Lipécz György*

A SZTOCHASZTIKUS KAPCSOLATOK MÉRÉSE ÉS A SZÓRÁSNÉGYZET-FELBONTÁS

A sztochasztikus kapcsolatok témakörében alapvető szerepet játszik a szórásnégyzet, és a szorossági mutatók egy jellegzetes csoportjában a szórásnégyzet-felbontás (a variancia dekompozíciója). Cikkünk e mutatók közös szemléleti és algebrai alapját elemzi. Igyekszünk olyan módon kifejteni mondandónkat, hogy az a statisztikaoktatásban is közvetlenül hasznosítható legyen.

Felhasználjuk, hogy a csoportokra bontott sokaság esetén a teljes variancia felírása a külső és belső szórásnégyzet összegeként egyszerűen és szemléletesen visszavezethető a szórásnégyzet-felbontás elemi esetére. Ezáltal a minőségi ismérv sztochasztikus hatása, az ezt mérő mutatószám minimális formalizmussal levezethető és tartalma szemléletessé tehető.

Ezután a szórásnégyzet-felbontás szükséges és elégséges feltételeit *általános* esetben fogalmazzuk meg, és ennek speciális eseteként mutatjuk be a regressziós függvényekre vonatkozóan a szórásnégyzet-felbontás lehetőségét és az illesztés jóságát mérő varianciahányadost.

Végezetül megmutatjuk, hogy a nem-lineáris regressziós függvényekre milyen feltételek mellett végezhető el a szórásnégyzet-felbontás, illetve, ha nem végezhető el, akkor milyen egyszerű lineáris transzformációval tehetők erre alkalmassá, miközben a regresszió továbbra is nem-lineáris marad.

A szórásnégyzet-felbontás elemi esete

Egy egyszerű alapösszefüggésről lesz az alábbiakban szó, amit nem szoktak a szórásnégyzet-felbontás elemi esetének minősíteni, de számunkra kiinduló pontul szolgál az összetettebb esetek elemzéséhez.

Legyen egy n elemű y_i adatsorunk $\bar{y}$ átlaggal, és egy A konstans. Az i-edik adat eltérése az A-tól, az $(y_i - A)$ különbség felbontható két részre,

- az y_i-nek az $\bar{y}$-tól való eltérésére,
- és az $\bar{y}$-nak az A-tól való eltérésére.

Azaz: $(y_i - A) = (y_i - \bar{y}) + (\bar{y} - A)$

* *főiskolai tanár, Általános Vállalkozási Főiskola*

Az $\sum (y_i - A)^2$ eltérésnégyzet-összeg szintén felbontható két részre, két négyzetösszeg összegére:

$$\sum (y_i - A)^2 = \sum (y_i - \bar{y} + \bar{y} - A)^2 = \sum (y_i - \bar{y})^2 + \sum (\bar{y} - A)^2 \qquad (1)$$

A **bizonyításhoz** csak azt kell belátni, hogy az alábbi

$$\sum (y_i - \bar{y} + \bar{y} - A)^2 = \sum \left[(y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - A) + (\bar{y} - A)^2 \right]$$

azonosságban a jobboldali másik tag, tehát a $2\sum (y_i - \bar{y})(\bar{y} - A)$ szorzatösszeg értéke azonosan nulla. Ez abból következik, hogy az $(\bar{y} - A)$ konstans, tehát kiemelhető, $2(\bar{y} - A)\sum (y_i - \bar{y})$, a $\sum (y_i - \bar{y})$ összege pedig azonosan 0, mert az átlagtól való eltérések összege mindig 0.

Megjegyzések

1. Jelöljük *d*-vel az $(\bar{y} - A)$ konstanst, ezzel a jelöléssel a gondolatmenet és a tétel kicsit más formában:

$$\sum (y_i - A)^2 = \sum (y_i - \bar{y} + \bar{y} - A)^2 = \sum [(y_i - \bar{y}) + d]^2 = \sum (y_i - \bar{y})^2 + 2d\sum (y_i - \bar{y}) + nd^2$$

De $2d\sum (y_i - \bar{y}) = 0$, ezért $\sum (y_i - A)^2 = \sum (y_i - \bar{y})^2 + \sum d^2 = \sum (y_i - \bar{y})^2 + nd^2$

2. Ezt többnyire úgy szokták megfogalmazni, hogy ha a szórásnégyzetet az átlag helyett egy annál *d*-vel nagyobb számmal számoljuk ki (a *d* negatív is lehet), akkor ez a „torzított" szórásnégyzet d^2-tel lesz nagyobb, mint az eredeti helyes érték.

$$\frac{\sum (y_i - (\bar{y} + d))^2}{n} \equiv \frac{\sum (y_i - \bar{y})^2}{n} + d^2 = \sigma^2 + d^2 \qquad (2)$$

3. Ebből – az analízis igénybevétele nélkül is – azonnal látható, hogy az $\sum (y_i - A)^2$, és az $\frac{\sum (y_i - A)^2}{n}$ kifejezés az $A = \bar{y}$ választás esetén lesz a legkisebb.

Szórásnégyzet-felbontás csoportokra osztott adatok esetén (A H-négyzet mutató logikája)

A fenti (1) ill. (2) összefüggést közvetlenül alkalmazhatjuk, ha a szórásnégyzet-felbontást csoportokra bontott adatok esetén szeretnénk elvégezni. Ez a kiindulás megkönnyíti a csoportképző ismérv sztochasztikus hatásának elemzését, értelmezését.

Legyen egy csoportokra bontott adathalmazunk. Jelöljük az adatokat továbbra is *y*-nal. A j-edik csoport létszáma n_j, átlaga $\bar{y}_j$, szórásnégyzete (a j-edik részátlagtól való eltérésekből számolva) σ_j^2. A főátlag $\bar{y}$.

Ha a *csoportok közötti* különbségektől el akarunk tekinteni, tehát csak a *csoportokon belüli* eltéréseket akarjuk jellemezni, végezzük el a következő gondolatkísérletet. Egy-egy csoport minden adatából vonjunk le egy konstanst, amely által az adott csoport átlaga a főátlaggal lesz egyenlő. Azaz: a j-edik csoport minden adatából vonjunk le $d_j = \bar{y}_j - \bar{y}$ értéket. Így most minden csoport átlaga egyenlő lesz egymással is, a főátlaggal is, miközben a csoporton *belüli* különbségek nem változtak. Számoljuk ki ennek a módosított adathalmaznak a szórásnégyzetét. Mivel a σ_j^2 értékek az adatmódosítással nem változtak, a módosított halmaz szórásnégyzete a σ_j^2 értékekből így számolható

$$\sigma_B^2 = \frac{1}{n}\sum_{j=1}^{m} n_j \sigma_j^2 \qquad (3)$$

Ez most már csak a csoporton belüli eltéréseket tükrözi, mert az adatok transzformációjával a csoportközi eltéréseket eltüntettük. Ezért ez a *belső szórásnégyzet*.

Most nézzük meg, mennyivel távolodtunk el a valódi szórásnégyzettől. Vegyük a j-edik eltolt csoportot. Ha most visszatoljuk a helyére, de a csoportátlag helyett továbbra is a főátlagtól való eltérések négyzetével számolunk, akkor a fenti (2) összefüggés szerint a csoport szórásnégyzeténél d_j^2-tel nagyobb értéket kapunk. Az *m* csoportra:

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{m}\left(n_j\sigma_j^2 + n_j d_j^2\right) = \sum_{j=1}^{m}\frac{n_j}{n}\sigma_j^2 + \sum_{j=1}^{m}\frac{n_j}{n}d_j^2 \qquad (4)$$

Itt a $\sum_{j=1}^{m}\frac{n_j}{n}d_j^2$ tag a külső szórásnégyzet.

Most induljunk ki ellenkezőleg:

Transzformáljuk úgy az adatokat, hogy minden egyes csoportban az adatokat a csoportátlaggal tesszük egyenlővé. Így a *csoporton belüli eltérésektől tekintünk el*. A főátlag ugyanaz marad. Számoljuk ki ennek a transzformált sokaságnak a szórásnégyzetét, ami nyilván csakis a csoportok közötti különbségeket tükrözheti. Ez nem más, mint a külső szórásnégyzet:

$$\sigma_K^2 = \frac{1}{n}\sum_{j=1}^{m} n_j d_j^2$$

Mennyivel torzítottuk el a teljes valódi szórásnégyzetet?

Most (1) alapján mondhatjuk, hogy ha a j-edik csoportban az elemeket visszatoljuk az eredeti helyükre, az elemek főátlagtól való eltéréseinek négyzetösszege $n_j\sigma_j^2$-tel fog növekedni.

Az *m* csoportra a szórásnégyzet növekedése $\frac{1}{n}\sum_{j=1}^{m} n_j\sigma_j^2$. Ez nem más, mint a belső szórásnégyzet.

Ennek a sztochasztikus kapcsolatnak a mérésére szolgáló hányados, a H-négyzet mutató:

$$H^2 = \frac{\sigma_K^2}{\sigma_K^2 + \sigma_B^2}$$

Ennek logikája az, hogy a két rész, amelyre a szórásnégyzetet felbontottuk, két elkülöníthető hatást tükröz: az egyik a csoportképző ismérv hatását, a másik az összes többi – véletlennek tekinthető – hatást.

- A külső szórásnégyzet a csoportok közötti különbségeket tükrözi, azzal a fikcióval, hogy mi lenne, ha az egyes csoportokon belül az adatok között nem lenne különbség.
- A belső szórásnégyzet viszont kizárólag a csoportokon belüli eltéréseket tükrözi, azzal a fikcióval, hogy mi lenne, ha a csoportátlagok között nem lenne különbség.
- A kétfajta szórásnégyzet összege kiadja a teljes szórásnégyzetet. A fentiek alapján pedig feltehető, hogy minél nagyobb a külső szórásnégyzet aránya a teljes szórásnégyzeten belül, annál nagyobb a csoportképző ismérv hatása a többi (véletlen) hatáshoz képest.

A szórásnégyzet-felbontás általános esetben

A fentiekben a teljes szórásnégyzetet úgy bontottuk két részre, hogy a főátlagtól való távolságot a részátlaggal választottuk ketté. Most azt az esetet nézzük meg, amikor minden egyes y_i adatnak az y átlagtól való távolságát egy a_i értékkel osztjuk két részre, ezekből számolunk eltérésnégyzet-összegeket, és azt vizsgáljuk, milyen feltételek mellett végezhető a szórásnégyzet-felbontás.

Legyen egy y_i és egy a_i adatsorunk. (i=1, 2…n). A

$$\sum (y_i - \bar{y})^2 = \sum (y_i - a_i)^2 + \sum (a_i - \bar{y})^2$$

szórásnégyzet-felbontás szükséges és elégséges feltétele, hogy (5)

$\sum (y_i - a_i)(a_i - \bar{y}) = 0$, vagy kicsit átalakítva.

$\sum (y_i - a_i)a_i - \sum (y_i - a_i)\bar{y} = 0$ legyen

A bizonyításhoz a baloldalt alakítsuk át:

$$\sum (y_i - \bar{y})^2 = \sum [(y_i - a_i) + (a_i - \bar{y})]^2 = \sum (y_i - a_i)^2 + \sum (a_i - \bar{y})^2 + 2\sum (y_i - a_i)(a_i - \bar{y})$$

Azaz, az utolsó tagnak kell 0-nak lennie.
A 2-vel való osztás után, kissé átalakítva:

$$\sum (y_i - a_i)a_i - \sum (y_i - a_i)\bar{y} = 0.$$

Megjegyzés

Statisztikai elemzésekben sokszor kézenfekvően adódik az

$$\sum \left(y_i - a_i\right) a_i = 0 \qquad \text{(6-a)}$$

$$\sum \left(y_i - a_i\right) = 0 \qquad \text{(6-b)}$$

feltételek fennállása, amelyek elégségesek, bár nem szükségesek.

A szórásnégyzet-felbontás lehetősége a regressziós függvényeknél

A regressziós függvény fogalma a szokásos és általánosan használt definíció szerint (elvileg természetesen más definíció is elképzelhető lenne) a *feltételes várható érték* fogalmára épül. Ha adott két valószínűségi változó a hozzájuk tartozó együttes eloszlással együtt, akkor a $y = M\left(Y \middle| X = x\right)$ függvényt nevezik *regressziós függvénynek,* ahol *X* és *Y* a valószínűségi változókat jelölik, és *M* a várható érték jele. A regressziós függvény x-hez tartozó értéke tehát az adott x-értékhez tartozó feltételes eloszlás várható értéke.

Ez a definíció eleve biztosítja a varianciafelbontás lehetőségét, de a statisztikai gyakorlatban erre vonatkozóan nem adódik tényleges számítási és értékelési feladat, hiszen szokásosan csak egy véges minta áll rendelkezésre, a megfigyelt $(x_i; y_i)$ értékpárok véges halmaza, aminek felhasználásával kívánjuk becsülni a regressziós függvény paramétereit.

Ha az egyes x-értékekhez elegendő y-érték áll rendelkezésre, akkor a minta pontjaira úgy illeszthetünk regressziós függvényt, hogy – követve az elméleti definíciót – a regressziós függvény i-edik értékét az x_i-hez tartozó y-értékek átlagaként (feltételes várható értékeként) határozzuk meg. Ez az ún. *empirikus* regressziófüggvény. Ez esetben a szórásnégyzet-felbontást (a mintabeli variancia regressziós dekompozícióját) a csoportokra bontott adathalmazra megfogalmazott összefüggések biztosítják. Egy x-értékhez tartozó y-értékek halmaza alkot ez esetben egy csoportot, és így külső és belső szórásnégyzetet tudunk számolni. A pontoknak a *görbe körüli* szóródása, illetve varianciája adja a belső szórásnégyzetet, a *görbe pontjainak az y átlaga körüli* varianciája a külső szórásnégyzetet. A kettő összege a teljes szórásnégyzet. Ezekből meghatározható a szórásnégyzet-hányados, amely mutatja az illeszkedés jóságát, és úgy értelmezhető, hogy megadja, a regressziós görbe milyen arányban magyarázza a tapasztalati y-értékek alakulását.

A leggyakoribb regressziós függvénytípus, az $\hat{y}_i = b_0 + b_1 x_i$ lineáris regresszió függvény esetén az (5) tételben szereplő feltételt, ill. a (6-a) és (6-b) feltételeket a *normálegyenletek* biztosítják. Az $\hat{y}_i$ értékek az a_i értékeknek feleltetendők meg.

A b_1 szerinti parciális derivált 0-val egyenlővé téve és egyszerűsítve: $\sum \left(y_i - \hat{y}_i\right) x_i = 0$.

Ez ekvivalens a $\sum \left(y_i - \hat{y}_i\right) \hat{y}_i = 0$ egyenlettel, a (6-a) előfeltevéssel.

A b_0 szerinti parciális derivált 0-val egyenlővé téve és egyszerűsítve: $\sum \left(y_i - \hat{y}_i\right) = 0$, ez a (6-b) előfeltevés.

A szórásnégyzet-felbontás feltételei tehát teljesülnek a lineáris regresszióra.

A nem lineárisan illesztett függvények esetében a szórásnégyzet-felbontás általában nem végezhető el, pl. az *exponenciális* és *logisztikus* függvények esetén, ha nem tartalmaznak additív konstanst. Az

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k$$

k-ad fokú *polinomiális illesztés* esetén teljesül, mert – könnyen belátható módon – teljesül a (6-a) és a (6-b) feltevés.

A regresszió lineáris igazítása

A becsléshez használt *nem-lineáris* regressziós függvények tehát általában nem teszik lehetővé a szórásnégyzet-felbontást. Van azonban egy egyszerű lehetőség, hogy úgy transzformáljuk a regressziós függvényt, hogy az alkalmassá váljon a szórásnégyzet-felbontásra, és közben a reziduális eltérésnégyzet-összeg még csökkenjen is – vagy maradjon változatlan. Ez *bármilyen* regressziós függvény esetén használható, amennyiben a regressziós függvény lineáris transzformációja is elfogadható regresssziós függvényként, tehát hogy az x és y kapcsolatára vonatkozó előzetes ismereteink ezt nem zárják eleve ki.

Definíciónk a következő:

Legyen az eredeti $(x_i; y_i)$ adatokra illesztett – tetszőleges módszerrel előállított – regresszió $\hat{y} = \hat{y}(x)$. Ennek *lineáris igazítása* az $\hat{\hat{y}} = c_0 + c_1 \hat{y}(x)$, ahol a c_0 és a c_1 olyan értékek, amely mellett a $SS_e = \sum_{i=1}^{n} [y_i - c_0 - c_1 \hat{y}(x_i)]^2$ kifejezés minimális. (7)

A lineáris igazítás néhány fontos tulajdonsága:

1. Az $\hat{\hat{y}}$ tehát egy összetett függvény: $\hat{\hat{y}} = \hat{\hat{y}}[\hat{y}(x)]$.

2. A lineáris igazítás *sokváltozós esetre is alkalmazható*, a transzformáció szempontjából mindegy, hogy az eredeti $\hat{y}_i$ -értékek egy skalár- vagy egy vektorváltozó függvényeként adódtak.

3. A fenti (7) definícióban szereplő SS_e kifejezés c_0 és c_1 változókra nézve másodfokú folytonos függvény, nem-negatív, így *alulról korlátos*, *van tehát minimuma*. (Parciális deriváltjaik elsőfokú fv-ek, a szélsőértékhelyek tehát könnyen meghatározhatók.)

4. Mivel $c_0 = 0$ és a $c_1 = 1$ értékpárra visszakapjuk az eredeti regresszió eltérés-négyzet összegét, $SS_e = \sum_{i=1}^{n} [y_i - \hat{y}(x_i)]^2$, ezért biztosak lehetünk benne, hogy ennél *nagyobb* négyzetösszeget a lineáris igazítással nem kaphatunk. Ez tehát a minimalizálással kapható négyzetösszeg felső korlátja. A lineáris igazítás *vagy javítja az illeszkedést, vagy változatlanul* hagyja.

5. Lineárisan igazított regressziós függvény esetén a *szórásnégyzet-felbontás lehetősége* könnyen belátható, ha az (5) tételben ill. a (6-a) és (6-b) feltételekben az a_i-értékek helyébe az $\hat{y}_i$-értékeket írjuk.

A lineáris igazítás egyszerű művelete után a függvényünk továbbra is hasonló jellegű marad, mint az eredeti regressziós függvény (ha eredetileg nem volt lineáris, továbbra se lesz az), miközben az illeszkedése javult vagy legalábbis nem romlott. A változók közötti sztochasztikus kapcsolat pedig most már korrekten kifejezhető a megfelelő varianciahányadossal, amely a regressziós függvény varianciáját a teljes varianciához viszonyítja.

Felhasznált irodalom

Hajdú Ottó (2003): *Többváltozós statisztikai számítások*. Budapest, Központi Statisztikai Hivatal.

Hunyadi László (1992): *A varianciafelbontásról.* Statisztikai Szemle, 12.

Köves Pál – Parnitzky Gábor (1981): *Általános statisztika*. Budapest, Közgazdasági és Jogi Kiadó.

Móri F. Tamás – Székely J. Gábor (szerk.) (1986): *Többváltozós statisztikai analízis*. Budapest, Műszaki Könyvkiadó.

Rényi Alfréd (1973): *Valószínűségszámítás*. Budapest, Tankönyvkiadó.

Weisstein, Eric W.: *Nonlinear Least Squares Fitting*. http://mathworld.wolfram.com