A DIADIKUS ADATELEMZÉS MÓDSZERTANÁNAK EGY KRITIKAI VIZSGÁLATA: A KETTŐS ADATBEVITEL ÉS FELCSERÉLHETŐ ESET[1]

DOBOS IMRE
Budapesti Corvinus Egyetem

A dolgozatban egy korábban ismertetett új módszertan, az ún. diadikus adatelemzés matematika-statisztikai alapjait gondolom újra. A szerző már vizsgálta, hogy a bevezetett módszertan a klasszikus statisztikai módszertanhoz képest információ növekedéssel jár-e. Kísérletet teszek a diadikus adatelemzés matematikai struktúrájának korrekciójára.

Kulcsszavak: matematikai statisztika, korrelációelemzés, diadikus adatelemzés

1 Bevezetés

A diadikus adatelemzéshez hasonló adatfelvétel ismert a páros minták elméletéből, így maga a módszertan nem teljesen ismeretlen a statisztikus társadalom számára. A páros minták módszere azzal a kérdésfeltevéssel él, hogy a párosan felvett minta két összetartozó párja azonos eloszlással, várható értékkel és szórással rendelkezik-e. Az ilyen kérdések a társadalomtudományok széles körében, mint a szociológia, pszichológia, vagy az orvostudományok egy része, már nem elegendőek, mert a párosan felvett minta összetartozó elemeire, azaz diádjaira több változó (kérdés) megválaszolását teszi szükségessé. (Kenny et al. (2006))

Az előbbi gondolatmenet miatt a páros minták módszere alkalmatlan arra, hogy a megfigyelések közötti sztochasztikus kapcsolatokat, összefüggéseket jellemezze, de nem is az a célja. (Vincze-Varbanova (1993))

Dolgozatom célja, hogy a klasszikus statisztika problémakörét rávetítsem a diadikus adatelemzésre, és azt vizsgáljam, hogy az eddig kifejlesztett módszertani megoldások mennyire tekinthetőek kielégítőnek matematikai értelemben. A klasszikus statisztikában a változók közötti kapcsolatokat leíró fontosabb módszerek az alábbiak:

- korrelációelemzés,
- ok-okozati kapcsolatok elemzése,
- regresszió elemzés, stb.

[1]Beérkezett: 2017. február 4. Dobos Imre a Budapesti Corvinus Egyetem egyetemi tanára. E-mail: imre.dobos@uni-corvinus.hu. A szerző köszöni az OTKA K 115542 és a Dortmundi Műszaki Egyetem (Németország) Gambrinus Fellowship Programme-ja támogatását.

A diadikus adatelemzés egyik első állomása az egyes változók adatpárjainak homogenitásvizsgálata, azaz annak eldöntése, hogy az összetartozó adatpár elemei azonos sztochasztikus jellemzőkkel bírnak-e. A homogenitásvizsgálat ebben az esetben nem két minta eloszlásának az azonosságát célozza, hanem azt vizsgálja, hogy a páros lekérdezésben részt vevő diádok válaszadói azonos válaszokat adnak-e az adott kérdésre. Ezt a feladatot a klasszikus statisztika az ANOVA-táblák elemzésével végezheti el. A diadikus adatelemzés az ún. páros adatbevitel (double entry) módszerének bevezetésével a korrelációelemzést javasolja, mint megoldást erre. (Gelei-Dobos-Sugár (2014), Gelei-Sugár (2016))

A diadikus adatelemzéssel foglalkozó tanulmányokban alkalmazott módszertanok matematikai háttere az esetek nagy részében nem tisztázott teljesen. A változók közötti kapcsolatok szorosságának mérésére a diadikus adatelemzés eddig kifejlesztett módszertana nem ad általánosan elfogadható megoldást.

A dolgozat második fejezetében a diadikus adatelemzés felcserélhető esetét állítom a vizsgálat középpontjába, belátva, hogy a pontos adatelemzéshez még további feltételezésekkel szükséges élni. A következő részben a diadikus adatelemzés homogenitásvizsgálatán keresztül a fontosabb statisztikai mérőszámokat állítom elő. A negyedik fejezetben kísérletet teszek a korrelációs fogalmak pontosítására, azok alapadatokra történő visszavezetésével. Az ötödik rész a diadikus adatelemzésben használt regressziós modelleket veszi górcső alá, belátva, hogy a kettős adatbevitel módszere információveszteséget okozhat, majd összegezem elemzéseimet.

2 Az adatfelvételről: a felcserélhető eset

A diadikus adatelemzés két mintatípust különböztet meg: a felcserélhető (exchangeable case) és a nem felcserélhető, azaz megkülönböztethető (distinguishable case) megfigyelésből álló párokat. (Gonzalez-Griffin (2000)) A nem felcserélhető esetben a diádban szereplő objektumok aszimmetrikus helyzetben vannak, míg a felcserélhető esetben teljesen egyenrangúak a diádba bekerült megfigyelések. A fejezet további részében csak a felcserélhető esettel foglalkozom.

Tegyük fel, hogy három diád került a mintánkba, amit az *1. táblázatban* szemléltetek. Mivel felcserélhető volt az adatfelvétel, ezért nem tudunk a szereplőink (adataink) között semmiféle különbséget tenni, azaz felcserélhetőek a diádon belüli adatfelvételek.

Megfigyelések	1. változó (X)	
	1. adat (X_1)	2. adat (X_2)
1. számú pár	x_{11}	x_{12}
2. számú pár	x_{21}	x_{22}
3. számú pár	x_{31}	x_{32}

1. táblázat. A diadikus adatelemzés három diád esetén

Ha a megfigyelésünk felcserélhető, amit feltételeztem, akkor a következő, *2. táblázat* is egy lehetséges induló táblázat, amit úgy nyerünk, hogy az 1. diádban felcseréltük a megfigyelésünket. Ezt szemlélteti a 2. táblázat.

Megfigyelések	1. változó (X)	
	1. adat (X'_1)	2. adat (X'_2)
1. számú pár	x_{12}	x_{11}
2. számú pár	x_{21}	x_{22}
3. számú pár	x_{31}	x_{32}

2. táblázat. A diadikus adatelemzés három diád esetén az első diád elemeinek felcserélése után

Az előbb szemléltetett eljárást még további hat alkalommal folytathatjuk, azaz felcserélhető esetben $2^3 = 8$ különböző indulótáblázatunk lehet. Ezt általánosítva, ha n darab diád áll rendelkezésre a vizsgálatokhoz, akkor 2^n különböző induló táblázat áll rendelkezésre, mivel nem tudunk a diád elemei között különbséget tenni.

A teljesség kedvéért soroljuk fel a *3. táblázatban* adódó nyolc mintát.

	Minta 1	Minta 2	Minta 3	Minta 4	Minta 5	Minta 6	Minta 7	Minta 8
1.pár	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{12}, x_{11})	(x_{12}, x_{11})	(x_{12}, x_{11})	(x_{12}, x_{11})
2.pár	(x_{21}, x_{22})	(x_{21}, x_{22})	(x_{22}, x_{21})	(x_{22}, x_{21})	(x_{21}, x_{22})	(x_{21}, x_{22})	(x_{22}, x_{21})	(x_{22}, x_{21})
3.pár	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})

3. táblázat. A diadikus adatelemzés három diád esetén előálló minták felsorolása

A fentiek következménye, hogy olyan módszert kell az adatelemzéshez keresni, amely az előbbiekben előálló problémát kezelni tudja. Hogy ez a rendezési probléma milyen nehézséget okoz, azt egy korábbi dolgozatból származó adatállományon szemléltetem. (Gelei-Dobos, 2016)

Most csak egy változót vizsgálok (mennyire ismerik egymást a válaszadók), és az adatfelvétel véletlenszerűen kialakult sorrendjét veszem a diádokban. Tételezzük fel azt is, hogy a feladat annak a vizsgálata, hogy – páros mintát feltételezve – a két „minta" átlaga azonos-e. Az is feltételezhető, hogy a két minta összefügg, ezért a páros minták átlagára vonatkozó próba alkalmazható.

A következő lépésben aztán cseréljük fel a diádok elemeit úgy, hogy az első oszlopban a diád válaszai közül a kisebb értékű, míg a második oszlopban a magasabb értékű elem szerepeljen. Az SPSS 22 programcsomag a következő eredményt adta a két esetben.

	Páros különbség					t-teszt	Szignifikancia
	Átlag	Szórás	Sztenderd hiba	95%-os konfidencia intervallum			
				alsó	felső		
1.minta	0,07865	1,79788	0,19058	−0,30008	0,45738	0,413	0,681
2.minta	1,13483	1,39146	0,14749	0,84172	1,42795	7,694	0,000

4. táblázat. A két vizsgálat összevetése, páros minták tesztje, $df = 88$

A *4. táblázatból* azonnal leolvasható, hogy az adatfelvitel során kapott mintában a diádokban szereplő válaszadók lényegében azonos választ adtak, mivel az átlagok eltérése szignifikánsan nem utasítható el. Míg a másik, újrarendezett mintában, ahol nagyság szerint rendeztük a válaszokat, az eredmény az, hogy szignifikánsan el kell utasítani az átlagok egyezőségét.

Ez a vizsgálat tehát azt támasztja alá, hogy a felcserélhető esetben az adatok felcserélhetőségi problémájával foglalkoznunk kell. Erre lehet egy válasz a kettős adatbevitel (double enrty). Ez azonban nem oldja meg az előbbi problémát, csak az adatállományt növeli kétszeresére, de ekkor is a lehetséges „minták" száma 2^n számú lesz.

A fentiek miatt olyan adatelemzési módszert kell találni, ami tökéletesen független az adatok felviteli sorrendjétől. Az ilyen operáció lehet a diádon belüli adatok összegének és/vagy különbségük abszolút értékének a megragadása, mert az minden egyes diádra állandó, függetlenül a felvitel sorrendjétől. (Ezzel később foglalkozom még a korreláció kapcsán!)

Itt jegyezem meg, hogy a nem felcserélhető esetben ez a probléma nem áll fenn, mert a diadikus, páros mintavétel esetén az oszlopok egyértelműen meghatározottak az (aszimmetrikus) szerepek rögzítésével.

A további elemzésemben abból indulok ki, hogy már rögzítettek, hogy a páros adatokban melyik válasz kerül az első ill. a második helyre, ezzel a megkülönböztethető és felcserélhető eseteket nem kell külön elemezni.

3 A homogenitásvizsgálat és kettős adatbevitel módszere

A kettős adatbevitel során a diád tagjai által adott összes választ egy diádok szerint rendezett vektorba töltjük fel; valamint egy új, másik vektort is konstruálunk, amiben az előbbi vektor szereplő diádelemeket felcseréljük. (Bővebben magyarul lásd (Gelei-Dobos-Sugár, 2014).) Ez azt is jelenti, hogy az eddig n elemű vektorokból $2n$ eleműekké transzformáltuk adatállományunkat.

Tételezzük most fel, hogy a diádokon belül a válaszadók sorrendjét rögzítettük, vagyis nem áll fenn az előbb vázolt felcserélhetőségi probléma. Jelölje most két változóra adott rendes adatfelvétel értékeit (x_1, x_2) és (y_1, y_2), valamint a kettős adatbevitel értékeit (X, X') és (Y, Y'). Mivel az (X, X') és (Y, Y') értékeket az eredeti adatainkból nyertük, ezért – az adatok átrendezhetőségét feltételezve – azt kapjuk, hogy

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad X' = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \text{valamint} \quad Y' = \begin{bmatrix} y_2 \\ y_1 \end{bmatrix},$$

ami azt mutatja, hogy az új változók a régiekből úgy származtathatóak, hogy a diád két megfigyelésvektorát egymás alá helyezzük, csak fordított sorrendben. Arra a kérdésre keresem a választ, hogy a kettős adatbevitel bevezetése az elemzésbe mennyire árnyalja a statisztikai vizsgálatokat, egyáltalán, információtöbbletet nyerhető-e vele.

Először a két módszer közötti összefüggéseket mutatom meg a klasszikus statisztika olyan mérőszámain keresztül, mint az átlag, a szórásnégyzet és a kovariancia. A torzítottság problémáját elkerülendő, tegyük fel, hogy most a vektorok az alapsokaságot képviselik, így – a számításokat megkönnyítendő – a vektorok elemszámával kell a variancia-kovariancia számításakor számolni. Természetesen mintát torzítatlanságot feltételezve is hasonló eredményeket kapnánk.

Először számítsuk ki az átlagot a két változóra mindkét esetben. Ekkor

$$E(X) = E(X') = \frac{E(x_1) + E(x_2)}{2} \qquad \text{és} \qquad E(Y) = E(Y') = \frac{E(y_1) + E(y_2)}{2},$$

ami egyszerű számolással belátható. Ez azt is jelenti, hogy a kettős adatbevitellel nyert új változók átlagai megegyeznek az eredeti elemek átlagának átlagával. Másként is megragadható ez, mégpedig azzal, hogy egy adott kérdésre adott összes válasz átlaga a kettős adatbevitellel nyert X és Y vektor átlaga.

A szórásnégyzetek kiszámítása sem nehéz, de türelmet igényel:

$$\operatorname{var}(X) = \operatorname{var}(X') = \frac{\operatorname{var}(x_1) + \operatorname{var}(x_2)}{2} + \left(\frac{E(x_1) - E(x_2)}{2}\right)^2$$

és

$$\operatorname{var}(Y) = \operatorname{var}(Y') = \frac{\operatorname{var}(y_1) + \operatorname{var}(y_2)}{2} + \left(\frac{E(y_1) - E(y_2)}{2}\right)^2.$$

Már csak a kovarianciák meghatározása maradt hátra

$$\operatorname{cov}(X, X') = \operatorname{cov}(x_1, x_2) - \left(\frac{E(x_1) - E(x_2)}{2}\right)^2$$

és

$$\operatorname{cov}(Y, Y') = \operatorname{cov}(y_1, y_2) - \left(\frac{E(y_1) - E(y_2)}{2}\right)^2,$$

valamint

$$\operatorname{cov}(X, Y) = \operatorname{cov}(X', Y') =$$
$$= \frac{\operatorname{cov}(x_1, y_1) + \operatorname{cov}(x_2, y_2)}{2} + \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}$$

és

$$\operatorname{cov}(X, Y') = \operatorname{cov}(X', Y) =$$
$$= \frac{\operatorname{cov}(x_1, y_2) + \operatorname{cov}(x_2, y_1)}{2} - \frac{[E(x_1) - E(x_2)] \cdot [E(y_1) - E(y_2)]}{4}.$$

Azonnal meg kell jegyezni, hogy a kettős adatbevitel lényegesen csökkenti a rendelkezésre álló információmennyiséget azzal, hogy az új változók átlagai, szórásnégyzetei, de kovarianciái közül számos azonos. Az új (X, X') és (Y, Y')

változók sztochasztikus mérőszámaiból nem tudjuk az előbbi szimmetriák miatt a (x_1, x_2) és (y_1, y_2) valószínűségi változók megfelelő mutatóit kiszámítani. Ez azt jelenti, hogy a logikai kapcsolat a két adathalmaz között egyirányú, azaz (x_1, x_2) és (y_1, y_2) változók egyértelműen meghatározzák az (X, X') és (Y, Y') változókat, viszont megfordítva ez nem igaz. Az információvesztés tehát ebből az aszimmetriából származik.

A fentiekből az is következik, hogy csak néhány esetben tudunk a változókra az új és régi kovarianciák között relációt felállítani. Ezek az esetek pedig a következők:

$$\operatorname{cov}(X, X') \le \operatorname{cov}(x_1, x_2) \quad \text{és} \quad \operatorname{cov}(Y, Y') \le \operatorname{cov}(y_1, y_2) \ ,$$

valamint a szórásnégyzetekre, amely szintén a kovarianciának egy speciális esete

$$\operatorname{var}(X) \ge \frac{\operatorname{var}(x_1) + \operatorname{var}(x_2)}{2} \ge \sqrt{\operatorname{var}(x_1)} \cdot \sqrt{\operatorname{var}(x_2)}$$

és

$$\operatorname{var}(Y) \ge \frac{\operatorname{var}(y_1) + \operatorname{var}(y_2)}{2} \ge \sqrt{\operatorname{var}(y_1)} \cdot \sqrt{\operatorname{var}(y_2)} \ .$$

Ha feltételezzük, hogy a diád párjai közel azonosan válaszolnak, vagyis a szereplők válaszainak átlaga közel azonos, amit az alábbi módon írhatunk:

$$\max\{|E(x_1) - E(x_2)|; |E(y_1) - E(y_2)|\} \le \varepsilon \ ,$$

ahol ε tetszőlegesen kicsi pozitív szám, akkor az alapadatok ismeretében az alábbi közelítések adhatók a kettős adatbevitellel nyert valószínűségi változókra:

$$\operatorname{var}(X) = \operatorname{var}(X') \sim \frac{\operatorname{var}(x_1) + \operatorname{var}(x_2)}{2} \ ,$$

$$\operatorname{var}(Y) = \operatorname{var}(Y') \sim \frac{\operatorname{var}(x_1) + \operatorname{var}(x_2)}{2} \ ,$$

$$\operatorname{cov}(X, X') \sim \operatorname{cov}(x_1, x_2) \ , \qquad \operatorname{cov}(Y, Y') \sim \operatorname{cov}(y_1, y_2) \ ,$$

$$\operatorname{cov}(X, Y) = \operatorname{cov}(X', Y') \sim \frac{\operatorname{cov}(x_1, y_1) + \operatorname{cov}(x_2, y_2)}{2}$$

valamint

$$\operatorname{cov}(X, Y') = \operatorname{cov}(X', Y) \sim \frac{\operatorname{cov}(x_1, y_2) + \operatorname{cov}(x_2, y_1)}{2} \ .$$

Az előbbi összefüggések elemi matematikai módszerekkel igazolhatóak, ettől itt eltekintek. A varianciákról azt lehet megállapítani, hogy az X változó szórásnégyzete nagyobb, mint az őt alkotó két vektor (változó) szórásának szorzata. Ez információvesztést jelenthet.

Mivel $\operatorname{cov}(X, Y)$ és $\operatorname{cov}(X, Y')$ kovarianciák esetén az átlagok szorzatai a jobb oldalon pozitívak és negatívak is lehetnek, ezért nagyságrendi becslés

nem adható a régi és új változók kovarianciáinak nagyságrendi viszonyáról, viszont az könnyen megállapítható, hogy

$$\operatorname{cov}(X,Y)+\operatorname{cov}(X,Y')=\operatorname{cov}(X,Y+Y')=$$
$$\frac{\operatorname{cov}(x_1,y_1)+\operatorname{cov}(x_1,y_2)+\operatorname{cov}(x_2,y_1)+\operatorname{cov}(x_2,y_2)}{2}=\frac{\operatorname{cov}(x_1+x_2,y_1+y_2)}{2}\,,$$

ami a variancia-kovariancia algebra alkalmazásával kapható meg.

A két korrelációt az alábbi képletekkel határozhatjuk meg:

$$r(X,X')=\frac{\operatorname{cov}(X,X')}{\operatorname{var}(X)}=\frac{\operatorname{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2}+\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}=$$
$$=\frac{\sqrt{\operatorname{var}(x_1)}\cdot\sqrt{\operatorname{var}(x_2)}\cdot r(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2}+\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}$$

és

$$r(Y,Y')=\frac{\operatorname{cov}(Y,Y')}{\operatorname{var}(Y)}=\frac{\operatorname{cov}(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2}+\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}=$$
$$=\frac{\sqrt{\operatorname{var}(y_1)}\cdot\sqrt{\operatorname{var}(y_2)}\cdot r(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2}+\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}\,.$$

Itt használható fel az, hogy a két pár új változó szórásnégyzete megegyezik. Ha feltételezzük újra, hogy a párok válaszainak átlaga közel azonos, akkor ezek a korrelációk a következő módon közelíthetők:

$$r(X,X')\sim\frac{\sqrt{\operatorname{var}(x_1)}\cdot\sqrt{\operatorname{var}(x_2)}}{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2}}\cdot r(x_1,x_2)\le r(x_1,x_2)$$

és

$$r(Y,Y')\sim\frac{\sqrt{\operatorname{var}(y_1)}\cdot\sqrt{\operatorname{var}(y_2)}}{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2}}\cdot r(y_1,y_2)\le r(y_1,y_2)\,.$$

Ez már sejtetni engedi, hogy a diadikus adatelemzés homogenitásvizsgálatát a szokásos ANOVA-táblákon kívül az eredeti, induló adatállományon is el lehet végezni, nem szükséges az új változók bevezetése. Nevezetesen $r(x_1,x_2)$ és $r(y_1,y_2)$ korrelációkon keresztül is mérhető, hogy a diádban szereplőknek, egy adott kérdésre adott válaszai egyeznek-e, vagy sem, azaz lineárisan összefüggnek-e.

Az elvégzett számítások a megkülönböztethető esetben is teljesülnek, így a javasolt módszer abban az esetben is alkalmazható. A következő rész a változók közötti lineáris kapcsolatokat elemzi a korrelációk segítségével.

4 Lineáris kapcsolat vizsgálata korreláció-elemzéssel diadikus adatokra

A diadikus adatelemzés ötféle korrelációs együtthatót határoz meg. (Griffin-Gonzalez (1995), Gonzalez-Griffin (1999), Gonzalez-Griffin (2000)) Ezeket az előbb említett dolgozat alapján elemezem, és bemutatom az előbbi szerzők által javasolt korrelációknak a gyengeségét.

A válaszadó belső korrelációját ($R(X,Y)$) a diadikus adatelemzés a következő képlettel határozza meg, ami átírható az alapadatokra:

$$r(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)} \cdot \sqrt{\operatorname{var}(Y)}} =$$

$$= \frac{\frac{\operatorname{cov}(x_1,y_1)+\operatorname{cov}(x_2,y_2)}{2} + \frac{[E(x_1)-E(x_2)][E(y_1)-E(y_2)]}{4}}{\sqrt{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2} + \left(\frac{E(x_1)-E(x_2)}{2}\right)^2} \cdot \sqrt{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2} + \left(\frac{E(y_1)-E(y_2)}{2}\right)^2}} =$$

$$= \frac{\frac{\sqrt{\operatorname{var}(x_1)\operatorname{var}(y_1)}\cdot r(x_1,y_1)+\sqrt{\operatorname{var}(x_2)\operatorname{var}(y_2)}\cdot r(x_2,y_2)}{2} + \frac{[E(x_1)-E(x_2)]\cdot[E(y_1)-E(y_2)]}{4}}{\sqrt{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2} + \left(\frac{E(x_1)-E(x_2)}{2}\right)^2} \cdot \sqrt{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2} + \left(\frac{E(y_1)-E(y_2)}{2}\right)^2}} .$$

A kovarianciák a képlet számlálójában azt mérik, hogy milyen irányú sztochasztikus kapcsolat van a diádban szereplő párok saját válaszai között, tehát ennyiben ez valóban egy „belső", de fogalmazhatunk úgy is, hogy individuális korrelációt mutat.

Ha újra feltételezzük a várható értékek közel azonos voltát, valamint a szórásnégyzetek is közel esnek egymáshoz

$$\max\{|\operatorname{var}(x_1) - \operatorname{var}(x_2)|; |\operatorname{var}(y_1) - \operatorname{var}(y_2)|\} \leq \eta ,$$

ahol η tetszőlegesen kicsi pozitív szám, akkor erre a korrelációra is adható egy közelítés

$$r(X,Y) \sim \frac{1}{2} \cdot \frac{\sqrt{\operatorname{var}(x_1)\operatorname{var}(y_1)} \cdot r(x_1,y_1) + \sqrt{\operatorname{var}(x_2)\operatorname{var}(y_2)} \cdot r(x_2,y_2)}{\sqrt{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2}} \cdot \sqrt{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2}}} \leq$$

$$\leq \frac{1}{2} \cdot \big[r(x_1,y_1) + r(x_2,y_2)\big] .$$

A keresztkorrelációk a következő módon határozhatóak meg:

$$r(X,Y') = \frac{\operatorname{cov}(X,Y')}{\sqrt{\operatorname{var}(X)}\cdot\sqrt{\operatorname{var}(Y')}} =$$

$$= \frac{\frac{\operatorname{cov}(x_1,y_2)+\operatorname{cov}(x_2,y_1)}{2} - \frac{[E(x_1)-E(x_2)][E(y_1)-E(y_2)]}{4}}{\sqrt{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2} + \left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2} + \left(\frac{E(y_1)-E(y_2)}{2}\right)^2}} =$$

$$= \frac{\frac{\sqrt{\operatorname{var}(x_1)\operatorname{var}(y_2)}\cdot r(x_1,y_2)+\sqrt{\operatorname{var}(x_2)\operatorname{var}(y_1)}\cdot r(x_2,y_1)}{2} - \frac{[E(x_1)-E(x_2)]\cdot[E(y_1)-E(y_2)]}{4}}{\sqrt{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2} + \left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2} + \left(\frac{E(y_1)-E(y_2)}{2}\right)^2}} .$$

Az alapadatok kovarianciája erre a korrelációra azt mutatja, hogy a diádban szereplők válaszai a pár másik kérdésre adott válaszaival milyen sztochasztikus kapcsolatban van. Adható erre is egy lokális közelítés, az előbbi gondolatmenetet követve:

$$r(X,Y') \sim \frac{1}{2}\cdot\frac{\sqrt{\operatorname{var}(x_1)\operatorname{var}(y_2)}\cdot r(x_1,y_2)+\sqrt{\operatorname{var}(x_2)\operatorname{var}(y_1)}\cdot r(x_2,y_1)}{\sqrt{\frac{\operatorname{var}(x_1)+\operatorname{var}(x_2)}{2}}\cdot\sqrt{\frac{\operatorname{var}(y_1)+\operatorname{var}(y_2)}{2}}} \leq$$

$$\leq \frac{1}{2}\cdot\left[r(x_1,y_2)+r(x_2,y_1)\right] .$$

Vizsgáljuk most a diád szintű korrelációt! Ennek a képlete az

$$r_m(X,X',Y,Y') = \frac{r(X,Y)+r(X,Y')}{\sqrt{1+r(X,X')}\cdot\sqrt{1+r(Y,Y')}}$$

kifejezéssel írható le. (Giffin-Gonzalez, 1995) Átírható a fenti korreláció a varianciákkal és kovarianciákkal. Ekkor kisebb átalakításokkal

$$r_m(X,X',Y,Y') = \frac{\operatorname{cov}(X,Y)+\operatorname{cov}(X,Y')}{\sqrt{\operatorname{var}(X)+\operatorname{cov}(X,X')}\cdot\sqrt{\operatorname{var}(Y)+\operatorname{cov}(Y,Y')}}$$

alakot kapjuk. Felhasználva, hogy $\operatorname{var}(X) = \operatorname{cov}(X,X)$, ami természetesen az Y vektorra is teljesül, valamint elemi kovariancia algebrával azt kapjuk, hogy

$$r_m(X,X',Y,Y') = \frac{\operatorname{cov}(X,Y+Y')}{\sqrt{\operatorname{cov}(X,X+X')}\cdot\sqrt{\operatorname{cov}(Y,Y+Y')}} .$$

Ez utóbbi kifejezés – a kovarianciák kiszámítása után – az alapadatokra írható át, ami

$$r_m(X,X',Y,Y') = \frac{\frac{1}{2}\operatorname{cov}(x_1+x_2,y_1+y_2)}{\sqrt{\frac{1}{2}\operatorname{var}(x_1+x_2)}\cdot\sqrt{\frac{1}{2}\operatorname{var}(y_1+y_2)}} = r(x_1+x_2,y_1+y_2) .$$

Ez az utóbbi eredmény azt jelenti, hogy a diád szintű korreláció egy tényleges korreláció, amely két újonnan bevezetett változó közötti korrelációt úgy értelmez, hogy a diádok megfigyeléseinek összegével azonosítja azt. Érdekes módon az $r_m(X, X', Y, Y')$ kifejezés nem egyezik meg egy hagyományos Pearson-féle korrelációval az új adatokra nézve, mert a számlálóban szereplő kovariancia azt feltételezné, hogy a nevezőben lévő kovarianciák helyett a $\sqrt{\text{var}(X)} \cdot \sqrt{\text{var}(Y + Y')}$ kifejezés álljon. Ha valaki veszi a fáradtságot, és végigszámolja a valódi korrelációt, akkor az alábbiakat kapja:

$$r(X, Y + Y') = \frac{\text{cov}(X, Y + Y')}{\sqrt{\text{var}(X)} \cdot \sqrt{\text{var}(Y + Y')}} =$$

$$= \frac{\frac{1}{2}\,\text{cov}(x_1 + x_2, y_1 + y_2)}{\sqrt{\frac{\text{var}(x_1)+\text{var}(x_2)}{2} + \left(\frac{E(x_1)-E(x_2)}{2}\right)^2} \cdot \sqrt{\frac{1}{2}\,\text{var}(y_1 + y_2)}},$$

ami nem egyezik a kapott $r(x_1 + x_2, y_1 + y_2)$ korrelációval, de nagyon jól közelíti azt.

Ezek után térjünk rá a Gelei-Dobos-Sugár (2014) dolgozatban is bemutatott legproblémásabb korrelációs definíciók vizsgálatára, azaz az egyéni és páros szintű korrelációk elemzésére. Az egyéni szintű korreláció javasolt képlete:

$$r_i(X, X', Y, Y') = \frac{r(X, Y) - r(X, Y')}{\sqrt{1 - r(X, X')} \cdot \sqrt{1 - r(Y, Y')}}$$

Átírható ez is a varianciák-kovarianciák segítségével:

$$r_i(X, X', Y, Y') = \frac{\text{cov}(X, Y) - \text{cov}(X, Y')}{\sqrt{\text{var}(X) - \text{cov}(X, X')} \cdot \sqrt{\text{var}(Y) - \text{cov}(Y, Y')}} =$$

$$= \frac{\text{cov}(X, Y - Y')}{\sqrt{\text{cov}(X, X - X')} \cdot \sqrt{\text{cov}(Y, Y - Y')}}.$$

Mielőtt tovább alakítanánk az előbbi kifejezést, itt is határozzuk meg a Pearson-i értelemben vett korrelációt, azaz számítsuk ki a tényleges korrelációt:

$$r(X, Y - Y') = \frac{\text{cov}(X, Y - Y')}{\sqrt{\text{var}(X)} \cdot \sqrt{\text{var}(Y - Y')}} =$$

$$= \frac{\frac{1}{2}\,\text{cov}(x_1 - x_2, y_1 - y_2) + \frac{[E(x_1)-E(x_2)]\cdot[E(y_1)-E(y_2)]}{4}}{\sqrt{\frac{\text{var}(x_1)+\text{var}(x_2)}{2} + \left(\frac{E(x_1)-E(x_2)}{2}\right)^2} \cdot \sqrt{\text{var}(y_1 - y_2) + (E(y_1) - E(y_2))^2}},$$

ami nem más, mint amit az irodalom javasol.

Folytassuk a javasolt korreláció visszavezetését az alapadatokra. A kifejezés nagy hasonlóságot mutat a diád szintű korrelációval, a különbség az előjelek ellentétessége. A korreláció további vizsgálata során helyettesítsük a legutolsó képletbe az alapadatainkat:

$$r_i(X,X',Y,Y') = \frac{\operatorname{cov}(x_1-x_2,y_1-y_2)+[E(x_1)-E(x_2)]\cdot[E(y_1)-E(y_2)]}{\sqrt{\operatorname{var}(x_1-x_2)+[E(x_1)-E(x_2)]^2}\cdot\sqrt{\operatorname{var}(y_1-y_2)+[E(y_1)-E(y_2)]^2}} =$$
$$= \frac{\sqrt{\operatorname{var}(x_1-x_2)}\cdot\sqrt{\operatorname{var}(y_1-y_2)}\cdot r(x_1-x_2,y_1-y_2)+[E(x_1)-E(x_2)]\cdot[E(y_1)-E(y_2)]}{\sqrt{\operatorname{var}(x_1-x_2)+[E(x_1)-E(x_2)]^2}\cdot\sqrt{\operatorname{var}(y_1-y_2)+[E(y_1)-E(y_2)]^2}} .$$

Ez az összefüggés azt mutatja, hogy az egyéni szintű korreláció egy felső korlátja a változók közötti azon korreláció, amikor a párok válaszainak különbségei közötti korrelációt határozzuk meg, természetesen abszolút értékben vizsgálva.

Adjunk becslést erre a korrelációra, feltételezve, hogy a diádok párjai várható értéke közel esik egymáshoz a két kérdésre, vagyis változóra:

$$r_i(X,X',Y,Y') \sim r(x_1-x_2,y_1-y_2) ,$$

ami azt jelenti, hogy ez a korreláció a diádok közötti individuális hatást mérheti valóban.

Tekintsük végül a páros szintű korrelációt. Ennek a képlete:

$$r_d(X,X',Y,Y') = \frac{r(X,Y')}{\sqrt{r(X,X')}\cdot\sqrt{r(Y,Y')}} .$$

Azonnal meg kell jegyezni, hogy ez a fajta korreláció nem szigorúan vett korreláció, mert a négyzetgyök alatti kifejezések negatív értéket is felvehetnek. Ez azt is jelentheti, hogy a diád párjai teljesen ellentétes választ adtak, amivel ez akár negatívvá válhat. Ettől most eltekintek, feltételezve a gyök alatti nemnegativitást. A formula a korreláció definícióját alkalmazva alakítható tovább:

$$r_d(X,X',Y,Y') = \frac{\operatorname{cov}(X,Y')}{\sqrt{\operatorname{cov}(X,X')}\cdot\sqrt{\operatorname{cov}(Y,Y')}} =$$
$$= \frac{\frac{\operatorname{cov}(x_1,y_2)+\operatorname{cov}(x_2,y_1)}{2} - \frac{[E(x_1)-E(x_2)][E(y_1)-E(y_2)]}{4}}{\sqrt{\operatorname{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\operatorname{cov}(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}} =$$
$$= \frac{\frac{\sqrt{\operatorname{var}(x_1)\operatorname{var}(y_2)}\cdot r(x_1,y_2)+\sqrt{\operatorname{var}(x_2)\operatorname{var}(y_1)}\cdot r(x_2,y_1)}{2} - \frac{[E(x_1)-E(x_2)]\cdot[E(y_1)-E(y_2)]}{4}}{\sqrt{\operatorname{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\operatorname{cov}(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}} .$$

Itt azonnal látható, hogy ha

$$\operatorname{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2 < 0$$

és/vagy

$$\operatorname{cov}(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2 < 0 ,$$

akkor ez a fajta korreláció nem állítható elő. Ez az eredmény arra utal, hogy a páros szintű korreláció inkább a párt alkotó személyek közötti keresztkorrelációval mutat hasonlóságot. A kifejezésünk számlálójában található kovarianciát elemezve azonnal látható, hogy a „helyes" korreláció ekkor – a már korábban meghatározott – keresztkorreláció $r(X, Y')$. A közelítés, azaz annak a feltételezése, hogy a diád tagjai hasonlóan válaszolnak, szintén erre utal, ugyanis ekkor a kovariancia közel varianciává válik a várható értékek és a szórások közel egyezése miatt.

Foglaljuk össze a javasolt korrelációs fogalmakat, és azoknak az alapadatainkkal való kapcsolatát. Ezt az *5. táblázat* mutatja be.

A korreláció neve	$(X, X'), (Y, Y')$ kettős adatbevitel $(x_1, x_2), (y_1, y_2)$ alapadatok
Csoporton belüli korreláció	$r(X, X') = \frac{\mathrm{cov}(X,X')}{\mathrm{var}(X)} = \frac{\mathrm{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}{\frac{\mathrm{var}(x_1)+\mathrm{var}(x_2)}{2}+\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}$
A válaszadó belső korrelációja	$r(X, Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\cdot\sqrt{\mathrm{var}(Y)}} = \frac{\frac{\mathrm{cov}(x_1,y_1)+\mathrm{cov}(x_2,y_2)}{2}+\frac{[E(x_1)-E(x_2)][E(y_1)-E(y_2)]}{4}}{\sqrt{\frac{\mathrm{var}(x_1)+\mathrm{var}(x_2)}{2}+\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\frac{\mathrm{var}(y_1)+\mathrm{var}(y_2)}{2}+\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}}$
A párt alkotó személyek közötti keresztkorreláció	$r(X, Y') = \frac{\mathrm{cov}(X,Y')}{\sqrt{\mathrm{var}(X)}\cdot\sqrt{\mathrm{var}(Y)}} = \frac{\frac{\mathrm{cov}(x_1,y_2)+\mathrm{cov}(x_2,y_1)}{2}-\frac{[E(x_1)-E(x_2)][E(y_1)-E(y_2)]}{4}}{\sqrt{\frac{\mathrm{var}(x_1)+\mathrm{var}(x_2)}{2}+\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\frac{\mathrm{var}(y_1)+\mathrm{var}(y_2)}{2}+\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}}$
Diád szintű korreláció	$r_m(X, X', Y, Y') = \frac{r(X,Y)+r(X,Y')}{\sqrt{1+r(X,X')}\cdot\sqrt{1+r(Y,Y')}} = r(x_1 + x_2, y_1 + y_2)$
Egyéni szintű korreláció	$r_i(X, X', Y, Y') = \frac{r(X,Y)-r(X,Y')}{\sqrt{1-r(X,X')}\cdot\sqrt{1-r(Y,Y')}} = \frac{\mathrm{cov}(x_1-x_2,y_1-y_2)+[E(x_1-E(x_2)]\cdot[E(y_1-E(y_2)]}{\sqrt{\mathrm{var}(x_1-x_2)+[E(x_1)-E(x_2)]^2}\cdot\sqrt{\mathrm{var}(y_1-y_2)+[E(y_1)-E(y_2)]^2}}$
Páros szintű korreláció	$r_d(X, X', Y, Y') = \frac{r(X,Y')}{\sqrt{r(X,X')}\cdot\sqrt{r(Y,Y')}} = \frac{\frac{\mathrm{cov}(x_1,y_2)+\mathrm{cov}(x_2,y_1)}{2}-\frac{[E(x_1)-E(x_2)][E(y_1)-E(y_2)]}{4}}{\sqrt{\mathrm{cov}(x_1,x_2)-\left(\frac{E(x_1)-E(x_2)}{2}\right)^2}\cdot\sqrt{\mathrm{cov}(y_1,y_2)-\left(\frac{E(y_1)-E(y_2)}{2}\right)^2}}$

5. táblázat. A korrelációk és meghatározásuk a páros adatbevitel és az alapadatok segítségével

Az előbbi korrelációkat lokálisan is közelítettük, foglaljuk most össze ezeket az eredményeinket is. Ezt a *6. táblázatban* mutatjuk be. Ezzel a korrelációs vizsgálatokat befejeztem.

A korreláció neve	Közelítések
Csoporton belüli korreláció	$r(X, X') \sim r(x_1, x_2)$
A válaszadó belső korrelációja	$r(X, Y) \sim \frac{1}{2}\left[r(x_1, y_1) + r(x_2, y_2)\right]$
A párt alkotó személyek közötti keresztkorreláció	$r(X, Y') \sim \frac{1}{2}\left[r(x_1, y_2) + r(x_2, y_1)\right]$
Diád szintű korreláció	$r_m(X, X', Y, Y') = r(x_1 + x_2, y_1 + y_2)$
Egyéni szintű korreláció	$r_i(X, X', Y, Y') = r(x_1 - x_2, y_1 - y_2)$
Páros szintű korreláció	$r_d(X, X', Y, Y') \sim \frac{1}{2}\left[r(x_1, y_2) + r(x_2, y_1)\right]$

6. táblázat. A páros adatbevitelű korrelációk és közelítése az alapadatok segítségével

5 Regressziószámítás diadikus adatokkal: ICC és APIM modell

A lineáris kapcsolatok elemzése után áttérek az ok-okozati tényezők vizsgálatára. Ebben az esetben azt vizsgálom, hogy a függetlennek választott változók milyen hatással vannak a függőnek választott változókra. A klasszikus statisztikában a független változók megválasztása egyszerűbbnek tűnik a diadikus adatelemzéssel szemben. A diadikus adatelemzés során ugyanis figyelembe kell venni az egyéni és páros hatásokat is. A diadikus adatelemzés regresszió vizsgálata ezért már egy független és egy függő változó esetén is több tényező figyelembevételével történhet meg. Ezek a tényezők a következők:

- cselekvő hatás (actor effect),
- partner hatás (partner effect) és
- kölcsönös hatás (mutual effect).

Ezen tényezők számának ismeretében építhetőek fel a diadikus adatelemzés regressziós modelljei. Ezen modellekből kettőt ismertetek (Gonzalez, 2010, Gelei-Dobos-Sugár, 2014). Az első modell, amelyet az irodalom ICC (Intraclass Correlation Coefficient) modellként ismer, csak a cselekvő és partner hatást építi be a regressziós modellbe. A másik modell mindhárom, azaz cselekvő, partner és kölcsönös hatást is kezeli. E modelltípust az irodalom Actor-Partner Interdependence Model-nek, röviden APIM modellnek nevezi. Az alábbiakban röviden ismertetem a modelleket. Először az ICC modellt vizsgálom meg kritikusabban. Nem az a célom, hogy a regresszió paramétereit előállítsam, hanem annak az elemzése, hogy a javasolt lineáris modell valóban teljesen leírja-e a diadikus változók közötti kapcsolatokat.

Az ICC modell tehát csak a párok egymásra hatását képezi le. A modell matematikai formája:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X' + \varepsilon \,,$$

ahol az X és X' a kettős adatbevitel során nyert független változók, Y a függő változó, míg ε a hiba. A β_0, β_1 és β_2 értékek a regressziós együtthatók.

Átírható a modell az alapadatokra. Ennek a formája ekkor az alábbi módon alakul:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \beta_2 \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} ,$$

ahol az 1 az összegző vektor, azaz olyan n-elemű vektor, amelynek minden eleme egy, valamint ε_1 és ε_2 a becslés hibája. (Eltekintek most attól, hogy legkisebb négyzetek módszerével, vagy maximum likelihood stb. módszerrel végezzük a paramétereink becslését.) Bontsuk szét elemeire ezt a becslést:

$$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon_1 ,$$
$$y_2 = \beta_0 \cdot 1 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1 + \varepsilon_2 .$$

Már ebből a felírásból is világos, hogy a második egyenletben ugyanazok a regressziós együtthatók szerepelnek, mint az első egyenletben, ezért a kettős adatbevitellel nyert becslés csak pontatlanul becsli a pár második tagjának válaszait az y_2 adatainkra.

A fentiek miatt pontosabb becslést ad az alábbi javaslat:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \varepsilon_{11} ,$$
$$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \varepsilon_{21} ,$$

vagyis a korábbi három együttható helyett most hatot kell becsülni, igaz, hogy ebben az esetben a két becslőfüggvény két független egyenletre esik szét, azokat nem köti össze a közös együttható. Az ε_{11} és ε_{21} értékek a becslés hibái.

Könnyen belátható, hogy a javasolt becslés kisebb hibát is ad és a paraméterek is pontosabban leírják a lineáris összefüggéseket; feltéve, hogy mindkét modellt azonos módszerrel becsüljük. Feltételezzük most, hogy a modellek együtthatói optimálisak, azaz $(\beta_0, \beta_1, \beta_2)$, $(\beta_{01}, \beta_{11}, \beta_{21})$ és $(\beta_{02}, \beta_{12}, \beta_{22})$ optimalizálják a becslőfüggvényeiket. Legyenek ugyanis a második modell becslőfüggvényei $f_1(\beta_{01}, \beta_{11}, \beta_{21})$ és $f_2(\beta_{02}, \beta_{12}, \beta_{22})$, ahonnan azonnal látjuk, hogy az első modell becslőfüggvénye ugyanazzal a módszerrel nem lesz más, mint

$$f_1(\beta_0, \beta_1, \beta_2) + f_2(\beta_0, \beta_2, \beta_1) .$$

Mivel $f_1(\beta_{01}, \beta_{11}, \beta_{21})$ és $f_2(\beta_{02}, \beta_{12}, \beta_{22})$ optimális együtthatókat adnak, ezért teljesül

$$f_1(\beta_{01}, \beta_{11}, \beta_{21}) \le f_1(\beta_0, \beta_1, \beta_2) \quad \text{és} \quad f_2(\beta_{02}, \beta_{12}, \beta_{22}) \le f_2(\beta_0, \beta_2, \beta_1) ,$$

vagyis

$$f_1(\beta_{01}, \beta_{11}, \beta_{21}) + f_2(\beta_{02}, \beta_{12}, \beta_{22}) \le f_1(\beta_0, \beta_1, \beta_2) + f_2(\beta_0, \beta_2, \beta_1) = f(\beta_0, \beta_1, \beta_2) .$$

Ez azt is jelenti, hogy az alapadatokra átírt lineáris modellünk pontosabb becslést nyújt. Most áttérek az APIM modell vizsgálatára!

Az APIM modell csak kissé különbözik az ICC modelltől. Az APIM modell nem csak a párok egymásra hatását képezi le, de figyelembe veszi a párok kölcsönös egymásra hatását is. A modell matematikai formája tehát

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X' + \beta_3 \cdot X \cdot X' + \varepsilon \ ,$$

ahol a β_0, β_1 és β_2 értékeket teljesen hasonlóan definiálható, mint az ICC modellben, és ε a becslés hibája. Az egyedüli eltérés az, hogy a kölcsönös hatást is beépítjük a modellbe a $\beta_3 \cdot X \cdot X'$ kifejezés szerepeltetésével. Az $X \cdot X'$ szorzat esetünkben új változó, a pár mindkét szereplőjének a kölcsönös, együttesen kifejtett hatását mutatja a cselekvő Y változójára.

Ekkor is átírhatjuk az alapadatokra a modellt:

$$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_1 \ ,$$

$$y_2 = \beta_0 \cdot 1 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_2 \ .$$

A $\langle x_1 \cdot x_2 \rangle$ kifejezés azt a vektort jelöli, amely az x_1 és x_2 vektorok egyes elemei szerint szorozza össze az elemeket.

Ekkor a javasolt új függvényeink a következők lehetnek:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_{11} \ ,$$

$$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_{21} \ .$$

Az ICC modellre tett megfontolások itt is könnyen megtehetők, vagyis az utóbbi becslési javaslat pontosabb eredményre vezet, és ezzel jobban árnyalja az egyes (diadikus) változók közötti kapcsolatot.

6 Összegzés

Dolgozatban összefoglaltam a diadikus adatelemzésben eddig paradigmának tekintett kettős adatbevitelt és annak statisztikai következményeit. Beláttam, hogy felcserélhető esetben valamilyen konszenzust kell keresni az adatok kezelésében, mert a szerepek szimmetriája miatt a vizsgálható táblázatok száma a felvett adatok exponenciális függvénye. Javaslatom az, hogy olyan transzformációt hajtsunk végre az adatokon, ami ezt a szimmetriát megszünteti pl. az adatok összeadásával, és/vagy azok különbségének abszolút értékével, és a két adat aszimmetrikussá tételével, mint a megkülönböztethető esetben.

Rámutattam arra, hogy a diadikus adatelemzés homogénitásvizsgálata alapvetően az alapadatokból is végrehajtható, nincs szükség a kettős adatbevitelre.

Sikerült a diadikus adatelemzésben eddig alkalmazott korrelációs fogalmakat egyrészt tisztázni, másrészt azt valóban Pearson-féle korrelációs együtthatóra átalakítani. Azt is megmutattam, hogy a korrelációkat az alapadatokra is ki lehet számítani, nincs szükség azt a kettős adatbevitellel megnehezíteni.

Végül, beláttam azt is, hogy a javasolt ICC és APIM modellek is rontják a becslést a kettős adatbevitellel. Pontosabb becslést lehet elérni az alapadatokra elvégzett regressziókkal. További kutatásokkal azt kell tisztázni, hogy valós adatokon milyen eredményt adnak a javasolt változtatások.

Irodalom

1. Gelei, A. – Dobos, I. – Sugár, A. (2014): Bevezetés a diadikus adatelemzésbe – elmélet és alkalmazás, *Statisztikai Szemle*, 92. évf. 5. szám, 417–446.
2. Gelei, A. – Sugár, A. (2016): Diadikus jelenségek kutatási kihívása – a diadikus adatelemzés és a hagyományos statisztikai megoldások összehasonlítása, *Statisztikai Szemle*, 94. évf., 10. szám, 977–1003.
3. Gelei, A. – Dobos, I. (2016): Bizalom az üzleti kapcsolatokban, *Közgazdasági Szemle*, LXIII. évf., 3. szám, 330–349.
4. Gonzalez, R., – Griffin, D. (1999): The correlational analysis of dyad-level data in the distinguishable case. *Personal Relationships*, 6(4), 449–469.
5. Gonzalez, R. – Griffin, D. (2000): On the Statistics of Interdependence: Treating Dyadic Data with Respect; in: Ickes, W. – Duck, S. (2000) (ed.) *The Social Psychology of Personal Relationships;* John Wiley and Sons, Ltd., 181–213.
6. Griffin, D. – Gonzalez, R. (1995): Correlational Analysis of Dyad-Level Data in the Exchangeable Case, *Psychological Bulletin* 1995. Vol. 118, No. 3, 430-439.
7. Kenny, D. A. – Kashy, D. A. – Cook, W. L. (2006): *Dyadic data Analysis*; The Guilford Press, New York-London.
8. Vincze I. – Varbanova, M. (1993): *Nem paraméteres matematikai statisztika – Elmélet és alkalmazások*; Akadémiai Kiadó, Budapest.

A CRITICAL INVESTIGATION OF METHODS IN DYADIC DATA ANALYSIS: THE DOUBLE ENTRY AND EXCHANGEABLE CASES

The aim of the paper is to examine the mathematical statistics foundations of Dyadic Data Analysis. In a former paper it was investigated, whether the Dyadic Data Analysis significantly contributes to traditional statistical analysis and provides surplus in understanding statistical phenomenon, or not. Further, it is tried to correct some of the mathematical structure of Dyadic Data Analysis.

Keywords: Mathematical Statistics Correlational Analysis, Dyadic Data Analysis