Szigma, XX. (1987-88) 4. sz.

BRÓDY ANDRÁS

Az orvosolhatatlan kollinearitásról*

A probléma megfogalmazása: Jeölések és definíciók

Legyen az y függő változóra vonatkozó M számú mérés (megfigyelés) az $y = (y_1, \ldots, y_M)$ vektorba rendezve, a független (magyarázó) változók értékei pedig az $X = X_{MN} = \{x_{ik}\}$ mátrixba; x_{ik} a k-adik változó i-edik — az y_i „eredmény"-hez kapcsolódó — adatát jelenti. $N < M$, így több megfigyelésünk van, mint független változónk.

Ha feltételezhetjük, hogy a függő változó az $y = Xb$ lineáris alakban állítható elő, akkor a $b = (b_1, \ldots, b_N)$ vektort a legkisebb négyzetek módszerével

$$b = (X^T X)^{-1} X^T y \tag{1}$$

alakban számíthatjuk ki, ahol T a transzpozíció jele. Ismeretes ugyanis, hogy ez a legkisebb négyzetes eltérést, vagyis az $(y - Xb)^T (y - Xb)$ skaláris szorzat minimumát adja. Erről a skaláris szorzat b szerinti deriválásával győződhetünk meg. Lásd pl. MALINVAUD (1974) VI. fejezet.

Ismeretes az is, hogy $X^T X = (X^T X)^T$ szimmetrikus, úgynevezett Gram–féle mátrix; pozitív szemidefinit, sajátértékei tehát nemnegatív valós számok. Lásd pl. GANTMACHER (1965) IX. fej. 3.§.

Definíció: A független változók rendszere akkor *multikollineáris*, ha az adatok X mátrixából képzett $X^T X$ mátrix *szinguláris*.

A gyakorlatban egzakt multikollinearitással ritkán találkozunk. A szakirodalom azonban akkor is multikollinearitásról beszél, ha az nem egzakt, ha tehát az $X^T X$ mátrix egy vagy több sajátértéke nem zérus ugyan, de igen kicsi. Ez gazdasági adatok vizsgálatakor elég gyakran bekövetkezik. Az ilyen jellegű makroökonómiai megfigyelések létrejöttének elméleti és tapasztalati okait próbáltam összefoglalni „Kószálás Logaritmiában" című tanulmányomban (BRÓDY (1985)). Arra a következtetésre jutottam, hogy Logaritmiában, ahol az összes gazdasági folyamat szigorúan exponenciálisan növekszik, csupán egyetlen nagy sajátérték fogja a mátrixot jellemezni, s ha a többi sajátérték nem pontosan zérus, akkor ez csupán a mérési hibák következménye.

---

* Köszönöm ismeretlen lektoraimnak ismételt fáradozását. Több hibát, pontatlanságot, félreérthető fogalmazást sikerült véleményük figyelembevételével kiküszöbölni. Szerettem volna teljesebben is elfogadni fenntartásaikat — de ebben meggátolt az, hogy az ökonometriában szokásos és nem vitatott alapfeltevéseket nem tudom elfogadni. Talán képes leszek az eltéréseket egy másik tanulmányban részletesebben tárgyalni.

A mérési hibák hatásának és kezelésük módjának áttekintésével *Kőrösi Gábor* kollégám foglalkozik — rövidesen a Szigma számára is feldolgozza vizsgálódását — ezért e kapcsolódó kérdéskört itt a lehetőség szerint elkerülöm és csupán a multikollinearitással kapcsolatban felmerült javaslatok ismertetésére szorítkozom.

A korai orvoslási javaslatok

A jelenségkör felfedezője és elnevezője R. FRISCH (1934). Tőle származik kezelésének első javaslata is: válogassuk ki a magyarázó változóknak azt a kombinációját, amely a lehető legkevésbé kollineáris és a lehető legjobb eredményt adja. („Bunch-maps"). Ez a szellemes eljárás a kísérleti modellépítés vezérfonalává is vált s a mai számítógépes programok rendkívül egyszerűvé és gyorssá teszik alkalmazását. Lásd pl GAUDI (1986) ismertetését a BMDP csomag P2R lépésenkénti regressziószámításáról (i.m. 361 és kö oldal).

Bár így kétségtelenül elkerülhetjük a multikollinearitás fellépését és ráadásul kitűnő vizsgálati eszközt nyerünk, a megoldás nem veszélytelen. A kiválasztás sorrendje nem egyértelmű és csak további feltételezésekkel lehet azzá tenni, hiszen tudjuk, hogy a függőség vagy függetlenség csak vektorok rendszerére és nem egyes vektorokra vonatkozó sajátosság. Hozzá kell tenni, hogy éppen az eljárás mai „automatizáltsága" növeli a téves következtetéseknek azt a veszélyét, amire már *R. Frisch* célzott, amikor például a gólyák megfigyelt számának és a gyerekszületések adatainak szoros korrelációját kiszámította. Későbbi kutatók például a rádióhallgatás „okozta" mentális betegségekre mutattak rá: itt is világosan a Logaritmiában általános „együttnövekedés" délibábos hatása jelentkezik.

A javasolt eljárás tehát elkerüli a multikollinearitás fellépését, de nem ad tanácsot arra, hogy ha az egyszer már fellépett, akkor hogyan birkózzunk meg vele. Ökonometriai szakkönyvek, vagy például W. CORLETT (1987) további adatok bevonását tanácsolták. Természetesen további magyarázó változók bevonása — az X mátrix oszlopainak szaporítása — nem javíthat a helyzeten, mert ez a mátrix szingularitását nem tudja kiküszöbölni. Lehetséges azonban, hogy a megfigyelések számának növelése, azaz X sorainak szaporítása megoldást hoz. Ha azonban — mint Logaritmia esetében — tudjuk, hogy a multikollinearitás nem a rosszul lefolytatott megfigyeléseknek, a meg nem tervezett vagy befolyásolhatatlanul adott „kísérletnek" az eredménye, hanem éppen ez a dolgok szokásos rendje, akkor további mérésektől vagy megfigyelésektől sem remélhetünk javulást.

Az utolsó évtized felgyorsult és számítógép-segítette ökonometriai kutatásai másfajta javaslatokat hoztak. Mielőtt azonban ezeket megvizsgálnánk, számoljunk végig egy ilyen ismerten beteg modellt, hogy világosabban lássuk, milyen bajokat és problémákat okoz a multikollinearitás.

Számpélda

Példánkat KLEIN (1950) modelljéből merítjük, az adatokat THEIL (1971) is felhasználta a multikollinearitás bemutatására. A példa sem közgazdaságilag sem statisztikailag nem állja meg helyét, csupán illusztrációul szolgál.

1. táblázat
Az USA idősorai. 10^{12} *dollár, folyó áron*

Év	y Teljes fogyasztás	x_1 Profit-jövedelem	x_2 Bér-jövedelem
1921	41.9	12.4	28.2
1922	45.0	16.9	32.2
1923	49.2	18.4	37.0
1924	50.6	19.4	37.0
1925	52.6	20.1	38.6
1926	55.1	19.6	40.7
1927	56.2	19.8	41.5
1928	57.3	21.1	42.9
1929	57.8	21.7	45.3

A rutinszerűen elvégzett számítás az

$$y = -0.15x_1 + 1.43x_2$$

egyenlethez vezet. Ez szakmailag értelmetlen eredmény, mert a bérből nem lehet állandóan és jelentősen többet költeni ennek összegénél, a profit növekedése pedig aligha csökkentheti a fogyasztást.

Nem járunk jobban az elaszticitások számításával sem ha az $y = x_1^{b_1} x_2^{b_2}$ alakból indulunk ki, az adatok logaritmizálásával. Ez esetben a $b_1 = -0.49$ és $b_2 = 1.48$ értékeket nyerjük, ez hasonlóan és hasonló okból értelmetlen.

A számítást az ismert programcsomagok valamelyikével lefolytató kutatót persze az eredmény szakmai használhatatlanságán kívül az igen nagy szórások és az ebből következő igen széles konfidencia-intervallumok is figyelmeztetik a számítás megbízhatatlanságára. A korreláció ugyan megnyugtató eredményt ad, $r^2 = 0.999$, de ha a korrelációs mátrixot vizsgáljuk, akkor a b együtthatók közti szintén magas negatív korreláció már utal a számítás problematikus voltára.

A problémát nyilván az $X^T X$ mátrix sajátosságai okozzák. A mátrix értéke, egy tizedesre

$$\begin{pmatrix} 3250.6 & 6575.9 \\ 6575.9 & 13331.3 \end{pmatrix}.$$

E mátrix két sajátértéke 16576.4 és 5.5, az előbbi kereken háromezerszerese az utóbbinak. A mátrix az alábbi két diádból áll, ismét egy tizedesre kerekítve:

$$\begin{pmatrix} 3246.2 & 6578.1 \\ 6578.1 & 13330.2 \end{pmatrix} \quad \text{és} \quad \begin{pmatrix} 4.4 & -2.2 \\ -2.2 & 1.1 \end{pmatrix}.$$

Mivel az inverz a sajátértékek reciprokával szorzott diádok összege, abban a második diád válik dominálóvá.

Az inverz értéke 3 tizedesre ugyanis

$$\begin{pmatrix} 0.145 & -0.072 \\ -0.072 & 0.035 \end{pmatrix}$$

Ez a lényegében a második, kisebb diád arányait és előjeleit tükrözi, elnyomva mindazt az információt, ami az első diádból származhatna.

A mátrix ilyen rossz tulajdonságát a *kondíciószám* fejezi ki: a mátrix legnagyobb és legkisebb sajátértékének hányadosa. Ez esetünkben mintegy 3.10^3 értékű. BELSLEY (1980. 114.o) utal rá, hogy egy 10^n nagyságú kondíciószám n szignifikáns jeggyel csorbíthatja az eredmény megbízhatóságát. Mivel alapadataink 3 jegyre voltak megadva, nem csodálkozhatunk, ha a számítás eredményében egyetlen szignifikáns jegy sem maradt.

Hozzá kell itt tennünk, hogy szakmailag az alapadatok három helyértékre terjedő megbízhatóságát is kétségbe kell vonnunk. Még ha a fogyasztás és bér adataiban esetleg meg is bíznánk — bár ezek elfogadása a mérés és a statisztika 1 ezrelékes toleranciáját engedi csak meg! — a profitok számbavételekor biztosan nagyobb hibák történtek. Ismerve az adóbevallások szokásos torzításait az első két jegy szignifikáns voltának elfogadása is a naivitás határán jár.

A kiinduló adatok kis változtatása — teszem azt, a tizedespont után álló jegyek törlése vagy kerekítése — éppen a második, kis sajátértékű diád jelentős megváltoztatásához vezetne, míg az első diád értéke aránylag stabil maradna.

Általános indoklás

Tegyük fel, hogy az adatokat a számítás előtt standardizáljuk. Ekkor az (1) egyenlet kifejezhető a három korrelációs együttható értékével. Legyen $\mathrm{Corr}(x_1, x_2) = \alpha$, $\mathrm{Corr}(x_1, y) = \beta$, és $\mathrm{Corr}(x_2, y) = \gamma$. Ezért

$$X^T X = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \quad \text{és} \quad X^T y = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}.$$

Az inverz értéke így

$$(X^T X)^{-1} = 1/(1-\alpha^2) \begin{pmatrix} 1 & -\alpha \\ \alpha & 1 \end{pmatrix}.$$

Ezak alapján

$$b_1 = (\beta - \alpha\gamma)/(1-\alpha^2)$$

$$b_2 = (\gamma - \alpha\beta)/(1-\alpha^2)$$

alakban számítható. Minél közelebb esik hát α^2 értéke 1-hez, annál nagyobb és így bizonytalanabb lesz az $X^T X$ mátrix inverze. E mátrix két sajátértéke $1+\alpha$ és $1-\alpha$,

és ha az utóbbi sajátérték zérushoz közelít, akkor reciproka a végtelenhez tart, s ezzel b_1 és b_2 bizonytalansága is minden határon túl növekszik.

A matrix inverze ugyanis

$$\frac{1}{1+\alpha}\begin{pmatrix}\sqrt{0.5} & \sqrt{0.5}\\ \sqrt{0.5} & \sqrt{0.5}\end{pmatrix} + \frac{1}{1-\alpha}\begin{pmatrix}\sqrt{0.5} & -\sqrt{0.5}\\ -\sqrt{0.5} & \sqrt{0.5}\end{pmatrix}$$

alakban írható fel a diádbontás után. Az első diád szorzója körülbelül 1/2 és nem különösebben érzékeny α értékének 1-hez való közeledésére. A második szorzó azonban $\alpha = 0.95$ esetén 20, $\alpha = 0.99$ esetén 100 és ha α értéke 1-hez tart, akkor végtelenné válik. Mindennek következtében $b_1 + b_2$ értéke mindig jól meghatározott marad, azaz stabilnak tekinthető, hiszen

$$b_1 + b_2 = (\beta + \gamma)(1 - \alpha)/(1 + \alpha)(1 - \alpha) = (\beta + \gamma)/(1 + \alpha)$$

és ez az érték az $\alpha = 1$ határesetben is véges. Amivel bajunk van, az a $b_1 - b_2$ különbség, mivel

$$b_1 - b_2 = (\beta - \gamma)(1 + \alpha)/(1 - \alpha)^2 = (\beta - \gamma)/(1 - \alpha)$$

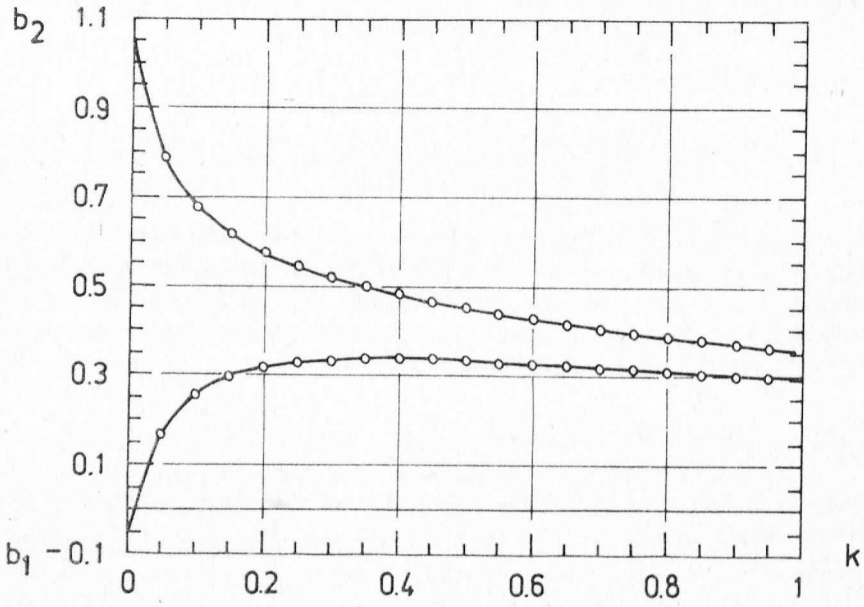
és ez az érték igen érzékeny arra, ha α közelít 1-hez.

A multikollinearitás tehát nem teszi lehetetlenné, hogy a regressziós együtthatók bizonyos lineáris függvényeit — jelen esetben összegét, mivel a nagyobbik sajátértékhez az (1,1) sajátvektor tartozik — a megfelelő biztonsággal becsüljük. Azonban az egyes együtthatókra mégis bizonytalan, sőt helytelen eredmények jönnek létre, mivel más lineáris függvényüket — jelen esetben különbségüket, mert a kisebbik sajátértékhez az $(1, -1)$ sajátvektor tartozik — csak bizonytalanul vagy egyáltalán nem tudjuk megbecsülni.

Újabb javaslatok

Látva a zérushoz közeli, kis sajátértékek zavaró szerepét, a multikollinearitás kezelését célzó javaslatok alapvetően kétfajta stratégiával dolgozhatnak. Az első, agresszív jellegű stratégia mesterségesen megnöveli a sajátértékeket. A második, defenzív stratégia — én ennek pártján állok — belenyugszik a tényekbe és kimenti azt az információt, amely akkor is megmarad, ha ezek a kis sajátértékek akár tovább is csökkennek vagy zérussá válnak.

A támadó jellegű stratégia alapirányzata az úgynevezett „ridge" regresszió, lásd pl. JUDGE (1985. 22.5.2. pont) vagy FOMBY (1984. 13.4.3. pont). Lényege, hogy egy k-értékű diagonális mátrix hozzáadásával megnöveljük a kiinduló mátrix sajátértékeit, s ezzel javítjuk a kondiciószámot. A kiinduló mátrix növelése azonban nyilván csökkenti ennek inverzét, s így becslésünk nem marad torzítatlan, „zsugorodni" fog, mégpedig annál inkább, minél erősebben növeljük a kiinduló mátrixot.



1. ábra. Ridge–regresszió

Az újabb statisztikai programcsomagok általában tartalmazzák a ridge–regressziót, ezért álljon itt egy grafikus kép a b értékek változására a sajátértékek k-val történő növelésének függvényében:

A ridge–becslés alkalmazása azon az (első hallásra meglepő) tételen alapul, hogy hatékonyabb becsléshez juthatunk k valamilyen (eleve nem ismert) pozitív értéke mellett, mint $k = 0$ esetén. A bizonyítást lásd pl SZÉKELY J.G. (1986) 133-4.o.

Mint az a grafikon alapján megállapítható, továbbra sem jutottunk azonban szakmailag elfogadható becsléshez (ezt a szakirodalom alapján a $0.2 < b_1 < 0.6$; $0.9 < b_2 < 1.1$ intervallumokban gyaníthatnánk, lásd pl. ERDŐS–MOLNÁR ,1982) a hatékonyság fokozása az ilyen típusú becslések esetén azonban a linearitás és a torzítatlanság feladását jelenti, s míg az elsőről még esetleg hajlandók is lennénk lemondani, a második elv feladása már igen fájdalmas.

Ez abból is belátható, hogy a sajátértékek k-val való növelése equivalens azzal az (elképzelt) megfigyeléssel, hogy a magyarázó változók külön–külön $\sqrt{k}$ nagyságú kilengése nem okoz változást a függő változó értékében. Ha ugyanis a ténylegesen megfigyelt értékeket kiegészítjük olyan fiktív megfigyelésekkel, amelyekben $y = 0$, x_1 vagy x_2 pedig $\sqrt{k}$ értékű, akkor éppen a ridge–regresszió egyenletéhez jutunk, amely a

$$b = (X^T X + kI)^{-1} X^T y$$

számítási előírással adható meg, ahol I az N-dimenziós egységmátrix.

Pszeudo-inverz és faktorbontás

Eddigi érvelésünk szerint a kisebbik sajátérték és a hozzá tartozó diád lehet puszta esetlegesség (modell vagy adathiba, esetleg kerekítés) eredménye. Ha a „támadás" helyett „védekezünk" akkor kézenfekvő, hogy ezt a zavaró diádot nem-létezőnek, a hozzá tartozó sajátértéket pedig zérusnak tekintsük.

Ez esetben azonban az $X^T X$ mátrix szinguláris válik, ezért a szokásos módon nem invertálható. Továbbra is létezik azonban úgynevezett pszeudo-inverze — lásd pl. BOLLA (1986. 382–3.o.) — amely úgy hozható létre, hogy csak a nem-zérus sajátértékek reciprokával számolunk, a zérus sajátértékeket (s a hozzájuk tartozó diádokat) figyelmen kívül hagyjuk.

Ha így folytatjuk le a számítást, akkor a

$$y = 0.53x_1 + 1.09x_2$$

becslést nyerjük. Ennek korrelációs együtthatója $r = 0.9993$, azaz $r^2 = 0.9986$ tehát csak lényegtelenül alacsonyabb az iméntinél, de még mindig gyanusan magas a közgazdász számára, aki — ismerve az adatok kiküszöbölhetetlen hibáját — semmiképp sem vár ilyen jó illeszkedést. Ugyanakkor az együtthatók számértékei a vizsgált időszak tekintetében szakmailag elfogadhatóbb képet adnak: a nagy válság előtti években a profitjövedelem minegy felét fordíthatták fogyasztásra, a másik fele felhalmozásra került. A munkabér pedig, a részletre és hitelre való vásárlási lehetőségének terjedésével, az életszínvonal állandó növekedését tapasztalva a jövedelemnél valamivel nagyobb fogyasztásra, tehát némi eladósodásra ösztönözte a bérből és fizetésből élőket. (Megjegyzendő itt az is, hogy ez az időszak rendkívül magas bevándorlással járt. E réteg ugyan kétségkívül szegény volt, de pótlólagos pénzeszközöket mindenképpen hozott magával. A modell feltűnő közgazdasági hibája, hogy a fogyasztás egyéb forrásait — az állam, társadalom, turisták, bevándorlók stb. pénzét — nem veszi figyelembe. Ezért aztán mindkét együttható becslése magasabb a valósnál.)

A becslés hibakorlátaira, konfidencia-intervallumaira stb. vonatkozó szokásos számításokat és próbákat azonban most nem tudjuk elvégezni, mivel kiléptünk a regressziószámítás szokásos elméletének kereteiből, s tulajdonképpen átléptünk a főkomponens — és faktoranalízis területére, hiszen csupán az $X^T X$ mátrix főkomponensével dolgoztunk. Itt a megválaszolható kérdések másként alakulnak.

A szinguláris dekompozíció

Kiindulva a Gram-féle mátrix $X^TX = RDR^T$ alakú ortogonális spektrálfelbontásából,[1] ahol $R = R^T = R^{-1}$ a sajátvektorok, D pedig a sajátértékek diagonális mátrixa, X alábbi diadikus felbontását végezhetjük el:

$$X = CD^{1/2}R^T \qquad \text{azaz} \qquad C = XRD^{-1/2}. \tag{2}$$

Ezt a felbontást nevezi a szakirodalom szinguláris dekompozíciónak, lásd BOLLA (1986) 376–7.o. A számítás egyszerűen gépesíthető.[2]

Az *1. táblázat* X adatmátrixának a (2) egyenlet szerinti felbontása két diádra, az eredményeket két tizedesre kerekítve, az alábbiakat adja:

$$X = 128.76 \begin{pmatrix} .24 \\ .28 \\ .32 \\ .32 \\ .34 \\ .35 \\ .36 \\ .37 \\ .39 \end{pmatrix} (.44;\ .9) + 2.35 \begin{pmatrix} -.58 \\ .39 \\ .05 \\ .44 \\ .40 \\ -.18 \\ -.26 \\ -.03 \\ -.25 \end{pmatrix} (.9; -.44)$$

A két diád közül az első, pozitív és viszonylag nagy abszolút értékű számokból álló diád „stabil", abban az értelemben, hogy a kiinduló adatok kisebb változása – például a tizedes értékek megváltoztatása, elhagyása vagy kerekítése — nem fogja érdemileg érinteni. A második diád „labilis" , azaz kis változtatások is erősen befolyásolhatják, esetleg zérussá tehetik.

Nyilván

$$C^TC = D^{-1/2}R^TX^TXRD^{-1/2} = D^{-1/2}R^TRDR^TRD^{-1/2} = 1 \tag{3}$$

ezért C a fenti előállítás alapján már normalizált (egymásra merőleges és egységnyi hosszú vektorokból álló) mátrix lesz. $C^TC = 1$, de természetesen általában $CC^T \neq 1$.

Mindezen jelölések mellett az (1) egyenlet a következőképpen alakul át:

$$b = (X^TX)^{-1}X^Ty = RD^{-1}R^TRD^{1/2}C^Ty = RD^{-1/2}C^Ty. \tag{4}$$

---

[1] Lásd EGERVÁRY (1953) és (1956)

[2] A programcsomagok általában csak a szimmetrikus mátrixok spektrálfelbontását tartalmazzák, de ennek alapulvételével tetszőleges téglalap-mátrix felbontható a fenti módon.

Ez a felbontás ad módot arra, hogy a számunkra oly fontos sajátértékeket és kihatásukat vizsgálhassuk. De az átalakításnak más előnyei is vannak.

A korrelációs együttható

A regresszió elvégzése alapján számított eredmény eltér az eredeti adatoktól:

$$y - Xb = y - X(X^TX)^{-1}X^Ty = y - CD^{-1/2}R^TRD^{-1/2}C^Ty = (1 - CC^T)y, \quad (5)$$

ahol a (2) és a (4) egyenletet vettük figyelembe.

Mármost $(1 - CC^T)^T = 1 - CC^T$, azaz szimmetrikus, és mivel $(1 - CC^T)^2 = 1 - 2CC^T + CC^TCC^T = 1 - CC^T$, ahol kihasználtuk a (3) egyenletet, ezért a mátrix idempotens is, azaz e mátrix úgynevezett projektor.

A számítás elvégzése után fennmaradó „megmagyarázatlan" szórásnégyzet ezek alapján

$$y^T(1 - CC^T)^T(1 - CC^T)y = y^Ty - y^TCC^Ty \quad (6)$$

amiből következik, hogy a „megmagyarázott" szórásnégyzet y^TCC^Ty és így a korrelációs együttható négyzete

$$r^2 = y^TCC^Ty/y^Ty \quad (7)$$

A CC^T mátrix tehát a független y változó elemeinek a hibákra gyakorolt hatását mutatja. Értelmezése különösen egyszerűvé válik, ha a számítást normalizált adatok alapján végezzük, mivel ekkor $y^Ty = 1$.

A számítás folytatása

Képezve a C álló téglalap alakú mátrix és a magyarázandó y változó szorzatát, az eredményt jelöljük a c vektorral:

$$C^Ty = c. \quad (8)$$

E jelöléssel az eredeti (1) illetőleg (4) feladatot az igen egyszerű

$$b = RD^{-1/2}c$$

alakban számíthatjuk.

Ez a c érték számpéldánkban egy tizedesre $(155.9, -1.8)$. A második diádhoz tartozó érték itt is kicsi, és bizonytalan. Azonban a $D^{-1/2}$ mátrixxal való szorzás után már az $(1.21 - 0.77)$ értékhez jutunk, itt a két diádhoz tartozó eltérő nagyságú és

biztonságú érték már közel került egymáshoz. Végül b értéke az R mátrixxal való szorzatból adódik, azaz

$$\begin{pmatrix} 0.44 & 0.9 \\ 0.9 & -0.44 \end{pmatrix} \begin{pmatrix} 1.21 \\ -0.77 \end{pmatrix} = \begin{pmatrix} -0.15 \\ 1.43 \end{pmatrix}$$

(a kisebb pontatlanság a kerekítésekből adódik). Ebben már teljesen elmosódott a bizonytalanság minden nyoma.

Mindezek alapján jogosultnak látszott az a felfogás, hogy a második diádot, amely a mérés pontatlanságának köszönheti bizonytalanságát, sőt egyáltalán létrejöttét is és mindenképpen igen esetleges, tekintsük nem létezőnek és *hagyjuk el* a számításból. Ez nem azt jelenti, hogy az amúgyis kicsi sajátértéket még kisebbé tesszük, mert ez éppen „végtelen" erőssé tenné a zavaró hatást. Az *egész* diádot kell nem létezőnek tekinteni. Ebben az esetben tehát csak az R mátrix első oszlopa a mérvadó és a

$$b = \begin{pmatrix} 0.44 \\ 0.9 \end{pmatrix} 1.21$$

összefüggésből a regresszió

$$y = 0.53x_1 + 1.09x_2$$

becslését nyertük.

További megfontolások

Már Egerváry utalt rá, SZÉKELY B. (1970, 1976) bizonyította és JÓZSA (1973) részletesen is tárgyalta, hogy a faktor-bontás az egy-egy diád leválasztása után megmaradó információt minimálissá teszi. Tehát az első diád (amelynek Logaritmiában oly nagy szerepe van) a maximális információt tartalmazza, amely a megfigyelt adatokból — lineáris módszerekkel — egyáltalán kinyerhető.

Meddig haladjunk az ilyen faktor-bontással? A közgazdász hajlamos lesz kialakult arányérzéke alapján (amely a statisztikai adatok megbízhatóságának ismeretéből táplálkozik) akkor megállni, amikor már csak a harmadik helyértéknek megfelelő vagy ennél kisebb számok jelennek meg a maradék-mátrixban. A matematikus — lásd REJTŐ (1986. 99.o.) — azt tudja ehhez hozzátenni, hogy a teljes varianciát az X^TX mátrix nyoma (diagonális elmeinek összege), azaz a sajátértékek összege adja meg, az tehát a (2) egyenlet alapján $\Sigma_i d_i$, számpéldánk szerint pedig 16581.9. Annyi faktort kell tehát kiválasztani, hogy a választott, mondjuk r számú faktor sajátértékeinek összege e variancia célul kitűzött hányadát megmagyarázza. Ezért a

$$\frac{d_1 + d_2 + \ldots + d_r}{\sum_i d_i}$$

hányadost kell mérvadónak tekinteni. Mi egyetlen faktort választottunk, s ez a 16576.4/16581.8 = 0.99967 hányadost adja. Ezért nem romlott a korrelációs

együttható érezhetően, mivel a teljes varianciának kevesebb mint 23 százezrelékét hanyagoltuk el.

Az ördög azonban a részletekben lakik: a kérdés az, hogy az így elhanyagolt diádot milyen valószínűséggel tekinthetjük véletlen mátrixok szorzatából adódó úgynevezett Wishart-mátrixnak? Sajnos erre, mint TUSNÁDY (1986. 22.o.) írja, még nincs igazán használható eljárás, nincs elfogadott próba.

Bármilyen kicsi a második faktor és bármilyen kicsi a szerepe a teljes varianciában, a közgazdász már ránézésre is gyanúsnak találja és nem tekintheti véletlennek, mert egyrészt határozott (és az irodalomban másutt már tárgyalt) ciklikus viselkedést mutat, másrészt — és ezen belül — a bér és a profit ellentétes irányú mozgását tételezi. A ciklus „súlya" az általános növekedésen belül ugyan viszonylag csekély (az X mátrix sajátértékeivel jellemezve 2.35/128.76, azaz kevesebb mint 2%-os), de semmiképp sem véletlen műve. Ezzel azonban ismét elérkeztünk kis számpéldánk sajátos korlátaihoz, amelyek a „modellezés" túlzásbavitt egyszerűsítéseinek és elhanyagolásainak következményei. A számpélda mentségére szolgáljon, hogy kizárólag az orvosolhatatlan „együttnövekedés" azaz multikollinearitás problémáját kívánta a lehető legegyszerűbben bemutatni.

(Beérkezett: 1988. augusztus 8-án.)

Irodalom

BELSLEY, D.A. – KUHN, E. – WELSH, R.E. (1980): *Regression Diagnostics.* Wiley, New-York.

BOLLA M. (1986): *Lineáris algebrai segédeszközök.* A MÓRI F.J. – SZÉKELY J.G. (1986) kötetben.

BRÓDY A. (1984): Barangolás Logaritmiában, *Szigma* XVII. 147–156.

CARLETT, W. (1987): Multicollinearity. Címszó a *The new Palgrave* c. közgazdasági enciklopédiában. Macmillan, London.

EGERVÁRY J. (1953): Mátrixfüggvények kanonikus előállításáról és annak néhány alkalmazásáról. *MTA III. Osztályának Közleményei.* II.1.sz. 417–458.

EGERVÁRY J. (1956): Az inverz mátrix általánosítása. *Az MTA Matematikai Kutatóintézetének Közleményei.* I. évf. 315–324.

ERDŐS P. – MOLNÁR F. (1982): *Válság és infláció a 70-es évek amerikai gazdaságában.* Közgazdasági és Jogi Könyvkiadó, Budapest.

FOMBY, T.B. – HILL, R.C. – JOHNSON, S.R. (1984): *Advanced Econometric Methods*, Springer, Berlin.

FRISCH, R. (1934): *Statistical confluence analysis by means of complete regression systems.* Oslo Universsity Institute of Economics. Oslo.

GAUDI J. (1986): A többváltozós statisztikai analizis számítógépes eljárásai. A MÓRI F.J. – SZÉKELY J.G. (1986) kötetben.

GANTMACHER, F.R. (1954): *Teoria Matric.* Goszizdat Techniko-Teoreticseszkoj Litteraturi, Moszkva.

JÓZSA S. (1973): A diád-módszer statisztikai expozíciója *Szigma.* 1. 63–72.

JUDGE, G.G. – GRIFFITHS, W.E. – HILL, R.C. – LÜTKEPOHL, H. – LEE, T.C. (1985): *The theory and practice of econometrics.* Wiley, New York.

KLEIN, L.R. (1950): *Economic Fluctuations in the United States*, 1921–1941. Wiley, New-York.

MALINVAUD, E. (1974): *Az ökonometria statisztikai módszerei.* Közgazdasági és Jogi Könyvkiadó, Budapest.
MÓRI F.J. – SZÉKELY J.G. (szerk) (1986): *Többváltozós statisztikai analízis.* Műszaki Könyvkiadó, Budapest.
REJTŐ L. (1986): A változók számának csökkentése: főkomponens és faktoranalízis. A MÓRI F.J. – SZÉKELY J.G. (1986) kötetben.
RIMLER, J. (1976): *Fejlődéselemzés ökonometriai módszerekkel.* Közgazdasági és Jogi Könyvkiadó, Budapest.
SZÉKELY B. (1970): Mátrixok egy speciális diadikus felbontása ennek néhány alkalmazása az összehasonlító elemzésben. *Szigma*, 4. 241–253.
SZÉKELY B. (1976): Mátrixok egy speciális diadikus felbontása. Rimler (1976) művében. 315–324.
SZÉKELY G. (1986): Regressziós modellek. A MÓRI F.J. – SZÉKELY J.G. (1986): kötetben.
THEIL, H. (1971): *Principles of Econometrics.* Wiley and Sons, New York.
TUSNÁDY G. (1986): A többdimenziós normális eloszlás. A MÓRI F.J. – SZÉKELY J.G. (1986): kötetben.

About Irremediable Multicollinearity

Economic time series are very often pervaded by a strong multicollinearity which bedevils multiple regression computations. Illustrating the phenomenon by a well known example, used already by Theil, the literature on remedial action is reviewed. By accepting, rather than fighting, multicollinearity one is led to exploiting the generalized inverse with small eigenvalues set to zero. This approach, utilizing results of Egerváry, maintains that the small eigenvalues do not carry any information and represent only errors of measurement or "noise". An exact test for "noise" is, alas, not yet established by statistical theory.