Hunyadi László
kandidátus, egyetemi tanár, a Statisztikai Szemle főszerkesztője
E-mail: laszlo.hunyadi@ksh.hu

A heteroszkedaszticitásról egyszerűbben

A heteroszkedaszticitás az ökonometriai modellezés egyik kulcsfogalma, és bár valójában meglehetősen egyszerű fogalom, talán bonyolult hangzású neve is hozzájárul ahhoz, hogy megértése és gyakorlati kezelése olykor elég nehézkes. Az oktatásban szerzett tapasztalatok alapján sokszor úgy tűnt, hogy a hallgatók itt adták fel a regressziós modell megismerését, itt jutottak el oda, hogy a továbbiakban meg se próbálták megérteni.

E rövid írás célja, annak bemutatása, hogy a heteroszkedaszticitás fogalmát, és a vele kapcsolatos fontosabb, elsősorban becslési vonatkozásokat a megszokottnál egyszerűbben is lehet tárgyalni, legalábbis, ami az alapokat illeti. A modell, amit itt használunk, igen egyszerű, és esetenként még ezt is csak speciális, könnyen kezelhető esetekre vizsgálom, hiszen mindez „csak" arra szolgál, hogy megértsük ezt a fontos fogalmat, és a vele kapcsolatos leglényegesebb tételeket. Ezért ez a kis írás nem helyettesítheti a heteroszkedaszticitás részletesebb ökonometriai leírásait, melyek a magyar nyelvű szakirodalomban is hozzáférhetők (*Kőrösi et al*. [1990], *Maddala* [2004], *Ramanathan* [2003]).

A heteroszkedaszticitás fogalma és egy egyszerű modellje

A heteroszkedaszticitás általában egy modellben a szórások különbözőségét jelenti. Sok modell esetén, főként az egyszerűség kedvéért, a különböző csoportok, kategóriák, változóértékek mögött meghúzódó sokaságok szórásai egyenlőségét feltételezik. Ez ritkán fedi a valóságot, de kényelmes feltételezés, többnyire leegyszerűsíti a modell szerkezetét, így a becslését, tesztelését stb. is. Az egyenlő szórások feltételezése, azaz a homoszkedaszticitás, nem természetes feltevés, hanem mesterséges egyszerűsítés (hasonlóan az idősorelemzés stacionaritás fogalmához). A heteroszkedaszticitás tehát nem hiba (mint ahogy azt sok könyv tárgyalja), hanem nyitás a valóság felé. Hibának csak abban az összefüggésben lehet tekinteni, amelyben – elsősorban didaktikai okokból – a lehető legegyszerűbb, homoszkedasztikus modellt tekintjük alapnak.

A heteroszkedaszticitás bemutatására és egyes kérdéseinek kezelésére szokásos lineáris regressziós modell helyett most egy egyszerű átlagbecslést tekintünk, amely persze felfogható olyan lineáris regressziónak is, ahol csak a tengelymetszet együtthatója különbözik 0-tól. Kiinduló pontunk az, hogy adott egy heterogénnek tekinthető sokaság, amelyben azonban a különböző, eltérő szórású csoportok közös várható értékkel rendelkeznek, és feladatunk ennek a közös várható értéknek a becslése. E feladat mögé különféle interpretációkat képzelhetünk. Egy ilyen feladat lehet az, hogy két vagy több ágazatban hasonlók az átlagfizetések, ám az eltérő vállalatszerkezet (például nagyság) miatt az ágazati szórások lényegesen eltérnek egymástól. Egy másik lehetséges feladat lehet az, hogy valamely műszaki cikk átlagárát kívánjuk becsülni, kombinált idősoros és keresztmetszeti adatokból. Ha feltételezzük az átlagár stabilitását (ami a gyártástechnológia fejlődése, illetve az ezzel párhuzamosan zajló műszaki fejlődés eredőjeként időben stabilnak tekinthető), akkor az időben változó (bővülő) választék következtében ugyancsak változó szórás tételezhető fel.

Ezekben az esetekben különböző szórású részsokaságokból veszünk mintákat, amelyek nagysága is változhat. Egy ilyen feladaton már jól megmutathatók a heteroszkedasztikus környezetben történő becslések tulajdonságai, de a még egyszerűbb tárgyalás érdekében a továbbiakban azt is feltételezzük, hogy minden részsokaságból egyetlen elemű mintát veszünk a becslés érdekében. Ekkor tehát feladatunk a következő.

Legyen $y_1, y_2, \ldots, y_n$ egy n-elemű FAE-minta, ahol feltételezzük, hogy $y_i \sim \left(\mu, \sigma_i^2\right)$ és σ_i ismert szórás. A cél a közös μ várható érték becslése.

Ez a feladat nem más, mint a legegyszerűbb átlagbecslés kiterjesztése nem egyenlő szórás (variancia) esetére. Hangsúlyozzuk, hogy a különböző mintanagyságok elvben semmit sem változtatnának a későbbieken, csak a jelölések válnának lényegesen bonyolultabbá, és így félő, hogy éppen az a pont sikkadna el a tárgyalásból, ami miatt ezt a kis elemzést végezzük. Ezért maradunk ennél a végletekig lecsupaszított, leegyszerűsített modellnél. A továbbiakban ezen a modellen keresztül mutatjuk meg a becslés fő problémáit, a lehetséges becslőfüggvényeket és azok tulajdonságait.

A hagyományos (OLS-) becslés

A legegyszerűbb becslőfüggvényt μ -re ebben az esetben is a közönséges legkisebb négyzetek (OLS) módszerével nyerhetjük a következő módon.

Legyen egyszerű modellünk

$$y_i = \mu + \varepsilon_i ,$$

ahol a modell végletesen egyszerű szerkezete folytán ε_i várható értékétől eltekintve ugyanolyan tulajdonságokkal rendelkezik, mint y_i. Ezért az OLS azt a $\hat{\mu}$-t keresi, amelyikre

$$g = \sum_{i=1}^{n} \left(y_i - \hat{\mu}\right)^2 \qquad /1/$$

minimális. Vegyük észre, hogy /1/ nem veszi figyelembe az eltérő varianciákat, azaz a heteroszkedasztikus környezetet.

A megoldás egyszerű szélsőérték-feladat megoldásaként közismert:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y} = \hat{\mu}_{OLS} ,$$

vagy a későbbi tárgyalással jobban összhangban lévő jelölésekkel:

$$\hat{\mu}_{OLS} = \sum_{i=1}^{n} w_i y_i , \quad \text{ahol } w_i = \frac{1}{n}. \qquad /2/$$

Ennek a becslőfüggvénynek a tulajdonságai azonnal adódnak:

– Torzítatlan, azaz $E(\hat{\mu}_{OLS}) = \mu$, hiszen $E(y_i) = \mu$ és $\sum_{i=1}^{n} w_i = 1$;

– Varianciája:

$$Var(\hat{\mu}_{OLS}) = \sum_{i=1}^{n} w_i^2 Var(y_i) = \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2} , \qquad /3/$$

ami alapján – rögtön belátjuk – lehet kisebb varianciájú torzítatlan becslőfüggvényt készíteni.

– A variancia becslése értelmetlen, hiszen nincs elegendő mintaelemünk a becsléshez. (Megjegyezzük, hogy ha az említett, súlyozott feladatot vizsgáltuk volna, akkor elvben lenne elegendő szabadságfokunk a varianciák becslésére, de a szokásos (OLS) varianciabecslés ekkor is értelmetlen, hiszen nem egy, hanem több variancia létezik.

Egy ad hoc becslőfüggvény jobb tulajdonságokkal

Ez az irodalomban nem szokásos becslőfüggvény a gyakorlatban sem használatos, itt csupán érdekességképp mutatjuk be annak demonstrálására, hogy a heteroszkedasztikus környezetben a hagyományos becslőfüggvény milyen gyengén teljesít. Ez a becslőfüggvény azon az ötleten alapul, hogy a nagyobb szórású réteghez tartozó mintaelem arányosan kisebb súlyt képviseljen a becslőfüggvényben, és fordítva. Ezért alakja:

$$\hat{\mu}_{AH} = \sum_{i=1}^{n} w_i y_i \text{, ahol } w_i = \frac{1/\sigma_i}{\sum_{i=1}^{n} 1/\sigma_i}. \qquad /4/$$

Mivel a konstrukciójánál fogva $\sum_{i=1}^{n} w_i = 1$, a /4/ becslőfüggvény is torzítatlan, varianciája pedig

$$Var(\hat{\mu}_{AH}) = \sum_{i=1}^{n} \frac{\left(1/\sigma_i^2\right) \cdot \sigma_i^2}{\left(\sum_{i=1}^{n} 1/\sigma_i\right)^2} = \frac{n}{\left(\sum_{i=1}^{n} 1/\sigma_i\right)^2}. \qquad /5/$$

Könnyen belátható, hogy az /5/ variancia kisebb vagy egyenlő a /3/ OLS-variancájával, azaz

$$\frac{n}{\left(\sum_{i=1}^{n} 1/\sigma_i\right)^2} \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2}. \qquad /6/$$

Az egyenlőtlenség igazolásához szorozzuk meg mindkét oldalt n-nel, majd vonjunk gyököt mindkét oldalon. Ekkor azt kapjuk, hogy

$$\frac{n}{\sum_{i=1}^{n} 1/\sigma_i} \leq \left(\frac{\sum_{i=1}^{n} \sigma_i^2}{n} \right)^{1/2}, \qquad /7/$$

ami pedig a harmonikus, és a négyzetes átlagok közti közismert nagyságrendi reláció.[1] Ekkor /7/ azt jelenti, hogy az OLS-becslőfüggvény ilyen feltételek mellett nem lehet minimális varianciájú, hiszen a „találomra” felvett becslőfüggvény is torzítatlan, de varianciája kisebb.

[1] Köszönet *Mihályffy Lászlónak* az itt nyújtott látszólag apró, de valójában nagyon hasznos segítségért.

További becslőfüggvények

Az ökonometriai szakirodalom további becslőfüggvényeket javasol a heteroszkedasztikus esetre. Ezek közül kiemelkedik a WLS, amelyik azon az elven nyugszik, hogy a varianciákkal fordítottan súlyozott eltérés-négyzetösszegeket minimalizálja:

$$g = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}}{\sigma_i} \right)^2 \to \min_{\mu} \qquad /8/$$

Szélsőérték-számítással azonnal adódik, hogy a $\frac{\partial g}{\partial \hat{\mu}} = -2\sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}}{\sigma_i} \right) \cdot \frac{1}{\sigma_i} = 0$ egyenlet megoldása adja a becslőfüggvényt:

$$\hat{\mu}_{WLS} = \sum_{i=1}^{n} w_i y_i \quad \text{és} \quad w_i = \frac{1/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2} . \qquad /9/$$

Ez a becslőfüggvény természetesen torzítatlan és varianciája:

$$Var(\hat{\mu}_{WLS}) = \sum_{i=1}^{n} \frac{\left(1/\sigma_i^4\right) \cdot \sigma_i^2}{\left(\sum_{i=1}^{n} 1/\sigma_i^2\right)^2} = \sum_{i=1}^{n} \frac{\left(1/\sigma_i^2\right)}{\left(\sum_{i=1}^{n} 1/\sigma_i^2\right)^2} = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2} . \qquad /10/$$

Úgy is becslőfüggvényhez juthatunk, ha a torzítatlan becslőfüggvények halmazán minimáljuk annak varianciáját. Ha mindez a lineáris becslőfüggvények terén történik, akkor jutunk el a legjobb torzítatlan, lineáris becslőfüggvényhez (BLUE). Esetünkben ez azt jelenti, hogy a következő korlátozott szélsőérték-feladatot kell megoldanunk:

$$\sum_{i=1}^{n} w_i^2 \sigma_i^2 \to \min_w, \text{ feltéve, hogy} \quad w_i \geq 0 \quad \text{és} \quad \sum_{i=1}^{n} w_i = 1 . \qquad /11/$$

A megoldást itt csak a legegyszerűbb ($n = 2$) esetre mutatjuk meg, mivel tetszőleges n esetére a megoldás kicsit hosszadalmas (*Hunyadi* [2000]). Ekkor

$$g = w_1^2 \sigma_1^2 + (1 - w_1)^2 \sigma_2^2 \to \min_{w_1}$$

és a szélsőérték létezésének elsőrendű feltétele

$$\frac{\partial g}{\partial w_1} = 2 w_1 \sigma_1^2 - 2(1 - w_1)\sigma_2^2 = 0 .$$

Ezt átrendezve w_1-re az adódik, hogy $w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$, és innen természetesen $w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$. További egyszerű átrendezés után ebből az kapható, hogy

$$w_1 = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}, \qquad /12/$$

ami analóg a /9/ WLS-súlyokkal. A hivatkozott tanulmány általánosságban is bizonyítja ezt, ezért azt állíthatjuk, hogy a WLS a legjobb lineáris torzítatlan becslőfüggvényt (BLUE) eredményezi. Megjegyezzük, hogy a /12/ súlyok alkalmazásának egy gyakori példája a portfolióelmélet, ahol két (vagy több) kockázatú (varianciájú) befektetés minimális kockázatú (optimális) portfolióját kívánjuk összerakni az összetétel (súlyok) alkalmas megválasztásával.

Eddig a változók eloszlására semmiféle feltételt nem kötöttünk ki, ezért ezek az eredmények eloszlástól függetlenek, az eloszlásra nézve robusztusak. Ha feltételezünk valamilyen ismert eloszlást a sokasági változóra, akkor a maximum likelihood alkalmazásával további, még erősebb eredményeket nyerhetünk. Ilyen esetekben kézenfekvő a normális eloszlás feltételezése, ami annyit jelent, hogy mintaelemeinkre a továbbiakban az $y_i \sim N(\mu, \sigma_i^2)$ feltételezést tesszük, és a független mintaelemek miatt a kovarianciamátrix diagonalitását is feltételezzük. Ekkor a likelihood függvény és a log-likelihood függvény rendre a következő lesz:

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma_1 \sigma_2 \cdots \sigma_n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma_i}\right)^2\right),$$

$$\log L = C - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma_i}\right)^2,$$

ahol a C konstansban foglaltuk össze a μ-től nem függő tagokat. A log-likelihood szélsőértékhelyén jutunk el a maximum likelihood (ML) becslőfüggvényhez. Az első derivált:

$$\frac{\partial \log L}{\partial \mu} = \left(\sum_{i=1}^{n}\frac{y_i - \mu}{\sigma_i}\right) \cdot \frac{1}{\sigma_i}, \qquad /13/$$

és ennek 0-helyén lehet szélsőérték. /13/-at megoldva az kapjuk, hogy

$$\hat{\mu}_{ML} = \sum_{i=1}^{n} w_i y_i \quad \text{és} \quad w_i = \frac{1/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2} \ . \qquad /14/$$

Mivel ezen a helyen a második derivált negatív, ez valóban ML-becslőfüggvény, és látható, hogy pontosan megegyezik a /9/ WLS-becslőfüggvénnyel. Ez természetesen azt is jelenti, hogy normális eloszlás esetén rendelkezik az ML nagymintás tulajdonságaival (konzisztens, aszimptotikusan hatásos és határeloszlása normális).

A likelihood függvény elemzése azonban még további következtetések levonását engedi meg. A második derivált ugyanis

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\sum_{i=1}^{n} \frac{1}{\sigma_i^2} ,$$

és ennek alapján az információs határ ennek várható értékének –1-szerese

$$I(\mu) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} ,$$

ami éppen reciproka a WLS-becslőfüggvény varianciájának:

$$I(\mu) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = \frac{1}{Var(\hat{\mu}_{WLS})} . \qquad /15/$$

Ez azt jelenti, hogy ha feltételezzük a sokasági változók normális eloszlását, akkor a közös várható értékre készített becslőfüggvény varianciája eléri a Cramér–Rao alsó határt, így a becslőfüggvény véges mintán is abszolút hatásos (MVUE) (*Hunyadi* [2001]).

Következtetések

A következtetéseket összefoglalva kirajzolódik az a néhány állítás, amelyek ebben, a legegyszerűbb esetben jellemzik a heteroszkedasztikus környezetben készített becsléseket.

– Heteroszkedasztikus környezetben a μ egyszerű OLS-becslőfüggvénye torzítatlan marad.

– A variancia OLS-becslése ebben az esetben értelmetlen.

– Könnyen lehet találni olyan torzítatlan becslőfüggvényt, amelyik varianciája kisebb, mint az OLS-becslőfüggvényé, tehát az elveszti hatásosságát.

– A súlyozott legkisebb négyzetekkel készült becslőfüggvény (WLS) torzítatlan és a legjobb lineáris tulajdonságokkal rendelkezik.

– Ha feltételezzük, hogy a változók normális eloszlást követnek, akkor a maximum likelihood módszer a WLS-sel megegyező becslőfüggvényt javasol. Ez nem csupán aszimptotikusan, de véges mintákban is legjobb (minimális varianciájú) torzítatlan becslőfüggvény, és nagy minták esetén örökli az ML tulajdonságait, azaz konzisztens, és határeloszlása normális.

Mint azt korábban említettük, a vizsgált modell az $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ lineáris regressziós modell speciális esete, ha $\beta_0 = \mu$ és $\beta_1 = 0$. A heteroszkedaszticitással kapcsolatos ökonometriai elemzések legegyszerűbb esetben egy ilyen regressziós modellből indulnak ki, de mivel a heteroszkedaszticitás a maradékváltozó jellemzője, az itt bemutatott tulajdonságok analógok az ökonometriából ismertekkel, értelemszerű általánosítások mellett.

Nem foglalkoztunk a heteroszkedaszticitás egy sor egyéb vonatkozásával, így azzal, hogy miként lehet kimutatni (grafikusan, tesztekkel), azzal, hogy miként lehet becsülni, és a becslésnek milyen tulajdonságai vannak akkor, ha a σ_i -k nem ismertek, milyen hatással vannak a különféle transzformációk a heteroszkedaszticitás kezelésére stb. Reméljük azonban, hogy elértük azt az alapvető célt, hogy egyszerű, könnyen áttekinthető feladaton, jórészt elemi eszközök segítségével megmutassuk a heteroszkedaszticitás jelenlétében alkalmazott és alkalmazható becslések tulajdonságait.

Irodalom

HUNYADI L. [2000]: A kétmintás *t*-próbáról. In: *Fél évszázad statisztika szolgálatában*. Tanulmánykötet Köves Pál tiszteletére. Központi Statisztikai Hivatal. Budapest.

HUNYADI L. [2001]: *Statisztikai következtetéselmélet közgazdászoknak.* Statisztikai módszerek a társadalmi és a gazdasági elemzésekben. Központi Statisztikai Hivatal. Budapest.

KŐRÖSI G. – MÁTYÁS L. – SZÉKELY I. [1990]: *Gyakorlati ökonometria*. Közgazdasági és Jogi Könyvkiadó. Budapest.

MADDALA, G. S. [2004]: *Bevezetés az ökonometriába*. Közgazdasági tankönyvek. Nemzeti Tankönyvkiadó. Budapest.

RAMANATHAN , R. [2003]: *Bevezetés az ökonometriába alkalmazásokkal*. Panem Kiadó. Budapest.