A korrelációs matrix alkalmazása többváltozós elemzésnél

JÓZSA SÁNDOR

Agrártudományi Egyetem, Keszthely

A gyakorlati problémák természete leginkább többváltozós, azaz nem írhatók le egy vagy két ismérv, jellemző („változó") számszerű megadásával. Az egyes ismérvek többnyire nem függetlenek egymástól, többé-kevésbé korreláltak. Egy-egy ilyen összetett kérdés elemzésére ismert és közhasználatúnak mondható néhány statisztikai módszer: a változók közötti páronkénti (lineáris) korrelációk meghatározása, esetleg parciális korrelációk számítása, továbbá regressziós függvények keresése és az ezekkel kapcsolatos statisztikai próbák ([4] 11.4).

Jelen dolgozatban a korrelációk egy általánosabb osztályát mutatom be, kiemelve a korrelációs matrixból történő számíthatóságukat, továbbá megmutatom, hogyan lehet e matrix segítségével többszörös regressziók együtthatóit számítani. A módszerek mélyebb megvilágításán kívül ezzel egyszerű számítási eljárások adódnak, melyeket számítógépes feldolgozásnál különösen előnyösnek vélek. Tételezzük fel, hogy a változók közötti kapcsolatok lineárisak — ellenkező esetben megfelelő transzformációkkal ez elérhető ([4] 14).

Az a tény, hogy a változók együttes normális eloszlása esetén a korrelációs matrix meghatározza a változó-rendszer szerkezetét, biztosíték arra, hogy pusztán a korrelációs matrixból megkaphatunk majdnem minden, a változók közötti összefüggésekre vonatkozó információt ([2] 234. o.).

Az eljárások könnyebb megértése érdekében egy példán vezetem végig a számításokat. A példa adatanyagát Ragasits István (Keszthely, Agrártudományi Egyetem, Növénytermesztéstani Tanszék) volt szíves rendelkezésemre bocsátani, melyért ezúton mondok köszönetet.

Nem lenne értelme az egyes állítások részletesebb bizonyításával terhelni az olvasót, az érdeklődők az irodalmi utalások nyomán rekonstruálhatják a bizonyításokat. Nem tekinthettem viszont el néhány eljárás matematikai indoklásától, melyekre nem találtam utalást az irodalomban, így ezeket, a rövidség kedvéért csak nagy vonalakban, függelékben csatolom.

Az elemzésbe bevont változók számát p-vel jelöljük. Korreláció alatt e dolgozatban a becsült értékét értjük. Az i.-ik és a j.-ik változó (közönséges) korrelációját (és az együtthatót is) jelöljük r_{ij}-vel. A korrelációs matrix jelölésében olykor feltüntetjük a szóbanforgó változókat vagy indexeiket is:

$$\mathbf{R} = \mathbf{R}_{x_1 x_2 \ldots x_p} = \mathbf{R}_{12\ldots p} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & & 1 \end{pmatrix}$$

A változók csoportjai közötti összefüggés mérése

Az $x_1, x_2, \ldots x_p$ változókat osszuk csoportokba és jelöljük e csoportokat $X_1, X_2, \ldots X_k$-val, pl. $X_1 = (x_1, x_3)$, $X_2 = (x_2, x_5, x_6)$ stb. E csoportok közötti kapcsolatot a következő kifejezéssel mérhetjük:

$$\mathbf{R}^2_{X_1X_2..X_k} = 1 - \frac{\det \mathbf{R}_{X_1X_2 \cdots Xk}}{\det \mathbf{R}_{X_1} \det \mathbf{R}_{X_2} \ldots \det \mathbf{R}_{Xk}} \tag{1}$$

Könnyen belátható, hogy e mérőszám a közönséges korrelációt adja, ha két változónk van:

$$\mathbf{R}^2_{x_1x_2} = 1 - \frac{\det \mathbf{R}_{x_1x_2}}{\det \mathbf{R}_{x_1} \det \mathbf{R}_{x_2}} = 1 - \frac{\det \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}}{\det[1] \det[1]} = 1 - \frac{1^2 - r^2}{1 \cdot 1} = r^2$$

(1) speciális eseteként kapjuk a többszörös korrelációt is, mégpedig akkor, ha a változókat két csoportra osztjuk és az egyik csoportban egyetlen változót hagyunk. Pl. x_1-nek az $(x_2, x_3 \ldots x_p)$ változókra vonatkozó többszörös korrelációs együtthatója:

$$R^2_{1\,(23\ldots p} = 1 - \frac{\det \mathbf{R}_{123\ldots p}}{\det \mathbf{R}_{23\ldots p}} \tag{1a}$$

Abban a speciális esetben, ha $X_1 = x_1$, $X_2 = x_2, \ldots X_p = x_p$, (1) véleményem szerint igen fontos összefüggésmérő számot eredményez, melyet tartalmára tekintettel, g l o b á l i s k o r r e l á c i ó - nak nevezek a továbbiakban, és R_g-vel jelölöm:

$$\mathbf{R}^2_g = \mathbf{R}^2_{1.2.\ldots p} = 1 - \det \mathbf{R}_{12\ldots p} \tag{1b}$$

Egy számolástechnikai megjegyzés: A fenti számításokhoz nem szükséges az r_{ij} korrelációs együtthatók tényleges meghatározása, elegendő az SP (eltérés sorozatösszeg) és SQ (eltérés négyzetösszeg) értékek kiszámítása. Jelöljük ugyanis az $SP_{x_ix_j}$ számokból álló matrixot **S**-sel Fennáll:

$$\det \mathbf{R}_{12\ldots p} = \frac{\det \mathbf{S}_{12\ldots p}}{SQ_{x_1} \cdot SQ_{x_2} \ldots SQ_{x_p}}$$

Ennek felhasználásával pl. (1a) így módosul:

$$\mathbf{R}^2_{1(23\ldots p,} = 1 - \frac{\det \mathbf{S}_{12\ldots p}}{SQ_{x1} \det \mathbf{S}_{23\ldots p}}$$

Az (1) alatt definiált mérőszám tartalmát a következő tulajdonságai érzékeltetik.

1. $0 \leqslant R^2 \leqslant 1$, mégpedig

a) $R^2 = 0$ akkor és csak akkor, ha az $X_1, X_2, \ldots X_k$ változócsoportok páronként korrelálatlanok. Két változócsoportot akkor nevezünk korrelálatlannak, ha az egyik csoportból vett bármelyik változó és a másik csoportból vett bármelyik változó között a korreláció zéró.

b) $R^2 = 1$ akkor és csak akkor, ha a változók között analitikus lineáris összefüggés áll fenn.

2. R^2 értéke nem csökken, ha a csoportfelosztást finomítjuk, azaz $R^2_{X_1X_2\ldots X_k} \leqslant R^2_{X_1X_2\ldots X_k'X_{k+1}'}$, ahol $X_k = \{X_k', X_{k+1}'\}$. Egyenlőség akkor áll, ha X_k' és X_{X+1}' korrelálaltlanok.

3. R^2 értéke nem csökken, ha újabb változókat csatolunk a változórendszerhez: $R^2_{X_1X_2\ldots X_k} \leqslant R^2_{X_1X_2\ldots X_kX_{k+1}}$. Egyenlőség akkor áll, ha X_{k+1} korrelálatlan a $X_1, X_2, \ldots X_k$ csoportok mindegyikével.

2. és 3.-ból következik, hogy az $x_1, x_2, \ldots x_p$ változókra számított R_g felső korlátot ad e változók közötti bármely (1) típusú korrelációra, beleértve e változóknak csak egy részére vonatkozó korrelációkat is. Speciálisan $R_g = R_{1(23\ldots p)}$ pontosan akkor áll, ha az $x_2, x_3, \ldots x_p$ változók páronként korrelálatlanok.

Megjegyezzük, hogy (1) nevezőjének valamely tényezője elvileg zérónak adódhat (a gyakorlatban ennek alig van esélye). Ilyenkor a számláló is szükségképpen zéró, az osztás nem végezhető el. Ennek ellenére $R^2_{n_1X_2\ldots X_k}$ értelmezhető, a következő módon. Annak az X_i csoportnak, melynek determinánsa zéró, egymás után kidobunk egy-egy tagját (1) jobb oldalából mindaddig, míg egy nem zéró determinánsú ilyen redukált csoportot kapunk. Ha az egy változóval csökkentett csoportok mindegyike zéró determinánst eredményez, változópárokat hagyunk el, és így tovább.

Konkrét elemzésnél a következő módon járhatunk el. A páronkénti korrelációs együtthatók kiszámítása után összeállítjuk a korrelációs matrixot. Ha a korrelációs együtthatók nem túl nagyok, esetleg kifejezetten kicsik, kiszámítjuk a globális korrelációt. (Ha a páronkénti együtthatók között 1-hez közeli értékű is van, R_g nem mondhat újat. Ilyenkor a magasan korrelált változópárok egyik tagját ideiglenesen kihagyhatjuk és a megmaradt változókat elemezhetjük.) Ha R_g zéróhoz közeli értékű, a változók közel állnak a páronkénti korrelálatlansághoz, ilyenkor az elemzést változónként külön kell végezni. Ha R_g közel 1, kereshetjük azt a változót (esetleg több is van ilyen), mely jó közelítéssel előállítható a többi lineáris függvényeként, azaz (1a) szerint számíthatjuk a többszörös korrelációkat. R_g nagyságától függetlenül érdekes lehet a változók bizonyos (esetleg a szakprobléma által definiált) csoportjai — különösen két csoport — közötti korrelációk meghatározása. Külön figyelmet érdemel az a tény, hogy az (1) alatt definiált összes lehetséges korreláció mindössze $2^p - p - 1$ determináns kiszámítását igényli.

Összetettebb kérdések kezelésére illusztrálásképpen megemlítem azt a problémát, melyet Sváb János és Wellisch Péter vetettek fel a közelmúltban. Adott nagyszámú változó (ismérv), melyek között többnyire magas korrelációk mutatkoznak. Kézenfekvő a gondolat, hogy ezen ismérvek egy részének mérése felesleges, a többi ismérv jól leírja ezeket. A feladat tehát az, hogy kiválasszuk azt a lehetőleg kevés számú ismérvet, melyek az összes figyelembe vett ismérv információtartalmának közel 100%-át tartalmazzák. Egyik megoldás nyers alapgondolata a következő lehet. Annak az információnak a mérésére, melyet

a változók valamely X csoportja az x_i változóról tartalmaz, elfogadjuk az $R^2_{x_iX}$ korrelációnégyzetet[1] (determinációs együttható).

Kimutatható, hogy két változócsoport — X_1 és X_2 — egymásról kölcsönösen tartalmazott információja (információelméleti értelemben)) $I(X_1, X_2) = -\log R_{x1x2}$, tehát R^2_{x1x2} valóban a szóbanforgó információ egy mérőszáma. Megjegyezzük, hogy e szám nem azonos a (bonyolultan számítható) kanonikus korrelációval (ld. [3], 288. old.) viszont tartalmuk hasonló.

Az X által tartalmazott átlagos információt az

$$I(X) = \frac{1}{p} \sum_{i=1}^{p} R^2_{x_iX}$$

mennyiséggel mérhetjük. $I(X)$ nyilván 0 és 1 közé esik. Feladatunk olyan X csoportot keresni, melyben kevés változó van és melyhez nagy I tartozik, valamint amelyre $R^2_{x_iX}$ elég nagy mindegyik i-re.

Az eljárás ebben a formában eléggé számolásigényes, de egyszerű meggondolásokkal számítógépre elfogadhatóvá redukálható. E modell természetesen még igen durva, legfeljebb jobb híján ajánlható. (A probléma szigorú matematikai megoldásáról nincs tudomásom.)

P é l d a. Réti csenkesz magfüves terméselemei közül négyet elemzünk, így a számítások kézzel is elvégezhetők, az eredmények könnyen áttekinthetők. A négy változó a következő: x_1 = bugahossz (mm), x_2 = padkaszám (db), x_3 = buga ág (db), x_4 = kalászkaszám (db). 25 mintából az alábbi eltérés négyzetösszegeket, eltérés szorzatösszegeket és korrelációkat kaptuk (a változók páronkénti pontdiagramjai elfogadhatóan lineáris kapcsolatot mutattak, transzformációt alkalmazni egyik változóra sem volt indokolt).

változó	SQ	átlag
x_1	213,83	19,68
x_2	65	12,28
x_3	180	17,60
x_4	1468	32,80

változópár	SP	r
x_1, x_2	73,41	0,62
x_1, x_3	62,24	0,32
x_1, x_4	411,22	0,73
x_4, x_3	59,80	0,55
x_2, x_4	139,40	0,45
x_3, x_4	297,00	0,58

A korrelációk a négy változó meglehetősen szoros összefüggésére utalnak. A változók korrelációs matrixa (a kis mintaszám miatt elegendő két tizedesjegy pontosság):

$$\mathbf{R} = \begin{pmatrix} 1{,}00 & 0{,}62 & 0{,}32 & 0{,}73 \\ 0{,}62 & 1{,}00 & 0{,}55 & 0{,}45 \\ 0{,}32 & 0{,}55 & 1{,}00 & 0{,}58 \\ 0{,}73 & 0{,}45 & 0{,}58 & 1{,}00 \end{pmatrix}$$

R harmadrendű aldeterminánsait gyorsan kiszámíthatjuk a „Sarrus-szabály" alkalmazásával ([2] 389. o.). Jelöljük az **R** első sorának és j-ik oszlopának elhagyásával maradó 3×3-as matrix determinánsát D_{1j}-vel. (Pl. $D_{12} = 0{,}32 \times 0{,}58 \times 0{,}73 + 0{,}62 \times 1{,}00 \times 1{,}00 + 0{,}55 \times 0{,}58 \times 0{,}45 - 0{,}55 \times$

$\times$ 0,32 $\times$ 0,62 — 0,45 $\times$ 1,00 $\times$ 0,73 — 0,58 $\times$ 0,58 $\times$ 1,00 = 0,2233.) A D determinánsokkal kapjuk:

$$\det \mathbf{R} = 1 \times D_{11} - r_{12}D_{12} + r_{13}D_{13} - r_{14}D_{14} =$$
$$= 0{,}4458 - 0{,}62 \times 0{,}2233 + 0{,}32 \times 0{,}1667 - 0{,}73 \times 0{,}3216 =$$
$$= 0{,}1259$$

Ebből: $R_g^2 = 1 - 0{,}1259 = 0{,}8741$ és $R_g = 0{,}93$.

Szignifikancia-vizsgálat nélkül is elfogadhatjuk, hogy a négy változó között a „globális" összefüggés igen szoros.

Szükség van még a következő determinánsokra:

$$\det \mathbf{R}_{234} = 0{,}4458 \qquad \det \mathbf{R}_{124} = 0{,}2875$$
$$\det \mathbf{R}_{134} = 0{,}2993 \qquad \det \mathbf{R}_{123} = 0{,}4289$$

továbbá

$$\det \mathbf{R}_{12} = 1 - r_{12}^2 = 0{,}6123$$
$$\det \mathbf{R}_{13} = 1 - r_{13}^2 = 0{,}8994 \quad \text{stb.}$$

A kapott $2^4 - 4 - 1 = 11$ determináns alapján bármely kívánt korreláció meghatározható. Az (1a) formula alkalmazásával kapjuk:

$$R_{1(234)}^2 = 1 - \frac{\det \mathbf{R}}{\det \mathbf{R}_{234}} = 1 - \frac{0{,}1259}{0{,}4458} = 0{,}7176$$

Hasonlóan adódnak a további determinációs együtthatók:

$$R_{2(134)}^2 = 0{,}5794 \quad R_{3(124)}^2 = 0{,}5621 \quad R_{4(123)}^2 = 0{,}7065$$

Látható, hogy az (x_2, x_3, x_4) csoport tartalmaz legtöbb információt a négy változóra vonatkozóan, mégpedig durván

$$I(x_2, x_3, x_4) = 1/4\ (3 + 0{,}7176) = 93\%\text{-ot.}$$

Kiszámíthatjuk az egy-egy változó és másik kettő közötti korrelációkat, pl.:

$$\mathbf{R}_{2(13)}^2 = 1 - \frac{\det \mathbf{R}_{123}}{\det \mathbf{R}_{13}} = 1 - \frac{0{,}4289}{1 - 0{,}32^2} = 0{,}5231$$

Hasonlóan kapjuk: $R_{4(13)}^2 = 0{,}6672$.

Az (x_1, x_3) csoport tehát durván

$$I(x_1, x_3) = 1/4\ (2 + 0{,}5231 + 0{,}6672) = 80\%$$

információt nyújt a négy ismérv együtteséről. A további változópárok információtartalma ennél kisebb.

Végül számítsuk ki a két-két pár közötti korrelációkat. A legtöbb információt tartalmazó (x_1, x_3) pár és az (x_2, x_4) pár korrelációjára adódik:

$$\mathbf{R}^2_{13.24} = 1 - \frac{\det \mathbf{R}_{1234}}{\det \mathbf{R}_{13} \det \mathbf{R}_{24}} = 1 - \frac{0{,}1259}{0{,}8994 \cdot 0{,}7975} = 0{,}8245$$

Hasonlóan: $R^2_{12.34} = 0{,}6901$, $R^2_{14.23} = 0{,}6136$.

Regressziós együtthatók meghatározása

Miután kiszámítottuk a többszörös korrelációs együtthatókat, meghatározhatók egyes relevansnak ítélt lineáris regressziók paraméterei. E célra szintén felhasználhatjuk a korrelációs matrixot a következő módon. Az

$$x_1 = b_{10} + b_{12}x_2 + b_{13}x_3 + \ldots + b_{1p}x_p$$

regresszióhoz a b_{12} együttható becsült értéke:

$$b_{12} = \frac{\sqrt{SQ_{x_1}}}{\sqrt{SQ_{x_2}}} b'_{12} \qquad \text{ahol} \quad b'_{12} = - \frac{\det \begin{pmatrix} 0 & (1\ 0 \ldots 0) \\ r_{21} & \\ \vdots & \mathbf{R}_{23\ldots p} \\ r_{p1} & \end{pmatrix}}{\det \mathbf{R}_{23\ldots p}} \qquad (2)$$

melyben SQ_{x_1} ill. SQ_{x_2} az x_1 ill. x_2 változóból számolt eltérés-négyzetösszeg. Hasonlóan kapható b_{13}, b_{14} stb., ha (2) számlálójában az (1 0 . . . 0) sorvektort a (0 1 0 . . . 0), (0 0 1 0 . . . 0) — és így tovább — sorvektorral cseréljük ki, és SQ_{x_2} helyett SQ_{x_3}, SQ_{x_4} stb. szerepel. Ha $\det \mathbf{R}_{23\ldots p} = 0$, a felírt regresszió nem egyértelmű (multikollinearitás, ld. [2] 308. o.).

Végül: $b_{10} = \bar{x}_1 - \sum_{k=2}^{p} b_{1k} \bar{x}_k$

A (2) formula is számítható az **S** matrixból:

$$b'_{12} = - \frac{\det \begin{pmatrix} 0 & (SQ_1\ 0 \ldots 0) \\ SP_{21} & \\ \vdots & \mathbf{S}_{23\ldots p} \\ SP_{p1} & \end{pmatrix}}{SQ_1 \cdot \det \mathbf{S}_{23\ldots p}}$$

P é l d a. Az előbbi példaanyagra számítsuk ki az

$$x_1 = b_{10} + b_{12}x_2 + b_{13}x_3 + b_{14}x_4$$

többszörös regresszió együtthatóit. (Az illeszkedés szorosságát ismerjük: $R^2_{1(234)} = 0{,}7176$.) A (2) képlet szerint:

$$b'_{12} = -\frac{\det\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.62 & 1.00 & 0.55 & 0.45 \\ 0.32 & 0.55 & 1.00 & 0.58 \\ 0.73 & 0.45 & 0.58 & 1.00 \end{pmatrix}}{\det \mathbf{R}_{234}}$$

A számláló első sora szerinti kifejtéssel adódik:

$$b'_{12} = D_{12}/\det \mathbf{R}_{234} = 0{,}2233/0{,}4458 = 0{,}5009, \text{ mellyel}$$

$$b_{12} = \sqrt{SQ_{x1}/SQ_{x2}} \cdot b'_{12} = \sqrt{213{,}83/65} \cdot 0{,}5009 = 0{,}91$$

Hasonlóan kapjuk a további együtthatókat:

$$b'_{13} = -D_{13}/\det \mathbf{R}_{234} = -0{,}1667/0{,}4458 = -0{,}3739\,, \quad \text{és}$$
$$b_{13} = -\sqrt{213{,}83/180} \times 0{,}3739 = -0{,}41$$
$$b'_{14} = D_{14}/\det \mathbf{R}_{234} = 0{,}3216/0{,}4458 = 0{,}7214\,, \quad \text{és}$$
$$b_{14} = \sqrt{213{,}83/1468} \times 0{,}7214 = 0{,}28$$

Végül:

$$b_{10} = \bar{x}_1 - (b_{12}\bar{x}_2 + b_{13}\bar{x}_3 + b_{14}\bar{x}_4) =$$
$$= 19{,}68 - (0{,}91 \times 12{,}28 - 0{,}41 \times 17{,}60 + 0{,}28 \times 32{,}80) = 6{,}54$$

A regresszió tehát:

$$x_1 = 6{,}54 + 0{,}91x_2 - 0{,}41x_3 + 0{,}28x_4$$

Az egyes regressziós együtthatók szignifikanciáját a többszörös korrelációk segítségével vizsgálhatjuk. Pl. a $b_{13} = -0{,}41$ együttható szignifikanciáját az

$$F_{1,21} = \frac{R_{1(234)} - R_{1(24)}}{1 - \mathbf{R}_{1(234)}} \times (25 - 4) = \frac{0{,}0670}{0{,}2935} \cdot 21 = 4{,}79$$

értékkel lehet ellenőrizni, mely itt 5%-os szinten szignifikáns.

A korrelációs matrix sajátértékeiről

A változók rendszerének mélyebb elemzésére érdemes kiszámítani **R** sajátértékeit: $\lambda_1, \lambda_2, \ldots \lambda_p$ ([2] 388. o.). E sajátértékek összege p. Az $x_1, x_2, \ldots x_p$ rendszer helyettesíthető p számú, páronként korrelálatlan változóval, mondjuk $y_1, y_2, \ldots y_p$, úgy, hogy y_i az $(x_1, x_2, \ldots x_p)$ rendszer információtartalmának 100 λ_i/p %-át foglalja magába. (Az y változókat főösszetevőknek nevezik, meghatározásukhoz **R** sajátvektoraira is szükség van. Bővebben ld. [1] 11. fej. és [3] 6. fej.). Ha tehát bizonyos, mondjuk k számú sajátérték zérónak tekinthető, akkor az $x_1, x_2, \ldots x_p$ rendszer mindössze $p - k$ változóval is leírható.[1]

E gondolatsorból kiindulva van remény a változók számának redukálásáról fentebb megfogalmazott probléma egzakt megoldására.

Szoros kapcsolat áll fenn a sajátértékek és a globális korreláció között, nevezetesen: $R_g^2 = 1 - \lambda_1 \cdot \lambda_2 \ldots \lambda_p$. Ebből belátható, hogy $R_g^2 = 1$ pontosan akkor, ha valamely $\lambda_i = 0$, és $R_g = 0$ pontosan akkor, ha mindegyik $\lambda_i = 1$, azaz ha **R** egységmatrix, tehát a változók páronként korrelálatlanok. Itt említem meg, hogy R^2 — speciálisan R_g — szignifikanciája a Wilks-próbával ellenőrizhető ([2], [1] 9. fej. és [5]).

P é l d a. Visszatérve a réti csenkesz terméselemeiről összeállított **R** matrixra, számítsuk ki sajátértékeit. A következő negyedfokú egyenletet (karakterisztikus polinom) kell megoldanunk:

$$\lambda^4 - 4\lambda^3 + 4{,}13\lambda^2 - 1{,}46\lambda + 0{,}13 = 0$$

melyben az együtthatók: 4 = a változók száma, $4{,}14 = \Sigma \det \mathbf{R}_{ij} = \binom{4}{2} - \Sigma r^2$, $1{,}46 = \Sigma \det \mathbf{R}_{ijk}$ és 0,13 = det **R**. Az egyenlet gyökeit iterációval határoztuk meg: $\lambda_1 = 2{,}64$, $\lambda_2 = 0{,}70$, $\lambda_3 = 0{,}53$, $\lambda_4 = 0{,}13$. Ellenőrizhető, hogy összegük 4, szorzatuk $0{,}13 = 1 - R_g^2$.

Eszerint a négy változóból másik — páronként korrelálatlan — négy változó képezhető (főösszetevők), melyek információ-tartalma rendre: 2,64/4 = 66%, 0,70/4 = 17,5%, 0,53/4 = 13,3% és 0,13/4 = 3,2%. Két változóba 66% + 17,5% = 83,5%-nyi információ sűríthető, három változóval pedig gyakorlatilag teljesen (96,8%-ban) leírható a négy terméselem.

Matematikai függelék

Az (1a) és (2) formulák igazolásának kulcspontja a következő azonosság (mely következik [1] Appendix 1. 5. tételéből):

$$\mathbf{a}^* \mathbf{Q}^{-1} \mathbf{b} = c - \frac{\det \begin{pmatrix} c & \mathbf{a}^* \\ \mathbf{b} & \mathbf{Q} \end{pmatrix}}{\det \mathbf{Q}} \tag{3}$$

melyben c tetszőleges szám, **a** és **b** tetszőleges vektorok, **Q** tetszőleges négyzetes matrix (rendje megegyezik **a** és **b** komponenseinek számával).

Legyen x_i' az x_i változó standardizáltja. x_1'-nek az $x_2', x_3', \ldots x_p'$ változókra vonatkozó regressziós együtthatóinak **b′** vektorára a normálegyenlet rendszer:

$$\mathbf{R}_{23\ldots p} \mathbf{b}' = \mathbf{r}_1 \quad \text{ahol} \quad \mathbf{r}_1 = (r_{12}, r_{13}, \ldots r_{1p})^*$$

amiből

$$\mathbf{b}' = \mathbf{R}_{23\ldots p}^{-1} \mathbf{r}_1$$

b′ komponenseit egységvektorokkal való szorzásával nyerjük, pl.:

$$b_{12}' = (1, 0, \ldots 0)\mathbf{b}' = (1, 0, \ldots 0)\, \mathbf{R}_{23\ldots p}^{-1} \mathbf{r}_1$$

(3) alkalmazásával innen azonnal adódik (2).

Az $R^2_{1(23\ldots p)}$ többszörös korrelációs együtthatót $\mathbf{b}'$ és a normálegyenlet jobb oldalának szorzata adja:

$$R^2_{1(23\ldots p)} = \mathbf{r}_1^* \mathbf{b}' = \mathbf{r}_1^* \mathbf{R}^{-1}_{23\ldots p} \mathbf{r}_1$$

Innen (1a)-t kapjuk, ha a (3) formulát $c = 1$, $\mathbf{a} = \mathbf{b} = \mathbf{r}_1$ és $\mathbf{Q} = \mathbf{R}_{23\ldots p}$ „szereposztás" mellett alkalmazzuk.

Összefoglalás

Többváltozós elemzés bázisát képezheti a korrelációs matrix. Segítségével a változók különböző csoportjai közötti kapcsolatok fedhetők fel, a közismert módszereknél általánosabb, könnyebben megjegyezhető, egyszerű számítási eljárásokkal.

Külön figyelmet érdemel a több változó közötti összefüggés egy speciális, szimmetrikus mérőszáma, a „globális" korrelációs együttható, mely az elemzés kezdetekor megfelelően orientál a változók közötti maximális korrelációról.

Egy matrixelméleti azonosság /(3)/ lehetővé teszi, hogy a többszörös regressziós együtthatókat is egyszerűen számíthassuk a korrelációs matrix felhasználásával.

A módszerek részletes bemutatása réti csenkesz magfüves terméselemeinek elemzésén történt, az eredmények szakmai megbeszélésének igénye nélkül.

Irodalom

[1] ANDERSON, T. W.: An Introduction to Multivariate Statistical Analysis. Wiley. London. 1958.
[2] Biometriai értelmező szótár. Mezőgazd. Kiadó Budapest, 1966.
[3] SEAL, H.: Multivariate Statistical Analysis for Biologists. Methuen & Co. London. 1964.
[4] SVÁB, J.: Biometriai módszerek a mezőgazdasági kutatásban. Mezőgazd. Kiadó. Budapest. 1967.
[5] WILKS, S. S.: Mathematical Statistics. Princeton, University Press. 1943.

Érkezett: **1971.** *március* **13.**

Application of Correlation Matrix for Multivariate Analyses

S. JÓZSA
University of Agricultural Sciences, Keszthely (Hungary)

Summary

Correlation matrix may be efficiently used for multivariate analyses. The formula (1) gives a measuring value for the correlations between the groups of variables. The special case of it is the determination coefficient (1a) for linear correlations. Another special case of (1) is the „global" correlation coefficient given in (1b), that can be very useful in getting first informations concerning the variables.

The identity of (3) makes the computation of multiple regression coefficients possible by the application of the correlation matrix, according to the formula (2).

The calculations were carried out on the analysis of the Festuca grass-seed yield elements. The yield elements involved were: x_1 = length of panicle, x_2 = nodes on the rachis, x_3 = panicle branches, x_4 = number of spikelets.

Anwendung der Korrelationsmatrix bei Multivariablenanalysen

S. JÓZSA
Universität für Agrarwissenschaften, Keszthely (Ungarn)

Zusammenfassung

Bei Multivariablenanalysen kann die Korrelationsmatrix vorteilhaft angewendet werden. Die Formel (1) ergibt Messzahlen, die den Zusammenhang zwischen den einzelnen Gruppen der Variablen kennzeichnen. Im Falle eines linearen Zusammenhanges gelangt man zu einem spezialen Fall von Formel (1), zum Determinationskoeffizienten. Einen weiteren spezialen Fall der Formel (1) ergibt der unter (1b) angegebene „globale" Korrelationskoeffizient, der zu der ersten Orientierung unter den Variablen von grossem Nutzen ist.

Durch Anwendung der Korrelationsmatrix können laut Formel (2) die mehrfachen Regressionskoeffizienten errechnet werden, was durch die Identität (3) ermöglicht wird.

Die Berechnungen wurden mit der Analyse der einzelnen Ertragskomponenten des Wiesen-Schwingels (Festuca pratensis) durchgeführt. Die in Betracht gezogenen Ertragskomponenten: x_1 = Rispenlänge, x_2 = Stufenzahl, x_3 = Rispenzweig, x_4 = Ährchenzahl.

Использование корреляционных матриц при многофакторном анализе

Ш. ЙОЖА
Аграрный Университет, Кестхей (Венгрия)

Резюме

При многофакторном анализе успешно можно использовать корреляционные матрицы. Формула (1) дает показатель зависимости между группами переменных, для которых специальным случаем является детерминационный коэффициент (1a) наряду с линейной зависимостью. Другой особый случай (1) это написанный под (1*b*) «глобальный» коэффициент корреляции, который очень важен для первой ориентировки между переменными.

Используя корреляционную матрицу, по формуле (2) можно рассчитать многократные регрессионные коэффициенты, что дает возможность для отождествления (3).

Расчеты провели при анализе структурных элементов урожая луговой овсяницы. Во внимание принимались: x_1 = длина метелки. x_2 = число колосковых подошв. x_3 = метелочки. x_4 = число колосков.