A zéróinflált és a hurdle-modellek egy lehetséges társadalomtudományi alkalmazása: roma ismerősök számának elemzése*

Vit Eszter,
az MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoport kutatási asszisztense
E-mail: vit.eszter@gmail.com

Jelen elemzés célja, hogy ismertesse az előfordulási gyakoriságokra alkalmazható modellek két speciális típusát, a zéróinflált (zero-inflated) és a hurdle- (gát-) modelleket, valamint bemutassa egy lehetséges társadalomtudományi alkalmazásukat. E kétkomponensű modellek abban az esetben javíthatják a becslések pontosságát, amikor a vizsgált adatokban a zérus értékek túlzott előfordulása a Poisson-modell túlszóródásához vezet.

A tanulmány az előfordulási gyakoriságok modellezésekor alkalmazott (Poisson- és negatív binomiális, zéróinflált Poisson- és zéróinflált negatív binomiális, hurdle Poisson- és hurdle negatív binomiális) modelleket veti össze egy adatfelvétel keretében megkérdezett személyek roma ismerőseinek számát elemezve. Az eredmények szerint egyrészt a kétkomponensű modellek növelik a becslés pontosságát, másrészt annak a kérdésnek a tanulmányozására is használhatók, hogy mely tényezők befolyásolják bizonyos személyek ismeretségének, egymással való kapcsolatba lépésének a lehetőségét. Azonban érdemes számításba venni azt is, hogy e modellek számos paraméter becslését igénylik, ami túlillesztésükhöz vezethet.

TÁRGYSZÓ:
Statisztikai módszertan.
Társadalomtudományi kutatás.
Előfordulási gyakoriság.

DOI: 10.20311/stat2018.07.hu0683

* A szerző ezúton fejezi ki köszönetét konzulensének, *Kmetty Zoltán*nak a tanulmány elkészítéséhez nyújtott értékes segítségéért.

A társadalomtudományokban gyakran előfordul a jelenségek széles körének – például az öngyilkosságok, a halálozások, a születések vagy a szakirodalmi hivatkozások mennyiségére gyakorolt hatás (*Moksony* [2006]) – vizsgálatakor, hogy az elemzés függő változója valamilyen esemény előfordulási gyakorisága (countváltozó). Előfordulási gyakoriság alatt azt értjük, hogy egy adott esemény hányszor következik be. E jelenségek diszkrét modellekkel írhatók le, ahol az előfordulási gyakoriságokat mérő valószínűségi változók minden esetben nemnegatív egész számok (*Hilbe* [2011]).

Jelen elemzés célja az előfordulási gyakoriságokra alkalmazható modellek két speciális típusának, a zéróinflált (zero-inflated) és a hurdle- (gát-) modellek egy lehetséges társadalomtudományi alkalmazásának a bemutatása az MTA–ELTE (Magyar Tudományos Akadémia – Eötvös Loránd Tudományegyetem) Peripato Összehasonlító Társadalmi Dinamika Kutatócsoport 2014. májusban, „Válság és innováció" címmel végzett adatfelvételében részt vett válaszadók roma ismerősei számának elemzésén keresztül. E kétkomponensű modellek alkalmazása akkor indokolt, ha a vizsgált adatainkban előforduló túl sok zérus érték[1] az eredeti modellünk túlszóródásához vezet.

Előfordulási gyakoriságot mérő kategoriális adatok esetén fontos különbség a binomiális vagy a multinomiális eloszlás segítségével modellezhető megfigyelésekhez képest, hogy az adatok előre nem meghatározott számú kísérletből származnak, így nincs felső határa sem a kísérletek, sem a megfigyelt gyakoriságok számának (*Agresti* [2002]). A jobbra ferde, aszimmetrikus eloszlású, diszkrét, véletlen bekövetkezésű megfigyelések modellezésére alapvetően a Poisson-eloszlás alkalmazható (*Moksony* [2006]).

Amikor az előfordulási gyakoriságokat mérő változót magyarázandó vagy függő változónak tekintjük, akkor a regressziós logika szerint vizsgálható, hogy hatnak-e bizonyos magyarázó változók a kitüntetett függő változó értékére, és amennyiben igen, miként. Előfordulási gyakoriságot mérő függő változó esetén a legegyszerűbb regressziós modell a Poisson-regresszió, ami az általánosított lineáris modellek közé sorolható (*Cameron–Trivedi* [1998]).

A Poisson-eloszlás függvénye a következő:

$$P(\lambda) = \frac{e^{-y}\lambda^{y}}{y!},\ y = 0,1\ldots,\lambda > 0,$$

[1] Itt és a továbbiakban a zérus és a nulla értéket szinonimaként használom.

ahol λ az eloszlás egyetlen paramétere, az ún. intenzitási paraméter, mely azt fejezi ki, hogy az adott eseményből átlagosan hány következik be egy intervallumban (*Cameron–Trivedi* [1998] 3. old.).

A Poisson-eloszlás általánosított lineáris modelljében a várható érték logaritmusát szokásos modellezni, mely bármilyen valós értéket felvehet. Tehát a Poisson-féle általánosított lineáris modell összekötő függvénye $g(\mu) = \ln(\mu)$, melynek inverze $\mu_i = e^{x_i\beta}$ (ahol x_i az x magyarázó változó értéke az i-edik megfigyelésre nézve, β pedig az x magyarázó változóhoz tartozó együttható) adja meg a modell által illesztett értéket (*Cameron–Trivedi* [1998], *Agresti* [2002]). A Poisson regressziós modell maximalizálandó loglikelihood függvénye x magyarázó változó esetén: $lnL(\beta, y) = \sum_{i=1}^{n} y_i(x_i^T\beta) - e^{x_i^T\beta} - \ln(y_i!)$ (*Cameron–Trivedi* [1998] 21. old.).

A Poisson-modell megfigyelésekre vonatkozó alapfeltevései közé tartozik, hogy az egyes események függetlenek egymástól és homogén eloszlásúak, továbbá, hogy a megfigyelések eloszlásának várható értéke és varianciája megegyezik egymással: $E(Y) = Var(Y) = \lambda$. Amennyiben adataink feltételes varianciája meghaladja a feltételes várható értéket, túlszóródásról, ha az alatt marad, alulszóródásról beszélhetünk (*Cameron–Trivedi* [1998]).[2] A túlszóródás egy speciális, jelen tanulmány középpontjában álló esete, amikor túl sok zérus érték figyelhető meg.

Ennek kapcsán fontos megemlíteni a nulla értékek két típusát: a mintavételi és a strukturális nulla értékeket. A mintavételi nulla értékek esetén a nem nulla érték előfordulása nem lehetetlen, a minta sajátossága viszont a nulla érték, tehát az, hogy a nulla érték valamilyen valószínűséggel előfordul. Ezzel szemben a strukturális nulla érték az, aminek az előfordulása elméleti szempontból is kizárt (*Agresti* [2002]).

A Poisson-modellben pozitív valószínűség tartozik a nulla érték előfordulásához (*Cameron–Trivedi* [1998]), amely:

$$p(0, \lambda) = \frac{e^{-\lambda}\lambda^0}{0!}, \text{ vagyis } e^{-\lambda}.$$

Mivel a Poisson-modell homogén eloszlást feltételez, nem különbözteti meg egymástól a zérus és a többi kimenetet generáló folyamatokat. Előfordulhat azonban, hogy tényleges adatainkat nem homogén eloszlás jellemzi, és a megfigyelt adatokban a nulla előfordulási gyakoriságok valószínűsége jelentősen eltér a Poisson-modell által feltételezettől. Ebben az esetben Poisson-modellünk torzított becslést ad. A homogén eloszlástól való eltérés megvalósulhat a Poisson-modell által feltételezettnél kevesebb és több zérus érték, vagy a zérus kimenet hiánya miatt is.

[2] A gyakorlatban a túlszóródás jelensége gyakrabban fordul elő (*Cameron–Trivedi* [1998]).

Amennyiben a megfigyeléseink egyáltalán nem tartalmaznak nulla értéket, akkor indokolt lehet zérócsonkolt modell alkalmazása (*Cameron–Trivedi* [1998]). A hurdle- és a zéróinflált modellekre a csonkolt modellekkel ellentétben akkor lehet szükség, amikor a megfigyelt adatok az eloszlás által feltételezettnél több nullát tartalmaznak, ami – mint már említettem – túlszóródáshoz vezet (*Hilbe* [2011]). Amellett, hogy e modellek képesek korrigálni a zérus értékek vártnál magasabb számából eredő túlszóródást, fontos megemlíteni azt a megközelítést is, ami az adatok keletkezésével kapcsolatos. A zéróinflált és a hurdle-modellek azon a feltevésen alapulnak, hogy eltérő folyamatok határozzák meg, hogy a vizsgált változó zérus vagy annál magasabb értéket vesz-e fel, és általában milyen előfordulási gyakoriság jellemzi.

A zérus értékek túlzott előfordulásából vagy hiányából fakadó túlszóródás a Poisson-modell túlszóródásának egy lehetséges, de nem kizárólagos esete. A továbbiakban a tanulmány először áttekinti a Poisson-modell túlszóródásának néhány főbb esetét, különös figyelmet szentelve a zérus értékekhez kapcsolódó problémákra és ezek orvoslására zéróinflált, illetve hurdle-modellek segítségével. Majd a már említett válaszadók roma ismerősei számának elemzésén keresztül mutat be példát e kétkomponensű modellek lehetséges társadalomtudományi alkalmazására.

1. A Poisson-modell túlszóródása

Egy Poisson regressziós modell esetén túlszóródásnak tekintjük, amikor a kimeneti vagy magyarázandó változónk szóródása nagyobb, mint annak várható értéke. A kutatási gyakorlatban az általánosított lineáris modellek keretei között illesztett Poisson regressziós modellek túlszóródásának megállapítására bevett mérőszám a modellszóródási statisztika és a hozzá tartozó szabadsági fokok száma. Ha ezek hányadosa nagyobb, mint 1, adataink túlszóródást mutatnak, ha kisebb annál, alulszóródást (amennyiben a hányados értéke 1, sem a túlszóródás, sem az alulszóródás esete nem áll fenn). A modell szóródása önmagában azonban nem tekinthető túlszóródásra vonatkozó statisztikai tesztnek (*Hilbe* [2011]).

Az adatokban megfigyelt túl- vagy alulszóródás azért jelent problémát, mert ekkor a Poisson-regresszió paraméterbecsléseinek standard hibája nem megbízható. A Poisson-regresszió együtthatóinak szórásszámítása ugyanis a Poisson-eloszlás előfeltételei alapján történik, így kisebb lesz, mintha a túlszóródás jelenségével számolnánk (*Agresti* [2002], *Hilbe* [2011], *Moksony* [2006]).

A túlszóródás jelensége mögött több, egymástól eltérő mechanizmus húzódhat meg. *Hilbe* ([2011] 142. old.) látszólagos (apparent overdispersion) és valódi túlszóródás (real overdispersion) kezelésére vonatkozó technikákat különböztet meg. Látszóla-

gos a túlszóródás akkor, ha a modellből hiányoznak fontos magyarázó változók, nem tartalmaz jelentős interakciós tagokat, az adatokban kiugró értékek vannak, vagy van olyan magyarázó változó, amelyet skálatranszformációnak kellene alávetni. Ennek egyik esete, amikor nem megfelelő annak a függvénynek a megválasztása, ami az előfordulási gyakoriságot mérő adatok elemzése esetén megteremti a lineáris kapcsolatot a magyarázandó és a magyarázó változók között az általánosított lineáris modell keretein belül. Szintén látszólagos túlszóródáshoz vezet, ha a megfigyelések függetlensége csoportos/klaszterezett mintavétel következtében nem teljesül, azonban többszintű modellek alkalmazásával ezt az elemzéskor nem veszik figyelembe (*Hilbe* [2011]). A túlszóródás esetei a kiugró értékek kezelésével, megfelelő interakciós tagok hozzáadásával, skálatranszformációval, az előbb leírt lineáris kapcsolatot megteremtő függvény helyes megválasztásával vagy többszintű elemzéssel kerülhetők el.

A látszólagos túlszóródás megszüntetésére szolgáló módszereken kívül rendelkezésre állnak „valódi" túlszóródásra alkalmazható technikák is. Valódi a túlszóródás akkor, amikor a látszólagos túlszóródás lehetséges okainak figyelembevételét és megszüntetését követően is nagyobb a magyarázandó változónk szóródása, mint annak várható értéke. Ekkor a fő problémát a standard hibák megbízhatatlansága jelenti, melyek korrigálására többféle módszer is létezik: az újraskálázás, a bootstrap- vagy a jacknife-módszer, a varianciarobusztus (az ún. „szendvics" [sandwich] variancia-) becslés stb. (*Hilbe* [2011]).

A Poisson-modell túlszóródása esetén a kutatási gyakorlatban a negatív binomiális modell alkalmazása a leggyakoribb

2. A negatív binomiális modell

A negatív binomiális eloszlásnak többféle definíciója létezik. Az előfordulási gyakoriságok modellezésével foglalkozó szakirodalom leginkább a negatív binomiális modell Poisson-eloszlásból származó megközelítését alkalmazza. A negatív binomiális eloszlás korrigálja az előfordulási gyakoriságok mint függő változók túlszóródását. A negatív binomiális regresszió esetén nem szükséges, hogy a feltételes variancia egyenlő legyen a feltételes várható értékkel. Az előbbi a várható érték valamilyen függvényeként modellezhető: $\omega_i = \omega(\mu_i, \alpha)$, ahol μ_i a feltételes átlag, α pedig a diszperziós paraméter (amit becsülni kell) (*Hilbe* [2011]).

A szakirodalom ezek alapján a negatív binomiális variancia általánosított függvényét leggyakrabban a következőképp határozza meg: $\omega_i = \mu_i + \alpha\mu_i^p$, ahol p adott konstans. Amennyiben $\alpha = 0$, $\omega_i = \mu_i$, tehát a feltételes variancia egyenlő a felté-

teles várható értékkel, a Poisson-modellnek megfelelő esettel állunk szemben, vagyis a Poisson-modell a negatív binomiális modellnek egy olyan speciális esete, ahol a diszperziós paraméter értéke nulla. Ennek megfelelően a Poisson- és a negatív binomiális modell egymásba ágyazottnak tekinthető (*Cameron–Trivedi* [1998], *Hilbe* [2011]). A p értéke általában két specifikus esetre szűkíthető, melyek a negatív binomiális eloszlás első és második változatai. $p = 1$ esetén a variancia és a várható érték között multiplikatív ($\omega_i = (1 + \alpha)\mu_i$), $p = 2$ esetén pedig négyzetes a kapcsolat ($\omega_i = \mu_i + \alpha\mu_i^2$) (*Cameron–Trivedi* [1998]).

A szakirodalomban a negatív binomiális modellre való hivatkozás általában az utóbbit ($p = 2$), a Poisson- és a gamma-keverékeloszlásból származtatott eloszlást takarja, ahol μ_i a Poisson-, $\alpha\mu_i^2$ pedig a gammavariancia. A heterogenitást vagy túlszóródást kifejező α valójában a két tényező indirekt kapcsolatát leíró $\frac{1}{\nu}$ inverze, melyből ν a keverékeloszlás gammaeloszlásának az alakparamétere. Amennyiben α (vagyis ν^{-1}) $\to 0$, vagy $\nu \to \infty$, akkor a negatív binomiális eloszlás a Poissonhoz tart (*Cameron–Trivedi* [1998], *Hilbe* [2011]).

A keverékeloszlás úgy értelmezhető, hogy feltesszük, Y Poisson-eloszlású λ várható értékkel, amely valamilyen gammaeloszlás szerint változik. Ekkor tehát λ gammaeloszlású, ν és μ paraméterekkel. A λ gammaeloszlás-függvénye:

$$f(\lambda;\nu,\mu) = \frac{\left(\frac{\nu}{\mu}\right)}{\Gamma(\nu)} \cdot e^{\frac{-\nu\lambda}{\mu}} \lambda^{\nu-1}, \text{ ahol } \lambda \geq 0.$$

$\Gamma(\nu)$ a gammafüggvény, $\nu > 0$ pedig az alakparaméter, mely azt befolyásolja, hogy az eloszlás milyen mértékben jobbra elnyúló. Ekkor a gammaeloszlású λ várható értéke és varianciája az előbbieknek megfelelően:

$$E(\lambda) = \mu, \ \text{var}(\lambda) = \frac{\mu^2}{\nu}.$$

Az utóbbi variancia a negatív binomiális eloszlás varianciájának a gammaeloszlásból származó része (*Agresti* [2002] 559–560. old.).

Amennyiben a modell α diszperziós paraméterét adott konstansként léptetjük be a becslésbe, akkor a negatív binomiális regressziós modell az általánosított lineáris modellek egy típusának tekinthető (*Hilbe* [2011]). Ha ez nem teljesül, vagyis a diszperziós paraméter értéke a regressziós együtthatókhoz hasonlóan az adatokból becsü-

lendő, akkor iteratív becsléssel végezhető el a maximum likelihood becslés a paraméterek értékeire (*Zeleis–Kleiber–Jackman* [2008]).

Az általánosított lineáris modellek keretei között a negatív binomiális regressziós modell összekötő függvénye

$$g(\mu) = -\ln\left(\left(\frac{1}{\alpha\mu}\right) + 1\right)$$

alakban írható fel. Ennek inverze (mely a modell által illesztett értékeket adja meg):

$$\mu = \frac{1}{\alpha(e^{x\beta} - 1)}.$$

A negatív binomiális modell maximalizálandó loglikelihood függvénye a következő:

$$\ln L\left(\beta_j; y, \alpha\right) = \sum_{i=1}^{n} y_i \ln\left(\frac{\alpha e^{x_i^T \beta}}{1 + \alpha e^{x_i^T \beta}}\right) - \frac{1}{\alpha}\ln\left(1 + \alpha e^{x_i^T \beta}\right) + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) -$$
$$- ln\Gamma\left(y_i + 1\right) - ln\Gamma\left(\frac{1}{\alpha}\right),$$

ahol α a már ismert diszperziós paraméter, mely v^{-1}-gyel egyenlő (*Hilbe* [2011] 191. old.).

A modellszóródási statisztika jelzi, hogy a Poisson-modellt túlszóródás (vagy alulszóródás) jellemzi, azonban azt, hogy a negatív binomiális regressziós modell valóban jobban illeszkedik-e az adatokhoz, mint a Poisson regressziós modell, különböző tesztek segítségével lehet eldönteni. A leggyakrabban erre a score-, a Lagrange-féle multiplikátor-, a Vuong-, valamint a határ likelihood hányados (boundary likelihood ratio) tesztet alkalmazzák (*Hilbe* [2011]).

A határ likelihood hányados teszt azt vizsgálja, hogy a diszperziós paraméter szignifikánsan eltér-e nullától.[3] A tesztstatisztika értéke azonos a hagyományos likelihood hányados próba esetén alkalmazott számítással: $LR = -2\left(L_P - L_{NB}\right)$, vagyis mínusz kétszer a Poisson-modell loglikelihoodjának és a negatív binomiális modell loglikelihoodjának a különbsége. A teszt a *p*-érték meghatározásának módjában tér el a hagyományos likelihood hányados próbáktól, mivel a határ likelihood

[3] A nullától való eltérés azonban csak „felfelé" tesztelhető, hiszen a negatív binomiális modell diszperziós paramétere nem vehet fel nullánál kisebb értéket, vagyis a teszt a Poisson-modell alulszóródásának tesztelésére nem alkalmas (*Cameron–Trivedi* [1998], *Hilbe* [2011]).

hányados teszt figyelembe veszi az $\alpha = 0$ határt, vagyis azt, hogy a negatív binomiális modell diszperziós paraméterének értéke nem lehet nullánál kisebb. A határ likelihood hányados teszt esetén a tesztstatisztika aszimptotikus eloszlásának egyik fele nulla, másik fele pedig nagyobb annál, egy szabadsági fokú khi-négyzet eloszlású. Határ likelihood hányados próba esetén tehát „a tesztstatisztikához tartozó *p*-érték annak valószínűségének a fele, hogy az egy szabadsági fokú khi-négyzet értéke nagyobb az összehasonlított modellek esetén számított likelihood hányados statisztikánál" (*Cameron–Trivedi* [1998], *Hilbe* [2011] 178. old.).

A teszt egymásba ágyazott, vagyis a Poisson- és a negatív binomiális modellek különböző párjainak (például a zéróinflált Poisson- és a zéróinflált negatív binomiális, a nullában csonkolt Poisson- és a nullában csonkolt negatív binomiális modellek) összehasonlítására használható, amikor is arról hozunk döntést, hogy a túlszóródási paraméter értéke szignifikánsan nagyobb-e nullánál (*Hilbe* [2011]).

Előfordulhat azonban, hogy a negatív binomiális modell alkalmazása esetén is túlszóródást figyelünk meg, illetve, hogy a Poisson-modellt túlszóródás, a negatív binomiálist pedig alulszóródás jellemzi. Ezekben az esetekben a túlszóródás azt jelenti, hogy a becsült modell varianciája meghaladja a modell nominális $\mu + \alpha\mu^2$ varianciáját (tehát $p = 2$, azaz kettes típusú negatív binomiális modellről van szó). A negatív binomiális modell túl- vagy alulszóródása többek között a túl kevés/túl sok zérus értékű előfordulási gyakorisághoz kötődő problémákból fakadhat. Mind a Poisson-, mind pedig a negatív binomiális modellnek léteznek olyan kiterjesztett modelljei (zéróinflált, zérócsonkolt és hurdle-modellek), melyek a túlszóródás túl kevés/túl sok nulla értékhez kötődő eseteit igyekeznek kezelni (*Hilbe* [2011]).

3. A zérus értékekhez kötődő problémák megoldása zéróinflált és hurdle-modellekkel

A zérus értékekhez kötődő túlszóródás egyik esete, amikor adataink egyáltalán nem tartalmaznak zérus értéket, hiszen a Poisson-modellben pozitív valószínűség tartozik a nulla kimenethez is. A zérus értékek hiánya a zérócsonkolt Poisson-modell segítségével orvosolható, melynek valószínűségeloszlás-függvénye: $P(Y = k|Y \neq 0) = \frac{e^{-\lambda}\lambda^k}{k!}$, ha $k > 0$, egyébként pedig 0 (*McDowell* [2003] 179. old.). A zérócsonkolt Poisson-modell túlszóródása esetén a csonkolt modell által feltételezett varianciához képest a zérócsonkolt negatív binomiális modell alkalmazása lehet a megoldás (*Hilbe* [2011]).

A zérus kimenethez kötődő túlszóródás egy másik típusa, amikor adataink a Poisson- vagy a negatív binomiális eloszlás által feltételezettnél több nullát tartalmaznak. Ebben az esetben a kevert eloszlásokat illesztő hurdle- és zéróinflált modell alkalmazható, melyek két komponensük pontos definícióját tekintve térnek el egymástól (*Hilbe* [2011]).

Általános formában ezek a véges kevert eloszlások úgy modellezhetők, hogy a megfigyelések két külön szakaszon mennek keresztül: először azt az átmenetet, mely a zérus és a nem nulla értéket felvevő megfigyeléseket választja el, majd az előfordulási gyakoriságokat modellezik (*Zorn* [1996]).

A zéróinflált modellek esetében van átfedés a keverékeloszlás két komponense között. Ekkor a nulla előfordulási gyakoriságok részei mind a bináris, mind pedig az előfordulási gyakoriságokat modellező folyamatnak, a cél pedig ezek elinflálása. A nulla értékek tehát két forrásból származhatnak, lehetnek „biztosan" zérus értékek, valamint keletkezhetnek a hagyományos előfordulási gyakoriságokat modellező folyamatból. A zéróinflált modell bináris komponense a nulla előfordulási gyakoriságokat becsli, vagyis a „sikeres" (egyes) kimenet azt az eseményt jelöli, amikor az előfordulási gyakoriság biztosan nulla, míg a zérus kimenet a nem nulla értéket.

A zéróinflált modellek feltevése szerint p_i annak a valószínűsége, hogy nulla kimenetet figyelünk meg, míg az $1 - p_i$ valószínűség az adott előfordulási gyakoriságokat modellező eloszlásból származó valószínűségi változót jelöli. Ekkor tehát a nulla előfordulási gyakoriság megfigyelésének valószínűsége:

$$\left(Y_i = 0\right) = p_i + \left(1 - p_i\right)e^{-\lambda_i},$$

ahol p_i az *i*-edik „biztosan zérus" nulla kimenet valószínűsége, $\left(1 - p_i\right)e^{-\lambda_i}$ pedig a Poisson-modellből származó *i*-edik nulla kimeneté. A $k \geq 0$ előfordulási gyakoriságok megfigyelésének valószínűsége a zéróinflált Poisson-modell keretei között (*Lambert* [1992] 3. old.):

$$P\left(Y_i = k\right) = \left(1 - p_i\right)\frac{e^{-\lambda_i}\lambda_i^k}{k!}, \text{ ahol } k = 0, 1, 2...$$

A zéróinflált modell bináris komponensét leggyakrabban logit- vagy probitfüggvénnyel modellezzük, de használható erre cauchit-, cloglog- és logfüggvény is. A nem bináris komponens modellezésére valamely előfordulási gyakoriságokat modellező (Poisson-, negatív binomiális, geometriai) regresszió illeszthető (*Cameron–Trivedi* [1998], *Hilbe* [2011]).

A zéróinflált modellekkel szemben a hurdle-modellekben a két komponens között nincs átfedés, a nulla kimenetek nem két, hanem egy folyamatból származnak, min-

den nulla előfordulási gyakoriság mögött strukturális ok feltételezhető. A bináris komponens a nulla és a pozitív előfordulási gyakoriságok közötti küszöb átlépésének valószínűségét modellezi, míg az ún. countkomponens a pozitív előfordulási gyakoriságokat. Vagyis a modell szerint, amennyiben átlépjük a zérus-nem zérus határvonalat, biztosan pozitív előfordulási gyakoriságokat figyelünk meg (*Hilbe* [2011]).

A hurdle-modellekben az első folyamatot (zérus-nem zérus határ átlépése) általában valamilyen (többek között logit, probit vagy log) bináris modell, utóbbit pedig zérócsonkolt (Poisson-, geometriai, negatív binomiális) modell segítségével becslik.[4]

Ekkor a nulla kimenet valószínűsége $P(Y = 0) = q_i$, ahol q_i annak a valószínűsége, hogy az *i*-edik megfigyelés nem lépi át a nulla és a pozitív kimenetek közötti „gátat". A pozitív előfordulási gyakoriságok valószínűsége (Poisson-modellt feltételezve) a nullában csonkolt modellek segítségével írható fel (*McDowell* [2003] 179. old.):

$$P(Y = k|Y \neq 0) = \frac{e^{-\lambda}\lambda^k}{k!},\ k > 0.$$

A hurdle-modellek esetén a modell specifikációja következtében külön történik a bináris komponens és a pozitív előfordulási gyakoriságok modellezése. E modellek két komponense között nincs átfedés, így azok loglikelihoodja a zéróinflált modellekkel ellentétben elkülöníthető, és külön-külön maximalizálható, a modell loglikelihoodja pedig a bináris modell (zérus-nem zérus átmenet) és a nullában csonkolt Poisson- vagy negatív binomiális komponens loglikelihoodjának összegeként áll elő: $\ln L = \ln\{L_1(\beta_1)\} + \ln\{L_2(\beta_2)\}$, ahol L_1 a bináris szakasz, L_2 pedig a zérócsonkolt Poisson-komponens likelihoodja (*McDowell* [2003] 179. old., *Hilbe* [2011] 356. old., *Cameron–Trivedi* [1998]).

Határ likelihood hányados és Vuong-tesztekkel vizsgálható, hogy a zéróinflált Poisson-modell esetén is fennáll-e a túlszóródás esete, vagyis a modell negatív binomiális párja a túl számos nulla előfordulási gyakoriság figyelembevétele mellett is jobban illeszkedik-e. A két próba a standard Poisson- és a negatív binomiális modellek esetén bemutatott módon működik a zéróinflált Poisson- és a zéróinflált negatív binomiális modellek összehasonlításakor is (*Hilbe* [2011]). A Vuong-teszt a zéróinflált (vagy a hurdle-) modellek túlszóródásának tanulmányozásakor pedig arra ad választ, hogy van-e szignifikáns különbség a zéróinflált Poisson- és a zéróinflált negatív binomiális (illetve a hurdle Poisson- és a hurdle negatív binomiális) modellek által illesztett értékek között (*Hilbe* [2011]).

[4] Azonban a bináris kimenetelű (zéró vagy pozitív előfordulási gyakoriságú) folyamat nem csupán bináris modellel becsülhető, hanem a jobbról „cenzorált" előfordulási gyakoriságok modelljeivel is. A küszöb egyébként nullától eltérő érték is lehet. A hurdle-modellek tehát a túl sok nulla mellett a túl kevés nulla előfordulási gyakoriság esetén is használhatók, azonban valamennyi nullának elő kell fordulnia az adatokban. Általában a túl sok nulla előfordulása esetén alkalmazzák (*Cameron–Trivedi* [1998], *Hilbe* [2011]).

A Vuong-teszt nem egymásba ágyazott modellek összehasonlítására is alkalmazható. A tesztstatisztikát így gyakran használják arra is, hogy megvizsgálják, statisztikailag szignifikánsan jobban illeszkedik-e a zéróinflált vagy a hurdle-modell a standard párjánál (például a zéróinflált Poisson-modell a standard Poissonnál, a zéróinflált negatív binomiális modell a negatív binomiálisnál). *Desmarais–Harden* [2013] azonban arra hívja fel a figyelmet, hogy a kétkomponensű modellekben jóval több paraméter becslésére van szükség, mint a standard Poisson- vagy a negatív binomiális modellekben. Amennyiben a becsült paraméterek számában megfigyelhető különbségekre nem korrigálunk, akkor a teszt „elfogult" lesz a kétkomponensű modellek irányába. Ezért az egy- és kétkomponensű modellek illeszkedésének Vuong-teszttel történő összehasonlításakor érdemes az információs kritériumok (általában az AIC [Akaike information criterion – Akaike információs kritérium] és a BIC [Bayesian information criterion – bayesiánus információs kritérium]) figyelembevételével korrigált tesztstatisztika-értékét alkalmazni.

A hurdle- és a zéróinflált modellek közötti választás a modellek illeszkedésének, reziduálisainak és illesztett értékeinek összevetésén túl leginkább azon alapul, hogy az adatok keletkezése mögött milyen folyamatot feltételezhetünk, tehát kizárólag strukturális nulla értékekről (hurdle-modellről) van-e szó, vagy strukturális és mintavételi nulla értékekről (zéróinflált modellről) egyaránt.

4. A roma ismerősök számának elemzése zéróinflált és a hurdle-modellek alkalmazásával

A zéróinfált és a hurdle-modellek gyakorlati alkalmazhatóságát egy kapcsolathálózati kutatási problémán keresztül mutatom be. (Jelen írás keretei között a kérdés szociológiai hátterét nem ismertetem részletesen, csak röviden hivatkozom a feldolgozott szakirodalomra.) Az elemzés során azt vizsgálom, milyen tényezők befolyásolják, hogy egy válaszadó hány (megítélése szerint) roma származású személyt ismer.

Az elemzésben ismertetett kérdésen kívül a zéróinflált és a hurdle-modellek alkalmasak lehetnek többek között biztosítási kárbejelentések és -események elemzésére is (*Boucher–Denuit–Guillen* [2007], [2009]; *Yip–Yau* [2005]), mivel a legtöbb biztosított nem tesz kárbejelentést. Ezeken kívül még egészségügyi (*Bohning et al.* [1999], *Wang et al.* [2003], *Rose et al.* [2006]) és kapcsolathálózati (*McPherson–Smith-Lovin–Brashears* [2009], *Cornwell–Cornwell* [2008], *Cornwell* [2011]) kutatásokban is találni példát e modellek alkalmazására.

McPherson–Smith-Lovin–Cook [2001] elmélete alapján az emberi kapcsolatok létrejöttének egyik fő meghatározó tényezője a homofília (hasonlóság) jelensége, vagyis az, hogy az emberek elsősorban magukhoz hasonló jellemzőkkel bíró embereket ismernek, velük barátkoznak. A szerzők az etnikai hovatartozást tartják a

homofília által leginkább meghatározott dimenziónak, vagyis úgy gondolják, azonos etnikai csoportba tartozó személyek között inkább létrejönnek kapcsolatok, mint eltérő etnikai csoportba tartozók esetén. *McPherson–Smith-Lovin–Cook* [2001] véleménye szerint a homofília rendező elve nem csupán az erős kötelékeket, hanem az ismerősi kapcsolatokat is áthatja. Az etnikai „választóvonalakat" tovább erősítheti, ha a különböző etnikai csoportok között az etnikai hovatartozáson kívül is jelentős különbségek vannak (például a magyarországi roma és nem roma népesség lakóhelye, iskolai végzettsége és munkaerőpiaci helyzete tekintetében).

Azonban nem minden tényező szempontjából egyformán erőteljes, és egy adott csoportképző változó tekintetében időben változhat is a baráti, ismerősi kapcsolatok homofíliája. Más-más társadalmi csoportokhoz tartozó emberek közötti (vagyis a bizonyos szempontból nem homofil) kapcsolatok létrejöttében fontos szerepe lehet olyan strukturális tényezőknek, melyek e kapcsolatok megvalósulásának lehetőségét befolyásolják. E tényezők közé tartozik például a társadalom demográfiai összetétele, az intézményekben megvalósuló szegregáció, a gazdasági egyenlőtlenségek mértéke (*Smith–McPherson–Smith-Lovin* [2014]).

Minél vegyesebb egy társadalom etnikai összetétele, annál nagyobb például az interetnikus kapcsolatok létrejöttének esélye. Ezzel szemben minél inkább átitatja a homofília elve a bejutást bizonyos intézményekbe, vagy minél nagyobbak a jövedelmi, státusbeli egyenlőtlenségek két csoport között, annál kevesebb csoportközi kapcsolat valósulhat meg (*Blum* [1985], *Smith–McPherson–Smith-Lovin* [2014]). *Feld–Carter* [1998] szerint (különösen az interetnikus) gyenge kötések létrejöttében fontos tényező, hogy van-e társadalmi tér a kötés létrejöttére.

A homofília elve, valamint a homofíliához vezető és az azt erősítő mechanizmusok alapján (*Blum* [1985], *Feld–Carter* [1998], *McPherson–Smith-Lovin–Cook* [2001], *Smith–McPherson–Smith-Lovin* [2014]) azt feltételezem, hogy azoknak van több (maguk által) roma származásúnak minősített ismerőse, akiknek társadalmi helyzetük folytán inkább van esélyük kapcsolatba kerülni romákkal. Magyarországon az önmagukat roma származásúnak valló személyek a teljes lakosságoz képest általánosságban alacsonyabb iskolai végzettséggel rendelkeznek, magasabb arányban élnek községekben és felülreprezentáltak az észak-alföldi és észak-magyarországi régiókban (*KSH* [2018a–t]).

Az előbbi feltevésen túl azt is valószínűsítem, hogy a roma származású emberek több roma származású embert ismernek, mint a nem roma származásúak, illetve azok, akik olyan földrajzi térségekben élnek, vagy olyan alacsony iskolai végzettségű csoportokba tartoznak, ahol a roma népesség aránya az országos adatot meghaladja, több roma származású személyt ismernek. E tényezőkön túl egy adott személy kapcsolathálózatának mérete is hatással lehet arra, hogy a válaszadónak hány roma ismerőse van.

Mint arról már volt szó, az MTA–ELTE Peripato Összehasonlító Társadalmi Dinamika Kutatócsoport „Válság és innováció" címmel végzett 2014. májusban szemé-

lyes megkérdezéssel adatfelvételt a felnőtt magyar lakosság körében (N = 1000), melyben több, a kapcsolathálózatok mérésére alkalmas kérdésblokk is helyet kapott. Jelen elemzés céljára ezek közül az összegző módszer vagy más néven méretgenerátor kérdései felelnek meg. A módszer segítségével egy válaszadó bizonyos csoportba tartozó ismerőseinek száma vagy akár teljes kapcsolathálózatának mérete becsülhető meg.

Más (például a név-, az erőforrás- vagy a pozíciógenerátor-) módszerekhez képest az összegző módszer nem csupán a közeli, személyes kapcsolatrendszer feltárására alkalmas, hanem arra is, hogy képet kapjunk a társadalmi választóvonalakról. Azt méri fel, hogy bizonyos társadalmi csoportokból van-e a válaszadónak ismerőse, és ha igen, akkor mennyi. Az egyes társadalmi csoportokba tartozó ismerősök száma így rávilágíthat arra, hogy azonosítható-e e csoportok tekintetében valamilyen társadalmi választóvonal. A kérdőívmodul elején a társadalmi csoportok mellett bizonyos nevű ismerősök számára is vonatkoznak kérdések, ami a válaszadó ismerőseiből álló kapcsolathálózat méretének meghatározásához nyújt segítséget („Kérem, mondja meg, hogy hány olyan embert ismer, akit … hívnak?") (*Kmetty–Koltai* [2015]).

A már említett összegző kérdésblokk „rákérdez" arra, hogy a válaszadónak hány ismerőse van a különböző társadalmi csoportokból (tehát hány olyan ember van, akinek tudja a nevét, és legalább egy pillanatra leállna beszélni vele, ha találkoznának). (Lásd a Függeléket.) Az elemzés függő változóját az „Önnek hány olyan ismerőse van: aki cigány származású?" kérdésre adott válaszok képezik, vagyis a válaszadón múlik, hogy kit tekint ismerősei közül roma származásúnak. A válaszadók átlagosan 7,16 roma származású ismerőssel rendelkeztek, de 25,5 százalékuknak egyáltalán nem volt ilyen ismerőse.

Az adatfelvételben résztvevők teljes kapcsolathálózatának méretét az összegző kérdésblokk elején található, ismert létszámú csoportokra (bizonyos nevű emberekre) vonatkozó kérdéseken keresztül becsültem meg a *Zheng–Salganik–Gelman* [2006] által ismertetett módszer segítségével. Az összegző módszer szerint a személy adott csoportba tartozó ismerőseinek számát megszorozzuk a csoport népességen belüli arányával,[5] ami a teljes ismerősi kapcsolathálózat méretének becslésére szolgál. Pontosabb eredményt kapunk, ha a különböző csoportokra vonatkozó becslések eredményét átlagoljuk, ezért jelen elemzéshez öt névre[6] vonatkozó ismerősszámot használtam fel (átlag = 249,3, SD = 255,9, N = 889).[7]

[5] Az adott keresztnévvel első névként rendelkezők számát a Belügyminisztérium Nyilvántartások Vezetéséért Felelős Helyettes Államtitkárság adatai alapján becsültem meg (Nyilvantarto.hu).

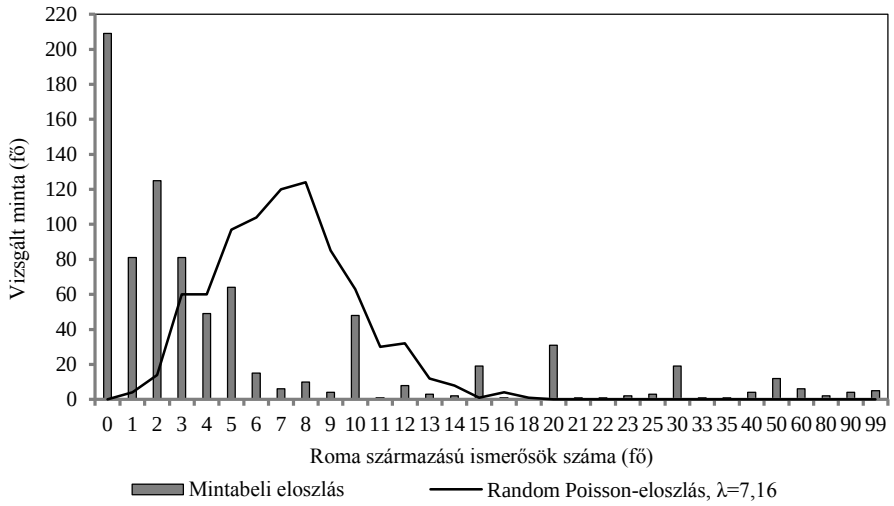
[6] A kérdőívben a következő (a neveket egyesével tartalmazó) kérdések vonatkoztak erre: „Kérem, mondja meg, hogy hány olyan embert ismer, akit Barbarának/Milánnak/Krisztiánnak/Juditnak/Sándornak hívnak?" Egy kérdést azonban nem vettem figyelembe („Kérem, mondja meg, hogy hány olyan embert ismer, akit Júliának hívnak?") a Júlia és Julianna nevű ismerősök megkülönböztetésének esetleges nehézsége miatt.

[7] Természetesen a módszer nem mentes a problémáktól. Nehézséget okozhat, hogy a Belügyminisztérium Nyilvántartások Vezetéséért Felelős Helyettes Államtitkárságának adatai nem teljes körűek, például az igazolványok, iratok nélkül élő emberek adatai nem szerepelnek a rendszerben. Az embereknek emellett nem feltétle-

Ha a kapcsolathálózat kialakulása a „véletlen műve”, Poisson-eloszlásúnak tételezhetjük fel azt, hogy adott társadalmi csoportból ki hány főt ismer; ha viszont strukturális tényezők is befolyásolják a kapcsolathálózat létrejöttét, akkor nem (*Kmetty–Koltai* [2015]).

Az egymintás Kolgomorov–Smirnov-próba eredménye szerint elvethetjük azt a nullhipotézist, hogy a cigány ismerősök száma a vizsgált mintában Poisson-eloszlású; a λ értéke pedig 7,16.[8]

1. ábra. Roma származású ismerősök mintabeli és random Poisson-eloszlása
(*N* = 819 fő)



Forrás: Itt és a továbbiakban saját számítás és készítés az MTA–ELTE Peripato Összehasonlító Társadalmi Dinamika Kutatócsoport „Válság és innováció” című kutatásának adatbázisa alapján.

A következőkben több, előfordulási gyakoriságok elemzésére alkalmas (Poisson-, kvázi-Poisson-, negatív binomiális, zéróinflált Poisson-, zéróinflált negatív binomiális, hurdle Poisson-, hurdle negatív binomiális) modell segítségével vizsgálom, milyen tényezők befolyásolják azt, hogy egy válaszadónak hány roma származású ismerőse van. A modelleket (Vuong-teszttel, határ likelihood hányados próbával) illeszkedés, az általuk adott becslések jellemzői (reziduálisok, prediktált valószínűségek), valamint az együtthatók hatása szempontjából vetem össze egymással. Az elemzést az R-programnyelv *glm* (Poisson-, kvázi-Poisson-regresszió), *glm.nb* (nega-

nül jut eszükbe minden ismerősük egy adatfelvételi szituációban, és az is előfordulhat, hogy az adott keresztnév kapcsán nem is gondolnak bizonyos ismerőseikre, mivel becenevükön szólítják őket.

[8] Az, hogy egy változó marginálisan nem Poisson-eloszlású, nem jelenti azt, hogy feltételesen ne lehetne az. Így a függő változó eloszlásának bemutatása kizárólag leíró statisztikai célokat szolgál.

tív binomiális regresszió), *zeroinfl* (zéróinflált Poisson- és negatív binomiális regresszió), illetve *hurdle* (hurdle Poisson- és hurdle negatív binomiális modell) függvényeinek a segítségével végeztem (*Jackman* [2017], *R Core Team* [2016], *Venables–Ripley* [2002]).

A kétkomponensű modellek esetén a paraméterbecsléshez a BFGS- (Broyden–Fletcher–Goldfarb–Shanno-) algoritmust alkalmaztam, mely egy kvázi-Newton-módszer, ami szerint a paraméterek maximum likelihood becsléséhez nem szükséges a második derivált közvetlen kiszámítása (*R Core Team* [2016]).

Illeszkedés szempontjából azonos változószettel rendelkező modelleket hasonlítok össze. (Lásd az 1. táblázatot.) Az általam tanulmányozott modellek függő változója minden esetben a válaszadó roma származású ismerőseinek száma, magyarázó változóik pedig a következők: a válaszadónak van-e roma származású rokona, mi a legmagasabb iskolai végzettsége, hol van a lakóhelye (településtípus és régió), valamint mekkora az ismerőseiből álló kapcsolathálózat becsült mérete.

Az egymásba ágyazott modelleket (azonos típusú Poisson- és negatív binomiális modellek) illeszkedés szempontjából határ likelihood hányados próbával és Vuong-teszttel, a nem egymásba ágyazottakat pedig *Desmarais–Harden* [2013] ajánlása alapján az AIC-statisztika értékével korrigált Vuong-teszttel hasonlítom össze.

1. táblázat

Azonos változószettel rendelkező modellek összehasonlítása illeszkedés szempontjából

Modell	Határ likelihood hányados próba	Vuong-teszt (vizsgált alternatív hipotézis: a második modell jobban illeszkedik az elsőnél)
Poisson vs. negatív binomiális	$-2 \cdot LR = 4018,\ p < 0,001$	$z = -5,873,\ p < 0,001$
Poisson vs. zéróinflált Poisson	–	$z = -6,821,\ p < 0,001$
Poisson vs. hurdle Poisson	–	$z = -6,870,\ p < 0,001$
Negatív binomiális vs. zéróinflált negatív binomiális		$z = -4,588,\ p < 0,001$
Negatív binomiális vs. hurdle negatív binomiális		$z = -5,292,\ p < 0,001$
Zéróinflált Poisson vs. zéróinflált negatív binomiális	$-2 \cdot LR = 3173,5,\ p < 0,001$	$z = -5,239,\ p < 0,001$
Hurdle Poisson vs. hurdle negatív binomiális	$-2 \cdot LR = 3186,7,\ p < 0,001$	$z = -5,216,\ p < 0,001$

A Vuong-tesztek alapján az egykomponensű modelleknél szignifikánsan jobb az azoknak megfelelő kétkomponensű (zéróinflált vagy hurdle-) modellek illeszkedése. Vagyis a kétkomponensű modellek jobb előrejelzést adnak akkor is, ha figyelembe vesszük, hogy több paramétert tartalmaznak.

A standardizált reziduálisok átlagos értéke a kétkomponensű negatív binomiális modelleknél esik a legközelebb nullához, vagyis átlagosan ezek adják a legpontosabb

becslést a roma származású ismerősök számára nézve. A reziduálisok szórása pedig ugyancsak ezekben, valamint standard párjaikban rendre kisebb, mint az egy- vagy kétkomponensű Poisson-modellekben. (Lásd a 2. ábrát.)

Ha kizárólag a zérus megfigyelt értékekre adott becslések pontosságát vizsgáljuk, megállapítható, hogy a standard negatív binomiális modell esetén összességében jóval nagyobbak a reziduálisok a nulla értékekre viszonyítva, mint a kétkomponensűekben. (Lásd a 3. ábrát.) Vagyis a standard negatív binomiális modell kevésbé pontos becslést ad a roma származású ismerősök nulla számára, mint a kétkomponensű párjai. A reziduálisok értékelésekor azonban azt is fontos mérlegelni, hogy a kétkomponensű modellek esetén magasabb a paraméterek száma, mint az egykomponensűekben, ezért a pontosabb becslés együtt járhat a modellek túlillesztésével.

Az együtthatók hatásnagysága és szignifikanciája tekintetében nagyon hasonlók egymáshoz az azonos típusú (Poisson- vagy negatív binomiális) egy- és kétkomponensű modellek előfordulási gyakoriságokat modellező countkomponensei. Így a terjedelmi korlátokat és az elemzés fókuszát figyelembe véve az együtthatók hatását csak a kétkomponensű modellekre nézve ismertetem részletesen.

A zéróinflált Poisson-, a hurdle Poisson-, illetve a zéróinflált negatív binomiális és a hurdle negatív binomiális modellek előfordulási gyakoriságokat modellező komponensében ugyanazon változóknak van szignifikáns hatása, és e hatások iránya, valamint nagyságrendje is megegyezik egymással. A zéróinflált és a hurdle Poisson-modellekben több a szignifikáns hatás, mint a negatív binomiális párjaikban, ami összefügghet azzal, hogy a Poisson-modell túlszóródás esetén hajlamos alulbecsülni a standard hibákat. A határ likelihood hányados próbák és a Vuong-tesztek alapján az azonos típusú Poisson- és negatív binomiális modellek közül minden esetben az utóbbi mutat szignifikánsan jobb illeszkedést, vagyis a roma származású ismerősök számára ható tényezők modellezésekor érdemes a túlszóródás jelenségét figyelembe venni.

A Poisson- és a hurdle negatív binomiális modellek zéróinflált párjuknál szintén több szignifikáns hatást mutatnak, ami azzal magyarázható, hogy a hurdle-modellek két komponensének illesztése egymástól függetlenül történik, míg a zéróinfláltak esetén van átfedés a két komponens között. Ugyanezen okból egyeznek meg a hurdle Poisson- és a hurdle negatív binomiális modellek bináris komponenseinek együtthatói is, hiszen e modellek csak a countkomponenseik tekintetében térnek el egymástól, mivel azok illesztése egymástól függetlenül történik. Vagyis a zéróinflált Poisson- és a zéróinflált negatív binomiális modellek bináris komponensében egy változó hatása arra nézve, hogy a válaszadónak van-e egyáltalán roma származású ismerőse csak abban az esetben szignifikáns, ha az az adott tényező roma ismerősök számára gyakorolt hatásán felül (countkomponens) is érvényesül.

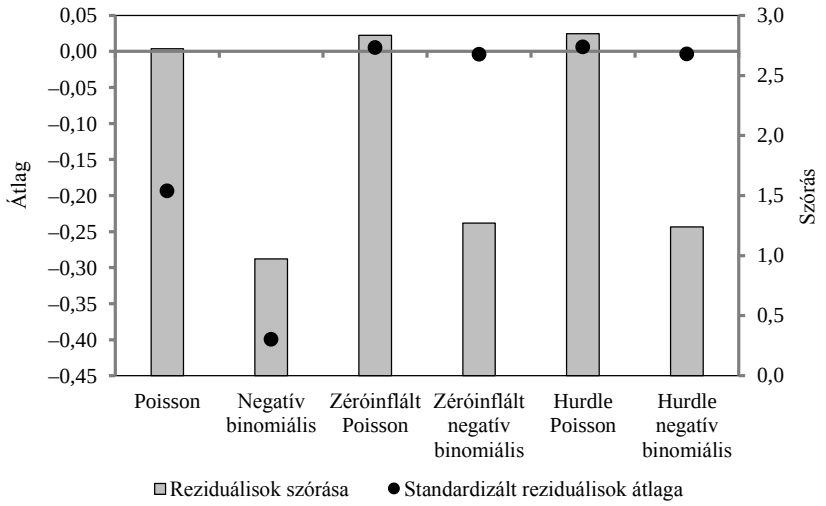
A Vuong-tesztek és a határ likelihood hányados próbák, továbbá a reziduálisok alapján a zéróinflált negatív binomiális és a hurdle negatív binomiális modellek mu-

tatják a legmegfelelőbb illeszkedést a vizsgált modellek közül, ezért ezek eredményeit részletesen is ismertetem. A zéróinflált negatív binomiális modell alapján, minden más tényezőt változatlannak tekintve, egy válaszadó által ismert roma személyek várható számának logaritmusa 1,151, míg a hurdle negatív binomiális modell szerint 1,607. A modellekben a konstans értéke azért különbözik, mert a hurdle-modell countkomponense csupán a pozitív előfordulási gyakoriságokat modellezi, és a két komponense között nincs átfedés. A countkomponensekben azonos tényezők gyakorolnak szignifikáns hatást a roma ismerősök várható számára. Mindkét modellben szignifikánsan növeli a roma ismerősök számának várható értékét, ha a válaszadónak van roma származású rokona, ha a közép-magyarországi régióval szemben Dél-Dunántúlon, Észak-Magyarországon, Dél-Alföldön vagy Észak-Alföldön él, illetve, ha nagy a kapcsolathálózata. A roma származású ismerősök várható számát ezzel szemben szignifikánsan csökkenti, ha a válaszadó nem községben, hanem megyeszékhelyen vagy városban lakik.

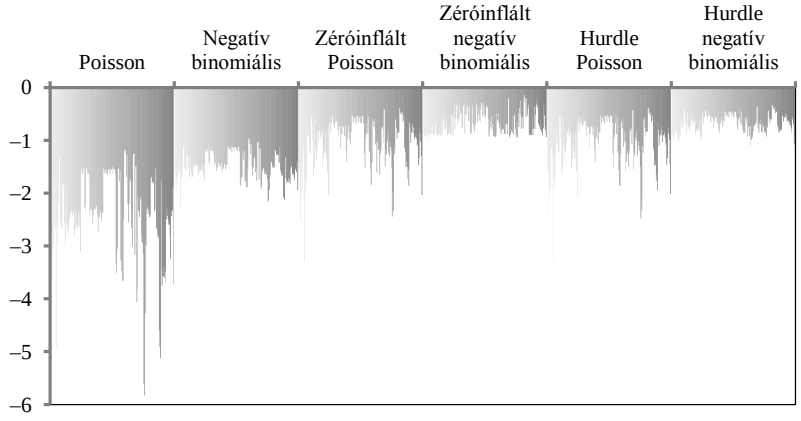
A zéróinflált és a hurdle negatív binomiális modell bináris komponensének konstansa ellentétes irányú: az előbbi szerint annak valószínűségének a logaritmusa, hogy egy válaszadónak biztosan nincsenek roma ismerősei, 2,869, míg az utóbbiban –0,93 annak esélyének a logaritmusa, hogy egy válaszadó több roma származású ismerősről is beszámol (tehát átlépi a zérus-nem zérus határvonalat). A hurdle negatív binomiális modell bináris komponense esetén tapasztalt nagyobb számú szignifikáns hatás azzal magyarázható, hogy a két komponens illesztése egymástól függetlenül történik (a countkomponens csak a pozitív gyakoriságokat modellezi), míg a zéróinflált modell esetén – mint arról már többször szó volt – a komponensek között van átfedés (a countkomponens zérus értékeket is modellez).

A zéróinflált negatív binomiális modell bináris komponense alapján a legfeljebb általános iskolai végzettséggel rendelkező és a városi válaszadók esetén a felsőfokú végzettségűekhez, valamint a községekben élőkhöz képest kisebb a valószínűsége annak, hogy nincs roma származású ismerősük. Elképzelhető tehát, hogy egy községben könnyebben létrejön olyan zárvány, mely meggátolja az eltérő etnikumúak bárminemű érintkezését. A hurdle negatív binomiális modell szerint ezzel szemben egyik iskolai végzettséget mérő kategória esetén sem szignifikáns az összehasonlítás, a községben élőkhöz képest viszont a fővárosiak esélye szignifikánsan nagyobb, míg a megyeszékhelyen élőké szignifikánsan kisebbb arra, hogy legyen legalább egy roma származású ismerősük. A zéróinflált negatív binomiális modellben a kapcsolathálózat egy egységnyi növekedése szignifikánsan csökkenti annak esélyét, hogy a válaszadónak ne legyen roma ismerőse, a hurdle-modell értelmezése szerint pedig növeli annak esélyét, hogy legalább egy roma ismerőse legyen. Mindkét modellben szignifikáns a régiók közül a közép-magyarországi régióhoz képest az észak-magyarországi lakóhely hatása, a hurdle negatív binomiális modellben pedig emellett még a nyugat-dunántúlié, a dél-dunántúlié és az észak-alföldié is.

2. ábra. A modellek standardizált reziduálisainak átlaga és szórása különböző modelltípusok esetén



3. ábra. A különböző modelltípusok reziduálisai a megfigyelt zérus értékekre nézve



Megjegyzés. Az ábra azt mutatja be, hogy az adott modell mekkora reziduálissal becsülte a zérus megfigyeléseket. Minél nagyobb a szürke terület, a modell annál pontatlanabb becslést adott rájuk.

A modellek értékelésekor a statisztikai mérőszámokon túl érdemes elméleti megfontolásokat is figyelembe venni. *Rose et al.* [2006] véleménye szerint amennyiben az adatfelvétel és a kutatástervezés a strukturális és a mintavételi nulla értékeket egyaránt lehetővé teszi, a zéróinflált modellek jelenthetik a megfelelő választást, ha azonban a kutatástervezés miatt az adatok csak mintavételi nullákat tartalmaznak, a hurdle-modellek.

E megfontolásokat figyelembe véve megállapítható, hogy a zéróinflált modellekben egyaránt szerepelhetnek olyan válaszadók, amelyeknek valamilyen strukturális ok miatt nincs lehetőségük romákkal találkozni, és olyanok is, amelyek esetén a nulla kimenetnek nincs strukturális oka. A hurdle-modellek ugyanakkor kizárólag a strukturális nulla előfordulási gyakoriságokat engedik meg. A vizsgált kérdés tekintetében strukturális ok lehet például, hogy egy válaszadó olyan településen lakik, ahol egyáltalán nem élnek romák, és ő pedig közvetlen lakókörnyezetét (például idős kora, betegsége miatt) sohasem hagyja el, vagy, ha egyáltalán nem ápol „ismerősi" kapcsolatokat (teljesen elzárkózik a külvilágtól/az emberi kapcsolatoktól, csupán a legközelebbi családtagjaival érintkezik). Úgy vélem, valóságszerűbb az a feltételezés, hogy a roma ismerőssel nem rendelkező válaszadók közül egyeseknek strukturális okokból, másoknak a véletlennek köszönhetően nincsenek roma származású ismerősei mintsem, hogy a kapcsolat hiányát minden esetben strukturális oknak tulajdonítsuk.

Elméletileg is indokolható, hogy miért illeszkednek jobban mind az egy-, mind pedig a kétkomponensű negatív binomiális modellek a Poisson-modelleknél, ugyanis nem életszerű, hogy az emberek azonos valószínűséggel, társadalmi helyzetüktől függetlenül ismernek bizonyos társadalmi csoportokba tartozó embereket (*Diprete et al*. [2011]).

Az elemzés eredményeinek értelmezésekor a következő elméleti megfontolásokat is figyelembe vettem: egyrészt nem egyértelmű az ismerősök válaszadó által történő etnikai besorolása, másrészt gyakran nem ugyanabba a társadalmi csoportba sorolják be az emberek saját magukat, mint a környezetük őket (bár *Kemény–Janky* [2006] eredményei szerint nagy az átfedés ezek között). *Ladányi–Szelényi* [2006], valamint *Csepeli–Simon* [2004] arra a következtetésre jutottak, hogy a külső környezet által alkotott besorolás sokszor a társadalmi státust kifejező tényezőkkel (munkaerőpiaci helyzettel, anyagi helyzettel, lakókörnyezettel stb.) függ össze. Vagyis a homofília elvén túl feltételezhetően azért is kevesebb a magasabb iskolai végzettségűek vagy az aktív munkaerőpiaci státusúak roma származású ismerőse, mert őket környezetük kevésbé minősíti roma származásúnak, mint az alacsonyabb társadalmi státusúakat. Az eredmények abban a keretben értelmezhetők, hogy a függő változó „külső besoroláson" alapul, etnikai önbesorolás pedig nem áll rendelkezésre. Az utóbbi helyett az vehető figyelembe, hogy van-e a válaszadónak roma származású családtagja.

Mindezeken túl azzal is számolni kell, hogy bár a kutatási kérdőívben szerepelt „felvezető" szöveg a vizsgált, ismerősök számára vonatkozó kérdések előtt, a válaszadók nem biztos, hogy egyformán ítélték meg, kit tekintenek ismerősüknek. Lehetséges, hogy ennek eldöntésekor különböző tényezőket vettek tekintetbe, és az sem biztos, hogy az adatfelvétel idején eszükbe jutott minden, (szerintük) az adott etnikai csoportba tartozó ismerősük. E problémák az ismerősök alkotta kapcsolathálózat méretének becslését is érintik.

2. táblázat

A kétkomponensű modellek eredményei
(N = 803 fő)

Megnevezés	Zéróinflált Poisson	Zéróinflált negatív binomiális	Hurdle Poisson	Hurdle negatív binomiális
	Countkomponens			
Konstans	1,605*** (0,09)	1,151*** (0,21)	1,149*** (0,09)	1,607*** (0,26)
Roma rokon (referencia: nincs)	1,043*** (0,04)	1,315*** (0,17)	1,354*** (0,04)	1,043*** (0,19)
Iskolai végzettség (referencia: főiskola/egyetem)				
Legfeljebb általános iskola	–0,126* (0,05)	–0,006 (0,16)	0,009* (0,05)	–0,122 (0,18)
Szakmunkásképző	0,050 (0,05)	0,110 (0,14)	0,134 (0,05)	0,053 (0,17)
Középiskola	–0,022 (0,09)	–0,040 (0,14)	0,001 (0,05)	–0,019 (0,16)
Településtípus (referencia: község)				
Főváros	–0,195*	–0,142 (0,20)	–0,254* (0,09)	–0,199 (0,24)
Megyeszékhely	–1,205*** (0,05)	–1,193*** (0,15)	–1,266*** (0,05)	–1,212*** (0,18)
Város	–0,786*** (0,03)	–0,706*** (0,11)	–0,816*** (0,03)	–0,786*** (0,13)
Régió (referencia: Közép-Magyarország)				
Közép-Dunántúl	–0,309** (0,10)	–0,102 (0,20)	–0,362** (0,10)	–0,321 (0,25)
Nyugat-Dunántúl	0,345*** (0,09)	0,407 (0,22)	0,451*** (0,09)	0,337 (0,26)
Dél-Dunántúl	0,674*** (0,09)	1,276*** (0,23)	1,227*** (0,09)	0,668*** (0,26)
Észak-Magyarország	1,316*** (0,08)	1,424*** (0,19)	1,345*** (0,08)	1,311*** (0,22)
Észak-Alföld	0,998*** (0,08)	1,294*** (0,19)	1,277*** (0,08)	0,995*** (0,22)
Dél-Alföld	0,627*** (0,09)	0,641*** (0,20)	0,691*** (0,09)	0,622** (0,24)
Kapcsolathálózat mérete	0,001*** (0,00)	0,001** (0,00)	0,001*** (0,00)	0,001*** (0,00)
Túlszóródási paraméter logaritmusa	–	0,00005	–	–0,255*
	Bináris komponens			
Konstans	0,963* (0,40)	2,869** (1,09)	–0,930* (0,38)	–0,930* (0,38)
Roma rokon (referencia: nincs)	–15,730 (88,83)	–15,734 (1274,78)	15,730 (84,90)	15,730 (84,90)
Iskolai végzettség (referencia: főiskola/egyetem)				
Legfeljebb általános iskola	–0,597 (0,38)	–2,075* (0,98)	0,474 (0,35)	0,474 (0,35)
Szakmunkásképző	–0,392 (0,31)	–1,312 (0,71)	0,352 (0,30)	0,352 (0,30)
Középiskola	0,061 (0,30)	–0,28 (0,71)	–0,079 (0,28)	–0,079 (0,28)
Településtípus (referencia: község)				
Főváros	–0,718* (0,35)	–1,557 (0,92)	0,691* (0,34)	0,691* (0,34)
Megyeszékhely	1,064*** (0,33)	1,217 (0,90)	–1,240*** (0,30)	–1,240*** (0,30)
Város	–0,227 (0,27)	–1,514* (0,72)	0,041 (0,25)	0,041 (0,25)
Régió (referencia: Közép-Magyarország)				
Közép-Dunántúl	–2,120*** (0,54)	–4,276* (1,95)	1,552*** (0,38)	1,552*** (0,38)
Nyugat-Dunántúl	–0,759 (0,40)	–1,839 (1,63)	0,804* (0,38)	0,804* (0,38)
Dél-Dunántúl	–3,040*** (0,70)	–5,785 (6,26)	3,045*** (0,63)	3,045*** (0,63)
Észak-Magyarország	–2,214*** (0,45)	–2,078* (0,94)	2,354*** (0,44)	2,354*** (0,44)
Észak-Alföld	–1,722*** (0,39)	–1,492 (0,83)	1,858*** (0,38)	1,858*** (0,38)
Dél-Alföld	–0,342 (0,34)	0,140 (0,73)	0,465 (0,32)	0,465 (0,32)
Kapcsolathálózat mérete	–0,005*** (0,00)	–0,031* (0,01)	0,005*** (0,00)	0,005*** (0,00)

Megjegyzés. * $p < 0,05$, ** $p < 0,01$, *** $p < 0,10$. Zárójelben a standardhiba-értékeket tüntettem fel.

5. Összegzés

A tanulmány egy, az előfordulási gyakoriságok modellezéséhez kapcsolódó problémával, a túl sok zérus értékből fakadó túlszóródás lehetséges kezelésével foglalkozik kétkomponensű modellek segítségével. A túlszóródás különböző eseteinek és kezelésüknek az áttekintését követően az MTA–ELTE Peripato Összehasonlító Társadalmi Dinamika Kutatócsoport által végzett „Válság és innováció" című adatfelvétel keretében megkérdezett válaszadók roma ismerőseinek számát elemzi példaként a kétkomponensű modellek lehetséges társadalomtudományi alkalmazására.

A vizsgált adatokon a negatív binomiális modell a Poisson-modell túlszóródásának egy jelentős részét korrigálni tudta, illeszkedés szempontjából (a Vuong-teszt, a határ likelihood hányados próba, illetve az AIC-statisztika értéke alapján) viszont a kétkomponensű negatív binomiális modellek valamivel jobbak, reziduálisaik kisebbek. Ismerősök számának elemzésekor és annak a válaszadóknak szóló kérdésnek a vizsgálatakor, hogy mely tényezők befolyásolják ismeretségeik kialakulását és másokkal való kapcsolatba lépésük lehetőségét, tehát lehet hozzáadott értéke a kétkomponensű modelleknek. Alkalmazásukkor ugyanakkor érdemes számításba venni, hogy becslésük más modellekhez képest bonyolultabb eljárás, több paraméter becslését igényli (e megállapítás különösen a kétkomponensű negatív binomiális modellekre igaz, ahol a paraméterek számát egyrészt a két komponens, másrészt a diszperziós paraméter is növeli). Ezen túl kis minták esetén felmerülhet az a probléma, hogy e modellek a becsült paraméterek magas száma miatt nem futnak le (*Atkins–Gallop* [2007]); és az általuk becsült túl sok paraméter a túlillesztésükhöz vezethet.

A vizsgált zéróinflált és hurdle Poisson-modellekben több szignifikáns hatás figyelhető meg, mint a negatív binomiális párjaikban, ami azzal magyarázható, hogy a Poisson-modell túlszóródás esetén hajlamos alulbecsülni a standard hibákat. Továbbá a hurdle Poisson- és a hurdle negatív binomiális modellek esetén a zéróinflált megfelelőjükhöz képest több a szignifikáns hatást kifejtő változó; ez azzal lehet összefüggésben, hogy a hurdle-modellek két komponensének illesztése egymástól függetlenül történik, míg a zéróinfláltaké között van átfedés.

Egyes tényezők minden vizsgált kétkomponensű modell szerint szignifikánsan befolyásolják a roma ismerősök számát. Amennyiben a válaszadónak van roma származású rokona, roma ismerőseinek várható száma szignifikánsan magasabb lesz, mint azoknak, akiknek nincs (az összes egyéb vizsgált tényezőre kontrollálva). Ugyanez igaz akkor is, ha a válaszadó (minden más tényezőt változatlanul hagyva) nem Közép-Magyarországon, hanem Dél-Dunántúlon, Észak-Magyarországon, Észak-, illetve Dél-Alföldön lakik. A megyeszékhelyeken vagy a városokban élőknek szignifikánsan kevesebb a roma ismerőse, mint a községbelieknek. Továbbá, a válaszadó becsült kapcsolathálózatának egy egységnyi növekedése szignifikánsan pozitív hatással van roma származású ismerőseinek várható számára.

Mind a négy vizsgált kétkomponensű modell bináris komponense szerint az, hogy a válaszadó Közép-Magyarország helyett Közép-Dunántúlon vagy Észak-Magyarországon él, szignifikáns hatást gyakorol arra, hogy van-e roma származású ismerőse. Emellett a válaszadó kapcsolathálózatának egy egységnyi növekedése szintén minden modellben szignifikáns hatást gyakorol annak esélyére, hogy van–e roma származású ismerőse. A bináris komponensek tekintetében azt, hogy van-e roma ismerőse a válaszadónak olyan strukturális tényezők befolyásolják, mint az összes ismerős száma vagy a roma származású emberekkel való találkozás fizikai (földrajzi) gátjának hiánya/megléte. Azokban a régiókban, ahol a népszámlálási adatok alapján magasabb a roma népesség aránya a teljes népességen belül, nagyobb az esélye, hogy valakinek van legalább egy roma származású ismerőse.

Függelék

Jelen tanulmányban az MTA–ELTE Peripato Összehasonlító Társadalmi Dinamika Kutatócsoport „Válság és innováció" című kutatásának kérdőívében szereplő összegző kérdésblokkból a következő kérdéseket használtam fel:

> Kérem, mondja meg, hogy hány olyan embert ismer, akit Barbarának hívnak?
> Kérem, mondja meg, hogy hány olyan embert ismer, akit Milánnak hívnak?
> Kérem, mondja meg, hogy hány olyan embert ismer, akit Krisztiánnak hívnak?
> Kérem, mondja meg, hogy hány olyan embert ismer, akit Juditnak hívnak?
> Kérem, mondja meg, hogy hány olyan embert ismer, akit Sándornak hívnak?
> Kérem, mondja meg, hogy hány olyan embert ismer, aki cigány származású?

Irodalom

AGRESTI, A. [2002]: *Categorical Data Analysis*. John Wiley & Sons. New York. http://dx.doi.org/10.1002/0471249688

ATKINS, D. C. – GALLOP, R. J. [2007]: Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of Family Psychology*. Vol. 21. No. 4. pp. 726–735. http://dx.doi.org/10.1037/0893-3200.21.4.726

BLUM, T. [1985]: Structural constraints on interpersonal relations: a test of Blau's macrosociological theory. *American Journal of Sociology*. Vol. 91. Issue 3. pp. 511–521. http://dx.doi.org/10.1086/228312

BM NYILVÁNTARTÁSOK VEZETÉSÉÉRT FELELŐS HELYETTES ÁLLAMTITKÁRSÁG [2018]: *Közérdekű utónevek*. http://www.nyilvantarto.hu/letoltes/statisztikak/kozerdeku_utonevek2014.xls

BÖHNING, D. – DIETZ, E. – SCHLATTMANN, P. – MENDONCA, L. – Kirchner, U. [1999]: The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society*: *Series A* (*Statistics in Society*). Vol. 162. No. 2. pp. 195–209. https://doi.org/10.1111/1467-985X.00130

BOUCHER, J. P. – DENUIT, M. – GUILLEN, M. [2009]: Number of accidents or number of claims? An approach with zero-inflated Poisson models for panel data. *Journal of Risk and Insurance*. Vol. 76. Issue 4. pp. 821–846. https://doi.org/10.1111/j.1539-6975.2009.01321.x

BOUCHER, J. P. – DENUIT, M. – GUILLÉN, M. [2007]: Risk classification for claim counts: a comparative analysis of various zeroinflated mixed Poisson and hurdle models. *North American Actuarial Journal*. Vol. 11. Issue 4. pp. 110–131. http://dx.doi.org/10.1080/10920277.2007.10597487

CAMERON, A. C. – TRIVEDI, P. K. [1998]: *Regression Analysis of Count Data*. Cambridge University Press. Cambridge. http://dx.doi.org/10.1017/CBO9780511814365

CORNWELL, B. [2011]: Age trends in daily social contact patterns. *Research on Aging*. Vol. 33. Issue 4. pp. 598–631. http://dx.doi.org/10.1177/0164027511409442

CORNWELL, E. Y. – CORNWELL, B. [2008]: Access to expertise as a form of social capital: an examination of race- and class-based disparities in network ties to experts. *Sociological Perspectives*. Vol. 51. Issue 4. pp. 853–876. http://dx.doi.org/10.1525/sop.2008.51.4.853

CSEPELI, GY. – SIMON, D. [2004]: Construction of Roma identity in Eastern and Central Europe: perception and self-identification. *Journal of Ethnic and Migration Studies*. Vol. 30. Issue 1. pp. 129–150. http://dx.doi.org/10.1080/1369183032000170204

DESMARAIS, B. A. – HARDEN, J. J. [2013]: Testing for zero inflation in count models: bias correction for the Vuong test. *The Stata Journal*. Vol. 13. No. 4. pp. 810–835.

DIPRETE, T. A. – GELMAN, A. – MCCORMICK, T. – TEITLER, J. – ZHENG, T. [2011]: Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology*. Vol. 116. Issue 4. pp. 1234–1283. http://dx.doi.org/10.1086/659100

FELD, S. L. – CARTER, W. C. [1998]: When desegregation reduces interracial contact: a class size paradox for weak ties. *American Journal of Sociology*. Vol. 103. No. 5. pp. 1165–1186. http://dx.doi.org/10.1086/231350

HILBE, J. M. [2011]: *Negative Binomial Regression*. Cambridge University Press. Cambridge. https://doi.org/10.1017/CBO9780511973420

JACKMAN, S. [2017]: pscl: *Classes and Methods for R Developed in the Political Science Computational Laboratory*. R package version 1.5.2. United States Studies Centre, University of Sydney. Sydney. https://github.com/atahk/pscl/

KEMÉNY, I. – JANKY, B. [2006]: Roma population of Hungary 1971–2003. In: *Kemény, I.* (ed.): *Roma of Hungary*. East European Monographs. Bradenton. pp. 70–225.

KMETTY Z. – KOLTAI J. A. [2015]: Kapcsolathálózatok mérése – elméleti és gyakorlati dilemmák, lehetőségek. *Socio.hu, Társadalomtudományi Szemle*. 4. sz. 34–49. old. http://dx.doi.org/10.18030/socio.hu.2015.4.34

KMETTY Z. – KOLTAI J. A. [2016]: Státuszelérés, társas támogatás, társadalmi törésvonalak – A kapcsolathálózati integráció aspektusai. *Socio.hu, Társadalomtudományi Szemle.* 3. sz. 1–21. old. http://dx.doi.org/10.18030/socio.hu.2016.3.1

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2014a]: *Népszámlálás 2011 – 9. Nemzetiségi adatok*. Online jelentés. http://www.ksh.hu/docs/hun/xftp/idoszaki/nepsz2011/nepsz_09_2011.pdf

KSH [2014b]: *Népszámlálás 2011 – 13. A népesség gazdasági aktivitása.* Online jelentés. http://www.ksh.hu/docs/hun/xftp/idoszaki/nepsz2011/nepsz_13_2011.pdf

KSH [2014c]: *Népszámlálás 2011 – 14. A népesség iskolázottsága*. Online jelentés. http://www.ksh.hu/docs/hun/xftp/idoszaki/nepsz2011/nepsz_14_2011.pdf

KSH [2018a]: *Területi adatok – Budapest.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/01/01_1_1_6_1.xls
KSH [2018b]: *Területi adatok – Bács-Kiskun megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/03/03_1_1_6_1.xls
KSH [2018c]: *Területi adatok – Baranya megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/02/02_1_1_6_1.xls
KSH [2018d]: *Területi adatok – Békés megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/04/04_1_1_6_1.xls
KSH [2018e]: *Területi adatok – Borsod-Abaúj-Zemplén megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/05/05_1_1_6_1.xls
KSH [2018f]: *Területi adatok – Csongrád megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/06/06_1_1_6_1.xls
KSH [2018g]: *Területi adatok – Fejér megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/07/07_1_1_6_1.xls
KSH [2018h]: *Területi adatok – Győr-Moson-Sopron megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/08/08_1_1_6_1.xls
KSH [2018i]: *Területi adatok – Hajdú-Bihar megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/09/09_1_1_6_1.xls
KSH [2018j]: *Területi adatok – Heves megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/10/10_1_1_6_1.xls
KSH [2018k]: *Területi adatok – Jász-Nagykun-Szolnok megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/16/16_1_1_6_1.xls
KSH [2018l]: *Területi adatok – Komárom-Esztergom megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/11/11_1_1_6_1.xls
KSH [2018m]: *Területi adatok – Nógrád megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/12/12_1_1_6_1.xls
KSH [2018n]: *Területi adatok – Pest megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/13/13_1_1_6_1.xls
KSH [2018o]: *Területi adatok – Somogy megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/14/14_1_1_6_1.xls
KSH [2018p]: *Területi adatok –Szabolcs-Szatmár-Bereg megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/15/15_1_1_6_1.xls
KSH [2018q]: *Területi adatok – Tolna megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/17/17_1_1_6_1.xls
KSH [2018r]: *Területi adatok – Vas megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/18/18_1_1_6_1.xls
KSH [2018s]: *Területi adatok – Veszprém megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/19/19_1_1_6_1.xls
KSH [2018t]: *Területi adatok – Zala megye.* http://www.ksh.hu/nepszamlalas/docs/tablak/teruleti/20/20_1_1_6_1.xls
LADÁNYI, J. – SZELÉNYI, I. [2006]: *Patterns of Exclusion: Constructing Gypsy Ethnicity and the Making of an Underclass in Transitional Societies of Europe*. East European Monographs. Bradenton.

LAMBERT, D. [1992]: Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. Vol. 34. Issue 1. pp. 1–14. http://dx.doi.org/10.2307/1269547

MCDOWELL, A. [2003]: From the help desk: hurdle models. *The Stata Journal*. Vol. 3. No. 2. pp. 178–184.

MCPHERSON, M. – SMITH-LOVIN, L. — BRASHEARS, M. E. [2009]: Models and marginals: using survey evidence to study social networks. *American Sociological Review*. Vol. 74. Issue 4. pp. 670–681. http://dx.doi.org/10.1177/000312240907400409

MCPHERSON, M. – SMITH-LOVIN, L. – COOK, J. M. [2001]: Birds of a feather: homophily in social networks. *Annual Review of Sociology*. Vol. 27. Issue 1. pp. 415–444. http://dx.doi.org/10.1146/annurev.soc.27.1.415

MOKSONY F. [2006]: A Poisson-regresszió alkalmazása a szociológiai és demográfiai kutatásban. *Demográfia*. 49. évf. 4. sz. 366–382. old.

R CORE TEAM [2016]: *The R Project for Statistical Computing*. R Foundation for Statistical Computing. Vienna. https://www.R-project.org/

ROSE, C. E. – MARTIN, S. W. – WANNEMUEHLER, K. A. – PLIKAYTIS, B. D. [2006]: On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*. Vol. 16. Issue 4. pp. 463–481. http://dx.doi.org/10.1080/10543400600719384

SMITH, J. A. – MCPHERSON, M. – SMITH-LOVIN, L. [2014]: Social distance in the United States: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004. *American Sociological Review*. Vol. 79. Issue 3. pp. 432–456. http://dx.doi.org/10.1177/0003122414531776

VENABLES, W. N. – Ripley, B. D. [2002]: *Modern Applied Statistics with S. Fourth Edition*. Springer. New York.

VUONG, Q. H. [1989]: Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*: *Journal of the Econometric Society*. Vol. 57. No. 2. pp. 307–333. http://dx.doi.org/10.2307/1912557

WANG, K. – LEE, A. H. – YAU, K. K. – CARRIVICK, P. J. [2003]: A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention*. Vol. 35. Issue 4. pp. 625–629. http://dx.doi.org/10.1016/S0001-4575(02)00036-2

YIP, K. C. – Yau, K. K. [2005]: On modeling claim frequency data in general insurance with extra zeros. *Insurance*: *Mathematics and Economics*. Vol. 36. Issue 2. pp. 153–163. http://dx.doi.org/10.1016/j.insmatheco.2004.11.002

ZEILEIS, A. – KLEIBER, C. – JACKMAN, S. [2008]: Regression models for count data in R. *Journal of Statistical Software*. Vol. 27. Issue 8. pp. 1–25.

ZHENG, T. – SALGANIK, M. J. – GELMAN, A. [2006]: How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association*. Vol. 101. No. 474. pp. 409–423. http://dx.doi.org/10.1198/016214505000001168.

ZORN, C. J. [1996]: Evaluating zero-inflated and hurdle Poisson specifications. *Midwest Political Science Association*. 18–20 April. pp. 1–16.

Summary

The aim of this paper is to introduce two types of count models (the zero-inflated and the hurdle models), and to present one of their possible applications in social sciences: analysing the number of Romany acquaintances of respondents that took part in one of the HAS–ELTE (Hungarian Academy of Sciences – Eötvös Loránd University) Peripato Comparative Social Dynamics Research Group's survey. Zero-inflated and hurdle two-component models can improve the accuracy of estimates if too many zero values lead to the Poisson's model overdispersion in data. The paper takes a brief look at some of the major cases of overdispersion in the Poisson model, paying particular attention to problems related to zero values and their remedies by means of zero-inflated or hurdle models. It compares the analysis of the number of Romany acquaintances in several count models (Poisson and negative binomial, zero-inflated Poisson and zero-inflated negative binomial, hurdle Poisson and hurdle negative binomial models). On the one hand, the results suggest that two-component models increase the accuracy of estimations and can be used to investigate which factors influence one's ability to know someone, to get in touch with that person. On the other hand, the estimation of two-component models requires the estimation of several parameters, which can lead to overdispersion.