a modell számunkra ismeretlen paramétereinek $(m+1)$ komponensű oszlopvektora;

m : a magyarázóváltozók (tényezőváltozók) száma;

$\tilde{\varepsilon}$: a hibatag n komponensű oszlopvektora.

A modellben szereplő ismeretlen paraméterek n megfigyelésből álló minta alapján történő becslőfüggvénye a legkisebb négyzetek módszere (LNM, a szokásos angol rövidítése OLS) szerint az alábbi.

$$(2) \quad \hat{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

A $\tilde{\beta}$ funkcionális operátor olyan hipersíkot eredményez, amely a legjobban illeszkedik a megfigyelések n -dimenziós „pontfelhőjéhez”.

Az eredményváltozó értékeinek becslését az $\hat{\mathbf{y}} = \tilde{\mathbf{X}}\hat{\beta}$ mátrixegyenlet adja, ahol

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \tilde{x}_{11} & \cdots & \tilde{x}_{1m} \\ 1 & \tilde{x}_{21} & & \tilde{x}_{2m} \\ \vdots & & & \\ 1 & \tilde{x}_{n1} & & \tilde{x}_{nm} \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix}.$$

A standard lineáris regressziós modell egy olyan fekete dobozként képzelhető el, amelynél a bemenetet a magyarázóváltozók alkotják, a kimenet pedig az eredményváltozó.

A modell feltételrendszerének fontos részét alkotják – többek között – az alábbiakban ismertetett feltételek is (PETRES – TÓTH, [2006]):

- a magyarázóváltozók lineárisan függetlenek. A feltétel nem teljesülése esetén az $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ mátrix nem invertálható, így a regressziós paraméterek meghatározása a (2) képlet alapján nem lehetséges.¹
- A hibatag nulla várható értékű, konstans $\tilde{\sigma}^2$ varianciájú, értékei páronként korrelálatlan valószínűségi változók, amelyek együttes eloszlása n -dimenziós normális eloszlás.

A regressziószámítás gyakorlati alkalmazásakor ügyelnünk kell arra, hogy a standard lineáris regressziós modellt ne használjuk, ha valamelyik feltétele szignifikánsan nem

¹ Ilyenkor a regressziós paraméterek valamilyen speciális lineáris kombinációja megbecsülhető, még ha az egyes paraméterek értékei nem is (THEIL, [1971]).

teljesül! Közgazdasági elemzéseknél ennek leggyakrabban három oka lehet (PETRES – TÓTH, [2006]):

- *multikollinearitás*: a magyarázóváltozók lineárisan nem függetlenek, vagy az együttmozgásuk statisztikailag jelentős, azaz szignifikáns;
- *autokorreláció*: a hibatag különböző megfigyelésekhez tartozó értékei együttmozognak;
- *heteroszkedaszticitás*: a hibatag szórásnégyzete nem állandó.

A standard lineáris regressziós modellben a becsült paraméterek varianciáit a

$$(3) \quad \text{Var}(\hat{\beta}) = \tilde{\sigma}^2 (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1}$$

alapján tudjuk kiszámítani.

A továbbiakban, a multikollinearitás jelenségére fókuszálunk, ugyanis lineáris regressziós modellek esetén ez a jelenség a redundancia egy fajtájaként értelmezhető.

I. 3. A multikollinearitás fogalma

A multikollinearitás fogalma a szakirodalomban látszólag egységes. Az egyes definíciók általában egy-egy szóban térnek el egymástól, de – mint látni fogjuk – ez jelentős tartalmi változást jelent.

A *multikollinearitás* fogalmát RAGNAR FRISCH vezette be. Olyan esetek leírására használta, amikor egy változó több összefüggésben szerepel. Ezekben a vizsgálataiban nem különböztette meg az eredményváltozót a magyarázóváltozóktól. Feltételezése szerint, minden változó mérése hibás, ez alapján kell megbecsülni a változók tényleges értékei közötti korrelációt (MADDALA, [2004]).

Nagyon felületes meghatározás az, hogy *a multikollinearitás a tényezőváltozók függetlenségének a hiánya*. Ezzel a meghatározással az a probléma, hogy nem derül ki egyértelműen az, hogy mit értünk a magyarázóváltozók függetlensége alatt. Netán ezek lineáris függetlenségét, netán statisztikai értelemben vett függetlenségét. Az 1. táblázat adatai és ezen tényezőváltozók korrelációs mátrixa (2. táblázat) alapján látható, hogy nem

szignifikáns korrelációs kapcsolatok esetén is lehetnek a tényezőváltozók lineárisan összefüggőek.

1. táblázat: Lineárisan függő tényezőváltozók nem szignifikáns korrelációs kapcsolatokkal

$z_i = x_i + y_i$	y_i	x_i
10	7	3
9	5	4
13	8	5
10	1	9
11	1	10
4	2	2

(forrás: saját szerkesztés)

2. táblázat: Lineáris korrelációs együtthatók

		$z_i = x_i + y_i$	y_i	x_i
$z_i = x_i + y_i$	r	1	.428	.517
	Szignifikancia		.397	.294
	n	6	6	6
y_i	r	.428	1	-.553
	Szignifikancia	.397		.256
	n	6	6	6
x_i	r	.517	-.553	1
	Szignifikancia	.294	.256	
	n	6	6	6

(forrás: saját szerkesztés)

Továbbá, nagyon erős korrelációs kapcsolatok esetén sem feltétlen lehet lineárisan összefüggő változókról beszélni. Viszont, az biztos, hogy bárhogyan is értik a függetlenséget, ennek hiánya esetén nem lesz minden korrelációs együttható nulla, azaz valamilyen mértékű együttmozgás létezik a tényezőváltozók között.

A standard lineáris regressziós modell egyik alapfeltétele a magyarázóváltozók lineáris függetlensége. Ezért, egyes forrásokban *multikollinearitás alatt a tényezőváltozók lineáris függetlenségének hiányát értik* (HERMAN – PINTÉR – RAPPAL – RÉDEI, [1994]). Ez gyakorlatilag azt jelenti, hogy valamelyik tényezőváltozó kifejezhető a többi tényezőváltozó nem triviális lineáris kombinációjaként. Ennek következtében az $\tilde{\mathbf{X}}\tilde{\mathbf{X}}$

mátrix nem invertálható, így a regressziós együtthatók (2) képlet szerinti becslése nem lehetséges. A továbbiakban ezt a megközelítést a multikollinearitás egy speciális esetének tekintem, melyet *extrém multikollinearitásnak* nevezünk. Ez az eset a gyakorlatban nem okoz különösebb problémát, mivel könnyen kezelhető.

Az empirikus elemzések során nagyon gyakran találkozhatunk az *extrém multikollinearitáshoz közeli esetekkel*, amikor is az $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ mátrix ugyan invertálható, de egyes becsült paraméterek varianciái nagymértékben növekednek a hibátag szórásnégyzetéhez képest. A multikollinearitással foglalkozó szakirodalmak döntő többsége ezzel az esettel foglalkozik. Azonban, előljáróban megjegyzem, hogy multikollinearitás alatt sokkal általánosabb jelenséget is lehetne érteni, mégpedig *a tényezőváltozók együttmozgását*. Természetesen ennek a meghatározásnak a speciális esetei mindenki számára visszaadnák azt a fogalmat, amit a multikollinearitás alatt ért.

Ha a modell specifikációja megfelelő, akkor a multikollinearitás csak a megfelelő információ hiányának következménye (KÖRÖSI – MÁTYÁS – SZÉKELY, [1990]). Empirikus vizsgálatoknál gyakran *komoly problémát jelent a multikollinearitás felismerése és okának megtalálása*, hiszen egyrészt *a multikollinearitás negatív következményei nem mindig lépnek fel*, másrészt *a multikollinearitást nem csak egy változó, hanem egy változócsoport is okozhatja*. Így sejthető, hogy *a multikollinearitás mérőszámai nem minden esetben jellemzik megfelelően ezt a jelenséget*.

A multikollinearitás mérőszámainak értelmezése sokszor meglehetősen szubjektív. Ugyanis a mérőszámok többsége arra ad választ, hogy a vizsgált adatállomány mennyire nem ideális, azaz milyen mértékben térünk el az „ideális esettől”, amikor is minden tényezőváltozó lineárisan független egymástól. Néhány mérőszám esetén nincs egyértelmű határ az „eltérés” káros mértékű jelzésére.

A multikollinearitás negatív hatásainak csökkentésére, illetve kiküszöbölésére gyakrabban használt módszerek sikeressége nagymértékben függhet a multikollinearitás pontos felismerésétől. Ezen módszerek többségének alkalmazása ugyan csökkenti, pontosabban – mint látni fogjuk – csökkentheti a multikollinearitás negatív következményeinek mértékét, de ez más negatív következményekkel – például jelentős információvesztéssel, az eredmények nem megfelelő értelmezhetőségével – járhat.

A redundancia egy új, lehetséges mérőszáma a PETRES-féle *Red*-mutató (PETRES – TÓTH, [2004]).

I. 4. A PETRES-féle *Red*-mutató

A *Red*-mutató definiálásakor a tényezőváltozók $\mathbf{R}$ korrelációs mátrixának λ_j ($j=1,2,\dots,m$) sajátértékeit alkalmazzuk. Mivel a korrelációs mátrix pozitív szemidefinit mátrix, ezért ennek sajátértékei nemnegatívak. A korrelációs mátrix sajátértékeinek száma és összege is megegyezik a magyarázóváltozók számával. Ebből következően sajátértékeinek számtani átlaga egy.

A *Red*-mutató az alábbi gondolatmeneten alapszik. Ha a magyarázóváltozók forrásául szolgáló adatállomány a $\hat{\beta}$ becslőfüggvény szempontjából redundáns, azaz nagymértékű az adatok együttmozgása, akkor nem mindegyik adat hordoz hasznos tartalmat. Minél kisebb a hasznos tartalmat hordozó adatok aránya, annál nagyobb a redundancia mértéke. Minél nagyobb mértékben szóródnak a sajátértékek, annál nagyobb mértékű az adatállományban szereplő magyarázóváltozók együttmozgása. Két szélsőséges eset létezik: minden sajátérték egyenlő egymással (azaz értékük egy), illetve egy sajátérték kivételével mindegyik sajátérték nullával egyenlő. A diszperzió mértékét számszerűsíthetjük a sajátértékek relatív szórásával vagy (ebben az esetben az ezzel egyenlő) szórásával.

$$(4) \quad v_\lambda = \frac{\sigma_\lambda}{\bar{\lambda}} = \frac{\sqrt{\frac{\sum_{j=1}^m (\lambda_j - \bar{\lambda})^2}{m}}}{\frac{\sum_{j=1}^m \lambda_j}{m}} = \frac{\sqrt{\frac{\sum_{j=1}^m (\lambda_j - \bar{\lambda})^2}{m}}}{\frac{m}{m}} = \sqrt{\frac{\sum_{j=1}^m (\lambda_j - 1)^2}{m}} = \sigma_\lambda$$

Különböző adatállományok redundanciájának összevethetősége végett a fenti mutatót normálni kell. Mivel a sajátértékek nemnegatívak, ezért a relatív szórásra vonatkozó $0 \leq v_\lambda \leq \sqrt{m-1}$ összefüggés² miatt, a normálás $\sqrt{m-1}$ értékével történik.

² A relatív szórás két szélső korlátjára (ha $x_i \geq 0$) felírhatjuk a $0 \leq v \leq \sqrt{N-1}$ összefüggést. Az alsó korlát $v=0$ minden esetben fennáll, ha $x_i = x$ ($i=1,2,\dots,N$). A felső korlát $v = \sqrt{N-1}$ csak akkor áll fenn, ha $x_i = 0$ ($i=1,2,\dots,N-1$) és $x_N = N \cdot \bar{x}$.

Az így kapott mutatót a redundancia mértékének számszerűsítésére használhatjuk, és segítségével a *Red*-mutatót az alábbiak szerint definiáljuk.

$$(5) \quad Red = \frac{v \lambda}{\sqrt{m-1}}$$

A redundancia hiánya esetén a fenti mutató értéke nulla, illetve nulla százalék, míg maximális redundancia esetén egy, illetve száz százalék.

A *Red*-mutató a vizsgált, adott méretű adatállomány redundanciáját méri. Két vagy több különböző méretű adatállomány redundanciájának összevetésekor a *Red*-mutatók alapján csak annyi állítható, hogy az egyes adatállományok mennyire redundánsak, de arra vonatkozó közvetlen kijelentés nem tehető, hogy ezek közül melyiknek van több hasznosítható adata.

I. 5. A disszertáció célja és a kutatási hipotézisek

A disszertáció céljának megfogalmazása végett röviden **összefoglalom az eddig ismertetett problémákat.**

Empirikus elemzéseknél gyakori eset, hogy a vizsgálat szempontjából nem minden adat hordoz hasznos tartalmat, azaz az adatállomány redundáns. Ez az eset a többváltozós lineáris regressziószámításnál a multikollinearitással magyarázható.

Ezért a regressziószámítás során fontos tudni a (2) becslőfüggvény szempontjából hasznos tartalmat hordozó adatok arányát, de probléma ennek a megfelelő mérése. Erre egy lehetőség a PETRES-féle *Red*-mutató.

Kérdéses, hogy mit jeleznek a multikollinearitás mérőszámai, illetve az, hogy a multikollinearitás jelenlétének negatív következményei hogyan csökkenthetőek.

A téma aktualitását az adja, hogy ezek a problémák a gazdasági elemzések során szinte kivétel nélkül jelentkeznek. Különösen igaz ez, ha a magyarázóváltozókban erős trend van, vagy ha túlságosan kevés információ áll rendelkezésre ahhoz, hogy a tényezőváltozóknak az eredményváltozóra gyakorolt hatását vizsgáljuk (KENNEDY, [2003]).

Ezért disszertációm **célja a *Red*-mutató tulajdonságainak vizsgálata, valamint más mérőszámokkal történő összehasonlítása, a többváltozós, lineáris regressziós modellen bemutatva.** Ennek megfelelően **értekezésem az alábbi felépítést követi.**

A II. fejezetben a multikollinearitással kapcsolatos szakirodalmat mutatom be. Ebben a fejezetben áttekintésre kerül a multikollinearitás számos ismert, illetve kevésbé ismert mérőszáma, detektálási módja, lehetséges következményei, valamint ezek negatív hatásának csökkentési lehetőségei.

A III. fejezetben ismertetem a kutatásaim során alkalmazott módszereket, illetve ezek eredményeit. Megvizsgálom a *Red*-mutató főbb tulajdonságait. Itt ismertetem más, hasonló vizsgálati módszerek eredményeit, összevetve az általam kapott eredményekkel.

A IV. fejezetben a kutatási tevékenységem és eredményeim értékelése következik.

Az V. fejezetben ismertetem a felhasznált irodalmak jegyzékét.

A VI. fejezetben találhatóak a disszertáció mellékletei. Az ábrák, a táblázatok jegyzéke, a hosszabb számítógépes elemzések kimenetei, és a publikációim felsorolása.

A disszertáció céljának eléréséhez az alábbiakban ismertetett problémaköröket vizsgálom, illetve feladatokat hajtom végre.

1. A multikollinearitás szakirodalmának áttekintése.

A szakirodalmi áttekintést a dolgozat II. fejezete tartalmazza. A multikollinearitás fogalmának előzetes tárgyalása után a fejezetben először a multikollinearitás gyakran emlegetett negatív következményeit foglalom össze. Majd áttekintésre kerül a multikollinearitás 18 ismert, illetve kevésbé ismert mérőszáma, detektálási, lokalizálási módja. A multikollinearitás felismerése és lokalizálása után 8 olyan módszert ismertetek, amelyek alkalmasak lehetnek a jelenség negatív hatásainak csökkentésére. A fejezet zárásaként egy empirikus példán szemléltetem a fejezetben bemutatott módszereket, eljárásokat.

2. A *Red*-mutató más módon történő kiszámítása.

A *Red*-mutató definíciója szerint a korrelációs mátrix sajátértékei alapján számítható ki. Felmerülhet a kérdés, hogy a sajátértékek ismerete nélkül kiszámítható-e a mutató

értéke, pusztán a tényezőváltozók korrelációs mátrixának elemei alapján. A III.1. fejezetben be fogom bizonyítani az alábbi állítást.

1. Állítás: *A Red-mutató kifejezhető a tényezőváltozók korrelációs mátrixa sajátértékeinek ismerete nélkül, pusztán a páronkénti korrelációs együtthatók négyzetes átlagaként.*

3. A multikollinearitás vizsgálati módszerének általánosítása

Úgy gondolom, hogy a multikollinearitás vizsgálatakor nem csak változópárok együttmozgása, hanem változócsoportok együttmozgása is problémát jelenthet. Ennek azonban még nincs részletesen kidolgozott szakirodalma. Úgy látom, hogy a problémára megoldást jelenthet a kanonikus korrelációelemzés használata, melynek egy speciális helyzete vizsgálható a Red-mutató segítségével. A III.1. fejezetben be fogom bizonyítani az alábbi állítást.

2. Állítás: *Tényezőváltozók két csoportja együttmozgásának vizsgálata egy-egy elemű csoportok esetén a Red-mutatóval, míg egy-(m-1) elemű csoportok esetén a VIF_j mutatók harmonikus átlagának segítségével lehetséges.*

4. A multikollinearitás új modellezési lehetőségének vizsgálata.

A multikollinearitás modellezésének egy módja a tényezőváltozók ortogonalitásának, azaz a tényezőváltozók tere „kifeszítettségének” vizsgálata. Jogos kérdés, hogy lehet-e másképpen modellezni a multikollinearitást. A III.2. fejezetben be fogom bizonyítani az alábbi állítást.

3. Állítás: *Új megközelítésként megalkotható a multikollinearitás elliptikus modellje a Red-mutató alapján.*

5. Valamilyen kapcsolat keresése a becsült regressziós paraméterek varianciái és a Red-mutató között.

Mivel a multikollinearitás egyik leggyakrabban említett negatív következménye a becsült regressziós paraméterek varianciáinak, illetve ezek felfújódásának növekedése,

ezért célszerű megvizsgálni a *Red*-mutató és a becsült regressziós paraméterek varianciáinak kapcsolatát. A III.3. fejezetben be fogom bizonyítani az alábbi állítást.

4. Állítás: *Megadható a Red-mutató egy olyan kritikus értéke, amely szükséges feltétele annak, hogy a becsült paraméterek varianciái ne legyenek végtelenek.*

6. A *Red*-mutató eloszlásának vizsgálata.

A III.4. fejezetben megpróbálom a *Red*-mutató empirikus eloszlásfüggvényét elkészíteni, illetve az elméleti eloszlását meghatározni.

7. A *Red*-mutató alkalmazási lehetőségeinek vizsgálata.

Érdekes kérdés, hogy a *Red*-mutató milyen területeken alkalmazható. A III.5. fejezetben be fogom bizonyítani az alábbi állítást.

5. Állítás: *A Red-mutató alapján kifejezhető a faktoranalízis során használt KMO-mutató.*

8. A *Red*-mutatóhoz hasonló mérőszám megadása.

Mivel a *Red*-mutató a tényezőváltozók korrelációs mátrixának sajátértékei alapján számított normált relatív szórás, ezért úgy gondolom, hogy a multikollinearitás mérhető a sajátértékek más szóródási mérőszámával is, melynek alapgondolata megegyezik a *Red*-mutató alapötletével. A III.6. fejezetben be fogom bizonyítani az alábbi állítást.

6. Állítás: *A Red-mutató definiálásának gondolatmenetén alapuló hasonló multikollinearitás mérőszám a tényezőváltozók korrelációs mátrixa sajátértékeinek GINI-együtthatója.*

A továbbiakban, a könnyebb megértés kedvéért az értekezésben szereplő nagyszámú képlet közül csak azok kerülnek számozásra, amelyekre hivatkozás történik.

II. A multikollinearitás szakirodalmának áttekintése

A kitűzött feladatoknak megfelelően először a multikollinearitás szakirodalmát foglalom össze. Ebben a fejezetben áttekintésre kerül a multikollinearitás számos ismert, illetve kevésbé ismert mérőszáma, detektálási módja, lehetséges következményei, valamint ezek negatív hatásának csökkentési lehetőségei.

II. 1. A multikollinearitás következményei

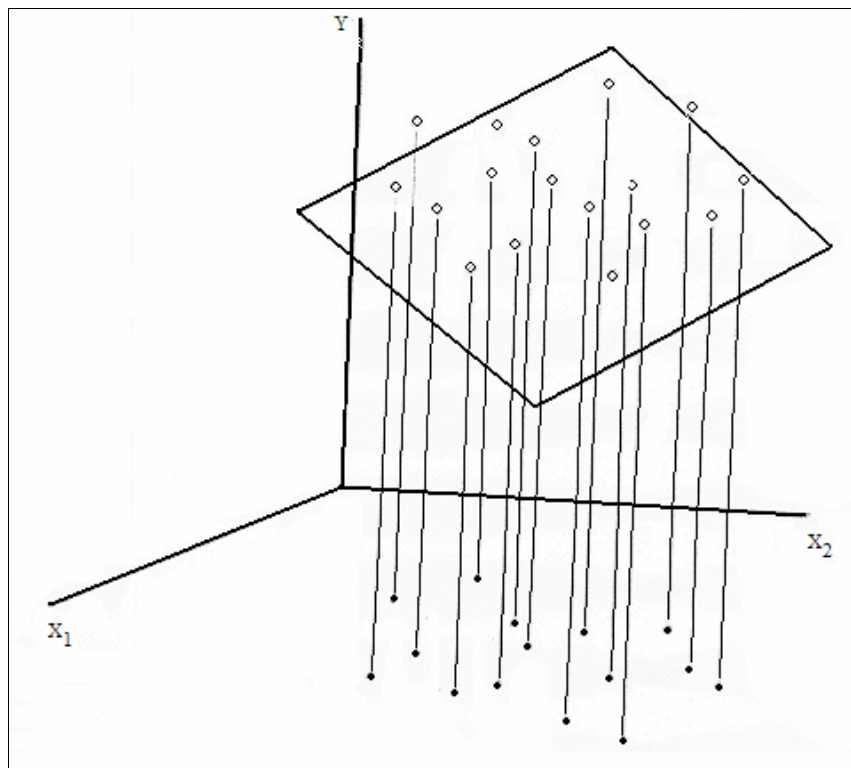
Sok helyen olvasható, hogy a multikollinearitásnak számos negatív következménye van. Mint a későbbiekben rávilágítok, a sokszor emlegetett negatív következmények nem mindig, csak bizonyos esetekben (*near multicollinearity*) jelentkeznek.

A multikollinearitás gyakran emlegetett következményei az alábbiak (KENNEDY, [2003]):

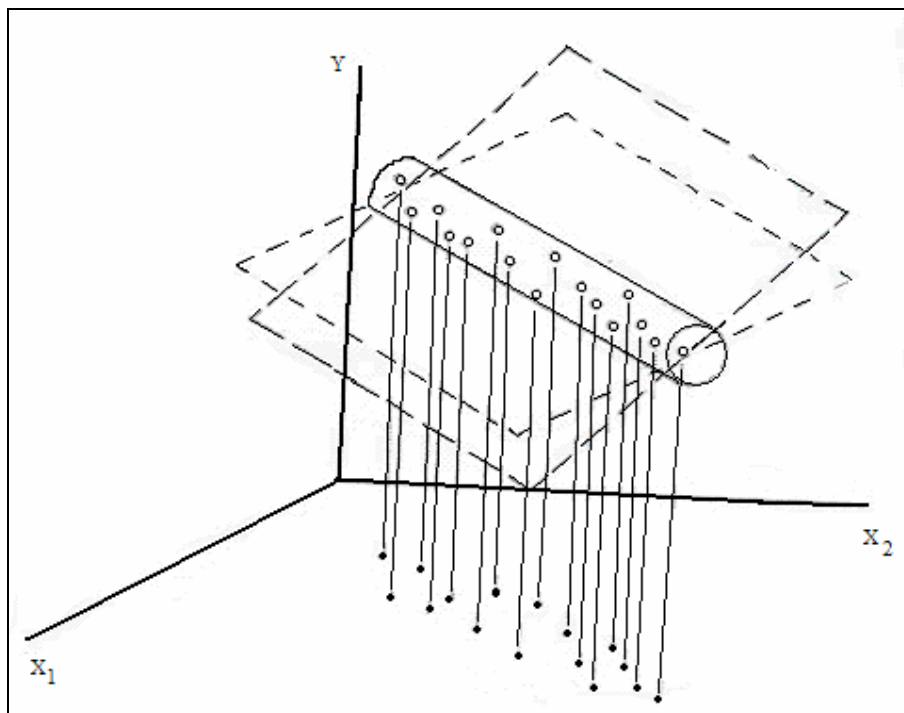
- a becslés és az előrejelzés torzítatlan marad.
- A regressziós együtthatók (3) képlettel adott standard hibái nőnek.
- Az egyes magyarázóváltozók szeparált hatásának vizsgálata értelmetlenné válik. Ugyanis, a becsült paraméterek szórásnégyzete (3) szerint nagy mértékben növekszik, melynek következtében a parciális F -próbák (vagy t -próbák) értelmüket veszítik, hiszen ezen próbafüggvényeknek az értékei nagyon alacsonyak lesznek.
- A regressziós paraméterek (2) képlettel adott becslése bizonytalanabbá, instabillá válik. Ezt szemlélteti két tényezőváltozó esetén az 1. ábra és a 2. ábra.

Az 1. ábra azt mutatja, hogy – a magyarázóváltozók statisztikailag jelentéktelen együttmozgása esetén – a becsült paraméterek varianciái, a jelentős együttmozgás esetén kiszámított szórásnégyzetekhez viszonyítva jóval kisebbek. Ez azért van, mert az első esetben az adatállománynak a „pontfelhője” az x_1 - x_2 síkvetületben minden dimenzióban szóródik és így az illesztett regressziós sík stabil. Míg a 2. ábra „pontfelhője” nem mindegyik dimenzióban szóródik az x_1 - x_2 síkvetületben, így a ráillesztett sík könnyen kibillen, azaz instabillá válik az illesztés. Magasabb dimenziószám esetén az ábrák vizsgálata minden x_i - x_j síkvetületben szükséges.

1. ábra: Stabil regressziós sík a magyarázóváltozók nem szignifikáns együttmozgása esetén ($m=2$)



2. ábra: Instabil regressziós sík szignifikáns multikollinearitás esetén ($m=2$)



(Mindkét ábra forrása: TRIČKOVIĆ, [1976])

Az eddigiek szerint, ha a használt modellel kizárólag előrejelzést szeretnénk készíteni, akkor nem jelent túlságosan nagy problémát a multikollinearitás jelenléte. Azonban a tényezőváltozók parciális hatásainak a vizsgálata értelmetlenné válik.

A következmények között találjuk azt, hogy a becsült regressziós paraméterek varianciái növekednek, illetve ezeknek magas értékük lesz. Ezzel az állítással kapcsolatosan két problémát lehet megfogalmazni. Egyrészt nem mindegyik variancia fog növekedni, másrészt pedig, mit értünk az alatt, hogy ezeknek magas értékük lesz. Erre mutatott rá MADDALA (2004). Olyan ellenpéldát adott, amelyben a magyarázóváltozók nagyon erős kapcsolatai ellenére a becsült paraméterek varianciái – a korábbi vizsgálati eredményekhez képest – alacsony értékűnek tűnnek.

A látszólagos ellentmondás abban rejlik, hogy számos irodalom elfelejti feltüntetni, hogy a varianciák növekedését ceteris paribus értjük. Ugyanis, ha megvizsgáljuk a (3) összefüggést, akkor láthatjuk, hogy a becsült paraméterek varianciái két tényezőtől függenek. Egyrészt, a hibatag varianciájától, másrészt a képletben szereplő $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ mátrix diagonális elemeitől. A MADDALA által adott ellenpéldában azért nem lesznek nagyok a becsült paraméterek varianciái, mert alacsony a hibatag varianciájának becsült értéke, azaz a reziduális szórásnégyzet. **Ezért, gyakorlatilag a becsült paraméterek varianciáinak nem az abszolút nagyságát kell nézni, hanem azt, hogy mekkora ezeknek**

$$(6) \quad \frac{Var(\hat{\beta}_i)}{\hat{\sigma}^2} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}_{ii}$$

„felfújódása” a hibatag varianciájához képest.

II. 2. A multikollinearitás felismerése, mérése, illetve mérőszámainak bírálata

A multikollinearitás detektálásának és mérésének számos módja ismert, azonban ezek közül kevés a széles körben elfogadott mivel, egyrészt a multikollinearitás detektálása sokszor nagyon nehéz feladat, másrészt a mutatók többségének értelmezése meglehetősen szubjektív. A mérőszámok, eljárások egy része általában csak detektálják a multikollinearitást, de – általában szintetikus jellegük miatt – nem lokalizálják a problémát.

Ezzel szemben a mérőszámoknak és eljárásoknak egy csoportja – több-kevesebb sikerrel – megpróbálja lokalizálni (általában egy változóhoz kötni) a multikollinearitás jelenségét (például a tényezőváltozók korrelációs mátrixának vizsgálata, a KLEIN-féle hüvelykujj-szabály, MASON és PERREAULT vizsgálata, WILKS F-tesztje, FRISCH sugárkéve-térképek módszere, a VIF_j -mutató, a tolerancia-mutató, BELSLEY variancia-dekompozíciós eljárása). Ezen módszerekkel a legfőbb probléma az, hogy a multikollinearitás lokalizálása vagy változópárok, vagy $1-k$ elemű változócsoportok együttmozgása alapján történik, azonban a multikollinearitást két tetszőleges számú tényezőváltozóból álló csoport együttmozgása is okozhatja.

Egy mutató hasznos tulajdonságai az alábbiak.

1. A mutató normált legyen, azaz értéke 0 és 1 közé essen.
2. A mutató szintetikus (átfogó) legyen.
3. A mutató értelmezése objektív legyen.

A továbbiakban e szempontok szerint is jellemzem a multikollinearitás néhány mutatóját.

A multikollinearitás felismerésének egy egyszerű módszere az, hogy a **tényezőváltozók korrelációs mátrixát vizsgálva**, nagyobbban tekintjük a multikollinearitás mértékét, ha a főátlón kívüli elemek abszolút értékei messzebb esnek nullától. A módszerrel több probléma van. A módszer nem határozza meg egyértelműen azt, hogy hány korrelációs együttható szignifikáns eltérése jelez multikollinearitást. Az 1. példában szereplő három tényezőváltozó lineárisan nem független egymástól, – azaz a modellben extrém multikollinearitás van – ennek ellenére a tényezőváltozók korrelációs mátrixában a korrelációs együtthatók nullától való különbözőségeiről nem tudjuk megállapítani, hogy azok jelentősek-e, vagy sem. Így az eljárás a multikollinearitás sikeres lokalizálására nem teljesen alkalmas.

A **KLEIN-féle hüvelykujjszabály** szerint akkor kell szignifikáns multikollinearitással számolni, ha a magyarázóváltozók korrelációs mátrixában létezik olyan korrelációs együttható, amelynek értéke közel van a többszörös korrelációs

együttható értékéhez. Ez a módszer meglehetősen szubjektíven értelmezi a közelség fogalmát (HERMAN – PINTÉR – RAPPAI – RÉDEY, [1994]).

MASON és PERREAULT azt javasolta, hogy a vizsgálatba vont eredményváltozó és m darab tényezőváltozó felhasználásával, a változók megkülönböztetése nélkül készítsük el az összes lehetséges $(m+1)$ -dimenziós regressziós modellt úgy, hogy mindegyik modellben az eredményváltozó eredetileg egy-egy magyarázóváltozó volt. Amennyiben ezen modelleknek a többszörös determinációs együtthatói kisebbek az eredeti szereposztású modell többszörös determinációs együtthatójánál, akkor a multikollinearitás nem jelent problémát a vizsgálat szempontjából (MASON – PERREAULT, [1991]).

Az M_1 szintetikus **mutató** a magyarázóváltozók és az eredményváltozó közötti korrelációs mátrixot használja. Ha a magyarázóváltozók egymástól függetlenek, akkor a többszörös determinációs együttható értéke megegyezik az eredményváltozó és a magyarázóváltozók közötti páronkénti korrelációs együtthatók négyzetösszegével. Ennek az összegnek az $r^2_{\tilde{y}, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m}$ többszörös determinációs együttható tényleges értékétől való eltérése a multikollinearitás jelenlétére utal.

$$M_1 = \sum_{i=1}^m r^2_{\tilde{y}\tilde{x}_i} - r^2_{\tilde{y}, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m}$$

A fő kérdés az, hogy mekkora eltérés jelez erős multikollinearitást (HERMAN – PINTÉR – RAPPAI – RÉDEY, [1994])?

Egy másik szintetikus mutató az

$$\mathbf{M} = r^2_{\tilde{y}, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m} - \sum_{j=1}^m (r^2_{\tilde{y}, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m} - r^2_{\tilde{y}, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{j-1}, \tilde{x}_{j+1}, \dots, \tilde{x}_m}),$$

aminek a többszörös determinációs együtthatóhoz közeli értéke jelentős multikollinearitást jelez (FÖRSTER – EGERMAYER, [1966]). Hátránya, hogy – a tényezőváltozók számának növekedésével – értékének meghatározása a többi mutató kiszámításához viszonyítva aránytalanul több számítást igényel. A „közelség” értelmezése szubjektív, ráadásul az M

értéke negatív is lehet, így semmiképp sem tekinthető normálnak. A képlet magyarázatának látszólag két megközelítést adhatunk.

Az egyik szerint, a képletet átrendezve láthatjuk, hogy az összefüggés a többszörös determinációs együtthatót bontja fel a tényezőváltozók közvetlen hatásaira, illetve az M által mért közvetett hatásra, tehát az eredményváltozó szórásnégyzetének a magyarázóváltozók által együttesen megmagyarázott hányadát bontjuk fel a tényezőváltozók által külön-külön és egy közösen meghatározott részre.

Két magyarázóváltozó esetén a fenti összefüggés szerkezete gyakorlatilag a szita-formula analógiája, az együttesen megmagyarázott részre, mint halmazra alkalmazva. Márpedig a szita-formula végeredménye nem lehet negatív előjelű. Tehát a képletnek az e fajta interpretációja félrevezető, ugyanis a magyarázóváltozók közvetlen hatásainak mértéke nem egyezik meg a képletben szereplő értékkel. A képletben közvetlen hatásként azt mérjük, hogy ha egy adott magyarázóváltozót utoljára kapcsolunk be a modellbe, akkor az mennyivel növeli meg a többszörös determinációs együttható értékét. THEIL ezeket a tényezőket, azaz a képlet összeadandó részeit az adott változónak a többszörös determinációs együtthatóhoz tartozó *növekményi hozzájárulásának* nevezte.

Pontosan ezek a növekmények jelentik a képlet másik magyarázatát. Ha az összes tényezőváltozó páronként független, akkor a többszörös determinációs együttható értéke pontosan megegyezik a növekmények összegével, tehát ekkor a mutató értéke nulla.

Egy újabb lehetséges vizsgálati módszer **a magyarázóváltozók ortogonalitásának vizsgálata**. Ha a magyarázóváltozók lineárisan függetlenek egymástól, akkor a modellben szereplő tényezőváltozók ortogonálisnak tekinthetők, ekkor a *tényezőváltozók korrelációs mátrixának determinánsa* egy. Minél jobban távolodunk ettől az esettől, a korrelációs mátrix determinánsának abszolút értéke egyre inkább nullához közelít. A korrelációs mátrix determinánsa megegyezik a mátrix sajátértékeinek szorzatával. Ez a módszer csak alacsony dimenziószám esetén használható megfelelően (FELLMAN, [1981]). A kérdés megint csak az, hogy mit jelent a nullához való közelség?

A **FARRAR – GLAUBER-féle vizsgálat** (FARRAR – GLAUBER, [1967]) szerint a korrelációs mátrix determinánsa megközelítőleg χ^2 - (*khi-négyzet*) *eloszlásúvá* transzformálásával az alábbi próbafüggvényt kapjuk.

$$\chi^2 = -(n-1 - \frac{1}{6}(2m+5)) \lg \det \mathbf{R}$$

A hipotézisvizsgálat nullhipotézise a magyarázóváltozók lineáris függetlensége, vagyis az, hogy a determináns abszolút értéke egy. Ennek a próbafüggvénynek a szabadságfoka $\frac{m(m-1)}{2}$. A nullhipotézis elfogadása nem jelenti automatikusan azt, hogy a multikollinearitás nincs jelen a modellben (HULYÁK, [1969]).

A **magyarázóváltozók korrelációs mátrixának inverzét vizsgálva** megállapítható, hogy a mátrix diagonális elemei egynél nem lehetnek kisebbek. Minél nagyobb az együttmozgás egy változó és a többi változó között, annál jobban eltérnek egytől a megfelelő diagonális elemek. Ez alapján egy parciális próbát lehet alkalmazni a multikollinearitás tesztelésére. WILKS kimutatta, hogy a diagonális elemek megközelítőleg $n-m$ és $m-1$ szabadságfokú F -eloszlásúvá transzformálhatóak, ami az alábbi próbafüggvényt eredményezi.

$$\omega_i = \frac{n-m}{m-1} (\mathbf{R}_{ii}^{-1} - 1)$$

A próba nullhipotézisének elvetése azt jelenti, hogy a vizsgált magyarázóváltozó és a többi tényezőváltozó között – adott szignifikanciaszint mellett – a multikollinearitás szignifikánsnak tekinthető (HULYÁK, [1969]).

A multikollinearitás jelenlétére gondolhatunk akkor is, amikor két tényezőváltozó közötti **parciális korrelációs együttható** értéke jelentősen eltér a két változó közötti korrelációs együttható értékétől. A parciális korrelációs együtthatók szignifikanciájának t -próba segítségével történő tesztelését is alkalmazhatjuk.

FRISCH sugárkéve-térképek módszere (*bunch maps*) a normált regressziós együtthatók ábrái alapján következtet a multikollinearitás jelenlétére. Az eljárás megfelelő rutin nélkül nagyon nehézkesen alkalmazható. A módszer nem különbözteti meg a magyarázóváltozókat az eredményváltozótól, tehát bemenetként adott $m+1$ darab változó. Ezután mindegyik változónak az átlagától való eltéréseire először $(m+1)$ -dimenziós lineáris regressziós modellt illesztünk úgy, hogy minden változó szerepeljen eredményváltozóként

is. Így kapunk $m+1$ darab $m+1$ változós lineáris regresszióegyenletet. Ezek mindegyikéből kifejezzük külön-külön az összes változót. Gyakorlatilag így mindegyik változó $m+1$ darab egyenlettel van felírva a többi változó segítségével. Ezek után iteráljuk az eljárást, vesszük az összes lehetséges m -dimenziós modellt, stb. Az iterációs eljárást két dimenzióig ismétljük. A kapott parciális regressziós együtthatókat az összehasonlíthatóság kedvéért normálnunk kell. A sugárkéve-térképekben ezeket a normalizált együtthatókat ábrázoljuk. A normalizált parciális regressziós együtthatók kifejezhetők a megfelelő korrelációs együtthatók adjungált mátrixának egy-egy megfelelő elemének hányadosaként. Ezen hányadosok számlálói, illetve nevezői lesznek a sugárkéve-térképeken ábrázolandó koordináták. Egy sugárkéve nem más, mint egy-egy változópár közötti, összes kapott – adott dimenziójú – együtthatóknak az ábrája.

A kévek zártóságából, meredekségéből és a sugarak hosszából kimutatható a multikollinearitás, illetve megállapítható, mely magyarázóváltozók lesznek hasznosak, károsak, illetve feleslegesek az eredményváltozó megmagyarázásának szempontjából. A kéve zártága azt mutatja, hogy a két változó között milyen szoros kapcsolat van. Minél rövidebb egy sugár, annál szorosabb a kapcsolat a többi változó között, ezért azok lesznek a legfontosabb változók, amelyekhez a leghosszabb sugarak tartoznak (CORRADI, [1967]).

A VIF_j (CHATTERJEE–PRICE, [1977]) nem szintetikus **mutató**, hiszen minden magyarázóváltozóra külön-külön kiszámítjuk, azaz ez a mutató valamelyik változóhoz próbálja kötni a multikollinearitást. Ez azért nem túl szerencsés, mert sok esetben a multikollinearitást nem egy változó okozza.

$$(7) \quad VIF_j = \frac{1}{1 - r_{\tilde{x}_j, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{j-1}, \tilde{x}_{j+1}, \dots, \tilde{x}_m}^2}$$

Ha a j -edik tényezőváltozó lineárisan független a többi magyarázóváltozótól, akkor a mutató értéke eggyel egyenlő. Extrém multikollinearitás esetén a mutató értéke végtelen. Az

$$(8) \quad x_{ij} = \frac{\tilde{x}_{ij} - \bar{\tilde{x}}_j}{\sqrt{n\sigma_j^2}}$$

szerint standardizált magyarázóváltozók esetén $(\mathbf{X}'\mathbf{X})_{jj}^{-1} = VIF_j$.³ A VIF_j mutató megmutatja a becsült regressziós együttható varianciája „felfújódásának” mértékét a hibatag varianciájához viszonyítva. Ennek értelmezése meglehetősen szubjektív abból a szempontból, hogy nincs egyértelmű küszöbszám a multikollinearitás káros voltának jelzésére. Egyes szerzők szerint a mutató öt és e feletti értéke jelez erős multikollinearitást. A VIF_j -mutató reciprokát **tolerancia-mutató**nak nevezzük. A T_j tolerancia-mutató megmutatja, hogy a j -edik magyarázóváltozó szórásnégyzetének mekkora része nem magyarázható együttesen a többi tényezőváltozóval. Ennek értéke nulla és egy közé esik. Minél nagyobb a multikollinearitás mértéke annál közelebb van a mutató értéke a nullához (KOVÁCS – PETRES – TÓTH, [2004]).

A VIF_j -mutató öthöz képest nagyon magas értéke miatt érdekes BOWERMAN példája. Az amerikai hadiflotta kórházainak 1979-es vizsgálatakor 17 kórház adatai alapján a havi munkaórák számára illesztett regressziós modell eredménye a 3. táblázatban látható (FENG-JENQ, [2006]).

3. táblázat: Példa a VIF_j magas értékére
A havi munkaórák becslése lineáris regressziós modellel

Változók	Becsült regressziós paraméterek	t -statisztika	VIF_j
Tengelymetszet	1962,482	1,832	–
Az ellátandó páciensek napi átlagos száma	–15,852	–0,162	9597,57
A havonta elvégzett röntgenvizsgálatok száma	0,056	2,631	7,94
Az ápolási napok száma egy hónapban	1,590	–0,514	8933,09
A körzethez tartozók száma (ezer fő)	–4,219	–0,588	23,29
Az ápolás átlagos időtartama (nap)	–394,314	–1,881	4,28

(forrás: FENG-JENQ, [2006])

³ Ugyanis, a magyarázóváltozók korrelációs mátrixa alapján felírható a $VIF_j = \mathbf{R}_{jj}^{-1}$ összefüggés.

Ekkor – a kizárólag az $x_{ij} = \frac{\tilde{x}_{ij} - \bar{\tilde{x}}_j}{\sqrt{n\sigma_j^2}}$ szerint standardizált változókra érvényes – $(\mathbf{X}'\mathbf{X}) = \mathbf{R}$

egyenlet figyelembevételével az $(\mathbf{X}'\mathbf{X})_{jj}^{-1} = VIF_j$ összefüggést kapjuk.

A táblázat adataiból megállapítható, hogy a VIF_j -mutató értéke az ápolás átlagos időtartamát leszámítva minden változó esetén nagyobb ötnél, azonban az értékek nagyságrendje között jelentős különbség mutatkozik. A multikollinearitásért elsősorban valószínűleg vagy az ellátandó páciensek napi átlagos száma, vagy az ápolási napok száma egy hónapban, vagy mindkét változó felelős. Ennek eldöntésére további vizsgálatokra lenne szükség. Most csak annyit állíthatunk, hogy nem tűnik célszerűnek ezt a két tényezőváltozót egyszerre ugyanabban a modellben szerepeltetni. Egyébként e két változó esetében a t -statisztika értéke is igen alacsony, azonban ezt a multikollinearitás jelenléte miatt nem értelmezhetjük megfelelően.

A multikollinearitás egyik legújabb mérőszáma a CURTO és PINTO által 2007-ben publikált **DEF-mutató** (CURTO – PINTO, [2007]). Amennyiben az

$$\tilde{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_{i,1} + \hat{\beta}_2 \tilde{x}_{i,2} + \dots + \hat{\beta}_m \tilde{x}_{i,m} + \tilde{e}_i$$

regressziós modellt standardizált változókra írjuk fel, akkor ez az egyenlet

$$y_i = \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \dots + \hat{\beta}_m x_{i,m} + \hat{\beta}_e e_i = \hat{y}_i + \hat{\beta}_e e_i$$

alakban írható fel, ahol a $\hat{\beta}_i$ a standardizált regressziós együtthatókat⁴ jelenti.

Ekkor

$$Var(\mathbf{y}) = Var(\hat{\mathbf{y}} + \hat{\beta}_e \mathbf{e}) = Var(\hat{\mathbf{y}}) + \hat{\beta}_e^2 Var(\mathbf{e}) + 2Cov(\hat{\mathbf{y}}; \mathbf{e}) = Var(\hat{\mathbf{y}}) + \hat{\beta}_e^2 + \sum_{i=1}^m \hat{\beta}_i \hat{\beta}_e r_{x_i e}.$$

A standardizált változók és a standardizált hibatag függetlenségének feltételezése mellett

$$Var(\mathbf{y}) = Var(\hat{\mathbf{y}}) + \hat{\beta}_e^2.$$

Ekkor az eredményváltozó eggyel egyenlő varianciáját két részre bonthatjuk fel:

- a tényezőváltozók által együttesen megmagyarázott $Var(\hat{\mathbf{y}})$ hányad, amit a többszörös determinációs együtthatóval mérünk;

⁴ Ez a terminológia azért félrevezető, mert a szakirodalom – kivétel nélkül – nem a regressziós együtthatók standardizált voltára utal, hanem arra, hogy standardizált változók szerepelnek a modellben.

- a tényezőváltozók által együttesen meg nem magyarázott hányad, ami gyakorlatilag

$$Var(\mathbf{y}) - r_{y, x_1, x_2, \dots, x_m}^2 = 1 - r_{y, x_1, x_2, \dots, x_m}^2.$$

Mivel a standardizált eredményváltozó a standardizált változók egy lineáris kombinációja, ezért

$$Var(\hat{\mathbf{y}}) = Var\left(\sum_{j=1}^m \hat{\beta}_j \mathbf{x}_j\right) = \sum_{j=1}^m \hat{\beta}_j^2 + \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \hat{\beta}_i r_{x_i x_j} \hat{\beta}_j.$$

Ezek szerint, a tényezőváltozók által együttesen megmagyarázott $Var(\hat{\mathbf{y}})$ variancia-hányad, és így speciálisan a többszörös determinációs együttható is két részből tevődik össze:

- a tényezőváltozók direkt hatásainak összege: $\sum_{i=1}^m \hat{\beta}_i^2$;
- a tényezőváltozók együttes hatása: $\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \hat{\beta}_i r_{ij} \hat{\beta}_j$.

Ezért, a

$$DEF = \frac{\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \hat{\beta}_i r_{ij} \hat{\beta}_j}{\sum_{i=1}^m \hat{\beta}_i^2 + \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \hat{\beta}_i r_{ij} \hat{\beta}_j}$$

mutató – a szerzők szerint – megmutatja, hogy a többszörös determinációs együttható hány százalékát teszi ki a tényezőváltozók együttes hatása. A szintetikus mutató egyhez közeli értéke erős multikollinearitást jelez.

Vizsgálataim szerint, a mutatóval kapcsolatban több hiányosság is felsorolható. A képletben a tényezőváltozók direkt hatásainak összege, és így a képlet számlálója negatív is lehet, így amellet, hogy százalékban sem fejezhető ki, gondot jelent az értelmezése is (VI. 3. melléklet).

A multikollinearitás mérőszámának egy családját alkotják az $\tilde{\mathbf{X}}' \tilde{\mathbf{X}}$ mátrix sajátértékeire épülő mutatók. A sajátértékek alkalmazásának alapötlete az, hogy a tényezőváltozók együttmozgását jelzik az $\tilde{\mathbf{X}}' \tilde{\mathbf{X}}$ kicsi sajátértékei (KENDALL, [1957];

SILVEY, [1969]). Az $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ mátrix spektrálfelbontása alapján a becsült regressziós paraméterek variancia-kovariancia mátrixa felírható a

$$Var(\hat{\boldsymbol{\beta}}) = \tilde{\sigma}^2 (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} = \tilde{\sigma}^2 \tilde{\mathbf{U}}\tilde{\Lambda}^{-1}\tilde{\mathbf{U}}'$$

formában, ahol $\tilde{\Lambda}$ az $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ mátrix sajátértékeinek diagonális mátrixa, $\tilde{\mathbf{U}}$ pedig a sajátértékekhez tartozó sajátvektorok mátrixa (THEIL, [1971]). Tehát a becsült regressziós együtthatók varianciái függenek az $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ mátrix sajátértékeitől:

$$(9) \quad Var(\hat{\beta}_j) = \tilde{\sigma}^2 \sum_{l=0}^m \frac{\tilde{u}_{jl}^2}{\tilde{\lambda}_l}.$$

A kondíciós szám⁵ négyzetgyökeként definiált

$$\gamma = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

gamma-mutató⁶ értéke egy, amennyiben a tényezőváltozók lineárisan függetlenek. Minél erőteljesebb a multikollinearitás jelenléte, a mutató értéke annál nagyobb lesz. Azonban a szignifikáns multikollinearitásnak nincs egyértelmű küszöbértéke, így értelmezése nem objektív. A mutató kiszámítása nem egységes, ugyanis használatos mind a normált $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ (BELSLEY–KUH–WELSCH, [1980]), mind pedig a standardizált $\mathbf{X}'\mathbf{X}$ mátrix, azaz a korrelációs mátrix (WICHERN–CHURCHILL, [1978]; CASELLA, [1980]) legnagyobb és legkisebb sajátértékének hányadosa is. Egyes szerzők szerint e mutató 30, míg mások szerint 15 feletti értéke jelez erős multikollinearitást.

⁵ A kondíciós szám értéke azt jelzi, hogy a mátrix elemeinek kicsiny (például tizednyi, századnyi) megváltozására hogyan változnak meg az inverz mátrix elemei. Ha ez a változás nagyságrendekkel nagyobb a mátrix elemeinek kicsiny megváltozásához képest, akkor a mátrix rosszul kondicionált.

⁶ Ezt a mutatót, illetve a négyzetét a különböző szakirodalmak más és más szerzők nevéhez kötik. Például WICHERN és CHURCHILL, CASELLA, BELSLEY.

A kondíciós számhoz tartalmilag hasonló, azonban nem szintetikus mérőszám a

$$CI = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}$$

kondíciós index. Adott mátrix esetében a legnagyobb kondíciós index megegyezik a gamma-mutató értékével. Egyes szerzők szerint e mutatók 30, míg mások szerint 15 feletti értéke jelez erős multikollinearitást. Ez a küszöbérték elég szubjektív, ráadásul, mint az alábbi példa (4. táblázat) is mutatja nem mindegy, hogy mely mátrix alapján számítjuk a sajátértékeket.

4. táblázat: Eltérő sajátértékek és kondíciós indexek az $X'X$ különböző alakjai alapján

$\tilde{X}\tilde{X}$		$\bar{X}\bar{X}$		$X'X$	
Sajátértékei	Kondíciós index	Sajátértékei	Kondíciós index	Sajátértékei	Kondíciós index
134339.800	1.000	4.563	1.000		
3483.839	6.210	0.315	3.808	2.216	1.000
112.169	34.607	0.092	7.028	1.627	1.167
35.924	61.152	0.028	12.730	0.149	3.853
1.325	318.472	0.001	56.408	0.007	17.308

(forrás: saját szerkesztés)

A kondíciós index és a (9) alapján BELSLEY és szerzőtársai (1980) egy eljárás adtak a multikollinearitás lokalizálására. A (9) képletet átrendezve

$$\frac{Var(\hat{\beta}_j)}{\tilde{\sigma}^2} = \sum_{l=0}^m \frac{\tilde{u}_{jl}^2}{\tilde{\lambda}_l}$$

látható, hogy az összefüggés a j -edik becsült regressziós paraméter varianciájának a hibátág szórásnégyzetéhez viszonyított „felfújódását” bontja tényezőkre. A képlet jobb oldalán szereplő kifejezés l -edik tagja azt mutatja meg, hogy a j -edik becsült regressziós paraméter varianciájának felfújódásából mekkora az l -edik sajátérték „részesedése”. A varianciák felfújódásának ezen részesedések szerinti megoszlását az 5. táblázat tartalmazza.

A szerzők szerint – normalizált mátrixra alkalmazva az eljárást – lineárisan független tényezőváltozók esetén a fenti elrendezés egy egységmátrix, továbbá ha valamely kicsiny

sajátérték legalább két varianciából magas (legalább 0,8) részesedési arányú, akkor megállapítható, hogy a tényezőváltozók mely csoportja(i) okolhatóak a multikollinearitásért.

5. táblázat: A varianciák „felfújódásának” felbontása

Sajátérték	$Var(\hat{\beta}_0)$	$Var(\hat{\beta}_1)$	...	$Var(\hat{\beta}_m)$
$\tilde{\lambda}_0$	$\frac{\tilde{u}_{00}^2}{\tilde{\lambda}_0}$ $\frac{\sum_{l=0}^m \tilde{u}_{0l}^2}{\tilde{\lambda}_l}$	$\frac{\tilde{u}_{10}^2}{\tilde{\lambda}_0}$ $\frac{\sum_{l=0}^m \tilde{u}_{1l}^2}{\tilde{\lambda}_l}$	...	$\frac{\tilde{u}_{m0}^2}{\tilde{\lambda}_0}$ $\frac{\sum_{l=0}^m \tilde{u}_{ml}^2}{\tilde{\lambda}_l}$
$\tilde{\lambda}_1$	$\frac{\tilde{u}_{01}^2}{\tilde{\lambda}_1}$ $\frac{\sum_{l=0}^m \tilde{u}_{0l}^2}{\tilde{\lambda}_l}$	$\frac{\tilde{u}_{11}^2}{\tilde{\lambda}_1}$ $\frac{\sum_{l=0}^m \tilde{u}_{1l}^2}{\tilde{\lambda}_l}$	...	$\frac{\tilde{u}_{m1}^2}{\tilde{\lambda}_1}$ $\frac{\sum_{l=0}^m \tilde{u}_{ml}^2}{\tilde{\lambda}_l}$
⋮	⋮	⋮		⋮
$\tilde{\lambda}_m$	$\frac{\tilde{u}_{0m}^2}{\tilde{\lambda}_m}$ $\frac{\sum_{l=0}^m \tilde{u}_{0l}^2}{\tilde{\lambda}_l}$	$\frac{\tilde{u}_{1m}^2}{\tilde{\lambda}_m}$ $\frac{\sum_{l=0}^m \tilde{u}_{1l}^2}{\tilde{\lambda}_l}$	...	$\frac{\tilde{u}_{mm}^2}{\tilde{\lambda}_m}$ $\frac{\sum_{l=0}^m \tilde{u}_{ml}^2}{\tilde{\lambda}_l}$

(forrás: BELSLEY–KUH–WELSCH, [1980])

Egy sajátértéket akkor tekintenek a szerzők alacsonynak, ha a hozzátartozó kondíciós index legalább 10. A 3. táblázatban szereplő normált $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ mátrix alapján végrehajtva az eljárást (6. táblázat) megállapíthatjuk, hogy a két legkisebb sajátértékhez tartozó kondíciós index nagyobb, mint a szerzők által megadott határ. Mivel a legkisebb sajátérték (0,001) a harmadik és a negyedik becsült regressziós együttható varianciájának felfújódásából is több mint 99 százalékban részesedik, ezért a harmadik és a negyedik tényezőváltozó nem tekinthetőek egymástól függetlennek. Mivel a második legkisebb sajátérték (0,028) az első és a második becsült regressziós együttható varianciájának felfújódásából is több mint 82 százalékban részesedik, ezért az első és a második tényezőváltozó nem tekinthetőek egymástól függetlennek.

6. táblázat: Variancia-dekompozíció az $\hat{X}'\hat{X}$ mátrix alapján

Sajátérték	Kondíciós index	Variancia-hányad				
		$Var(\hat{\beta}_0)$	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	$Var(\hat{\beta}_3)$	$Var(\hat{\beta}_4)$
4.563	1.000	0.0049175	0.0021241	0.0021822	0.000126	0.000126
0.315	3.808	0.0004181	0.0323175	0.0445732	0.002462	0.002211
0.092	7.028	0.9511659	0.0268882	0.0369498	0.001091	0.001596
0.028	12.730	0.0018491	0.9242759	0.8274794	1.86E-05	0.000309
0.001	56.408	0.0416493	0.0143943	0.0888153	0.996303	0.995759

(forrás: saját szerkesztés)

A sajátértékekre épülő multikollinearitás mérőszámok egy speciális csoportját alkotják a tényezőváltozók korrelációs mátrixának, azaz a standardizált $\mathbf{X}'\mathbf{X}$ sajátértékeire épülő mutatók. Mivel egy $\mathbf{R}_{k \times k}$ korrelációs mátrix négyzetes, szimmetrikus, pozitív szemidefinit, ezért a sajátértékeire általánosan az alábbi tulajdonságok érvényesek (RÓZSA, [1991]).⁷

1. A sajátértékek száma és összege is megegyezik a változók számával. Tehát, ha csak a tényezőváltozók $\mathbf{R}_{m \times m}$ korrelációs mátrixát tekintjük, akkor a λ_i sajátértékek száma és összege is m .
2. A korrelációs mátrix sajátértékei nem negatívak. Az első két tulajdonságból és a mátrix típusából következik, hogy a korrelációs mátrixok sajátértékei a $[0 ; m]$ intervallumba esnek, továbbá a legkisebb sajátérték legfeljebb 1, míg a legnagyobb legalább 1.

3. Az $\mathbf{R}_{m \times m}$ mátrix nyoma megegyezik a sajátértékeinek összegével, azaz

$$tr(\mathbf{R}) = \sum_{i=1}^m \lambda_i .$$

4. Az $\mathbf{R}_{m \times m}$ sajátértékeinek szorzata megegyezik a mátrix determinánsával.
5. Ha az $\mathbf{R}_{m \times m}$ mátrix sajátértékei λ_i , akkor az $\mathbf{R}_{m \times m}^N$ sajátértékei λ_i^N ($N \in \mathbb{Z}$).
Tehát a korrelációs mátrix inverzének sajátértékei a korrelációs mátrix sajátértékeinek reciprokai.

⁷ Természetesen a tulajdonságok egy része általánosabb mátrixokra is érvényes. Itt csak a kutatás szempontjából fontos tulajdonságokat említettem meg.

Ha a (8) szerint standardizált változókat vizsgálunk, akkor a standardizált változókhoz tartozó becsült paraméterek variancia-kovariancia mátrixa felírható az

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \text{Var}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \mathbf{R}^{-1} = \sigma^2 \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}'$$

formában is a korrelációs mátrix spektrálfelbontása alapján, ahol $\mathbf{\Lambda}$ a korrelációs mátrix sajátértékeinek diagonális mátrixa, $\mathbf{U}$ pedig a sajátértékekhez tartozó sajátvektorok mátrixa. Ez utóbbi, illetve a loading változókat tartalmazó $\mathbf{A}$ főkomponenssúly-mátrix tulajdonságainak⁸ figyelembevételével a j -edik standardizált magyarázóváltozóhoz tartozó paraméter becslésének szórásnégyzete az alábbi.

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \sum_{l=1}^m \frac{u_{jl}^2}{\lambda_l} = \sigma^2 \sum_{l=1}^m \frac{a_{jl}^2}{\lambda_l^2}$$

Ebből a varianciák összegére a következő összefüggést⁹ kapjuk:

$$(10) \quad \sum_{j=1}^m \text{Var}(\hat{\beta}_j) = \sigma^2 \sum_{l=1}^m \frac{1}{\lambda_l}.$$

Ezek szerint a varianciák összegének a hibatag szórásnégyzetéhez viszonyított „felfújódásának” mértékét végső soron a sajátértékek befolyásolják: ha legalább egy túl közel van nullához, akkor nagy mértékben növekszik a becsült paraméterek szórásnégyzeteinek átlaga. Az, hogy legalább egy λ közel esik-e nullához, egyértelműen az adatállomány adatainak együttmozgásától, azaz a multikollinearitás mértékétől függ (KOVÁCS – PETRES – TÓTH, [2004]).

A továbbiakban részletesen áttekintem a sajátértékekre épülő mutatókat.

⁸ Az $a_{kl} = u_{kl} \sqrt{\lambda_l}$ főkomponenssúlyok megadják a magyarázóváltozók és a főkomponensek közötti lineáris korrelációs együtthatót: $a_{kl} = r_{\tilde{x}_k c_l} = r_{x_k c_l}$. A főkomponenssúlyok oszloponkénti négyzetösszege λ_j , a soronkénti négyzetösszege egy. Oszloppáronkénti szorzatösszegük nulla, sorpáronkénti szorzatösszegük a megfelelő két magyarázóváltozó lineáris korrelációs együtthatója (PETRES – TÓTH, [2006]).

⁹ Az összefüggés egyszerűbben is megkapható az alábbi módon.

$$\sum_{j=1}^m \text{Var}(\hat{\beta}_j) = \sum_{j=1}^m \sigma^2 (\mathbf{X}'\mathbf{X})_{jj}^{-1} = \sigma^2 \text{tr}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \text{tr}(\mathbf{R}^{-1}) = \sigma^2 \sum_{l=1}^m \frac{1}{\lambda_l}$$

FELLMAN **L-mutatój**ának

$$L = \sum_{i=1}^m \frac{1 - \lambda_i}{\lambda_i^2}$$

nullához közeli értékei jelentéktelen multikollinearitást jeleznek. Minél nagyobb a mutató értéke, annál erősebb a multikollinearitás mértéke (FELLMAN, [1981]).

LAWLESS (1978) a multikollinearitás mérésére egy „másik” **M₁-mutatót** használt (FELLMAN, [1981]).

$$M_1 = \sum_{i=1}^m \frac{1}{\lambda_i}$$

Ennek a mutatónak az előnye a *gamma-mutatóhoz* képest az, hogy az összes sajátértéket figyelembe veszi.

THISTED (1980) egyszerre két mutatót is javasolt. Az egyik az **mci multikollinearitás-index**, a másik pedig a **pmci tervező multikollinearitás-index** (FELLMAN, [1981]).

$$mci = \sum_{i=1}^m \frac{\lambda_{\min}^2}{\lambda_i^2}$$

$$pmci = \sum_{i=1}^m \frac{\lambda_{\min}}{\lambda_i}$$

THISTED az *mci indexet* becslések, míg a *pmci indexet* előrejelzések készítésekor ajánlotta használni. A két indexről könnyen igazolható, hogy

$$1 < mci \leq pmci \leq m .$$

A két index értéke pontosan akkor egyezik meg, ha minden sajátérték megegyezik, azaz mindegyik értéke 1, ekkor mindkét index értéke *m*. Minél jobban közelít a nullához a legkisebb sajátérték, a mutatók értékei annál jobban közelítenek egyhez. THISTED állítása szerint az indexek egyhez közeli értékei szignifikáns multikollinearitást jeleznek. Azonban,

ez az állítás cáfolható. FELLMAN rámutatott arra, hogy ha egy olyan speciális korrelációs mátrixot tekintünk, amiben a tényezőváltozók korrelációs mátrixának minden főátlón kívüli eleme a , akkor a két indexre szigorúbb alsó határt adhatunk.¹⁰ Ekkor

$$m-1 < mci \leq pmci.$$

Például, három magyarázóváltozó esetén mindkét index értéke kettőnél nagyobb lesz. Márpedig, például ha $a=0,9$; akkor az erős multikollinearitás ellenére, a két index értéke meg sem közelíti az egyet. THISTED mérőszámai csak akkor adnak megfelelő képet a multikollinearitás mértékéről, ha legfeljebb egy darab nullához közeli sajátérték van.

FELLMAN példájával kapcsolatosan felmerül az a kérdés, hogy tetszőleges $0 < a \leq 1$ paraméter mellett egyáltalán léteznek-e ilyen speciális korrelációs mátrixok. A korrelációs mátrix szerkezetére vonatkozóan nincsenek általános érvényű eredmények. Azonban három, négy és öt változó esetére igen. Ezeknek a feltételeknek a megadása négy és öt változó esetén nagyon bonyolult, míg három változó esetén könnyen megadható: egy 3×3 -as korrelációs mátrixot az

$$\mathbf{R} = \begin{pmatrix} 1 & \cos \alpha & \cos \beta \\ \cos \alpha & 1 & \cos \gamma \\ \cos \beta & \cos \gamma & 1 \end{pmatrix}$$

alakban paraméterezzük¹¹, ahol $\alpha, \beta, \gamma \in [0;1]$ a változók, mint vektorok által bezárt szögek cosinus értékei. $\mathbf{R}$ pontosan akkor lesz korrelációs mátrix, ha egyszerre teljesülnek az alábbi feltételek (BOLLA – KRÁMLI, [2005]).

¹⁰ Ekkor a korrelációs mátrix sajátértékei: $\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = 1-a$; $\lambda_m = 1+(m-1)a$.

¹¹ A korrelációs mátrix ilyen típusú paraméterezésének alapja az, hogy standardizált változókat (zéróátlagú változókat), mint vektorokat tekintve, a változók által bezárt szög cosinusa megegyezik a lineáris korrelációs együttható értékével, a vektorok skaláris szorzatának kétféle felírásából következően.

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^n x_i y_i = |\mathbf{x}| \cdot |\mathbf{y}| \cdot \cos(\mathbf{x}'; \mathbf{y}) \rightarrow \cos(\mathbf{x}'; \mathbf{y}) = \frac{\sum_{i=1}^n x_i y_i}{|\mathbf{x}| \cdot |\mathbf{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}} = r_{xy}$$

$$\begin{aligned}
\alpha + \beta &\geq \gamma \\
\beta + \gamma &\geq \alpha \\
\alpha + \gamma &\geq \beta \\
\alpha + \beta + \gamma &\leq 2\pi
\end{aligned}$$

Megjegyzem, hogy a felvetett kérdés végül is lényegtelen, ugyanis az ellenpéldában kizárólag a sajátértékek nagysága lényeges és nem a korrelációs mátrix szerkezete. Mint a későbbiekben látni fogjuk, létezik ilyen „sajátérték-szett”. Azt viszont már nem állíthatjuk, hogy egy sajátérték-szethez csak egy korrelációs mátrix tartozik, hiszen egy tetszőleges $\mathbf{A}$ négyzetes mátrix és egy $\mathbf{B}^{-1}\mathbf{A}\mathbf{B}$ transzformáltjának a sajátértékei megegyeznek. Egy későbbi kutatás témája lehetne a korrelációs mátrixok egymásba transzformálhatóságának vizsgálata.

A VINOD (1976), WICHERN és CHURCHILL (1978) által adott *ISRM index* értéke 0 és $m(m-1)$ közé esik.

$$ISRM = \sum_{i=1}^m \left(\frac{\frac{m}{\sum_{j=1}^m \lambda_j} - 1}{\lambda_i} \right)^2$$

Az index kifejezhető az eddigi mutatók segítségével is.

$$ISRM = \sum_{i=1}^m \left(\frac{m}{\lambda_i M_1} - 1 \right)^2 = \frac{m^2 mci}{pmci^2} - m$$

Minél jobban távolodik a mutató értéke a nullától, annál erősebb a multikollinearitás mértéke. Azonban, az *mci* indexnél bemutatott példával ezt az állítást is cáfolhatjuk. Minél közelebb van az a paraméter értéke egyhez, annál nagyobb a multikollinearitás mértéke, viszont az *ISRM index* értéke $\frac{m}{m-1}$ -hez tart (FELLMAN, [1981]).

Összefoglalva megállapítható, hogy a sajátértékek reciprokait használó indikátorok nagy hátránya, hogy értelmezésük szubjektív, azaz nincs egy olyan egyértelmű küszöbszám, ami már erős multikollinearitást jelez. Továbbá ezen mutatók értékei főleg csak a legkisebb sajátértéktől függenek.

Az eddig ismertetett tesztek és mutatók összefoglalását a 7. táblázat tartalmazza.

7. táblázat: A multikollinearitás mérőszámainak és tesztjeinek összefoglaló táblázata

Mutató	Szintetikus-e?	Normált-e?	Eredményének értelmezése objektív-e?	Van-e egyértelmű különbértéke a káros multikollinearitásnak?	Cél
DEF-mutató	Igen	Igen	Nem	Nem	Detektálás
FELLMAN L - mutatója	Igen	Nem	Nem	Nem	Detektálás
gamma- mutató	Igen	Nem	Nem	Nem	Detektálás
ISRM-mutató	Igen	Nem	Nem	Nem	Detektálás
CI kondíciós index	Igen	Nem	Nem	Nem	Detektálás
M_1 -mutató	Igen	Nem	Nem	Nem	Detektálás
MAHAYAN és LAWLES M_1 mutatója	Igen	Nem	Nem	Nem	Detektálás
M -mutató	Igen	Nem	Nem	Nem	Detektálás
Red-mutató	Igen	Igen	Igen	Nem	Detektálás
THISTED mci és $pmci$ mutatója	Igen	Nem	Nem	Nem	Detektálás
VIF_j -mutató	Nem	Nem	Igen	Nem	Detektálás Lokalizálás
Tolerancia- mutató	Nem	Igen	Igen	Nem	Detektálás Lokalizálás
FARRAR– GLAUBER teszt	Igen	Nem	–	–	Detektálás
WILKS teszt	Nem	Nem	–	–	lokalizálás

(forrás: saját szerkesztés)

II. 3. A multikollinearitás negatív következményeinek csökkentése

A multikollinearitás jelenségének felismerése után felmerül az a kérdés, hogy hogyan szüntethetőek meg, illetve hogyan csökkenthetőek a multikollinearitás káros következményei.

Abban az esetben, ha a multikollinearitás jelenléte nem jelent problémát az adott vizsgálat szempontjából – például előrejelzések esetén – akkor a multikollinearitás mértéke csökkentésének érdekében semmit sem kell tennünk, a modell használható marad.

Ha a multikollinearitás problémát jelent, akkor megoldás lehet néhány **tényezőváltozó elhagyása**. Mivel a változók elhagyása után a regressziós paramétereket újra kell becsülni, ezért a paraméterek becsült értéke attól is függ, hogy mely változót, illetve változókat hagytuk el a modellből.

A magyarázóváltozók elhagyásával kapcsolatosan több probléma fogalmazható meg.

- Egyrészt, a változók elhagyása mindig információvesztéssel jár. Előfordulhat, hogy bizonyos változók elhagyásával nagymértékű lesz ez a veszteség.
- Másrészt, a vizsgálat szempontjából releváns változók elhagyása ugyan csökkentheti a multikollinearitás negatív következményeit, de fogyatékosná teszi az alkalmazott modellt. Ekkor a regressziós együtthatók becsült értékeinek értelmezése nem lesz valóságos.
- Harmadrészt, honnan tudjuk, hogy melyik magyarázóváltozót kell elhagyni? Ennek meghatározásához elengedhetetlen a multikollinearitás pontos lokalizálása. Ugyan bizonyos mutatók a multikollinearitást magyarázóváltozókhoz próbálják kötni, de ahogyan már korábban hangsúlyoztam, ezért a jelenségért nem mindig egy változó okolható, így többnyire nehézkes a lokalizálás.

Általában az abszolút értékben legkisebb t -értékkel rendelkező paraméterhez tartozó tényezőváltozót hagyjuk el, de létezik olyan eljárás is, amelyben mindig a legnagyobb VIF_j értékkel rendelkező változót vesszük ki a modellből. A változók elhagyásának végrehajtásánál figyelni kell arra, hogy a változókat kizárólag egyesével szelektáljuk. A statisztikai szoftverek többsége tartalmaz olyan modellépítési módszereket, úgynevezett *stepwise* eljárásokat, amelyek a változók szelektálásánál figyelembe veszik a t -értékeket, valamint – általában – a *tolerancia-mutató* változónkénti értékét is (HUNYADI – MUNDRUCZÓ – VITA, [1997]):

- a *forward* eljárás végrehajtása során először az a tényezőváltozó kerül a modellbe, amelyiknek az eredményváltozóval vett korrelációs együtthatójának abszolút értéke a legnagyobb. Ezután lépésenként növeljük a modell magyarázóerejét a tényezőváltozók beléptetésével egészen addig, amíg a reziduális négyzetösszeg szignifikánsan csökken.
- Az előzőn túl, a *stepwise* eljárás használatakor is a magyarázóváltozók egyesével lépnek be a modellbe, de ha az újabb belépők hatására (egy adott lépésben) a már bent levő változók közül valamelyikhez tartozó t -érték adott szint alá csökken, akkor az kilép a modellből.
- A *backward* módszer használatakor kezdetben minden tényezőváltozó beléptetésre kerül, majd a már bent levő változók közül azt hagyjuk el, amelyikhez a legkisebb t -érték tartozik.

A változószelekcióra vonatkozó állítás második fele azonban vitatható, ugyanis ezek eljárások elsősorban a modellek magyarázóerejének meghatározására optimalizáltak. Ebből következően előfordulhat hogy az eljárások végrehajtása után olyan modellt kapunk eredményül, amelyben már nincs szignifikáns multikollinearitás. Az azonban nem állítható, hogy az így kapott modell a multikollinearitás negatív hatásainak csökkentése szempontjából optimális lesz.

Az állítással kapcsolatosan egy – SOKAL és ROHLF által adott – érdekes ellenpéldát szeretnék megemlíteni (FENG-JENQ, [2006]). Klimatológusok a 70-es évek végén a levegő minőségének előrejelzésére egy elemzés során 41 amerikai nagyváros adatait vizsgálták. Az egyik részvizsgálat során a kéndioxid koncentrációját, mint eredményváltozót, hét magyarázóváltozó függvényében elemezték. Ekkor a 8. táblázatban szereplő lineáris regressziós modellt kapták.

A 8. táblázat adatai alapján megállapítható, hogy a multikollinearitás szempontjából népesség számát és a gyárak számát egyidejűleg nem célszerű a regressziós modellben szerepeltetni, ugyanis öthöz képest túlságosan nagy a VIF_j -mutatók értéke e két változó esetében. Ugyanakkor látható, hogy ennek ellenére mindkét változónál a t -statisztika abszolút értéke nem kisebb a kritikus értékhez (2,035) képest.

8. táblázat: A kéndioxid koncentrációjának becslése lineáris regressziós modellel

Változók	Becsült regressziós paraméterek	<i>t</i> -statisztika	<i>VIF_j</i>
Tengelymetszet	112,159	2,338	–
A népesség száma 1979-ben (ezer fő)	–0,039	–2,564	14,342
A legalább 20 főt foglalkoztató gyárak száma	0,064	4,008	14,883
Évi átlaghőmérséklet (F)	–1,282	–2,032	3,783
Évi átlagos szélesség	–3,222	–1,747	1,262
Évi átlagos csapadékmennyiség (inch)	0,497	1,340	3,465
Az esős napok évi átlagos száma	–0,048	–0,292	3,463
Porkoncentráció (ppm)	0,233	0,319	1,279

(forrás: FENG-JENQ, [2006])

Mi történik akkor, ha a regressziós modellt stepwise algoritmussal építjük fel? Ekkor a 9. táblázat adatait kapjuk.

9. táblázat: A kéndioxid koncentrációjának becslése lineáris regressziós modellel stepwise algoritmus alkalmazásával

Változók	Becsült regressziós paraméterek	<i>t</i> -statisztika	<i>VIF_j</i>
Tengelymetszet	26,325	6,855	–
A legalább 20 főt foglalkoztató gyárak száma	0,082	5,609	11,434
A népesség száma 1979-ben (ezer fő)	–0,057	–3,959	11,434

(forrás: FENG-JENQ, [2006])

A 9. táblázat adatai alapján látható, hogy a stepwise algoritmus mind a népesség számát, mind a gyárak számát szerepelteti magyarázóváltozóként, azaz a multikollinearitás jelensége nem szűnt meg. Ebből következően a stepwise algoritmus nem minden esetben szünteti meg a multikollinearitást.

Egy újabb megoldást jelenthet a **megfigyelések számának, a minta elemszámának növelése**. Ennél a módszernél a fő problémát az jelenti, hogy a minta elemszámának növelésével a változók közötti korreláció akárhogy változhat, így az is

előfordulhat, hogy egyáltalán nem tudjuk csökkenteni a multikollinearitás negatív következményeit. Idősorok vizsgálata esetén egy másik probléma is jelentkezik: nincs lehetőség a megfigyelések számának növelésére (MADDALA, [2004]).

Egy hasonló jellegű megoldás a **külső információk felhasználása**. Például fogyasztáselemzéseknél egyszerre keresztmetszeti és idősoros adatokat is felhasználnak. Például, TOBIN kutatásaiban a jövedelmi elaszticitások becslését keresztmetszeti, míg az ár rugalmassági együtthatókat idősoros adatok alapján számította ki. (PETRES – TÓTH, [2006]).

Habár általában a multikollinearitás negatív következményeit nem csökkenti, de technikailag – főleg akkor, amikor a korrelációs mátrix invertálása nehézségekbe ütközik – megoldást jelenthet az általánosított inverz mátrix, más néven a **MOORE – PENROSE inverz** alkalmazása. Az $\mathbf{X}_{n \times (m+1)}^+$ mátrix az $\mathbf{X}_{(m+1) \times n}$ mátrix általánosított inverze, ha teljesülnek az alábbi feltételek.

$$\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$$

$$\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$$

$$(\mathbf{X}\mathbf{X}^+)' \mathbf{X} = \mathbf{X}\mathbf{X}^+$$

$$(\mathbf{X}^+\mathbf{X})' = \mathbf{X}^+\mathbf{X}$$

A MOORE – PENROSE inverz segítségével megoldható az (1) egyenlet. Ekkor

$$\hat{\boldsymbol{\beta}}^* = \tilde{\mathbf{X}}^+ \tilde{\mathbf{y}} = \tilde{\mathbf{X}}^+ \tilde{\mathbf{X}} \tilde{\boldsymbol{\beta}} + \tilde{\mathbf{X}}^+ \tilde{\boldsymbol{\varepsilon}} .$$

A módszer használata egy hagyományos LNM becslést jelent (HEINCZINGER, [1983]).

Gyakran alkalmazott eljárás a standardizált tényezőváltozók mesterséges, ortogonális, azaz egymástól lineárisan független változóba, úgynevezett **főkomponensekbe** történő transzformálása. Ez az eljárás gyakorlatilag megegyezik az általánosított inverz módszer alkalmazásával. A főkomponensek a standardizált tényezőváltozók lineáris kombinációi, tehát a főkomponensek $\mathbf{Z}$ mátrixa felírható a $\mathbf{Z} = \tilde{\mathbf{X}}\mathbf{U}$ alakban, ahol $\mathbf{U}$ a korrelációs mátrix sajátértékeihez tartozó sajátvektorok

mátrixa. Mivel $\mathbf{U}^{-1} = \mathbf{U}'$, így $\tilde{\mathbf{X}} = \mathbf{Z}\mathbf{U}'$. Ezért az (1) egyenlet felírható az alábbi formában is.

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}} = \mathbf{Z}\mathbf{U}'\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}} = \mathbf{Z}\boldsymbol{\gamma} + \tilde{\boldsymbol{\varepsilon}}$$

A főkomponensek használata abban az értelemben megoldása a multikollinearitás problémájának, hogy a korrelált tényezőváltozók helyett magyarázóváltozókként korrelálatlan főkomponenseket használunk. Habár a főkomponensek a tényezőváltozók lineáris kombinációi, ennek ellenére mégsem érdemes a főkomponensek segítségével felírt regressziós modellbe ezeket a lineáris kombinációkat behelyettesíteni, mivel ekkor a multikollinearitás ugyanúgy megmarad a tényezőváltozók között. A főkomponensek használatának fő hátránya az, hogy nem tudjuk minden esetben megadni a főkomponensek jelentését. A jelentések megadásához – többek között – a loading változókat kell vizsgálnunk. Ezek segítségével megállapítható az, hogy mely főkomponensek mely tényezőváltozókval állnak szorosabb kapcsolatban. A főkomponensek jelentését ezen tényezőváltozók figyelembevételével kell megadni.

A HOERL és KENNARD (1970) nevéhez fűződő *ridge-regresszió* gyakorlatilag egy torzító paraméter ($k > 0$) alkalmazását jelenti: az $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ mátrixhoz hozzáadjuk az egységmátrix k -szorosát. Ekkor a regressziós paraméterek – a (2) egyenlet helyett – az alábbi formában becsülhetők.

$$\hat{\tilde{\boldsymbol{\beta}}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I})^{-1} \tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

A módszer legkényesebb kérdése a torzító paraméter megválasztása. Amennyiben a torzító paraméter értékét mindenhatáron túl növeljük, a becsült paraméterek értékei nullához tartanak. A regressziós paramétereknek a pozitív torzító paraméter használatával kapott becslése torzított ugyan, de bizonyítható, hogy mindig létezik olyan ($0 < k < 1$) érték, amelyre a becslés hatásosabb lesz, mint a legkisebb négyzetek módszerén alapuló becslés. HOERL azt javasolta, hogy a k értékét oly módon válasszuk meg, hogy az a lehető legkisebb érték legyen úgy, hogy az együtthatók becslései stabilak legyenek, azaz k változására a regressziós paraméterek becslései csak nagyon kis mértékben változzanak meg, vagyis stagnáljanak. Ezzel csak az a probléma, hogy a stabilitás fogalmát nagyon szubjektíven értelmezték. A k paraméter becslésére számos javaslat született. Ezek közül csak két, HOERL által is alkalmazott technikát említek meg.

A becsült regressziós együtthatókat a torzító paraméter függvényében ábrázolva megkapjuk a *ridge-görbét*.¹² A görbe alapján megsejthető k megfelelő értéke. Ez az eljárás bár nagyon szemléletes, nem egzakt.

HOERL javaslata szerint k megfelelő értékét az alábbi képlettel kaphatjuk meg.

$$k = m \frac{s_e^2}{\sum_{i=0}^m \hat{\beta}_i^2}$$

Adott k érték mellett a multikollinearitás szignifikáns voltára következtethetünk abból, ha a torzító paraméter kicsiny változására a becsült regressziós paraméterek nagyon megváltoznak, azaz instabil a becslés (HEINCZINGER, [1983]).

A ridge-regresszió alkalmazásával kapcsolatban, a stabilitás szubjektív meghatározásán kívül, más probléma is felmerül.

- Egyrészt, a módszer nem független a változók mértékegységeitől, illetve a lineáris transzformációjuktól. A mérési egységekből adódó probléma kiküszöbölhető úgy, hogy minden változót standardizálunk az eljárás előtt.
- Másrészt, a torzító paraméter alkalmazása miatt kérdéses a regressziós paraméterek helyes értelmezhetősége.

A felmerülő problémák miatt MADDALA nem is javasolja általános esetben a multikollinearitás problémájának megoldására a ridge-regressziót. MADDALA szerint olyan helyzetekben érdemes a ridge-regressziót alkalmazni, amikor a regressziós együtthatókról van valamilyen – például az előjeleikre, összegükre, négyzetösszegükre – előzetes információnk.

Az általánosított legkisebb négyzetek módszerére épülő *nested estimate*, azaz az *egymásba ágyazott becslések eljárás* során a regressziós együtthatókat lépésenként, egyesével becsüljük meg. Az eljárás végén adódó modellt *nested regression*, azaz *egymásba ágyazott regressziós modellnek* nevezzük. Az eljárás során először kiválasztjuk azokat a tényezőváltozókat, amelyek szignifikáns kapcsolatban állnak az

¹² A k torzító paraméter értékét a hatásosságra vonatkozó állítás miatt a (0;1) tartományon kell ábrázolni.

eredményváltozóval. A többi tényezőváltozót eleve kizárjuk a modellből. Ezután csak a megmaradt tényezőváltozókat használhatjuk.

Az első iteráció során válasszuk ki azt a tényezőváltozót, amelyiknek a legerősebb a kapcsolata az eredményváltozóval, azaz azt a tényezőváltozót szerepeltetjük, amelyikkel az eredményváltozó lineáris korrelációs együtthatója abszolút értékben a legnagyobb. Legyen ez a változó $\tilde{x}_1$. E két változó alapján alkalmazhatjuk az

$$\hat{y}_i = \hat{\beta}_{0,1} + \hat{\beta}_{1,1}\tilde{x}_{i,1} + \tilde{\varepsilon}_{i,1}$$

lineáris regressziós modellt, ahol a regressziós paraméterek második alsóindexe jelöli azt, hogy az adott paraméter hányadik iterációs lépésben adódik.

A második iterációban a megmaradt tényezőváltozók közül válasszuk ki azt, amelyik az $\tilde{\varepsilon}_{i,1} = \tilde{y}_i - \hat{y}_i$ hibataggal a legerősebben korrelál. Legyen ez a változó $\tilde{x}_2$. Ekkor felírhatjuk az

$$\hat{\varepsilon}_{i,1} = \hat{\beta}_{0,2} + \hat{\beta}_{1,2}\tilde{x}_{i,2} + \tilde{\varepsilon}_{i,2}$$

lineáris regressziós modellt. Ekkor t -próbával tesztelnünk kell a kapott $\hat{\beta}_{1,2}$ regressziós együtthatót ($H_0 : \tilde{\beta}_{1,2} = 0$). Ha a hipotézisvizsgálat során a nullhipotézist elfogadjuk, akkor az eljárás végeredménye az első iteráció során kapott regressziós modell lesz. Ellenkező esetben a 2. iterációban kapott egyenletet behelyettesítjük az 1. iteráció végén kapott regressziós egyenletbe:

$$\hat{y}_i = \hat{\beta}_{0,1} + \hat{\beta}_{1,1}\tilde{x}_{i,1} + \hat{\beta}_{0,2} + \hat{\beta}_{1,2}\tilde{x}_{i,2} + \tilde{\varepsilon}_{i,2} = \hat{\beta}_{0,1} + \hat{\beta}_{0,2} + \hat{\beta}_{1,1}\tilde{x}_{i,1} + \hat{\beta}_{1,2}\tilde{x}_{i,2} + \tilde{\varepsilon}_{i,2},$$

majd következik a 3. iteráció.

Általánosan a k -adik iteráció során az előző iterációban megmaradt tényezőváltozók közül válasszuk ki azt, amelyik az $\tilde{\varepsilon}_{i,k-1} = \tilde{y}_i - \hat{y}_i$ hibataggal a legerősebben korrelál. Legyen ez a változó $\tilde{x}_k$. Ekkor felírhatjuk az

$$\hat{\varepsilon}_{i,k-1} = \hat{\beta}_{0,k} + \hat{\beta}_{1,k}\tilde{x}_{i,k} + \tilde{\varepsilon}_{i,k}$$

lineáris regressziós modellt. Ekkor t -próbával tesztelnünk kell a kapott $\hat{\beta}_{1,k}$ regressziós együtthatót ($H_0 : \tilde{\beta}_{1,k} = 0$). Ha a hipotézisvizsgálat során a nullhipotézist elfogadjuk,

akkor az eljárás végeredménye a $(k-1)$. iterációban kapott regressziós modell lesz. Ellenkező esetben a k . iteráció során kapott egyenletet behelyettesítjük az előző iteráció végén kapott regressziós egyenletbe:

$$\hat{y}_i = \sum_{j=1}^k \hat{\beta}_{0,j} + \sum_{j=1}^k \hat{\beta}_{1,j} \tilde{x}_{i,j} + \tilde{\varepsilon}_{i,k},$$

majd, amennyiben maradt még tényezőváltozó, következik a $(k+1)$. iteráció, ellenkező esetben az eljárás végeredménye a k . iterációban kapott regressziós modell lesz (FENG-JENQ, [2006]).

Látható, hogy az eljárás lefuttatásával párhuzamosan lehetőség van a modell dimenziószámának csökkentésére.

Ha az eljárás során minden iterációs lépésben a k -adik hibatag független a k -adik tényezőváltozótól, akkor a multikollinearitás nem jelentkezik az eljárás végén kapott regressziós modellben.

Összefoglalva elmondható, hogy nincs olyan eljárás a multikollinearitás negatív következményeinek csökkentésére, mely általános érvényű lenne, azaz – más szempontból – minden eljárásnak lehetnek káros mellékhatásai.

II. 4. Az ismertett mutatók és eljárások empirikus szemléltetése

A fejezet zárásaként egy empirikus példán fogom szemléltetni a multikollinearitás mérőszámait, felismerési módjait és a negatív hatásainak csökkentésére bemutatott eljárásokat. Mivel fogyasztáselemzések során a multikollinearitás szinte kivétel nélkül mindig jelen van, ezért az alábbi példát vettem.

Az 1 főre jutó évi marhahúsfogyasztást (y , kg/fő), mint eredményváltozót vizsgáltam 1990 és 2004 között az alábbi tényezőváltozók függvényében:

- egy főre jutó éves reáljövedelem 2004-es árszínvonalon (ezer Ft/fő);
- sertéshús, comb, csont és csülök nélkül (Ft/kg);
- marhahús, rostélyos, csontos (Ft/kg);
- sertésszár (Ft/kg);

- napraforgó-étolaj (Ft/liter);
- kenyér, fehér (Ft/kg);
- sör, hazai világos (Ft/0,5 liter);

A vizsgálat időtartományát az indokolta, hogy egyrészt csak rendszerváltozás utáni kiskereskedelmi évi átlagárakat akartam felhasználni, másrészt a fogyasztások vizsgálatánál 2004 után módszertanváltás következett be.

Az alapadatok 1996-tól a Központi Statisztikai Hivatal adat-tábláiból, míg a korábbi évekre vonatkozó adatok a Magyar statisztikai évkönyvekből származnak.¹³

Mivel az árak a különböző években más és más árszínvonalon vannak megadva, ezért ezek összehasonlíthatósága végett, az elemzés megkezdése előtt az adatokat deflálni kellett. Az elemzésben minden ár 2004-es árszínvonalon van megadva.

A vizsgált modellel a céloom elsősorban az, hogy szemléltessem az előző alfejezetek mérőszámait és mutatóit.

Az adattáblák és a számítógépes kimenetek a méretük miatt a VI. 4. mellékletben találhatóak. A dolgozatban csak azoknak a tábláknak a közlésére szorítkozom, melyek a kitűzött célok szempontjából fontosak.

A számítógépes elemzéshez az SPSS 13.0 verzióját, illetve a Microsoft Excel programot használtam.

A korrelációs mátrix (VI. 4. melléklet) felírása után az adatokra lineáris regressziós modellt illeszttem. Az illesztett modellben a multikollinearitás vizsgálatának és kezelésének módszereit ugyanolyan sorrendben szemléltetem, mint amilyenben a fejezetben bemutatam ezeket.

A magyarázóváltozók korrelációs mátrixában a sör és a sertészsír ára közötti $-0,758$ értékű korrelációs együttható esik a legközelebb a többszörös korrelációs együttható $0,957$ értékéhez. Az elemző szubjektív döntése az, hogy ez a két érték mennyire van közel egymáshoz. Ezért a KLEIN-féle hüvelykujjszabály szerint valószínűsíthető, hogy a multikollinearitás jelenlétével számolni kell a modellben.

¹³ Az elektronikus adatforrásokból az adatok 2007. július 20-án lettek letöltve.

MASON és PERREAULT módszere szerint is a multikollinearitás erős voltára következtethetünk, ugyanis a modellben – amelyben az eredményváltozó az egy főre jutó reáljövedelem, a tényezőváltozó pedig a többi 7 változó – a többszörös determinációs együttható értéke 0,948. Ez az érték pedig nagyobb az eredeti modell determinációs együtthatójának értékénél (0,917).

Az M_1 -mutató értékét a magyarázóváltozók és az eredményváltozó közötti korrelációs együtthatók (10. táblázat) alapján határozhatjuk meg. Mivel a magyarázóváltozók és az eredményváltozó közötti korrelációs együtthatók négyzetösszege eltér a többszörös determinációs együttható (0,917) értékétől, azaz $M_1=0,648$ ezért a mutató értéke multikollinearitást jelez.

10. táblázat: A magyarázóváltozók és az eredményváltozó közötti korrelációs együtthatók

	Korrelációs együtthatók
Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	– 0.400
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0.713
Marhahús, rostélyos, csontos (Ft/kg)	0.450
Sertésszár (Ft/kg)	– 0.042
Napraforgó-étolaj (Ft/liter)	0.458
Kenyér, fehér (Ft/kg)	0.664
Sör, hazai világos (Ft/0,5 liter)	0.201
<i>Négyzetösszeg</i>	1.565

(forrás: saját szerkesztés)

Az M -mutató kiszámításához szükségünk van az összes, 6 tényezőváltozós modell többszörös determinációs együtthatójára. Ezeket az adatokat tartalmazza a 11. táblázat.

$M=0,917-0,127=0,79$. Ezek szerint, a modellben erős a multikollinearitás.

Mivel a tényezőváltozók korrelációs mátrixának determinánsa: 0,005 nullához közeli, ezért sejthetjük a multikollinearitás szignifikáns voltát.

Viszont, a korrelációs mátrix determinánsának tesztelésekor a próbafüggvény értéke:
 $\chi^2 = 24,93$.

Ekkor ötszázalékos szignifikanciaszint mellett elfogadjuk a nullhipotézist, azaz a magyarázóváltozók lineárisan függetleneknek tekinthetők. Ez azonban nem jelenti azt, hogy a modellben nem kell számolnunk a multikollinearitás fellépésével.

11. táblázat: Az M -mutató kiszámításának mellékszámításai

Kimaradó tényezőváltozó	Többszörös determinációs együttható	A tényezőváltozó növekményi hozzájárulása a többszörös determinációs együtthatóhoz
Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	0,907	0,01
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,909	0,008
Marhahús, rostélyos, csontos (Ft/kg)	0,902	0,015
Sertézsír (Ft/kg)	0,916	0,001
Napraforgó-étolaj (Ft/liter)	0,917	0,000
Kenyér, fehér (Ft/kg)	0,826	0,091
Sör, hazai világos (Ft/0,5 liter)	0,915	0,002
<i>Összesen</i>	–	0,127

(forrás: saját szerkesztés)

A magyarázóváltozók korrelációs mátrixának inverzét vizsgálva, a diagonális elemek segítségével felírhatjuk az ω próbafüggvény értékeit, amelyeket a 12. táblázat tartalmaz.

12. táblázat: Az ω -próbafüggvény értékei

Tényezőváltozó	ω_i	Szignifikanciaszint
Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	21,503	0,001
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	8,314	0,013
Marhahús, rostélyos, csontos (Ft/kg)	11,380	0,007
Sertézsír (Ft/kg)	5,773	0,028
Napraforgó-étolaj (Ft/liter)	8,390	0,013
Kenyér, fehér (Ft/kg)	3,381	0,071
Sör, hazai világos (Ft/0,5 liter)	26,445	0,001

(forrás: saját szerkesztés)

A próba nullhipotézisét a *kenyér árának* kivételével minden tényezőváltozó esetén elvetjük, azaz, az adott magyarázóváltozó és a többi között szignifikáns a multikollinearitás.

A 13. táblázat alapján, összehasonlítva az eredményváltozó, azaz az egy főre jutó marhahúsfogyasztás és a tényezőváltozók közötti lineáris korrelációs együtthatókat a parciális korrelációs együtthatókkal, megállapítható, hogy ezek értékei a *sertéshús* és az *étolaj árának* esetében jelentősen eltérnek egymástól. Ebből is sejthető a multikollinearitás szignifikáns volta.

13. táblázat: A tényezőváltozók és eredményváltozó közötti korrelációs és parciális korrelációs együtthatók

Tényezőváltozó	Korrelációs együtthatók	Parciális korrelációs együtthatók
Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	– 0.400	–0,318
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0.713	0,289
Marhahús, rostélyos, csontos (Ft/kg)	0.450	0,390
Sertészsír (Ft/kg)	– 0.042	–0,096
Napraforgó-étolaj (Ft/liter)	0.458	0,007
Kenyér, fehér (Ft/kg)	0.664	0,723
Sör, hazai világos (Ft/0,5 liter)	0.201	0,141

(forrás: saját szerkesztés)

A *kenyér árának* kivételével minden tényezőváltozó esetén a VIF_j értékek nagyobbak ötnél, így megint csak szignifikáns multikollinearitást feltételezhetünk (14. táblázat).

A normált tényezőváltozók alapján számított kondíciós indexek (50,80; 129,93; 278,32) és gamma-mutató (278,32) értékei alapján megállapíthatjuk, hogy szignifikáns multikollinearitással kell számolnunk. Továbbá a becsült regressziós paraméterek felfújódását vizsgálva sejthetjük, hogy a multikollinearitásért a *jövedelem*, a *marhahús ára* és a *sör árának* a modellben való együttes szerepeltetése okolható (15. táblázat).

14. táblázat: A tolerancia-mutató és a VIF_j -mutató értékei

Tényezőváltozó	Tolerancia-mutató	VIF_j -mutató
Egy főre jutó éves jövedelem (2004-es árszínvonalon e Ft)	0,058	17,127
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,138	7,235
Marhahús, rostélyos, csontos (Ft/kg)	0,105	9,535
Sertésszár (Ft/kg)	0,188	5,330
Napraforgó-étolaj (Ft/liter)	0,137	7,292
Kenyér, fehér (Ft/kg)	0,283	3,536
Sör, hazai világos (Ft/0,5 liter)	0,048	20,834

(forrás: saját szerkesztés)

15. táblázat: A kondíciós indexek és a varianciák felfűződése

Sajátértékek	Kondíciós index	Részesezés a varianciák felfűződéséből							
		Konstans	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	Marhahús, rostélyos, csontos (Ft/kg)	Sertésszár (Ft/kg)	Napraforgó-étolaj (Ft/liter)	Kenyér, fehér (Ft/kg)	Sör, hazai világos (Ft/0,5 liter)
7,9151	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,0382	14,39	0,00	0,00	0,08	0,00	0,00	0,00	0,00	0,00
0,0226	18,70	0,00	0,00	0,01	0,00	0,13	0,00	0,00	0,00
0,0114	26,33	0,00	0,01	0,06	0,00	0,00	0,10	0,00	0,00
0,0090	29,68	0,00	0,00	0,03	0,03	0,00	0,01	0,12	0,00
0,0031	50,80	0,00	0,01	0,15	0,15	0,02	0,00	0,18	0,01
0,0005	129,93	0,68	0,15	0,00	0,01	0,15	0,28	0,07	0,00
0,0001	278,32	0,31	0,83	0,67	0,80	0,70	0,60	0,62	0,99

(forrás: saját szerkesztés)

A korrelációs mátrix sajátértékeire épülő mutatók meghatározásához szükséges részeredményeket a 16. táblázat tartalmazza.

A korrelációs mátrix sajátértékeire épülő mutatók értékeit pedig a 17. táblázat tartalmazza.

16. táblázat: A tényezőváltozók korrelációs mátrixának sajátértékei

	Sajátértékek	A sajátértékek reciprokai	A sajátértékek reciprokainak négyzetei
1	2,615	0,382	0,146
2	1,814	0,551	0,304
3	1,147	0,872	0,761
4	0,929	1,076	1,158
5	0,273	3,657	13,376
6	0,205	4,885	23,866
7	0,017	59,464	3535,992
Összesen	7,000	70,889	3575,603

(forrás: saját szerkesztés)

17. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei

Mutató	Érték
γ	12,471
L	3504,715
M_I	70,889
mci	1,011
$pmci$	1,192
$ISRM$	27,865
Red	0,359

(forrás: saját szerkesztés)

Látható, hogy az L , az M_I , az $ISRM$ mérőszámok értékei – a maguk módján – a nullához képest távolinak, míg az mci és a $pmci$ értékei egyhez közelinek mondhatóak, így ezek erős multikollinearitást jeleznek. Azonban, mint látható, az egyes mutatók értékei más és más nagyságrendűek, így mindegyiknél mást és mást jelent a „távoli” kifejezés. **Ebből kifolyólag ezeknek a mutatóknak az értékei egymással közvetlenül nem összehasonlíthatók.** A γ -mutató értéke egyhez képest, nem olyan magas, mint amennyire várnánk. A Red -mutató értelmezése szerint a vizsgált adatállományban, – a regressziós együtthatók becslőfüggvénye alkalmazásának szempontjából – a hasznos tartalmat hordozó adatok aránya 64,1 százalék a maximális információtartalmú, adott méretű adatállományhoz viszonyítva.

A DEF -mutató kiszámításához szükségünk van a standardizált regressziós együtthatók becsült értékeire. Ezeket a számítógépes kimenet *Regressziós paraméterek* táblázat *Standardizált regressziós paraméterek* oszlopa tartalmazza (VI. 4. melléklet). Ekkor $DEF = 0,78$; ami szignifikáns multikollinearitásra utal.

A multikollinearitás jelenlétének felismerése után, kérdés, hogy hogyan lehet ennek negatív következményeit csökkenteni.

Mivel idősoros adatok alapján történik az elemzés, ezért a megfigyelések számát nem tudjuk növelni. Külső információkat nem tudunk felhasználni, mivel ezek nem állnak rendelkezésre.

A tényezőváltozók szelektálásának hatását három módon vizsgáltam. Kipróbáltam mind az SPSS stepwise, mind a backward szelekcióját, majd harmadik módszerként pedig a VIF_j -mutató értékeit használtam a tényezőváltozók körének kialakítására.

1. Az SPSS stepwise modellépítési eljárása két modellt is felkínál, de a probléma az, hogy gazdaságilag egyik modell sem túl értelmes, mert a vizsgálat szempontjából releváns változók maradtak ki mind a két esetben. A VIF_j -mutatók szerint mindkét modell megfelelő a multikollinearitás szempontjából, ha elfogadjuk azt a feltevést, hogy a mutató öt feletti értéke jelez erős multikollinearitást (18. táblázat).

18. táblázat: VIF_j értékek a stepwise eljárás eredményeül adódó regressziós modellekben

Modell	Változó	VIF_j
1	Konstans	1,00
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	
2	Konstans	1,02
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	
	Kenyér, fehér (Ft/kg)	1,02

(forrás: saját szerkesztés)

A korrelációs mátrix sajátértékeire épülő mutatók alapján is ugyanerre a megállapításra juthatunk (19. táblázat).

19. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei a 2. modellben

Mutató	Érték
γ	1,159
L	0,090
M_I	2,044
mci	1,554
$pmci$	1,744
$ISRM$	0,043
Red	0,146

(forrás: saját szerkesztés)

2. A tényezőváltozók elhagyásának vizsgálatához másodszor az SPSS backward modellépítési eljárását használtam.

20. táblázat: VIF_j értékek a backward eljárás eredményéül adódó modellekben

Modell	Változó	VIF_j
1	Konstans	
	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	3,422
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	4,973
	Marhahús, rostélyos, csontos (Ft/kg)	4,104
	Sertészsír (Ft/kg)	4,202
	Kenyér, fehér (Ft/kg)	1,764
	Sör, hazai világos (Ft/0,5 liter)	8,484
2	Konstans	
	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	2,622
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	3,426
	Marhahús, rostélyos, csontos (Ft/kg)	2,915
	Kenyér, fehér (Ft/kg)	1,404
	Sör, hazai világos (Ft/0,5 liter)	2,479
3	Konstans	
	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	1,588
	Marhahús, rostélyos, csontos (Ft/kg)	1,036
	Kenyér, fehér (Ft/kg)	1,399
	Sör, hazai világos (Ft/0,5 liter)	1,971

(forrás: saját szerkesztés)

Az eljárás eredményeként – az eredeti modellen túl – három másik modellt kapunk eredményül. Az első modellből a *napraforgó-étolajár*, a másodikból a *napraforgó-étolajár* mellett a *sertésszár* is, míg a harmadik modellből ezen két tényező mellett a *sertéshúsár* is hiányzik.

Az eljárás eredménye szerint ugyan minden modellben egyre kisebb mértékű a multikollinearitás, de sajnos – a harmadik modell kivételével – a VIF_j -mutató értékei között még mindig előfordul ötnél nagyobb, azaz a modellben még mindig szignifikáns a multikollinearitás (20. táblázat). Ugyanerre következtethetünk a korrelációs mátrix sajátértékeire épülő mutatók értékei alapján (21. táblázat).

21. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei

Mutató	1. modell	2. modell	3. modell
γ	6,780	3,950	2,404
L	376,175	61,320	7,913
M_l	26,949	12,847	5,993
mci	1,046	1,129	1,207
$pmci$	1,373	1,585	1,765
$ISRM$	13,982	11,178	9,937
Red	0,356	0,335	0,289

(forrás: saját szerkesztés)

- Harmadszor, a tényezőváltozókat a VIF_j -mutató értékei alapján szelektáltam. Első lépésben a *sörárát* hagytam el a modellből, mivel ehhez a változóhoz tartozott a mutató legnagyobb értéke. Ekkor az első lépés után olyan modellt kaptam, amelyben – a mutató szerint – a multikollinearitás mértéke már nem káros (22. táblázat). Azonban, a kapott modellben nem mindegyik tényezőváltozóhoz tartozó t -érték megfelelő, azaz bizonyos tényezőváltozók szerepeltetése felesleges a modellben.

A multikollinearitás már nem káros mértékére következtethetünk a korrelációs mátrix sajátértékeire épülő mutatók értékei alapján (23. táblázat).

22. táblázat: A VIF_j értékek és a tényezőváltozók szignifikanciái a szelekciós eljárás eredményéül adódó regressziós modellben

Változó	Szignifikancia	VIF_j
Konstans	0,265	–
Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	0,220	3,254
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,098	2,763
Marhahús, rostélyos, csontos (Ft/kg)	0,142	2,566
Sertésszár (Ft/kg)	0,312	1,650
Napraforgó-étolaj (Ft/liter)	0,623	2,969
Kenyér, fehér (Ft/kg)	0,001	1,206

(forrás: saját szerkesztés)

23. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei

Mutató	Érték
γ	4,029
L	53,971
M_I	14,408
mci	1,460
$pmci$	2,105
$ISRM$	5,858
Red	0,331

(forrás: saját szerkesztés)

Összességében elmondható, hogy a változószelekció után kapott modellek, vagy nem szüntették meg a multikollinearitás szignifikáns voltát, vagy az eredményváltozó szempontjából életszerűtlenné váltak a modellek: releváns változók maradtak ki a modellből.

A főkomponensanalízis használata standardizált tényezőváltozókon történik. A számítógépes kimenet szerint (VI. 4. melléklet) a tényezőváltozók három főkomponenssel helyettesíthetők jelentős információtartalom-veszteség nélkül, hiszen három sajátérték értéke lesz egynél nagyobb és ezek együttesen 79,7 százalékát teszik ki a sajátértékek összegének. Tehát a három főkomponens együttesen a tizenhárom tényezőváltozó információtartalmának 79,7 százalékát őrzi meg. A fő probléma a főkomponensek tartalmának megadása, mivel az első és a második főkomponensekre olyan változók illeszkednek jól, amelyeknek nincs közös vonása (24. táblázat).

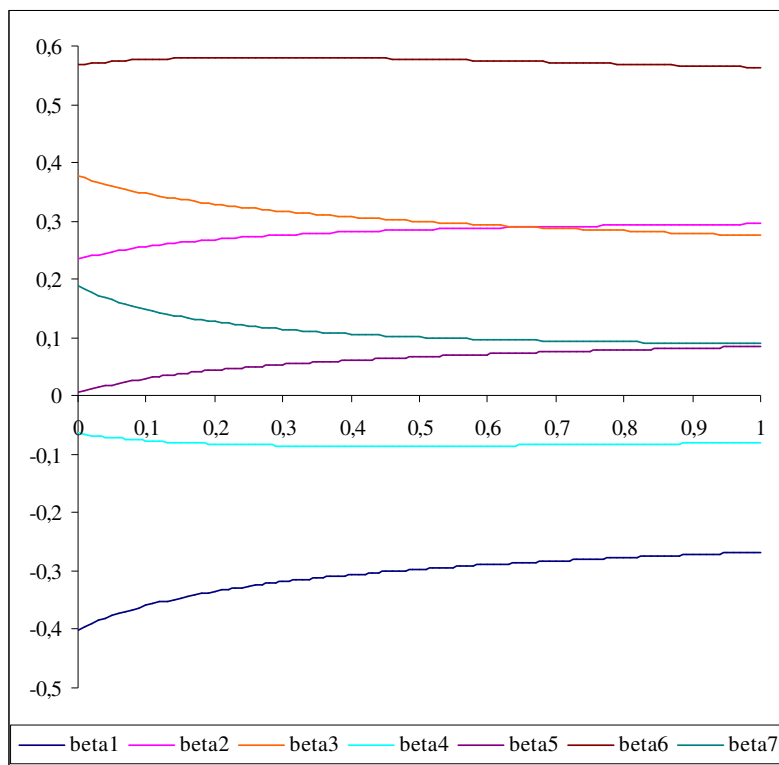
24. táblázat: Főkomponenssúly-mátrix rotálás után

Változók	Főkomponensek		
	1	2	3
Zscore: Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	–,412	,758	–,371
Zscore: Sertéshús, comb, csont és csülök nélkül (Ft/kg)	,890	–,046	,076
Zscore: Marhahús, rostélyos, csontos (Ft/kg)	,807	,104	–,419
Zscore: Sertésszár (Ft/kg)	–,024	–,845	–,126
Zscore: Napraforgó-étolaj (Ft/liter)	,758	–,218	,198
Zscore: Kenyér, fehér (Ft/kg)	,025	,188	,834
Zscore: Sör, hazai világos (Ft/0,5 liter)	–,024	,875	,408

(forrás: saját szerkesztés)

Szintén standardizált változókon alkalmaztam a ridge-regressziót.

3. ábra: Ridge-görbe



(forrás: saját szerkesztés)

A ridge-görbe (3. ábra) alapján sejthető, hogy a torzító paraméter értékét nincs értelme 0,2-nél kisebbnek választani, mert ekkor – értékének kicsiny változásához képest – a becült

regressziós együtthatók közül néhány jóval nagyobb mértékben változik, ez pedig a multikollinearitás jelenlétére utal. Azonban megjegyzem, hogy a diagram a két tengely léptékének különbözősége miatt csalóka lehet.

Összegzésül elmondható, hogy bármilyen pozitív értéket is választunk a torzító paraméternek, a kapott regressziós együtthatókat nem tudjuk helyesen értelmezni.

A nested estimate eljárás során a tényezőváltozók lehetséges körét tekintve csak a sertéshús és a kenyér ára jöhet szóba magyarázóváltozóként.

Az első iterációban kiválasztjuk a sertéshús árát, mert ennek a legerősebb a kapcsolata az eredményváltozóval. Mivel az egy főre jutó marhahúsfogyasztás és a sertéshús árára illesztett modellben a regressziós együttható t -tesztje során a nullhipotézist elvetjük, ezért következik a második iteráció.

A második iterációban, a megmaradt tényezőváltozót, azaz a kenyér árát választjuk. Ebben az iterációban is elvetjük a t -próba nullhipotézisét, ezért, mivel a használható tényezőváltozók köre elfogyott az eljárás véget ért. Tehát az eljárás végeredménye a második iteráció után egymásba ágyazott regressziós modell lesz (25. táblázat).

25. táblázat: A nested estimate eljárás után kapott regressziós modell

Tényezőváltozó	Becsült regressziós paraméterek
Konstans	$-0,09-9,957 = -9,966$
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,05
Kenyer, fehér (Ft/kg)	0,59

(forrás: saját szerkesztés)

Az eljárás ugyanazokat a változókat tette a modellbe, mint a stepwise algoritmus 2. modellje. Azonban, itt is probléma, hogy a vizsgálat szempontjából releváns változók kerültek ki a modellből. Viszont, káros multikollinearitás nem lép fel a modellben, de a két változó között gyenge együttmozgás létezik.

A fejezet összefoglalásaként az alábbi megállapításokat tehetjük:

- a multikollinearitás negatív következményeként leggyakrabban említik a becsült paraméterek varianciáinak növekedést, azonban ezeknek nem az abszolút nagyságát kellene nézni, hanem azt, hogy mekkora ezeknek a „felfújódása” a hibatag varianciájához képest.
- A multikollinearitás detektálásának és mérésének számos módja ismert, azonban ezek közül kevés a széles körben elfogadott mivel, egyrészt a multikollinearitás detektálása sokszor nagyon nehéz feladat, másrészt a mutatók többségének értelmezése meglehetősen szubjektív. A mérőszámok, eljárások egy része általában csak detektálják a multikollinearitást, de – általában szintetikus jellegük miatt – nem lokalizálják a problémát. Ezzel szemben a mérőszámoknak és eljárásoknak egy csoportja – több-kevesebb sikerrel – megpróbálja lokalizálni a multikollinearitást.
- A sajátértékek reciprokait használó indikátorok nagy hátránya, hogy értelmezésük szubjektív, azaz nincs egy olyan egyértelmű küszöbszám, ami már erős multikollinearitást jelez. A mutatók értékei egymással nem összehasonlíthatóak. Továbbá ezen mutatók értékei főleg csak a legkisebb sajátértéktől függnék.
- A bemutatott mérőszámok más és más szempontból jellemzik a multikollinearitást.
- Nincs olyan eljárás a multikollinearitás negatív következményeinek csökkentésére, mely általános érvényű lenne, azaz – más szempontból – minden eljárásnak lehetnek káros mellékhatásai.
- A ismertetett és alkalmazott mutatók, gondolatok, algoritmusok összegzéseként elmondható, hogy az említett mérőszámok és eljárások nem általános érvényűek abban az értelemben, hogy a multikollinearitás jelenségét csak speciális esetekben jellemzik, illetve kezelik megfelelően.

III. A PETRES-féle *Red*-mutató vizsgálata

A szakirodalom áttekintése után, ebben a fejezetben részletesen ismertetem vizsgálataimat és ezek eredményeit. Mivel az egyes területeken más és más módszertan került alkalmazásra, ezért ezeket az adott vizsgálatoknál fogom részletesen ismertetni. Elöljáróban annyit, hogy a vizsgálatok egy része elméleti megfontolásokon alapszik, míg egy másik részéhez különböző mintákat kellett létrehozni és ezek eredményeit elemezni. Az elemzésekhez az SPSS 13.0; illetve Microsoft Excel programokat használtam. A geometriai ábrázolás elkészítése a Derive 6.0 segítségével történt.

Először kísérletet teszek a *Red*-mutató meghatározására a tényezőváltozók korrelációs mátrixának sajátértékei nélkül.

III. 1. A *Red*-mutató számszerűsítése a sajátértékek ismerete nélkül

Vizsgálatom eredeti célja az volt, hogy a korrelációs mátrix sajátértékeit valamilyen egyszerű képlettel kifejezzem a korrelációs mátrix elemeiből. Ezáltal a *Red*-mutató egy olyan kiszámítási módját kaptam volna, amelyiknél nem szükséges a sajátértékek ismerete. A korrelációs mátrix sajátértékei a mátrix karakterisztikus egyenletének gyökei, és ezek meghatározására nincs ismert, általános érvényű összefüggés. Már a 3×3 -as korrelációs mátrixok sajátértékeire vonatkozó összefüggések is összetettek. Ezért – a *Red*-mutató képletéből adódóan – a sajátértékek négyzetösszegét próbáltam meg kifejezni közvetlenül a korrelációs mátrix elemei segítségével.

Mivel

$$(11) \quad \text{tr}(\mathbf{R}^2) = \sum_{i=1}^m \mathbf{R}_{ii}^2 = \sum_{i=1}^m \lambda_i^2,$$

ezért határozzuk meg a korrelációs mátrix négyzetének főátlójában álló elemeket. Felhasználva a korrelációs mátrix szimmetrikus voltát azt kapjuk, hogy:

$$(12) \quad \mathbf{R}_{ii}^2 = (\mathbf{R}\mathbf{R})_{ii} = \sum_{j=1}^m r_{ij}r_{ji} = \sum_{j=1}^m r_{ij}^2,$$

azaz a korrelációs mátrix négyzetének főátlójában lévő elemek a korrelációs mátrix megfelelő sorában álló elemek négyzetösszegeként adódik. Mivel a korrelációs mátrix főátlójában minden elem egy, ezért a (12) képlet az alábbi formában is felírható.

$$(13) \quad \mathbf{R}_{ii}^2 = (\mathbf{R}\mathbf{R})_{ii} = \sum_{j=1}^m r_{ij} r_{ji} = 1 + \sum_{\substack{j=1 \\ j \neq i}}^m r_{ij}^2$$

A (13) képletet (11) képletbe behelyettesítve azt kapjuk, hogy

$$(14) \quad \sum_{i=1}^m \lambda_i^2 = \sum_{i=1}^m \mathbf{R}_{ii}^2 = \sum_{i=1}^m \left(1 + \sum_{\substack{j=1 \\ j \neq i}}^m r_{ij}^2 \right) = m + \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m r_{ij}^2 .$$

A *Red*-mutató (5) képletét átrendezve kapjuk:

$$(15) \quad Red = \frac{v_{\lambda}}{\sqrt{m-1}} = \frac{\sigma_{\lambda}}{\sqrt{m-1}} = \frac{\sqrt{\frac{\sum_{i=1}^m \lambda_i^2}{m} - (\bar{\lambda})^2}}{\sqrt{m-1}} = \frac{\sqrt{\frac{\sum_{i=1}^m \lambda_i^2}{m} - 1}}{\sqrt{m-1}} = \sqrt{\frac{\sum_{i=1}^m \lambda_i^2 - m}{m(m-1)}} .$$

A (14) képlet (15) képletbe való behelyettesítésével adódóan

$$(16) \quad Red = \sqrt{\frac{\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m r_{ij}^2}{m(m-1)}} .$$

A (16) képlet azt jelenti, hogy sikerült igazolnom az 1. állításomat, mely szerint a *Red*-mutató kiszámítható a tényezőváltozók korrelációs mátrixa főátlón kívüli elemeinek négyzetes átlagaként is.

Az összefüggés abból a szempontból érdekes, hogy a *Red*-mutató egy olyan négyzetes átlag, amely – a definíciójából következően – százalékban is kifejezhető.

A (16) képlet szerint a *Red*-mutatóval mérni lehet a tényezőváltozók átlagos együttmozgásának mértékét.¹⁴ A mutató definíciójából és a (16) képletből következik,

¹⁴ A II. 4. alfejezet empirikus példáján *Red*=0,359. Ez azt jelenti, hogy a vizsgált adatállományban a tényezőváltozók átlagos együttmozgásának mértéke 35,9 százalék.

hogy a mutató előnye a többi sajátértékekre épülő mutatóval szemben az, hogy úgy veszi figyelembe az összes sajátértéket, hogy értékét minden sajátérték azonos súllyal befolyásolja, továbbá figyelembe veszi a tényezőváltozók összes páronkénti együttmozgását is, így a *Red*-mutató mindenképpen pozitív elmozdulást jelent a multikollinearitás eddigi kutatásához képest. A mutató segítségével megkülönböztethetők az extrém multikollinearitás különböző esetei is, hiszen a mutató akkor is használható, ha valamelyik sajátérték nulla.

Azonban megjegyzem, hogy a multikollinearitás vizsgálatakor nem csak változópárok együttmozgása, hanem változócsoporthoz együttmozgása is problémát jelenthet, ennek azonban még nincs részletesen kidolgozott szakirodalma. Ezek a vizsgálatok későbbi kutatásaim feladata lesz. **Ere megoldást jelenthet a kanonikus korrelációelemzés használata**, ahol valamilyen korrelációs együtthatók négyzetes átlaga szerepel az *RI redundancia-indexben* is, de alkalmazási körét és tartalmát tekintve ez teljesen más, mint a *Red*-mutató.

A *redundancia-indexet* a *kanonikus korrelációelemzés* során alkalmazzuk. A kanonikus korrelációelemzés a lineáris korrelációvizsgálat általánosításának tekinthető. A kanonikus korrelációelemzés során adott az $x_1, x_2, \dots, x_p$ és $y_1, y_2, \dots, y_q$ ($q \leq p$) két standardizált változócsoporthoz. A feladat az, hogy mindkét változócsoporthoz helyettesítsük a változók különböző u_t, z_t ($t=1,2,\dots,q$) lineáris kombinációival úgy, hogy az u_t, z_t kanonikus változópárok közötti r_t korrelációs együttható maximális legyen.¹⁵ Ezeket a korrelációkat kanonikus korrelációknak nevezzük. A kanonikus változók közötti korrelációs mátrix szerkezete az alábbi.

¹⁵ A kanonikus korrelációelemzés efféle megközelítése gyakorlatilag kettős faktoranalízisnek tekinthető, mivel két változóhalmazt azon faktorait keressük, amelyek maximálisan korrelálnak egymással. A kanonikus korrelációelemzés másfajta megközelítése az, hogy változók egy csoportjával próbáljuk a függőváltozók egy csoportját megmagyarázni, azonban ez nem a megfigyelt változókon keresztül történik, hanem a magyarázóváltozók azon lineáris kombinációja segítségével, amely maximálisan megmagyarázza a függőváltozókat, azok lineáris kombinációján keresztül (FÜSTÖS – KOVÁCS – MESZÉNA – SIMONNÉ, [2004]).

$$\mathbf{R} = \begin{array}{c|ccc|ccc} & u_1 & \dots & u_q & z_1 & \dots & z_q \\ \hline u_1 & 1 & 0 & 0 & r_1 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & 0 & \ddots & 0 \\ u_q & 0 & 0 & 1 & 0 & 0 & r_q \\ \hline z_1 & r_1 & 0 & 0 & 1 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & 0 & \ddots & 0 \\ z_q & 0 & 0 & r_q & 0 & 0 & 1 \end{array}$$

Ekkor az y változók szórásnégyzetét a z_t kanonikus változó átlagosan

$$r_{yz_t}^2 = \frac{\sum_{i=1}^q r_{y_i z_t}^2}{q}$$

mértékben, míg az u_t kanonikus változó

$$RI_{yz_t} = r_{yz_t}^2 r_{z_t u_t}^2$$

mértékben magyarázza (HAJDU, [2003]).

Tehát, a kanonikus korrelációelemzések során az eredeti változók és az ezeket helyettesítő valamelyik kanonikus változó közötti korrelációs együtthatók négyzetes átlagának négyzete használatos. Ezzel szemben a *Red*-mutató képletében a tényezőváltozók közötti korrelációs együtthatók négyzetes átlaga szerepel. A kanonikus korrelációelemzéseknél használatos négyzetes átlag inkább a VIF_j -mutatókkal hozható kapcsolatba.

A kanonikus korrelációelemzés speciális esete az, amikor az eredményváltozók csoportja egy változóból áll. Ekkor az egyetlen kanonikus korreláció nem más, mint a többszörös korrelációs együttható. Ekkor, a j -edik tényezőváltozót különvéve, a többitől a kanonikus korreláció négyzete pontosan $r_{x_j \cdot x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_m}^2$ lesz. Ezt minden lehetséges kombinációra elkészítve – felhasználva a VIF_j (7) képletét – kiszámíthatjuk azt, hogy az egyes tényezőváltozók varianciái átlagosan

$$(17) \quad \frac{\sum_{j=1}^m r_{x_j \cdot x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_m}^2}{m} = \frac{\sum_{j=1}^m \left(1 - \frac{1}{VIF_j}\right)}{m} = 1 - \frac{\sum_{j=1}^m \frac{1}{VIF_j}}{m} = 1 - \frac{1}{VIF_{j_H}}$$

mértékben magyarázhatóak a többi tényezőváltozóval együttesen, ahol $\overline{VIF_j}_H$ a VIF_j - mutatók harmonikus átlaga.

A (17) képlet négyzetgyöke megadja az egyes tényezőváltozóknak a többi tényezőváltozó csoportjával való együttmozgás átlagos mértékét, mellyel a multikollinearitás okainak ismételten csak egy speciális csoportja vizsgálható. A vizsgálatot a későbbiekben általánosítani kell a tényezőváltozók – minden lehetséges módon előállított – két tetszőleges csoportja átlagos együttmozgásának mérésére. **Ennek egyik speciális esete az egy-egy elemű csoportok vizsgálata, mely a Red-mutatóval lehetséges, illetve a másik az egy-(m-1) elemű csoportok vizsgálata, amely a (17) képlettel lehetséges. Ez pontosan a 2. állításom teljesülését jelenti.**

III. 2. A multikollinearitás elliptikus modellje

Felmerülhet az a kérdés, hogy a multikollinearitás hogyan modellezhető. A tényezőváltozókat, mint vektorokat ábrázolva sejtéseket fogalmazhatunk meg a multikollinearitás jelenlétére vonatkozóan. Amennyiben az ábrázolt vektorok ortogonálisak, azaz a tényezőváltozók tere maximálisan kifesztett, akkor nincs multikollinearitás a modellben. Minél kisebb a tér kifesztettsége, annál nagyobb a multikollinearitás mértéke.

A Red-mutató definíciójából kiindulva megadható a multikollinearitás egy más fajta modellje is. A Red-mutató (5) képletét átrendezve az alábbi összefüggéseket kapjuk.

$$\text{Red} = \frac{v_\lambda}{\sqrt{m-1}} = \frac{\sigma_\lambda}{\sqrt{m-1}} = \sqrt{\frac{\sum_{i=1}^m (\lambda_i - 1)^2}{m(m-1)}} = \sqrt{\frac{\sum_{i=1}^m (\lambda_i - 1)^2}{m(m-1)}}$$

$$(18) \quad \sum_{i=1}^m (\lambda_i - 1)^2 = (\sqrt{m(m-1)} \text{Red})^2$$

A (18) egyenlet egy olyan gömb egyenlete, melynek sugara $\sqrt{m(m-1)} \text{Red}$, továbbá középpontjának minden koordinátája egy. Abban az esetben, ha a változók átlagos együttmozgása nulla, azaz nincs együttmozgás a tényezőváltozók között, akkor a gömb

arra az egyetlen pontra redukálódik, melynek minden koordinátája egy. Minél nagyobb mértékű a változók átlagos együttmozgása, annál nagyobb lesz a gömb sugara, azaz annál nagyobb mértékű a gömb „felfújódása”.

Abban az esetben, ha a változók átlagos együttmozgása egy, azaz minden tényezőváltozó-páros közötti korrelációs együttható abszolút értéke egy, akkor a gömb sugara $\sqrt{m(m-1)}$.

Természetesen a gömbök nem minden pontja jelent létező korrelációs struktúrát, hiszen a gömbökön a sajátértékek olyan kombinációi is megtalálhatóak, amelyek korrelációs mátrixok esetén nem lehetségesek. Kérdés, hogy a gömbök mely pontjai jelentenek létező korrelációs struktúrát? **Ezeket a sajátérték-kombinációkat a továbbiakban röviden csak „lehetséges sajátértékeknek” fogom nevezni.** A „lehetséges sajátértékek” vizsgálatához figyelembe kell vennünk a korrelációs mátrix sajátértékeinek tulajdonságait. Mivel a sajátértékek összege megegyezik a tényezőváltozók számával, azaz a gömb dimenziójával, ezért a „lehetséges sajátértékek” biztosan a (18) egyenlettel adott gömbök és (19) metszetein helyezkednek el.

$$(19) \quad \sum_{i=1}^m \lambda_i = m$$

A továbbiakban az általánosság megszorítása nélkül felteszem, hogy

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m = \lambda_{\min}$$

A (19) képletből a legkisebb sajátértéket kifejezve és a (18) egyenletbe behelyettesítve az alábbi egyenletet kapjuk.

$$\sum_{i=1}^{m-1} (\lambda_i - 1)^2 + \left(m - \sum_{i=1}^{m-1} \lambda_i - 1 \right)^2 = m(m-1) \operatorname{Re} d^2$$

Az egyenletet rendezve az alábbi egyenleteket kapjuk.

$$(20) \quad \begin{aligned} & 2 \sum_{i=1}^{m-1} \lambda_i^2 - 2m \sum_{i=1}^{m-1} \lambda_i + 2 \sum_{i=1}^m \sum_{\substack{j=1 \\ j>i}}^{m-1} \lambda_i \lambda_j + m^2 - m = m(m-1) \operatorname{Re} d^2 \\ & \sum_{i=1}^{m-1} \lambda_i^2 - m \sum_{i=1}^{m-1} \lambda_i + \sum_{i=1}^m \sum_{\substack{j=1 \\ j>i}}^{m-1} \lambda_i \lambda_j + \frac{m(m-1)}{2} = \frac{m(m-1)}{2} \operatorname{Re} d^2 \end{aligned}$$

A (20) egyenlet azt jelenti, hogy a „lehetséges sajátértékeket” – adott Red érték mellett – egy $(m-1)$ -dimenziós ellipszoid tartalmazza. Speciálisan három tényezőváltozó esetén az ellipszisek valamely pontjai jelentik a „lehetséges sajátértékeket”. A modell elliptikus elnevezése a görbék jellegéből adódik. Látható, hogy a (20) egyenlet alapján a sajátértékek számához képest eggyel alacsonyabb dimenzióban kapjuk meg a sajátértékek reprezentációját.

Ha a tényezőváltozók száma három, akkor a (20) egyenlet az alábbi formában írható fel.

$$(21) \quad \lambda_1^2 + \lambda_2^2 - 3\lambda_1 - 3\lambda_2 + \lambda_1 \lambda_2 + 3 = 3 \operatorname{Re} d^2$$

Három tényezőváltozó esetén a „lehetséges sajátértékek” tartományának körülhatárolása – a (19) képletén túl – további három feltétel megadásával lehetséges.

- A sajátértékek közötti relációt figyelembe véve: $\lambda_1 \geq \lambda_2$.
- A sajátértékek közötti relációt figyelembe véve: $\lambda_2 \geq \lambda_3 = 3 - \lambda_1 - \lambda_2$, ezért $\lambda_2 \geq \frac{3 - \lambda_1}{2}$.
- Továbbá $\lambda_1 + \lambda_2 \leq 3$. Ez a feltétel már tartalmazza a $\lambda_1 + \lambda_3 \leq 3$ és $\lambda_2 + \lambda_3 \leq 3$ feltételeket is.

A különböző Red értékek melletti szintvonalak közül néhány szemléltetését az 4. ábra tartalmazza. Az ábrázolás a két legnagyobb sajátérték függvényében történik.

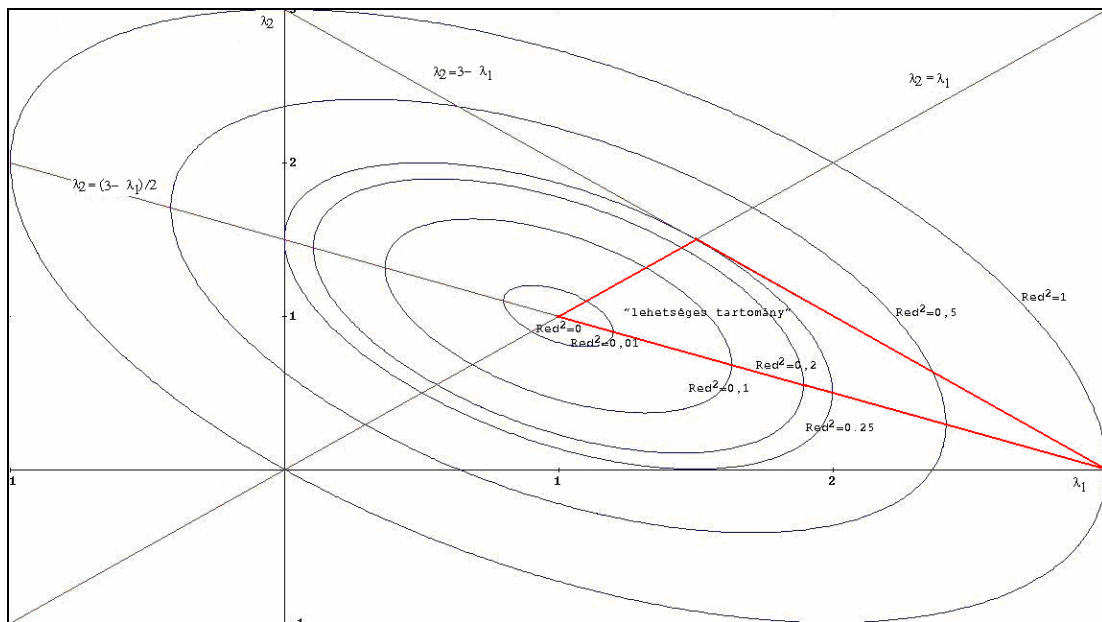
Tehát, három dimenzió esetén a „lehetséges sajátértékek” a 4. ábra háromszögében találhatóak. Az extrém multikollinearitás eseteit az ellipszisek és a $\lambda_2 = 3 - \lambda_1$ egyenes metszéspontjai adják. Ebből is látható, hogy az extrém multikollinearitás különböző esetei is megkülönböztethetők a Red -mutató segítségével.

Magasabb dimenziók esetében – a fenti gondolatmenet alapján – a „lehetséges sajátértékek” ábrázolása nem lehetséges a feltételek magas száma miatt. Ezért magasabb dimenziókban csak azt állíthatjuk biztosan, hogy a változók átlagos együttmozgásának növekedésével a vizsgált m -dimenziós gömb sugara nő. Továbbá, a

Red-mutató rögzített értéke mellett a „lehetséges sajátértékek” egy $(m-1)$ -dimenziós ellipszoid felület részén helyezkednek el. Ez a 3. állításom teljesülését jelenti.

Megjegyzem, hogy hasonló ábrázolás a szakirodalomban a lineáris korrelációs együtthatókra vonatkozóan létezik. Ezek egy elliptópot alkotnak (BOLLA – KRÁMLI, [2005]). Magasabb dimenziókban az ábrázolás ilyen megközelítési módja is nehézkes.

4. ábra: A multikollinearitás elliptikus modellje három tényezőváltozó esetén



(forrás: saját szerkesztés)

A következőkben három tényezőváltozó esetén bemutatom az ellipszisek néhány sajátosságát.

Először megvizsgálom azt, hogy egy adott ellipszisen a „lehetséges sajátértékek” milyen tulajdonsággal bírnak. Az empirikus tapasztalatok szerint, adott *Red* érték mellett, a λ_1 sajátérték növekedése a λ_2 sajátérték nagyobb mértékű csökkenésével jár együtt, ezért a legkisebb sajátérték is növekedni fog, mivel a sajátértékek összege három. Ezt szemlélteti a VI. 6. mellékletben közölt táblázat.

A következőkben megvizsgálom, hogy rögzített *Red* érték mellett, az egyes sajátértékek milyen értékeket vehetnek fel.

Rögzített Red érték mellett, a λ_1 sajátérték legnagyobb értéke a sajátértékek lehetséges tartományának alsó, míg legkisebb értéke a sajátértékek lehetséges tartományának felső határán helyezkedik el.

Ezek szerint, a λ_1 sajátérték legnagyobb értékét a Red -mutató rögzített értéke mellett a

$$\left. \begin{aligned} \lambda_1^2 + \lambda_2^2 - 3\lambda_1 - 3\lambda_2 + \lambda_1 \lambda_2 + 3 &= 3 \operatorname{Re} d^2 \\ \lambda_2 &= \frac{3 - \lambda_1}{2} \end{aligned} \right\}$$

egyenletrendszer megoldása adja.

Ekkor

$$\lambda_1^2 + \left(\frac{3 - \lambda_1}{2} \right)^2 - 3\lambda_1 - 3 \frac{3 - \lambda_1}{2} + \lambda_1 \frac{3 - \lambda_1}{2} + 3 = 3 \operatorname{Re} d^2,$$

azaz

$$4\lambda_1^2 + (3 - \lambda_1)^2 - 12\lambda_1 - 6(3 - \lambda_1) + 2\lambda_1(3 - \lambda_1) + 12 = 12 \operatorname{Re} d^2,$$

melyet rendezve az alábbi egyenlet adódik.

$$3\lambda_1^2 - 6\lambda_1 + 3 = 12 \operatorname{Re} d^2$$

$$\lambda_1^2 - 2\lambda_1 + 1 = 4 \operatorname{Re} d^2$$

$$(\lambda_1 - 1)^2 = (2 \operatorname{Re} d)^2$$

Mivel a legnagyobb sajátérték egynél nem lehet kisebb, ezért

$$\lambda_1 = 1 + 2 \operatorname{Re} d,$$

továbbá

$$\lambda_2 = \frac{3 - \lambda_1}{2} = 1 - \operatorname{Re} d;$$

$$\lambda_3 = 3 - \lambda_1 - \lambda_2 = 1 - \operatorname{Re} d.$$

A λ_1 sajátérték legkisebb értékének meghatározásakor, a Red -mutató rögzített értéke mellett, két eset lehetséges, attól függően, hogy a tartomány felső határai közül melyiket

kell figyelembe venni. Mivel a $\lambda_2 = 3 - \lambda_1$ és a $\lambda_2 = \lambda_1$ határok metszéspontja $\lambda_1 = \lambda_2 = 1,5$; $\lambda_3 = 0$ és ekkor $Red=0,5$; ezért a két eset az alábbi lesz.

A) Ha $Red \leq 0,5$; akkor a

$$\left. \begin{aligned} \lambda_1^2 + \lambda_2^2 - 3\lambda_1 - 3\lambda_2 + \lambda_1 \lambda_2 + 3 &= 3Red^2 \\ \lambda_2 &= \lambda_1 \end{aligned} \right\}$$

egyenletrendszert kell megoldani.

Ekkor

$$\lambda_1^2 + \lambda_1^2 - 3\lambda_1 - 3\lambda_1 + \lambda_1 \lambda_1 + 3 = 3Red^2;$$

$$3\lambda_1^2 - 6\lambda_1 + 3 = 3Red^2;$$

$$\lambda_1^2 - 2\lambda_1 + 1 = Red^2;$$

$$(\lambda_1 - 1)^2 = (Red)^2.$$

Mivel a legnagyobb sajátérték egynél nem lehet kisebb, ezért

$$\lambda_1 = 1 + Red;$$

$$\lambda_2 = \lambda_1 = 1 + Red;$$

$$\lambda_3 = 3 - \lambda_1 - \lambda_2 = 1 - 2Red.$$

B) Ha $Red \geq 0,5$; akkor a

$$\left. \begin{aligned} \lambda_1^2 + \lambda_2^2 - 3\lambda_1 - 3\lambda_2 + \lambda_1 \lambda_2 + 3 &= 3Red^2 \\ \lambda_2 &= 3 - \lambda_1 \end{aligned} \right\}$$

egyenletrendszert kell megoldani.

$$\lambda_1^2 + (3 - \lambda_1)^2 - 3\lambda_1 - 3(3 - \lambda_1) + \lambda_1(3 - \lambda_1) + 3 = 3Red^2$$

$$\lambda_1^2 - 3\lambda_1 + 3 = 3Red^2$$

Ekkor, mivel a legnagyobb sajátérték egynél nem lehet kisebb, ezért

$$\lambda_1 = \frac{3 + \sqrt{12Red^2 - 3}}{2};$$

$$\lambda_2 = \frac{3 - \sqrt{12 \operatorname{Re} d^2 - 3}}{2};$$

$$\lambda_3 = 0.$$

Összefoglalva elmondható, hogy a $\operatorname{Re} d$ -mutató rögzített értéke mellett a sajátértékek az alábbi tartományokban mozoghatnak.

Amennyiben $\operatorname{Re} d \leq 0,5$; akkor

$$\lambda_1 \in [1 + \operatorname{Re} d; 1 + 2 \operatorname{Re} d];$$

$$\lambda_2 \in [1 - \operatorname{Re} d; 1 + \operatorname{Re} d];$$

$$\lambda_3 \in [1 - 2 \operatorname{Re} d; 1 - \operatorname{Re} d].$$

Ha $\operatorname{Re} d \geq 0,5$; akkor

$$\lambda_1 \in [\frac{3 + \sqrt{12 \operatorname{Re} d^2 - 3}}{2}; 1 + 2 \operatorname{Re} d];$$

$$\lambda_2 \in [1 - \operatorname{Re} d; \frac{3 - \sqrt{12 \operatorname{Re} d^2 - 3}}{2}];$$

$$\lambda_3 \in [0; 1 - \operatorname{Re} d].$$

Ez az eredmény – bár a gyakorlati jelentősége kicsi – arra használható, hogy a tényezőváltozók korrelációs mátrixának elmeiből, a (16) képlett kiszámított $\operatorname{Re} d$ -mutató értéke alapján megbecsülhetjük a korrelációs mátrix sajátértékeit. Érdekesebb a kérdés megfordítása, azaz a – számítógépes kimeneten – megjelenő sajátértékek közül egy ismerete alapján, hogyan becsülhető meg a $\operatorname{Re} d$ -mutató értéke.

A sajátértékekre most megadott intervallumokból következik, hogy tetszőleges sajátértékhez meg lehet állapítani, hogy a $\operatorname{Re} d$ -mutató értéke milyen határok közé fog esni.

Például, ha $\lambda_1 \leq 1,5$; akkor

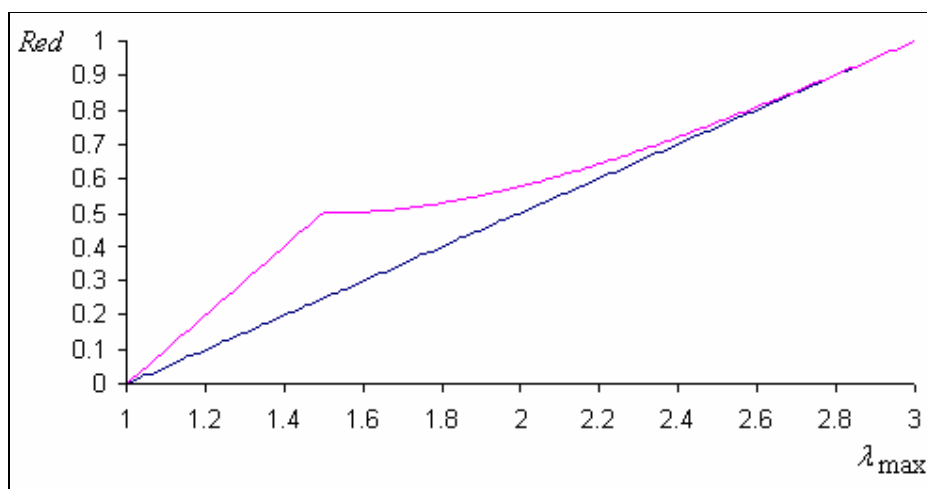
$$\operatorname{Re} d \in [\frac{\lambda_1 - 1}{2}; \lambda_1 - 1],$$

továbbá, ha $\lambda_1 \geq 1,5$; akkor

$$\operatorname{Re} d \in [\frac{\lambda_1 - 1}{2}; \sqrt{\frac{(2\lambda_1 - 3)^2 + 3}{12}}].$$

A legnagyobb sajátérték függvényében a *Red*-mutató lehetséges értékeinek tartományát szemlélteti az 5. ábra.

5. ábra: A *Red*-mutató lehetséges legkisebb és legnagyobb értéke a legnagyobb sajátérték függvényében három tényezőváltozó esetén



(forrás: saját szerkesztés)

A 4. ábra alapján látható, hogy a változók együttmozgásának növekedésére a „lehetséges sajátértékek” tartományán az ellipszisek balról jobbra tolódnak.

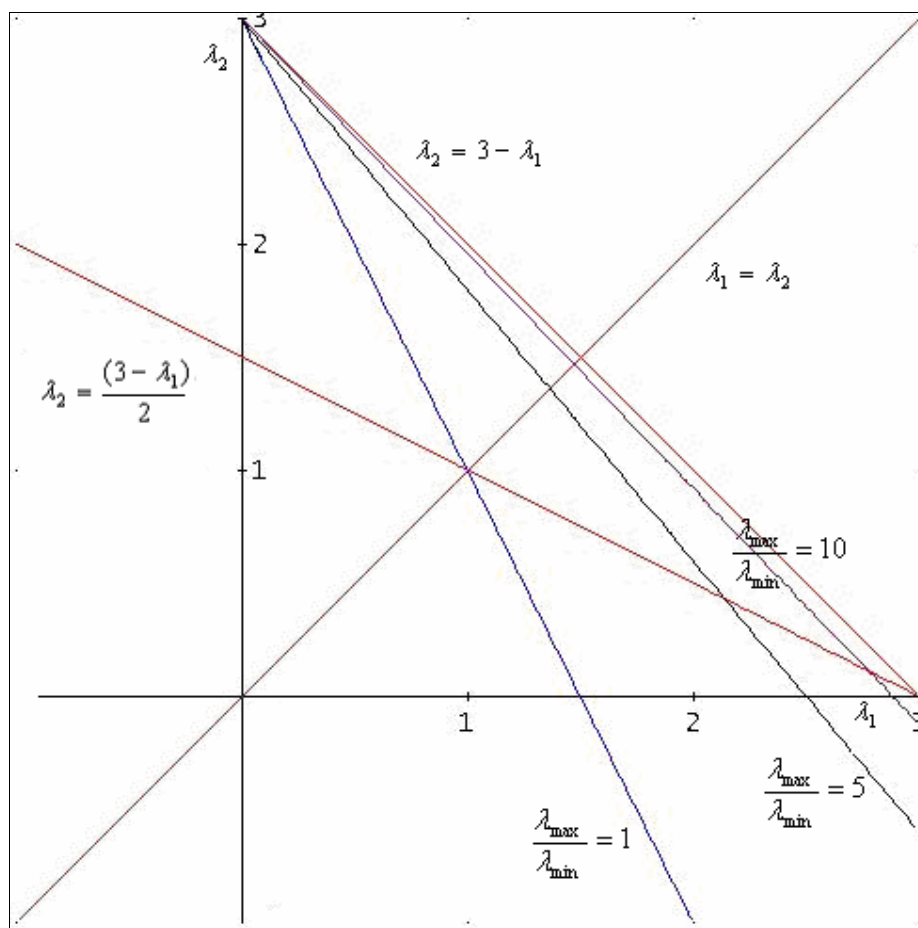
A legnagyobb és a legkisebb sajátértékek hányadosait vizsgálva megállapítható, hogy a hányadosok rögzített értéke mellett, a lehetséges sajátérték-kombinációk egy egyenesen helyezkednek el. Ezeknek az egyeneseknek közös pontja a (0;3) pont, továbbá ezek a multikollinearitás mértékének növekedésére egyre jobban közelítenek a tartomány $\lambda_2 = 3 - \lambda_1$ határához (6. ábra).

A sajátérték legnagyobb és legkisebb értékének hányadosát megbecsülhetjük a *Red*-mutató rögzített értéke mellett. Ehhez a „lehetséges sajátértékek” tartományának határán kell meghatároznunk az ellipszisek és az egyenesek metszéspontjait. Nyilvánvaló, hogy a tartomány alsó határán határozhatjuk meg a hányados minimális értékét, míg a felső határán a maximális értékét.

Mivel a Red -mutató rögzített értéke mellett az alsó határon $\lambda_1 = 1 + 2\text{Re } d$ és $\lambda_3 = 1 - \text{Re } d$, ezért

$$\frac{\lambda_1}{\lambda_3} \geq \frac{1 + 2\text{Re } d}{1 - \text{Re } d}.$$

6. ábra: A legnagyobb és a legkisebb sajátérték hányadosának viselkedése három tényezőváltozó esetén



(forrás: saját szerkesztés)

Ha $\text{Re } d \geq 0,5$; akkor a legnagyobb és legkisebb sajátérték hányadosának nincs felső korlátja, hiszen ekkor a legkisebb sajátérték nulla is lehet. Ha $\text{Re } d < 0,5$; akkor a maximális érték a $\lambda_1 = \lambda_2$ határ segítségével állapítható meg. Ekkor $\lambda_1 = 1 + \text{Re } d$ és $\lambda_3 = 1 - 2\text{Re } d$; ezért

$$\frac{\lambda_1}{\lambda_3} \leq \frac{1 + \text{Re } d}{1 - 2 \text{Re } d}.$$

Összefoglalva elmondható, hogy ha $\text{Re } d < 0,5$; akkor

$$\frac{\lambda_1}{\lambda_3} \in \left[\frac{1 + 2 \text{Re } d}{1 - \text{Re } d} ; \frac{1 + \text{Re } d}{1 - 2 \text{Re } d} \right],$$

ha $\text{Re } d \geq 0,5$; akkor

$$\frac{\lambda_1}{\lambda_3} \in \left[\frac{1 + 2 \text{Re } d}{1 - \text{Re } d} ; +\infty \right).$$

Fordítva: ha a sajátérték legnagyobb és legkisebb értékének hányadosa rögzített, véges érték, akkor

$$\text{Re } d \in \left[\frac{\frac{\lambda_1}{\lambda_3} - 1}{1 + 2 \frac{\lambda_1}{\lambda_3}} ; \frac{\frac{\lambda_1}{\lambda_3} - 1}{\frac{\lambda_1}{\lambda_3} + 2} \right],$$

ellenkező esetben $\text{Re } d \geq 0,5$.

Későbbi vizsgálataim tárgyát fogják képezni az ellipszisek további tulajdonságainak vizsgálata, illetve a vizsgáltat kiterjesztése a „magasabb dimenziókba”.

III. 3. A *Red*-mutató és a regressziós paraméterek varianciái

Mivel a multikollinearitás egyik leggyakrabban említett negatív következménye a becsült regressziós paraméterek varianciáinak, illetve ezek felfújódásának növekedése, ezért célszerű megvizsgálni a *Red*-mutató és a becsült regressziós paraméterek varianciáinak kapcsolatát.

A becsült paraméterek varianciáinak összege akkor véges, ha a sajátértékek minimuma pozitív. Adott m mellett a sajátértékek minimuma akkor nulla, ha valamelyik magyarázóváltozó lineárisan függ a magyarázóváltozók egy részrendszerétől. Egy korábbi hipotézisem szerint, a fenti gondolatmenet alapján, a *Red*-mutató értéke akkor minimális, ha $m-1$ darab tényezőváltozó ortogonális, azaz lineárisan korrelálatlanok, és egy tényezőváltozó lineárisan függ valamelyik magyarázóváltozótól. Ekkor a *Red*-mutató értéke:

$$(22) \quad Red_c = \sqrt{\frac{2}{m(m-1)}}.$$

Tehát, ha egy adatállomány redundanciájának mértéke kisebb a Red_c kritikus redundancia-értéknél, akkor a lineáris regressziós modell illesztése után kapott becült paraméterek szórásnégyzetei biztosan végesek. Ha egy adatállomány redundanciájának mértéke nagyobb a Red_c kritikus redundancia-értéknél, akkor a lineáris regressziós modell illesztése után kapott becült paraméterek szórásnégyzeteiről nem állíthatjuk biztosra, hogy végesek (KOVÁCS – PETRES – TÓTH, [2004]).

Azonban, a (22) képlet nem érvényes. A multikollinearitás elliptikus modelljének szemléltetéséből következően, három tényezőváltozó esetén már a $Red=0,5$ értéknél is ($\lambda_1 = \lambda_2 = 1,5$; $\lambda_3 = 0$) lehet a varianciák összege végtelen. Az eredeti gondolatmenetben az a hiba, hogy a megadott esetben a képlet számlálója lehet kettőnél kisebb is. Erre jó példa az 1. táblázat adatai alapján felírt korrelációs mátrixból számított Red érték:

$$\begin{aligned} Red &= \sqrt{\frac{0,42796^2 + 0,516842^2 + (-0,55254)^2 + 0,42796^2 + 0,516842^2 + (-0,55254)^2}{3 \cdot 2}} = \\ &= \sqrt{\frac{1,51115}{3 \cdot 2}} = 0,502. \end{aligned}$$

Az $m=3$ esetből kiindulva – amikor is $\lambda_1 = \lambda_2 = 1,5$; $\lambda_3 = 0$ és a Red_c kritikus redundancia-érték $\frac{1}{m-1}$, azaz 0,5 – a következő állítást lehet általánosan megfogalmazni:

$$Red_c = \frac{1}{m-1},$$

és ekkor

$$\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = \frac{m}{m-1}; \lambda_m = 0.$$

Ugyanis, ha $\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = \frac{m}{m-1}$, $\lambda_m = 0$, akkor

$$Red = \sqrt{\frac{\sum_{i=1}^m \lambda_i^2 - m}{m(m-1)}} = \sqrt{\frac{\sum_{i=1}^{m-1} \left(\frac{m}{m-1}\right)^2 - m}{m(m-1)}} = \sqrt{\frac{\frac{m}{m-1} - 1}{m-1}} = \sqrt{\frac{1}{(m-1)^2}} = \sqrt{\frac{1}{(m-1)(m-1)}}$$

Ha a nem nulla sajátértékek eltávolodnak a $\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = \frac{m}{m-1}$, $\lambda_m = 0$ esettől, akkor

$$\begin{aligned} \sum_{i=1}^{m-1} (\lambda_i - a_i)^2 &= \sum_{i=1}^{m-1} \lambda_i^2 + \sum_{i=1}^{m-1} a_i^2 + 2 \sum_{i=1}^{m-1} \lambda_i a_i = \\ &= \sum_{i=1}^{m-1} \lambda_i^2 + \sum_{i=1}^{m-1} a_i^2 + 2 \frac{m}{m-1} \sum_{i=1}^{m-1} a_i = \sum_{i=1}^{m-1} \lambda_i^2 + \sum_{i=1}^{m-1} a_i^2 > \sum_{i=1}^{m-1} \lambda_i^2. \end{aligned}$$

Mivel a sajátértékek összege mindig m , ezért $\sum_{i=1}^{m-1} a_i = 0$. Ez pedig azt jelenti, hogy ekkor a

Red -mutató értéke nagyobb lesz, mint a $\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = \frac{m}{m-1}$, $\lambda_m = 0$ esetben.

Ebből következik, hogy $Re d_c = \frac{1}{m-1}$, azaz ha a Red -mutató értéke kisebb ennél a kritikus értéknél, akkor a becslt paraméterek varianciái biztosan végesek. Ahhoz, hogy valamely becslt paraméter varianciája végtelen legyen szükséges, de nem elégséges feltétel a $Re d \geq Re d_c$ reláció teljesülése.

Az állítással kapcsolatosan megfogalmazható néhány következmény. Ha minden sajátérték egy, akkor $Red=0$. Amennyiben a sajátértékek távolodnak ettől az esettől, akkor a Red -mutató értéke növekszik. Speciális esetként eljuthatunk a

$\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = \frac{m}{m-1}$, $\lambda_m = 0$ lehetséges sajátérték-kombinációba. Ha ezt a triviális

következményt továbbgondoljuk, akkor megállapítható az, hogy a

$\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = \frac{m}{m-1}$, $\lambda_m = 0$ esettől távolodva eljuthatunk a

$\lambda_1 = \lambda_2 = \dots = \lambda_{m-2} = \frac{m}{m-2}$, $\lambda_{m-1} = \lambda_m = 0$ esethez is. Ebből következik, hogy a két zéró

sajátérték létezésének szükséges feltétele: $Re d \geq Re d_{c,2}$, ahol $Re d_{c,2}$ a két nulla sajátérték létezésének kritikus értéke. A fenti gondolatmenetet $Re d_{c,2}$ értékre felhasználva azt kapjuk, hogy

$$Re d_{c,2} = \sqrt{\frac{\sum_{i=1}^m \lambda_i^2 - m}{m(m-1)}} = \sqrt{\frac{\sum_{i=1}^{m-2} \left(\frac{m}{m-2}\right)^2 - m}{m(m-1)}} = \sqrt{\frac{\frac{m}{m-2} - 1}{m-1}} = \sqrt{\frac{2}{(m-1)(m-2)}}.$$

Általánosítva: k darab sajátérték zéró voltának szükséges feltétele az alábbi formában írható fel.

$$\operatorname{Re} d \geq \operatorname{Re} d_{c,k} = \sqrt{\frac{\sum_{i=1}^m \lambda_i^2 - m}{m(m-1)}} = \sqrt{\frac{\sum_{i=1}^{m-k} \left(\frac{m}{m-k}\right)^2 - m}{m(m-1)}} = \sqrt{\frac{\frac{m}{m-k} - 1}{m-1}} = \sqrt{\frac{k}{(m-1)(m-k)}}.$$

Tehát megadtam a Red-mutató egy olyan kritikus értékét, amely szükséges feltétele annak, hogy a becsült paraméterek varianciái ne legyenek végtelenek, illetve annak, hogy k darab sajátérték zéró legyen.

Ezen állítást szemlélteti a 26. táblázat.

26. táblázat: A Red-mutató kritikus értékei, mint k darab sajátérték zéró voltának szükséges feltétele

m	k								
	1	2	3	4	5	6	7	8	9
2	0,707								
3	0,408	0,816							
4	0,289	0,500	0,866						
5	0,224	0,365	0,548	0,894					
6	0,183	0,289	0,408	0,577	0,913				
7	0,154	0,239	0,327	0,436	0,598	0,926			
8	0,134	0,204	0,274	0,354	0,456	0,612	0,935		
9	0,118	0,178	0,236	0,298	0,373	0,471	0,624	0,943	
10	0,105	0,158	0,207	0,258	0,316	0,387	0,483	0,632	0,949

(forrás: saját szerkesztés)

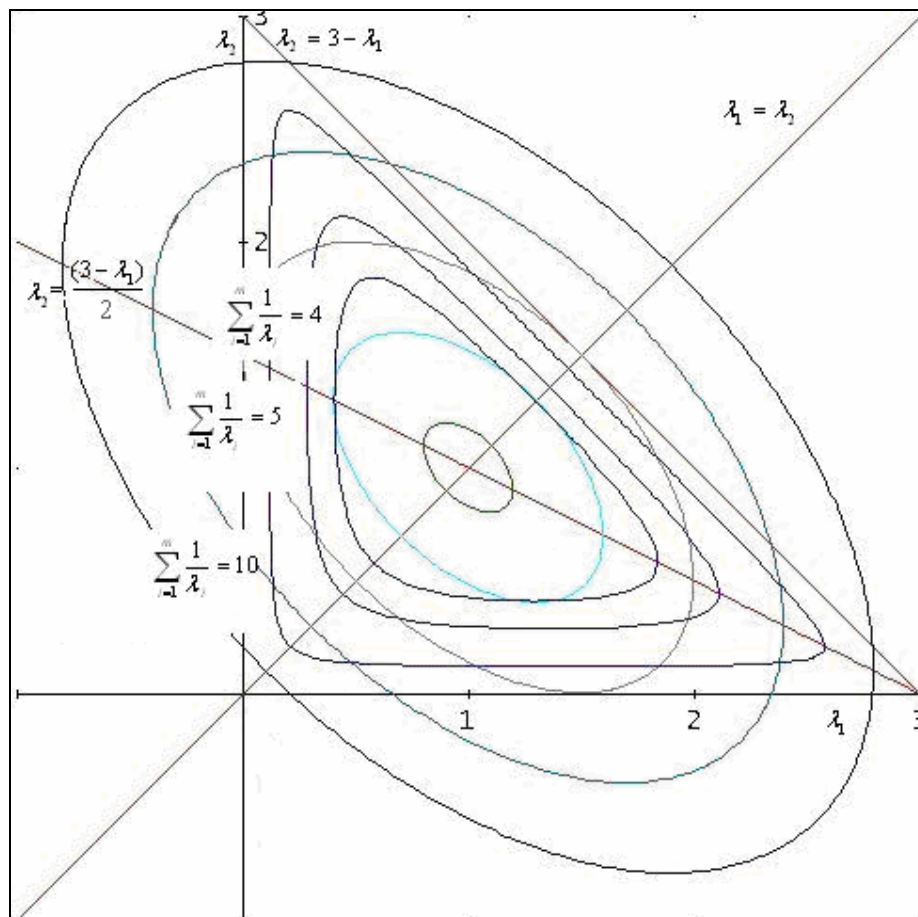
Ezek alapján varianciák „felfújódásának” szempontjából, a II.4. alfejezetben bemutatott példa különböző méretű modelljeiben kiszámított *Red* értékekből mit állapíthatunk meg? Az eredeti modellben ($m=7$) kiszámított *Red* érték (0,359) az adott dimenzióban nem tekinthető alacsonynak, hiszen a mutató 0,154 feletti értéke már problémát jelezhet. Egyedül a stepwise eljárás eredményéül adódó 2. modellről állíthatjuk a *Red-mutató* értéke (19. táblázat) alapján azt, hogy a modellben a becsült regressziós paraméterek varianciái végesek lesznek.

A következőkben megvizsgálom, hogy a különböző *Red* értékek függvényében mekkora a varianciák „felfújódásának” mértéke a három tényezőváltozót tartalmazó modellekben. Mivel a *Red*-mutató egy szintetikus mérőszám, ezért segítségével a varianciák összegének „felfújódását” tudjuk elemezni.

Az empirikus tapasztalatok szerint a *Red*-mutató rögzített értéke mellett, minél jobban távolodunk a „lehetséges sajátértékek” tartományának alsó határától, annál nagyobb lesz a sajátértékek reciprokösszege. Ezt szemlélteti a VI. 6. mellékletben közölt táblázat.

A 7. ábra tartalmazza az ellipsziseket, illetve a sajátértékek reciprokösszegét, abban az esetben, amikor ez az összeg 4, 5, illetve 10.

7. ábra: A sajátértékek reciprokösszegének viselkedése ($m=3$)



(forrás: saját szerkesztés)

Ennek segítségével a Red -mutató rögzített értéke mellett becsülhető a becsült paraméterek varianciáinak összege. Megállapítható, hogy a mutató egyes értékeitől kezdve minimum, illetve maximum mekkora „felfújódásra” számíthatunk. Mivel ennek nagysága a sajátértékek reciprokösszegétől függ, ezért ezt az összeget kell becsülnünk. Mivel a Red -mutató rögzített értéke mellett a sajátértékek reciprokösszege a lehetséges tartomány alsó határán a legkisebb, ezért a sajátértékek nagyságára vonatkozó korábbi állításokat figyelembe véve azt kapjuk, hogy

$$\sum_{i=1}^3 \frac{1}{\lambda_i} \geq \frac{1}{1-Red} + \frac{1}{1-Red} + \frac{1}{1+2Red} = \frac{3+3Red}{(1-Red)(1+2Red)} = 3 \frac{1+Red}{(1-Red)(1+2Red)}.$$

Ha $\lambda_1 < 1,5$; azaz $Red < 0,5$; akkor a sajátértékek reciprokösszege felülről korlátos. Ezért a sajátértékek nagyságára vonatkozó korábbi állításokat figyelembe véve azt kapjuk, hogy

$$\sum_{i=1}^3 \frac{1}{\lambda_i} \leq \frac{1}{1+Red} + \frac{1}{1+Red} + \frac{1}{1-2Red} = \frac{3-3Red}{(1+Red)(1-2Red)} = 3 \frac{1-Red}{(1+Red)(1-2Red)}.$$

Ha $\lambda_1 \geq 1,5$; azaz $Red \geq 0,5$; akkor a sajátértékek reciprokösszege végtelen is lehet.

Összefoglalva: a becsült paraméterek varianciái összegének a „felfújódása” a hibatag szórásnégyzetéhez viszonyítva a

$$\left[3 \frac{1+Red}{(1-Red)(1+2Red)} ; 3 \frac{1-Red}{(1+Red)(1-2Red)} \right]$$

tartományba esik, ha $Red < 0,5$; ellenkező esetben

$$\left[3 \frac{1+Red}{(1-Red)(1+2Red)} ; +\infty \right)$$

közé esik. Például, ha $Red=0,8$; akkor a varianciák összege a hibatag szórásnégyzetének legalább tízszerese.

Tehát, ha előre meghatározzuk, hogy mekkora „felfújódást” tekintünk megfelelőnek a vizsgálataink szempontjából, akkor egy határszámot kaphatunk a Red -mutatóra vonatkozóan. Ez a határszám azt mutatja meg, hogy ha a Red -mutató értéke ennél az

értéknél kisebb, akkor a varianciák „felfújódása” biztosan megfelel a varianciák összegével kapcsolatos elvárásainknak.

Végezetül – érdekességképpen – megvizsgálom, hogy milyen összefüggés van a sajátértékek harmonikus átlaga és a becsült paraméterek számtani átlaga között.

A (10) képlet szerint a varianciák értékét, azaz a varianciáknak a hibatag szórásnégyzetéhez viszonyított „felfújódásának” mértékét végső soron – ceteris paribus – a sajátértékek befolyásolják: ha legalább egy túl közel van nullához, akkor nagy mértékben növekszik a becsült paraméterek szórásnégyzeteinek átlaga.

A (10) egyenletet átrendezve az

$$\frac{1}{\sum_{l=1}^m \frac{1}{\lambda_l}} \cdot \sum_{j=1}^m \text{Var}(\hat{\beta}_j) = \sigma^2$$

összefüggést kapjuk, ami az $\frac{m}{m}$ törttel beszorozva az alábbi összefüggéshez vezet.

$$\frac{m}{\sum_{l=1}^m \frac{1}{\lambda_l}} \cdot \frac{\sum_{j=1}^m \text{Var}(\hat{\beta}_j)}{m} = \sigma^2$$

Ebben a kifejezésben

$$\frac{m}{\sum_{l=1}^m \frac{1}{\lambda_l}}$$

a tényezőváltozók **R** korrelációs mátrixának spektrálfelbontásával kapott sajátértékek harmonikus átlaga,

$$\frac{\sum_{j=1}^m \text{Var}(\hat{\beta}_j)}{m}$$

pedig a becsült paraméterek szórásnégyzeteinek számtani átlaga.

Ezek szerint felírható az alábbi összefüggés.

$$(23) \quad \bar{\lambda}_h \cdot \overline{Var(\hat{\beta}_j)} = \sigma^2$$

Mivel a (9) szerint standardizált változók esetében

$$\sum_{j=1}^m VIF_j = \sum_{l=1}^m \frac{1}{\lambda_l},$$

ezért

$$\frac{\overline{Var(\hat{\beta}_j)}}{\overline{VIF_j}} = \sigma^2.$$

Ezt a (23) összefüggésbe behelyettesítve azt kapjuk, hogy

$$\bar{\lambda}_h \cdot \overline{VIF} = 1.$$

Tehát, a sajátértékek harmonikus átlagának és a VIF_j -mutatók számtani átlagának szorzata egy.

III. 4. A *Red*-mutató eloszlásának vizsgálata

Vizsgálataim során megpróbáltam a *Red*-mutató empirikus eloszlásfüggvényét elkészíteni, illetve az eloszlást meghatározni. A *Red*-mutató eloszlásának tanulmányozása céljából két lehetőség állt rendelkezésemre: vagy a magyarázóváltozók közötti lineáris korrelációs együtthatókkal, vagy a korrelációs mátrix sajátértékeivel számítom ki a *Red* összes lehetséges értékeit. Vizsgálataim során az utóbbiak, azaz a sajátértékek segítségével történik az elemzés. Ennek két oka van: egyrészt sokkal alacsonyabb dimenzióban kell dolgozni, másrészt így tudom biztosítani azt a nagyon fontos követelményt, hogy kizárólag valós struktúrákat vizsgáljak.

Az empirikus vizsgálatok során felmerülő fontos probléma, hogy folytonos változók – mint amilyenek a sajátértékek, illetve a *Red* is – esetén hogyan vizsgálható az összes lehetséges érték eloszlása. Ennek gyakorlati megoldása az, hogy elkészítjük az empirikus sűrűségfüggvényeket úgy, hogy közben növeljük a megfigyelt változók értékeinek pontosságát. A *Red*-mutató értékei eloszlásának vizsgálatára saját magam által, Pascal nyelven megírt eljárást alkalmaztam (VI. 7. melléklet).

Az adott algoritmus figyelembe veszi a sajátértékekkel kapcsolatos, a kutatás szempontjából fontos jellemzőket:

- a sajátértékek nem negatívak, maximális értékük megegyezik a magyarázóváltozók számával;
- a sajátértékek csökkenő sorrendben kapjuk, függetlenül a magyarázóváltozók sorrendjétől;
- a korrelációs mátrix determinánsa megegyezik a sajátértékek szorzatával, mely értéke egynél nem nagyobb;
- a sajátértékek összege megegyezik a magyarázóváltozók számával.

Az adatelemzést nehezíti, hogy már három magyarázóváltozós modellek esetén is – az adatok pontosságától függően – nagyméretű adatállományok jönnek létre. Ez változónként több tíz-százmilliós megfigyelésnagyságot eredményez, ami a feldolgozás szempontjából gyakorlati gondot jelent. Ezért a programon belül az algoritmust egy másik algoritmussal kellett bővíteni, amelynek kimenete már a *Red*-mutató empirikus eloszlása. Mivel ennek a programnak a futási ideje is nagyon hosszú, akár több óra, több nap volt, ezért – kérésünkre – TÉBI GÁBOR programtervező matematikus optimalizálta a kódot és egy grafikus felülettel látta el a programot. Az ő módosításainak köszönhetően a hat tényezőváltozót tartalmazó modellek esetében már egy óránál kevesebb a program futási ideje, ha az adatokat kéttizedes pontossággal kezeljük.

A készített algoritmus lényege az, hogy adott pontosság mellett előállítjuk az összes lehetséges sajátérték kombinációt. Például, abban az esetben, ha az adatokat kéttizedes pontossággal kezeljük, akkor a program által generált „lehetséges sajátértékek” számát a 27. táblázat tartalmazza.

Az algoritmus által generált adatállományt egyszerűen át lehet adni akár az Excel, akár az SPSS programnak. A programok segítségével előállítható – kéttizedes pontossággal – a *Red*-mutató empirikus eloszlása, amelyet speciális esetekben a 8.-10. ábrák szemléltetnek.

Természetesen, az adatok kéttizedes pontosságú kezelése az eloszlásfüggvénynek nem a legfinomabb közelítést jelenti. A vizsgálataim során arra törekedtem, hogy az empirikus eloszlásfüggvény jellegét meg tudjam állapítani magasabb dimenziókban is, ezért és a program futási ideje miatt, nem az adatok pontosságának, hanem a tényezőváltozók számának növelése mellett döntöttem.

Három magyarázóváltozó esetén úgy tűnt, hogy az eloszlás „kétoldali eloszlású” (*two-tailed distribution*) lesz. A kimenet alapján azt a sejtést fogalmaztam meg, hogy a bal oldali szár egy egyenes lesz, míg a jobb oldali szár, valamilyen más, folytonos függvény.

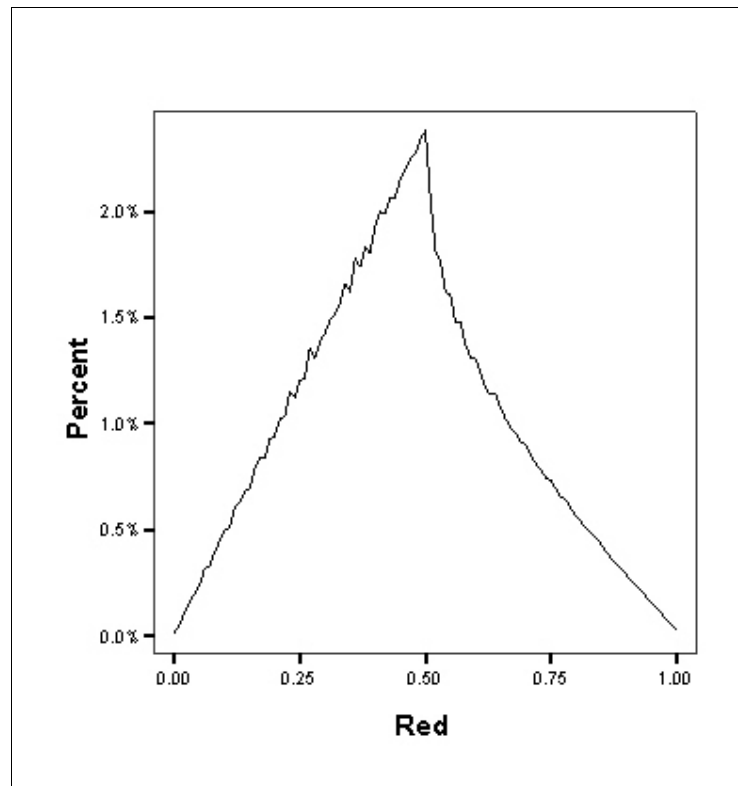
Ez a sejtés a négy tényezőváltozót tartalmazó modellek alapján máris cáfolható, mivel a bal oldali szár már nem egyenesként viselkedik. Magasabb dimenziókban pedig megszűnik a „két szár” metszéspontjának „élessége”.

27. táblázat: Az előállított „lehetséges sajátérték” kombinációk száma kéttizedes pontossággal

Tényezőváltozók száma (<i>m</i>)	„lehetséges sajátérték”kombinációk száma („sajátérték <i>m</i>-esek”)
3	7 651
4	461 312
5	23 030 612
6	981 355 696
7	36 498 678 901

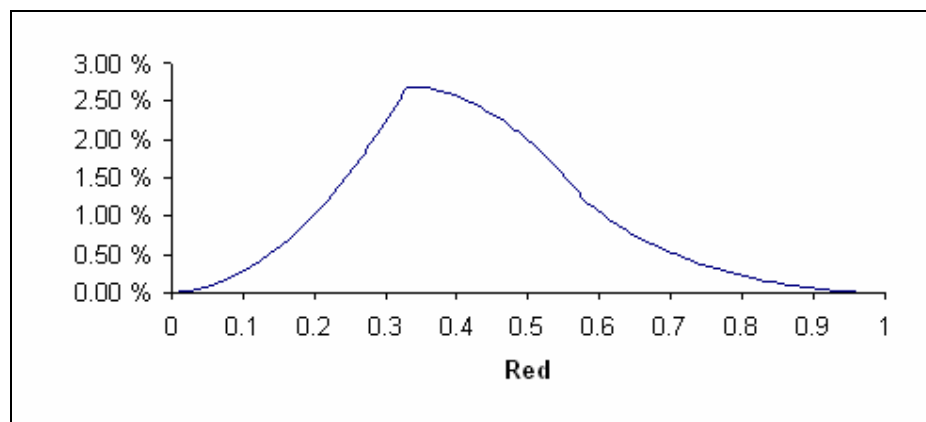
(forrás: saját szerkesztés)

8. ábra: Az empirikus eloszlás grafikus ábrázolása ($m=3$)



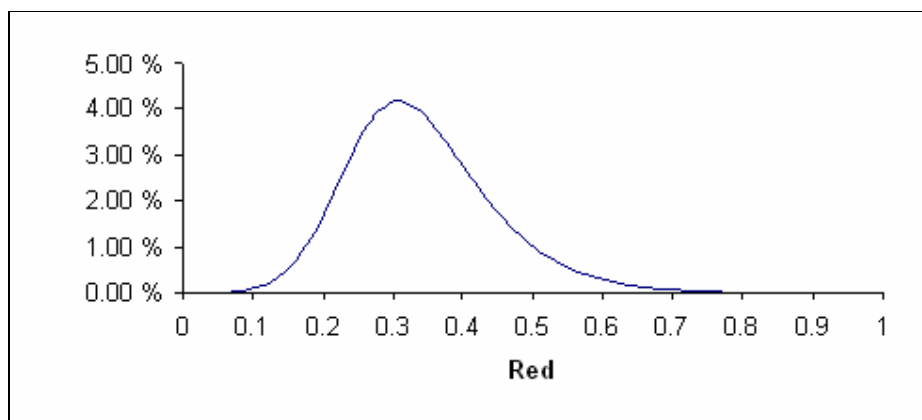
(forrás: saját szerkesztés)

9. ábra: Az empirikus eloszlás grafikus ábrázolása ($m=4$)



(forrás: saját szerkesztés)

10. ábra: Az empirikus eloszlás grafikus ábrázolása ($m=7$)



(forrás: saját szerkesztés)

Az ábrák alapján a II. 4. alfejezetben bemutatott példa különböző méretű modelljeiben kiszámított *Red* értékekből mit állapíthatunk meg? Az eredeti modellben ($m=7$) kiszámított *Red* érték (0,359), továbbá a backward szelekció eredményéül adódó 3. modellben ($m=4$) kiszámított *Red* érték (0,289) az adott dimenzióban közel van a mutató móduszához. További következtetések levonásához még több vizsgálatra van szükség.

Az empirikus eloszlásfüggvény meghatározásának újabb vizsgálataiban már elsősorban a *Red*-mutató négyzetének eloszlását próbálom először megvizsgálni. Ehhez pedig a mutató definíciója szerint a sajátértékek négyzetösszegének az eloszlását kell megvizsgálni. Ezek részletes vizsgálata a közeljövőben fog megtörténni.

A *Red*-mutató eloszlásának beazonosítására teszteltem az SPSS 13.0 programba beépített eloszlástípusokat. Sajnos a *Red*-mutató eloszlásának meghatározása nem sikerült. Az ábrák alapján úgy tűnik, hogy a tényezőváltozók számának növekedésével az eloszlás egyre inkább aszimmetrikus lesz.

III. 5. A *Red*-mutató és a *KMO*-mutató kapcsolata

Egy érdekes kérdés, a *Red*-mutató alkalmazásának a lehetősége. Sejtésem szerint a mutató alkalmazható a faktoranalízis során.

A faktoranalízis alkalmazásakor a célunk az, hogy a nagyszámú korrelált változó közötti összefüggéseket megmagyarázzuk, az eredetnél kevesebb – korrelálatlan – látens faktor segítségével, a megfigyelt változók közös és egyedi tulajdonságainak szétválasztásával.

Ezáltal feltérképezhető a változók közötti kapcsolat, illetve csökkenthető a változók száma az adathalmaz további elemzésében.

A faktoranalízis alkalmazásának akkor van értelme, ha a változók redundáns módon tartalmaznak információt, és mögöttük egy látens struktúra húzódik meg. Az előbbit a BARTLETT-féle teszt, az utóbbit a KAISER – MEYER – OLKIN (*KMO*) mutató¹⁶ segítségével vizsgálhatjuk:

$$KMO = \frac{\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m r_{ij}^2}{\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m r_{ij}^2 + \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m r_{x_i x_j, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_m}^2},$$

ahol r_{ij} a páronkénti korrelációs együtthatókat, $r_{x_i x_j, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_m}$ a parciális korrelációs együtthatókat¹⁷ jelöli.

Mivel az $r_{x_i x_j, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_m}$ parciális korrelációs együttható definiálása úgy történik, hogy mind az i -edik, mind a j -edik tényezőváltozóra, mint eredményváltozóra a többi $m-2$ tényezőváltozóval lineáris regressziós modellt illesztünk, ezért a parciális korrelációs együtthatók a két modell reziduuma közötti korrelációs együtthatók. Éppen ezért, a parciális korrelációs együtthatók a két változónak a többi változó által, lineáris regresszióval meg nem magyarázott részének korrelációja (ÉLTETŐ – MESZÉNA – ZIERMANN, [1982]).

A *KMO* képlete azon a gondolaton alapszik, hogy két változó között a tényleges korreláció felbontható egy, a többi változó által lineáris regresszióval megmagyarázható, és egy lineáris regresszióval meg nem magyarázható részre. Az előbbit a páronkénti korrelációs együtthatók, míg utóbbit a parciális korrelációs együtthatók mérik. Ezek szerint a *KMO*-mutató megmutatja, hogy az összes változópár közötti korreláció mekkora hányada magyarázható lineáris regresszióval.

¹⁶ Ortogonális tényezőváltozók esetén a mutató értékét egynek tekintjük.

¹⁷ A parciális korrelációs együtthatók az $r_{x_i x_j, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_m} = -\frac{\mathbf{R}_{ij}^{-1}}{\sqrt{\mathbf{R}_{jj}^{-1} \mathbf{R}_{ii}^{-1}}}$ képlettel

számíthatók ki a korrelációs mátrix inverzének megfelelő elemei alapján.

A mögöttes struktúra létét akkor feltételezhetjük, ha nem csak változópárok vannak egymással kapcsolatban, hanem sztochasztikus kapcsolatok rendszere jelentkezik. A *KMO*-mutató 0 és 1 közé eső számot ad. A *KMO*-mutató egy – egyébként szubjektív – értelmezése a 28. táblázatban található. A *KMO* egyhez közeli értékei tehát azt jelzik, hogy a faktorelemzés alkalmazása eredményes lehet.

Definiáljuk a parciális korrelációs együtthatók mátrixát úgy, hogy a mátrix főátlójában minden elem egy, azon kívül pedig minden ij -edik elem az i -edik és a j -edik változó közötti parciális korrelációs együtthatót adja. Ez alapján a parciális korrelációs együtthatók mátrixára is kiszámíthatjuk a *Red*-mutató Red_p értékét.

Ekkor látható, hogy a *KMO* képletében szereplő mennyiségek kifejezhetők a *Red*-mutatóval (5. állítás).

$$(24) \quad KMO = \frac{m(m-1) \operatorname{Re} d^2}{m(m-1) \operatorname{Re} d_p^2 + m(m-1) \operatorname{Re} d^2} = \frac{\operatorname{Re} d^2}{\operatorname{Re} d_p^2 + \operatorname{Re} d^2}$$

28. táblázat: Útmutató a *KMO* értelmezéséhez

<i>KMO</i>	<i>Minősítés</i>
– 0,5)	elfogadhatatlan
[0,5 ; 0,6)	szánalmas
[0,6 ; 0,7)	mérsékelt
[0,7 ; 0,8)	közepes
[0,8 ; 0,9)	dicséretes
[0,9 ; 1,0]	csodálatos

(forrás: KOVÁCS – PETRES – TÓTH, [2006])

A (24) képlet átrendezéséből adódóan

$$(\text{Re } d_p^2 + \text{Re } d^2) KMO = \text{Re } d^2,$$

$$\text{Re } d_p^2 = \frac{1 - KMO}{KMO} \text{Re } d^2 = \left(\frac{1}{KMO} - 1 \right) \text{Re } d^2.$$

Ez azt jelenti, hogy a parciális korrelációs együtthatók négyzetes átlaga nem kisebb a korrelációs együtthatók négyzetes átlagánál. Ennek részletes jelentése későbbi vizsgálataim tárgyát fogják képezni.

III. 6. A multikollinearitás egy másik mérési lehetősége

Mivel a *Red*-mutató a tényezőváltozók korrelációs mátrixának sajátértékei alapján számított normált relatív szórás, ezért úgy gondolom, hogy a multikollinearitás mérhető a sajátértékek más szóródási mérőszámával is, melynek alapgondolata megegyezik a *Red*-mutató alapötletével. A *Red*-mutatóhoz hasonló gondolatmenet alapján a redundancia mértékét a tényezőváltozók korrelációs mátrixa sajátértékeinek GINI-együtthatójával is lehetne jellemezni. Minél nagyobb egy adatállományban a redundancia mértéke a sajátértékek annál jobban távolodnak egymástól. Két szélsőeset van: ha minden sajátérték megegyezik, akkor $G=0$, ha egy sajátérték kivételével mindegyik nulla, akkor $G=2$. A mutató értéke nulla és kettő közé esik.

A sajátértékek GINI-együtthatója (25) alakban írható fel.

$$(25) \quad G = \frac{\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m |\lambda_i - \lambda_j|}{m(m-1)}$$

Mivel

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m = \lambda_{\min},$$

(25) kiszámítása más módon is történhet. Ehhez azt kell figyelembe venni, hogy ezért

melyik sajátérték hányszor, illetve milyen előjellel szerepel a kifejezésben. A képletben minden sajátértékpár kétszer fordul elő úgy, hogy a kettő közül a nagyobb sajátérték mindig kétszer szerepel pozitív, a kisebb pedig kétszer negatív előjellel. Minden sajátérték annyiszor fordul elő pozitív előjellel, ahányszor nála kisebb sajátértékkel van párosítva a

kifejezésben, illetve annyiszor fordul elő negatív előjellel, ahányszor nála nagyobb sajátértékkel van párosítva a kifejezésben. Ezért,

a λ_1 sajátérték együtthatója: $(m-1)2-0$;

a λ_2 sajátérték együtthatója: $(m-2)2-1 \cdot 2$;

a λ_3 sajátérték együtthatója: $(m-3)2-3 \cdot 2$;

a λ_i sajátérték együtthatója: $(m-i)2-(i-1) \cdot 2 = 2(m-2i+1)$.

Ekkor azt kapjuk, hogy

$$(26) \quad G = \frac{\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m |\lambda_i - \lambda_j|}{m(m-1)} = \frac{\sum_{i=1}^m 2(m-2i+1)\lambda_i}{m(m-1)}.$$

Mivel a sajátértékek összege m , ezért

$$\begin{aligned} G &= 2 \frac{(m+1) \sum_{i=1}^m \lambda_i}{m(m-1)} - 4 \frac{\sum_{i=1}^m i \lambda_i}{m(m-1)} = 2 \frac{(m+1)}{(m-1)} - 4 \frac{\sum_{i=1}^m i \lambda_i}{m(m-1)} = \\ &= 2 \frac{(m+1)}{(m-1)} - \frac{4}{(m-1)} - 4 \frac{\sum_{i=1}^m (i-1) \lambda_i}{m(m-1)} = 2 - 4 \frac{\sum_{i=1}^m (i-1) \lambda_i}{m(m-1)}. \end{aligned}$$

A (26) képletben szereplő együtthatók szabályosan viselkednek. Az i -edik és az $(m-i+1)$ -edik sajátértékek együtthatói egymás ellentettje (29. táblázat).

Az elliptikus modellhez hasonlóan, a GINI-együtthatók a „lehetséges sajátértékek” tartományán a redundancia növekedésével szintén jobbra tolódnak, azonban az együttható rögzített értéke mellett a sajátértékek egy egyenesen helyezkednek el (11. ábra). Ha „lehetséges sajátértékek” tartományán egyszerre ábrázoljuk az elliptikus jellegű görbéket és ezeket az egyeneseket, akkor azt az érdekességet lehet észrevenni, hogy rögzített Red

érték mellett az elliptikus görbe és a $G=2Red$ egyenesek a $\lambda_1 = \lambda_2$ és a $\lambda_2 = \frac{3-\lambda_1}{2}$

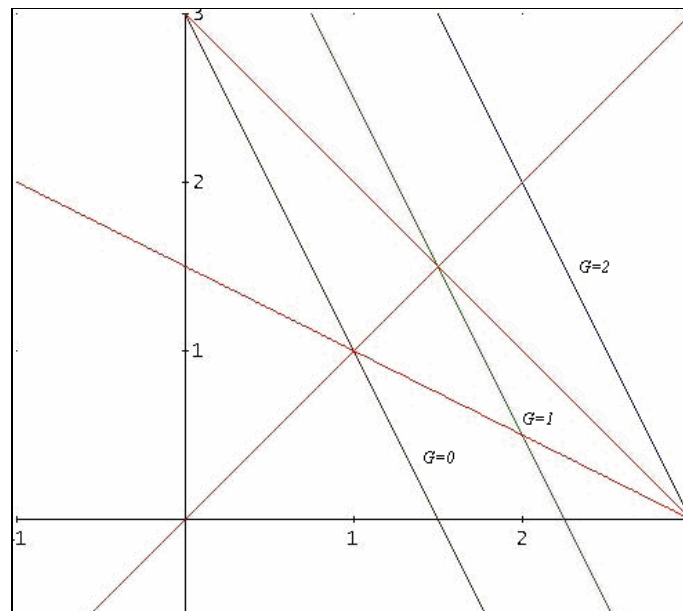
határvonalakon metszik egymást (12. ábra).

29. táblázat: A sajátértékek együtthatói a GINI-együttható képletében

i	m				
	2	3	4	5	10
1	2	4	6	8	18
2	-2	0	2	4	14
3		-4	-2	0	10
4			-6	-4	6
5				-8	2
6					-2
7					-6
8					-10
9					-14
10					-18

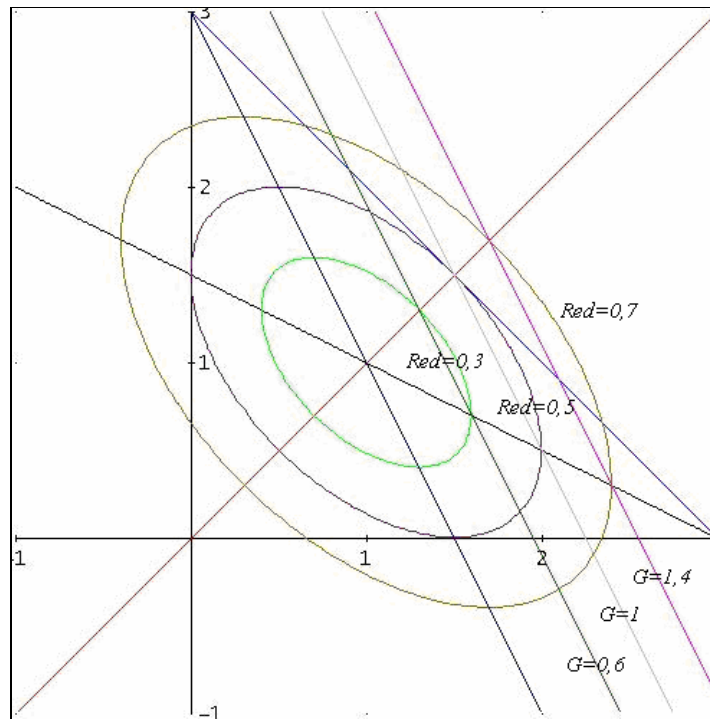
(forrás: saját szerkesztés)

11. ábra: A GINI-együtthatók azonos értékei három tényezőváltozó esetén



(forrás: saját szerkesztés)

12. ábra: A *Red*-mutató és a GINI-együttható kapcsolata ($m=3$)



(forrás: saját szerkesztés)

A fejezetben elért fontosabb eredményeim az alábbiak szerint foglalhatóak össze.

- A *Red*-mutató kifejezhető a sajátértékek ismerete nélkül, pusztán a korrelációs mátrix elemei alapján.
- A *Red*-mutató megadja a tényezőváltozók páronkénti együttmozgásának átlagos mértékét.
- A *Red*-mutató segítségével megkülönböztethetők az extrém multikollinearitás különböző esetei is.
- A *Red*-mutató alapján elkészíthető a multikollinearitás elliptikus modellje.
- A mutató segítségével mérni lehet, hogy a becsült paraméterek varianciáinak összege legalább milyen mértékben „fújódik fel” a hibtag szórásnégyzetéhez viszonyítva.
- A *Red*-mutató alapján megadható a szükséges feltétele annak, hogy a tényezőváltozók korrelációs mátrixában k db zéró sajátérték legyen.

- **Kombinatorikus úton elkészíthető a *Red*-mutató empirikus eloszlásfüggvénye.**
- **A *Red*-mutató segítségével kifejezhető a faktoranalízis során használt *KMO*-mutató.**

IV. A kutatás eredményeinek értékelése, további kutatási irányok

IV. 1. Értékelés, tézisek

A disszertáció céljából (a *Red*-mutató tulajdonságainak vizsgálata, valamint más mérőszámokkal történő összehasonlítása, a többváltozós, lineáris regressziós modellen bemutattva) adódó feladatok elvégzése egészében véve sikeresnek mondható.

Először összefoglaltam a multikollinearitás ismert és kevésbé ismert szakirodalmát. A disszertációban a multikollinearitás 18 mérőszáma, négy nem metrikus detektálási módja, továbbá negatív következményeinek csökkentésére használt 8 eljárás került bemutatásra. A bemutatott mérőszámok, mutatók alkalmazásának előnyei mellett szót ejtettem ezek hátrányairól is. Összességében megállapítottam, hogy a jelenleg használt mutatók általánosan nem, csak bizonyos esetekben jellemzik megfelelően a multikollinearitás mértékét. A mérőszámok értékeinek értelmezése általában szubjektív módon kezeli a „közelség” és a „nagyság” fogalmát, ezért ezek értékei egymással közvetlenül nem összehasonlíthatóak. Az ismertetett eljárások pedig nem minden esetben csökkentik a multikollinearitás negatív következményeinek mértékét. Pontosabban, ha csökkentik is, általában más negatív következményekkel kell szembe néznünk.

A *Red*-mutatót sikerült kifejeznem – a sajátértékek ismerete nélkül – a tényezőváltozók korrelációs mátrixának főátlón kívüli elemeinek négyzetes átlagaként. Ez azt jelenti, hogy a mutató nem csak a (2) becslőfüggvény szempontjából hasznos tartalmat hordozó adatok arányát mutatja, hanem a tényezőváltozók együttmozgásának átlagos mértékét is. Ezt az eredményt több nemzetközi konferencián elismerték, illetve rangos nemzetközi szaklapban is hivatkoztak erre.

Megállapítottam, hogy a multikollinearitást nem csak változók, hanem változócsoportok is okozhatják. Mivel ennek nincs bőséges szakirodalma, ezért a későbbiekben vizsgálni fogom a változócsoportok együttmozgásának hatását.

Megállapítottam, hogy ennek egyik speciális esete a *Red*-mutató segítségével, míg egy másik speciális esete a *VIF_j*-mutatók harmonikus átlagának segítségével mérhető.

Új megközelítésként, megalkottam a multikollinearitás elliptikus modelljét. A változók átlagos együttmozgása mértékének növekedésével párhuzamosan, a „lehetséges sajátértékek”, egy nagyobb sugarú m -dimenziós gömbön helyezkednek el. A „lehetséges sajátértékek” az m -dimenziós gömbnek egy metszetén helyezkednek el úgy, hogy rögzített *Red* érték mellett ezek egy $(m-1)$ -dimenziós ellipszoidon helyezkednek el.

Sajnos, minél nagyobb a modell dimenziószáma, annál több feltételt kell megadni a „lehetséges sajátértékek” tartományának körülhatárolásához, vizsgálatához. Ezért ennek a tartománynak, illetve az elliptikus görbéknek a részletes vizsgálata csak három magyarázóváltozó esetén történt meg. Egy sajátérték függvényében behatároltam a *Red*-mutató lehetséges értékeit, illetve a *Red*-mutató értékétől függően meg tudtam adni az egyes sajátértékek lehetséges értékeit. Összehasonlítottam, hogy a „lehetséges sajátértékek” tartományát hogyan „járják be” az ellipszisek, illetve a sajátértékek legnagyobb és legkisebb értékének azonos értékű hányadosait tartalmazó egyenesek.

A későbbiekben megpróbálom a modellt továbbfejleszteni, illetve a vizsgálatot magasabb dimenziókra kiterjeszteni.

Mivel a *Red*-mutató egy szintetikus mutató, ezért a becsült regressziós paraméterek varianciáihoz külön-külön nem lehet kötni. Megállapítottam, hogy a becsült regressziós paraméterek varianciáinak nem az abszolút nagyságát kell vizsgálni, hanem ezeknek a hibatag szórásnégyzetéhez viszonyított „felfújódását”. Ezek összege, átlaga pedig a sajátértékek reciprokösszegétől függ. Bebizonyítottam, hogy a sajátértékek harmonikus átlagának és a becsült regressziós paraméterek számtani átlagának szorzata megegyezik a hibatag szórásnégyzetével, illetve a sajátértékek harmonikus átlagának és a *VIF_j*-mutatók számtani átlagának szorzata egy.

Egy korábbi állításom cáfolása után, megadtam a *Red*-mutatók egy olyan kritikus értékét, amely szükséges feltétele annak, hogy a becsült paraméterek varianciái ne legyenek végtelenek, illetve olyan kritikus értékeket, amelyek szükséges feltételei annak, hogy a zéró sajátértékek száma k darabnál kevesebb legyen. Mivel ennek így önmagában kevés gyakorlati jelentősége van, ezért ez további részletes vizsgálatokat

igényel. Azonban, ezeket a vizsgálatokat három tényezőváltozó esetén az elliptikus modell használatával elvégeztem. Megfigyeltem, hogy a „lehetséges sajátértékek” tartományának alsó határától távolodva a sajátértékek reciprokösszege nő. Ez alapján a *Red*-mutató függvényében megadtam a becsült paraméterek varianciái összegének a hibatag szórásnégyzetéhez viszonyított „felfújódásának” legkisebb és – ha lehetett – legnagyobb mértékét. Ez alapján a *Red*-mutató függvényében megadható egy olyan kritikus érték, amely szükséges feltétele annak, hogy a becsült paraméterek varianciáinak összege egy előre adott mértéknél jobban ne „fújódjanak fel” a hibatag szórásnégyzetéhez viszonyítva.

A *Red*-mutató eloszlásának vizsgálata során elkészítettem néhány dimenzióban az empirikus eloszlásfüggvényt. Az elemzés során csak létező korrelációs struktúrákat vizsgáltam. A „lehetséges sajátértékek” generálásához és a *Red*-mutató eloszlásának elkészítéséhez egy saját magam által írt algoritmust használtam. Az elemzést nehezítette, hogy a legenerált sajátértékek száma már durvább közelítés esetén is több százezer, több százmillió lehet. **A *Red*-mutató eloszlásának azonosítása sikertelen volt.**

További vizsgálatok elvégzéséhez nagy teljesítményű számítógépekre lenne szükség.

Javasoltam a *Red*-mutató egy alkalmazási lehetőségét.

A *Red*-mutató alapján kifejezhető a faktoranalízis során használt *KMO*-mutató. Ezek alapján megállapítottam, hogy a parciális korrelációs együtthatók átlagos együttlmozgása nem lehet kisebb a korrelációs együtthatók átlagos együttlmozgásánál.

Megadtam a multikollinearitás egy másik lehetséges mérőszámát, ami a *Red*-mutatóval megegyező gondolatmeneten alapul. Ez a mutató a sajátértékek **GINI-együtthatója**. Megadtam a mérőszám egy könnyen kezelhető kiszámítási módját. Ez a mérőszám, szemben a *Red*-mutatóval, nem normált, illetve százalékban sem fejezhető ki. Megvizsgáltam a mutató viselkedését – három tényezőváltozó esetén – a „lehetséges sajátértékek” tartományán. A mutató viselkedése további részletes vizsgálatokat igényel.

Összefoglalva értekezésem az alábbi téziseket tartalmazza.

1. Tézis: A Red-mutató kifejezhető a tényezőváltozók korrelációs mátrixa sajátértékeinek ismerete nélkül, pusztán a páronkénti korrelációs együtthatók négyzetes átlagaként.

2. Tézis: Tényezőváltozók két csoportja együttmozgásának vizsgálata egy-egy elemű csoportok esetén a Red-mutatóval, míg egy-(m-1) elemű csoportok esetén a VIF_j mutatók harmonikus átlagának segítségével lehetséges.

3. Tézis: Új megközelítésként megalkotható a multikollinearitás elliptikus modellje a Red-mutató alapján.

4. Tézis: Megadható a Red-mutató egy olyan kritikus értéke, amely szükséges feltétele annak, hogy a becsült paraméterek varianciái ne legyenek végtelenek.

5. Tézis: A Red-mutató alapján kifejezhető a faktoranalízis során használt KMO-mutató.

6. Tézis: A Red-mutató definiálásának gondolatmenetén alapuló hasonló multikollinearitás mérőszám a tényezőváltozók korrelációs mátrixa sajátértékeinek GINI-együtthatója.

IV. 2. Kutatási irányok a jövőre vonatkozóan

Dolgozatom zárásaként a jövőben tervezett kutatási irányokat foglalom össze a dolgozat felépítésében szereplő sorrendben.

1. Nagyon fontos gyakorlati probléma a multikollinearitás negatív következményeinek csökkenthetősége. Ezért
 - a. egyrészt megszeretném vizsgálni azt, hogy a ridge-regresszióban alkalmazott torzító paraméterre lehetséges-e valamilyen optimális becslés a Red-mutató értéke alapján.

- b. Másrészt egy változószelekciós eljárást szeretnék készíteni a *Red*-mutató értéke alapján, úgy, hogy a mutatót tényezőváltozónként, parciálisan is definiálom, mint egy adott tényezőváltozóknak az összes többi tényezőváltozóval vett átlagos együttmozgása.
2. Szeretném folytatni a multikollinearitás kiterjesztésének vizsgálatát, azaz két vagy több tetszőleges számú tényezőváltozóból álló csoport együttmozgása hogyan mérhető, illetve a jelenségnek milyen negatív következményei vannak.
 3. A későbbiekben az elliptikus modell további tulajdonságaira szeretnék fényt deríteni mind három tényezőváltozó esetén, mind pedig a magasabb dimenziókban.
 4. Szeretném mélyebben megvizsgálni a *Red*-mutató, illetve a parciálisan definiálandó *Red*-mutató kapcsolatát a becsült regressziós paraméterek „felfújódásával”.
 5. A *Red*-mutató elméleti eloszlásának, illetve az empirikus eloszlásának meghatározása egy komoly jövőbeni feladatot jelenthet.
 6. A *Red*-mutató hipotetikus értékére vonatkozóan valamilyen statisztikai tesztet szeretnék készíteni.
 7. A *Red*-mutató alkalmazási körét mind elméleti módszerek, mind közgazdasági vizsgálatok során szeretném gyarapítani.

V. Irodalomjegyzék

- [1] BELSLEY D. A. – KUH E. – WELSCH R. E. [1980]: Regression Diagnostics, Identifying Influential Data and Sources of Collinearity, John Wiley and Sons, New York.
- [2] BOLLA M. – KRÁMLI A. [2005]: Statisztikai következtetések elmélete, Typotex Kiadó, Budapest.
- [3] CASELLA G. [1980]: Minimax ridge estimation. *The Annals of Statistics*, 8. évfolyam 5. szám, 1036.-1056. oldal.
- [4] CHARLOTTE M. – WILLIAM P. [1991]: Collinearity, power, and interpretation of multiple regression analysis, *Journal of Marketing Research*, 28. évfolyam, August, 268.-280. oldal.
- [5] CHATTERJEE S. – PRICE B. [1977]: Regression Analysis by Example, John Wiley and Sons, New York.
- [6] CORRADI E. [1967]: A multikollinearitás vizsgálata Frisch „sugárkéve-térképek” módszerével, *KSH Statisztikai és Matematikai módszerek Közgazdasági Alkalmazásának laboratóriumának 6. számú munkaanyaga*, Központi Statisztikai Hivatal, Budapest.
- [7] CURTO J. D. – PINTO J. C. [2007]: New multicollinearity indicators in linear regression models, *International Statistical Review*, 75. évfolyam, 1. szám, International Statistical Institute, Voorburg, The Netherlands, 114.-121. oldal.
- [8] ÉLTETŐ Ö. – MESZÉNA GY. – ZIERMANN M. [1982]: Sztochasztikus módszerek és modellek, Közgazdasági és Jogi Könyvkiadó, Budapest.
- [9] FARRAR D. E – GLAUBER R. R. [1967]: Multicollineanty in regression analysis: the problem revisited, *Review of Economic and Statistics* 49. évfolyam, 92.-107. oldal.
- [10] FEKETE Z. – ZALAY M. [2000]: Többváltozós függvények analízise, 2. kiadás, Műszaki Könyvkiadó, Budapest.
- [11] FENG-JENQ L. [2006]: Solving multicollinearity in the process of fitting regression model using the nested estimate procedure, *Quality & Quantity* online , Springer, 10 oldal, letöltve 2007.január 6-án.

- [12] FELLMAN J. [1981]: LESKINEN'S preliminary orthogonalizing ridge estimator and a new measure of multicollinearity, *Swedish School of Economics and Business Administration 75. számú munkanyaga*, Swedish School of Economics and Business Administration, Helsinki.
- [13] FÖRSTER E. – EGERMAYER F. [1966]: Korrelations- und Regressionsanalyse, Verlag der Wirtschaft, Berlin.
- [14] FÜSTÖS L. – KOVÁCS E. – MESZÉNA GY. – SIMONNÉ M. N. [2004]: Alakfelismerés (Sokváltozós statisztikai módszerek), Új Mandátum Kiadó, Budapest.
- [15] HAJDU O. [2003]: Többváltozós statisztikai számítások, Központi Statisztikai Hivatal, Budapest.
- [16] HAVASI É. [2007]: Az indikátorok, indikátorrendszerek jellemzői és statisztikai követelményei, *Statisztikai Szemle*, 85. évfolyam 8. szám, Központi Statisztikai Hivatal, Budapest, 677.-689. oldal.
- [17] HEINCZINGER M. [1983]: A multikollinearitás felismerése, mérése és kiszűrése, *Statisztikai szemle*, 61. évfolyam, 7. szám, 741.-761. oldal.
- [18] HERMAN S. – PINTÉR J. – RAPPAI G. – RÉDEY K. [1994]: Statisztika II., JPTE, Pécs.
- [19] HOERL, A. E. – KENNARD, R. [1970]: Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12. évfolyam 1. szám, 55.-67. oldal.
- [20] <http://portal.ksh.hu>
- [21] <http://www.ecostat.hu>
- [22] <http://www.mathworld.com>
- [23] <http://www.xycoon.com>
- [24] HULYÁK K. [1969]: A multikollinearitás feltárása és elemzése, *KSH Statisztikai és Matematikai módszerek Közgazdasági Alkalmazásának laboratóriumának 9. számú munkanyaga*, Központi Statisztikai Hivatal, Budapest.
- [25] HUNYADI L. – MUNDRUCZÓ GY. – VITA L. [1997]: Statisztika, Aula Kiadó, Budapest.
- [26] HUNYADI L. [2001]: Statisztikai következtetéselemélet közgazdászoknak, Központi Statisztikai Hivatal, Budapest.
- [27] JÁNOSA A. [2005]: Adatelemzés számítógéppel, Perfekt Kiadó, Budapest.

- [28] KATONA T. – LENGYEL I. (szerk.) [1999]: Statisztikai ismerettár (fogalmak, képletek, módszerek EXCEL és SPSS alkalmazásokkal), JATEPress, Szeged.
- [29] KENDALL M. G. [1957]: A Course in Multivariate Analysis, Griffin, London.
- [30] KENNEDY P. [2003]: A Guide to econometrics, 5. kiadás, MIT, Cambridge.
- [31] KETSKEMÉTY L. – IZSÓ L. [2005]: Bevezetés az SPSS programrendszerbe, ELTE Eötvös Kiadó, Budapest.
- [32] KÖRPÁS ATTILÁNÉ (szerk.) [1997]: Általános statisztika II., Nemzeti Tankönyvkiadó, Budapest.
- [33] KOVÁCS P. – PETRES T. – TÓTH L. [2004]: Adatállományok redundanciájának mérése, *Statisztikai Szemle*, 82. évfolyam 6-7 szám, Központi Statisztikai Hivatal, Budapest, 595.-604. oldal.
- [34] KOVÁCS P. – PETRES T. – TÓTH L. [2005]: A new measure of multicollinearity in linear regression models, *International Statistical Review (ISR)*, 73. évfolyam 3. szám, International Statistical Institute, Voorburg, The Netherlands, 405.-412. oldal.
- [35] KOVÁCS P. – PETRES T. – TÓTH L. [2006]: Válogatott fejezetek Statisztikából, többváltozós statisztikai módszerek, JATEPress, Szeged.
- [36] KOVÁCS P. – LAMPERTNÉ A. I. – PETRES T. [2005]: A multikollinearitás mérése lineáris regressziós modellekben, *A Dunaújvárosi Főiskola Közleményei XXVII/II.*, Dunaújváros, 355.-365. oldal.
- [37] KOVÁCS P. – PETRES T. [2006]: A PETRES-féle *Red*-mutató eloszlásának vizsgálata, *A Dunaújvárosi Főiskola Közleményei XXVIII/II.*, Dunaújváros, 521.-530. oldal.
- [38] KOVÁCS P. – PETRES T. – LUKOVICS M. [2006]: A PETRES-féle *Red*-mutató alkalmazásának lehetőségei, *A Dunaújvárosi Főiskola Közleményei XXVIII.*, Dunaújváros, 304.-316. oldal.
- [39] KÖRÖSI G. – MÁTYÁS L. – SZÉKELY I. [1990]: Gyakorlati ökonometria, Közgazdasági és Jogi Könyvkiadó, Budapest.
- [40] LAWLESS J. F. [1978]: Ridge and related estimation procedures: Theory and practice. *Communication in Statistics – Theory and Methods*, 7. évfolyam 2. szám, 139.-164. oldal.
- [41] LEINDLER L. [1997]: Analízis, Nemzeti Tankönyvkiadó, Budapest.
- [42] MADDALA GS. [2004]: Bevezetés az ökonometriába, Nemzeti Tankönyvkiadó, Budapest.

- [43] Magyar statisztikai évkönyv 1994, Központi Statisztikai Hivatal, Budapest.
- [44] Magyar statisztikai évkönyv 1996, Központi Statisztikai Hivatal, Budapest.
- [45] Magyar statisztikai évkönyv 2003, Központi Statisztikai Hivatal, Budapest.
- [46] MALEVERGNE Y. – SORNETTE D. [2004]: Collective origin of the coexistence of apparent random matrix theory noise and factors in large sample correlation matrices, *Physica A*, 331.évfolyam, 3.-4. szám, 660.-668. oldal.
- [47] MÓRI F. T. – SZÉKELY J. G. (szerk.) [1986]: Többváltozós statisztikai analízis, Műszaki Könyvkiadó, Budapest.
- [48] MUNDRUCZÓ GY. [1981]: Alkalmazott regressziószámítás, Akadémiai kiadó, Budapest.
- [49] PETRES T. – TÓTH L. [2004]: Piaci információk és a multikollinearitás, *A szociális identitás, az információ és a piac*, SZTE Gazdaságtudományi Kar Közleményei, JATEPress, Szeged.
- [50] PETRES T. – TÓTH L. [2006]: Statisztika, Központi Statisztikai Hivatal, Budapest.
- [51] RAMANATHAN R. [2002]: Bevezetés az ökonometriába, alkalmazásokkal, Panem Kiadó, Budapest.
- [52] RAPPAI G. [2001]: Üzleti statisztika Excellel, Központi Statisztikai Hivatal, Budapest.
- [53] RÓZSA P. [1991]: Lineáris algebra és alkalmazásai, Tankönyvkiadó, Budapest.
- [54] SAJTOS L. – MITEV A. [2007]: SPSS kutatási és adatelemzési kézikönyv, Alinea Kiadó, Budapest.
- [55] SILVEY S. D. [1969]: Multicollinearity and Imprecise Estimation, *Journal of Royal Statistical Society, Series B*, 31. évfolyam, 539.-552. oldal.
- [56] SZÉKELYI M. – BARNA I. [2003]: Túlélőkészlet az SPSS-hez, Többváltozós elemzési technikákról társadalomkutatók számára, Typotex Kiadó, Budapest.
- [57] THEIL, H. [1971]: Principles of Econometrics, John Wiley and Sons, New York.
- [58] THISTED R. A. [1980]: Comment on Smith and Campbell, *Journal of the American Statistical Association*, 75. évfolyam, 81.-86. oldal.
- [59] TRIČKOVIĆ V. [1976]: Teorijski modeli i metodi kvantitativnog istraživanja tržišta, Institut za ekonomiku industrije, Beograd.

- [60] VINOD H. D. [1976]: Application of new ridge regression methods to a study of Bell system scale economies, *Journal of the American Statistical Association*, 71. évfolyam december, 835.-841. oldal.
- [61] WICHERN D. W. – CHURCHILL G. A. [1978]: A comparison of ridge estimators, *Techometrics* 20. évfolyam 3. szám, 301.-311. oldal.

VI. Mellékletek

VI. 1. Ábrajegyzék

1. ábra: Stabil regressziós sík a magyarázóváltozók nem szignifikáns együttmozgása esetén ($m=2$).....	14
2. ábra: Instabil regressziós sík szignifikáns multikollinearitás esetén ($m=2$).....	14
3. ábra: Ridge-görbe.....	51
4. ábra: A multikollinearitás elliptikus modellje három tényezőváltozó esetén.....	61
5. ábra: A <i>Red</i> -mutató lehetséges legkisebb és legnagyobb értéke a legnagyobb sajátérték függvényében három tényezőváltozó esetén.....	65
6. ábra: A legnagyobb és a legkisebb sajátérték hányadosának viselkedése három tényezőváltozó esetén	66
7. ábra: A sajátértékek reciprokösszegének viselkedése ($m=3$)	71
8. ábra: Az empirikus eloszlás grafikus ábrázolása ($m=3$).....	77
9. ábra: Az empirikus eloszlás grafikus ábrázolása ($m=4$).....	77
10. ábra: Az empirikus eloszlás grafikus ábrázolása ($m=7$).....	78
11. ábra: A GINI-együtthatók azonos értékei három tényezőváltozó esetén.....	83
12. ábra: A <i>Red</i> -mutató és a GINI-együttható kapcsolata ($m=3$)	84

VI. 2. Táblajegyzék

1. táblázat: Lineárisan függő tényezőváltozók nem szignifikáns korrelációs kapcsolatokkal	6
2. táblázat: Lineáris korrelációs együtthatók.....	6
3. táblázat: Példa a VIF_j magas értékére A havi munkaórák becslése lineáris regressziós modellel.....	21
4. táblázat: Eltérő sajátértékek és kondíciós indexek az $\mathbf{X}'\mathbf{X}$ különböző alakjai alapján ..	25
5. táblázat: A varianciák „felfújódásának” felbontása.....	26
6. táblázat: Variancia-dekompozíció az $\hat{\mathbf{X}}\hat{\mathbf{X}}$ mátrix alapján	27
7. táblázat: A multikollinearitás mérőszámainak és tesztjeinek összefoglaló táblázata	32
8. táblázat: A kéndioxid koncentrációjának becslése lineáris regressziós modellel.....	35
9. táblázat: A kéndioxid koncentrációjának becslése lineáris regressziós modellel stepwise algoritmus alkalmazásával	35
10. táblázat: A magyarázóváltozók és az eredményváltozó közötti korrelációs együtthatók	42
11. táblázat: Az M -mutató kiszámításának mellékszámításai	43
12. táblázat: Az ω -próbafüggvény értékei.....	43
13. táblázat: A tényezőváltozók és eredményváltozó közötti korrelációs és parciális korrelációs együtthatók	44
14. táblázat: A tolerancia-mutató és a VIF_j -mutató értékei.....	45
15. táblázat: A kondíciós indexek és a varianciák felfújódása.....	45
16. táblázat: A tényezőváltozók korrelációs mátrixának sajátértékei.....	46
17. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei.....	46
18. táblázat: VIF_j értékek a stepwise eljárás eredményéül adódó regressziós modellekben	47
19. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei a 2. modellben	48
20. táblázat: VIF_j értékek a backward eljárás eredményéül adódó modellekben	48
21. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei.....	49
22. táblázat: A VIF_j értékek és a tényezőváltozók szignifikanciái a szelekciós eljárás eredményéül adódó regressziós modellben.....	50
23. táblázat: A korrelációs mátrix sajátértékeire épülő mutatók értékei.....	50
24. táblázat: Főkomponenssúly-mátrix rotálás után	51
25. táblázat: A nested estimate eljárás után kapott regressziós modell	52
26. táblázat: A Red-mutató kritikus értékei, mint k darab sajátérték zéró voltának szükséges feltétele	70
27. táblázat: Az előállított „lehetséges sajátérték” kombinációk száma kéttizedes pontossággal	76
28. táblázat: Útmutató a KMO értelmezéséhez	80
29. táblázat: A sajátértékek együtthatói a GINI-együttható képletében	83

VI. 3. Példa a DEF-mutató hiányosságára

$\tilde{y}$	$\tilde{x}_1$	$\tilde{x}_2$	e
40	30	5	-1.6532
40	25	6	1.2447
35	20	7	-0.8575
40	22	8	1.4858
35	15	9	0.97069
32	10	10	0.86854
30	10	11	-2.2012
35	12	12	0.14206

y	x_1	x_2	e
1.150308	1.727558	-1.52753	-1.25929
1.150308	1.007742	-1.09109	0.948145
-0.244	0.287926	-0.65465	-0.65316
1.150308	0.575853	-0.21822	1.131801
-0.244	-0.43189	0.218218	0.739419
-1.08059	-1.15171	0.654654	0.661606
-1.63832	-1.15171	1.091089	-1.67674
-0.244	-0.86378	1.527525	0.108214

Korrelációs mátrix

		y	x_1	x_2	e
y	r	1			
	Szignifikancia				
x_1	r	0.898265	1		
	Szignifikancia	0.002436			
x_2	r	-0.753053	-0.93461	1	
	Szignifikancia	0.03102	0.000665		
e	r	0.366083	-4.1E-07	1.51E-10	1
	Szignifikancia	0.372456	0.999999	1	

Regressziós statisztika

r értéke	0,931
r -négyzet	0,866
Korrigált r -négyzet	0,677
Standard hiba	0,423

Varianciaanalízis

	Négyzetösszeg	Szabadságfok	F	Szignifikancia
Regresszió	6,928	2	19,385	0,004
Maradék	1,072	6		
Összesen	8,000	8		

Regressziós paraméterek

Változó	Standardizált regressziós paraméterek	t -érték	Szig.
x_1	1,537	3,658	0,010
x_2	0,683	1,627	0,155

Ekkor a DEF -mutatóban szereplő felbontás az alábbi lesz.

$$\sum_{i=1}^m \hat{\beta}_i^2 = 1,537^2 + 0,683^2 = 2,830$$

$$\sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \hat{\beta}_i r_{ij} \hat{\beta}_j = 2 \cdot 1,537 \cdot (-0,9346) \cdot 0,683 = -1,964$$

$$\sum_{i=1}^m \hat{\beta}_i^2 + \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \hat{\beta}_i r_{ij} \hat{\beta}_j = 0,866$$

$$DEF = \frac{-1,964}{0,866} = -2,268$$

Tehát, a kapott eredmény egyáltalán nincs összhangban a szerzők állításával.

**VI. 4. A II. 4. alfejezet empirikus példájának deflált adatai 2004-es
árszínvonalon, illetve a számítógépes elemzés néhány kimenete**

Év	Egy főre jutó éves jövedelem (2004-es árszínvonalon, ezer Ft)	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	Marhahús, rostélyos, csontos (Ft/kg)	Sertésszár (Ft/kg)	Napraforgó-étolaj (Ft/liter)	Kenyér, fehér (Ft/kg)	Sör, hazai világos (Ft/0,5 liter)	1 főre jutó marhahús fogyasztás (kg/fő)
1990	641	1725,37	1265,74	338,00	392,45	149,20	116,67	6,50
1991	630	1305,18	997,74	265,10	415,92	183,31	128,78	7,40
1992	608	1467,20	937,65	286,15	370,20	184,13	130,69	7,40
1993	579	1244,65	925,03	371,89	347,83	203,81	122,59	7,90
1994	594	1276,79	1020,82	347,38	332,15	188,93	113,66	8,00
1995	562	1299,32	1111,04	392,82	367,25	168,28	104,36	6,80
1996	558	1125,52	1006,63	352,72	386,40	170,41	110,97	5,10
1997	563	1175,44	903,92	364,27	369,33	166,96	111,30	4,80
1998	584	1098,24	894,81	431,93	367,06	153,31	109,09	4,30
1999	588	868,16	901,11	300,85	392,53	151,86	117,47	4,10
2000	614	1018,04	878,98	287,31	317,48	154,80	116,76	4,30
2001	643	1187,40	900,35	376,97	288,20	166,01	114,13	3,90
2002	684	982,19	925,32	369,69	318,28	165,16	118,13	4,30
2003	717	823,25	947,85	270,51	302,46	166,14	124,61	4,10
2004	747	878,00	959,00	310,00	278,00	178,00	124,00	3,90

Korrelációs mátrixok

Korrelációs mátrix

		marhahús- fogyasztás	jövedelem	sertés_ár	marha_ár	sertészsír_ár	étolaj_ár	kenyér_ár	sör_ár
marhahús- fogyasztás	<i>r</i>	1,000	-0,400	0,713	0,450	-0,042	0,458	0,664	0,201
	<i>Szig.</i>		<i>0,139</i>	<i>0,003</i>	<i>0,092</i>	<i>0,882</i>	<i>0,086</i>	<i>0,007</i>	<i>0,473</i>
jövedelem	<i>R</i>	-0,400	1,000	-0,365	-0,023	-0,416	-0,642	-0,048	0,526
	<i>Szig.</i>	<i>0,139</i>		<i>0,181</i>	<i>0,936</i>	<i>0,123</i>	<i>0,010</i>	<i>0,865</i>	<i>0,044</i>
sertés_ár	<i>r</i>	0,713	-0,365	1,000	0,684	0,141	0,501	0,147	-0,012
	<i>Szig.</i>	<i>0,003</i>	<i>0,181</i>		<i>0,005</i>	<i>0,617</i>	<i>0,057</i>	<i>0,602</i>	<i>0,966</i>
marha_ár	<i>r</i>	0,450	-0,023	0,684	1,000	0,042	0,365	-0,112	-0,169
	<i>Szig.</i>	<i>0,092</i>	<i>0,936</i>	<i>0,005</i>		<i>0,882</i>	<i>0,181</i>	<i>0,691</i>	<i>0,548</i>
sertészsír_ár	<i>r</i>	-0,042	-0,416	0,141	0,042	1,000	-0,022	-0,090	-0,758
	<i>Szig.</i>	<i>0,882</i>	<i>0,123</i>	<i>0,617</i>	<i>0,882</i>		<i>0,939</i>	<i>0,751</i>	<i>0,001</i>
étolaj_ár	<i>r</i>	0,458	-0,642	0,501	0,365	-0,022	1,000	-0,085	-0,067
	<i>Szig.</i>	<i>0,086</i>	<i>0,010</i>	<i>0,057</i>	<i>0,181</i>	<i>0,939</i>		<i>0,763</i>	<i>0,813</i>
kenyér_ár	<i>r</i>	0,664	-0,048	0,147	-0,112	-0,090	-0,085	1,000	0,427
	<i>Szig.</i>	<i>0,007</i>	<i>0,865</i>	<i>0,602</i>	<i>0,691</i>	<i>0,751</i>	<i>0,763</i>		<i>0,113</i>
sör_ár	<i>r</i>	0,201	0,526	-0,012	-0,169	-0,758	-0,067	0,427	1,000
	<i>Szig.</i>	<i>0,473</i>	<i>0,044</i>	<i>0,966</i>	<i>0,548</i>	<i>0,001</i>	<i>0,813</i>	<i>0,113</i>	

Tényezőváltozók korrelációs mátrixa (determináns=0,005)

	jövedelem	sertés_ár	marha_ár	sertészsír_ár	étolaj_ár	kenyér_ár	sör_ár
jövedelem	1,000	-0,365	-0,023	-0,416	-0,642	-0,048	0,526
sertés_ár	-0,365	1,000	0,684	0,141	0,501	0,147	-0,012
marha_ár	-0,023	0,684	1,000	0,042	0,365	-0,112	-0,169
sertészsír_ár	-0,416	0,141	0,042	1,000	-0,022	-0,090	-0,758
étolaj_ár	-0,642	0,501	0,365	-0,022	1,000	-0,085	-0,067
kenyér_ár	-0,048	0,147	-0,112	-0,090	-0,085	1,000	0,427
sör_ár	0,526	-0,012	-0,169	-0,758	-0,067	0,427	1,000

Tényezőváltozók korrelációs mátrixának inverze

	jövedelem	sertés_ár	marha_ár	sertészsír_ár	étolaj_ár	kenyér_ár	sör_ár
jövedelem	17,127	8,563	-11,083	-5,766	9,997	5,917	-17,001
sertés_ár	8,563	7,235	-7,399	-4,147	4,062	2,616	-9,653
marha_ár	-11,083	-7,399	9,535	4,710	-6,293	-3,637	12,050
sertészsír_ár	-5,766	-4,147	4,710	5,330	-2,868	-2,645	8,755
étolaj_ár	9,997	4,062	-6,293	-2,868	7,292	3,594	-9,490
kenyér_ár	5,917	2,616	-3,637	-2,645	3,594	3,536	-6,967
sör_ár	-17,001	-9,653	12,050	8,755	-9,490	-6,967	20,834

Regressziószámítás; változószelekció: enter eljárás

Regressziós statisztika

r értéke	0,957
r -négyzet	0,917
Korrigált r -négyzet	0,834
Standard hiba	0,654

Variansciaanalízis

	Négyzetösszeg	Szabadságfok	F	Szignifikancia
Regresszió	32,934	7,000	11,014	0,003
Maradék	2,990	7,000		
Összesen	35,924	14,000		

Regressziós paraméterek

Változó	Regressziós paraméterek	Standardizált regressziós paraméterek	t -érték	Szig.	VIF_j	Parciális korreláció
Tengelymetszet	-9,324		-1,151	0,287		
Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	-0,011	-0,401	-0,888	0,404	17,127	-0,097
Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,002	0,234	0,798	0,451	7,235	0,087
Marhahús, rostélyos, csontos (Ft/kg)	0,006	0,377	1,120	0,300	9,535	0,122
Sertészsír (Ft/kg)	-0,002	-0,064	-0,254	0,807	5,330	-0,028
Napraforgó-étolaj (Ft/liter)	0,000	0,005	0,018	0,986	7,292	0,002
Kenyér, fehér (Ft/kg)	0,059	0,567	2,767	0,028	3,536	0,302
Sör, hazai világos (Ft/0,5 liter)	0,040	0,188	0,377	0,717	20,834	0,041

Regressziószámítás; változószelekció: stepwise eljárás

Regressziós statisztika

Modell	r értéke	r - négyzet	Korrigált r - négyzet	Standard hiba
1	0,713	0,508	0,471	1,166
2	0,910	0,828	0,800	0,717

Varianciaanalízis

Modell		Négyzetösszeg	Szabadságfok	F	Szignifikancia
1	Regresszió	18,261	1,000	13,440	0,003
	Maradék	17,663	13,000		
	Összesen	35,924	14,000		
2	Regresszió	29,758	2,000	28,959	0,000
	Maradék	6,166	12,000		
	Összesen	35,924	14,000		

Regressziós paraméterek

	Változó	Regressziós paraméterek	Standardizált regressziós paraméterek	t -érték	Szig.	VIF_j	Parciális korreláció
1	Konstans	-0,009		-0,006	0,996		
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,005	0,713	3,666	0,003	1,000	0,713
2	Konstans	-9,534		-4,286	0,001		
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,004	0,629	5,204	0,000	1,022	0,832
	Kenyér, fehér (Ft/kg)	0,060	0,572	4,730	0,000	1,022	0,807

Regressziószámítás; változószelekció: backward eljárás

Regressziós statisztika

Modell	r értéke	r - négyzet	Korrigált r - négyzet	Standard hiba
eredeti	0,957	0,917	0,834	0,654
1	0,957	0,917	0,854	0,611
2	0,957	0,916	0,869	0,580
3	0,951	0,905	0,867	0,584

Varianciaanalízis

Modell		Négyzetösszeg	Szabadságfok	F	Szignifikancia
eredeti	Regresszió	32,934	7,000	11,014	0,003
	Maradék	2,990	7,000		
	Összesen	35,924	14,000		
1	Regresszió	32,934	6,000	14,684	0,001
	Maradék	2,990	8,000		
	Összesen	35,924	14,000		
2	Regresszió	32,901	5,000	19,590	0,000
	Maradék	3,023	9,000		
	Összesen	35,924	14,000		
3	Regresszió	32,508	4,000	23,791	0,000
	Maradék	3,416	10,000		
	Összesen	35,924	14,000		

Regressziós paraméterek

Modell	Változó	Regressziós paraméterek	Standardizált regressziós paraméterek	t-érték	Szig.	VIF_j
1	Konstans	-9,324		-1,151	0,287	
	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	-0,011	-0,401	-0,888	0,404	17,127
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,002	0,234	0,798	0,451	7,235
	Marhahús, rostélyos, csontos (Ft/kg)	0,006	0,377	1,120	0,300	9,535
	Sertészsír (Ft/kg)	-0,002	-0,064	-0,254	0,807	5,330
	Napraforgó-étolaj (Ft/liter)	0,000	0,005	0,018	0,986	7,292
	Kenyér, fehér (Ft/kg)	0,059	0,567	2,767	0,028	3,536
	Sör, hazai világos (Ft/0,5 liter)	0,040	0,188	0,377	0,717	20,834
2	Konstans	-9,325		-1,231	0,253	
	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	-0,011	-0,408	-2,162	0,063	3,422
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,002	0,231	1,016	0,339	4,973
	Marhahús, rostélyos, csontos (Ft/kg)	0,006	0,382	1,847	0,102	4,104
	Sertészsír (Ft/kg)	-0,002	-0,062	-0,296	0,775	4,202
	Kenyér, fehér (Ft/kg)	0,059	0,565	4,168	0,003	1,764
	Sör, hazai világos (Ft/0,5 liter)	0,042	0,195	0,655	0,531	8,484

Regressziós paraméterek

Modell	Változó	Regressziós paraméterek	Standardizált regressziós paraméterek	<i>t</i>-érték	Szig.	<i>VIF_j</i>
3	Konstans	–11,300		–3,339	0,009	
	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	–0,012	–0,435	–2,778	0,021	2,622
	Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,001	0,194	1,082	0,308	3,426
	Marhahús, rostélyos, csontos (Ft/kg)	0,007	0,415	2,511	0,033	2,915
	Kenyér, fehér (Ft/kg)	0,057	0,547	4,772	0,001	1,404
	Sör, hazai világos (Ft/0,5 liter)	0,058	0,269	1,764	0,112	2,479
4	Konstans	–12,167		–3,669	0,004	
	Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	–0,015	–0,541	–4,405	0,001	1,588
	Marhahús, rostélyos, csontos (Ft/kg)	0,009	0,558	5,622	0,000	1,036
	Kenyér, fehér (Ft/kg)	0,058	0,554	4,805	0,001	1,399
	Sör, hazai világos (Ft/0,5 liter)	0,074	0,343	2,506	0,031	1,971

Regressziószámítás; változószelekció: VIF_j –mutató értékei alapján

Regressziós statisztika

r értéke	0,957
r -négyzet	0,915
Korrigált r -négyzet	0,851
Standard hiba	0,618

Varianciaanalízis

	Négyzetösszeg	Szabadságfok	F	Szignifikancia
Regresszió	32,873	6,000	14,366	0,001
Maradék	3,051	8,000		
Összesen	35,924	14,000		

Regressziós paraméterek

Változó	Regressziós paraméterek	Standardizált regressziós paraméterek	t -érték	Szig.	VIF_j
Tengelymetszet	-7,637		-1,198	0,265	
jövedelem	-0,007	-0,247	-1,332	0,220	3,254
sertés_ár	0,002	0,321	1,875	0,098	2,763
marha_ár	0,004	0,269	1,628	0,142	2,566
sertészsír_ár	-0,005	-0,143	-1,079	0,312	1,650
étolaj_ár	0,003	0,091	0,512	0,623	2,969
kenyér_ár	0,066	0,630	5,569	0,001	1,206

Főkomponensanalízis

Kommunalitások

Változó	Kommunalitások
Zscore: Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	0,882
Zscore: Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,801
Zscore: Marhahús, rostélyos, csontos (Ft/kg)	0,836
Zscore: Sertészsír (Ft/kg)	0,731
Zscore: Napraforgó-étolaj (Ft/liter)	0,662
Zscore: Kenyér, fehér (Ft/kg)	0,731
Zscore: Sör, hazai világos (Ft/0,5 liter)	0,932

Megőrzött varianciarányad

Főkomponens	Sajátérték	Variancia százaléka	Sajátérték	Variancia százaléka	Kumulált százalék
1,000	2,615	37,363	2,615	37,363	37,363
2,000	1,814	25,910	1,814	25,910	63,273
3,000	1,147	16,379	1,147	16,379	79,652
4,000	,929	13,277			
5,000	,273	3,906			
6,000	,205	2,924			
7,000	,017	0,240			

Főkomponenssúly-mátrix

Változó	Főkomponens		
	1,000	2,000	3,000
Zscore: Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	-0,785	0,077	0,510
Zscore: Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,650	0,612	0,065
Zscore: Marhahús, rostélyos, csontos (Ft/kg)	0,536	0,486	0,559
Zscore: Sertészsír (Ft/kg)	0,592	-0,600	-0,143
Zscore: Napraforgó-étolaj (Ft/liter)	0,666	0,451	-0,125
Zscore: Kenyér, fehér (Ft/kg)	-0,201	0,414	-0,720
Zscore: Sör, hazai világos (Ft/0,5 liter)	-0,675	0,680	-0,120

Rotált főkomponenssúly-mátrix

Változó	Főkomponens		
	1,000	2,000	3,000
Zscore: Egy főre jutó éves jövedelem (2004-es árszínvonalon ezer Ft)	-0,412	0,758	-0,371
Zscore: Sertéshús, comb, csont és csülök nélkül (Ft/kg)	0,890	-0,046	0,076
Zscore: Marhahús, rostélyos, csontos (Ft/kg)	0,807	0,104	-0,419
Zscore: Sertészsír (Ft/kg)	-0,024	-0,845	-0,126
Zscore: Napraforgó-étolaj (Ft/liter)	0,758	-0,218	0,198
Zscore: Kenyér, fehér (Ft/kg)	0,025	0,188	0,834
Zscore: Sör, hazai világos (Ft/0,5 liter)	-0,024	0,875	0,408

Nested estimate eljárás

1. iteráció

Regressziós statisztika

<i>r</i> értéke	<i>r</i> – négyzet	Korrigált <i>r</i> – négyzet	Standard hiba
0,713	0,508	0,471	1,166

Varianciaanalízis

	Négyzetösszeg	Szabadságfok	<i>F</i>	Szignifikancia
Regresszió	18,261	1	13,440	0,003
Maradék	17,663	13		
Összesen	35,924	14		

Regressziós paraméterek

Változó	Regressziós paraméterek	Standardizált regressziós paraméterek	<i>t</i> –érték	Szig.
Tengelymetszet	–0,0088		–0,006	0,996
sértés_ár	0,00475	0,713	3,666	0,003

Hibatag az első iteráció végén (hibatag1)

–1,679
1,215
0,446
2,002
1,949
0,642
–0,233
–0,770
–0,903
–0,011
–0,523
–1,726
–0,353
0,202
–0,258

2. iteráció: függőváltozó: hibatag1

Korrelációs mátrix

	Hibatag1	Kenyér_ár
Hibatag1	1,000	0,798
Kenyér_ár	0,798	1,000

Regressziós statisztika

<i>r</i> értéke	<i>r</i> -négyzet	Korrigált <i>r</i> - négyzet	Standard hiba
0,798	0,637	0,609	0,702

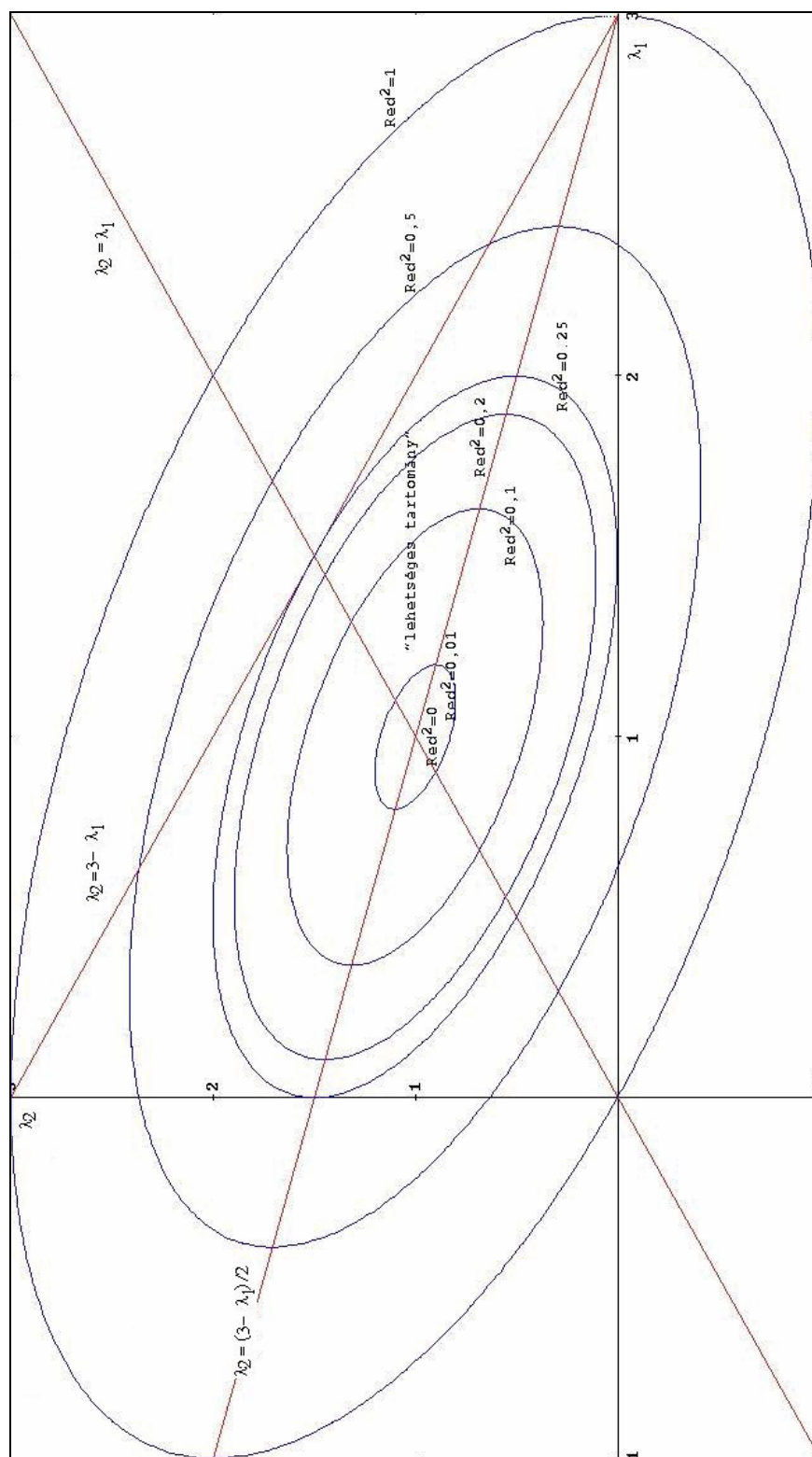
Varianciaanalízis

	Négyzetösszeg	Szabadságfok	<i>F</i>	Szignifikancia
Regresszió	11,250	1	22,806	0,000
Maradék	6,413	13		
Összesen	17,663	14		

Regressziós paraméterek

Változó	Regressziós paraméterek	Standardizált regressziós paraméterek	<i>t</i> -érték	Szig.
Tengelymetszet	-9,957	2,093	-4,758	0,000
kenyér_ár	0,059	0,798	4,776	0,000

VI. 5. A multikollinearitás elliptikus modellje három tényezőváltozó esetén



**VI. 6. A tényezőváltozók korrelációs mátrixa sajátértékeinek és
reciprokösszegüknek a viselkedése három tényezőváltozó esetén, rögzített
Red érték mellett**

<i>Red</i>	λ_1	λ_2	λ_3	A sajátértékek reciprokösszege
0	1	1	1	3,000000
0,1	1,1	1,1	0,8	3,068182
	1,11	1,08965	0,80035	3,068084
	1,12	1,07856	0,80144	3,067776
	1,13	1,06662	0,80338	3,067241
	1,14	1,05369	0,80631	3,066459
	1,15	1,03956	0,81044	3,065411
	1,16	1,02392	0,81608	3,064079
	1,17	1,00624	0,82376	3,062446
	1,18	0,9855	0,8345	3,060493
	1,19	0,95908	0,85092	3,058202
	1,2	0,9	0,9	3,055556
0,2	1,2	1,2	0,6	3,333333
	1,21	1,18983	0,60017	3,333098
	1,22	1,17931	0,60069	3,332377
	1,23	1,16842	0,60158	3,331147
	1,24	1,15713	0,60287	3,329387
	1,25	1,14542	0,60458	3,327076
	1,26	1,13325	0,60675	3,324192
	1,27	1,12059	0,60941	3,320715
	1,28	1,10739	0,61261	3,316627
	1,29	1,09359	0,61641	3,311909
	1,3	1,07913	0,62087	3,306544
	1,31	1,06392	0,62608	3,300515
	1,32	1,04785	0,63215	3,293808
	1,33	1,03077	0,63923	3,286407
	1,34	1,01248	0,64752	3,2783
	1,35	0,99271	0,65729	3,269476
	1,36	0,971	0,669	3,259925
	1,37	0,94662	0,68338	3,249636
	1,38	0,91817	0,70183	3,238604
	1,39	0,88197	0,72803	3,226822
	1,4	0,8	0,8	3,214286

<i>Red</i>	λ_1	λ_2	λ_3	A sajátértékek reciprokösszege
0,3	1,3	1,3	0,4	4,038462
	1,31	1,28989	0,40011	4,037918
	1,32	1,27955	0,40045	4,036264
	1,33	1,26896	0,40104	4,033467
	1,34	1,25814	0,40186	4,029497
	1,35	1,24705	0,40295	4,024323
	1,36	1,23569	0,40431	4,01792
	1,37	1,22405	0,40595	4,010265
	1,38	1,21212	0,40788	4,001337
	1,39	1,19987	0,41013	3,991117
	1,4	1,1873	0,4127	3,979592
	1,41	1,17437	0,41563	3,966749
	1,42	1,16108	0,41892	3,952581
	1,43	1,14739	0,42261	3,937082
	1,44	1,13327	0,42673	3,920251
	1,45	1,11869	0,43131	3,90209
	1,46	1,10362	0,43638	3,882606
	1,47	1,08799	0,44201	3,861807
	1,48	1,07177	0,44823	3,839706
	1,49	1,05487	0,45513	3,816319
	1,5	1,03723	0,46277	3,791667
	1,51	1,01872	0,47128	3,765772
	1,52	0,99923	0,48077	3,738661
	1,53	0,97857	0,49143	3,710363
	1,54	0,9565	0,5035	3,680912
	1,55	0,93267	0,51733	3,650343
	1,56	0,90655	0,53345	3,618693
	1,57	0,87725	0,55275	3,586004
	1,58	0,84304	0,57696	3,552319
	1,59	0,79947	0,61053	3,517683
	1,6	0,7	0,7	3,482143
0,4	1,4	1,4	0,2	6,428571
	1,41	1,38992	0,20008	6,426588
	1,42	1,37966	0,20034	6,420578
	1,43	1,36923	0,20077	6,410468
	1,44	1,35862	0,20138	6,396199
	1,45	1,34782	0,20218	6,377729
	1,46	1,33683	0,20317	6,355035
	1,47	1,32565	0,20435	6,328113

<i>Red</i>	λ_1	λ_2	λ_3	A sajátértékek reciprokösszege
	1,48	1,31426	0,20574	6,296977
	1,49	1,30265	0,20735	6,261663
	1,5	1,29083	0,20917	6,222222
	1,51	1,27878	0,21122	6,178727
	1,52	1,2665	0,2135	6,131268
	1,53	1,25397	0,21603	6,079951
	1,54	1,24118	0,21882	6,024903
	1,55	1,22812	0,22188	5,966262
	1,56	1,21477	0,22523	5,904184
	1,57	1,20113	0,22887	5,838834
	1,58	1,18718	0,23282	5,770393
	1,59	1,17289	0,23711	5,699049
	1,6	1,15826	0,24174	5,625
	1,61	1,14325	0,24675	5,548449
	1,62	1,12784	0,25216	5,469605
	1,63	1,112	0,258	5,38868
	1,64	1,09569	0,26431	5,305889
	1,65	1,07889	0,27111	5,221445
	1,66	1,06154	0,27846	5,135563
	1,67	1,04358	0,28642	5,048451
	1,68	1,02497	0,29503	4,960317
	1,69	1,00561	0,30439	4,871363
	1,7	0,98541	0,31459	4,781784
	1,71	0,96426	0,32574	4,691768
	1,72	0,94199	0,33801	4,601496
	1,73	0,91842	0,35158	4,511141
	1,74	0,89325	0,36675	4,420866
	1,75	0,86609	0,38391	4,330827
	1,76	0,83633	0,40367	4,241168
	1,77	0,80295	0,42705	4,152023
	1,78	0,76395	0,45605	4,06352
	1,79	0,7142	0,4958	3,975773
	1,8	0,6	0,6	3,888889
0,5	1,5	1,5	0	végtelen
	1,51	1,48993	6,7E-05	14900,66
	1,52	1,47973	0,00027	3700,658
	1,53	1,46939	0,00061	1633,987
	1,54	1,4589	0,0011	913,1494
	1,55	1,44827	0,00173	580,6452
	1,56	1,4375	0,0025	400,641

<i>Red</i>	λ_1	λ_2	λ_3	A sajátértékek reciprokösszege
	1,57	1,42657	0,00343	292,4737
	1,58	1,41548	0,00452	222,5079
	1,59	1,40423	0,00577	174,703
	1,6	1,39282	0,00718	140,625
	1,61	1,38124	0,00876	115,4972
	1,62	1,36949	0,01051	96,45062
	1,63	1,35755	0,01245	81,67859
	1,64	1,34543	0,01457	69,99751
	1,65	1,33312	0,01688	60,60606
	1,66	1,32062	0,01938	52,94616
	1,67	1,3079	0,0221	46,61956
	1,68	1,29498	0,02502	41,33598
	1,69	1,28184	0,02816	36,8798
	1,7	1,26847	0,03153	33,08824
	1,71	1,25486	0,03514	29,8365
	1,72	1,241	0,039	27,02768
	1,73	1,22688	0,04312	24,5856
	1,74	1,21249	0,04751	22,44971
	1,75	1,19782	0,05218	20,57143
	1,76	1,18285	0,05715	18,91138
	1,77	1,16756	0,06244	17,4374
	1,78	1,15194	0,06806	16,12302
	1,79	1,13597	0,07403	14,94629
	1,8	1,11962	0,08038	13,88889
	1,81	1,10286	0,08714	12,93542
	1,82	1,08568	0,09432	12,07289
	1,83	1,06804	0,10196	11,29025
	1,84	1,04989	0,11011	10,57808
	1,85	1,03121	0,11879	9,928296
	1,86	1,01193	0,12807	9,333931
	1,87	0,992	0,138	8,78896
	1,88	0,97134	0,14866	8,288148
	1,89	0,94987	0,16013	7,826931
	1,9	0,92749	0,17251	7,401316
	1,91	0,90406	0,18594	7,007796
	1,92	0,87941	0,20059	6,643282
	1,93	0,85332	0,21668	6,305047
	1,94	0,82547	0,23453	5,990671
	1,95	0,79542	0,25458	5,698006
	1,96	0,76249	0,27751	5,425138
	1,97	0,72554	0,30446	5,170358

<i>Red</i>	λ_1	λ_2	λ_3	A sajátértékek reciprokösszege
	1,98	0,68234	0,33766	4,932134
	1,99	0,62717	0,38283	4,709093
	2	0,5	0,5	4,5
0,6	2,1	0,86533	0,03467	30,47619
	2,11	0,83987	0,05013	21,61408
	2,12	0,8131	0,0669	16,64817
	2,13	0,78475	0,08525	13,47397
	2,14	0,7545	0,1055	11,27131
	2,15	0,72186	0,12814	9,654305
	2,16	0,68608	0,15392	8,417508
	2,17	0,64592	0,18408	7,441486
	2,18	0,59894	0,22106	6,652069
	2,19	0,53888	0,27112	6,000769
	2,2	0,4	0,4	5,454545
0,7	2,34853	0,65147	0	végtesen
	2,35853	0,61365	0,02782	38,00166
	2,36853	0,57137	0,0601	18,81041
	2,37853	0,52227	0,09921	12,41525
	2,38853	0,46063	0,15084	9,219072
	2,39853	0,35632	0,24515	7,302469
	2,4	0,3	0,3	7,083333
0,8	2,582	0,41626	0,00174	577,7354
	2,592	0,34239	0,06561	18,5482
	2,6	0,2	0,2	10,38462
0,9	2,797	0,19146	0,01154	92,25431
	2,798	0,17446	0,02754	42,40569
	2,799	0,15245	0,04855	27,51576
	2,8	0,1	0,1	20,35714
1	3	0	0	végtesen

VI. 7. A *Red*-mutató empirikus eloszlásfüggvényének meghatározásához használt algoritmus

```

program sajátert;
uses crt;
const m=3;
var
    red,l1,l2,l3,tmp:real;
    redf:text;
    pontos,pontos2,i,j,k:longint;
begin;
    clrscr;
    assign(redf,'red.txt');
    rewrite(redf);
    write('pontossag=');
    readln(pontos2);    pontos:=1;
    for i:=1 to pontos2 do pontos:=10*pontos;
    for i:=m*pontos downto pontos do
        begin
            for j:=m*pontos-i downto 0 do
                begin
                    if i>=j then
                        begin
                            k:=m*pontos-i-j;
                            if ((k>=0) and (k<=j))
                                then
                                    begin
                                        if ((i/pontos)*(j/pontos)*(k/pontos)<=1) then
                                            begin
                                                l1:=i/pontos;
                                                l2:=j/pontos;
                                                l3:=k/pontos;
                                                red:=sqrt((l1*l1+l2*l2+l3*l3-m)/(m*(m-1)));
                                                writeln(redf,red:10:pontos2);
                                            end;
                                        end;
                                    end;
                                end;
                            end;
                        end;
                    end;
                end;
            end;
        end;
    close(redf);
    while not keypressed do;
end.

```

VI. 8. Publikációim, konferencia-előadásaim

Lektorált folyóiratban, szaklapban megjelent tudományos cikkek

- [1] SZONDI I. – KOVÁCS P. – IDOVICA B. [2002]: A családok helyzete Szeged város lakótelepein, *ACTA JURIDICA ET POLITICA*, Tomus LXII. Fasc. 18., Szeged, 30 oldal.
- [2] FÜLÖP V. – SZONDI I. – KOVÁCS P. [2003]: Lakáscélú állami támogatások és egyéb, a lakáshoz jutást segítő ellátási formák, *PUBLICATIONES DOKTORANDUM JURIDICORUM*, Tomus II. Fasc. 6. Szeged, 30 oldal.
- [3] KOVÁCS P. – SZONDI I. [2003]: Úton az információs társadalom felé, *ACTA JURIDICA ET POLITICA*, Tomus LXIII. Fasc. 13., Szeged, 20 oldal.
- [4] KOVÁCS P. – PETRES T. – TÓTH L. [2004]: Adatállományok redundanciájának mérése, *Statisztikai Szemle*, KSH, Budapest, 82. évfolyam 6.-7. szám, 595.-604. oldal.
- [5] GYÉMÁNT R. – PETRES T. – KOVÁCS P. [2005]: A Szandzsák, az egyedülálló vallási régió, *Területi statisztika*, KSH, 8. (45.) évfolyam 3. szám 278.-287. oldal.
- [6] KOVÁCS P. – PETRES T. – TÓTH L. [2005]: A new measure of multicollinearity in linear regression models, *International Statistical Review (ISR)*, Volume 73 Number 3, International Statistical Institute, Voorburg, The Netherlands, 405.-412. oldal.
- [7] KOVÁCS P.: A multikollinearitás vizsgálata lineáris regressziós modellekben, [2008]: *Statisztikai Szemle*, KSH, Budapest, 86. évfolyam 1 szám, 38.-67. oldal.
- [8] LUKOVICS M. – KOVÁCS P. [2008]: Eljárás a területi versenyképesség mérésére, *Területi statisztika*, KSH, 11. (48.) évfolyam, 20 oldal, (megjelenés alatt).

Oktatási segédanyagok, jegyzetek, könyvfejezetek

- [1] KOVÁCS P. – PETRES T. [2004]: Statisztika Feladatgyűjtemény (közgazdász hallgatók számára), SZTE GTK, 120 oldal.
- [2] KOVÁCS P. – PETRES T. [2004]: Statisztika Feladatgyűjtemény, Dunaújvárosi Főiskola, Dunaújváros, 284 oldal.
- [3] KOVÁCS P. – PETRES T. [2004]: Statisztika Képletgyűjtemény, Dunaújvárosi Főiskola, Dunaújváros, 50 oldal.

- [4] KATONA T. – KOVÁCS P. – PETRES T. [2006]: Általános statisztika, tankönyv, JATEPress, Szeged, 225 oldal.
- [5] KOVÁCS P. – PETRES T. [2006]: Általános Statisztika Feladatgyűjtemény (joghallgatók részére), JATEPress, Szeged, 2005, 132 oldal.
- [6] KOVÁCS P. – PETRES T. – TÓTH L. [2006]: Válogatott fejezetek Statisztikából, Többváltozós statisztikai módszerek, JATEPress, Szeged, 167 oldal.
- [7] KOVÁCS P. [2006]: Általános statisztikai alapismeretek, EU távoktatás elektronikus jegyzet, 80 oldal.
- [8] KOVÁCS P. — PETRES T. [2007]: Tanulási útmutató a főiskolák és egyetemek Általános statisztika című tantárgyához, Dunaújvárosi Főiskola, Dunaújváros, 190 oldal.
- [9] KOVÁCS P. — PETRES T. [2008]: Szoftverek alkalmazása az üzleti életben: statisztikai programok, Dunaújvárosi Főiskola, Dunaújváros, 83 oldal.
- [10] KOVÁCS P. — PETRES T. [2008]: Tanulási útmutató a Szoftverek alkalmazása az üzleti életben: statisztikai programok tantárgyához, Dunaújvárosi Főiskola, Dunaújváros, 78 oldal.
- [11] VILMÁNYI M. – KOVÁCS P. [2008]: Egyetemi-ipar együttműködések teljesítménye és lehetséges vizsgálati módszere, *Kérdőjelek a régiók gazdasági fejlődésében* (szerk. LENGYEL I. – LUKOVICS M.), JATEPress, Szeged, 25 oldal, (megjelenés alatt).
- [12] KOVÁCS P. [2008]: Az információs társadalom szerinti területi egyenlőtlenségek mérése, *Kérdőjelek a régiók gazdasági fejlődésében* (szerk. LENGYEL I. – LUKOVICS M.), JATEPress, Szeged, 11 oldal, (megjelenés alatt).

Idegen-nyelvű konferencia-kiadványok

- [1] KOVÁCS P. – SZONDI I. [2006]: E-europe- E-Hungary, Ungarn auf der Schwelle in die EU, A Pólay Elemér Alapítvány Könyvtára, sorozatszerkesztő: Balogh Elemér, Szeged, 29.-48. oldal.
- [2] KOVÁCS P. – PETRES T. [2006]: A New Measure of Multicollinearity in Linear Regression Models, International Conference Applied Statistics (2006, Ribno, Slovenia), Program and Abstract, Statistical Society of Slovenia, Ljubljana.

- [3] KOVACS P. – LUKOVICS M. [2006]: Classifying Hungarian sub-regions by their competitiveness, Globalization Impact on Regional and Urban Statistics, 25th SCORUS Conference on Regional and Urban Statistics Research, Wroclaw, Poland, <http://www.scorus2006.ae.wroc.pl>, 12 oldal.
- [4] KOVÁCS P. – PETRES T. [2007]: Measure of Multicollinearity with a New, Original Indicator (PETRES' *Red*) in Linear Regression Models, International Conference on Mathematics & Statistics, Athens Institute for Education Research, Athens, (KIADÁS ALATT)

Magyarnyelvű konferencia-kiadványok

- [1] KOVÁCS P. – LAMPERTNÉ A. I. – PETRES T. [2005]: A multikollinearitás mérése lineáris regressziós modellekben, A Dunaújvárosi Főiskola Közleményei XXVI/II., Dunaújváros, 355.-365. oldal.
- [2] KOVÁCS P. [2005]: Statisztikai mintákat generáló algoritmusok, A Dunaújvárosi Főiskola Közleményei XXVI/II., Dunaújváros, 347.-354. oldal.
- [3] KOVÁCS P. [2005]: Az informatika alkalmazása a közgazdasági képzésben, Informatika a felsőoktatásban konferencia 2005 CD-melléklete, Debreceni Egyetem Informatikai Kar, Debrecen, 6 oldal.
- [4] KOVÁCS P. [2005]: Az informatika oktatása és lehetőségei a jogászképzésben, Informatika a felsőoktatásban konferencia 2005 CD-melléklete, Debreceni Egyetem Informatikai Kar, Debrecen, 6 oldal.
- [5] KOVÁCS P. – PETRES T. [2006]: A PETRES-féle *Red*-mutató eloszlásának vizsgálata, A Dunaújvárosi Főiskola Közleményei XXVII/II., Dunaújváros, 2006, 521.-530. oldal.
- [6] KOVÁCS P. – PETRES T. – LUKOVICS M. [2006]: A PETRES-féle *Red*-mutató alkalmazásának lehetőségei, A Dunaújvárosi Főiskola Közleményei XXVIII., Dunaújváros, 304.-316. oldal.

Egyéb publikációk

- [1] KOVÁCS P. [2006]: A statisztika oktatásának és oktatásmódszertanának reformálása a saját gyakorlatomban, „A felsőoktatás szerkezeti és tartalmi fejlesztése” tárgyú

Humánerőforrás-fejlesztési Operatív Program (HEFOP 3.3.) „Partnerközpontú önértékelési modell megalkotása és továbbképzések a felsőoktatási intézmények humánerőforrásainak fejlesztéséért” tanulmányainak CD gyűjteménye, Dunaújváros, 52 oldal.

Idegennyelvű konferencia-előadások

- [1] KOVÁCS P. – SZONDI I.: eEurope, eHungary, Társadalmi és gazdasági kihívások az Eu-csatlakozás küszöbén, angol nyelvű, nemzetközi konferencia előadás, Szeged, 2004. június 12.
- [2] KOVÁCS P. – LUKOVICS M.: Classifying Hungarian sub-regions by their competitiveness, Globalization Impact on Regional and Urban Statistics, 25th SCORUS Conference on Regional and Urban Statistics Research, Wroclaw, Poland, 2006. augusztus 30.-szeptember 1.
- [3] KOVÁCS P. – PETRES T.: A new measure of multicollinearity in linear regression models, Applied Statistics 2006 International Conference, Ribno (Bled), Slovenia, 2006. szeptember 17.-20.
- [4] KOVÁCS P. – PETRES T.: Measure of Multicollinearity with a New, Original Indicator (PETRES' *Red*) in Linear Regression Models, International Conference on Mathematics & Statistics, ATINER, 2007. június 11.

Magyarnyelvű konferencia-előadások

- [1] KOVÁCS P. – LAMPERTNÉ A. I. – PETRES T.: A multikollinearitás mérése lineáris regressziós modellekben, DUF Közgazdasági szimpózium, Dunaújváros, 2004. november
- [2] KOVÁCS P.: Statisztikai mintákat generáló algoritmusok, DUF Informatikai szimpózium, Dunaújváros, 2004. november
- [3] KOVÁCS P.: Az informatika alkalmazása a közgazdasági képzésben, Informatika a felsőoktatásban konferencia 2005, B szekció, Debreceni Egyetem Informatikai Kar, Debrecen, 2005. augusztus 24.-26.

- [4] KOVÁCS P.: Az informatika oktatása és lehetőségei a jogászképzésben, Informatika a felsőoktatásban konferencia 2005, F szekció, Debreceni Egyetem Informatikai Kar, Debrecen, 2005. augusztus 24.-26.
- [5] KOVÁCS P. – PETRES T.: A PETRES-féle *Red*-mutató eloszlásának vizsgálata, Magyar Tudomány Hete a Dunaújvárosi Főiskolán Közgazdasági és menedzsment Konferencia, Dunaújváros, 2005. november 22.
- [6] KOVÁCS P. – PETRES T. – LUKOVICS M.: A PETRES-féle *Red*-mutató alkalmazásának lehetőségei, Magyar Tudomány Hete a Dunaújvárosi Főiskolán Közgazdasági és menedzsment Konferencia, Dunaújváros, 2006. november 16.
- [7] KOVÁCS P. – PETRES T.: A PETRES-féle *Red*-mutató ismertetése, VI. Természet-, Műszaki- és Gazdaságtudományok Alkalmazása Nemzetközi Konferencia, Szombathely, 2007. május 18.
- [8] VILMÁNYI M. – KOVÁCS P.: Egyetemi-ipar együttműködések teljesítménye és lehetséges vizsgálati módszere, "*Kérdőjelek a régiók gazdasági fejlődésében*" Konferencia, Szeged, 2007. november 12.
- [9] KOVÁCS P. : Az információs társadalom szerinti területi egyenlőtlenségek mérése, "*Kérdőjelek a régiók gazdasági fejlődésében*" Konferencia, Szeged, 2007. november 13.