KÖZLEMÉNYEK

A POISSON-REGRESSZIÓ ALKALMAZÁSA A SZOCIOLÓGIAI ÉS DEMOGRÁFIAI KUTATÁSBAN

MOKSONY FERENC

Mind a szociológiai, mind a demográfiai elemzésekben sűrűn találkozunk olyan függő változókkal, amelyek valamilyen esemény *előfordulási gyakoriságát* fejezik ki. A deviáns viselkedéssel foglalkozó kutatók például vizsgálják a társadalmi változásoknak az öngyilkosságok számára gyakorolt hatását; a demográfusok kutatják, hogy miként befolyásolják a különféle környezeti ártalmak a halálozások vagy a születési rendellenességek számát; a tudományszociológia művelői pedig elemzik azokat a tényezőket, amelyek a szakmai publikációkra történő hivatkozások számát meghatározzák. Mindezek során a kutatók sokszor a hagyományos lineáris regresszióelemzés módszerét alkalmazzák. Ez bizonyos esetekben elfogadható eredményre vezet, különösen akkor, ha a vizsgált jelenség átlagos előfordulási gyakorisága viszonylag *nagy*, ilyenkor ugyanis a függő változó eloszlása rendszerint nem tér el túlságosan a lineáris regresszió által feltételezett normális eloszlástól.

Más a helyzet akkor, ha a vizsgált jelenség statisztikai értelemben *ritka* előfordulású esemény – mint amilyen például a születési rendellenesség, a halálos kimenetelű közlekedési baleset vagy éppen az öngyilkosság. Ilyenkor a változó eloszlása többnyire meglehetősen *ferde*: a megfigyelések jelentős része a legalacsonyabb értékek körül tömörül, s innen a magasabb értékek felé haladva a gyakoriság meredeken csökken. Jól szemlélteti ezt az I. ábra, amelyen a Magyarország községeiben 1990 és 1995 között elkövetett öngyilkosságok gyakorisági eloszlása látható.[1] A normális eloszlás szimmetrikus haranggörbéje helyett a rajz egy erősen aszimmetrikus, jobbra ferde eloszlást mutat: közel 800 faluban a vizsgált hat év során egyáltalán nem történt öngyilkosság, s további csaknem 600-ban is csupán egyetlen esetet jegyeztek fel. Olyan település viszont, ahol 20-nál több öngyilkosságot követtek el, mindössze 38 akadt. Az ilyen aszimmetrikus, jobbra ferde eloszlások általában jól közelíthetők a diszkrét eloszlások egyik fontos típusával, a *Poisson-eloszlással*.[2]

A Poisson-eloszlás egyik fontos tulajdonsága, hogy esetében az *átlag egyenlő a varianciával*. Ez könnyen belátható, ha a Poisson-eloszlást úgy tekintjük, mint a binomiális eloszlás szélső esetét – mégpedig olyan szélső esetét, amikor a két lehetséges

[1] A rajzon azoknak a településeknek az adatai szerepelnek, amelyek 1990-ben községnek minősültek, s ezt a jogállásukat egészen 1995-ig megőrizték.

[2] Amint a vizsgált esemény egyre gyakoribbá válik, azaz amint az eloszlás átlaga emelkedik, a Poisson-eloszlás ferdesége fokozatosan csökken, és alakja mind közelebb kerül a normális eloszláséhoz (lásd pl. Sváb 1981. 498., 19.4–2. ábra). A Poisson-eloszlásnak ezért a *ritka* előfordulású események vizsgálatában van különös jelentősége (vö. Land – McCall 1996).

Demográfia, 2006. 49. évf. 4. szám 366–382.

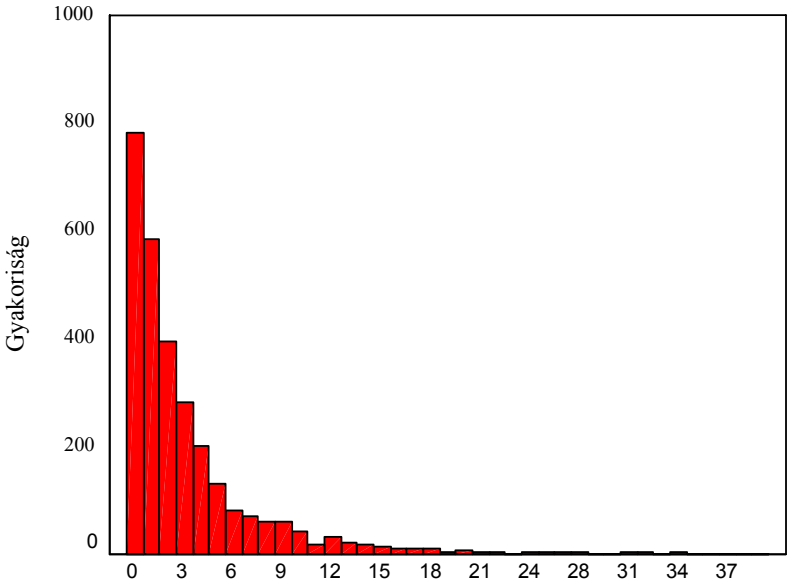
kimenetel (pl. öngyilkos/nem öngyilkos, rendellenességgel született/nem rendellenességgel született) egyikének a valószínűsége rendkívül kicsi. Egy binomiális eloszlású változó átlaga

$$\bar{x} = np,$$

varianciája pedig

$$s^2 = np(1-p),$$

ahol n a megfigyelések száma és p a két lehetséges kimenetel egyikének (pl. az öngyilkosság vagy a születési rendellenesség bekövetkezésének) a relatív gyakorisága. Mi történik a varianciával, amint p értéke egyre kisebb és kisebb lesz, vagyis amint az egyik kimenetel relatív gyakorisága egyre csökken? A következmény az, hogy az $(1–p)$ szorzótényező fokozatosan tart 1-hez, ennek nyomán pedig s^2, azaz a variancia fokozatosan tart np-hez, ami nem más, mint az átlag.



I. Az öngyilkosságok számának gyakorisági eloszlása Magyarország községeiben, 1990–1995
Frequency distribution of suicide in Hungarian villages, 1990–1995

Hogyan érinti az átlagnak és a varianciának ez az egyenlősége a hagyományos lineáris regresszióelemzés alkalmazhatóságát? A regresszióelemzés során abból indulunk ki, hogy a függő változó átlaga a magyarázó változók valamilyen függvénye, vagyis az átlag együtt mozog ezeknek a változóknak az értékeivel. Poisson-eloszlású változók

esetében ez egyszersmind azt is jelenti, hogy a változó *varianciája* is együtt mozog a magyarázó változókkal. Épp ez a probléma: a hagyományos lineáris regresszió egyik alapfeltevése ugyanis, hogy a függő változó szórása a magyarázó változó minden egyes értéke vagy kategóriája esetén ugyanakkora (ezt hívják a statisztikai szakirodalomban homoszkedaszticitásnak). Ez a feltevés az, ami Poisson-eloszlású függő változóknál – az átlag és a variancia azonossága miatt – rendszerint nem teljesül.

A hagyományos lineáris regresszió tehát – mint látjuk – több szempontból sem igazán alkalmas az olyan típusú függő változók vizsgálatára, amelyek ritka előfordulású események gyakoriságát fejezik ki. Az egyik probléma az eloszlás ferdesége, a másik pedig a homoszkedaszticitás hiánya. Hogyan tudnánk orvosolni ezeket a problémákat? Az egyik lehetséges megoldás a *függő változó transzformálása*. Ennek a műveletnek a célja, hogy az új, átalakított függő változó eloszlásának sajátosságai jobban megfeleljenek a lineáris regresszió követelményeinek, s így azután a transzformált adatokra a hagyományos legkisebb négyzetek módszerét használhassuk. Poisson-eloszlású változók esetén a leggyakrabban alkalmazott átalakítás a négyzetgyök-transzformáció. Ez az eljárás két szempontból is hasznos. Egyrészt ismeretes, hogy egy Poisson-változó négyzetgyökének a varianciája független az átlagától (Chatterjee – Price 1977. 39), következésképpen ezzel az átalakítással elkerülhetjük a homoszkedaszticitás követelményének a megsértését. Másrészt ily módon a függő változó eloszlása közelebb kerül a normális eloszláshoz, ami a szignifikanciateszt alkalmazásának egyik feltétele.

Bár a függő változó átalakítása egyszerű és széles körben használt módszer – egy korábbi munkámban (Moksony 1995) magam is ezt alkalmaztam –, az újabb kutatásokban egyre inkább teret hódít egy másik megoldás. Ennek lényege, hogy ne a függő változó eloszlását igazítsuk – a transzformáció révén – a lineáris regresszió követelményeihez, hanem fordítva: *a regressziós modellt idomítsuk a függő változó eloszlásának a sajátosságaihoz*. A hagyományos regressziónak az a fajta módosítása, amelynek nyomán ez az elemzési módszer alkalmassá válik az előfordulási gyakoriságot kifejező függő változók vizsgálatára, a Poisson-regresszió.

Az általánosított lineáris modell

A Poisson-regresszió az *általánosított lineáris modellek* (Hoffman 2004; Agresti 1996. 4. fejezet) családjába tartozik. Ezek a modellek két ponton lazítják a hagyományos lineáris regresszió esetében alkalmazott megkötéseket, nagymértékben kiszélesítve ezzel a regresszióelemzéssel vizsgálható jelenségek körét. Először is, míg a hagyományos regresszió magát a függő változó átlagát írja le a magyarázó változók lineáris függvényeként, addig az általánosított lineáris modell esetében az átlag valamilyen *függvénye vagy transzformáltja* (pl. logaritmusa) tölti be ezt a szerepet. Az általánosított lineáris modell is megőrzi tehát a linearitás feltevését, csak éppen azt az átlag helyett annak átalakított formájára vonatkoztatja. Magát az átalakítást, vagyis az átlagból annak transzformáltját létrehozó műveletet vagy hozzárendelést *link function*-nak, azaz kapcsolati függvénynek nevezik. (Az átlag és a transzformáltja közötti kapcsolat természetesen lehet az egyszerű azonosság is; ez esetben maga az átlag lesz a magyarázó változók lineáris függvénye, vagyis ekkor a hagyományos lineáris regresszióhoz jutunk. A hagyományos lineáris regresszió tehát nem más, mint az általánosított lineáris modell speciális esete.)

Másodszor, míg a hagyományos lineáris regresszió feltevése szerint a függő változó normális eloszlású, addig az általánosított lineáris modellben *a függő változó eloszlása ettől eltérő típusú is lehet*.[3] Ezek közül leggyakrabban a binomiális és a Poisson-eloszlást használják a kutatók; az előbbi alkalmazása a dichotóm, az utóbbié pedig az előfordulási gyakoriságot kifejező függő változókra terjeszti ki a regresszióelemzés hatókörét.

Az *általánosított* lineáris modellt meg kell különböztetni az *általános* lineáris modelltől (Fennessey 1968; Cohen 1968), amely a szűkebb értelemben vett lineáris regresszió mellett a varianciaelemzést és a kovarianciaelemzést foglalja magában. E három statisztikai módszer lényegében csupán a magyarázó változók fajtája vagy mérési szintje tekintetében különbözik egymástól: a szorosan vett lineáris regresszió esetében mindegyik magyarázó változó numerikus, a varianciaelemzés esetében mindegyik kategoriális, míg a kovarianciaelemzés esetében egy részük numerikus, más részük viszont kategoriális. Picit leegyszerűsítve azt mondhatjuk, hogy míg az általános lineáris modell a *magyarázó* változók típusa szempontjából bővíti a hagyományos lineáris regressziót – a numerikus változók mellett a kategoriálisakat is „beengedve" az elemzésbe –, addig az általánosított lineáris modell még egy lépéssel tovább megy, és a *függő* változóval – annak típusával, eloszlásával – kapcsolatos korlátokat is feloldja.

Az eddig elmondottakat foglalja össze az 1. táblázat, amely az általánosított lineáris modell körébe tartozó modellfajtákat tartalmazza. Ebből a táblázatból jól látható, hogy a Poisson-regresszió alapvetően két dologban tér el a hagyományos lineáris regressziótól. Egyrészt abban, hogy ezúttal nem a függő változó átlagát, hanem annak természetes alapú *logaritmusát* írjuk le a magyarázó változók lineáris függvényeként; másrészt pedig abban, hogy a függő változót nem normális eloszlásúnak, hanem – a ritka előfordulású események sajátosságait figyelembe véve – *Poisson-eloszlásúnak* tekintjük.

1. Az általánosított lineáris modell körébe tartozó modelltípusok
Generalized linear models

Modelltípus	Transzformáció (kapcsolati függvény)	A függő változó eloszlása	A magyarázó változó fajtája
Szűk értelemben vett lineáris regresszió	Azonosság	Normális	Numerikus
Varianciaelemzés	Azonosság	Normális	Kategoriális
Kovarianciaelemzés	Azonosság	Normális	Numerikus és kategoriális
Logisztikus regresszió	Logit	Binomiális	Numerikus és kategoriális
Poisson-regresszió	Logaritmus	Poisson	Numerikus és kategoriális

Forrás: Agresti 1996. 97 (4.5. táblázat).

[3] Ezek az eloszlással kapcsolatos feltevések a függő változónak nem az ún. marginális, hanem a feltételes eloszlására vonatkoznak, ahol is a feltételt a magyarázó változó adott értéke jelenti.

A Poisson-regresszió modellje és az együtthatók értelmezése

A Poisson-regresszió modellje a következő:

$$\ln \lambda = \beta_0 + \beta_1 X \qquad (1),$$

ahol λ a függő változó átlaga vagy várható értéke, X a magyarázó változó, β_0 és β_1 pedig regressziós együtthatók. Az *(1)* egyenlet mindkét oldalának az antilogaritmusát véve a Poisson-regresszió multiplikatív formájához jutunk:

$$\lambda = \exp(\beta_0 + \beta_1 X) = \exp(\beta_0) * \exp(\beta_1 X) \qquad (2).$$

Ez az átalakítás azért hasznos, mert megkönnyíti a regressziós együtthatók – mindenekelőtt a magyarázó változó hatását kifejező β_1 – értelmezését. A Poisson-regresszió eredeti, log-lineáris formájában – az *(1)* egyenletben – a β_1 együttható jelentése lényegében azonos a hagyományos lineáris regresszióban szereplő β_1 együttható jelentésével. Eszerint β_1 azt mutatja meg, hogy mennyivel változik – nő vagy csökken – a függő változó átlagos értéke, amint a magyarázó változó értéke 1 egységgel emelkedik. A Poisson-regresszió esetében a függő változó az előfordulási gyakoriság logaritmusa, ennek megfelelően a β_1 együttható azt fejezi ki, mennyivel nő vagy csökken a vizsgált jelenség átlagos előfordulási gyakoriságának – például az öngyilkosságok átlagos számának – a logaritmusa, amint a magyarázó változó értéke 1 egységgel emelkedik.

Ez az értelmezés azonban nem igazán „felhasználóbarát": a kutató általában nem logaritmusokban gondolkodik. A Poisson-regresszió multiplikatív formájának – a (2) egyenletnek – az előnye éppen az, hogy ott a függő változó maga az átlagos gyakoriság, nem pedig annak logaritmusa. Ebből az egyenletből látható, hogy $\exp(\beta_1)$, vagyis a β_1 együttható antilogaritmusa a magyarázó változó multiplikatív hatását tükrözi: azt fejezi ki, hányszorosára nő vagy csökken a vizsgált jelenség átlagos gyakorisága – például az öngyilkosságok átlagos száma –, amint a magyarázó változó 1 egységgel emelkedik. Hogy ez világos legyen, tegyük föl, hogy X értéke éppen a. Ekkor

$$(\lambda \mid X = a) = \exp(\beta_0 + \beta_1 a) = \exp(\beta_0) * \exp(\beta_1 a),$$

ahol a függőleges vonal ($\mid$) a feltételt jelöli, vagyis $(\lambda \mid X = a)$ a vizsgált jelenség átlagos gyakorisága akkor, ha $X = a$. Növeljük most meg a magyarázó változó értékét 1 egységgel, azaz legyen ezúttal $X = a + 1$. Ekkor

$$(\lambda \mid X = a+1) = \exp\left[\beta_0 + \beta_1 (a+1)\right] = \exp(\beta_0) * \exp(\beta_1 a) * \exp(\beta_1).$$

Végül vegyük a fenti két egyenlet hányadosát:

$$\frac{(\lambda \mid X = a+1)}{(\lambda \mid X = a)} = \frac{\exp(\beta_0) * (\beta_1 a) * \exp(\beta_1)}{\exp(\beta_0) * (\beta_1 a)} = \exp(\beta_1).$$

Mindebből látható, hogy amint a magyarázó változó értéke a-ról $(a+1)$-re, azaz pontosan 1 egységgel emelkedett, a függő változó átlaga (λ) valóban éppen $\exp(\beta_1)$-szeresére változott.

Ezt a változást célszerű lehet százalékos formában kifejezni:

$$\text{százalékos változás} = 100 * [\exp(\beta_1) - 1].$$

Ha például a függő változó az öngyilkosságoknak egy adott településen megfigyelt száma, a magyarázó változó a munkanélküliek százalékos aránya ugyanezen a településen, a β_1 együttható értéke pedig, mondjuk, 0,055, akkor $\exp(\beta_1) = \exp(0{,}055) = 1{,}057,$ és $100 * (\exp(\beta_1) - 1 = 100 * (1{,}057 - 1) = 5{,}7$, vagyis a munkanélküliség minden 1 százalékpontos emelkedése átlagosan 5,7%-kal növeli az öngyilkosságok számát. Ugyanígy, ha a magyarázó változó a településtípus, a falvakat 0-val, a városokat 1-gyel kódoljuk, a β_1 együttható értéke pedig, mondjuk, –0,35, akkor $\exp(\beta_1) = \exp(-0{,}35) = 0{,}70,$ és $100 * (\exp(\beta_1) - 1) = 100 * (0{,}70 - 1) = -30$, vagyis a falvakhoz viszonyítva a városokban átlagosan 30%-kal alacsonyabb az öngyilkosságok száma.[4]

A rizikónépesség nagyságának bevonása a modellbe

A Poisson-regresszió idáig tárgyalt modelljének egyik fogyatékossága, hogy nem veszi figyelembe a *rizikónépesség eltérő nagyságát*, s ennek az eltérésnek a függő változóra gyakorolt hatását. Márpedig nyilvánvaló, hogy – mondjuk – egy népesebb településen pusztán a nagyobb lélekszám miatt várhatóan nagyobb a vizsgált jelenség – például az öngyilkosság – előfordulási gyakorisága, mint ott, ahol csak kevesen élnek. A népesség számában mutatkozó különbséget egyfajta standardizálással építhetjük be a regressziós modellbe, mégpedig úgy, hogy a függő változó átlagát elosztjuk a rizikónépességgel:

$$\ln\left(\frac{\lambda}{n}\right) = \beta_0 + \beta_1 X \qquad (3),$$

ahol *n* a rizikónépesség, például egy város vagy egy falu lakóinak a száma. A Poisson-regressziónak ebben a módosított modelljében tehát a vizsgált jelenség abszolút gyakorisága helyett az előfordulási *arányát* tekintjük a magyarázó változó függvényének. A *(3)* egyenlet az alábbi formában is felírható:

$$\ln(\lambda) - \ln(n) = \beta_0 + \beta_1 X \qquad (4).$$

A *(4)* egyenletet átalakítva pedig a következőt kapjuk:

$$\ln(\lambda) = \ln(n) + \beta_0 + \beta_1 X \qquad (5).$$

[4] Fontos megjegyezni, hogy kétértékű magyarázó változó – a példában a településtípus (város és falu) – esetén az együttható antilogaritmusa csak akkor tükrözi a magyarázó változó multiplikatív hatását, ha *dummy* kódolást alkalmazunk, vagyis ha az egyik kategóriát 1-gyel, a másikat pedig 0-val jelöljük. Csak ebben az esetben egyenlő ugyanis az 1 egységnyi változás a két kategória közötti távolsággal. Hatáskódolás esetén például – azaz amikor az egyik kategóriát 1-gyel, a másikat pedig –1-gyel jelöljük – a kategóriák közti távolság nem 1, hanem 2 egység, ennek megfelelően a multiplikatív hatás nagysága nem $\exp(\beta_1)$, hanem $\exp(2 * \beta_1) = \exp(\beta_1)^2$.

A Poisson-regresszionak ez a modellje – mint látható – abban tér el az eredetitől, hogy szerepel benne egy külön magyarázó változó, ln(*n*), ami a rizikónépesség nagyságát tükrözi, s aminek az együtthatóját automatikusan 1-nek vesszük. Ezt a magyarázó változót általában kiegyenlítő vagy kiegyensúlyozó (az angol nyelvű szakirodalomban *offset*) változónak nevezik. Amennyiben a rizikónépesség nagysága minden megfigyelés esetében azonos, akkor ln*(n)* állandó, és így beépíthető a β_0 konstansba, aminek eredményeként az eredeti modellhez, vagyis az (*1*) egyenlethez jutunk.

A „túlszórás"

A rizikónépesség nagyságának figyelembevétele mellett a Poisson-regresszió kiinduló modellje egy másik szempontból is sokszor kiegészítésre szorul. Az előfordulási gyakoriságot kifejező változók sajátosságairól szólva említettem már, hogy a Poisson-eloszlás esetében az átlag és a variancia egyenlő egymással. A Poisson-regresszió alapértelmezésben erre az azonosságra épül: a számítások során azt feltételezzük, hogy a függő változó átlaga és varianciája valóban egyforma. A gyakorlatban azonban ez a feltevés számos esetben tévesnek bizonyul: *a variancia meghaladja az átlagot*. Ezt a jelenséget hívják a módszertani szakirodalomban „overdispersion"-nek, azaz „túlszórás"-nak (Agresti 1996. 92–3; Le 1998. 226–228).

A „túlszórás" hátterében rendszerint két ok valamelyike áll (King 1989. 766–769). Egyrészt szinte minden empirikus kutatásban előfordul, hogy egy vagy több, a modellben szereplő független változókkal nem korreláló *magyarázó változó kimarad* az elemzésből – vagy azért, mert nem is gondolunk rájuk, vagy pedig azért, mert bár eszünkbe jutnak, nem tudunk róluk adatokat gyűjteni. Tegyük fel például, hogy az öngyilkosságok településenkénti gyakoriságát a gazdasági fejlettség és az adott régió sajátos kultúrája egyaránt befolyásolja, mi azonban e két tényező közül csupán az elsőt vonjuk be a vizsgálatba. Mi történik ebben a helyzetben? Ha rögzítjük a gazdasági fejlettség szintjét, akkor az ehhez a szinthez tartozó települések földrajzilag – és kulturálisan – *heterogének* lesznek; olyan részsokaságok keverékei, amely részsokaságok mindegyikének megvan a maga saját, az adott régióra jellemző – és a többiétől eltérő – öngyilkossági gyakorisága. Vagyis a függő változónak a gazdasági fejlettség adott értékéhez tartozó átlaga nem lesz konstans – ahogyan azt a Poisson-eloszlás feltételezi -, hanem annyi különböző átlagunk lesz, ahány különböző régiónk van. Ennek a fajta heterogenitásnak – a jelenség régióról régióra változó gyakoriságának – a következtében az öngyilkosság feltételes – azaz a gazdasági fejlettség meghatározott szintjéhez tartozó – eloszlásának a szórása nagyobb lesz annál, mint amire a Poisson-eloszlás alapján számítani lehetne. A „túlszórás" forrása itt tehát végső soron *a modellből kihagyott magyarázó változók által előidézett „többletszóródás"* – hasonlóan ahhoz, ahogyan a hagyományos lineáris regresszió esetében is a kihagyott változók hatása a függő változó nagyobb reziduális szórásában csapódik le, s az újabb magyarázó változók bevonásának egyik indoka épp ennek a reziduális szórásnak a csökkentése (Moksony 1999. 185).[5]

[5] A hagyományos lineáris regressziónál azonban mindez nem okoz problémát, hiszen az normális eloszlásra alapít, amely esetében – a Poisson-eloszlásétól eltérően – az átlag és a szórás független egymástól.

A másik ok, ami szintén „túlszóráshoz" vezethet, a *megfigyelések közötti függőség*. A Poisson-eloszlás egyik alapfeltevése, hogy a megfigyelések függetlenek egymástól: egy esemény – pl. öngyilkosság – bekövetkezése nem befolyásolja egy másik esemény – egy másik öngyilkosság – bekövetkezésének a valószínűségét. Mi történik, ha ez a feltétel nem teljesül; ha pl. egy esemény előfordulása *növeli* egy másik előfordulásának az esélyét? Ebben az esetben a Poisson-eloszlás alapján vártnál *nagyobb lesz a magas és az alacsony – vagyis a szélső – értékek előfordulási gyakorisága*, ennek eredményeként pedig *emelkedik* a változó *szórása*. Ha – mondjuk – egy iskola diákjai egymás viselkedését utánozzák, akkor az öngyilkossági kísérletek véletlenszerű eloszlása helyett várhatóan azt tapasztaljuk, hogy míg hónapokig egyetlen eset sem történik, addig – a mintakövetés hatására – néhány nap vagy hét leforgása alatt többen is megpróbálnak véget vetni életüknek. A deviáns viselkedés szakirodalmában jól ismert az önpusztításnak ez a fajta utánzás révén történő terjedése, s ennek nyomán a jelenség idő- vagy térbeli „sűrűsödése", ún. „öngyilkossági klaszterek" kialakulása (pl. Phillips 1974; Phillips – Carstensen 1986; Gould et al. 1990). A mi szempontunkból mindebből a lényeg most az, hogy a mintakövetés, a megfigyelések közötti függőség ugyanúgy növeli a függő változó szórását, mint a magyarázó változók kihagyása, s ugyanúgy a Poisson-eloszlás egyik alapvető feltevésének – az átlag és a variancia azonosságának – a megsértéséhez vezet.

Hogyan befolyásolja a „túlszórás" a Poisson-regresszióelemzés eredményét? A legsúlyosabb következmény, hogy bár maguk a regressziós együtthatók torzítatlanok maradnak, az együtthatók becsült *standard hibái a ténylegesnél kisebbek* lesznek, s ennek az alulbecslésnek a következtében a szignifikanciatesztek is hamis – mégpedig a valóságosnál kedvezőbb – képet mutatnak. „Túlszórás" esetén tehát az indokoltnál könnyebben kapunk statisztikailag szignifikáns eredményt. Mindennek az oka, hogy a standard hibák meghatározása során az átlagot és a varianciát azonosnak vesszük – miközben ez utóbbi igazából nagyobb az előbbinél. Vagyis a Poisson-eloszlás „túlszórás" esetén tévesnek bizonyuló feltevését alkalmazva a standard hibákat lényegében kényszerítjük arra, hogy alacsonyabbak legyenek a ténylegesnél.

Miként orvosolhatjuk ezt a problémát? Az egyik lehetőség *a standard hibák utólagos kiigazítása*, felfelé korrigálása. Ennek lényege, hogy a standard hibákat a „túlszórás" mértékét jelző ún. diszperziós paraméter négyzetgyökével szorozzuk. Magát ezt a paramétert a hagyományos khí-négyzet-mutató és a hozzá tartozó szabadságfok hányadosával becsülhetjük (Tiao 1994. 72). Ennek alapja, hogy „túlszórás" hiányában – feltéve, hogy maga a regressziós modell helyesen specifikált – a khí-négyzet várható értéke egyenlő a szabadságfokkal, következésképpen a kettő hányadosa 1 (Agresti 1996. 93). Az 1-nél nagyobb értékek „túlszórásra" utalnak, az 1-nél kisebbek pedig az ennél – legalábbis a társadalomtudományban – ritkábban előforduló „alulszórást" jelzik, vagyis azt a helyzetet, amikor a függő változó varianciája kisebb, mint az átlag.[6]

[6] A standard hibáknak ezzel az utólagos kiigazításával kapcsolatban egyetlen kérdés merül fel csupán, mégpedig az, hogy melyik khí-négyzetet használjuk: a jól ismert Pearson-félét vagy inkább az ún. Likelihood-ratio khí-négyzetet. Bár e két mutató értéke rendszerint meglehetősen közel áll egymáshoz, következésképpen a döntésnek általában nincs túlságosan nagy jelentősége, elméleti szempontból a Pearson-féle khí-négyzet használata az indokoltabb (Allison 2001. 223).

A másik lehetséges megoldás a Poisson-modell módosítása, mégpedig úgy, hogy a függő változónak a magyarázó változó adott értékéhez – illetve több magyarázó változó esetén értékkombinációjához – *tartozó átlagát nem konstansnak tekintjük, hanem változónak*, amelynek értékei véletlenszerűen szóródnak egy középponti érték körül. Technikailag ez azt jelenti, hogy az eredeti

$$\lambda = \exp(\beta_0 + \beta_1 X) \qquad (2)$$

modellről áttérünk a

$$\tilde{\lambda} = \exp(\beta_0 + \beta_1 X + e) \qquad (6)$$

modellre, ahol $\tilde{\lambda}$ az a véletlen változó, amely a függő változó korábban konstansként kezelt átlaga (λ) helyére lép, *e* pedig egy véletlen hibatényező, amely – egyebek között – az elemzésből kihagyott magyarázó változók hatását – az általuk előidézett heterogenitást vagy „többletszórást" – képviseli. A két modell – a (2) és a (6) egyenlet – közötti alapvető különbség, hogy míg az elsőben a magyarázó változó (*X*) rögzített értékéhez a függő változónak *egyetlen egy* átlaga tartozik (λ), addig a másodikban az átlagoknak $(\tilde{\lambda})$ egy egész *eloszlása*; annak megfelelően, hogy a hibatényezőben (*e)* összefoglalt kihagyott változók most rendre véletlenszerűen „eltérítik" az átlagot attól a szinttől, ami *X* adott értékéből – a modell szisztematikus részéből – következne (Long 1997. 230–231; lásd még Gardner et al. 1995. 399–400).

A Poisson-regresszionak ez a módosítása tehát a Poisson-eloszlás átlagának *konstansból változóvá* tételében lép túl az eredeti modellen. Az eredeti modellben a függő változó konkrét értéke (*Y*) véletlen változó volt, amely Poisson-eloszlást követve szóródott az átlag (λ) körül. Maga ez az átlag azonban nem változó volt, hanem konstans – a magyarázó változó értéke egyértelműen meghatározta az átlag értékét is. A módosított modellben a függő változó konkrét értéke továbbra is véletlen változó, és továbbra is Poisson-eloszlást követve szóródik az átlag körül. Most azonban már ez az átlag is véletlen változó, amely az *e* hibatényező hatására ingadozik egy középponti érték körül. Mivel a hiba várható értéke – feltételezés szerint – nulla, ez a középponti érték azonos az eredeti modell konstansként kezelt átlagával, azaz

$$E(\tilde{\lambda}) = \lambda .$$

De milyen az eloszlása ennek a változóvá tett átlagnak? A leggyakoribb – és matematikai szempontból a legcélszerűbb – feltételezés az, hogy $\tilde{\lambda}$ Gamma-eloszlású. A Poisson-regresszió módosított modelljében tehát végső soron két véletlen változó két különböző eloszlásával, a két eloszlás kombinációjával van dolgunk. Van egyfelől a függő változó konkrét értéke (*Y*), amely Poisson-eloszlást követve szóródik az átlag $(\tilde{\lambda})$ körül, másfelől az átlag, amely szintén véletlen változó és Gamma-eloszlású (Agresti 2002. 559–560.; Land – McCall 1996). E kétféle eloszlás kombinációjának eredménye az ún. negatív binomiális eloszlás – ezért nevezik a Poisson-regresszió itt bemutatott módosítását *negatív binomiális regressziónak*.

A negatív binomiális regresszió előnye, hogy – a hagyományos Poisson-regresszióval ellentétben – *nem követeli meg az átlag és a variancia azonosságát*, ha-

nem lehetővé teszi, hogy az utóbbi meghaladja az előbbit. Míg ui. a Poisson-regresszió esetében

$$\mathrm{E}(Y) = \mathrm{var}(Y) = \lambda,$$

addig a negatív binomiális regresszió esetében

$$\mathrm{E}(Y) = \lambda \quad \text{és} \quad \mathrm{var}(Y) = \lambda(1+\alpha\lambda) = \lambda + \alpha\lambda^2,$$

ahol α a „túlszórás" mértékét jelző ún. diszperziós paraméter (dispersion parameter). Látható, hogy amennyiben nincsen „túlszórás", akkor α = 0, és $\mathrm{var}(Y) = \lambda + \alpha\lambda^2 = \lambda,$ vagyis ekkor a variancia egyenlő az átlaggal – miként a Poisson-regresszió esetében –, azaz a negatív binomiális regresszió ekkor a hagyományos Poisson-regresszióra egyszerűsödik.

Egy példa

Megkönnyítheti az eddig elmondottak megértését, ha befejezésként egy konkrét példával is szemléltetjük a Poisson-regresszió alkalmazását.

Egy nemrég befejezett kutatásomban a társadalmi-gazdasági fejlődésnek, illetve – ami az érem másik oldala – az elmaradottságnak az öngyilkosságra gyakorolt hatását vizsgáltam településszintű adatok segítségével. Az elemzés az 1990 és 1995 közötti időszakot ölelte fel, és azokra a magyarországi településekre terjedt ki, amelyek 1990-ben községnek minősültek, és ezt a jogállásukat a vizsgált időszakban végig megőrizték. Ezeknek a településeknek a száma 2869 volt. Ebben a csaknem 3000 faluban 1990 és 1995 között összesen 9237 öngyilkosságot követtek el.

A vizsgálat két fő lépésből állt. Először főkomponens-elemzés segítségével létrehoztam az elmaradottság vagy – ahogyan gyakran nevezik – a depriváció összefoglaló mutatóját,[7] majd ezt követte a Poisson-regresszióelemzés, amelyben az öngyilkosságok száma volt a függő változó, a főkomponens-elemzés eredményeként kapott főkomponenspontszám pedig a magyarázó változó. A regresszióelemzés során a következő modellel dolgoztam:

$$\ln(\lambda_i) = \ln(n_i) + \beta_0 + \beta_1 X_i,$$

[7] A főkomponens-elemzésbe az 1990 és 1995 közötti évekre vonatkozó TSTAR adatbázisokban szereplő területi ismérvek közül azokat vontam be, amelyek a települések fejlettségét, illetve elmaradottságát, depriváltságát, valamint az azzal együtt járó társadalmi-demográfiai folyamatokat leginkább tükrözték. Ezek az ismérvek a következők voltak: 1000 lakosra jutó személygépkocsik száma, 1992–1995; 1 háztartásra jutó villamosenergia-fogyasztás, 1993–1995 (1000 kwh); közlekedési, oktatási és egészségügyi infrastruktúra 1993-ban (intézmények száma); munkanélküliek a 18–59 évesek százalékában, 1993–1994; szociális jövedelempótló támogatásban részesülők a lakónépesség százalékában, 1993–1995; rendszeres szociális segélyben részesülők a lakónépesség százalékában, 1993–1995; eltartottsági hányados (a 18 év alatti és 59 év feletti népességnek a 18–59 éves népességhez viszonyított aránya); vándorlási egyenleg az 1990–1995 közötti átlagos év végi lakónépesség százalékában.

ahol λ az öngyilkosságok 1990 és 1995 közötti átlagos száma az *i*-edik községben, n_i a népesség száma ugyanebben az időszakban ugyanezen a településen, X_i ennek a településnek a főkomponenspontszáma, β_0 és β_1 pedig regressziós együtthatók. Ebben a modellben – amint azt korábban már említettem – a népességszám (*n*) ún. kiegyenlítő (offset) változó, amelynek feladata a rizikónépesség eltérő nagyságából adódó különbségek kiszűrése, s amelynek együtthatóját az elemzés során automatikusan 1-nek veszszük. A modell együtthatóinak meghatározására a STATA programcsomagot használtam.

A Poisson-regresszió eredményeit a 2. táblázat tartalmazza. A táblázatból megállapítható, hogy a főkomponenspontszám emelkedésével – vagyis a depriváció mértékének a növekedésével – párhuzamosan az öngyilkosságok átlagos száma is emelkedik, s ez a hatás statisztikailag is erősen szignifikáns. A regressziós együtthatóból (0,1148) látható, hogy a főkomponenspontszám 1 egységnyi növekedése nyomán az öngyilkosságok számának a logaritmusa átlagosan mintegy 0,1-del emelkedik. Célszerűbb azonban az együttható helyett annak antilogaritmusára fordítani a fő figyelmet, ez ui. – amint azt korábban már említettük – a magyarázó változónak magára a függő változó átlagára (nem pedig az átlag logaritmusára) gyakorolt hatását fejezi ki, megkönnyítve ezzel az eredmények tartalmi értelmezését. A regressziós együttható antilogaritmusa esetünkben exp(0,1148) = 1,122, ami azt jelenti, hogy a főkomponenspontszám 1 egységnyi emelkedése átlagosan 12,2%-kal növeli az öngyilkosságok számát. Ezek az eredmények arra mutatnak, hogy *a település szintű depriváció fokozza az öngyilkosság veszélyét***.**

2. A depriváció hatása az öngyilkosságra. A Poisson-regresszióelemzés eredményei
The impact of deprivation on suicide. Poisson regression results

Változó	Együttható	Standard hiba	z-érték	Együttható antilogaritmusa
Főkomponenspontszám	0,1148*	0,0114	10,11	1,122
Konstans	–7,7462	0,0110	–706,17	

n = 2869 ; * p < 0,001

A „túlszórás" vizsgálata és hatásának kiküszöbölése

Amint azt korábban már említettem, „túlszórás" esetén – vagyis amikor a függő változó varianciája meghaladja annak átlagát – a Poisson-regresszió alulbecsli a standard hibákat, s ennek nyomán a szignifikanciateszt is torzított – a valóságosnál kedvezőbb – eredményt ad. Éppen ezért fontos, hogy mielőtt túlságosan messzire mennénk az adatainkból levont következtetések megfogalmazásával, megállapítsuk az esetleges „túlszórás" mértékét, s amennyiben szükséges, kiküszöböljük annak torzító hatását.

Első lépésként az öngyilkosságok ténylegesen megfigyelt eloszlását összehasonlítottam a Poisson-modellből számított eloszlással. Ez utóbbi azt mutatja meg, mekkorák lennének az öngyilkosság egyes előfordulási gyakoriságaihoz tartozó valószínűségek akkor, ha az eloszlás átlaga – a valóságos helyzetnek megfelelően – 3,22 lenne, és a Poisson-eloszlás feltételei maradéktalanul teljesülnének. Az eredményt a 3. táblázat

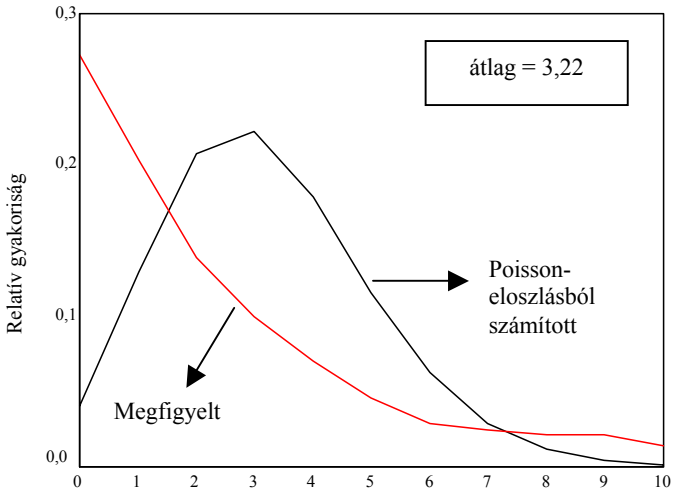
tartalmazza. Ebből a táblázatból megállapítható, hogy a legalacsonyabb és a legmagasabb – vagyis a szélső – értékek esetében a megfigyelt relatív gyakoriságok meghaladják a Poisson-eloszlás alapján várt valószínűségeket, a középső értéktartományban viszont alatta maradnak azoknak. Ezzel összhangban a tényleges eloszlás varianciája (22,72) – a Poisson-eloszlás alapfeltevésével ellentétben – lényegesen nagyobb, mint az átlaga (3,22).

3. A községi öngyilkosságok ténylegesen megfigyelt és a Poisson-eloszlásból számított eloszlása
Observed and expected (based on Poisson distribution) number of suicides

Öngyilkosságok száma	Megfigyelt relatív gyakoriság	Poisson-eloszlásból számított valószínűség
0	0,273	0,040
1	0,203	0,129
2	0,138	0,207
3	0,099	0,222
4	0,070	0,179
5	0,045	0,115
6	0,029	0,062
7	0,024	0,028
8	0,021	0,011
9	0,021	0,004
10	0,014	0,001

A 3. táblázatban látott tendencia még jobban kivehető a II. ábrán, amely vonaldiagram formájában jeleníti meg a kétféle – megfigyelt és feltételezett – eloszlást. A rajzból világosan kitűnik, hogy a vízszintes tengely jobb és bal szélén – tehát a legmagasabb és a legalacsonyabb előfordulási gyakoriságoknál– a ténylegesen megfigyelt eloszlást tükröző görbe a Poisson-eloszlásnak megfelelő másik görbe fölött, középen – a mérsékelten gyakori előfordulásoknál– viszont az alatt halad. A 3. táblázat és a II. ábra összességében tehát arra mutat, hogy valóban számolni kell a „túlszórás" veszélyével.

Ezt erősíti meg a khí-négyzet-statisztika és a szabadságfok hányadosa. Amint azt korábban már említettem, „túlszórás" hiánya esetén e hányados értéke 1, a „túlszórás"-ra pedig az 1-nél nagyobb értékek utalnak. Esetünkben a Pearson-féle khí-négyzet és a szabadságfok hányadosa 1,42, a Likelihood-ratio khí-négyzet és a szabadságfok hányadosa pedig 1,52. Mindkét mutató megkérdőjelezi a Poisson-regresszió egyik alapfeltevését, az átlag és a variancia azonosságát, indokolt tehát az idáig tárgyalt eredmények felülvizsgálata.



II. A községi öngyilkosságok ténylegesen megfigyelt és a Poisson-eloszlásból számított eloszlása
Observed and expected (based on Poisson distribution) number of suicides

A tanulmány egy korábbi részében a „túlszórás" problémájának kétféle megoldásáról beszéltünk. Az egyik volt a standard hibák utólagos kiigazítása, felfelé korrigálása, a másik pedig a Poisson-regresszió helyett a negatív binomiális regresszió alkalmazása. Arról is szó volt akkor, hogy a standard hibák kiigazításához a khí-négyzet mutató és a szabadságfok hányadosát használjuk: a standard hibákat e hányados négyzetgyökével szorozzuk. A korrekció eredményét a 4. táblázat tartalmazza. Mint látható, a kiigazítás nyomán a standard hibák valamelyest emelkedtek, ennek nyomán pedig – mivel az együtthatók értéke nem változott – a z-értékek kissé csökkentek. A főkomponenspontszám hatása azonban még így is – a standard hibák felfelé korrigálása után is – statisztikailag messze szignifikáns maradt.

4. Poisson-regresszió: a standard hibák kiigazítása
Poisson regression: corrected standard errors

Változó	Együttható	Eredeti standard hiba	Pearson-féle khi-négyzet alapján korrigált standard hiba	Likelihood-ratio khi-négyzet alapján korrigált standard hiba
Főkomponens-pontszám	0,1148	0,0114	0,0140	0,0135
		(10,11)	(8,50)	(8,20)
Konstans	–7,7462	0,0110	0,013	0,014
		(–706,17)	(–593,65)	(–573,22)

Megjegyzés: A standard hiba alatt zárójelben lévő szám a z-érték, azaz a regressziós együttható és a standard hiba hányadosa.

A standard hibák utólagos módosítása mellett megvizsgáltam a másik megoldást, a *negatív binomiális regressziót* is. Ennek eredményei az 5. táblázatban szerepelnek. A főkomponenspontszám együtthatója nagyon hasonló a Poisson-regresszióból kapott együtthatóhoz, és azt mutatja, hogy a depriváció mértékének az emelkedése növeli az öngyilkosság gyakoriságát. Amint az a regressziós együttható antilogaritmusából megállapítható, a főkomponenspontszám 1 egységnyi emelkedése mintegy 11%-kal növeli az öngyilkosságok átlagos számát. Az együttható standard hibája ugyanakkor valamelyest nagyobb nem csupán az eredeti, korrigálatlan standard hibánál, hanem a korrigált értéknél is. Ezzel együtt azonban a főkompoenspontszám hatása statisztikailag továbbra is erősen szignifikáns: a z-érték, vagyis az együttható és a standard hiba hányadosa 6,74, a hozzá tartozó szignifikanciaszint pedig jóval alatta marad az általában használt 5%-nak.

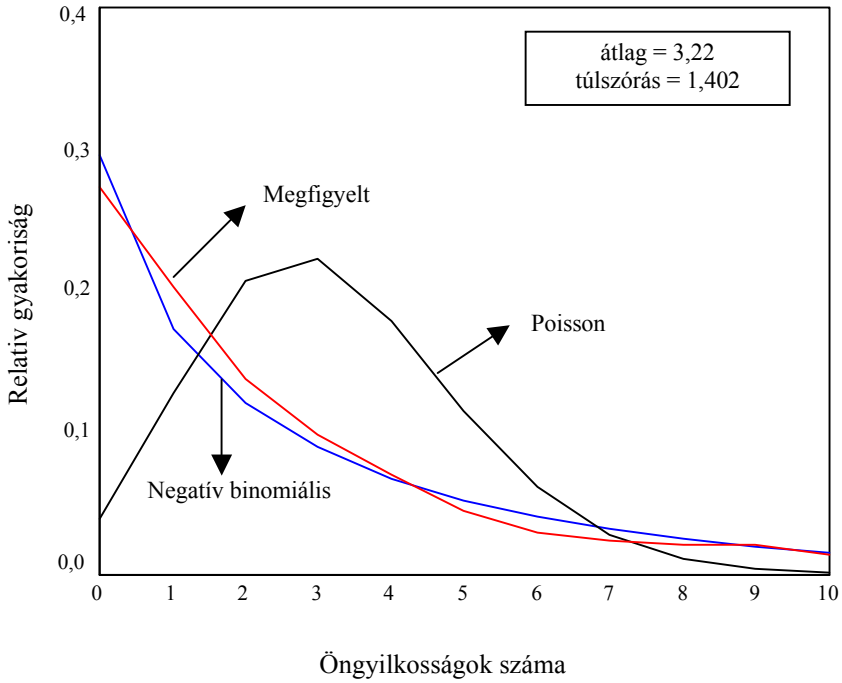
5. A depriváció hatása az öngyilkosságra. A negatív binomiális regresszióelemzés eredményei
The impact of deprivation on suicide. Results from negative binomial regression

Változó	Együttható	Standard hiba	z-érték	Együttható antiloga-ritmusa
Főkomponenspontszám	0,1034*	0,0153	6,74	1,109
Konstans	–7,7816	0,0144		
Alpha	0,1462*	0,0130		

n = 2869 ; * $p < 0{,}001$

Bár a kétféle – egyrészt a Poisson-, másrészt a negatív binomiális – regresszióelemzés meglehetősen hasonló eredményt adott, összességében mégis a negatív binomiális regresszió illeszkedik jobban a vizsgált adatokhoz. Jól látható ez a III. ábrán, amely egymás mellett mutatja az öngyilkosságok gyakoriságának ténylegesen megfigyelt, illetve a Poisson- és a negatív binomiális regresszió alapján becsült eloszlását. Míg a Poisson-eloszlás – amint azt már korábban a II. ábra kapcsán is megállapítot-

tuk – a valóságosnál kisebbnek feltételezett szórásnak megfelelően jelentősen alulbecsli a szélső, vagyis a nagyon alacsony és a nagyon magas gyakoriságokat, a nagyobb szórású negatív binomiális eloszlás ezeken a részeken is jól követi a ténylegesen megfigyelt adatokat.



III. A községi öngyilkosságok ténylegesen megfigyelt és a Poisson-, ill. a negatív binomiális regresszió alapján becsült eloszlása
Observed and expected number of suicide, based on Poisson and negative binomial regression

HIVATKOZÁSOK

Agresti, A. (1996): *An introduction to categorical data analysis*. New York etc.: Wiley.
Agresti, A. (2002): *Categorical data analysis*. 2nd edition. New York etc.: Wiley.
Allison, P.D. (2001): *Logistic regression using the SAS system. Theory and application*. Cary, NC: SAS Institute.
Chatterjee, S. – Price, B. (1977): *Regression analysis by example*. New York: Wiley.
Cohen, J. (1968): Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70: 426–443.

Fennessey, J. (1968): The general linear model: a new perspective on some familiar topics. *American Journal of Sociology*, 74: 1–27.
Gardner, W. et al. (1995): Regression analyses of counts and rates: Poisson, overdispersed Poisson and Negative Binomial models. *Psychological Bulletin*, 118: 392–404.
Gould, M.S. et al. (1990): Time-space clustering of teenage suicide. *American Journal of Epidemiology*, 131: 71–78.
Hoffman, J. P. (2004): *Generalized linear models*. Boston, etc.: Pearson Education Inc.
King, G. (1989): Variance specification in event count models: from restrictive assumptions to a generalized estimator. *American Journal of Political Science*, 33: 762–784.
Land, K.C. – McCall, P.L. (1996): A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models. *Sociological Methods & Research*, 24: 387–443.
Le, Ch. T. (1998): *Applied categorical data analysis*. New York etc.: Wiley.
Liao, T.F. (1994): *Interpreting probability models. Logit, probit, and other generalized linear models*. Thousand Oaks, etc.: Sage.
Long, J.S. (1997): *Regression models for categorical and limited dependent variables*. Thousand Oaks etc.: Sage.
Moksony Ferenc (1995): A fejlődés ára vagy az elmaradottság átka? Az öngyilkosság alakulása Magyarország községeiben. *Szociológiai Szemle*, No. 2. 73–84.
Phillips, D.P. (1974): The Influence of Suggestion on Suicide: Substantive and Theoretical Implications of the Werther Effect. *American Sociological Review*, 39: 340–354.
Phillips, D.P. – Carstensen, L.L. (1986): Clustering of teenage suicides after television news stories about suicide. *New England Journal of Medicine*, 315: 685–89.
Sváb János (1981): *Biometriai módszerek a kutatásban*. Budapest: Mezőgazdasági Kiadó.

Tárgyszavak:

Általánosított lineáris modell
Deviáns viselkedés
Öngyilkosság
Poisson-regresszió
Ritka események statisztikai elemzése

THE USE OF POISSON REGRESSION IN SOCIOLOGY AND DEMOGRAPHY

Abstract

This paper gives an introduction to Poisson regression, a statistical method that is particularly useful when the dependent variable describes the number of occurrences of some rare event such as suicide. After pointing out why ordinary linear regression is

inappropriate for dependent variables of this sort, the author goes on to present the basic Poisson regression model and shows how it fits in the broad class of generalized linear models. Then he turns to discussing a major problem of Poisson regression known as overdispersion and explains how the negative binomial regression can help solve this problem. The paper ends with a detailed empirical example, drawn from the author's own research on suicide.