# VZORČNE IN STRUKTURNE NIČLE V KONTINGENČNIH TABELAH

K.Košmelj
Biotehniška fakulteta, Univerza v Ljubljani

Povzetek

Namen tega članka je pokazati, kakšne probleme povzročajo pri analizi loglinearnih modelov vzorčne in strukturne ničle v kontingenčnih tabelah in nakazati, kako te probleme rešujemo.

ključne besede: loglinearni modeli, vzorčne ničle, strukturne ničle

Abstact

How to deal with sampling and structural zeros in log-linear models?

key words: loglinear models, sampling zeroes, structural zeroes

## 1.UVOD

Kontingenčne tabele lahko vsebujejo dve vrsti 'ničel':

- **vzorčne** ničle (ang. random zeros, sampling zeros), ki nastanejo, kadar je število enot v vzorcu premajhno, da bi našli enote z določeno kombinacijo ravni faktorjev. Če bi število enot v vzorcu dovolj povečali, bi vzorčne ničle lahko izginile. Vzorčne ničle se pojavijo zaradi vzorčne variabilnosti in zaradi relativno majhnega števila enot v vzorcu v primerjavi s številom celic v tabeli. Pričakovane frekvence za celice z vzorčnimi ničlami so pozitivne.

- **strukturne** ničle (ang. fixed zeros), ki jih dobimo tedaj, kadar ne morejo obstajati enote z določeno kombinacijo ravni faktorjev. Npr.: spol (M, Ž), vrsta raka (grlo, dojka, pljuča, itd): moški z rakom na dojkah ne obstaja. Pričakovane frekvence za celice s strukturnimi ničlami so nič.

### 1.1 Vzorčne ničle

Z loglinearnim modelom lahko ocenimo le toliko parametrov, kolikor je neničelnih frekvenc v kontingenčni tabeli.

a) **Če je število parametrov, ki jih želimo oceniti, manjše ali enako številu pozitivnih frekvenc v kontingenčni tabeli, vzorčne ničle ohranimo kot 0 v podatkih.**

b) **V primeru, ko hočemo oceniti več parametrov kot je pozitivnih frekvenc** (angl. overparametrised model), so dobljeni rezultati narobe:
- nekateri parametri so neocenjeni (ang. aliased)
- število stopenj prostosti je prevelikq
- standardne napake nekaterih parametrov so prevelike
- residuali in pričakovane frekvence so napačne.

Do takih pojavov pride lahko v naslednjih primerih:
- **maksimalni model** (saturirani model): če obstaja vsaj ena vzorčna ničla, je število parametrov večje od števila pozitivnih frekvenc.
- **vsaj ena robna vsota je 0:** tako vrstica oz. stolpec ne vsebuje nobene informacije, zato se lahko zgodi, da se določeni kontrasti ne dajo oceniti. Pri ocenjenjevanju parametrov pride do deljenja z 0, kar povzroči konvergenčne probleme. V takih primerih je potrebno vrstice (stolpce), ki imajo robno vsoto 0, odstraniti iz tabele ali jim pri modeliranju dati utež nič in ustrezno popraviti stopnje prostosti.

Poglejmo primer. V tabeli 1 je prikazano število volilcev, ki so volili za določeno stranko leta 1970 glede na njihovo izbiro leta 1964. Podatki so za Švedsko in so bili dobljeni z anketo. Zanimina nas, ali rezultati volitev 1970 odvisni od rezultatov leta 1964.

Tabela 1: število volilcev glede na njihove odločitve za stranke leta 1964 in 1970 na Švedskem. V oklepajih so pričakovane frekvence in standardizirani ostanki pod predpostavko o neodvisnosti izbire leta 1964 in leta 1970. (Vir: Linsey, str.79)

| | Komunist. | Soc.dem. | Center | Ljudska | Konzerv. |
|---|---|---|---|---|---|
| Komunist. | 22 | 27 | 4 | 1 | 0 |
| | (1.5) | (27.6) | (11.8) | (8.2) | (4.9) |
| | (16.9) | (-0.1) | (-2.3) | (-2.5) | (-2.2) |
| Soc.dem. | 16 | 861 | 57 | 30 | 8 |
| | (26.5) | (497.4) | (212.6) | (147.4) | (88.0) |
| | (-2.0) | (16.3) | (-10.7) | (-9.7) | (-8.5) |
| Center | 4 | 26 | 248 | 14 | 7 |
| | (8.2) | (153.0) | (65.4) | (45.3) | (27.1) |
| | (-1.5) | (-10.3) | (22.6) | (-4.7) | (-3.9) |
| Ljudska | 8 | 20 | 61 | 202 | 11 |
| | (8.2) | (154.0) | (65.8) | (45.7) | (27.3) |
| | (-0.1) | (-10.8) | (-0.6) | (23.0) | (-3.1) |
| Konzerv. | 0 | 4 | 31 | 32 | 140 |
| | (5.6) | (105.9) | (45.3) | (31.4) | (18.7) |
| | (-2.4) | (-9.9) | (-2.1) | (0.1) | (28.0) |

V tabeli sta dve vzorčni ničli: noben v anketo vključeni volilec ni svojega glasu dal pri prvih volitvah komunistični stanki, pri drugih pa konzervativni stranki in obratno.

Minimalni model vključuje faktorja v70 in v64. Vzorčni ničli pustimo kot 0 v podatkih, pričakovane frekvence pri teh ničlah bodo pozitivne.

| parametri v modelu | devianca | SP |
|---|---|---|
| v70 + v64 | 2259.9 | 16 |

Pričakovane frekvence in standardizirani ostanki za ta model so prikazani v tabeli 1. Očitno je model nesprejemljiv, kar se kaže:
- v preveliki vrednosti za devianco (za sprejemljiv model velja, da je vrednost deviance približno enaka SP);
- iz primerjave opazovanih in pričakovanih frekvenc in iz analize ostankov.

Če vzorčnih ničel v podatkih ne bi bilo, bi z dodajanjem člena za interakcijo v70.v64 dobili maksimalni model, za katerega velja: devianca = 0, SP = 0. Vrednost razlike za devianco pri zaporednih dveh modelih (1. model : v70 + v64 z SP1; 2.model: v70 + v64 + v64.v70 z SP2) meri pomembnost dodanega člena v64.v70. Teorija pove, da je v primeru, ko je dodani člen v modelu nepomemben, razlike devianc porazdeljena po $\chi^2$ porazdelitvi z SP = SP1-SP2. Ker je 2259 bistveno večje od kritične vrednosti za $\chi^2$ z SP = 16, ugotovimo, da je interakcija močno statistično značilna. Na osnovi tega sklepamo, da so se volilci v letu 1970 odločali v veliki meri tako kot v letu 1964.

Ker pa sta v podatkih dve vzorčni ničli, bo parametrov v maksimalnem modelu več kot pozitivnih frekvenc v kontingenčni tabeli in nastopili bodo problemi:

| parametri v modelu | devianca | SP |
|---|---|---|
| + V70.V64 | ni konvergence | 1 |

Devianca se ne da izračunati, število stopenj prostosti je 1 namesto 0. V rezultatih je nekaj pozitivnih ostankov, 8 prevelikih standardnih napak parametrov, en parameter pa je ostal neocenjen.

Če bi nas zanimale vrednosti parametrov za maksimalni model, bi morali iz podatkov izločiti ti dve ničli in ustrezno popraviti stopnje prostosti.

1.2 Strukturne ničle

Strukture ničle ne povzročajo nobenih problemov pri analizi loglinearnih modelov, če jih izločimo iz tabele. S tem zmanjšamo

število opazovanih frekvenc in tabele niso več simetrične. Strukturne ničle lahko izločimo iz podatkov tudi tako, da jim damo pri modeliranju utež 0.

## 3. PRIMER

Podatki so iz epidemiološke študije (Rojnik), ki je bila zasnovana kot študija primerov s kontrolami. Študijski primeri so bile ženske, ki so zanosile prvo leto po porodu in se odločile za abortus. Kontrole pa so bile ženske v prvem letu po porodu, ki niso bile noseče in so ustrezale še nekaterim drugim kriterijem izbire. V študijo je bilo vključenih 112 primerov in 214 kontrol. Za primere so podatke zbrali na ginekološki kliniki, kontrole pa so anketirali v otroških dispanzerjih v času vakcinacije otrok.

Avtorica študije je definirala 4 vzorce dojenja: nič, polno dojenje, polno + delno dojenje in delno dojenje. Podatki za vzorec dojenja so bili retrospektivni in 'cenzurirani' (ang. censored data) v trenutku zanositve (za primere) oz. ankete (za kontrole). Spremenljivki stanje dojenja (polno doji, delno doji, ne doji) in stanje amenoreje (amenoroična, ni amenoroična) odražata trenutno stanje ob zanositvi oziroma ob anketi.

Želimo odgovore na naslednja vprašanja:
- kako vzorec dojenja vpliva na stanje dojenja in na stanje amenoreje ob zanositvi (za primere) oziroma ob anketi (za kontrole)?
- ali sta stanje dojenja in stanje amenoreje ob zanositvi oziroma ob anketi povezani?
- v čem so bistvene razlike med primeri in kontrolami glede na te relacije ?

Podatki za primere oziroma kontrole so zapisani v trorazsežnih kontingenčnih tabelah, ki jih določajo faktorji:
- vzorec dojenja (DOJENJE)
- stane dojenja ob zanositvi (DOJZAN) oz. ob anketi (DOJANK)
- stanje amenoreje ob zanositvi (AMEZAN) oziroma ob anketi (AMEANK).

Spremenljivko DOJENJE štejemo za pojasnjevalno spremenljivko, stanja ob zanositvi oziroma anketi DOJZAN oz. DOJANK ter AMEZAN oz. AMEANK pa za odzivne spremenljivke. Glede na vprašanja zgoraj želimo za primere izvrednotiti dvofaktorske interakcije DOJENJE.DOJZAN, DOJENJE.AMEZAN ter DOJZAN.AMEZAN ter trofaktorsko interakcijo DOJENJE.DOJZAN.AMEZAN. Analogno za kontrole.

V tabeli 2 so prikazani podatki za 112 primerov, v tabeli 3 pa za 214 kontrol. Pri določenem vzorcu dojenja so možna le nekatera stanja dojenja ob zanositvi oziroma anketi, kar se odraža v pojavu strukturnih ničel, ki so v označene z -. Najprej bomo predstavili rezultate za primere, nato za kontrole. Analizo smo opravili s statističnim paketom GLIM.

3.1 Primeri

Tabela 2: Število žensk in pričakovane vrednosti po vzorcih dojenja glede na stanje dojenja in stanje amenoreje ob zanositvi. Pričakovane vrednosti so za model DOJENJE + DOJZAN + AMEZAN.

| Vzorci dojenja | Stanje ob zanositvi: Amenoreja | Način dojenja: polno | delno | nič | skupaj | Skupaj |
|---|---|---|---|---|---|---|
| nič | da | - | - | 0<br>0.3 | 0 | 4 (4%) |
| | ne | - | - | 4<br>3.7 | 4 | |
| polno | da | 2<br>0.4 | - | 1<br>1.1 | 3 | 22 (20%) |
| | ne | 4<br>5.6 | - | 15<br>14.9 | 19 | |
| polno + delno | da | - | 1<br>1.5 | 4<br>4.1 | 5 | 79 (71%) |
| | ne | - | 21<br>19.9 | 53<br>53.7 | 74 | |
| delno | da | - | 0<br>0.1 | 0<br>0.4 | 0 | 7 (6%) |
| | ne | - | 1<br>1.7 | 6<br>4.8 | 7 | |
| | Skupaj | 6<br>(5%) | 23<br>(21%) | 83<br>(74%) | 112<br>(100%) | 112(100%) |

V tabeli 2 je 14 vrednosti (11 pozitivnih frekvenc in 3 vzorčne ničle z dvema robnima vsotama 0) in 10 strukturnih ničel. Z modelom torej ne bomo mogli oceniti več kot 11 parametrov.

Poskusimo najti optimalni model. Poglejmo hierarhično zaporedje modelov.

| parametri v modelu | devianca | SP | štev.podatkov |
|---|---|---|---|
| povprečje | 221.71 | 13 | 14 |
| + DOJENJE | 127.78 | 10 | 14 |
| + DOJZAN | 103.72 | 8 | 14 |
| + AMEZAN | 6.0967 | 7 | 14 |
| + DOJENJE.DOJZAN | 5.4174 | 6 | 14 |

Samo en parameter za interakcijo DOJENJE.DOJZAN je ocenjen, 5 pa ne (kar je posledica dejstva, da lahko v tabeli DOJENJE x DOJZAN ocenimo skupno le 6 parametrov, od tega pa smo jih 5 že). Ko dodamo interakcijo DOJENJE.AMEZAN, hočemo oceniti še 6 parametrov, kar pa zaradi vzorčnih ničel z ničelnimi robnimi vsotami ne gre. Ko le-te izločimo iz tabele, se število podatkov zmanjša na 11 in dobimo pravilne rezultate

| | | | |
|---|---|---|---|
| + DOJENJE.AMEZAN | 2.803 | 2 | 11 |

Ko v model damo še zadnjo dvofaktorsko interakcijo, dobimo maksimalni model

| | | | |
|---|---|---|---|
| + DOJZAN.AMEZAN | 0 | 0 | 11 |

Najboljši model je DOJENJE + DOJZAN + AMEZAN, ki ima devianco 6.1 približno enako SP = 7. Pričakovane vrednosti za ta model se močno prilegajo opazovanim frekvencam (glej tabelo 2). Dobili smo model **popolne neodvisnosti**, saj so vse dvofaktorske interakcije neznačilne, trofaktorska interakcija pa se ne da oceniti. Ocene parametrov in njihove standardne napake v oklepajih so:

| | | |
|---|---|---|
| -2.234 | (0.7714) | povprečje |
| 1.384 | (0.5590) | DOJE(2) |
| 2.672 | (0.5166) | DOJE(3) |
| 0.2484 | (0.6302) | DOJE(4) |
| -0.0268 | (0.5370) | DOJZ(2) |
| 0.9808 | (0.4786) | DOJZ(3) |
| 2.565 | (0.3669) | AMEZ(2) |

Izračunajmo oceno za verjetnost, da 'primer' (ženska, ki zanosi prvo leto po porodu in se odloči za abortus) zanosi v času polnega dojenja in je amenoroična:

$p_{ijk} = \mu_{ijk}/N = 0.4286/112 = 4.10\ E{-}3$

da zanosi v času polnega dojenja in ni več amenoroična

$p_{ijk} = \mu_{ijk}/N = 5.5714/112 = 50.10\ E{-}3$

Torej na 1000 'primerov' pričakujemo 4, ki zanosijo v času polnega dojenja in amenoreje, in 50, ki zanosijo v času polnega dojenja in niso več amenoroične. Analogni izračuni so prikazani v tabeli 2a.

Tabela 2a: pričakovano število 'primerov' na 1000 'primerov' glede na stanje dojenja in stanje amenoreje ob zanositvi.

| | STANJE DOJENJA | | | |
|---|---|---|---|---|
| AMENOROIČNA | polno | delno | nič | Σ |
| da | 4 | 14 | 49 | 67 |
| ne | 50 | 193 | 688 | 931 |
| Σ | 54 | 207 | 737 | 998 |

V skupini 'primerov' pričakujemo približno 4 krat več žensk, ki ob zanositvi delno dojijo v primerjavi s številom, ki ob zanositvi polno dojijo, neodvisno od stanja amenoreje. Nastop menstruacije poveča verjetnost 'biti primer' približno 13 krat ne glede na stanje dojenja. Slednje izhaja iz interpretacije parametra AMEZ(2)= log (amenoroična/ni amenoroična)= 2.565).

**Sklep**: vzorci dojenja nimajo nobenega vpliva na to, v kakšnem stanju dojenja in amenoreje 'primer' zanosi. Večina 'primerov' zanosi, ko ne dojijo več in se jim je menstruacija že povrnila. Prav povratek menstruacije bistveno poveča verjetnost zanositve po porodu, ne glede na vzorec dojenja, in opozarja, da se na morebitni kontraceptivni učinek dojenja ne gre (več) zanašati. Pričakujemo lahko izredno malo žensk v skupini primerov, ki zanosijo v času polnega dojenja in so amenoroične (0.4%).

3.2 Kontrole

Tabela 3: Število žensk po vzorcih dojenja glede na stanje dojenja in stanje amenoreje ob anketi.

| Vzorci dojenja | Stanje ob anketi: Amenoreja | Način dojenja: polno | delno | nič | skupaj | Skupaj |
|---|---|---|---|---|---|---|
| nič | da | - | - | 0 | 0 | 6 (3%) |
| | ne | - | - | 6 | 6 | |
| polno | da | 15 | - | 0 | 15 | 34 (16%) |
| | ne | 4 | - | 15 | 19 | |
| polno + delno | da | - | 20 | 8 | 28 | 165 (77%) |
| | ne | - | 35 | 102 | 137 | |
| delno | da | - | 0 | 0 | 0 | 9 (4%) |
| | ne | - | 0 | 9 | 9 | |
| Skupaj | | 19 (9%) | 55 (26%) | 140 (65%) | 214 (100%) | 214(100%) |

V tabeli 3 je od 14 frekvenc 5 vzorčnih ničel. Torej ne bo mogoče mogoče oceniti več kot 9 parametrov modela. Poglejmo hierarhično zaporedje modelov.

| paramètri v modelu | devianca | SP | štev. podatkov |
|---|---|---|---|
| povp | 412.97 | 13 | 14 |
| +DOJENJE | 179.30 | 10 | 14 |
| +DOJANK | 154.73 | 8 | 14 |
| +AMEANK | 72.789 | 7 | 14 |
| +DOJENJE.DOJANK | 65.728 | 5 | 12 |
| +DOJENJE.AMEANK | 47.947 | 2 | 10 |
| +DOJANK.AMEANK | 0 | 0 | 9 |

Pri dodajanju interakcijskih členov je bilo potrebno izločiti vzorčne ničle in s tem zmanjšati število podatkov. Vrednosti nekaterih parametrov interakcijskih členov so neocenjene (število neocenjenih členov se da ugotoviti iz dvorazsežnih tabel).

Najboljši model je zadnji, kar je maksimalni model. Ocene parametrov in njihove standardne napake so:

```
-2.076    (0.7862)    povprečje
 4.784    (0.8275)    DOJE(2)
 2.833    (0.4201)    DOJE(3)
 0.4055   (0.5270)    DOJE(4)
 2.2388   (0.7012)    DOJANK(2)
 1.3228   (0.5627)    DOJANK(3)
 2.546    (0.3672)    AMEANK(2)
 aliased              DOJE(2).DOJANK(2)
 aliased              DOJE(2).DOJANK(3)
 aliased              DOJE(3).DOJANK(2)
 aliased              DOJE(3).DOJANK(3)
 aliased              DOJE(4).DOJANK(2)
 aliased              DOJE(4).DOJANK(3)
-3.867    (0.6719)    DOJE(2).AMEANK(2)
 aliased              DOJE(3).AMEANK(2)
 aliased              DOJE(4).AMEANK(2)
-1.9806   (0.4619)    DOJANK(2).AMEANK(2)
 aliased              DOJANK(2).AMEANK(2)
```

Gre za model **popolne odvisnosti**. Interpretacija interakcij je naslednja:
- interakcija med vzorcem dojenja in stanjem dojenja ob anketi se ne da izvrednotiti (pa tudi vsebinsko pri kontrolah ni zanimiva).

Ostali dve interakciji sta močno statistično značilni:
- vzorec dojenja vpliva na stanje amenoreje
- stanje amenoreje in stanje dojenja ob anketi sta povezani.

**Sklep**: pri 'kontrolah' ugotovimo, da način dojenja vpliva na stanje amenoreje in da sta stanje amenoreje in stanje dojenja močno povezani.

**Primerjava primerov in kontrol**: Dobili smo dva 'ekstremna' modela: neodvisni za primere in odvisni za kontrole. Ugotovimo lahko, da določene relacije med spremenljivkami, ki jih pri primerih ne opazimo, pri kontrolah obstojajo. To dejstvo si lahko razlagamo na dva načina: teh relacij pri primerih nikoli ni bilo ali pa jih v času zanositve ni več. Kaj torej povzroči, da se ženske po porodu lahko znajdejo v skupini 'primerov'? Delni odgovor je dejstvo, da vzorec dojenja nima (več) nobenega vpliva na stanje amenoreje. Prvi znak, da se je rodna sposobnost vrnila, je povratek menstruacije. Tedaj se verjetnost, da se ženska znajde v skupini primerov, močno poveča, ne glede na način in stanje dojenja.

## 4. LITERATURA

Dobson A.: Introduction to Statistical Modelling, Chapman and Hall, London, 1986

Fienberg S.E.: The Analysis of Cross-Classified Categorical Data, MIT Press, 1987

Krzanowski W.J.: Principles of Multivariate Analysis, A User's Perspective, Oxford University Press, 1988

Lindsey J.K.: The Analysis of Cross-Classified Categorical Data using GLIM, Lecture Notes in Statistics 56, Springer-Verlag, 1989

McCullagh P., Nelder J.A.: Generalized Linear Models, Chapman and Hall, London, 1983

Rojnik B.: Vzorci dojenja pri ženskah, ki zanosijo zgodaj po porodu, magistrsko delo, Medicinska fakulteta, Univerza v Ljubljani, Ljubljana, 1989

Priloga 1 : dvorazsežne tabele za primere

| Vzorci dojenja | Način dojenja: polno | delno | nič | skupaj |
|---|---|---|---|---|
| nič | - | - | 4 | 4 (4%) |
| polno | 6 | - | 16 | 22 (20%) |
| pol+del | - | 22 | 57 | 79 (71%) |
| delno | - | 1 | 6 | 7 (6%) |
| skupaj | 6 (5%) | 23 (21%) | 83 (74%) | 112 (100%) (100%) |

| Vzorci dojenja | amenoreja: da | ne | skupaj |
|---|---|---|---|
| nič | 0 | 4 | 4 (4%) |
| polno | 3 | 19 | 22 (20%) |
| pol+del | 5 | 74 | 79 (71%) |
| delno | 0 | 7 | 7 (6%) |
| skupaj | 8 (7%) | 104 (93%) | 112 (100%) (100%) |

| stanje dojenja | amenoreja: da | ne | skupaj |
|---|---|---|---|
| polno | 2 | 4 | 6 (5%) |
| delno | 1 | 22 | 23 (21%) |
| nič | 5 | 78 | 83 (74%) |
| skupaj | 8 (7%) | 104 (93%) | 112 (100%) (100%) |

Priloga 2 : dvorazsežne tabele za kontrole

| Vzorci dojenja | Način dojenja: polno | delno | nič | skupaj |
|---|---|---|---|---|
| nič | - | - | 6 | 6 (3%) |
| polno | 19 | - | 15 | 34 (16%) |
| pol+del | - | 55 | 110 | 165 (77%) |
| delno | - | 0 | 9 | 9 (4%) |
| skupaj | 19 (9%) | 55 (26%) | 140 (65%) | 214 (100%) |

| Vzorci dojenja | amenoreja: da | ne | skupaj |
|---|---|---|---|
| nič | 0 | 6 | 6 (3%) |
| polno | 15 | 19 | 34 (16%) |
| pol+del | 28 | 137 | 165 (77%) |
| delno | 0 | 9 | 9 (4%) |
| skupaj | 43 (20%) | 171 (80%) | 214 (100%) |

| stanje dojenja | amenoreja: da | ne | skupaj |
|---|---|---|---|
| polno | 15 | 4 | 19 (9%) |
| delno | 20 | 35 | 55 (26%) |
| nič | 8 | 132 | 140 (74%) |
| skupaj | 43 (20%) | 171 (80%) | 214 (100%) |