Statisztikai Szemle

Közzététel: 2019. március 8.

A tanulmány címe:

Logisztikus regressziós együtthatók összehasonlítása

Szerzők:

Bartus Tamás, a Budapesti Corvinus Egyetem egyetemi tanára

E-mail: tamas.bartus@uni-corvinus.hu;

Kisfalusi Dorottya, a Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpont Szociológiai Intézetének tudományos segédmunkatársa

E-mail: kisfalusi.dorottya@tk.mta.hu;

Koltai Júlia, a Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpont Szociológiai Intézetének tudományos munkatársa, az Eötvös Loránd Tudományegyetem egyetemi adjunktusa

E-mail: koltai.julia@tk.mta.hu

DOI: 10.20311/stat2019.3.hu0221

Logisztikus regressziós együtthatók összehasonlítása

Bartus Tamás,
a Budapesti Corvinus Egyetem egyetemi tanára
E-mail: tamas.bartus@uni-corvinus.hu

Kisfalusi Dorottya,
a Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpont Szociológiai Intézetének tudományos segédmunkatársa
E-mail: kisfalusi.dorottya@tk.mta.hu

Koltai Júlia,
a Magyar Tudományos Akadémia Társadalomtudományi Kutatóközpont Szociológiai Intézetének tudományos munkatársa,
az Eötvös Loránd Tudományegyetem egyetemi adjunktusa
E-mail: koltai.julia@tk.mta.hu

Az utóbbi években egyre több figyelmet kapott az a probléma, hogy az egymásba ágyazott modellspecifikációkban szereplő, illetve a különböző részmintákra vonatkozó logisztikus (és más nemlineáris) regressziós együtthatók nem hasonlíthatók össze, mivel a különböző modellspecifikációkban és a különböző részmintákban eltér a nem megfigyelhető reziduális szórás. A tanulmányban a szerzők bemutatják a probléma megoldására kidolgozott módszereket és szimuláció segítségével vizsgálják azok hatékonyságát. Az egymásba ágyazott modellek együtthatói összehasonlíthatóvá tehetők az együtthatók y-standardizálásával vagy a többváltozós modell (nem közvetlenül az egyváltozós modellhez, hanem) egy speciális, kváziegyváltozós modellhez való hasonlításával. A különböző részmintákra vonatkozó becslések összehasonlítására kidolgozott módszerek – a csoport-interakciók arányosságának tesztelése, valamint a heterogenitást tartalmazó logisztikus regressziós modellek – azonban nem adnak érdemi megoldást a problémára. A tanulmány ennek a kudarcnak az elemzésével zárul.

TÁRGYSZÓ:
Logisztikus regressziós együtthatók.
y-standardizálás.
KHB-módszer.

DOI: 10.20311/stat2019.3.hu0221

Az utóbbi években egyre több figyelmet kapott az a probléma, hogy a logisztikus (és más nemlineáris) regressziós modellek együtthatói sok tekintetben különböznek a lineáris regressziós modellek együtthatóitól. Ennek következtében egyes kutatási kérdések megválaszolása nemlineáris modellek használata esetén bonyolultabb, mint a lineáris regresszióelemzés esetén. Lineáris regressziós modellek becslésével például vizsgálni tudjuk, hogy eltérő-e egy-egy magyarázó változó függő változóra gyakorolt hatása a megfigyelési egységek különböző csoportjaiban. A paraméterbecslések alapján így többek között össze tudjuk hasonlítani, hogy az iskolai végzettség miként hat a fizetésekre különböző társadalmi csoportokban (például a férfiak és a nők tekintetében), különböző országokban vagy különböző adatfelvételi időpontokban. Ugyanez az összehasonlítás logisztikus regressziós modellekben problémás lehet, ha a reziduális variancia mértéke társadalmi csoportonként, országonként vagy időpontonként különbözik (*Allison* [1999], *Keele–Park* [2005], *Long–Mustillo* [2017], *Mood* [2010], *Williams* [2009]).

Továbbá a lineáris regressziós modellekkel szemben a logisztikus regressziós modellek paraméterbecsléseit abban az esetben is befolyásolják a modellből kihagyott magyarázó változók, ha azok függetlenek a modellbe bevont magyarázó változóktól (*Mood* [2010]). Emiatt az egymásba ágyazott modellek[1] (*nested models*) paraméterbecsléseinek összehasonlításával kevésbé tudunk az összefüggések mögött álló magyarázó mechanizmusokra következtetni. Tegyük fel például, hogy az előléptetések során tapasztalható nemi különbségeket szeretnénk magyarázni. Ha lineáris regresszióelemzés során azt látnánk, hogy az iskolai végzettség bevonásával nő/csökken a nem hatása a függő változóra, arra következtethetnénk, hogy a nők és a férfiak összetétele eltér az iskolai végzettségüket tekintve, és a nemi különbségeket részben elfedi/magyarázza az iskolai végzettségben tapasztalható összetételbeli különbség. Logisztikus regresszióelemzés során a nem változó paraméterbecslésének két modell közötti összehasonlításából nem következtethetünk egyértelműen ugyanezekre a mechanizmusokra, mivel további magyarázó változók bevonásával akkor is növekedhet egy változó paraméterbecslése, ha a magyarázó változók között nincs összefüggés.

A tanulmányban először azt ismertetjük, hogy a paraméterbecslések összehasonlítása során mi okozza az eddig bemutatott nehézségeket. A probléma gyökere az, hogy a logisztikus regressziós együtthatók egy látens változós modellben értelmez-

[1] Egymásba ágyazott modellek alatt azt értjük, amikor egy modellbe fokozatosan egyre több magyarázó változót építünk be, így az újabb modellek tartalmazzák azokat a változókat is, amelyek a korábbi modellünkben szerepeltek.

hető strukturális együtthatók[2] és a reziduális szórás hányadosai. Az alapvető identifikációs probléma az, hogy a strukturális hatások és a reziduális szórás külön-külön nem becsülhetők. Amit becsülni tudunk, csak ennek a két tényezőnek a hányadosa. Eltérő modellspecifikációkban, illetve mintákban a reziduális szórás is eltér, emiatt az ilyenkor kapott becslések nem tükrözik a strukturális hatások különbségeit. A tanulmány fő kérdése természetesen az, hogy milyen módszerekkel lehet „közös nevezőre hozni" a logisztikus regressziós együtthatókat annak érdekében, hogy következtetni tudjunk a strukturális együtthatók valós eltérésére.

Ezt követően az egymásba ágyazott modellek összehasonlításával foglalkozunk. Az együtthatók összevetésére két módszert dolgoztak ki. Az y-standardizálás módszere (*Winship–Mare* [1984]) lehetővé teszi a reziduális szórások hányadosának becslését. Ezáltal a strukturális hatások hányadosa is becsülhetővé válik. A KHB-(Karlson–Holm–Breen-) módszer (*Karlson–Holm–Breen* [2012]) lényege abban áll, hogy a kontrollváltozókat felbontja egy szisztematikus és egy véletlen komponensre, a többváltozós becsléseket pedig a véletlen komponenseket tartalmazó, kváziegyváltozós becslésekkel hasonlítja össze. Mindkét módszer sikeresen megoldja az összehasonlíthatósági problémát, a KHB-módszer azonban valamivel pontosabb a másiknál.

Végül az eltérő részmintákra vonatkozó becslések összehasonlításával foglalkozunk. A különböző részmintákban becsült logisztikus regressziós együtthatók összehasonlítása első látásra egyszerűnek tűnik, ha a csoportspecifikus regressziók helyett a teljes mintára vonatkozó, csoport-interakciókkal kibővített regressziós modellt becsüljük. A csoportspecifikus hatások összehasonlítása ekkor az interakciós hatások előjelének és szignifikanciájának megítéléséből áll. Az alapvető identifikációs probléma miatt azonban a csoport-interakciók jelenléte nemcsak a strukturális hatások eltérésével magyarázható, hanem azzal is, hogy a csoportokban eltér a reziduális szórás. A problémára kétfajta megoldás ismert. Az elsőt *Allison* [1999] javasolta, és statisztikai tesztek végzésével jár. A másik pedig a *Williams* [2009] által kidolgozott, heterogenitást tartalmazó logisztikus regressziós modell. A két eljárás bemutatása után amellett érvelünk, hogy ezek a módszerek nem adnak valódi megoldást a problémára. Majd a tanulmány végén azt próbáljuk értelmezni, hogy miért nem oldható meg az összehasonlíthatósági probléma az eltérő részmintákra vonatkozó modellek esetében.

[2] Strukturális együtthatónak nevezzük azokat az együtthatókat, amelyek megmutatják, hogy miként befolyásolják a megfigyelt magyarázó változók a látens függő változó által leírt jelenséget.

1. A probléma

A logisztikus regressziós modellnek két megfogalmazása létezik (*Buis* [2016], *Hajdu* [2004]). Az egyszerűbb és valószínűleg népszerűbb megfogalmazás egy valószínűségi modell:

$$\text{logit}\left(y = 1\right) = \beta x\,, \qquad /1/$$

ahol y a diszkrét függő változó ($y = 1$ a „siker”, $y = 0$ a „kudarc”), a logit függvény az argumentumban szereplő esemény esélyének logaritmusa, βx pedig az együtthatók és a magyarázó változók lineáris kombinációja. A β együtthatók a logisztikus regresszió paraméterei. A jobb oldalon ugyan nem szerepel hibatag, a megfigyelt kimenet és a magyarázó változók közötti kapcsolat sztochasztikus jellegére az utal, hogy a bal oldalon nem a megfigyelt kimenet, hanem a bekövetkezési esély logaritmusa szerepel.

A másik, látens változós megfogalmazás (*Fülöp* [2002]) abból a feltevésből indul ki, hogy létezik egy folytonos változó, y^*, amely a szisztematikus és a véletlen komponensek összege:

$$y^* = \alpha x + \sigma\varepsilon\,, \qquad /2/$$

ahol αx a magyarázó változók és az együtthatók lineáris kombinációja (tehát $\alpha x = \sum \alpha x$), ε a nulla várható értékű és $\pi/\sqrt{3}$ szórású eloszlást követő (standard) logisztikus eloszlásfüggvény[3], σ pedig a reziduum tényleges szórása. Fontos felhívni a figyelmet arra, hogy a reziduális szórás mértéke nem 1, hanem $\pi/\sqrt{3}$. A σ paraméter egész pontosan azt fejezi ki, hogy a tényleges reziduális szórás hányszorosa a standard szórásának. Az α együtthatókat strukturális együtthatóknak szokás nevezni, mivel ezek ragadják meg, hogy miként befolyásolják a megfigyelt magyarázó változók azt a jelenséget (például attitűdöt, hajlamot, hasznosságot), amiről a folytonos függő változó segítségével gondolkodunk.

A folytonos függő változót azonban nem figyeljük meg, az „pusztán” elméleti konstrukció, aminek segítségével a diszkrét függő változó bekövetkezése levezethető. A megfigyelt diszkrét függő változó (sikeres-e: igen/nem) és a látens y^* változó (sikerre való hajlam) között a mérési modell teremt kapcsolatot:

$$\begin{aligned} y &= 1, \text{ ha } y^* > 0; \\ y &= 0, \text{ ha } y^* \leq 0. \end{aligned} \qquad /3/$$

[3] Fontos hangsúlyzni, hogy a standard normális eloszlással ellentétben a standard logisztikus eloszlásfüggvény szórása nem egységnyi (lásd például *Hunyadi* [2004]).

A /2/ egyenletben definiált látens változós lineáris regressziós modell és a /3/ egyenletben definiált mérési modell a következőképpen határozza meg a siker valószínűségét:

$$P(y = 1) = P(\alpha x + \sigma\varepsilon > 0) = P\left(\varepsilon > -\alpha x/\sigma\right) = 1 - F\left(-\alpha x/\sigma\right).$$

A logisztikus eloszlásfüggvény szimmetrikus, a siker valószínüsége tehát:

$$P(y = 1) = F(\alpha x/\sigma). \qquad /4/$$

Mivel az /1/ és /4/ egyenletek ugyanazt a valószínűséget definiálják, a közvetlen valószínűségi modell β együtthatói és a látens változós lineáris regressziós modell α együtthatói között a következő kapcsolat áll fenn (*Allison* [1999]):

$$\beta = \alpha/\sigma. \qquad /5/$$

Az /5/ egyenlet azt állítja, hogy a logisztikus regressziós együtthatók nem azonosak a látens változós lineáris regresszió együtthatóival. Másképp fogalmazva: a logisztikus regresszió képtelen a strukturális együtthatók identifikálására, mivel a becsülhető β együtthatók a strukturális α együtthatók és az ismeretlen σ szórás hányadosai.

A logisztikus regresszió látens változós értelmezése azt vonja maga után, hogy a logisztikus valószínűségi modell együtthatói nem értelmezhetők úgy, mint a lineáris regressziós modell együtthatói, hiszen – a lineáris regressziós együtthatóktól eltérően – a logisztikus regressziós együtthatók a strukturális együtthatók és az ismeretlen szórások hányadosai. Ebből az következik, hogy ugyanannak a változónak eltérő modellekben vagy eltérő mintákban számolt logisztikus regressziós becsléseinek különbsége:

$$\beta_1 - \beta_2 = \alpha_1/\sigma_1 - \alpha_2/\sigma_2. \qquad /6/$$

Ez a skálatorzítás problémája: a becslésekben mutatkozó pozitív (vagy negatív) különbség nem vonja maga után, hogy a strukturális együtthatók különbsége is pozitív (vagy negatív). Könnyen belátható, hogy az $\alpha_1 > \alpha_2$ egyenlőtlenségből csak akkor következik a $\beta_1 > \beta_2$ egyenlőtlenség, ha teljesül az

$$\alpha_1/\alpha_2 > \sigma_1/\sigma_2 \equiv s \qquad /7/$$

egyenlőtlenség (s a nem megfigyelt reziduális szórások hányadosa).

A probléma két összefüggésben merül fel. Az egyik egy adott változó különböző, egymásba ágyazott specifikációkban becsült együtthatóinak összehasonlítása. Vegyük az alábbi modelleket:

$$y^* = \alpha_1 x + \varepsilon_1;$$
$$y^* = \alpha_2 x + \gamma z + \varepsilon_2.$$

Lineáris regressziónál az $\alpha_1 - \alpha_2$ különbség rendszerint zérustól eltérő, aminek az elhanyagolt változókból fakadó torzítás az oka: az, hogy az *x* és a *z* változók valamilyen mértékben korrelálnak egymással, és a *z* kontrollváltozók hatása zérustól eltérő.

Logisztikus regressziónál az $\alpha_1 - \alpha_2$ különbség helyett csak a /6/ egyenletben definiált $\beta_1 - \beta_2 = \alpha_1/\sigma_1 - \alpha_2/\sigma_2$ különbséget tudjuk megfigyelni. Ha az egy- és a többváltozós egyenletben azonos lenne a reziduális szórás, a logisztikus regressziós együtthatók különbségét ismét csak az elhanyagolt változókból fakadó torzítással magyarázhatnánk. A reziduális szórások azonban nem azonosak, mivel az egyváltozós modellben a *z* változó nem a szisztematikus komponensnek, hanem a hibatagnak a része. Ha tehát a *z* változók együtthatói eltérnek nullától, akkor teljesülnie kell a $\sigma_1 > \sigma_2$ egyenlőtlenségnek. Emiatt lehetséges, hogy a $\beta_1 - \beta_2$ különbséget alapvetően nem az elhanyagolt változókból fakadó torzítás (az $\alpha_1 - \alpha_2$ különbség), hanem a reziduális szórások nem megfigyeltő különbsége magyarázza.

Logisztikus regressziós modellekben adott *x* magyarázó változó paraméterbecslése tehát abban az esetben is változik, ha a modellhez további, *x*-szel nem korreláló változókat adunk hozzá. A korreláció hiánya miatt a strukturális együtthatók különbsége nem változik. Az új változók hozzáadásával viszont csökken a reziduális szórás, ami az /5/ egyenlet alapján növeli az identifikálható logisztikus regressziós együttható abszolút értékét. A logisztikus regressziós modellek paraméterbecsléseit tehát akkor is befolyásolják a modellből kihagyott magyarázó változók, ha azok függetlenek a modellbe bevont magyarázó változóktól. Előfordulhat tehát, hogy az egyváltozós modell becsült együtthatója akkor is kisebb a többváltozós modellben szereplő becsült együtthatónál, ha a további bevont magyarázó változók függetlenek a közös *x* változótól (*Mood* [2010]).

A másik kontextus a különböző részmintákban kapott becslések összehasonlítása. Ha két csoportban (például a férfiak és a nők között) eltérést is találunk egy független változó hatása között, nem lehetünk abban biztosak, hogy ténylegesen van eltérés, hiszen a két alcsoport modelljeiben más és más lehet a meg nem magyarázott rész (*Allison* [1999]). Például csupán a regressziós együtthatók nagysága alapján nem mondhatjuk, hogy a férfiaknál az iskolázottság jobban számít a siker elérésében, mint a nőknél, hiszen a két érték nem összevethető. Első látásra a probléma kiküszöbölhető lenne, ha az elemzést a férfiak és a nők összevont mintáján végeznénk, és a

modellbe bevonnánk a nem és a vizsgált magyarázó változó interakcióját (*Allison* [1999]). Ez a megoldási javaslat azonban figyelmen kívül hagyja azt – a tanulmányunk 3. fejezetében részletesen kifejtett – problémát, hogy az interakciós hatás a reziduális szórásban tapasztalható, nemek közötti különbséggel is magyarázható.

2. Az egymásba ágyazott modellek együtthatóinak összehasonlítása

Az egymásba ágyazott modellek együtthatóinak összehasonlítására két módszert dolgoztak ki: az y-standardizálást (*Mood* [2010], *Winship–Mare* [1984]) és a KHB-módszert (*Karlson–Holm–Breen* [2012]).[4] A módszerek célja az, hogy az eltérő strukturális hatásokat „közös nevezőre hozzuk", azaz ugyanazzal a reziduális szórással skálázzuk. A tanulmányban később látni fogjuk, hogy a „közös nevezőre hozás" szoros kapcsolatban van egy dekompozícós eljárással. A becsült logisztikus regressziós együtthatók – /6/ egyenletben definiált – különbsége ugyanis felírható a következő formában:

$$\begin{aligned}\beta_1 - \beta_2 &= \left(\alpha_1 - \alpha_2\right)\big/\sigma_2 + \alpha_1\left(1/\sigma_1 - 1/\sigma_2\right) \\ &= \left(s\beta_1 - \beta_2\right) + \beta_1\left(1 - s\right),\end{aligned} \qquad /8/$$

ahol s a szórások hányadosa: $s = \sigma_1/\sigma_2$. Az együtthatók becsülhető különbségét tehát felbonthatjuk két tényező összegére. Az első tényező a strukturális együtthatók átskálázott különbsége: az $s\beta_1$ együttható skálája ugyanis megegyezik a β_2 együtthatóéval. A második tényező az eltérő skálák különbségét türközi, azaz felfogható a skálatorzítás mérőszámaként. A probléma megoldásának kulcsa tehát az $s = \sigma_1/\sigma_2$ hányados becslése.

[4] A magyarázó változók különböző modellek közötti összehasonlításának egy további, valószínűségek összehasonlításán alapuló lehetséges eszköze a marginális hatások számítása. Logisztikus regressziós modellekben a hatások nagyságát értelmezhetjük oly módon, hogy a látens haszon, illetve a siker valószínűségének változását vizsgáljuk (*Angrist* [2001]). Az esélyhányadossal szemben azonban egy magyarázó változó hatása a siker valószínűségére nem konstans, és többféle mérőszámmal is megragadható (*Long* [1997]). Az egyik megoldás a marginális hatások számítása. Egy változó marginális hatása a feltételes várható értékek különbségeit jelenti, és az esélyhányadossal szemben a marginális hatások számításakor nem csupán az adott változó paraméterbecslését vesszük figyelembe, hanem a többi változó értékét és paraméterbecslését is (*Bartus* [2003]). Az átlagos marginális hatások összehasonlíthatók, azok értékét nem befolyásolja jelentősen a skálatorzítás, amely a modellbe bevont magyarázó változókkal nem korreláló változók kihagyásából fakad (*Mood* [2010]).

2.1. Az y-standardizálás

Az y-standardizálás módszerét *Winship* és *Mare* [1984] javasolta a skálatorzítás kiküszöbölésének céljából. A módszer részletes kifejtése *Karlson* [2015] tanulmányában található. Képzeljük el, hogy a látens változós modellt lineáris regresszióval becsüljük, a regressziós becsléseket – a strukturális együtthatókat – pedig elosztjuk a függő változó szórásával. Ez a művelet az y-standardizálás. Mivel a függő változó varianciája a szisztematikus és a véletlen komponensek varianciáinak összege, a *j*-edik magyarázó változó y-standardizált strukturális együtthatója:

$$\alpha_j \Big/ \sqrt{Var(\alpha x) + \sigma^2 \pi^2 / 3}. \qquad /9/$$

A σ paraméter most explicit módon azt fejezi ki, hogy hányszorosa a tényleges reziduális szórás a $\pi/\sqrt{3}$ mértékegységnek.

A logisztikus regressziós együtthatók a lineáris regresszióval becsült strukturális együtthatók és a σ hányadosai. Ha a /9/-ben szereplő tört számlálóját és nevezőjét is megszorozzuk σ-val, majd – az /5/ egyenletet követve – az α együtthatókat a $\sigma\beta$ szorzatra cseréljük, akkor a σ-val való egyszerűsítés után a

$$\beta_j \Big/ \sqrt{Var(\beta x) + \pi^2 / 3} \qquad /10/$$

hányadost kapjuk, amelyben a β együtthatók a logisztikus regressziós együtthatók. A hányadosban már nem szerepel σ, ezért az a becsült együtthatókból könnyen kiszámolható. A skálafüggő nyers logisztikus regressziós becslésektől eltérően az y-standarizált együtthatók függetlenek a nem megfigyelt reziduális szórástól. Az y-standardizálás módszere tehát lehetővé teszi a logisztikus regressziós együtthatók összehasonlítását.

Tételezzük fel, hogy az egymásba ágyazott modelleket ugyanazon a mintán becsüljük. Emiatt a látens függő változó varianciája mindegyik modellspecifikációban változatlan marad. Az egyszerűség kedvéért vegyünk egy egyváltozós és egy többváltozós regressziós modellt. Ha a látens változó megfigyelhető lenne, e modellek lineáris regressziós becslése után teljesül a

$$Var(\alpha_1 x) + \sigma_1^2 = Var(\alpha_2 x + \gamma z) + \sigma_2^2 \qquad /11/$$

egyenlőség. A látens változókat azonban nem lehet megfigyelni. A logisztikus regressziós becsléket úgy kapjuk, ha a lineáris regressziós becsléseket elosztjuk a meg-

felelő szórással. Szorozzuk meg a /11/ egyenlet mindkét oldalát $\pi^2/3\sigma_2^2$ -tel! A szorzás és némi algebra után a következő egyenlőséget kapjuk:

$$s^2\left[\, Var\left(\beta_1 x\right) + \pi^2/3\right] = Var\left(\beta_2 x + \delta z\right) + \pi^2/3\,, \qquad /12/$$

ahol $s^2 = \sigma_1^2/\sigma_2^2$. A reziduális szórások hányadosának ismeretében a strukturális együtthatók hányadosa is becsülhetővé válik:

$$\alpha_1/\alpha_2 = s\beta_1/\beta_2\,. \qquad /13/$$

Továbbá *s* ismeretében a /8/ egyenletben szereplő dekompozíciót is el lehet végezni.

2.2. A KHB-módszer

A KHB-módszer a *z* kontrollváltozók felbontásával teszi összehasonlíthatóvá az egymásba ágyazott modellek eredményeit. Az eljárás a következő. Vegyük az

$$y^* = \alpha_1 x + \varepsilon_1 \text{ és az}$$

$$y^* = \alpha_2 x + \gamma z + \varepsilon_2$$

látens változós modelleket. A módszer központi gondolata az, hogy a $z_k = \delta_k x + r_k$ lineáris regresszióval a z_k kontrollváltozót két, egymástól független komponensre bonthatjuk: a $\delta_k x$ szisztematikus komponensre és az r_k reziduális komponensre. A dekompozíció segítségével a többváltozós modellt felírhatjuk az

$$\begin{aligned} y^* &= \left[\, \alpha_2 + \sum \gamma_k \delta_k \right] x + \sum \gamma_k r_k + \varepsilon_2 \\ &= \alpha_2^T x + \sum \gamma_k r_k + \varepsilon_2 \end{aligned} \qquad /14/$$

formában. A /14/ egyenlet olyan regressziós modellt definiál, amelyben az érdeklődés középpontjában álló *x* változó mellett nem a természetes, hanem az *x*-től „függetlenített" kontrollváltozók szerepelnek. A $\sum \gamma_k \delta_k$ szorzatösszeg az elhanyagolt változókból fakadó torzítást méri, az egyenlet tehát kifejezi, hogy az egy- és a többváltozós becslések különbségét ez a torzítás magyarázza.

A /14/ egyenletben szereplő modellt kváziegyváltozós modellnek nevezhetjük a következő ok miatt. Az *r* változók függetlenek bármelyik *x* változótól, ezért az α_2^T együttható akkor is konzisztensen becsülhető lineáris regresszióval, ha az *r* változók nem szerepelnek a modellben. Az *r* változók elhanyagolása az egyváltozós modellhez vezet, ezért $\alpha_2^T = \alpha_1$.

A kváziegyváltozós modell másik fontos tulajdonsága, hogy abban a többváltozós modell reziduuma szerepel. Emiatt nemcsak az igaz, hogy a kváziegyváltozós modell magyarázó változójának strukturális együtthatója azonos a valódi egyváltozós modellével, hanem az is, hogy a kváziegyváltozós modell logisztikus regressziós becslése az egyváltozós strukturális (látens változós) együttható és a többváltozós modellhez tartozó szórás hányadosa: $\beta_2^T = \alpha_1 / \sigma_2$.

A kváziegyváltozós logisztikus regressziós modell együtthatója és a többváltozós modell megfelelő együtthatója tehát ugyanazt a skálát, reziduális szórást feltételezi. A kváziegyváltozós és a többváltozós együtthatók ezáltal összehasonlíthatók, különbségük a /8/ egyenletben szereplő $\left(\alpha_1 - \alpha_2\right)/\sigma_2$ tényezővel azonos. Ezzel az összehasonlítással így becsülni tudjuk a logisztikus regressziós együtthatók skálatorzítástól mentes különbségét és az elhanyagolt változókból fakadó torzítás nagyságát.

2.3. A két módszer közötti választás

A strukturális együtthatók hányadosa, illetve a strukturális együtthatók átskálázott különbsége (lásd a /8/ egyenlet jobb oldalának első komponensét) elvileg egyaránt becsülhető az y-standardizálás és a KHB-módszereivel. *Karlson*, *Holm* és *Breen* [2012] Monte-Carlo-szimulációval hasonlították össze a két módszert, és azt találták, hogy az y-standardizálás kevésbé megbízató, különösen akkor, amikor a kontrollváltozók eloszlása eltér a normálistól. A szóban forgó szimulációs vizsgálatot replikáltuk, azaz megismételtünk négyet az általuk közölt öt szimulációs vizsgálatból. A szimulált adatbázis 5000 megfigyelést tartalmaz. A folytonos látens függő változót az

$$y^* = 1 + \beta x + \gamma z + u$$

egyenlettel hozzuk létre. Az első vizsgálatban $\beta = 0{,}5$, $\gamma = 1$, a magyarázó változók pedig normális eloszlást követnek. A második vizsgálat csak abban tér el az elsőtől, hogy a magyarázó változók eloszlása lognormális. A harmadik és a negyedik vizsgálat rendre az elsőt és a másodikat ismétli meg a $\beta = 1$, $\gamma = 0{,}5$ paraméterek

mellett. A két magyarázó változó egymással korrelálhat: a korrelációs együtthatók értékei 0, 0,1, 0,3, 0,5, 0,7 és 0,9. A diszkrét magyarázó változó 1, ha a látens változó nagyobb a 0,5, 0,7, 0,9 vagy 0,95 értéket felvevő küszöbértéknél. Mindegyik vizsgálathoz tehát 20 forgatókönyv tartozik. Ezek mindegyike esetén 250 alkalommal ismételtük meg a kísérletet.

Karslonék a becslési hiba abszolút értékeinek átlagával vizsgálták a megbízhatóságot, és azt találták, hogy ez az átlag saját módszerüknél alacsonyabb. Ezt az előnyt azzal magyarázták, hogy az y-standardizálás érzékenyebben reagál arra, ha a reziduális eloszlás eltér a feltételezett logisztikus eloszlástól. Az egyváltozós modell hibatagjának eloszlása lényegesen különbözik a logisztikustól, ha a kontrollváltozók eloszlása aszimmetrikus.[5] Mi emellett azt is vizsgáltuk, hogy adott esetben az y-standarizálással vagy a KHB-módszerrel kapott becslés közelíti-e meg jobban a valós értéket. Azt találtuk, hogy az esetek durván 90 százalékában a KHB-módszer ad jobb becslést.

3. A különböző mintákban becsült együtthatók összehasonlítása

A különböző mintákban becsült regressziós együtthatók összehasonlításának problémája lényegesen nehezebb, mint az egymásba ágyazott modellek együtthatóié. Ebben a fejezetben először a problémát fejtjük ki, majd bemutatjuk az annak megoldására tett kísérleteket: a statisztikai teszteken nyugvó eljárásokat, valamint a heterogenitást tartalmazó modellt (*Williams* [2009], [2010]). Amellett érvelünk, hogy ezek a megoldási javaslatok korántsem tekinthetők valódi megoldásnak.

3.1. A nem megfigyelt heterogenitásból fakadó probléma

A különböző mintákban becsült logisztikus regressziós együtthatók összehasonlítása első látásra egyszerűnek tűnik. A csoportspecifikus

$$\begin{aligned} \text{logit}(y = 1) &= \beta_0 x, \text{ ha } z = 0; \\ \text{logit}(y = 1) &= \beta_1 x, \text{ ha } z = 1 \end{aligned} \qquad /15/$$

[5] *Karlson–Holm–Breen* [2012] az aszimmetrikus eloszlást lognormális eloszlással modellezték. Azt találták, hogy az y-standardizálás akkor a legkevésbé megbízható, ha a kontrollváltozó eloszlása lognormális.

logisztikus regressziós modellek ekvivalensek a teljes mintára vonatkozó, interakciókkal kibővített

$$\text{logit}\left(y = 1\right) = \beta_0 x + \left(\beta_1 - \beta_0\right) xz \qquad /16/$$

logisztikus regressziós modellel. (A z indikátorváltozó a csoportokat azonosítja.) A csoportspecifikus hatások összehasonlítása ekkor az interakciós hatások előjelének és szignifikanciájának megítéléséből áll.

Ez az eljárás azonban hibás, mert az interakciós hatások nagyságát a nem megfigyelt heterogenitás is befolyásolja (*Allison* [1999]). A probléma forrása, hogy az interakciós hatások a reziduális szórások eltérésével is magyarázhatók. Vegyük a következő látens változós modellt:

$$\begin{aligned} y_0^* &= \alpha x + \sigma_0 \varepsilon_0, \text{ ha } z = 0; \\ y_1^* &= \alpha x + \sigma_1 \varepsilon_1, \text{ ha } z = 1. \end{aligned} \qquad /17/$$

A két csoportban megegyeznek a strukturális hatások. Az egyetlen eltérés, hogy a két csoportban eltérő a hibatag szórása, emiatt a csoportspecifikus regressziós modellekben szereplő szórások sem azonosak. Ez tehát egy heteroszkedaszticitást tartalmazó strukturális modell.

El lehet-e dönteni, hogy a logisztikus regressziós együtthatók csoportok közötti eltérése – vagy a teljes mintán becsült modellben az interakciós hatás nagysága – a strukturális különbséggel vagy a heterogenitással magyarázható-e inkább? A kérdést először *Allison* [1999] vizsgálta. A probléma friss tárgyalása *Tutz* [2018] tanulmányában található.

Induljunk ki abból a hipotézisből, hogy a csoportokban eltér a reziduális szórás, de a csoportok között nincsenek strukturális különbségek. Ezt a hipotézist fejezi ki a /17/ egyenletben szereplő modell. Tudjuk, hogy a logisztikus regressziós együtthatók a strukturális együtthatók és a szórások hányadosai. Tetszőleges szórás reciproka a

$$\sigma^{-1} = \sigma_0^{-1} \exp\left(-\gamma z\right) \qquad /18/$$

multiplikatív egyenlettel modellezhető. A /17/ modell és a /18/ egyenlet a következő, heteroszkedaszticitást tartalmazó logisztikus regressziós modellt definiálják:

$$\begin{aligned} \text{logit}\left(y = 1\right) &= \beta_0 x, && \text{ha } z = 0; \\ \text{logit}\left(y = 1\right) &= \beta_0 x \exp\left(-\gamma\right), && \text{ha } z = 1, \end{aligned} \qquad /19/$$

ahol $\beta_0 = \alpha/\sigma_0$. A $\lambda = \exp(-\gamma)$ jelölést bevezetve a teljes mintára érvényes modell:

$$\text{logit}(y = 1) = \beta_0 x + (\lambda - 1)\beta_0 xz. \qquad /20/$$

A heteroszkedaszticitási feltevésből levezetett logisztikus regressziós modell is tartalmaz interakciós hatásokat.

A probléma tehát a következő. A magyarázó változók és a csoportot azonosító indikátorváltozó összes interakcióját tartalmazó logisztikus regressziós modell kétféleképpen értelmezhető: *1*. a csoportokban eltérő a magyarázó változók oksági hatása, de azonos a hibatagok szórása (*strukturális értelmezés*); illetve *2*. a csoportokban azonosak az oksági hatások, de eltérnek a reziduális varianciák (*heterogenitáson alapuló értelmezés*). A következő kérdés az, hogy miként választhatunk a riválisértelmezések közül.

3.2. A strukturális és a heterogenitáson alapuló értelmezések közötti választás

A strukturális értelmezés és a heterogenitáson alapuló értelmezés közötti választás abból a felismerésből indul ki, hogy a heterogenitási feltevésből levezetett, /20/ egyenletben szereplő modell valójában nem azonos a strukturális eltérésből levezetett, /16/ egyenletben szereplővel. Az előbbi ugyanis az utóbbi speciális esete (*Allison* [1999], *Rohwer* [2015], *Tutz* [2018]). A modell abban az értelemben speciális, hogy mindegyik x_j strukturális változóra igaz, hogy az $x_j z_k$ szorzatváltozó együtthatójának és az x_j változó együtthatójának a hányadosa konstans. A /16/ és a /20/ egyenletek kapcsolatát a

$$\beta_1 = \lambda\beta_0 \qquad /21/$$

azonosság írja le.

A strukturális értelmezés és a heterogenitáson alapuló értelmezés közötti választás logikája a következő (*Allison* [1999], *Williams* [2009], *Tutz* [2018]). Induljunk ki abból a nullhipotézisből, hogy a heterogenitáson alapuló értelmezés igaz, a strukturális értelmezés pedig hamis. Tehát a strukturális hatások mindkét részmintában azonosak, a csoportspecifikus reziduális szórások viszont eltérnek. Ez a nullhipotézis a /21/ egyenletben szereplő azonosságot vonja maga után, azaz mindegyik x_j strukturális változóra igaz, hogy az $x_j z_k$ szorzatváltozó együtthatójának és az x_j változó együtthatójának a hányadosa konstans. Ez egy, a paraméterekre vonatkozó nemlineá-

ris korlátozó feltevés. A strukturális és a heterogenitáson alapuló értelmezések közötti választás e korlátozó feltevés vizsgálatán alapul.

Hogyan vizsgálható ez a korlátozó feltevés? Az egyik lehetséges eljárás az, hogy a /16/ és a /20/ modellt egyaránt becsüljük, majd likelihoodarány-teszttel vizsgáljuk a /21/ egyenletben szereplő korlátozó feltevés hihetőségét (*Allison* [1999]).[6] Egy alternív eljárás, ha a /16/ egyenletben szereplő modell becslése után nemlineáris Wald-teszttel vizsgáljuk azt a hipotézist, miszerint a magyarázó változók interakciós és főhatásainak hányadosai megegyeznek.[7] *Tutz* [2018] egy ennél erősebb követelményt javasol: ha a heterogenitást tartalmazó modell a helyes modell, akkor a modellben az összes interakciós hatásnak szignifikánsnak kell lennie.

A szignifikanciateszteken alapuló választásnak két problémája van. Egyrészt az, hogy a nullhipotézis feltételezése szerint a strukturális hatások nem térnek el a két csoportban. Ha ez a feltevés sérül, akkor az interakciós hatásokat tévesen tulajdonítjuk a heterogenitás hatásának (*Williams* [2009]). Másrészt kérdéses a teszt statisztikai ereje, hiszen a szorzatváltozók és a szorzáshoz használt változók közötti multikollinearitás miatt várhatóan alacsony lesz a tesztstatisztika értéke. A szignifikanciateszt logikája miatt ekkor arra fogunk következtetni, hogy a nullhipotézist nem lehet elvetni. Ez pedig azt jelenti, hogy a magyarázó változók és a csoportképző változó interakciója nemcsak a strukturális eltérés, hanem a heterogenitási feltevéssel is magyarázható. Ez a következtetés természetesen a kiinduló problémát ismétli meg, de nem oldja meg azt. A probléma tehát csak nagy mintákban oldható meg.

3.3. A heterogenitást tartalmazó logisztikus regressziós modell

A strukturális értelmezés és a heterogenitáson alapuló értelmezés közötti választás korábban említett problémái motiválták a heterogenitást tartalmazó logisztikus regressziós modell kidolgozását (*Williams* [2009]). A modell egyrészt egyesíti a strukturális és a heterogenitáson alapuló értelmezéseket: a csoportok között tehát eltérhet mind a magyarázó változók oksági hatása, mind pedig a reziduumok szórása. Ennek a gondolatnak a kézenfekvő modellje:

$$\begin{aligned} y_0^* &= \alpha_0 x + \sigma_0 \varepsilon_0, \text{ ha } z = 0; \\ y_1^* &= \alpha_1 x + \sigma_1 \varepsilon_1, \text{ ha } z = 1. \end{aligned} \qquad /22/$$

[6] *Allison* [1999] a /20/ egyenletben szereplő modell becslésének technikai részleteivel is foglalkozik. A technikai probléma forrása az, hogy a maximumlikelihood-becslések nem engedik meg a paraméterekben a nemlináris korlátozó feltevéseket. A megoldás egy olyan likelihood-függvény maximalizálása, amelyben az együtthatók és a magyarázó változók nemlineáris, $(\alpha x)(\gamma z)$ formában megadható kombinációja szerepel.

[7] A Stata statisztikai programcsomagban ez a módszer könnyen kivitelezhető a nemlináris hipotézisek tesztelésére szolgáló *testnl* parancs segítségével.

Másrészt a reziduális szórás nemcsak a csoportképző változó, hanem más, akár a strukturális egyenlekben szereplő változókkal is modellezhető. A heterogenitást tartalmazó modell általános megfogalmazása:

$$\begin{aligned} \text{logit}(y = 1) &= (\alpha x)\sigma^{-1}, \\ \sigma^{-1} &= \exp(\gamma z). \end{aligned} \qquad /23/$$

A modell tehát két részre oszlik: az additív strukturális egyenletre (αx), és a multiplikatív variancia egyenletére $\left(\exp(\gamma z)\right)$, ami az ismeretlen szórást – a reziduumok varianciáját – modellezi. A varianciaegyenletben nem szerepel konstans, így a reziduális variancia egységnyi abban a részmintában, ahol az összes z változó zérus. A modell tehát beépíti az egyenletbe a reziduumok szórásának különbségére vonatkozó előfeltevést, és próbálja megbecsülni azt. A regressziós együtthatók nagyságának kiszámításakor így kompenzálja a minta alcsoportjainak különböző reziduális szórását, és ezáltal összehasonlíthatóvá teszi az alcsoportok együtthatóit (*Williams* [2009]).

Vajon tényleg megoldja ez a modell a strukturális és a heterogenitáson alapuló értelmezés problémáját? A kifejtés egyszerűségének kedvéért vegyük azt a problémát, hogy egy x változó hatásait szeretnénk összehasonlítani. Ha megengedjük, hogy x hatása és a reziduális variancia is eltérjen a csoportok között, akkor a következő regressziós modellt kapjuk:

$$\text{logit}(y = 1) = (\alpha_0 + \alpha_z z + \alpha_x x + \alpha_{xz} xz)\exp(\gamma_z z). \qquad /24/$$

A probléma az, vajon identifikálható-e az összes együttható. A modellben ugyanis 5 együttható szerepel, eggyel több, mint a közismerten identifikálható

$$\text{logit}(y = 1) = \beta_0 + \beta_z z + \beta_x x + \beta_{xz} xz$$

logisztikus regressziós modellben. A heterogenitást tartalmazó logisztikus regressziót kidolgozó tanulmányok (*Alvarez–Brehm* [1995]; *Williams* [2009], [2010]) nem foglalkoztak az identifikáció kérdéskörével. Egyértelműnek tűnik, hogy az együtthatók identifikálását a nemlineáris függvényforma teszi lehetővé – és nem valamilyen pótlólagos változó. A függvényformán alapuló identifikálás probléma, hiszen az nem megfigyelésre, hanem elméleti feltevésre támaszkodik.

3.4. Szimulációs eredmények

Az előző alfejezetben amellett érveltünk, hogy a strukturális és a heterogenitáson alapuló értelmezések szétválasztására a heteroszkedaszticitást explicit módon modellező, heterogenitást tartalmazó modellek sem alkalmasak. Állításunkat most egy szimulációs vizsgálattal támasztjuk alá. Elemzésünkben 1000 megfigyelésből álló, szimulált adatbázisokon vizsgáljuk az interakciós és a heterogenitást tartalmazó modellek becsléseit. Az adatbázisban először két egymással korreláló változót, *x*-et és *z*-t, illetve egy ezektől független *u* változót hozunk létre. A korrelációs együttható mindegyik szimulációban 0,2. A változók nullaátlagú és egységnyi szórású normális eloszlást követnek. Az *u* változót $\pi/\sqrt{3}$ szórású logisztikus eloszlású változóvá alakítjuk. Végül a *z* változóból létrehozzuk a diszkrét *D* változót: $D = 1(z > 0)$.

A folytonos látens függő változót az

$$y^* = -2 + aD + 0{,}25x + bDx + (sD + 1)u \qquad /25/$$

egyenlet adja. A megfigyelt diszkrét függő változó a látens változó dichotomizált változata. Az *a* és a *b* paraméterek a valódi interakciókat, az *s* paraméter pedig a heterogenitást (heteroszkedaszticitást) modellezi. A szimulációs vizsgálat során *a* 0 és 0,5, *b* 0 és 0,25, *s* pedig 0, 1 és 2 értékeket vesz fel.[8] A 8 különböző forgatókönyvben tehát vannak olyan esetek, amelyekben az adatok csak strukturális interakciókat tartalmaznak $(a = 0{,}5;\ b = 0{,}25;\ s = 0)$, illetve olyanok is, amelyekben az adatokat csak a heterogenitás jellemzi $(a = 0;\ b = 0;\ s = \{1,\ 2\})$. A szimulációt mindegyik forgatókönyv esetén 500 alkalommal végeztük el.

Mindegyik szimulációnál két modellt becsültünk. Az egyik a strukturális interakciókat tartalmazó

$$\text{logit}(y = 1) = \beta_0 + \beta_D D + \beta_x x + \beta_{Dx} Dx \qquad /26/$$

logisztikus regressziós modell, a másik a heterogenitást tartalmazó

$$\text{logit}(y = 1) = \left[\alpha_0 + \alpha_x x\right] / \exp(\gamma_D D) \qquad /27/$$

modell. A szimuláció során azt vizsgáltuk, mennyire térnek el a *D* és a *Dx* változók együtthatói azoktól az együtthatóktól, amelyeket akkor kapnánk, ha a látens változó

[8] Az értékek kiválasztásakor nem törekedtünk arra, hogy a kísérlet valamilyen tartalmi probléma kontextusában valósághű legyen. Mindenesetre a választott értékek kisebbek, mint amelyeket az egymásba ágyazott együtthatók szimulációs vizsgálatakor használtunk.

megfigyelhető lenne, az adatokat lineáris regresszióval elemeznénk, a lineáris regressziós becsléseket pedig elosztanánk a hibatag szórásával. A szokásos logisztikus regressziónál a *D* és a *Dx* együtthatóit közvetlenül, a heterogenitást tartalmazó regressziónál pedig e becsléseket úgy kapjuk meg, hogy az α_0, illetve az α_x együtthatókat megszorozzuk az $\exp(\gamma_D) - 1$ mennyiséggel.

Átlagos torzítások az interakciós változókat és a heterogenitást tartalmazó logisztikus regresziós modellekben

Szimuláció paramétere	Csoportképző változó főhatása		Csoportképző és magyarázó változók interakciója	
	logit	hetlogit	logit	hetlogit
a = 0; *b* = 0; *s* = 0	0,013	0,012	–0,006	–0,001
a = 0; *b* = 0; *s* = 1	1,018	0,017	–0,144	–0,130
a = 0; *b* = 0; *s* = 2	1,341	1,341	–0,176	–0,178
a = 0; *b* = 0,3; *s* = 0	0,007	0,020	0,124	–0,141
a = 0; *b* = 0,3; *s* = 1	1,012	1,010	–0,088	–0,257
a = 0; *b* = 0,3; *s* = 2	1,352	1,349	–0,140	–0,263
a = 0,5; *b* = 0,0; *s* = 0	0,224	0,224	–0,000	–0,071
a = 0,5; *b* = 0,0; *s* = 1	1,090	1,089	–0,122	–0,175
a = 0,5; *b* = 0,0; *s* = 2	1,386	1,385	–0,171	–0,201
a = 0,5; *b* = 0,3; *s* = 0	0,223	0,224	0,110	–0,249
a = 0,5; *b* = 0,3; *s* = 1	1,081	1,075	–0,079	–0,307
a = 0,5; *b* = 0,3; *s* = 2	1,381	1,378	–0,155	–0,287

Az eredmények azt mutatják, hogy egyik módszer sem képes torzítatlanul becsülni azokat a hatásokat, amelyeket akkor kapnánk, ha a látens változó megfigyelhető lenne, az adatokat lineáris regresszióval elemeznénk, a lineáris regressziós becsléseket pedig elosztanánk a hibatag szórásával. A szokásos logisztikus regressziós módszer torzítása nem szolgálhat meglepetéssel, hiszen ez a módszer csak az interakciós hatásokkal tudja reprezentálni a reziduális szórások heteroszkedaszticitását. Az viszont talán már joggal kelthet meglepetést, hogy a heterogenitást tartalmazó modell szintén torzított becsléseket ad, és nem képes csökkenteni a torzítások mértékét. Amikor a csoporttagság a strukturális hatást és a reziduális szórást is befolyásolja, a heterogenitást tartalmazó logisztikus regresszió mindig nagyobbat téved, mint az egyszerűbb, interakciókat tartalmazó logisztikus regresszió.

4. Összefoglalás

A tanulmányban azt vizsgáltuk, hogy miként lehet összehasonlíthatóvá tenni, „közös nevezőre hozni" akár az egymásba ágyazott modellekben szereplő, akár az eltérő részmintákból becsült logisztikus regressziós együtthatókat. A közös nevezőre hozás egyrészt azért szükséges, mert a logisztikus regressziós együttható egy – a látens változós modellben értelmezhető – strukturális együttható és a reziduális szórás hányadosa; másrészt, mert a különböző részmintákban vagy specifikációkban eltér a reziduális szórás. A probléma megoldására több módszert dolgoztak ki, tanulmányunkban ezeket tekintettük át.

Az összehasonlítás problémáját az egymásba ágyazott modellek összehasonlításakor sikerült megoldani. Ebben az esetben azonosítható az egy- és a többváltozós – vagy a bizonyos kontrollváltozókat nélkülöző, illetve felhasználó – modellekhez tartozó reziduális szórások hányadosa. Az y-standardizálás módszere ezt a szóráshányadost közvetlenül becsüli meg. A KHB-módszer ugyanakkor a többváltozós becsléseket kváziegyváltozós becslésekkel hasonlítja össze. Habár elvileg mindkét módszer eléri a kitűzött célt, a KHB-módszer valamivel pontosabbnak tűnik, helyesebben szólva, erre a szimulációk nem adtak ellenpéldát. Fontos megjegyezni, hogy a két módszer egyike sem ígéri az alapprobléma megoldását, tehát azt, hogy egy adott logisztikus regressziós együtthatóból kiszámítsák a strukturális hatást és a reziduális szórást. A módszerek célja a reziduális szórások hányadosának azonosítása, ami viszont lehetővé teszi a strukturális hatások hányadosának vagy közös skálán értelmezett különbségének az azonosítását.

Az eltérő részmintákra vonatkozó becslések összehasonlítása nehezebb feladatnak tűnik. E probléma onnan fakad, hogy adott magyarázó változó és a részmintát azonosító indikátorváltozó interakciós hatása kétféleképpen értelmezhető: *1*. a csoportokban eltérő a magyarázó változó hatása (strukturális értelmezés); illetve *2*. a csoportokban azonos a szóban forgó változó hatása, viszont eltérnek a reziduális varianciák (heterogenitáson alapuló értelmezés). A kétfajta értelmezés eltérő következményekkel jár: csak az utóbbi vonja maga után azt, hogy az interakciós és a főhatások hányadosa mindegyik magyarázó változónál ugyanaz. Habár ezt az arányossági feltevést statisztikai módszerekkel tesztelhetjük, a tesztnek véleményünk szerint alacsony a statisztikai ereje, és csak nagy mintákban van esély arra, hogy elutasítsuk a heterogenitáson alapuló értelmezést. A kétfajta értelmezés közötti választás másik megközelítését nyújta a heterogenitást tartalmazó logisztikus regressziós modell, ami explicit módon modellezi a reziduális varianciát. Amellett érveltünk, hogy e modell együtthatóinak identifikálása a függvényformán alapul, emiatt az eredményeket óvatosan kell kezelni. Állításunkat szimulációval vizsgáltuk. Azt találtuk, hogy a heterogenitást tartalmazó modellek együtthatói ugyanolyan mértékben torzítottak, mint a csoportinterakciós hatásokat is tartalmazó logisztikus regressziós modellek együtthatói.

A mintaspecifikus logisztikus regressziók összehasonlításának problémája azt feltételezi, hogy a statisztikai modellben el lehet választani a heterogenitásból fakadó hatásokat az okságinak gondolt strukturális hatásoktól. Mivel a problémát véleményünk szerint nem – vagy csak nagy mintákban – lehet megoldani, felmerülhet a kérdés, érdemes-e ragaszkodni a strukturális és a heterogenitáshatások megkülönböztetéséhez. *Keele* és *Park* [2005] vetették fel azt a problémát, hogy a heterogenitást tartalmazó modellben a varianciaegyenlet szerepét úgy is értelmezhetjük, hogy az bonyolultabbá teszi a magyarázó változó és a függő változó kapcsolatát definiáló függvényformát. Szintén a strukturális és a heterogenitáskomponensek szétválasztása ellen érvel *Buis* [2016]. A lineáris regresszióban a függő változó értékével kapcsolatos bizonytalanságot a hibatag képviseli. A logisztikus regressziós valószínűségi modell megfogalmazásában viszont közvetlenül a függő változó helyén szereplő valószínűség fejezi ki a bizonytalanságot. A magyarázó változók nemcsak a siker valószínűségét, várható értékét befolyásolják, hanem a siker bekövetkezésével kapcsolatos bizonytalanságot is, hiszen a diszkrét függő változó varianciája annak átlagától függ. Ennek az érvnek az elfogadása természetesen nem vonja maga után, hogy a statisztikai modell megalapozását szolgáló gondolatainkban nem lehetne különbséget tenni a strukturális és a bizonytalansági hatások között. Az érv csak azt bizonyítja, hogy az elméleti szempontból talán könnyen szétválasztható strukturális és bizonytalansági hatások szétválaszthatatlanul összefonódnak a logisztikus regressziós modellekben.

Irodalom

ALLISON, P. D. [1999]: Comparing logit and probit coefficients across groups. *Sociological Methods & Research*. Vol. 28. No. 2. pp. 186–208. https://doi.org/10.1177/0049124199028002003

ALVAREZ, R. M. – BREHM, J. [1995]: American ambivalence towards abortion policy: development of a heteroskedastic probit model of competing values. *American Journal of Political Science*. Vol. 39. No. 4. pp. 1055–1082. https://doi.org/10.2307/2111669

ANGRIST, J. D. [2001]: Estimation of limited dependent variable models with dummy endogenous regressors. *Journal of Business & Economic Statistics*. Vol. 19. Issue 1. pp. 2–28. https://doi.org/10.1198/07350010152472571

BARTUS T. [2003]: Logisztikus regressziós eredmények értelmezése. *Statisztikai Szemle*. 81. évf. 4. sz. 328–347. old.

BUIS, M. L. [2016]: *Logistic regression: When can we do what we think we can do?* 29 May. Working Paper. http://www.maartenbuis.nl/wp/odds_ratio_3.1.pdf

FÜLÖP P. [2002]: A bináris logit modellek használatának és tesztelésének eszközei. *Statisztikai Szemle*. 80. évf. 3. sz. 261–278. old.

HAJDU O. [2004]: A csődesemény logit-regressziójának kismintás problémái. *Statisztikai Szemle*. 82. évf. 4. sz. 392–422. old.

HUNYADI L. [2004]: A logisztikus függvény és a logisztikus eloszlás. *Statisztikai Szemle*. 82. évf. 10–11. sz. 991–1011. old.

KARLSON, K. B. – HOLM, A. – BREEN, R. [2012]: Comparing regression coefficients between same-sample nested models using logit and probit: a new method. *Sociological Methodology*. Vol. 42. Issue 1. pp. 286–313. https://doi.org/10.1177/0081175012444861

KARLSON, K. B. [2015]: Another look at the method of y-standardization in logit and probit models. *Journal of Mathematical Sociology*. Vol. 39. Issue 1. pp. 29–38. https://doi.org/10.1080/0022250X.2014.897950

KEELE, L. – PARK, D. K. [2005] *Difficult Choices: An Evaluation of Heterogenous Choice Models*. Working paper presented at the Annual Meeting of the Midwest Political Science Association. 7–10 April. Chicago.

LONG, J. S. [1997]: *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications. Thousand Oaks.

LONG, J. S. – MUSTILLO, S. A. [2017]: *Comparing Groups in Binary Regression Models Using Predictions*. Working Paper. https://pdfs.semanticscholar.org/9080/0f860ff738840e65d747d059a7f4f9ec6cfb.pdf

MOOD, C. [2010]: Logistic regression: Why we cannot do what we think we can do, and what we can do about it? *European Sociological Review*. Vol. 26. Issue 1. pp. 67–82. https://doi.org/10.1093/esr/jcp006

ROHWER, G. [2015]: A note on the heterogeneous choice model. *Sociological Methods & Research*. Vol. 44. No. 1. pp. 145–148. https://doi.org/10.1177/0049124114552750

TUTZ, G. [2018]: Binary response models with underlying heterogeneity: identification and interpretation of effects. *European Sociological Review*. Vol. 34. Issue 2. pp. 211–221. https://doi.org/10.1093/esr/jcy001

WILLIAMS, R. [2009]: Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research*. Vol. 37. Issue 4. pp. 531–559. https://doi.org/10.1177/0049124109335735

WILLIAMS, R. [2010]: Fitting heterogeneous choice models with OGLM. *The Stata Journal*. Vol. 10. No. 4. pp. 540–567. https://doi.org/10.1177/1536867X1001000402

WINSHIP, C. – MARE, R. D. [1984]: Regression models with ordinal variables. *American Sociological Review*. Vol. 49. No. 4. pp. 512–525. https://doi.org/10.2307/2095465

Summary

Recently, increasing attention has been devoted to the problem that estimated coefficients of logistic (and other non-linear) regression models cannot be compared across groups, samples, or nested model specifications due to the possible differences in the magnitude of unobserved heterogeneity. This study reviews methods which aim to solve this problem and investigates their effectiveness through simulation. Parameter estimates of nested model specifications can be made comparable using y-standardization or by comparing the estimates of the multivariate model to the estimates of a special, quasi-univariate model. Methods which aim to make coefficients comparable across groups and samples (such as testing the proportionality of interaction effects and heterogeneous choice models), however, do not provide adequate solutions for the problem. Causes behind this failure are discussed.