A multikollinearitás vizsgálata lineáris regressziós modellekben

Kovács Péter,
a Szegedi Tudományegyetem egyetemi adjunktusa
E-mail: pepe@eco.u-szeged.hu

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. A multikollinearitás felismerésének, mérésének és e jelenség káros következményeinek csökkentésének számos módja ismert. Azonban, kérdéses, hogy mit jeleznek a multikollinearitás mérőszámai, illetve az, hogy a multikollinearitás jelenlétének káros következményei hogyan, illetve milyen lehetséges mellékhatásokkal csökkenthetők. A tanulmányban összefoglalom, illetve véleményezem a multikollinearitás detektálásának, illetve mérésének közel húsz módját, valamint a multikollinearitás negatív hatásainak csökkentésére kidolgozott nyolc módszert.

TÁRGYSZÓ:
Főkomponenselemzés.
Algoritmusok, programok, számítási módszerek.

Mai globalizálódó világunkban egyre inkább 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. Többváltozós empirikus elemzéseknél a statisztikai módszerek közül leggyakrabban a regressziós modellt alkalmazzák, melynek legismertebb típusa a standard lineáris regressziós modell. Ez mátrixalgebrai jelöléssel az

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad /1/$$

formában írható fel.

A modellben szereplő ismeretlen paraméterek *n* megfigyelésből álló minta alapján történő becslőfüggvénye a legkisebb négyzetek módszere szerint a következő.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \,. \qquad /2/$$

A $\hat{\boldsymbol{\beta}}$ funkcionális operátor olyan hipersíkot eredményez, amely a legjobban illeszkedik a megfigyelések *n*-dimenziós „pontfelhőjéhez".

A regressziószámítás gyakorlati alkalmazásakor ügyelnünk kell arra, hogy a standard lineáris regressziós modellt ne használjuk, ha valamelyik feltétele nem teljesül. Közgazdasági elemzéseknél ennek leggyakrabban három oka lehet:

1. autokorreláció: a hibatagok együttmozgása szignifikáns;
2. heteroszkedaszticitás: a hibatag szórásnégyzete nem állandó;
3. multikollinearitás: a magyarázóváltozók együttmozgása statisztikailag jelentős, azaz szignifikáns. Lineáris regressziós modellek esetén ez a jelenség a redundancia egy fajtájaként értelmezhető.

A standard lineáris regressziós modellben a becsült paraméterek varianciáit a

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad /3/$$

alapján tudjuk kiszámítani.

Mivel az előző két összefüggésnél a hibatagok σ^2 varianciája számunkra ismeretlen, ezért ennek a becsült paraméterek varianciáit a

$$Var\left(\hat{\boldsymbol{\beta}}\right) = s_e^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad /4/$$

képlettel becsülhetjük.

1. A multikollinearitás fogalma

A multikollinearitás fogalma a szakirodalomban látszólag egységes. Az egyes meghatározások 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 becsülni 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 lineárisan függetlenségét, vagy statisztikai értelemben vett függetlenségét. 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, hogy a magyarázóváltozók egymástól lineárisan függetlenek legyenek. Ezért, egyes forrásokban multikollinearitáson a tényezőváltozók lineáris függetlenségének hiányát értik. Ez gyakorlatilag azt jelenti, hogy valamelyik tényezőváltozó kifejezhető a többi tényezőváltozó nem triviális lineáris kombinációjaként. Ennek következtében az $\mathbf{X}'\mathbf{X}$ 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 $\mathbf{X}'\mathbf{X}$ mátrix ugyan invertálható,

de egyes becsült paraméterek varianciái nagymértékben növekednek a hibatagok szórásnégyzetéhez képest. A multikollinearitással foglalkozó szakirodalmak döntő többsége ezzel az esettel foglalkozik. Azonban, elöljáróban megjegyzem, hogy multikollinearitáson 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 multikollinearitáson ért.

A multikollinearitás szignifikáns volta egy adottság és nem az alkalmazott modell hibája. 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 nemcsak 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 információveszteséggel, az eredmények nem megfelelő értelmezhetőségével) járhat.

2. A multikollinearitás következményei

A multikolinearitással foglalkozó tanulmányok, tankönyvfejezetek szinte kivétel nélkül megemlítik a multikollinearitás negatív következményeit. Mint a későbbiekben rávilágítok, a sokszor emlegetett negatív következmények nem mindig, csak bizonyos esetekben (near multicollinearity) jelentkeznek.

A multikollinearitás gyakran emlegetett következményei a következők.

– A becslés és az előrejelzés torzítatlan marad.

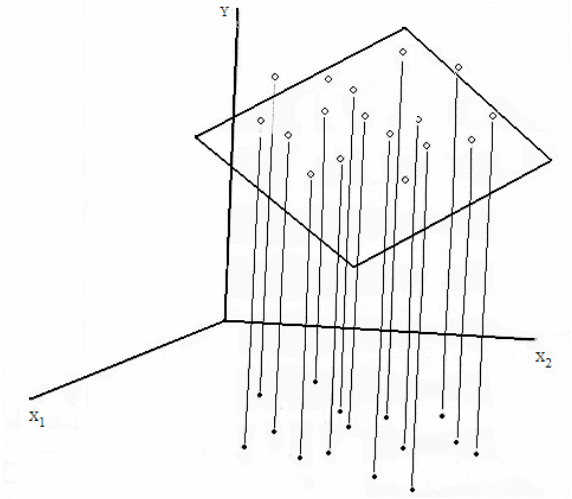
– A regressziós együtthatók /3/ képlettel adott standard hibái nőnek.

– Az egyes magyarázóváltozók szeparált hatásának vizsgálata értelmetlenné válik. Ugyanis, a becsült paraméterek szórásnégyzete /4/ szerint nagy mértékben növekszik, melynek következtében a parciális

F-próbák (vagy *t*-próbák) értelmüket vesztik, hiszen ezen próbafüggvényeknek az értékei nagyon alacsonyak lesznek.

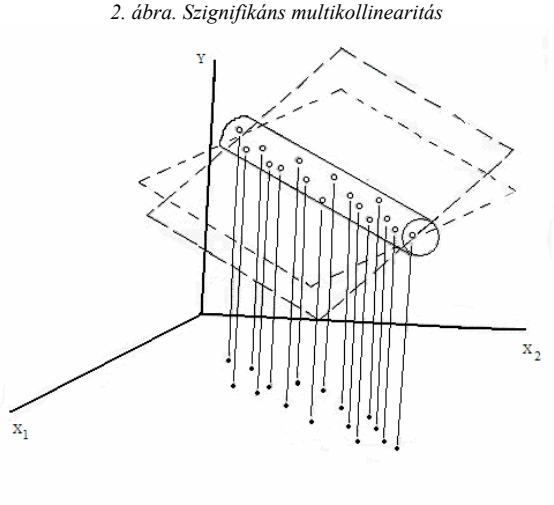
– A regressziós paraméterek /2/ képlettel adott becslése bizonytalanná, instabillá válik. Ezt szemlélteti az 1. és a 2. ábra.

1. ábra. A magyarázóváltozók nem szignifikáns együttmozgása



Forrás: *Tričković* [1976].

2. ábra. Szignifikáns multikollinearitás



Forrás: *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 vizsgálata értelmetlenné válik.

A 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 adatállomány „pontfelhője" minden dimenzióban szóródik, és így a ráillesztett sík stabil. Míg a 2. ábra „pontfelhője" nem mindegyik dimenzióban szóródik, így a ráillesztett sík könnyen kibillen, azaz instabillá válik az illesztés.

A következmények között találjuk azt, hogy a becsült regressziós paraméterek varianciái növekednek, illetve értékük magas lesz. Ezzel az állítással kapcsolatosan két problémát lehet megfogalmazni. Egyrészt nem mindegyik variancia fog nőni, másrészt pedig, mit értünk az alatt, hogy ezeknek magas értékük lesz. Erre mutatott rá *Maddala* [2004]. Olyan ellenpéldát ad, 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 feltűntetni, hogy a varianciák növekedését ceteris paribus értjük. Ugyanis, ha megvizsgáljuk a /3/ és a /4/ összefüggést, akkor láthatjuk, hogy a becsült paraméterek varianciái két tényezőtől függnek. Egyrészt, a hibatagok varianciájától, másrészt a képletben szereplő $(\mathbf{X}'\mathbf{X})^{-1}$ mátrix diagonális elemeitől. A *Maddala* [2004] által adott ellenpéldában azért nem lesznek nagyok a becsült paraméterek varianciái, mert alacsony a hibatagok varianciáinak becsült értéke, azaz a reziduális szórásnégyzet. Ezért, gyakorlatilag a becsült paraméterek varianciáinak nem az abszolút nagyságát kell ceteris paribus nézni, hanem azt, hogy mekkora ezeknek

$$\frac{Var\left(\hat{\beta}_i\right)}{\sigma^2} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{ii} \qquad /5/$$

inflálódása a hibatagok varianciájához képest.

3. 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 de-

tektálása sokszor nagyon nehéz feladat, másrészt a mutatók többségének értelmezése meglehetősen szubjektív.

Egy mutatóval szemben támasztott minimális elvárások a következők.

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

A továbbiakban e szempontok szerint is elemzem a multikollinearitás néhány mutatóját. A multikollinearitás felismerésének egy egyszerű módszere az, hogy a tényezőváltozók korrelációs mátrixát vizsgálva, nagyobbnak tekintjük a multikollinearitás mértékét, ha a főátlón kívüli elemek abszolút értékei messzebb esnek nullától. A módszerrel több probléma van. Az 1. táblázat korrelációs mátrixában a korrelációs együtthatók nullától való különbözőségeiről nem tudjuk megállapítani, hogy azok jelentősek-e, vagy sem. A módszer nem határozza meg egyértelműen azt, hogy hány korrelációs együttható szignifikáns eltérése jelez multikollinearitást.

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 (*Herman et al.* [1994]). Ez a módszer meglehetősen szubjektíven értelmezi a közelség fogalmát, abból a szempontból, hogy a közelség mindenkinek mást és mást jelent, azaz nincs olyan egyértelmű küszöbszám, amely alapján azt mondhatjuk, hogy egy korrelációs együttható közelinek tekinthető a többszörös korrelációs együttható értékéhez.

Mason és *Perreault* [1991] 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 $r^2_{y.x_1,x_2,\ldots,x_m}$ többszörös determinációs[2] együttható tényleges értékétől való eltérése a multikollinearitás jelenlétére utal.

[1] Ez az elvárás általánosságban nem követelmény, csak hasznos tulajdonság.

[2] Az $r^2_{y.x_1,x_2,\ldots,x_m}$ alsó indexében a pont után a tényezőváltozók felsorolása ezek lineáris kombinációja utal.

$$M_1 = \sum_{i=1}^{m} r_{yx_i}^2 - r_{y.x_1,x_2,\ldots,x_m}^2 .$$

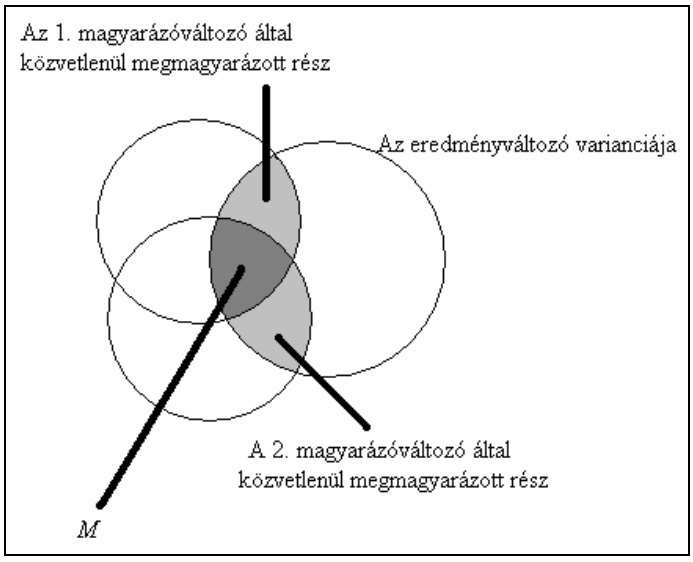
A fő kérdés az, hogy mekkora eltérés jelez erős multikollinearitást (*Herman et al.* [1994]).

Egy másik szintetikus mutató az

$$M = r_{y.x_1,x_2,\ldots,x_m}^2 - \sum_{j=1}^{m} \left(r_{y.x_1,x_2,\ldots,x_m}^2 - r_{y.x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}^2 \right), \qquad /6/$$

aminek a többszörös determinációs együtthatóhoz közeli értéke jelentős multikollinearitást jelez (*Förster–Egermayer* [1966]). A „közelség" értelmezése szubjektív, ráadásul az *M* értéke negatív is lehet. A /6/ képlet magyarázatának két főbb megközelítése ismert. Az egyik szerint, a /6/ 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. (Lásd a 3. ábrát.)

3. ábra. Az M-mutató illusztrációja



Forrás: Saját szerkesztés.

Két magyarázóváltozó esetén a tárgyalt összefüggés szerkezete gyakorlatilag a szitaformula analógiája, az együttesen megmagyarázott részre, mint halmazra alkalmazva. Márpedig a szitaformula végeredménye nem lehet negatív előjelű. Tehát a /6/ képletnek az e fajta interpretációja nem tökéletes, ugyanis a magyarázóváltozók közvetlen hatásainak mértéke nem egyezik meg a képletben szereplő értékkel. A /6/ 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* (1971) ezeket a tényezőket, azaz a /6/ képlet összeadandó részeit az adott változónak a többszörös determinációs együtthatóhoz tartozó növekményi hozzájárulásának nevezte. Pontosan ezek a növekmények jelentik a /6/ 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* (*Farrar et al.* [1970]) -féle vizsgálat szerint a korrelációs mátrix determinánsa megközelítőleg χ^2-(khi-négyzet) eloszlásúvá transzformálásával a következő próbafüggvényt kapjuk.

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

A hipotézisvizsgálat nullhipotézise a magyarázóváltozók lineárisan függetlensége, vagyis az, hogy a determináns abszolút értéke egy. Ennek a statisztikai próbának a szabadságfoka $\frac{m(m-1)}{2}$ (*Hulyák* [1969]). Meg kell jegyeznem, hogy a nullhipotézis elfogadása nem jelenti automatikusan azt, hogy nem lép fel a multikollinearitás a modellben.

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 (*Hulyák* [1969]), hogy a diagonális elemek megközelítőleg $n-m$ és $m-1$ szabadságfokú F-eloszlásúvá transzformálhatók, ami a következő próbafüggvényt eredményezi.

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

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 a 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, természetesen normális eloszlás feltételezése mellett.

Frisch sugárkévetérképek 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 változót $m+1$ darab egyenlettel írjuk fel a többi változó segítségével. Ezek után iteráljuk az eljárást, vesszük az összes lehetséges m-dimenziós modellt stb. Az iterációs eljárást két dimenzióig ismételjük. A kapott parciális regressziós együtthatókat az összehasonlíthatóság kedvéért normálnunk kell. A sugárkévetérképekben 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 sugárkévetérképeken ábrázolandó koordináták. Egy sugárkéve nem más, mint egy-egy változópár közötti, összes kapott – adott dimenziójú – együtthatók ábrája. A kévék zártságából, meredekségéből és a sugarak hosszából kimutatható a multikollinearitás, illetve megállapítható, mely magyarázóváltozók lesznek hasznosak, károsak, illetve feleslegesek az eredményváltozó magyarázatának szempontjából. A kéve zártsága 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 (Variance Inflation Factor – Varianciainflációs tényező) 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.

$$VIF_j = \frac{1}{1 - r^2_{x_j.x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}} . \qquad /7/$$

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

$$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{n\sigma_j^2}} \qquad /8/$$

szerint standardizált magyarázóváltozók esetén $\left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj} = VIF_j$.[3] A VIF_j-mutató megmutatja a $\hat{\beta}_j$ becsült regressziós együttható varianciája inflálódásának mértékét a hibatagok 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 toleranciamutatónak nevezzük. 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 Az 1. táblázatban látható (*Feng-Jenq* [2006]).

Az 1. 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 ápo-

[3] Ugyanis, a magyarázóváltozók korrelációs mátrixa alapján felírható a $VIF_j = \mathbf{R}^{-1}_{jj}$ összefüggés. Ekkor – a kizárólag az $\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{n\sigma_j^2}}$ szerint standardizált változókra érvényes – $\mathbf{X}'\mathbf{X} = \mathbf{R}$ egyenlet figyelembevételével az $(\mathbf{X}'\mathbf{X})^{-1}_{jj} = VIF_j$ összefüggést kapjuk.

lási 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.

1. táblázat

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].

A multikollinearitás mérőszámának egy családját alkotják a tényezőváltozók korrelációs mátrixának *sajátértékeire* épülő mutatók. 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 függnek.

Míg a VIF_j értékének meghatározása általában standardizált változókkal történik, addig a magyarázóváltozók egészére vonatkozó

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

gamma-mutató[4] értékének kiszámítása a magyarázóváltozók normált értékeivel történik. Ha a tényezőváltozók lineárisan függetlenek, akkor a mutató értéke eggyel egyenlő. Azonban a szignifikáns multikollinearitásnak nincs egyértelmű küszöbérté-

[4] Ezt a mutatót, illetve a négyzetét a különböző szakirodalmak más és más szerzők nevéhez kötik. Például *Wichern* és *Churchill, Casella, Belsley*. A mutató négyzete a kondiciószám, melynek értékei azt jelzik, 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.

ke, így értelmezése nem objektív. Egyes szerzők szerint e mutató 30 feletti értéke jelez erős multikollinearitást (*Kovács–Petres–Tóth* [2004]).

Fellman *L*-mutatójának

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

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

Mahayan és *Lawles* a multikollinearitás mérésére egy „másik" M_1-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 egyszerre két mutatót is javasolt. Az egyik az *mci* multikollinearitás-index, a másik pedig a *pmci* tervező multikollinearitás-index (*Fellman* [1981]).

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

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

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

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* [1981] rámutatott arra, hogy ha egy olyan speciális korrelációs mátrixot tekintünk, amiben a tényezőváltozók korrelációs mátrixának minden főátlón kívüli eleme α, akkor a két indexre szigorúbb alsó határt adhatunk.[5] Ekkor

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

[5] Ekkor a korrelációs mátrix sajátértékei: $\lambda_1 = \lambda_2 = \ldots = \lambda_{m-1} = 1 - a$; $\lambda_m = 1 + (m-1)a$.

Például, három magyarázóváltozó esetén mindkét index értéke kettőnél nagyobb lesz. Márpedig, például ha $\alpha = 0,9$; 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.

A *Vinod*, *Wichern* és *Churchill* által adott ISRM- (Index of Stability of Relative Magnitudes – Relatív terjedelem stabilitásának indexe) index értéke 0 és $m(m-1)$ közé esik (*Fellman* [1981]).

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

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]).

Mivel fogyasztáselemzések során a multikollinearitás szinte kivétel nélkül mindig jelen van, ezért például az 1 főre jutó évi marhahúsfogyasztást (**y**, kg/fő), mint eredményváltozót vizsgálva 1990 és 2004 között a következő tényezőváltozók[6] függvényében:

x_1 – egy főre jutó reáljövedelem indexe (2004=100,00%);
x_2 – sertéshús, comb, csont és csülök nélkül (Ft/kg);
x_3 – marhahús, rostélyos, csontos (Ft/kg);
x_4 – tojás (Ft/darab);
x_5 – pasztőrözött tej (Ft/liter);

[6] 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 árat 2004-es árszínvonalon adunk meg.

x_6 – sertészsír (Ft/kg);
x_7 – napraforgó-étolaj (Ft/liter);
x_8 – kenyér, fehér (Ft/kg);
x_9 – normál kristálycukor (Ft/kg);
x_{10} – narancs (Ft/kg);
x_{11} – sör, hazai világos (Ft/0,5 liter);
x_{12} – cigaretta, Sophianae, multifilteres, rövid, 20 db (Ft/csomag);
x_{13} – 1 főre évi jutó sertéshúsfogyasztás (kg/fő).

A sajátértékekre épülő mutatók értékeit a 2. táblázat tartalmazza.

2. táblázat

A sajátértékekre épülő mutatók értékei

Mutató	Érték
χ	47,756
L	221494,584
M_1	807,419
mci	1,675
pmci	2,216
ISRM	44,628

Forrás: Saját számítások.

Látható, hogy mind a γ egyhez képest, mind az *L,* az M_1, az *ISRM*-mérőszámok értékei – a maguk módján – a nullához képest távolinak mondhatók, így ezek erős multikollinearitást jeleznek. Azonban, 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. Az *mci* és a *pmci* értékei viszont nincsenek annyira közel az egyhez, mint amennyire várnánk. Ugyanis, az összes eddigi mérőszám nagyon erős multikollinearitást jelzett, ekkor nyilvánvalóan azt várnánk, hogy ennek a két indexnek az értéke egyhez közeli. Ezzel szemben, *pmci* = 2,216; tehát ezen indexek szerint ugyan létezhet multikollinearitás a modellben (az értékek eltávolodtak *m*-től), de ennek mértéke nem ítélhető meg objektíven.

Egy jogos kérdés az, hogyha ennyire szubjektív a sajátértékek reciprokaira épülő mutatók értelmezése, akkor miért próbálkoznak sokan ilyen típusú mutató megadásával?

Ugyanis, ha a /8/ szerint standardizált változókat vizsgálunk, akkor $\mathbf{X}'\mathbf{X} = \mathbf{R}$. A standardizált változókhoz tartozó becsült paraméterek variancia-kovariancia mátrixa felírható az

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

formában is a korrelációs mátrix spektrálfelbontása alapján, ahol $\boldsymbol{\Lambda}$ 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 tulajdonságainak[7] figyelembevételével a *j*-edik standardizált magyarázóváltozóhoz tartozó paraméter becslésének szórásnégyzete a következő.

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

Ebből a varianciák összegére a következő összefüggést[8] kapjuk:

$$\sum_{j=1}^{m} Var\left(\hat{\beta}_j\right) = \sigma^2\sum_{l=1}^{m}\frac{1}{\lambda_1}. \qquad /9/$$

Ezek szerint a varianciák értékét, azaz a varianciáknak a hibatagok szórásnégyzetéhez viszonyított inflálódásának mértékét végső soron – ceteris paribus – 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 multikollinearitás egyik legújabb mérőszáma a *Curto* és *Pinto* által 2007-ben publikált *DEF*- (Direct Effect Factor – Közvetlen hatás faktor) mutató (*Curto–Pinto* [2007]).

[7] 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.

[8] Az összefüggés egyszerűbben is megkapható a következő módon.

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

Amennyiben az

$$y_i = \hat{\beta}'_0 + \hat{\beta}'_1 x_{i,1} + \hat{\beta}'_2 x_{i,2} + ... + \hat{\beta}'_m x_{i,m} + e_i$$

regressziós modellt standardizált változókra írjuk fel, akkor ez az egyenlet

$$Y_i = \hat{\beta}_1 X_{i,1} + \hat{\beta}_2 X_{i,2} + ... + \hat{\beta}_m X_{i,m} + \hat{\beta}_e E_i = \hat{Y}_i + \hat{\beta}_e E_i$$

alakban írható fel, ahol a $\hat{\beta}_i$ a standardizált regressziós együtthatókat[9] jelenti. Ekkor[10]

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

A standardizált változók és a standardizált hibatag függetlenségének feltételezése mellett

$$Var(Y_i) = Var(\hat{Y}_i) + \hat{\beta}_e^2 .$$

Ekkor az eredményváltozó eggyel egyenlő varianciáját két részre bonthatjuk fel:

1. a tényezőváltozók által együttesen megmagyarázott $Var(\hat{Y}_i)$ hányad, amit a többszörös determinációs együtthatóval mérünk;

2. a tényezőváltozók által együttesen meg nem magyarázott hányad, ami gyakorlatilag $Var(Y_i) - r^2_{Y.X_1,X_2,...,X_m} = 1 - r^2_{Y.X_1,X_2,...,X_m}$.

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

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

[9] 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.

[10] Az összefüggés alapja az, hogy standardizált változók lineáris kombinációjának varianciája: $Var(y) = Var(\sum_{j=1}^{m} \beta_j x_j) = \sum_{j=1}^{m} \beta_j^2 + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \beta_i r_{x_i x_j} \beta_j$.

Ezek szerint, a tényezőváltozók által együttesen megmagyarázott $Var(\hat{Y}_i)$ varianciahányad, és így speciálisan a többszörös determinációs együttható is két részből tevődik össze:

1. a tényezőváltozók direkt hatásainak összege: $\sum_{i=1}^{m} \hat{\beta}_i^2$;

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 mutató egyhez közeli értéke erős multikollinearitást jelez.

Vizsgálataim szerint, a mutatóval kapcsolatban több hiányosság is felsorolható. A képlet számlálója negatív is lehet, így amellett, hogy százalékban sem fejezhető ki, gondot jelent az értelmezése is. Ennek szemléltetésére tekintsük a 3. táblázatban szereplő példát.

3. táblázat

Példa a DEF-mutató bírálatára

y_i	x_{i1}	x_{i2}
5	6	15
6	6	12
7	8	55
8	9	70
9	3	55
10	34	10
11	3	16
12	45	30

Forrás: Saját számítások.

A standardizált adatok korrelációs mátrixa a következő.

	Standardizált (y)	Standardizált (x_1)	Standardizált (x_2)
Standardizált (y)	1,000	0,602	–0,031
Standardizált (x_1)	0,602	1,000	–0,231
Standardizált (x_2)	–0,031	–0,231	1,000

Az illesztett modell főbb jellemzői a következők.

Modell	R	R^2	Korrigált R^2	A becslés standard hibája
1	0,612	0,374	0,166	0,85446711

Az illesztett modell együtthatói a következők.

	Nem standardizált együtthatók	Standardizált (x_1)	Standardizált együtthatók
Standardizált (x_1)	0,628	0,332	0,628
Standardizált (x_2)	0,114	0,332	0,114

Ekkor a *DEF*-mutatóban szereplő felbontás a következő lesz.

$$\sum_{i=1}^{m} \hat{\beta}_i^2 = 0,628^2 + 0,114^2 = 0,407 ,$$

$$\sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \hat{\beta}_i r_{ij} \hat{\beta}_j = 2 \cdot 0,628 \cdot (-0,231) \cdot 0,114 = -0,033 ,$$

$$\sum_{i=1}^{m} \hat{\beta}_i^2 + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \hat{\beta}_i r_{ij} \hat{\beta}_j = 0,374 ,$$

$$DEF = \frac{-0,033}{0,374} = -0,088 .$$

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

A redundancia egy új, lehetséges mérőszáma a Petres-féle *Red*-mutató (*Petres–Tóth* [2004]). A *Red*-mutató definiálásakor a tényezőváltozók **R** korrelációs mátrixának $\lambda_j \left(j = 1,2,...,m \right)$ sajátértékeit alkalmazzuk. A *Red*-mutató a következő gondolatmeneten alapszik. Ha a magyarázóváltozók forrásául szolgáló adatállomány a $\hat{\boldsymbol{\beta}}$ 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.

$$v_\lambda = \frac{\sigma_\lambda}{\bar{\lambda}} = \frac{\sqrt{\frac{\sum_{j=1}^{m}\left(\lambda_j - \bar{\lambda}\right)^2}{m}}}{\frac{\sum_{j=1}^{m}\lambda_j}{m}} = \frac{\sqrt{\frac{\sum_{j=1}^{m}\left(\lambda_j - \bar{\lambda}\right)^2}{m}}}{\frac{m}{m}} = \sqrt{\frac{\sum_{j=1}^{m}\left(\lambda_j - 1\right)^2}{m}} = \sigma_\lambda . \qquad /10/$$

Különböző adatállományok redundanciájának összevethetősége végett a mutatót normálni kell. Mivel a sajátértékek nemnegatívak, ezért a relatív szórásra vonatkozó $0 \le v_\lambda \le \sqrt{m-1}$ összefüggés[11] miatt, a normálás $\sqrt{m-1}$ értékével történik.

Az így kapott mutatót a redundancia mértékének számszerűsítésére használhatjuk, és segítségével a *Red*-mutatót a következők szerint határozzuk meg.

$$Red = \frac{v_\lambda}{\sqrt{m-1}} . \qquad /11/$$

A redundancia hiánya esetén a 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*-

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

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. A *Red*-mutató kiszámítható a korrelációs mátrix főátlón kívüli elemeinek négyzetes átlagaként is

$$Red = \sqrt{\frac{\sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} r_{ij}^2}{m(m-1)}}. \qquad /12/$$

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

A /12/ képlet szerint a *Red*-mutatóval mérni lehet a tényezőváltozók átlagos együttmozgásának mértékét. A mutató definíciójából és a /12/ képletből következik, 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.

4. A multikollinearitás negatív következményeinek csökkentése

Abban az esetben, ha a multikollinearitás jelenléte nem jelent problémát a vizsgálatok szempontjából – például előrejelzések esetén – akkor semmit sem kell tennünk. 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.

1. Egyrészt, a változók elhagyása mindig információvesztéssel jár. Előfordulhat, hogy bizonyos változók elhagyásával nagymértékű lesz ez a veszteség.

2. Másrészt, a vizsgálat szempontjából releváns változók elhagyása ugyan csökkentheti a multikollinearitás negatív következményeit, de

specifikációs torzítás lép fel az alkalmazott modellben. Ekkor a regressziós együtthatók becsült értékeinek értelmezése nem lesz valósághű.

3. Harmadrészt, honnan tudjuk, hogy melyik magyarázóváltozót kell elhagyni? Ugyan bizonyos mutatók a multikollinearitást magyarázóváltozókhoz próbálják kötni, de ahogyan már korábban hangsúlyoztam, ezért a jelenségért nem mindig egy változó okolható.

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, ún. *stepwise* eljárásokat, amelyek a változók szelektálásánál figyelembe veszik a *t*-értékeket, valamint – általában – a *toleranciamutató* változónkénti értékét is (*Hunyadi–Mundruczó–Vita* [1997]).

4. táblázat

A kéndioxid koncentrációjának becslése lineáris regressziós modellel

Változók	Becsült regressziós paraméterek	*t*-statisztika	VIF_j
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 szélsebesség	–3,222	–1,747	1,262
Évi átlagos csapadékmennyiség (hüvelyk)	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
A becslés stepwise algoritmus alkalmazásával			
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].

Ezzel kapcsolatosan egy – *Sokal* és *Rohlf* által adott – érdekes példát szeretnék megemlíteni (*Feng-Jenq* [2006]). Klimatológusok a 1970-es évek végén a levegő minőségének előrejelzésére egy elemzés során 41 amerikai nagyváros adatait vizsgálták. Az egyik részvizsgálat során a kéndioxid koncentrációját, mint eredményvál-

tozót, hét magyarázóváltozó függvényében elemezték. Ekkor a 4. táblázatban szereplő lineáris regressziós modellt kapták. A 4. táblázat adatai alapján megállapítható, hogy a 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 értéke nem kisebb a kritikus értékhez képest. Mi történik akkor, ha a regressziós modellt stepwise algoritmussal építjük fel? Ekkor a 4.táblázat második felének adatait kapjuk.

Az 4. 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 csak akkor tudja figyelembe venni a VIF_j-mutató értékét, ha ez valamelyik paraméter szórásnégyzetének olyan magas értékét jelzi, hogy a *t*-statisztika értéke alacsonyabb a kritikus értéknél.

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 akárhogyan 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*. Ez a módszer különösen fogyasztáselemzéseknél használatos, ahol is egyszerre keresztmetszeti és idősoros adatokat is felhasználnak. Például, *Tobin* kutatásaiban a jövedelmi elaszticitások becslését keresztmetszeti, míg az árrugalmassági 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 $\mathbf{X}^{+}_{nx(m+1)}$ mátrix az $\mathbf{X}_{(m+1)xn}$ mátrix általánosított inverze, ha teljesülnek a következő feltételek.

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

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

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

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

A Moore–Penrose inverz segítségével megoldható az /1/ egyenlet.

Ekkor

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

A módszer használata egy hagyományos LNM-becslést jelent (*Heinczinger* [1983]).

Gyakran alkalmazott eljárás a standardizált tényezőváltozók mesterséges, ortogonális, azaz egymástól lineárisan független változókba, 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 $\mathbf{Z} = \mathbf{XU}$ alakban, ahol $\mathbf{U}$ a korrelációs mátrix sajátértékeihez tartozó sajátvektorok mátrixa. Mivel $\mathbf{U}^{-1} = \mathbf{U}'$, így $\mathbf{X} = \mathbf{ZU}'$. Ezért az /1/ egyenlet felírható ilyen formában is.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{ZU}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,.$$

A *Hoerl* nevéhez fűződő *ridge-regresszió* (*Hoerl et al.* [1970]) gyakorlatilag egy torzító paraméter ($k>0$) alkalmazását jelenti: az $\mathbf{X}'\mathbf{X}$ mátrixhoz hozzáadjuk az egységmátrix k-szorosát. Ekkor a regressziós paraméterek – a /2/ egyenlet helyett – a következő formában becsülhetők.

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

A módszer legkényesebb kérdése a torzító paraméter megválasztása.[12] Amennyiben a torzító paraméter értékét minden határon 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 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. A k-érték megválasztásával az a probléma, hogy a stabilitás fogalmát nagyon szubjektíven értelmezték. A k-paraméter becslésére számos javaslat született. A 12. lábjegyzetben közölteknek megfelelően csak két, Hoerl által is alkalmazott technikát említek meg.

A becsült regressziós együtthatókat a torzító paraméter függvényében ábrázolva megkapjuk a *ridge-görbét*.[13] A görbe alapján megállapítható k megfelelő értéke. Ez

[12] Ennek a problémának igen gazdag szakirodalma van. Ennek részletes bemutatásától eltekintek.

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

az eljárás bár nagyon szemléletes, nem egzakt. Hoerl javaslata szerint *k* megfelelő értékét a következő 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.

1. 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.

2. Másrészt, a torzító paraméter alkalmazása miatt kérdéses a regressziós paraméterek helyes értelmezhetősége.

A felmerülő problémák miatt *Maddala* [2004] nem is javasolja általános esetben a multikollinearitás problémájának megoldására a ridge-regressziót. *Maddala* [2004] 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 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 x_1. E két változó alapján alkalmazhatjuk az

$$\hat{y}_i = \hat{\beta}_{0,1} + \hat{\beta}_{1,1} x_{i,1} + \varepsilon_{i,1}$$

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

A második iterációban a megmaradt tényezőváltozók közül válasszuk ki azt, amelyik az $\varepsilon_{i,1} = y_i - \hat{y}_i$ hibataggal a legerősebben korrelál. Legyen ez a változó x_2. Ekkor felírhatjuk az

$$\hat{\varepsilon}_{i,1} = \hat{\beta}_{0,2} + \hat{\beta}_{1,2} x_{i,2} + \varepsilon_{i,2}$$

lineáris regressziós modellt. Ekkor *t*-próbával tesztelnünk kell a kapott $\hat{\beta}_{1,2}$ regressziós együtthatót $\left(H_0 : \beta_{1,2} = 0\right)$. 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:

$$\hat{y}_i = \hat{\beta}_{0,1} + \hat{\beta}_{1,1} x_{i,1} + \hat{\beta}_{0,2} + \hat{\beta}_{1,2} x_{i,2} + \varepsilon_{i,2} = \hat{\beta}_{0,1} + \hat{\beta}_{0,2} + \hat{\beta}_{1,1} x_{i,1} + \hat{\beta}_{1,2} x_{i,2} + \varepsilon_{i,2},$$

majd következik a 3. iteráció.

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

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

lineáris regressziós modellt. Ekkor *t*-próbával tesztelnünk kell a kapott $\hat{\beta}_{1,k}$ regressziós együtthatót $\left(H_0 : \beta_{1,k} = 0\right)$. Ha a hipotézisvizsgálat során a nullhipotézist elfogadjuk, akkor az eljárás végeredménye a (*k*–1)-edik iterációban kapott regressziós modell lesz. Ellenkező esetben a *k*-adik iteráció során kapott egyenletet behelyettesítjük az előző iteráció végén kapott regressziós egyenletbe:

$$\hat{y}_i = \sum_{j=1}^{k} \hat{\beta}_{0,j} + \sum_{j=1}^{k} \hat{\beta}_{1,j} x_{i,j} + \varepsilon_{i,k},$$

majd, amennyiben maradt még tényezőváltozó, következik a (*k*+1)-edik iteráció, ellenkező esetben az eljárás végeredménye a *k*-adik 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.

5. A multikollinearitás vizsgálatának általánosítása

A multikollinearitás vizsgálatakor nem csak változópárok együttmozgása, hanem változócsoportok együttmozgása is problémát jelenthet, ennek azonban még nincs részletesen kidolgozott szakirodalma. Ezek a vizsgálatok későbbi kutatásaim feladatai lesznek. Erre megoldást jelenthet a kanonikus korrelációelemzés használata, ahol valamilyen korrelációs együtthatók négyzetes átlaga szerepel az *RI redundanciaindexben* is, de alkalmazási körét és tartalmát tekintve ez teljesen más, mint a *Red*-mutató.

A *redundanciaindexet* a *kanonikus korrelációelemzés* során alkalmazzuk. A kanonikus korrelációelemzés a lineáris korrelációvizsgálat általánosításának tekinthető. A kanonikus korrelációelemzés során adott az $x_1, x_2, ..., x_p$ és $y_1, y_2, ..., y_q$ ($q \le p$) két standardizált változócsoport. A feladat az, hogy mindkét változócsoportot helyettesítsük a változók különböző u_t, z_t ($t = 1,2,\ldots,q$) lineáris kombinációival úgy, hogy az u_t, z_t kanonikus változópáros közötti r_t korrelációs együttható maximális legyen.[14] Ezeket a korrelációkat kanonikus korrelációknak nevezzük. A kanonikus változók közötti korrelációs mátrix szerkezete a következő.

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

Ekkor az *y* változók szórásnégyzetét a z_t kanonikus változó átlagosan

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

[14] A kanonikus korrelációelemzés efféle megközelítése gyakorlatilag kettős faktoranalízisnek tekinthető, mivel két változóhalmaz 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 et al*. [2004]).

mértékben, míg az u_t kanonikus változó

$$RI_{yz_t} = r^2_{yz_t} r^2_{z_t u_t}$$

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 $r^2_{x_j.x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}$ lesz. Ezt minden lehetséges kombinációra elkészítve – felhasználva a VIF_j /7/ képletét – kiszámíthatjuk azt, hogy az egyes tényezőváltozók varianciái átlagosan

$$\frac{\sum_{j=1}^{m} r^2_{x_j.x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}}{m} = \frac{\sum_{j=1}^{m}\left(1-\frac{1}{VIF_j}\right)}{m} = 1-\frac{\sum_{j=1}^{m}\frac{1}{VIF_j}}{m} = 1-\frac{1}{\overline{VIF}_{j\,H}} \qquad /13/$$

mértékben magyarázhatók a többi tényezőváltozóval együttesen, ahol $\overline{VIF}_{jH}$ a VIF_j-mutatók harmonikus átlaga. A /13/ 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 módón 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 /13/ képlettel lehetséges.

*

A tanulmányban a multikollinearitás 17 mérőszáma, négy nem metrikus detektálási módja, továbbá negatív következményeinek csökkentésére használt 8 eljárás került bemutatásra. Összességében megállapítható, hogy a jelenleg használt mutatók általánosan nem, csak bizonyos esetekben jellemzik megfelelően a multikollinearitás mértékét. Az ismertetett eljárások pedig nem minden esetben csökkentik a multikollinearitás ká-

ros következményeinek mértékét. Pontosabban, ha csökkentik is, általában más negatív következményekkel kell szembenéznünk. A multikollinearitást nem csak változók, hanem változócsoportok is okozhatják. A változócsoportok hatása vizsgálatának egyik speciális esete a *Red*-mutató segítségével, míg egy másik speciális esete a VIF_j-mutatók harmonikus átlagának segítségével mérhető.

Irodalom

BOLLA M. – KRÁMLI A. [2005]: *Statisztikai következtetések elmélete*. Typotex Kiadó. Budapest.

MASON, CH. – PERREAULT, W. [1991]: Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*. 28. évf. 8. sz. 268–280. old.

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.

BELSLEY, D. A. – KUH, E. –WELSCH, R. E. [1980]: *Regression diagnostics, identifying influential data and sources of collinearity*. Wiley. New York.

FARRAR D E – GLAUBER R R. [1967]: Multicollineanty in regression analysis: the problem revisited. *Review of Economic and Statistics* 49. sz. 92–107. old.

FÖRSTER, E. – EGERMAYER, F. [1966]: *Korrelations- und Regressionsanalyse*. Verlag der Wirtschaft. Berlin.

FENG-JENQ L. [2006]: Solving multicollinearity in the process of fitting regression model using the nested estimate procedure. *Quality & Quantity* online.
http://springer.om.hu/content/j58255j05450u607/fulltext.pdf

FÜSTÖS L. ET AL.[2004]: *Alakfelismerés (Sokváltozós statisztikai módszerek)*. Új Mandátum Kiadó. Budapest.

MADDALA, GS. [2004]: *Bevezetés az ökonometriába*. Nemzeti Tankönyvkiadó. Budapest.

HAJDU O. [2003]: *Többváltozós statisztikai számítások*. Központi Statisztikai Hivatal. Budapest.

HEINCZINGER M. [1983]: A multikollinearitás felismerése, mérése és kiszűrése, *Statisztikai szemle*. 61. évf. 7. sz. 741–761. oldal.

HERMAN S. ET AL. [1994]: *Statisztika II*. JPTE. Pécs.

HOERL, A. E. – KENNARD, R. [1970]: Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12. évf. 1. sz. 55–67. old.

HULYÁK K. [1969]: A multikollinearitás feltárása és elemzése. *KSH Statisztikai és Matematikai módszerek Közgazdasági Alkalmazásának laboratóriumának 9. számú munkaanyaga*. Központi Statisztikai Hivatal. Budapest.

HUNYADI L. – MUNDRUCZÓ GY. – VITA L. [1997]: *Statisztika.* Aula Kiadó. Budapest.

HUNYADI L. [2001]: *Statisztikai következtetéselmélet közgazdászoknak*. Központi Statisztikai Hivatal. Budapest.

FELLMAN, J. [1981]: Leskinen's preliminary orthogonalizing ridge estimator and a new measure of multicollinearity. *Swedish School of Economics and Business Administration 75. számú munkaanyaga*. Swedish School of Economics and Business Administration. Helsinki.

CURTO, J. D. – PINTO, J. C. [2007]: New multicollinearity indicators in linear regression models. *International Statistical Review*. 75. évf. 1. sz. 114–121. old.

KORPÁS A.-NÉ (szerk.) [1997]: *Általános statisztika II*. Nemzeti Tankönyvkiadó. Budapest.

KOVÁCS P. – PETRES T. – TÓTH L. [2004]: Adatállományok redundanciájának mérése. *Statisztikai Szemle*. 82. évf. 6–7 sz. 595–604. old.

KOVÁCS P. – PETRES T. – TÓTH L. [2006]: *Válogatott fejezetek Statisztikából, többváltozós statisztikai módszerek.* JATEPress. Szeged.

PETRES T. – TÓTH L. [2004]: Piaci információk és a multikollinearitás. *A szociális identitás, az információ és a piac*. SZTE Gazdaságtudományi Kar Közleményei. JATEPress. Szeged.

PETRES T. – TÓTH L. [2006]: *Statisztika*. Központi Statisztikai Hivatal. Budapest.

RAMANATHAN, R. [2002]: *Bevezetés az ökonometriába, alkalmazásokkal*. Panem Kiadó. Budapest.

THEIL, H. [1971]: *Principles of econometrics*. Wiley. New York.

TRIČKOVIĆ, V. [1976]: *Teorijski modeli i metodi kvantitativnog istraživanja tržišta*. Institut za ekonomiku industrije. Beograd.

Summary

Huge database with lot of data very often means little information. In linear regression models it is because collinearity of variables. This collinearity is in fact a kind of redundancy of database. A lot of indicator, detection way and methods for decreasing of the deleterious effect of multicollinearity are known. But the means and the side effect of there are questionable. In the study near 20 indicators and 8 methods are examined.

It can be proved, that the currently used indicators of multicollinearity just in some special case indicate well the measure of multicollinearity. The mentioned methods not always decrease the deleterious effect of multicollinearity or conduce to other deleterious effect.

The cause of the multicollinearity could be not only a variable but group of variables. The effect of the group of variable could be examined with the Red-indicator in a special case, and in another special case with the harmonic means of VIF_j indicators.