A parciális korrelációs együttható értelmezési problémái a többdimenziós normalitás feltételének sérülése esetén

Vargha András
egyetemi tanár, az MTA doktora, Károli Gáspár Református Egyetem Pszichológiai Intézete, ELTE Pszichológiai Intézete
E-mail: vargha_andras@kre.hu

A parciális korrelációt sokan és gyakran használják olyan esetekben, amikor két kvantitatív változó kapcsolatából ki akarják szűrni egy vagy több kvantitatív változó lineáris hatását. A parciális korreláció értékét szokásosan úgy értelmezik, hogy milyen lenne a vizsgált két változó kapcsolata akkor, ha a kiszűrt változókat állandó szinten tartanánk (feltételes korreláció).

A szerző arra hívja fel a figyelmet – elméleti megfontolások és konkrét példák segítségével –, hogy ha nem teljesül a parciális korreláció linearitásra vonatkozó alkalmazási feltétele (amit a többdimenziós normalitás biztosít), akkor az említett értelmezés nem tekinthető helytállónak, vagyis ilyenkor megnő a téves következtetés esélye a parciális korrelációs együtthatóval kapcsolatban. Olyan szélsőséges eset is előfordulhat, hogy a parciális korreláció erős pozitív kapcsolatot jelez, miközben a feltételes korreláció –1-hez közeli negatív érték. E probléma kezelésének egyik lehetséges egyszerű módja, hogy nemlineáris összefüggések fellépte esetén a kiszűrendő változó alkalmas függvényét (például négyzetét) is kiszűrjük.

A tanulmány kitér arra a speciális esetre is, amikor két változó korrelációját egy harmadik változó értékskálájának korlátozása mellett számítjuk ki.

TÁRGYSZÓ:
Korrelációszámítás.

Empirikus adatok elemzésekor néha meglepő korrelációkkal találkozhatunk. Ha kiszámítjuk a korrelációt a budapesti taxisok napi jövedelme és a Duna napi vízállása között egy teljes év viszonylatában, a kapott magas pozitív érték alapján bizonyára eltöprengünk azon, hogy milyen fura kapcsolat van a két változó között. Egy kis fejtörés után könnyen juthatunk arra a következtetésre, hogy a magas korreláció fellépte bizonyos közvetítő vagy háttérváltozók hatásának köszönhető. Ilyen háttérváltozó lehet például a napi csapadékmennyiség. Az esős napokon ugyanis egyaránt megnő a Duna vízállása és a taxi igénybevételének a valószínűsége, ami azonban nem jelenti azt, hogy e két tényező között bármilyen közvetlen kapcsolat lenne.

A leírt szituációt általánosítva kérdezhetjük a statisztikustól: mit tegyünk, ha egy *X* és egy *Y* változó közötti olyan kapcsolat érdekel bennünket, ami akkor állna fenn, ha nem hagynánk, hogy egy *X*-szel és *Y*-nal egyaránt korreláló *Z* változó kifejtse a hatását? Erre a kérdésre találták ki a statisztikában a parciális korrelációs együtthatót, melynek egyik ismert képlete a páronkénti korrelációk segítségével írható fel a következőképpen (*Pedhazur* [1982] 103. old., *Vincze* [1968] 256–257. old.):

$$\rho_{XY.Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}. \qquad /1/$$

Ebben az /1/ formulában az elméleti parciális korrelációt fejezzük ki a páronkénti korrelációk segítségével, de ugyanez az összefüggés érvényes az empirikus parciális korreláció és a páronkénti korrelációk között is.

A parciális korrelációs együttható érvényes alkalmazásával kapcsolatban a következő feltételeket szokták megfogalmazni (lásd például *Garson* [2009]):

– kvantitatív (legalább intervallumskálájú) *X*, *Y*, *Z* változók;
– csak lineáris típusú összefüggések léteznek az *X*, *Y*, *Z* változók között;
– *X* és *Y* között ugyanolyan jellegű és szintű kapcsolat van a *Z* változó teljes értéktartományában.

Megjegyzendő, hogy ha a vizsgált változók együttes eloszlása többdimenziós normális, akkor ezek a feltételek szükségképpen fennállnak (*Tabachnik–Fidell* [2001] 72. old.).

A parciális korreláció alkalmazása rendkívül népszerű az empirikus kutatásokban. Például a Web of Sciences cikkarchívumának keresőjében a „partial correlation” ki-

fejezéshez 2010. október 20-án 9589 olyan cikk került listázásra, amelyek mind 2000 és 2010 között jelentek meg.

Aggasztónak tűnik a parciális korreláció ilyen széles körű használata, ha figyelembe vesszük, hogy milyen szigorúak az érvényes alkalmazás előbbiekben megfogalmazott feltételei. Például a társadalomtudományokban a normális eloszlás inkább tekinthető ritka kivételnek, mint általános szabályosságnak (*Micceri* [1989], illetve *Vargha* [2003a]) és a változók közötti gyakori nemlineáris összefüggések (például az izgalmi szint és a mentális teljesítmény, a vérnyomás és a jó közérzet között stb.) is arra figyelmeztetnek, hogy a parciális korreláció mérlegelés nélküli, automatikus használata esetenként komoly bajok forrása lehet.

Ezen körülmények boncolgatása nem tűnik mindennaposnak a tudományos publikációkban. Például a „partial correlation interpretation" kifejezésre a Web of Sciences archívumából már csak 202 tétel jött elő, amelyek közül csupán 3 foglalkozott magának a parciális korrelációnak a jelentésével, értelmezésével.

Brillinger [2001] *John Tukey* álláspontját követve hangsúlyozza a keresztmetszeti adatokból számított korrelációs és parciális korrelációs együtthatók korlátait és helyettük az idősorelemzésből ismert koherencia, valamint parciális koherencia mutatók használatát javasolja. *Rae* és *Carretta* [2006] a mérési hiba hatását tekinti át a különböző statisztikai mutatók és próbák esetén. *Cramer* [2003] tanulmánya pedig azt boncolgatja, hogy a parciális korrelációs együttható előjele és nagysága miként függ a vizsgálatba bevont *X*, *Y*, *Z* változók páronkénti közönséges korrelációinak mintázatától. Cramer megemlíti például, hogy ha a *Z* változó ugyanolyan irányú, de szorosabb kapcsolatban van az *X*, *Y* változókkal, mint emezek egymással, akkor az $r_{xy.z}$ parciális korrelációs együttható mindig ellentétes előjelű lesz, mint az eredeti – nulladrendű – r_{xy} korrelációs együttható, ami alapjaiban érinti az *X* és az *Y* változó közötti kapcsolat értelmezését.

Kérdésfeltevésünk aktualitását erősíti, hogy az áttekintett tanulmányok egyike sem foglalkozik azzal, hogy az alkalmazási feltételek sérülése milyen hatással van a parciális korrelációs együttható jelentésére, értelmezésére. A jelen tanulmány célja kettős:

> *a*) elméleti levezetésekkel kimutatni, hogy az alkalmazási feltételek számottevő sérülése esetén nem érvényes a parciális korrelációs együttható hagyományos értelmezése;
>
> *b*) gyakorlati útmutatást adni ahhoz, hogy e feltételek sérülése esetén az ismertebb statisztikai programcsomagok (például SPSS és ROPstat) eszköztára segítségével hogyan lehet a szakmai célnak megfelelő statisztikai mutatót készíteni.

Tanulmányunk első részében a parciális korrelációs együttható matematikai definícióját és néhány elméleti vonását tekintjük át. Ezután matematikai levezetésekkel

kimutatjuk, hogy ha a *Z* változó nemlineáris módon (például kvadratikusan) hat *X*-re és *Y*-ra, akkor a parciális korreláció hogyan válhat téves következtetés forrásává. Végül tanulmányunk harmadik részében arra teszünk javaslatot, hogy az ismert korrelációs és regressziós technikák alkalmazásával a gyakorlatban miként kerülhetjük el a téves értelmezés csapdáját.

1. A parciális korrelációs együttható matematikai definíciója

Tegyük fel, hogy egy *X* és egy *Y* kvantitatív változó közötti lineáris kapcsolat foglalkoztat bennünket, melyet konkrét statisztikai elemzésekben szokásosan a Pearson-féle r_{xy} korrelációval mérünk (az ennek megfelelő elméleti korreláció szokásos jele: ρ_{XY}). Ha *X* és *Y* együttjárását elemezve felmerül, hogy egy harmadik, *Z*-vel jelölt kvantitatív változó, mely hat *X*-re és *Y*-ra, befolyásolhatja azok r_{xy}-nal mért korrelációját, akkor elgondolkodhatunk azon, mekkora r_{xy}-ban az a rész, amely *X* és *Y* közvetlen, *Z*-től nem függő együttjárásának az eredménye. Ennek a részleges, „parciális" kapcsolatnak a mérésére találták ki a matematikai statisztikában a parciális korrelációs együtthatót a következő gondolatmenet szerint.

1. Határozzuk meg a *Z* változó *X*-re vonatkozó lineáris regressziós becsléseként az *X* változónak azt a részét, amely lineárisan függ *Z*-től $\left(X_Z\right)$. Ekkor úgy vélhetjük, hogy ha *X*-ből elhagyjuk (kivonjuk) ezt a *Z*-től függő X_Z összetevőt, akkor ami marad, már nem függ *Z*-től, vagyis $X_{mar} = X - X_Z$ az *X* változónak az a része, amely nem függ lineárisan *Z*-től.

2. Hasonló logikával határozzuk meg *Y*-ban a *Z*-től lineárisan függő Y_Z összetevőt, s ennek segítségével a *Z*-től lineárisan nem függő $Y_{mar} = Y - Y_Z$ összetevőt.

3. Ezen *Z*-től lineárisan nem függő X_{mar}, Y_{mar} összetevők közötti Pearson-féle korrelációt nevezzük parciális korrelációnak:

$$r_{xy.z} = r\left(X_{mar}, Y_{mar}\right)$$

(*Pedhazur* [1982] 97–104. old.).

Ha ugyanezeket a lépéseket az elméleti regressziós modellben hajtjuk végre, akkor a $\rho_{XY.Z}$ elméleti parciális korrelációs együtthatóhoz jutunk (*Vargha* [2007a] 300–314. old.).

Az $r_{xy.z}$ $\left(\rho_{XY.Z}\right)$ parciális korrelációs együtthatót úgy szokták tekinteni, mint az r_{xy} (ρ_{XY}) korrelációnak azt a részét, amelyből a Z változó lineáris hatása ki van szűrve. Kiszámításának egyik egyszerű módja a tanulmányunk elején felírt /1/ formula alkalmazása, melyhez csupán az X, Y, Z változók között páronként kiszámított korrelációk szükségesek.

Ezen gondolatmenet általánosításával természetesen több kvantitatív változó hatását is ki lehet szűrni X és Y kapcsolatából, de ennek technikai részleteire itt most nem térünk ki. Ezzel kapcsolatban csak annyit jegyzünk meg, hogy több változó kiszűrése esetén a végső parciális korreláció nem függ a kiszűrések sorrendjétől, tehát például $r_{xy.zuv} = r_{xy.vuz} = r_{xy.zvu}$ stb.

A parciális korrelációs együttható értelmezésével kapcsolatban alapvetően fontos, hogy ha teljesül az X, Y, Z változókra a többdimenziós normális eloszlás feltétele, akkor az $r_{xy.z}$ parciális korrelációs együttható becslés lesz arra, hogy mekkora lenne az elméleti korreláció X és Y között, ha a Z változót bármely konkrét z pontban rögzítenénk:

$$r_{xy.z} \approx \rho\left(X, Y \middle| Z = z\right).$$

Alkalmazási feltételeinek teljesülése esetén tehát a parciális korrelációs együttható valóban azt mutatja (méri), hogy a Z változó fixálásakor (ezzel érjük el azt, hogy Z ne fejthesse ki hatását X-re és Y-ra) mekkora lesz a korreláció X és Y között. Ez utóbbi korrelációt feltételes korrelációnak nevezzük. Az X, Y, Z változók együttes eloszlásának többváltozós normalitása azt biztosítja, hogy egyrészt közöttük csak lineáris típusú összefüggések léphetnek fel (emiatt a Pearson-féle r teljesen adekvát mérőszáma a páronkénti kapcsolatoknak), másrészt az X és az Y közti összefüggés Z bármely rögzített értéke esetén ugyanakkora lesz.

Ha viszont a normalitási feltétel nem teljesül, a parciális korrelációs együttható nem feltétlenül jelzi azt, hogy mekkora a korreláció X és Y között, ha Z-t rögzítjük, vagyis állandó szinten tartjuk. A parciális korreláció és a feltételes korreláció tehát nem feltétlenül fog megegyezni, ami ilyen esetben megkérdőjelezi a parciális korreláció hagyományos értelmezésének a jogosságát. A következőkben ezt fogjuk elméleti levezetésekkel igazolni.

2. A parciális korrelációs együttható értelmezésének problémája nemlineáris összefüggések felléptekor

Jelen fejezetben mesterségesen konstruált változók felhasználásával, elméleti levezetéssel mutatjuk meg, hogy ha *X*, *Y* és *Z* között nemlineáris összefüggések vannak (ilyenkor a többváltozós normalitás feltétele szükségképpen sérül), akkor a parciális korreláció és a feltételes korreláció értéke akár óriási mértékben is különbözhet egymástól.

Legyen *U*, *V* és *Z* normális eloszlású, egymástól független változó! Az egyszerűség kedvéért legyenek standardizált alakban (0 átlaggal és 1 szórással). Definiáljuk ezek segítségével először az

$$X0 = aZ + cU \quad \text{és} \quad Y0 = aZ - cU + cV \qquad /2/$$

változót, ahol *a* és *c* tetszőleges pozitív szorzótényezők. $X0$ és $Y0$ kifejezése egyaránt tartalmaz egy *a* egység súlyú pozitív (aZ), valamint egy *c* súlyú, de ellentétes előjelű (cU, illetve $-cU$) közös komponenst. $Y0$-t kiegészíti még egy ugyancsak *c* súlyú független összetevő (cV) is.

Matematikailag igazolható (a bizonyítást lásd a Függelék F1. pontjában), hogy az *a*, *c* paraméterek segítségével a következőképpen írható fel az $X0$ és $Y0$ közötti elméleti korrelációs együttható:

$$\rho_{X0Y0} = \frac{a^2 - c^2}{\sqrt{a^2 + c^2}\sqrt{a^2 + 2c^2}}. \qquad /3/$$

$X0$ és $Y0$ kapcsolatának előjele és szorossága *a* és *c* viszonyától függ. Ha $a > c$, akkor az $X0$ és $Y0$ közötti elméleti korreláció pozitív, ha pedig $a < c$, akkor ez a korreláció negatív lesz. Rögzített *a* érték mellett a kapcsolat szorossága *c* növelésével gyengébb, csökkentésével pedig erősebb lesz. Például $a = 5$ és $c = 1$ érték mellett $\rho(X0, Y0) = 0{,}906$, $a = 5$ és $c = 3$ esetén pedig $\rho(X0, Y0) = 0{,}418$. (Lásd az 1. táblázatot.)

Mivel $X0$ és $Y0$ az egymástól páronként független *U*, *V*, *Z* standard normális eloszlású változók lineáris kombinációja, együttes eloszlásuk igazolhatóan többdimenziós normális eloszlású lesz (*Rényi* [1968]), ami miatt a $\rho_{X0Y0.Z}$ elméleti parciális korrelációnak meg kell egyeznie a $\rho(X, Y|Z = z)$ feltételes korrelációval bármely *z* szám esetén. Ez a közös $\rho_{X0Y0|Z}$ feltételes korrelációs érték a következőképpen ha-

tározható meg. Mivel Z rögzítése esetén az $X0$-ban és $Y0$-ban egyaránt megtalálható aZ összetevő konstans, a korreláció csak a maradék részek (cU és $cV - cU$) viszonyától függ. Emiatt

$$\rho_{X0Y0|Z} = \rho\left(cU, cV - cU\right) = \frac{-Cov\left(cU, cU\right)}{\left[D\left(cU\right)D\left(cV - cU\right)\right]} = \frac{-1}{\sqrt{2}}, \qquad /4/$$

ami nem függ a és c értékétől, és három tizedesre kerekítve –0,707-tel egyenlő.[1]

De mi történik akkor, ha $X0$-hoz és $Y0$-hoz hozzáadunk egy Z-től nemlineárisan függő összetevőt? Például

$$X = X0 + bZ^2 \quad \text{és} \quad Y = Y0 + bZ^2 \qquad /5/$$

esetén X és Y kvadratikusan (parabolikusan) függ Z-től. Emiatt X, Y és Z együttes eloszlása nem lehet normális, továbbá az sem garantált, hogy az X és Y közti, Z hatását kiszűrő parciális korrelációs együttható $\left(\rho_{XY.Z}\right)$ meg fog egyezni a Z rögzítése mellett kiszámított X és Y közti feltételes korrelációval $\left(\rho_{XY|Z}\right)$. Ennek kimutatásához először is meghatároztuk X és Y között a korrelációt, mely a Függelék F2. levezetése alapján a következőképpen írható fel a, b és c függvényében:

$$\rho_{XY} = \frac{a^2 + 2b^2 - c^2}{\sqrt{a^2 + 2b^2 + c^2}\sqrt{a^2 + 2b^2 + 2c^2}}. \qquad /6/$$

Ezután a ρ_{XZ}, ρ_{YZ} korrelációkat is meghatározva (lásd Függelék F1-et), az /1/ formulába való behelyettesítéssel és egyszerű algebrai átalakításokkal kaphatjuk meg a $\rho_{XY.Z}$ parciális korrelációs együttható képletét:

$$\rho_{XY.Z} = \frac{2b^2 - c^2}{\sqrt{2b^2 + c^2}\sqrt{2b^2 + 2c^2}}. \qquad /7/$$

Végül a $\rho_{XY|Z}$ feltételes korreláció meghatározásához azt vegyük figyelembe, hogy Z rögzítése esetén X és Y között pontosan ugyanolyan lesz a korreláció, mint $X0$ és $Y0$ között, vagyis –0,707 (vö. /4/ és /5/).

[1] A /4/ formulában *Cov* a kovariancia, *D* pedig a szórás operátorát jelöli.

A feltételes korreláció tehát láthatóan nem függ *a*, *b* és *c* értékétől, de a parciális korreláció igen, aminek a konkrét szemléltetésére a következő három paraméterkombinációra kiszámítottuk $\rho_{XY.Z}$ értékét. (Lásd az 1. táblázatot.)

$$a)\ a=5,\ c=1,\ b=3;$$

$$b)\ a=5,\ c=2,\ b=2;$$

$$c)\ a=5,\ c=3,\ b=1.$$

1. táblázat

Az X0 és az Y0, illetve az X és az Y változó közötti közönséges (ρ_{XY}) és parciális ($\rho_{XY.Z}$) korrelációk

X változó	*Y* változó	Közönséges korreláció (ρ_{XY})	Parciális korreláció $(\rho_{XY.Z})$	Feltételes korreláció $(\rho_{XY\|Z})$
$X0(c=1)$	$Y0(c=1)$	0,906	–0,707	–0,707
$X0(c=2)$	$Y0(c=2)$	0,679	–0,707	–0,707
$X0(c=3)$	$Y0(c=3)$	0,418	–0,707	–0,707
$X(b=3,\ c=1)$	$Y(b=3,\ c=1)$	0,944	0,872	–0,707
$X(b=2,\ c=2)$	$Y(b=2,\ c=2)$	0,745	0,289	–0,707
$X(b=1,\ c=3)$	$Y(b=1,\ c=3)$	0,447	–0,472	–0,707

Az 1. táblázat alapján levonható következtetések:

– $X0$ és $Y0$, illetve *X* és *Y* között a közönséges ρ_{XY} korrelációs együttható a *c* paraméter értékének növelésével csökken, ahogy ezt már korábban is megállapítottuk (vö. /3/ formula), ugyanis *c* szorzótényezője az $X0$-ban és $Y0$-ban, illetve *X*-ben és *Y*-ban ellentétes együtthatójú *U* összetevőnek, valamint az $Y0$, illetve *Y* egyediségét képviselő *V* összetevőnek (vö. /2/ egyenletek).

– Ha a *Z* változótól csak lineárisan függő $X0$ és $Y0$ változó korrelációjából kiszűrjük a *Z* változót, a kapott $\rho_{XY.Z}$ értékek pontosan megegyeznek a feltételes korreláció –0,707-es értékével. Ebben az esetben tehát a parciális korrelációs együttható valóban azt mutatja, hogy milyen a kapcsolat $X0$ és $Y0$ között, ha a *Z* változó értékét állandó szinten tartjuk.

– Ugyanez a szabályszerűség azonban nem figyelhető meg abban az esetben, amikor *X*-ben és *Y*-ban megjelenik a *Z* változó kvadratikus

hatása. A probléma természetesen ott a legsúlyosabb, ahol a kvadratikus komponens *b* szorzótényezője a legnagyobb $(b=3)$. Itt a parciális korrelációs együttható értéke 0,872, ami igen erős közvetlen pozitív kapcsolatot jelez *X* és *Y* között *Z* kiszűrése után, miközben a *Z*-re vonatkozó feltételes korreláció –0,707-es értéke jól mutatja, hogy *Z* rögzítésekor *X* és *Y* erős negatív kapcsolatban van egymással. Bár kisebb mértékű, de még mindig erősen félrevezető információt nyújt a parciális korrelációs együttható $b=2$ érték mellett ($\rho_{XY.Z} \approx 0,3$, miközben $\rho_{XY|Z} \approx 0,7$). A kvadratikus komponens legkisebb szorzótényezője $b=1$ esetén is 0,2-nél nagyobb eltérés van a parciális korreláció és a feltételes korreláció értéke között.

Mindezek az eredmények egyértelműen bizonyítják, hogy a parciális korrelációs együttható értelmezésekor minden esetben mérlegelni kell, hogy alkalmazási feltételei teljesülnek-e, különben könnyen juthatunk téves következtetésekre.

Kvantitatív változók korrelációs elemzései során gyakori, hogy két változó (*X* és *Y*) kapcsolatát egy harmadik (*Z*) változó értéktartományának bizonyos szűkebb övezetében vizsgáljuk. Például szakmailag érdekes lehet, hogy milyen kapcsolatban van az öngyilkosságban elhunytak és a bejelentett munkanélküliek száma 1998 és 2002, vagy 2002 és 2006 között, illetve 2006 után. Ha az ilyen övezetek szélességét a 0-hoz közelítjük, az *X* és *az Y* közötti korreláció a feltételes korrelációt adja meg *Z* adott értéke – a felső vagy az alsó végpont rögzítése – mellett.

Az ilyen típusú kérdések tisztázására a feltételes korreláció fogalmát általánosítjuk. Kiszámításához a feltételes várható érték formuláit vesszük alapul (*Vincze* [1968]), képletét normális eloszlású változók esetén a következő formulákkal adhatjuk meg.

A /2/ egyenletekkel megadott $X0$ és $Y0$ változó közötti korreláció a standard normális eloszlású *Z* változó tetszőleges $(Z<z)$ alakú résztartománya esetén:

$$\rho(X0,Y0 \mid Z<z)=\frac{a^2V_Z(z)-c^2}{\sqrt{a^2V_Z(z)+c^2}\sqrt{a^2V_Z(z)+2c^2}}, \qquad /8/$$

ahol $V_Z(z)$ a *Z* változó varianciáját jelöli a $(Z<z)$ résztartományon, mely a következőképpen határozható meg:

$$V_Z(z)=Var(Z|Z<z)=1-z\frac{f(z)}{F(z)}-\left(\frac{f(z)}{F(z)}\right)^2. \qquad /9/$$

Ebben a formulában $f(z)$ és $F(z)$ a standard normális eloszlás sűrűség-, illetve eloszlásfüggvényének értéke a *z* helyen.

Hasonlóképpen az $X0$ és $Y0$ változó közötti korreláció a standard normális eloszlású *Z* változó tetszőleges $(z_1 \le Z \le z_2)$ alakú résztartománya esetén így számítható ki:

$$\rho(X0, Y0 \mid z_1 \le Z \le z_2) = \frac{a^2 V_Z(z_1, z_2) - c^2}{\sqrt{a^2 V_Z(z_1, z_2) + c^2}\sqrt{a^2 V_Z(z_1, z_2) + 2c^2}}, \qquad /10/$$

ahol $V_Z(z_1, z_2)$ a *Z* változó varianciáját jelöli a $(z_1 \le Z \le z_2)$ résztartományon, mely a következőképpen határozható meg:

$$V_Z(z_1, z_2) = Var(Z \mid z_1 \le Z \le z_2) = 1 - \frac{z_2 f(z_2) - z_1 f(z_1)}{F(z_2) - F(z_1)} - \left(\frac{f(z_2) - f(z_1)}{F(z_2) - F(z_1)}\right)^2. \qquad /11/$$

Megjegyezzük, hogy $z = -\infty$, illetve $z = \infty$ esetén a $zf(z)$ szorzat 0-val egyenlő.

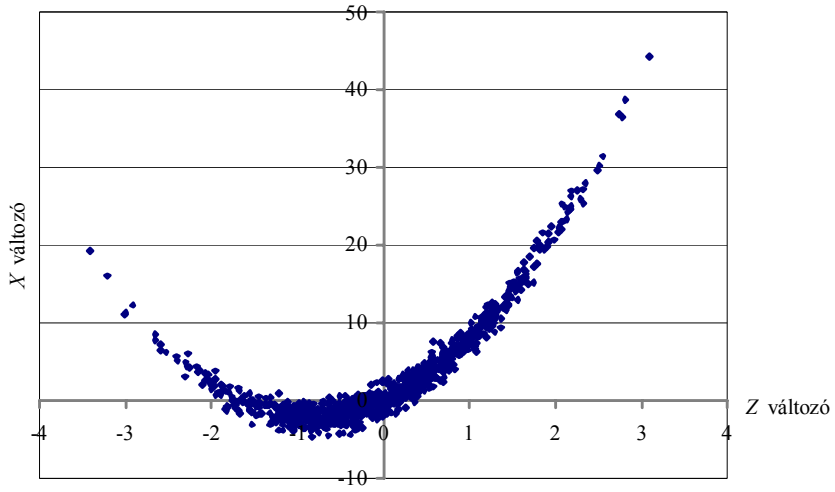
Az összefüggések részletes levezetését nem ismertetjük, de a bizonyítások logikáját és egyes lépéseket szemléltetésképpen a Függelék F3-ban bemutatjuk. Ezen formulák segítségével egyszerűen kiszámíthatók a feltételes korrelációk az $X0$ és $Y0$ változók tetszőleges lineáris és egyszerűbb nemlineáris (vö. /5/) transzformáltjaira is.

3. Hogyan kerülhetjük el a téves következtetések csapdáját?

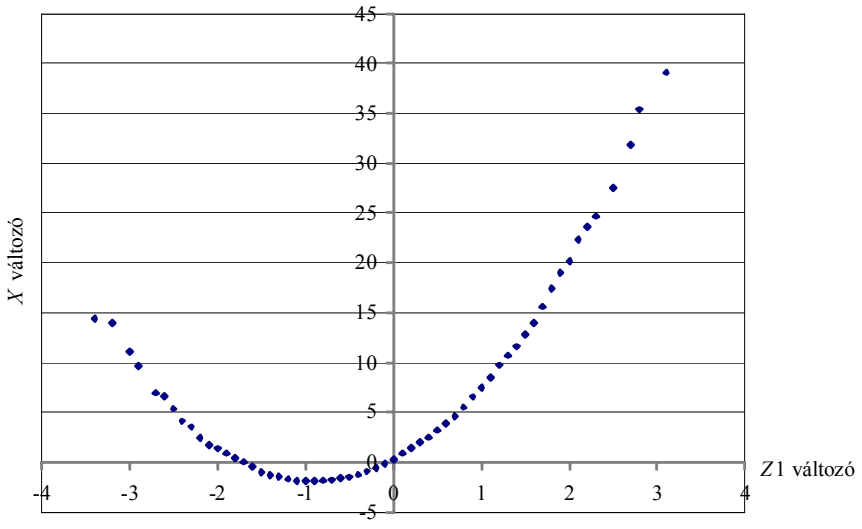
A parciális korrelációs együttható értelmezése akkor válik problematikussá, ha értéke nem egyezik meg a feltételes korrelációéval. Ez utóbbi pedig akkor következhet be legnagyobb eséllyel, ha a *Z* változó nemlineáris összefüggésben van *X*-szel és/vagy *Y*-nal. Hogy lehet a nemlineáris összefüggéseket felderíteni? Nyilván nincs értelme mindig, minden esetben nemlineáris kapcsolatok után kutatni, különösen sok változó vizsgálata esetén, mert az nagyon bonyolítaná a statisztikai adatfeldolgozást. Ugyanakkor a viszonylag kevés változóval operáló vizsgálatokban vagy olyan esetekben, amikor szakmai érvek szólnak a nemlineáris kapcsolatok lehetősége mellett, a nemlineáris összefüggések felderítése alapvetően fontos feladat. A továbbiakban bemutatunk néhány elemzési módszert, amellyel ezt megtehetjük.

Két változó összefüggésének milyenségéről sok esetben jó képet nyújt azok egyszerű pontdiagramja. Például az X változó Z-től való nemlineáris függése, $b = 3$ és $c = 1$ értéke mellett (1000 véletlen megfigyelés alapján) az 1. ábra pontdiagramján szépen kirajzolódik.

1. ábra. A Z és az X változó kétváltozós pontdiagramja b = 3 és c = 1 esetén



2. ábra. Az X változó simított nemlineáris regressziós becslése
és az egy tizedesre kerekített Z változó (Z1) kétváltozós pontdiagramja b = 3 és c = 1 esetén



Ha a változók között a kapcsolat nem olyan erős, mint az 1. ábrán látható esetben, a diagram pontjai annyira szóródhatnak, hogy nehézkes az összefüggés kiolvasása az ábráról. Ilyen esetben hasznos lehet a két változó között egy simított nemlineáris regressziós elemzést végezni (mozgó átlagos módszerrel), mely képes a véletlen ingadozások jelentős részének kiszűrésére és a kapcsolat fő tendenciáinak a kidomborítására. Ilyen elemzésre képes például a ROPstat „Korreláció, egyszerű regresszió" modulja, ha itt a „Lokális korreláció, nemlineáris regresszió" opcióra klikkelünk (www.ropstat.com).

Például az X változó $(b = 3 \text{ és } c = 1)$ Z-től való nemlineáris függésének felderítésére először egyszerűsítettük a Z változót értékeinek egytizedesre való kerekítésével ($Z1$), majd elvégeztük a simított nemlineáris regressziós elemzést a ROPstatban. A kapott regressziós becslés függését a $Z1$ változótól az Excelben elkészített pontdiagram jól szemlélteti. (Lásd a 2. ábrát.) Az 1. és a 2. ábra összehasonlítása mutatja, hogy a véletlen ingadozások kiszűrése milyen jól kiemeli a valódi összefüggést a két változó között.

A simított nemlineáris regresszió a mozgó átlag módszerével úgy szűri ki a véletlen ingadozások jelentős részét, hogy a független változó (itt $Z1$) minden z értéke esetén a z-hez tartozó regressziós becslést a z körüli szomszéd értékekhez tartozó függő változó (jelen esetben X) értékeinek átlagaként határozza meg. A programban beállítható, hogy a z körüli „szomszédsági övezet" mennyire legyen szűk, illetve tág. A program kiszámítja a simított nemlineáris regresszió által magyarázott varianciahányadot (nemlineáris determinációs együtthatót) is, mely a 2. ábrán bemutatott esetben 0,966, vagyis közel 100 százalékos lett.

Ha van konkrét elképzelésünk a változók nemlineáris függésének a típusáról, alkalmazhatjuk az SPSS nemlineáris regressziós modulját (Analysis/Regression/Nonlinear), konkrét elképzelés híján pedig a program görbeillesztő modulját (Analysis/Regression/Curve Estimation). Ez utóbbiban egyidejűleg több lehetséges függési típus (lineáris, kvadratikus, harmadfokú, logaritmikus, exponenciális stb.) is megvizsgálható és összevethető egymással.

Amennyiben sikerült meghatározni a nemlineáris függés jellegét, akkor nincs más dolgunk, minthogy a parciális korreláció számítása során a szűrendő változó adott függvényét is kiszűrjük. A 2. ábrán bemutatott esetben a függés egyértelműen parabolikus jellegű, ami jelzi számunkra, hogy ha X és Y kapcsolatából Z hatását teljesen ki akarjuk szűrni, akkor Z mellett a Z^2 változót is ki kell szűrnünk. Elvégezve ezt az elemzést a ROPstatban, a b és c paraméterértékek mindhárom kombinációja esetén –0,713-at kaptunk, ami csak igen kis mértékben különbözik az elméleti feltételes korreláció –0,707-es értékétől.

4. Értékelés

A függő változó (változók) eloszlásának normalitása számos statisztikai eljárás alkalmazásának feltétele. A teljesség igénye nélkül idetartozik az egy- és a kétmintás *t*-próba, az egy- és a többszempontos varianciaanalízis, a Pearson-féle korrelációval végzett korrelációs és regressziós elemzések, a faktoranalízis stb. A normalitás sérülése nem vezet mindig súlyos következményekhez (*Vargha* [2001], [2003b]), de a társadalomtudományok kutatóinak jó tudniuk, hogy mikor kell komolyan venniük ezt az alkalmazási feltételt. Például az egymintás *t*-próbánál, ha a mintanagyság nem haladja meg a 10-et, erősen nem normális eloszlású változók esetén jelentősen sérül a próba érvényessége (*Vargha* [2003b]).

Jelen tanulmány a parciális korrelációs együttható esetében veszi górcső alá a normalitási feltétel sérülésének a hatását. Mesterségesen szerkesztett változók segítségével meggyőzően kimutattuk, hogy ha az *X* és az *Y* változó kapcsolatából kiszűrendő *Z* változó nemlineáris összefüggésben van *X*-szel és *Y*-nal (ilyenkor *X*, *Y* és *Z* együttes eloszlása bizonyosan nem lehet normális), akkor a parciális korrelációs együttható esetenként teljesen mást mér, mint amit várunk tőle, illetve ahogy értelmezni szokták az értékét, ami erősen megnöveli az adatokból levont téves következtetések esélyét. Például cikkünk egyik változópárja esetében az $r_{xy.z}$ parciális korrelációs együttható értéke 0,875 volt, miközben a *Z* változó bármely rögzített értéke mellett –0,7 körüli erős negatív kapcsolatban volt egymással *X* és *Y*.

Ilyen anomália fellépéséhez nem kellett valami különösen kacifántos példát konstruálni. Mindössze annyit tettünk, hogy *X*-be és *Y*-ba beépítettünk egy sima kvadratikus *Z* hatást, valamint egy olyan lineáris összetevőt, mely *X*-re és *Y*-ra ellentétes hatást fejt ki (vö. /2/ egyenletek). Tekintve, hogy a társadalomtudományok kutatásainak változói között a kvadratikus jellegű (U vagy fordított U alakú) kapcsolatok nem tekinthetők fehér hollónak, a kutatóknak adatfeldolgozásaik során ezzel a lehetőséggel is számolniuk kell.

Z markáns kvadratikus hatása *X*-re és/vagy *Y*-ra (lásd például a 2. ábrát) azért kavarja meg annyira a dolgokat, mert ilyen esetben *X* és *Z*, illetve *Y* és *Z* között a *Z* változó értéktartományának különböző részeiben ellentétes (hol pozitív, hol negatív) kapcsolat van, ami átöröklődik *X* és *Y* kapcsolatára is.

Tanulmányunkban több olyan módszert is megemlítettünk, amelyekkel a nemlineáris összefüggések felderíthetők. Az egyik ilyen módszer a simított nemlineáris regresszió volt, mely egyszerűen futtatható a MiniStat programcsomag Windows változatában, a ROPstatban. A ROPstat (lásd *Vargha* [2007a], illetve www.ropstat.com) nehézség nélkül be tud olvasni az SPSS-ből *.por formátumban, vagy az Excelből szövegfájl formában elmentett (tabulátorral formattált) adatfájlokat.

Egyszerű módszert javasoltunk nemlineáris kapcsolatok esetén a parciális korrelációs együttható korrekciójára. Ez a korrekció mindössze abból áll, hogy ha feltételezhető a kvadratikus jellegű hatás fellépte, akkor Z mellett parciáljuk ki (szűrjük ki) a Z^2 változót is X és Y kapcsolatából. Ez végrehajtható bármely statisztikai programcsomagban (SPSS, ROPstat, Statistica stb.), csak előtte egy egyszerű transzformációval létre kell hozni Z^2-et Z segítségével.

Végül szeretnénk felhívni a figyelmet arra, hogy a parciális korrelációs együtthatók logikailag nagyon hasonlítanak a többszörös lineáris regresszió standardizált regressziós együtthatóira. Ez utóbbiakat egyesek eleve úgy értelmezik, mint parciális korrelációs együtthatókat az egyes független változók és a függő változó között, ha kiszűrjük a többi független változó hatását (lásd például *Bryman–Cramer* [2008]). Ennek az értelmezésnek a hibás voltáról bárki meggyőződhet, ha kiszámítja az említett parciális korrelációkat valamilyen korrelációs rutinban, s összeveti azokat a többszörös lineáris regresszió eredménylistáján megjelenő standardizált regressziós együtthatókkal. A standardizált regressziós együtthatók mindössze azt jelzik, hogy a függő változó várhatóan mekkorát változik szórásléptékben, ha az egyes függő változók értékét 1 szórásnyival megnöveljük, miközben a többi függő változót állandó szinten tartjuk (*Pedhazur* [1982] 247. old.).

Függelék

F1. A ρ_{X0Y0} korrelációs és a $\rho_{X0Y0.Z}$ parciális korrelációs együttható meghatározása a cikk /2/ egyenleteinek kikötése mellett tetszőleges pozitív a, b, c paraméterekre a következő.

A korrelációs együttható definíciója szerint (lásd például *Vincze* [1968]):

$$\rho_{X0Y0} = \frac{Cov(X0,Y0)}{D(X0)D(Y0)}. \qquad /F1/$$

A /2/ egyenletek és a korreláció lineáris operáció volta miatt

$$Cov(X0,Y0) = Cov(aZ+cU, aZ-cU+cV) = a^2Cov(Z,Z) - acCov(Z,U) + acCov(Z,V) + \\ + caCov(U,Z) - c^2Cov(U,U) + c^2Cov(U,V).$$

Tekintve, hogy U, V, Z egymástól függetlenek,

$$Cov(Z,U) = Cov(Z,V) = Cov(U,Z) = Cov(U,V) = 0\,,$$

továbbá U, V, Z standard volta miatt

$$Cov(Z,Z) = Cov(U,U) = 1.$$

Következésképpen

$$Cov(X0,Y0) = a^2 - c^2.$$

Most rátérünk $D(X0)$ és $D(Y0)$ meghatározására. A /2/ formula és a variancia tulajdonságai miatt – Z és U függetlenségét is figyelembe véve – kapjuk, hogy:

$$Var(X0) = Var(aZ + cU) = a^2 Var(Z) + c^2 Var(U) = a^2 + c^2.$$

Hasonló levezetéssel kapjuk, hogy

$$Var(Y0) = Var(aZ - cU + cV) = Var(aZ) + Var(-cU) + Var(cV) = a^2 + 2c^2.$$

Mindezek alapján $\rho(X0,Y0)$ fenti /F1/ képletébe helyettesítve kapjuk az igazolni kívánt /3/ formulát.

A $\rho_{X0Y0.Z}$ parciális korrelációs együttható meghatározásához az /1/ képletet használjuk, s ehhez szükségünk van $\rho(X0,Y0)$ mellett még a $\rho(X0,Z)$, $\rho(Y0,Z)$ korrelációkra is. Az előbbiekkel analóg gondolatmenetet követve

$$\rho(X0,Z) = \frac{Cov(X0,Z)}{D(X0)D(Z)} = \frac{a^2}{\sqrt{a^2 + c^2}} \text{ és}$$

$$\rho(Y0,Z) = \frac{Cov(Y0,Z)}{D(Y0)D(Z)} = \frac{a^2}{\sqrt{a^2 + 2c^2}}.$$

A $\rho(X0,Y0)$, $\rho(X0,Z)$, $\rho(Y0,Z)$ korrelációk így kapott kifejezéseit behelyettesítve az /1/ formulába kapjuk, hogy

$$\rho_{X0Y0.Z} = \frac{-c^2}{\sqrt{2c^4}} = \frac{-1}{\sqrt{2}} = -0,707, \qquad /F2/$$

ami egyben a feltételes korrelációs együttható értéke is $X0$ és $Y0$ között a Z változó rögzítése mellett.

F2. A továbbiakban a ρ_{XY} korrelációs és a $\rho_{XY.Z}$ parciális korrelációs együttható határozzuk meg a cikk /2/ és /5/ egyenleteinek kikötése mellett tetszőleges pozitív a, b, c paraméterekre.

Az F1. pontban alkalmazott utat követve $\rho_{XY} = \rho(X,Y)$-hoz a $Cov(X,Y)$, $D(X)$ és $D(Y)$ összetevőket határozzuk meg először. A /2/, /5/ egyenletek és a kovariancia tulajdonságai alapján, felhasználva azt is, hogy U, V, Z egymástól függetlenek:

$$Cov(X,Y) = Cov\left(aZ + bZ^2 + cU, aZ + bZ^2 - cU + cV\right) =$$
$$= a^2 Cov(Z,Z) + b^2 Cov\left(Z^2,Z^2\right) + abCov\left(Z,Z^2\right) + ba\left(Z^2,Z\right) - c^2 Cov(U,U) =$$
$$= a^2 + 2b^2 + 2abCov\left(Z,Z^2\right) - c^2.$$

Itt felhasználtuk, hogy $Cov\left(Z^2,Z^2\right) = Var\left(Z^2\right) = 2,$ mivel Z^2 1 szabadságfokú khi-négyzet-eloszlást követ (*Vincze* [1968]). De

$$Cov\left(Z,Z^2\right) = E\left(Z^3\right) - E(Z)E\left(Z^2\right) = 0,$$

mert a normális eloszlású változók páratlan fokszámú momentumai 0-k (*Rényi* [1968]), ami miatt

$$Cov(X,Y) = a^2 + 2b^2 - c^2.$$

Ugyanakkor

$$Var(X) = Var(aZ) + Var\left(bZ^2\right) + Var(cU) = a^2 + b^2 Var\left(Z^2\right) + c^2 = a^2 + 2b^2 + c^2.$$

Hasonlóképpen

$$Var(Y) = Var(aZ) + Var\left(bZ^2\right) + Var(cU) + Var(cV) = a^2 + 2b^2 + c^2 + c^2 = a^2 + 2b^2 + 2c^2.$$

Mindezek alapján már egyszerű behelyettesítéssel adódik a ρ_{XY} korrelációs együtthatóra vonatkozó /6/ formula.

A $\rho_{XY.Z}$ parciális korrelációs együttható meghatározásához az /1/ képletet használjuk, s ehhez szükségünk van ρ_{XY} mellett még a ρ_{XZ}, ρ_{YZ} páronkénti korrelációkra is. Az F1. pontban leírtakkal analóg gondolatmenetet követve:

$$Cov(X,Z) = Cov\left(X0 + bZ^2, Z\right) = Cov(X0,Z) + bCov\left(Z^2,Z\right) = Cov(X0,Z) = a,$$

mivel Z^2 és Z korrelálatlan egymással (lásd korábban). Hasonlóképpen kapjuk, hogy

$$Cov(Y,Z) = Cov(Y0,Z) = a,$$

így

$$\rho_{XZ} = \frac{a}{D(X)} = \frac{a}{\sqrt{a^2 + 2b^2 + c^2}} \quad \text{és} \quad \rho_{YZ} = \frac{a}{D(Y)} = \frac{a}{\sqrt{a^2 + 2b^2 + 2c^2}}.$$

Mindezek alapján egyszerű behelyettesítéssel adódik a $\rho_{XY.Z}$ parciális korrelációs együtthatóra vonatkozó /7/ formula igazsága.

F3. A következőkben a /8/–/11/ összefüggések bizonyítását foglaljuk össze vázlatosan.
A /8/ és a /10/ formula feltételes korrelációs együtthatóját egyaránt egy

$$\rho(X0,Y0 \mid feltétel) = \frac{Cov(X0,Y0 \mid feltétel)}{D(X0 \mid feltétel)D(Y0 \mid feltétel)}$$

formájú képlet segítségével határozzuk meg. Mivel a kovariancia lineáris operátor, az $X0$-t és az $Y0$-t definiáló /2/ egyenletek az F1. pontban ismertetett módon felbonthatók elemi komponenseikre. Ebből adódik /8/ és /10/ jobb oldalának számlálója, azt is felhasználva, hogy U, V, Z egymástól független és standard

$$Cov\left(U,U \mid feltétel\right) = Cov\left(U,U\right) = Var\left(U\right) = 1 \text{ és}$$
$$Cov\left(U,V \mid feltétel\right) = Cov\left(U,V\right) = 0.$$

Emiatt /8/ és /10/ levezetéséhez alapvetően $Var\left(Z \mid feltétel\right)$ alakú varianciák meghatározására van szükség. Például $V_Z\left(z\right)$ esetén ehhez a következő utat követhetjük.

A variancia definíciója miatt

$$V_Z\left(z\right) = Var\left(Z \mid Z < z\right) = E\left(Z^2 \mid Z < z\right) - E^2\left(Z \mid Z < z\right). \qquad /F3/$$

Itt a jobb oldalon

$$E\left(Z \mid Z < z\right) = \frac{1}{P(Z < z)} \int_{-\infty}^{z} sf(s)ds = F^{-1}(z)\left[f(s)\right]_{-\infty}^{z} = -\frac{f(z)}{F(z)}, \qquad /F4/$$

ahol $f\left(z\right)$ és $F\left(z\right)$ a standard normális eloszlás sűrűség-, illetve eloszlásfüggvényének értéke a z helyen. /F4/ levezetésénél felhasználtuk, hogy deriváltja:

$$f'\left(s\right) = -sf\left(s\right) \qquad /F5/$$

bármely s helyen.

Az $E\left(Z^2 \mid Z < z\right)$ komponensre parciális integrálással az alábbi összefüggést kapjuk:

$$E\left(Z^2 \mid Z < z\right) = \frac{1}{P(Z < z)} \int_{-\infty}^{z} s^2 f(s)ds.$$

A jobb oldali integrált $e(z)$-vel jelölve kapjuk:

$$e\left(z\right) = \int_{-\infty}^{z} f(s)ds + \int_{-\infty}^{z} f''(s)ds = F(z) + \left[f'(s)\right]_{-\infty}^{z} = F(z) + f'(z) = F(z) - zf(z)$$

/F5/ miatt és mert könnyen beláthatóan $f'\left(-\infty\right) = 0$. Mindebből már egyszerűen adódik a /9/ összefüggés. A /11/ formula hasonló gondolatmenettel vezethető le.

Irodalom

BRILLINGER, D. R. [2001]: Does Anyone Know When the Correlation Coefficient is Useful? A Study of the Times of Extreme River Flows. *Technometrics.* 43. évf. 3. sz. 266–273. old.

BRYMAN, A. – CRAMER, D. [2008]: *Quantitative Sata Analysis with SPSS 14, 15 & 16: A Guide for Social Scientists.* Psychology Press. London.

CRAMER, D. [2003]: A Cautionary Tale of Two Statistics: Partial Correlation and Standardized Partial Regression. *Journal of Psychology.* 137. évf. 5. sz. 507–511. old.

GARSON, G. D. [2009]: *Partial Correlation.* http://faculty.chass.ncsu.edu/garson/PA765/partialr.htm#assume.

MICCERI, T. [1989]: The Unicorn, the Normal Curve, and Other Improbable Creatures. *Psychological Bulletin.* 105. évf. 1. sz. 156–166. old.

PEDHAZUR, E. J. [1982]: *Multiple Regression in Behavioral Research. (Second Edition.)* Holt, Rinehart and Winston. Chicago.

RAE, M. J. – CARRETTA, T. R. [2006]: The Role of Measurement Error in Familiar Statistics. *Organizational Research Methods.* 9. évf. 1. sz. 99–112. old.

RÉNYI A. [1968]: *Valószínűségszámítás.* Tankönyvkiadó. Budapest.

TABACHNICK, B. G. – FIDELL, L. S. [2001]: *Using Multivariate Statistics*. Allyn and Bacon. Boston.

VARGHA A. [2001]: Érvényes-e a kétmintás *t*-próba nem normális eloszlások esetén? *Pszichológia.* 21. évf. 1. sz. 83–105. old.

VARGHA A. [2003a]: *Mi történik, mit tegyünk, ha változónk nem normális eloszlású? Számítógépes statisztikai elemzések, ordinális csoportösszehasonlító modellek.* MTA doktori értekezés. Budapest.

VARGHA A. [2003b]: Robusztussági vizsgálatok az egymintás *t*-próbával. *Statisztikai Szemle.* 81. évf. 10. sz. 872–890. old. http://www.ksh.hu/statszemle_archive/2003/2003_10/2003_10_872.pdf

VARGHA A. [2007a]: *Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal.* Pólya Kiadó. Budapest.

VARGHA A. [2007b]: *A ROPstat statisztikai menürendszere.* http://www.ropstat.com/.

VINCZE I. [1968]: *Matematikai statisztika ipari alkalmazásokkal*. Műszaki Könyvkiadó. Budapest.

Summary

The partial correlation is a frequently used coefficient for assessing the bivariate correlation of two quantitative variables by eliminating the influence of one or more other variables. It is generally interpreted as the correlation under the condition that the variables to be eliminated are fixed (not allowed to vary and influence the dependent variables), which is called in the statistical literature as conditional correlation.

The present paper convincingly shows, by means of theoretical derivations and practical examples, that under the violation of the assumption of multivariate normality (frequently due to nonlinear relationships among the variables investigated) the usual interpretation of the partial correlation

coefficient can be basically incorrect. There may be an extreme case where the value of the partial correlation coefficient is highly positive, close to 1, whereas the conditional correlation is a large negative value. To heal this problem the paper suggests partialling out certain function (in the simplest case the square) of the variables whose effects are to be eliminated if nonlinear relationships are likely to occur.

The paper discusses also the special case where the correlation of two variables is computed by a restriction of the range of a third variable.