Ritka események előfordulási gyakoriságának elemzése regressziós módszerekkel – Különböző statisztikai szoftverek összehasonlítása

Virág Katalin[1], Boda Krisztina[1], Nyári Tibor[1]
[1]SZTE ÁOK TTIK Orvosi Fizikai és Orvosi Informatikai Intézet
virag.katalin@med.u-szeged.hu, boda.krisztina@med.u-szeged.hu,
nyari.tibor@med.u-szeged.hu
6720 Szeged Korányi fasor 9.

Összefoglaló: **Epidemiológiai kutatások során gyakran események előfordulási gyakoriságát vizsgálják (pl. megbetegedések, halálozások száma). Ezen cikk olyan regressziós modelleket mutat be, ahol a függő változó valamilyen ritka esemény bekövetkezési gyakorisága vagy a népességhez viszonyított aránya.**

Bevezetés

Ha a függő változó valamilyen ritka esemény előfordulási gyakorisága, akkor az eloszlása általában jól közelíthető a Poisson- vagy a negatív binomiális eloszlások valamelyikével. A népességhez viszonyított arányok esetén gyakori a túlszóródás („overdispersion"), vagyis a variancia meghaladja az átlagot; sérül a hagyományos Poisson-regresszió feltétele, miszerint a várható érték és a variancia megegyezik ([1] – [7]).

Célkitűzés

Illusztrálni szeretnénk, hogy túlszóródás esetén a hagyományos Poisson-regressziós modell nem jól illeszkedik az adatokhoz.

További célunk különböző statisztikai szoftverek használatával illesztett regressziós modellek összehasonlítása.

Módszer

Általánosított lineáris modellek

A Poisson- és a negatív binomiális regressziók az általánosított lineáris modellek családjába tartoznak. Az általánosított lineáris modellek az egyszerű lineáris modellek általánosításai: a függő változó eloszlása eltérhet a normális eloszlástól (akár diszkrét eloszlást is követhet); a függő változó várható értéke helyett annak valamilyen függvényét írják le a magyarázó változók lineáris függvényeként (egy „link" függvény segítségével); valamint megengedik a variancia átlagtól való függését.

Gyakorisági adatok regressziója: Poisson-regresszió

Ha a függő változó valamilyen ritka esemény előfordulási gyakorisága (pl. adott populációban adott időtartam alatt előforduló új megbetegedések száma), akkor a Poisson-modell használható. Ha *Y* Poisson eloszlású $\mu > 0$ várható értékkel, akkor:

$$\mathrm{P}(Y = k) = \frac{\mu^k e^{-\mu}}{k!} \quad (k = 0, 1, 2, \ldots); \quad \mathrm{E}(Y) = \mathrm{Var}(Y) = \mu.$$

A lineáris modell (logaritmikus link függvény használatával):

Y: függő változó, a központi idegrendszerre ható érsérülések okozta halálozások évenkénti száma, $Y \sim \mathrm{Poisson}(\mu)$; n: év eleji népesség az adott évben; $\boldsymbol{X}$: magyarázó változók vektora; $\boldsymbol{\beta}$: regressziós együtthatók vektora.

$$\ln(\mu) = \ln(n) + \boldsymbol{X}^T\boldsymbol{\beta}$$

$$\mathrm{E}(Y) = \mu = ne^{\boldsymbol{X}^T\boldsymbol{\beta}}$$

Túlszóródás („overdispersion")

Poisson eloszlás esetén a várható érték és a variancia megegyezik, az ún. diszperziós paraméter 1-gyel egyenlő. Ha a megfigyelt gyakoriságokat a kockázatnak kitett népességhez viszonyítjuk, akkor gyakran a variancia meghaladja a várható értéket, vagyis túlszóródás figyelhető meg.

A Poisson-regresszió lehetséges általánosításai túlszóródás esetén:

- **Robusztus módszer:** a kovariancia mátrixot robusztus („szendvics") módszerrel becsüljük, miközben a várható érték becslése változatlan.
- **Kvázi-Poisson modell:** a diszperziós paramétert a modellből becsüljük. A regressziós együtthatók változatlanok, a standard hibák nagyobbak.
- **Negatív binomiális modell:** keverék Poisson-eloszlás, ahol feltételezzük, hogy a függő változó várható értéke Gamma-eloszlást követ.

Adatok

A Központi Statisztikai Hivatal adatai alapján vizsgáltuk a központi idegrendszerre ható érsérülések okozta magyarországi halálozási rátát, 1979 és 2010 között (BNO-kód: 1979-1995.: 430-438; 1996-2010.: I60-I69).

Szoftverek

A különböző regressziós modelleket az Intézetünkben elérhető SAS ([8]), SPSS ([9]), STATA ([10]), valamint az ingyenes R ([11]) statisztikai szoftvercsomagok segítségével illesztettük. A főbb paramétereket az 1. sz. táblázat tartalmazza.

1. sz. táblázat

Modell	R	SAS	STATA	SPSS
Poisson	*stats* csomag **glm( )** fggv.: family = poisson	**Proc genmod** dist = poisson	**poisson**	**genlin** distribution = poisson link = log
Robusztus	*sandwich* csomag **sandwich( )** fggv.	**repeated** parancs	**robust** paraméter	**criteria covb** = robust
Kvázi-Poisson	family = quasipoisson	**scale** paraméter	**glm** **scale** paraméter	**criteria scale**
Negatív binomiális	*MASS* csomag **glm.nb( )** fggv.	dist = nb	**nbreg**	distribution = negbin(mle) link = log

Eredmények

Ha a halálozási rátát csak a 40 év felettiek körében vizsgáljuk nemek szerint, akkor a diszperziós paraméter becslésére 22-t kapunk, vagyis a variancia jelentősen meghaladja az átlagot. A regressziós együtthatók becsléseit és a standard hibákat a 2. sz. táblázat tartalmazza. Az egyszerű Poisson-regressziós modell esetén a nemek között szignifikáns különbség állapítható meg 5%-os szinten, míg a többi modell esetén ez a különbség eltűnik. A hagyományos Poisson-regressziós és a negatív binomiális regressziós modellek likelihood-hányados próbával történő összehasonlítása során azt kapjuk, hogy a negatív binomiális regresszió szignifikánsan jobban illeszkedik az adatokhoz ($2(\ln(L_1)-\ln(L_2))=1106$; sz.f.=1; $p<0{,}001$).

Következtetések

Az ingyenesen elérhető R statisztikai szoftver segítségével illesztett modellek a többi szoftverrel megegyező eredményt adnak.

Ritka események bekövetkezési gyakoriságának, illetve a népességhez viszonyított arányának modellezése során gyakran megfigyelhető túlszóródás esetén az adatokhoz a hagyományos Poisson-regressziónál jobban illeszkedő modellek az R rendszerben csupán néhány paraméter változtatásával egyszerűen elérhetők.

2. sz. táblázat

Együttható	Poisson	Robusztus	Kvázi-Poisson	Negatív binomiális
Intercept	-5,192 (0,0026)	-5,192 (0,0117)	-5,192 (0,0122)	-5,185 (0,0135)
Év	-0,020 (0,0001) $p < 0{,}001$	-0,020 (0,0008) $p < 0{,}001$	-0,020 (0,0007) $p < 0{,}001$	-0,021 (0,0007) $p < 0{,}001$
Nem	0,019 (0,0025) $p < 0{,}001$	0,019 (0,0113) $p = 0{,}09895$	0,019 (0,0120) $p = 0{,}124$	0,019 (0,0123) $p = 0{,}126$

Kitekintés

Ha az adatok között sok nulla szerepel, akkor a „Zero-inflated”; ha egyáltalán nincs nulla, akkor a „Zero-truncated” modellek használata is felmerülhet. Ha trendvizsgálat során az adatsorban töréspontok vannak, akkor a „joinpoint” regressziós módszer alkalmazható.

R-forráskódok

Poisson-regresszió:
summary(mPois <- glm(Gyakorisag ~ Ev + relevel(Nem, 2) + offset(log(Nepesseg)), family = poisson))
deviance(mPois)/df.residual(mPois)
Robusztus módszer:
require(sandwich)
require(lmtest)
coeftest(mPois, vcov = sandwich)
Kvázi-Poisson-regresszió:
summary(mQPois <- glm(Gyakorisag ~ Ev + relevel(Nem, 2) + offset(log(Nepesseg)), family = quasipoisson))
Negatív binomiális regresszió:
require(MASS)
summary(mNB <- glm.nb(Gyakorisag ~ Ev + relevel(Nem, 2) + offset(log(Nepesseg)), link = log))

Köszönetnyilvánítás

A kutatás a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosító számú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program című kiemelt projekt keretében zajlott. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

[1] Agresti. Categorical Data Analysis, Second Edition. Hoboken, New Jersey, John Wiley & Sons, Inc., 2002, pp. 115-164, 385-387, 559-563.
[2] C. Cameron. Advances in Count Data Regression Talk for the Applied Statistics Workshop, March 28, 2009. http://cameron.econ.ucdavis.edu/racd/count.html.
[3] Data Analysis Examples. UCLA: Statistical Consulting Group. from http://www.ats.ucla.edu/stat/dae/ (accessed October 12, 2013).
[4] J. Dobson. An Introduction to Generalized Linear Models, Second Edition. Boca Raton, Florida, Chapman & Hall/CRC, 2002, pp. 151-170.
[5] S. Everitt. Modern Medical Statistics: A Practical Guide. London, Arnold, 2003, pp. 1-20.
[6] Zeileis, C. Kleiber, and S. Jackman. „Regression Models for Count Data in R”, Journal of Statistical Software, vol. 27(8), pp. 1-25, 2008.
[7] Pedan. Analysis of Count Data Using the SAS System. Proceedings of the Twenty-Sixth Annual SAS® Users Group International Conference, Cary, NC: SAS Institute Inc., 2001.
[8] SAS 9.2. for Windows, SAS Institute Inc., Cary, NC, USA
[9] IBM SPSS Statistics for Windows, Version 21.0., IBM Corp., Armonk, NY, USA
[10] Stata Statistical Software: Release 8, StataCorp. 2003, College Station, TX: StataCorp LP
[11] R 2.15.3: A Language and Environment for Statistical Computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria)