SZAK-MA

Szakmai közéleti diskurzusra reflektáló rovatunkban folytatjuk a Módszeresen előadás-sorozatról szóló beszámolók közlését. Ugyancsak beszámolunk a 2017. novemberében a 20. Század Hangja Archívum és Kutatóműhely 10. születésnapja alkalmából tartott kétnapos konferenciáról.

Beszámoló a Módszeresen rendezvénysorozatról

Szociológiai Szemle 28(1): 105–112.

Minden, amit tudni akartál a logisztikus regresszióról, de sohasem merted megkérdezni

Máté Fanni

https://doi.org/10.51624/SzocSzemle.2018.1.5

Bevezető

Az ELTE TáTK és az MTA TK szervezésében zajló Módszeresen vitasorozat[1] a társadalomkutatások módszereinek ismét – vagy még mindig – aktuális elméleti kérdéseit, problémáit feszegeti. A sorozat[2] keretén belül Kisfalusi Dorottya és Koltai Júlia a logisztikus regresszió alkalmazásáról, értelmezhetőségéről tartottak vitaindító előadást. A szerzők saját kutatásaik során tapasztalták meg a módszer alkalmazásának korlátait, és kerestek azokra lehetséges megoldásokat. A vita felkért hozzászólói Németh Renáta és Bartus Tamás voltak, a beszélgetést ezúttal is Janky Béla moderálta. A téma relevanciáját mutatja, hogy tudományos kutatások során a lineáris regressziós modellek alkalmazása mellett gyakori a logisztikus regresszió használata is, azonban gyakran figyelmen kívül hagyják értelmezhetőségének és használhatóságának korlátait – miként ez az egyik korábbi Módszeresen-vita témájául szolgáló szignifikanciateszt esetében is előfordul.

A modell felépítéséből adódó korlátok, problémák ismertetéséhez röviden bemutatjuk magát a logisztikus regressziós modellt, amelyben egy dichotóm függő és egy vagy több független változó szerepel.

1 Előadásait lásd részletesen a következő honlapon: http://co.o-o-o.hu/hu/home/m%C3%B3dszeresen

2 Az első három előadásról készült összefoglaló a Szociológiai Szemle 2017. évi első számában olvasható.

Tegyük fel, hogy a logisztikus regresszió függő változójaként szereplő dichotóm változót egy nem mérhető, folytonos változó határoz meg: ha ennek értéke meghalad egy küszöbértéket, a kétértékű változó értéke 1 lesz, különben 0. A logisztikus regresszió függő változójára tekinthetünk tehát valójában egy látens, elméleti folytonos változó empirikusan megfigyelhető indikátoraként (Allison 1999). Ha ezt elfogadjuk, akkor minden kétértékű változó egy hajlandóságot, potenciált, vagy akár kockázatot mérő folytonos változó kettéosztása annak egy küszöbértékénél.[3] Például a migrációt vizsgálva a költözés tényét rögzítő kétértékű változó mögött meghúzódik a vizsgált személy hajlandóságát vagy döntési folyamatát leíró folytonos változó, amely eredménye a költözés vagy a maradás. Valójában a különböző kutatások során célunk tehát ezt a látens potenciált, hajlandóságot, vagyis a folytonos változót megragadni, és azt mérni, hogy bizonyos tényezők milyen hatással vannak rá. Azaz e látens lineáris modell regressziós együtthatóit kívánjuk becsülni, hiszen – a példánál maradva – arra vagyunk kíváncsiak, milyen szerepet játszik a migrációs hajlandóságban például az életkor vagy a munkaerőpiaci helyzet. Azonban a lineáris modell együtthatóit nem, csak a rendelkezésünkre álló kétértékű változóhoz tartozó modell együtthatóit tudjuk megbecsülni.

Az egyik lényeges kérdés tehát az, hogy mennyire ragadhatók meg a valóban felmérni kívánt hatások a rendelkezésünkre álló logisztikus regressziós modellel? A kérdés megválaszolásához vizsgáljuk meg a modelleket: a lineáris modellben szereplő függő változó varianciájának bizonyos hányadát megmagyarázzák az általunk bevont magyarázó változók, de nem a teljes varianciát – ez a meg nem magyarázott rész a regressziós egyenlet hibatagja. Pontosan ez a hibatag lesz a probléma forrása a logisztikus modellben: mivel nem ismerjük az elméleti látens folytonos változó varianciáját, így nem ismerjük azt sem, mekkora ennek a meg nem magyarázott része. A logisztikus regressziós modellben, ahol e folytonos változó helyett annak indikátorára adunk becslést, ez a hibatag standard logisztikus eloszlásúnak feltételezett, ebből következően varianciája 3,29 (Mood 2010). Azonban a látens lineáris modellben szereplő hibatag varianciája a legtöbb esetben nem ennyi, így szükség van egy együtthatóra, amely ezen modell hibatagjának varianciájához igazítja a logisztikus regressziós modell hibájának fix értékű varianciáját (Allison 1999; Mood 2010). Ezáltal viszont az eredetileg keresett lineáris regressziós együtthatókat is módosítjuk, éppen akkora mértékben, amennyi a lineáris modell hibatagjának korrekciós tényezője. Ebből következően a logisztikus regresszió együtthatói függnek a modell által meg nem magyarázott résztől (a reziduálisok varianciájától), azaz az együtthatók nem csak az adott változó hatását mutatják, hanem nagyságuk attól is függ, hogy mekkora a modell reziduálisa. Ennek folyományaként két olyan logisztikus regressziós modell együtthatóit, amelyeknek eltérő a meg nem magyarázott része, nem hasonlíthatjuk össze, hiszen nem tudhatjuk: tényleg különbség van-e az egyes vál-

3 Ez a küszöbérték is csak átvitt értelemben értendő, hiszen nem feltétlenül egzakt, mérhető folytonos változókról beszélünk.

tozók hatása között, vagy csak az eltérő meg nem magyarázott résznek köszönhető az együtthatók különbözősége.

Ez a tulajdonság több problémát is eredményez a logisztikus regressziós modellben szereplő független változók hatásainak vizsgálatakor. Egyfelől (1) nem hasonlíthatók tehát össze azonos struktúrájú, de különböző mintabeli csoportokra vonatkozó modellek együtthatói, (2) a modellben szereplő interakciók értelmezése is problémás, és (3) új változó bevonásakor sem csak amiatt változhatnak meg az együtthatók, mert az új változót kontroll alatt tartva megváltozik a már a modellben szereplő változók hatása, hanem a meg nem magyarázott rész eltérése miatt is. Az alábbiakban részletesebben sorra vesszük ezen problémákat és a rájuk adott megoldási lehetőségeket.

Nehézségek a logisztikus regresszióval kapcsolatban

Az együtthatók varianciától való függésének következménye tehát egyrészt az, hogy ha a minta két vagy több csoportjára külön-külön végzünk azonos felépítésű logisztikus regressziós elemzést (például férfiak és nők között vizsgáljuk ugyanazon, előléptetésre ható tényezőket), a két modell együtthatóinak eltérése esetén nem tudhatjuk biztosan, hogy az eltérés valóban a független változók hatásának különbözősége miatt van-e, avagy amiatt, hogy a két csoportban a modell változói által meg nem magyarázott heterogenitás eltér.

Ugyanígy az interakciók együtthatóinak értelmezése is problémássá válik – mivel az a struktúra, amiben egy kétértékű változót minden független változóval interakcióba léptetünk, ekvivalens azzal, mintha e változó két kategóriájában két külön regressziós modellt vizsgálnánk (Allison 1999) – amely modelleknél viszont a fentiek alapján nem hasonlíthatók össze az együtthatók.

Szintén problémás a független változók hatását összevetni egymásba ágyazott modelleknél, azaz amikor az eredeti modellt különböző változók bevonásával bővítjük annak érdekében, hogy minél pontosabb becslést adjunk. Gyakori ez a módszer akkor, amikor azt feltételezzük, hogy létezik egy olyan változó, amely az eredeti, általunk vizsgált kapcsolatra hatással van, és célunk ennek a változónak a hatását kiszűrni, kontroll alatt tartani. Mivel lineáris regresszió során a függő változó varianciája állandó, új változó bevonásával csökkenteni tudjuk a reziduális, azaz a modell által meg nem magyarázott varianciát – vagyis valóban pontosabb becslést adhatunk. Lineáris regresszió esetén tehát, ha az újonnan bevont változó korrelálatlan a modellben szereplő változókkal, a hatás mértéke ugyanakkora lesz a többváltozós modellben, mint a kevesebb változót tartalmazóban, míg logisztikus regresszió esetén ez az összefüggés nem áll fenn. Mivel logisztikus regresszió esetén a reziduális variancia értéke adott, további változók bevonásával nem tudjuk csökkenteni azt. Így új változók bevonása nem a hibatagot csökkenti, hanem a megmagyarázott varianciát – azaz összességében a függő változó varianciáját – növeli. Emiatt logisz-

tikus regresszióná l a hibatag korrekciós tényezőjétől, így a meg nem magyarázott hányadtól akkor is függnek a modell együtthatói, ha a modellben szereplő változók korrelálatlanok a modellből kimaradó változókkal.

Ez könnyen belátható az alábbi példán keresztül: tegyük fel, hogy a szavazási hajlandóságot kívánjuk mérni, amit egy kétértékű változóval képezünk le, amelynek értéke 1, ha a kérdezett elmegy szavazni, és 0, ha nem. Elsőként egyedül a válaszadók nemét vonjuk be független változóként a modellbe. Ebben az esetben azt tapasztalhatnánk például, hogy minden férfinál 50 százalék, és minden nőnél 40 százalék annak a valószínűsége, hogy valaki részt vesz a választáson. A modellt az életkor változóval bővítve azt láthatnánk, hogy a férfiaknál a becsült valószínűség 45 és 55 százalék, a nőknél pedig 35 és 45 százalék között mozog. Így a választási részvétel becsült valószínűsége az egy független változós modellben 40 és 50 százalék lehet, míg a bővebb modellben 35 és 55 százalék között mozog – azaz a kibővített modellben a függő változó varianciája nagyobb (Williams 2016).

Amennyiben nem bővítjük a modellt, hanem bizonyos, a függő változóra valóban hatással bíró független változót kihagyunk a modellből, az kétféleképpen is befolyásolhatja a modellben szereplő változók együtthatóit. Egyrészt befolyásolhatja az együtthatókat – a lineáris regressziós modellhez hasonlóan – az elhanyagolt változóból fakadó torzítás, másrészt a reziduális varianciából eredő torzítás is. Előbbinél az együtthatók nagyobbak és kisebbek is lehetnek a valós hatáshoz viszonyítva, attól függően, hogy a kihagyott változó hogyan korrelál a modellben szereplő változókkal. A reziduális varianciából adódó torzítás pedig alacsonyabb együtthatókat eredményezhet az egyváltozós modellben, mint amiket többváltozós modellben kapnánk a függő változó alacsonyabb varianciája miatt (Mood 2010; Williams 2016).

Összefoglalva tehát: téves a logisztikus regressziós modell során a modellben szereplő független változók együtthatóit, azaz hatását két hasonló felépítésű – akár egymásba ágyazott – modellben összevetni egymással, mivel az együtthatók értéke függ a modell által meg nem magyarázott résztől (amelynek nagyságát viszont nem ismerjük).

A fentieken túl a logisztikus regresszió alkalmazását problematikussá teszi, hogy az eredmények interpretálása során a modellek együtthatóit a könnyebb érthetőség érdekében gyakran esélyhányadosokká alakítják, így a modellek ezen mérőszámok gyengeségeit is hordozzák. Nevezetesen például azt, hogy az esélyhányados felülről nem korlátos, ezért arról, hogy a megfigyelt erősség mennyire van közel a determinisztikus összefüggéshez, nem tudunk megállapítást tenni. Ezt a problémát kiküszöbölendő gyakran használják az esélyhányadosok logaritmusát, amellyel viszont a mérőszám könnyű interpretálhatósága veszik el. Az esélyhányados emellett nem alkalmas olyan keresztt áblák esetén, ahol az egyik cellában nulla a gyakoriság, hiszen ebben az esetben erős, determinisztikus kapcsolatot mutat akkor is, ha valójában nem ilyen jellegű két változó kapcsolata (Bartus 2003a).

A modell korlátainak ismertetése után felmerül a kérdés, hogy mennyire jelentős az együtthatók össze nem vethetőségének problémája a mindennapi kutatási gyakorlatban?

Az egymásba ágyazott logisztikus regressziós modellek esetén értelemszerűen nem kell számolnunk az együtthatók változásának értelmezési problémáival, amennyiben az egyre bővebb modellek esetén nem hasonlítjuk össze az együtthatókat, csak a végső – minden magyarázó változót tartalmazó – modellben értelmezzük azokat, s a kevesebb magyarázó változós modelleknél csak illeszkedésstatisztikákat közlünk.

A logisztikus regressziós modellben szereplő együtthatók változásának problémája szintén nem jelentős, ha a látens függő változó varianciája a modellek között nem tér el jelentősen, és ha az együtthatók csökkennek a többváltozós modellben. Előbbit sajnos nem tudhatjuk biztosan, utóbbi esetben pedig arról van szó, hogy csupán alulbecsüljük az együtthatók csökkenésének mértékét – de ennek ismerete is lényeges az eredmények helyes értelmezéséhez. Ha azonban az együtthatók magasabbak a többváltozós modellben, akkor tisztában kell lennünk vele, hogy nem feltétlenül szupresszorhatásról beszélünk, vagyis nem biztos, hogy olyan változót találtunk, amely elfedte a már modellben szereplő változók hatását, hanem az eltérés abból is adódhat, hogy a meg nem magyarázott variancia tér el a modellek között (Williams 2016).

Mit tehetünk akkor, ha a kutatás során ezekbe a problémákba ütközünk?

A szerzők ötféle módszert mutattak be, melyek megoldást jelenthetnek a felvetett problémákra. Ezek a következők voltak: a (1) heterogenous choice model, az (2) y-standardizálás, az (3) átlagos marginális hatás, a (4) lineáris valószínűségi modell alkalmazása és a (5) KHB-módszer. Az alábbiakban ezen módszerek rövid összefoglalását közöljük.

Ha a minta két vagy több csoportjában kívánjuk összehasonlítani az együtthatókat (ahogy a fenti példában a férfiakra és nőkre külön illesztett modell esetén), akkor megoldást nyújthat az ún. heterogenous choice model, amelynek célja az együtthatók változásából kiszűrni azt a részt, amely a meg nem magyarázott variancia eltéréséből adódik, azaz a módszer a heteroszkedaszticitást kívánja kontroll alatt tartani. A regressziós modell két részből áll: a számlálóban található, ún. choice equationből és a nevezőben található variance equationből. Utóbbi célja, hogy modellezze a reziduális szórást, és ezáltal korrigálja a regressziós együtthatókat a meg nem magyarázott rész szempontjából (Williams 2009).

Az y-standardizálás során az együtthatókat úgy tesszük összehasonlíthatóvá, hogy minden modell esetén elosztjuk az együtthatókat a látens változó becsült szórásával, mely szórásra adott becslésünk a logitok szórásának és a hibatag szórásának

összegeként áll elő. Így ugyan kiküszöböljük a meg nem magyarázott hányadból fakadó különbséget, viszont az együtthatók értelmezése megváltozik: azt mutatják meg, hogy a magyarázó változók egy egységnyi növekedésével a feltételezett látens változó értéke hány szórásegységgel változik (Mood 2010; Williams 2016).

A marginális hatások kiszámításával is összehasonlíthatjuk a változók hatását. Ennél a módszernél minden esetben kiszámoljuk, mennyi az *y*, függő változó bekövetkezésének valószínűsége a független változó adott értéke mellett, majd vesszük a két valószínűség különbségét, hogy megkapjuk a marginális hatást (Bartus 2003b). Kategoriális változóknál a marginális hatás jelentése az, hogy hogyan változik a függő változó bekövetkezési valószínűsége, ha a független változó értéke nulláról egyre nő, míg folytonos változók esetén az, hogy mennyivel változik egy esemény bekövetkezésének valószínűsége, ha a magyarázó változó végtelenül kis mennyiséggel növekszik (Bartus 2003b). Mivel több – a kategóriák közti – összehasonlítás is lehetséges, felmerül, hogy miként fejezhető ki egyetlen számmal a marginális hatás? Erre kétféle megoldás is adódik: vagy átlagoljuk a marginális hatásokat, vagy pedig a független változót rögzítjük egy adott értéken. Azonban, ahogyan a későbbiekben látjuk majd, ez a mérőszám sem alkalmas minden esetben a változók hatásának mérésére.

Általánosan alkalmazható megoldásként felmerül, hogy ha eredendően is lineáris regressziós együtthatókra vagyunk kíváncsiak, egyszerűen alkalmazzunk lineáris regressziós modellt úgy, hogy a függő változónk nem folytonos, hanem dichotóm. Ekkor a függő változó 1-es értékének valószínűségét a független változók lineáris függvénye adja. Ezzel a módszerrel szemben több kritika fogalmazódik meg. Az egyik kritika arról szól, hogy nem véletlenül alkalmazunk logisztikus modellt és nem lineárist: mivel a modell egy valószínűséget becsül, a becslés eredményeképpen 0 és 1 közötti számokat kellene kapnunk, ez azonban a lineáris modellnél nem garantálható. Szintén kritikaként jelenik még meg, hogy a logisztikus modellben a hibatagok sem normális eloszlásúak.

Az eddig bemutatott módszerek azonban mind problémásak lehetnek, ha a modellekben, amelyeket vizsgálunk, a hibatagok megoszlása nagyon különböző. Ezekre az esetekre alkották meg a KHB-módszert, melynek lényege, hogy a látens lineáris modellben az egyik független (*x*) változónak csak azt a részét szerepeltetjük, ami független a másik (*z*) független változótól. Ezt oly módon érhetjük el, hogy a két változó – *x*, mint függő és *z*, mint független – közötti lineáris regressziós modell reziduálisát szerepeltetjük az eredeti modellben (Kohler et al. 2011).

Az előadók által bemutatott és a fentiekben vázolt lehetőségek sem problémamentesek, és – ahogyan arra a szerzők felhívták a figyelmet – nincs egy univerzálisan alkalmazható megoldás. Azonban jól használva a módszereket, adott kutatási helyzetben alkalmasak lehetnek a logisztikus regresszió egyes hiányosságainak kezelésére.

Az előadást követő vita tapasztalatai

A kutatók, társadalomtudósok között zajló eszmecsere egyrészt a logisztikus regresszió mögött húzódó látens lineáris modell értelmezésére irányult, s arra, mit tehet a kutató annak érdekében, hogy elkerülje a logisztikus regresszió problémáit. Emellett a beszélgetés kitért arra is, hogy az előadásban vázolt megoldási javaslatok közül melyiket mikor érdemes használni.

A kétértékű függő változóra illeszthető lineáris modell mellett és ellene is felsorakoztattak szempontokat a felkért hozzászólók. Ha például a vizsgált változó az S alakú telítődési görbe szerint van hatással a függő változóra, akkor nem alkalmazható a lineáris valószínűségi modell, mivel a hatás nem lineáris. Azonban, hogyha a görbének csak egy szűk intervallumát vesszük a kutatás során – egy olyan szakaszt, ahol a görbe lineáris –, a lineáris valószínűségi modell e feltétele teljesül. Illetve, ha csak kategoriális független változókkal dolgozunk a lineáris modellben, a becsült valószínűség inkább benne marad a nullától egyig terjedő intervallumban, míg folytonos változók esetén erre nem számíthatunk.

Míg az átlagos marginális hatás alkalmazása adekvát lehet olyan esetekben, amikor a cél egy valószínűség megbecslése, addig bizonyos kutatások során (például eset-kontroll vizsgálatokban, ahol a marginális gyakoriságok a kutató által rögzítettek) nehezebb interpretálhatósága mellett is az esélyhányados használata indokolt.

Ezek mellett pedig természetesen a kutatás tervezése során átgondolandó az is, mely mérőeszköz az, amely a kutatás célját tekintve a független változók hatásának kifejezésére a legalkalmasabb.

A vitasorozat ezen előadása is elérte azt a célját, hogy felhívja a figyelmet gyakran használt statisztikai módszerek alkalmazhatóságának kérdéseire, s kritikai gondolkodásra sarkallja mind az egyetemi hallgatókat, mind a gyakorló kutatókat a tanult módszerekkel, jelen esetben a logisztikus regresszióval kapcsolatban. Végső konklúzióként megfogalmazható – és ez a gondolat nem csak ebben a témában, hanem általában a kutatásmódszertani dilemmák esetén megállja a helyét –, hogy az egyes módszerek határaira, problémáira nyújtott megoldások nem feltétlenül általánosak: mindig a kutatási célnak megfelelően válasszunk megoldást a felmerülő problémára, és ne csak újonnan alkalmazott módszereknél mélyedjünk el az eszköz működésének megismerésében, hanem olyan esetekben is, amikor már rutinszerűen használjuk azt.

Az előadás alapján készült (és szimulációkkal kiegészített) tudományos cikk várhatóan 2018 végén-2019 elején jelenik meg Bartus Tamás, Kisfalusi Dorottya és Koltai Júlia szerzőségével.

Irodalom

Allison, P. D. (1999): Comparing logit and probit coefficients across groups. *Sociological Methods and Research, 28*(2): 186–208.

Bartus T. (2003a): Oksági kapcsolatok erejének mérése kontingenciatáblákban: az esélyhányados problémái és a hatásnagyság. *Szociológiai Szemle, 13*(2): 42–58.

Bartus T. (2003b): Logisztikus regressziós eredmények értelmezése. *Statisztikai Szemle, 81*(4): 328–347.

Kohler, U. – Karlson, K. B. – Holm, A. (2011): Comparing coefficients of nested nonlinear probability models. *The Stata Journal, 11*(3): 420–438.

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, 26*(1): 67–82.

Williams, R. (2009): Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods and Research, 37*(4): 531–559.

Williams, R. (2016): Comparing logit & probit coefficients between nested models. Working Paper, March 1.