Liet. matem. rink, **45***, spec. nr.*, 2005, 354–362

Statistinė struktūrų analizė: kai kurios jos taikymo problemos

Marijus RADAVIČIUS (MII), J. ŽIDANAVIČIŪTĖ (VGTU)
el. paštas: mrad@ktl.mii.lt, jurz@fm.vtu.lt

1. Įvadas

Bet kuri sudėtinga sistema dažnai funkcionuoja kaip vientisa, sudėtinga struktūra, kurios tam tikras būsenas stengiamasi nusakyti ryšiais tarp atskirų tos sistemos dalių. Dažnai tai vadinama terminu „struktūrų analize“ (*Structure Learning*) [20, 17]. Minėta sistema, kurios struktūrą stengiamasi įvertinti, yra abstrakti sąvoka. Ją galima sukonkretinti, imant bet kurią sritį, kurioje yra sukauptas tam tikras duomenų kiekis, nusakantis atskirų sistemos dalių sąveiką bei visos sistemos funkcionavimą. Šiuo metu, esant ypač išvystytoms duomenų kaupimo technologijoms, yra galimybė tuos didelius informacijos kiekius panaudoti sudėtingų sistemų (struktūrų) identifikavimui. Bet dėl didelio įvairių duomenų kiekio, o taip pat dėl didelės potencialiai galimų struktūrų įvairovės bei parametrų, kuriuos reikia įvertinti skaičiaus, struktūrų identifikavimo uždavinys yra komplikuotas ne vien tik statistiniu, bet ir šiuolaikinių kompiuterinių resursų bei programinės įrangos požiūriu.

Šiame darbe nagrinėjami realūs duomenys iš Lietuvos Žmogaus Genetikos Centro kaupiamos duomenų bazės LIRECA. Parodoma, kad jau parenkant modelį trijų kokybinių kintamųjų sąveikai aprašyti susiduriama su tipinėmis sudėtingų sistemų problemomis [5, 6, 7, 8, 9, 10]: net aukšto lygio standartinė programinė įranga, šiuo atveju sistema SAS, nepajėgi išspręsti pilno, nesupaprastinto uždavinio; trūksta efektyvių metodų, padedančių „atrasti“ duomenyse struktūras ir jas pavaizduoti grafiškai; netinka tradiciniai modelio parinkimo kriterijai.

Antrame skyrelyje pateikta trumpa struktūrinės analizės modelių apžvalga, skiriant specialų dėmesį Log-tiesiniams modeliams ir jų vizualizavimui. Trečiame aprašomas atliktas statistinis tyrimas ir aptariami gauti rezultatai, gale suformuluotos išvados.

2. Struktūrinės analizės modelių apžvalga

Priklausomai nuo keliamų tikslų, struktūrų analizėje naudojami įvairūs modeliai: paslėptojo kintamojo (*latent variable*) [13] ir paslėptieji Markovo modeliai (hidden Markov model, HMM) [17], struktūrinių lygčių modeliai (structural eqations, SEM) [14], grafiniai modeliai (*graphical models*) [17, 18], Log-tiesiniai (log-linear) [1, 2, 3, 4, 18] ir kiti [19, 17]. Šis skirstymas gana sąlyginis, išvardintos modelių klasės yra tarpusavyje persipynusios.

Paslėptojo kintamojo modelio idėja – pagal stebimo dydžio Y, kuris priklauso ir nuo X, reikšmes atstatyti nestebimas X reikšmes ir įvertinti jų tikimybines charakteristikas. Analogiškas uždavinys yra ir **paslėptuose Markovo modeliuose** tik Y yra atsitiktinis procesas, o X yra nestebima Markovo grandinė. Šio tipo modeliams priskiriami klasterinė, faktorinė, logistinė analizės [1, 2, 4], modeliai su cenzūruotais duomenim. Paslėptojo kintamojo modelis yra tipinis **struktūrinių lygčių**, taikomų ne Gauso sistemoms aprašyti, modelis.

Grafiniai modeliai (*pavadiniams yra kilęs iš matematinio termino „grafas“*) – neatskiriama visų struktūrinių modelių dalis, nes grafai geriausiai tinka struktūros vidinių sąryšių pavaizdavimui. Sistemos (struktūros) elementai arba kintamieji tapatinami su grafo viršūnėmis, ir dvejų viršūnių sujungimas reiškia tiesioginę priklausomybę tarp atitinkamų sistemos elementų. Nesujungtos grafo viršūnės yra nepriklausomos arba sąlyginai nepriklausomos, kai yra žinomos kitų viršūnių būsenos. Priklausomybė tarp kintamųjų gali būti nusakoma parametriniu sąlyginiu skirstiniu arba dar kitaip vadinama potencialine funkcija (*potential function*). Grafo jungčių aibė ir sąlyginiai skirstiniai drauge apibrėžia bendrą visų grafo kintamųjų tikimybinį skirstinį (*joint probability distribution*), nusakantį visos sistemos funkcionavimą. Paprastai grafo jungčių aibė vadinama grafo struktūra, o sąlyginių skirstinių parametrai tiesiog grafo parametrais.

Yra du tipai grafinių modelių: kryptiniai (*directed*) ir nekryptiniai (*undirected*). Kryptiniams grafiniams modeliams priskiriami Bajeso tinklai (*BNs*) [17], belief networks, priežastiniai modeliai (*causal models*) [19] ir kt. Nekryptiniai modeliai žinomi kaip Markovo tinklai (*Markov Networks* arba *Markov random fields (MRFs)*), **Log-tiesiniai modeliai** (*Log-linear model*) [1, 2, 3, 4]) ir kt. Log-tiesinius modelius aptarsime plačiau:

Log-tiesiniai modeliai. Kai kalbama apie log-tiesinius modelius, paprastai turima omenyje daugiamačiai Puasono arba multinominiai tikimybiniai modeliai, aprašantys sudėtingus kokybinių požymių tarpusavio sąryšius ir skirti kryžminių dažnių lentelių (*contingency tables, cross-tabs*) daugiamatei statistinei analizei. Šiuose modeliuose nėra kintamųjų klasifikavimo į aiškinamuosius ir aiškinančiuosius. Juose visi kintamieji traktuojami kaip aiškinantieji, o aiškinamuoju kintamuoju laikomas konkrečios kokybinių kintamųjų reikšmių (būsenų) kombinacijos stebėtas dažnis. Tuo šie modeliai primena koreliacinę ir faktorinę analizes.

Tačiau log-tiesinių modelių klasei priklauso ne vien tik kokybinių kintamųjų modeliai, logistinė (binominė) ir Puasono regresija, o tam tikra prasme ir multiplikatyvūs ekonometriniai modeliai, yra atskiri jos atvejai [3]. Log-tiesiniai modeliai gali aprašyti labai sudėtingas nagrinėjamų kintamųjų sąveikas, ne tik porines, bet ir aukštesnio lygio, todėl šiuos modelius ne visada pavyksta aprašyti kintamųjų sąveikų grafu, tam reikia papildomų sąlygų [3, 4]. Modeliai, kurie tenkina šias sąlygas, vadinami *grafiniais log-tiesiniais modeliais* ir yra atskiras nekryptinių grafinių modelių atvejis.

Turint didelį skaičių kokybinių kintamųjų su tokiu pat dideliu skaičiumi kategorijų kiekvienam jų, iš karto parinkti vieną sudėtingą log-tiesinį modelį, kuris aprašytų bendrus ryšius tarp visų kintamųjų, bei įvertinti to modelio parametrus yra pakankamai sudėtinga. Situaciją dar labiau apsunkina nedidelis stebėjimų skaičius lentelės ląstelėse, kuris atsiranda lentelę vis labiau smulkinant pagal joje esančius kokybinius kintamuosius.

Kadangi log-tiesiniais modeliais siekiama aprašyti visų stebėtų dažnių bendrą tikimybinį skirstinį, tai, didėjant kintamųjų skaičiui, jų tarpusavio sąryšių galimų struktūrų aibė labai greitai auga, eksponentiškai auga ir modelio parametrų kiekis, kuris taip pat priklauso ir nuo nagrinėjamų kintamųjų galimų kategorijų skaičiaus. Todėl net ir didelėms imtims daugelis ląstelių dažnių lentelėse gali būti tuščios. Tai seniai žinoma išretintų (*sparse*) lentelių problema, kuriai literatūroje skirta daug dėmesio [5, 6, 7, 8, 9, 10, 11], bet galutinio sprendimo kol kas nerasta.

Trimačio log-tiesinio modelio identifikacijos (parinkimo) problema. Duomenų bazėje LIRECA nuo 1993 m. kaupiami duomenys apie visus Lietuvos naujagimius, turinčius įgimtas raidos anomalijas (ĮRA). Šiame tyrime naudojami trys kintamieji, kurie laikomi kokybiniais: A nusako ĮRA rūšį, jų yra 17-ka ($I = 17$), R yra Lietuvos rajono identifikatorius ($J = 45$), M yra naujagimio gimimo metai, apimantys dešimt metų laikotarpį ($K = 10$). Uždavinys yra sudaryti šių kintamųjų tarpusavio ryšių modelį.

Trijų kintamųjų dažnių lentelei (*three-way contingency table*) prisotintas (*saturated*) log-tiesinis modelis užrašomas taip:

$$\log(m_{ijk}) = \mu + \lambda_i^A + \lambda_j^R + \lambda_k^M + \lambda_{ij}^{AR} + \lambda_{ik}^{AM} + \lambda_{jk}^{RM} + \lambda_{ijk}^{ARM},$$
$$i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K,$$

kur m_{ijk} yra dažnių lentelės ląstelės (i, j, k) prognozuojamas vidutinis dažnis (stebėto dažnio n_{ijk} vidurkis $m_{ijk} = \mathrm{E}n_{ijk}$). Multinominio modelio atveju$m_{ijk} = \pi_{ijk}N$, kur π_{ijk} yra tikimybė, kad stebėjimas pateks į ląstelę(i, j, k), N – imties dydis.

Kad modelis būtų identifikuojamas, būtina apibrėžti papildomas sąlygas modelio parametrams $\mu, \lambda_i^A, \lambda_j^R, \lambda_k^M, \lambda_{ij}^{AR}, \lambda_{ik}^{AM}, \lambda_{jk}^{RM}, \lambda_{ijk}^{ARM}$.

Pažymėkime $\lambda_{+jk}^{ARM} = \sum_{i=1}^{I} \lambda_{ijk}^{ARM}$. Analogiškai apibrėžiami ir kiti dydžiai: λ_{i+k}^{ARM}, π_{i+k} ir kt.

Tada reikalaujama, kad

$$\lambda_+^A = \lambda_+^R = \lambda_+^M = 0, \quad \lambda_{+j}^{AR} = \lambda_{i+}^{AR} = 0, \quad \lambda_{+k}^{AM} = \lambda_{i+}^{AM} = 0,$$
$$\lambda_{+k}^{RM} = \lambda_{j+}^{RM} = 0, \quad \lambda_{+jk}^{ARM} = \lambda_{i+k}^{ARM} = \lambda_{ij+}^{ARM} = 0,$$

su dar viena papildoma sąlyga multinominio skirstinio atveju: $m_{+++} = N$. Toliau laikysime, kad stebėtų dažnių skirstinys yra Puasono.

Log-tiesiniuose modeliuose galima išskirti kelių tipų nepriklausomybes tarp kintamųjų: tarpusavio (*mutual),* bendra *(joint),* marginalinė *(marginal)* ir sąlyginė (*conditional*). Tarpusavio nepriklausomybė tarp trijų prieš tai minėtų kokybinių kintamųjų galioja, kai $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$. Tokią nepriklausomybę atitinka modelis: $\log(m_{ijk}) = \mu + \lambda_i^A + \lambda_j^R + \lambda_k^M$, t.y. visi sąveikų parametrai lygūs nuliui. Kintamasis R yra bendrai (*jointly*) nepriklausomas nuo AM, kai $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$. Tokia nepriklausomybė užrašoma $\log(m_{ijk}) = \mu + \lambda_i^A + \lambda_j^R + \lambda_k^M + \lambda_{jk}^{AM}$. Kintamieji A ir R yra sąlyginai nepriklausomi, kai žinomas M, jeigu $\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$. Log-tiesinis modelis šiam atvejui – $\log(m_{ijk}) = \mu + \lambda_i^A + \lambda_j^R + \lambda_k^M + \lambda_{ik}^{AM} + \lambda_{jk}^{RM}$. Kintamieji A ir R yra marginaliai (*marginally*) nepriklausomi, kai $\pi_{ij+} = \pi_{i++}\pi_{+j+}$

Hipotezėms apie log-tiesinio modelio parametrų ir sąveikų statistinį reikšmingumą, o taip pat parinkto modelio suderinamumą (*Goodness of Fit*) tikrinti naudojama tikėtinumo santykiu paremta statistika

$$G^2 = 2\sum n_{ijk} \log(n_{ijk}/\hat{m}_{ijk}).$$

Čia $\hat{m}_{ijk}$ yra stebėtų dažnių ląstelėje $(i.j.k)$ prognozė, naudojant pasirinktą log-tiesinį modelį. Kai pasirinktas modelis yra teisingas, statistika G^2 turi (prie tam tikrų sąlygų) asimptotinį chi-kvadrat skirstinį su laisvės laipsnių skaičiumi, lygiu ląstelių kiekio dažnių lentelėje ir į modelį įtrauktų parametrų skaičiaus skirtumui. Ši statistika, lyginant su Pirsono chi-kvadrat statistika χ^2 turi svarbų privalumą, nes yra adityvi ir gali būti taikoma lygiai taip pat, kaip kvadratų sumų statistikos dispersinėje analizėje [2, 3, 4]. Tačiau išretintose lentelėse G^2 skirstinio aproksimacija chi-kvadrat skirstiniu yra netiksli. Kai ląstelių lentelėje yra labai daug, natūralu tikėtis, kad tas skirstinys bus artimas normaliajam. Prie tam tikrų reguliarumo sąlygų taip ir yra [7, 11], bet bendru atveju gali būti tinkamesnės aproksimacijos kitais skirstinias (log-normaliuoju [8], gama [10]) ir skirstinių mišiniais [6, 9]. Praktikoje vietoje aproksimacijų teoriniais skirstiniais taikomas bootstrap‘o metodas. Darbe [12] parodyta, kad parametrinio bootstrap‘o metodas tinka ir labai išretintoms lentelėms, tačiau pagal prasmę jis atitinka parametrinį testą ir tuo esminiai skiriasi nuo suderinamumo kriterijaus G^2, kuris lygina pasirinktą parametrinį modelį su pilnu (neparametriniu) log-tiesiniu modeliu. Taigi, prarandamas vienas iš svarbiausių log-tiesinių modelių privalumų. Kaip pastebėta [10], ar lentelė išretinta ar ne, priklauso ne vien nuo to, kiek joje yra ląstelių su mažu dažniu, o ir nuo to, koks yra pasirinktas (bazinis) modelis. Todėl nesunku sukonstruoti (dirbtinę) situaciją, kai parametrinis bootstrap‘as duos neteisingą rezultatą. Todėl šiame darbe buvo taikomas ir neparametrinio bootstrap‘o metodas. Deja, taikomas tiesiogiai, be papildomo „glodinimo“, jis duoda paslinktus rezultatus.

3. Statistinis tyrimas

LIRECA duomenų bazės pagrindu (3533 stebėjimai) buvo tiriamos Lietuvos naujagimių įgimtų raidos anomalijų (ĮRA) bendros tendencijos ir teritorinio pasiskirstymo ypatumai. Minėtoje duomenų bazėje yra didelis skaičius įvairių kintamųjų, susijusių su kiekvienu pacientu, bet vartotojui lengviausiai interpretuojami Lietuvos rajonai (jų yra 45), metai (10) ir ĮRA rūšys (17). Pilnas (*saturated*) log-tiesinis modelis turi 7650 parametrų. SAS procedūroms CATMOD ar GENMOD [21], kuriose realizuoti log-tiesiniai modeliai, šis parametrų vertinimo uždavinys, pasirodo, neįveikiamas. Modelį supaprastinus iki antros eilės sąveikų, lieka 1314 parametrų, bet CATMOD taip pat jų neįvertina. Tą problemą tenka spręsti grupuojant tam tikras kategorijas arba tiesiog neįtraukiant į modelį vienų ar kitų kintamųjų. Ją dar labiau apsunkina tai, kad rajonai labai skiriasi gyventojų skaičiumi, o tuo pačiu ĮRA dažniais. Lentelėje greta ląstelių su dideliu dažniu yra daug ląstelių su dažniais jose mažesniais už 5. Šiuo atveju tikrasis G^2 statistikos skirstinys gali ypač nukrypti nuo teorinio.

Todėl, analizuojant duomenis, buvo parenkama keletas log-tiesinių modelių. Pradžioje kintamųjų *anomalija*, *rajonai* ir *metai* reikšmės tam tikru būdu apjungiamos, sudarant naujus kintamuosius su mažesniu reikšmių kiekiu ir tuo pačiu sumažinant modelio parametrų skaičių. Parinkus adekvatų log-tiesinį modelį (visi modelio parametrai yra statistiškai reikšmingi ir tikėtinumo santykio testas G^2 neatmeta modelio), pagal jo prognozuotas tikimybes kiekvienai ląstelei generuojama 200 naujų multinominių dydžių (dažnių lentelių) su minėtomis tikimybėmis. Kiekvienai naujai lentelei vėl vertinamas tas pats log-tiesinis modelis ir išsaugoma *Likelihood Ratio* testo naudojamos statistikos G^2 reikšmė, kuri, esant teisingai nulinei hipotezei, turi chi-kvadrat skirstinį, kurio laisvės laipsnių skaičius priklauso nuo parametrų neįtrauktų į pilną *(saturated)* modelį kategorijų skaičiaus. Ši hipotezė tikrina, ar modelio parametrai prie neįtrauktų į modelį kintamųjų yra lygūs nuliui.

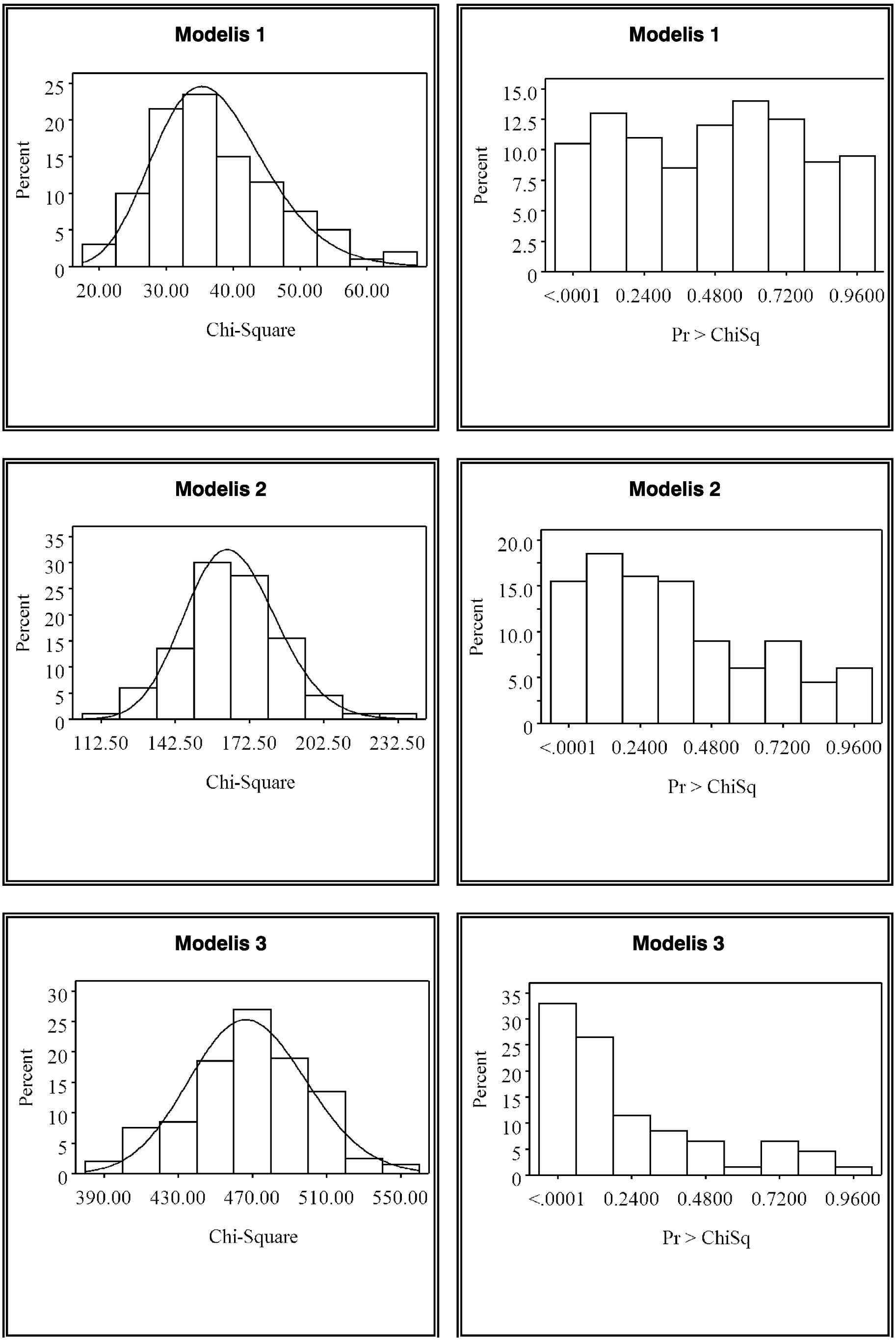
Pradinis grupavimas yra toks: rajonai suskirstyti į tris grupes: *didelius,* kurie apjungia Vilnių, Kauną, Panevėžį, Klaipėdą ir Šiaulius (bendrai duomenų bazėje jie pasitaiko 1716 kartų), *vidutinius* (viso 688 stebėjimai) ir *mažus* (viso 1129 stebėjimai); ĮRA taip pat suskirstytos į tris grupes: pirmąją sudaro didelę dalį duomenų bazėje užimančios *širdies ir širdies kraujagyslių anomalijos* (*S-SK*) (viso 989 stebėjimai), antrąją – taip labai dažnai pasitaikančios *DVD* (viso 824) ir trečiąją *vidutinės* anomalijos (*Virškinimo, Chromosominės, Kvėpavimo, Nervinio vamzdelio ir Urogeninės* – viso 1720 stebėjimai). Likusios, labai retos, nenagrinėjamos. Tokiu būdu gautoje lentelėje minimali dažnių reikšmė ląstelėje yra 9, o maksimali lygi 104. Esant reikšmingoms visoms kintamųjų sąveikoms, log-tiesinio modelio parametrų skaičius būtų lygus 90 (tiek ląstelių yra dažnių lentelėje).

Įvertintas modelis (1 lentelė) suderintas ($G^2 = 44.18,\ p = 0.1643$) ir rodo, kad rajonų ir ĮRA grupės yra tiesiogiai susiję tarpusavyje ($p < 0.001$). Be to, ši sąveika dar priklauso ir nuo metų, t.y kiekvienais metais rajonų ir ĮRA grupių sąveika statistiškai reikšmingai skiriasi ($p = 0.0044$). Likusios [RM] ir [AM] sąveikos yra statistiškai nereikšmingos. Šiame modelyje nėra kintamųjų, kuriems galiotų sąlyginė nepriklausomybė. Parinktą log-tiesinį modelį galima būtų užrašyti taip: $\log(m_{ijk}) = \mu + \lambda_i^A + \lambda_j^R + \lambda_k^M + \lambda_{ij}^{[RA]} + \lambda_{ijk}^{[RAM]}$. Pastebėkime, kad šis modelis nėra grafinis.

Antras modelis gaunamas smulkinant pradinę dažnių lentelę: rajonų grupės paliekamos tos pačios (*dideli, vidutiniai ir maži rajonai*), o vidutinių ĮRA grupė išskaidoma į atskiras ĮRA, apjungtas pirmame modelyje. Dėl smulkesnio lentelės skaidymo padidėja ląstelių skaičius lentelėje, o tuo pačiu sumažėja ir jose esantys dažniai, nors parametrų skaičius lieka tas pats: rajonai (*vidutiniai, dideli* ir *maži)*, anomalijos (*S-SK*, *DVD* ir *vidutinės)* bei *metai*. Tiesiog dabar įvertinti tiems patiems parametrams

1 lentelė. Log-tiesinis modelis su stambiausiu lentelės smulkinimu

Source	**DF**	**ChiSq**	**ProbChiSq**
rajonai	2	368,09	<.0001
anomal	2	289,48	<.0001
metai	9	20,08	0,0174
rajonai*anomal	4	34,93	<.0001
rajonai*anomal*metai	36	62,11	0.0044
Likelihood Ratio	**36**	**44,18**	**0,1643**



1 pav. σ^2 statistikos ir jos p-reikšmių histogramos.

2 lentelė

Testas	Modelis 1	Modelis 2	Modelis 3
Kolmogorov-Smirnov	0,053	0,25	0,155
Cramer-von Mises	0,02	0,25	0,179
Anderson-Darling	0,014	0,25	0,110

yra daugiau ląstelių, bet su mažesniu skaičiumi dažnių kiekvienoje iš jų. Šiuo atveju jau turime 210 ląstelių. Visų parametrų įvertiniai ir jų p-reikšmės nesikeičia, tačiau G^2 testas rodo, kad šiuo atveju sudarytas modelis nėra adekvatus ($G^2 = 1119.01$, $p < 0.001$).

Trečias modelis parenkamas pradinę dažnių lentelę smulkinant ne tik pagal ĮRA, bet ir pagal rajonus, atskirai išskiriant *didžiuosius* rajonus (*Vilnius, Kaunas, Klaipėda, Panevėžys ir Šiauliai*), o likusius *vidutinius* ir *mažus* rajonus paliekant apjungtus kaip ir prieš tai modelyje. Dabar jau lentelėje yra 490 ląstelių, bet parametrų skaičius lieka tas pats.Vėl gi, modelis nėra adekvatus ($G^2 = 1795, 1$, $p < 0.001$).

Gautų trijų modelių su skirtingu ląstelių, bet vienodu modelio parametrų skaičiumi adekvatumui patikrinti taip pat buvo naudojamas parametrinis bootstrap‘as (žr. 2 lentelę), atliekant 200 generacijų pagal tų modelių pagrindu prognozuotas tikimybes. Kiekvienoje generacijoje vertinamas tas pats log-tiesinis modelis, išsaugoma G^2 statistikos reikšmė bei ją atitinkanti p-reikšmė ir nagrinėjamas gautas šių statistikų empirinis skirstinys.

Kaip buvo minėta anksčiau, skaitoma, kad esant teisingam log-tiesiniam modeliui, G^2 statistika turi chi-kvadrat skirstinį. Visais trim bootstrap‘o atvejais tik kai kurie testai atmetė šią hipotezę, bet, lyginant p reikšmės skirstinį, matosi, kad didėjant ląstelių skaičiui lentelėje p-reikšmių empirinis skirstinys darosi vis labiau netolygus (žr. 1 pav.). Tai rodo, kad lentelių išretinimas visų pirma atsiliepia G^2 statistikos skirstinio uodegoms. Parametriniu bootstrap‘o metodu įvertintos modelių adekvatumo p-reikšmės yra atitinkmai 0.215, < 0.005 ir < 0.005.

Palyginimui su parametriniu bootstrap‘u šiame tyrime naudojamas ir neparametrinis bootstrap‘as. Kadangi dėl išaugusio laisvės laipsnių skaičiaus jis turi ryškų poslinkį į dešinę, tai jį prasminga taikyti tik antrojo ir trečiojo modelio lentelėms, kai nulinė hipotezė atmetama. Minėtoms skirtingo skaidymo lentelėms pagal stebėtus santykinius dažnius buvo generuota 200 naujų atsitiktinių lentelių ir kiekvienai iš jų apskaičiuota tikėtinumo santykio statistika G^2, matuojanti neatitikimą tarp pradinių (stebėtųjų) ir generuotų dažnių. Gauti rezultatai patvirtino, kad modeliai nėra adekvatūs: abiem atvejais $p < 0.005$.

4. Išvados

Atliktas tyrimas parodė, kad sąryšių struktūros vertinimas ir tarp trijų kokybinių kintamųjų jau gali būti gana sudėtingu uždaviniu. Naudojant standartines SAS procedūras CATMOD ir GENMOD dėl didelio parametrų skaičiaus (7650) modelyje nepavyko jiems parinkti log-tiesinio modelio, kuris aprašytų smulkiausio skaidymo dažnių lentelę (pagal visus rajonus, anomalijų rūšis ir dešimt metų). Dėl to teko mažinti parametrų skaičių kokiu nors būdu apjungiant lentelės ląsteles. Bet tai nėra tinkamas sprendimas, nes apjungimo būdas įtakoja suderinamumo kriterijaus reikšmę, o tuo pačiu ir tai, koks modelis bus pripažintas suderintu. Pirmasis modelis, parinktas pagal gerokai sustambintą pradinę dažnių lentelę, buvo atmestas šiek tiek mažiau sustambintoms lentelėms net naudojant labai konservatyvų neparametrinio bootstrap‘o testą.

Parametrinio bootstrap‘o metodu gauta G^2 statistikos skirstinio aproksimacija skiriasi nuo klasikinės aproksimacijos chi-kvadrat skirstiniu visų pirma skirstinio uodegoje, bet vidutiniškai išretintose lentelėse duoda palyginamus rezultatus. Tačiau parametrinis bootstrap‘as iš esmės yra *parametrinis* testas. Neparametrinis bootstrap‘as, kuris būtų neparametrinio kriterijaus analogas, deja, turi ryškų poslinkį. Jį taikant reiktų naudoti tam tikrą lentelių „suglodinimo“ procedūrą.

Literatūra

1. M.E. Stokes, C.S. Davis, G.S. Koch, *Categorical Data Analysis Using the SAS*(*R*) *System*, SAS Institute, Cary, NC (2001).
2. A. Agresti, *Categorical Data Analysis*, John Wiley &Sons, New York (1990).
3. R. Christensen, *Log-Linear Models*, New York (1990).
4. T.J. Santer, D.E. Duffy, *The Statistical Analysis of Discrete Data*, New York (1989).
5. P. Burman, On some testing problems for sparse contingency tables, *Journal of Multivariate Analysis*, **88**(1), 1–18 (2004).
6. W.G. Cochran, The χ^2 test of Goodness-of-Fit, *Annals of Mathematical Statistics*, **23**, 315–345 (1952).
7. K.J. Koehler, Goodness-of-fit tests for log-linear models in sparse contingency tables, *JASA*, **81**, 483–493 (1986).
8. B.P. Lawal, J.G. Upton, An approximation of the distribution of χ^2 test of goodness-of-fit statistics for use with small expectations, *Biometrika*, **67**, 447–453 (1980).
9. J.K. Yarnold, *Tools for Statistical Inference*: *Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd ed., Springer-Verlag, New York (1970).
10. M.-Y. Hu, *Model Checking for Incomplete High Dimensional Categorical Data*, Dissertation, University of California, Los Angeles (1999). `[http://citeseer.ist.psu.edu/cache/papers/cs/20551/http:zSzzSzwww.stat.ucla.eduzSzthesesszSzmingyi.pdf/hu99model.pdf]`.
11. U. Muller, G. Osius, Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, *Statistics*, **37**(2), 119–143 (2003).
12. M. von Davier, Bootstrapping goodness-of-fit statistics for sparse categorical data - results of a Monte Carlo study, *Methods of Psychological Research Online*, **2**(2) (1997). `http://www.mpr-online.de/issue3/art5/article.html]`.
13. J.C. Loehlin, *Latent Variable Models*: *an Introduction to Factor, Path, and Structural Analysis*, Erlbaum, Hillsdale, New York (1997).
14. R. Mueller, *Basic Principles of Structural Equation Modeling*, Springer-Verlag, New York (1996).
15. M. Friendly, *User's Guide for MOSAICS Version 3.6*, Psychology Department, York University. `[http://euclid.psych.yorku.ca/SCS/mosaics.pdf]`.
16. M. Friendly, Mosaic displays for loglinear models, in: *ASA Meetings* (*Statistical Graphics Section*): *Proceedings of the Statistical Graphics Section*, Psychology Department, York University (1992), pp. 61–68.
17. K. Murphy, *An Introduction to Graphical Models*, Technical Report, Intel Research Technical Report (2001).

18. L.M. Koehly, S.M. Goodreau, M. Morris, *The Link between Exponential Random Graph Models and Loglinear Models for Networks*, Center for Studies in Demography and Ecology Working Paper No.03-05, University of Washington.
19. J.B. Tenenbaum, T.L. Griffiths, Structure learning in human causal induction, *Advances in Neural Information Processing Systems*, **13**, 59–65 (2001).
20. Peter von Rohr, *GM Seminar. Learning Structure from Data* (2002).
21. SAS Institute Inc. 2004. SAS/STAT 9.1 *User's Guide*, Cary, SAS Institute Inc.
22. SAS Institute Inc. 2004. SAS/IML 9.1 *User's Guide*, Cary, SAS Institute Inc.

SUMMARY

M. Radavičius, J. Židanavičiūtė. Structure learning: some testing problems

The work is based on data about the prevalence of congenital anomalies among newborns in Lithuania. The log-linear model is used to assess dependence structure of a subset of categorical variables. It is shown that fitting the log-linear model with just three categorical variables can be a rather complicated task due to large number of unknown parameters and cells in the contingency table. The classical chi-squre test and the bootstrap technique are compared for testing goodness-of-fit. The results demonstrate that the number of cells of even nonsparse contingency tables has significant impact on the tail distribution of the likelihood ratio statistics.

Keywords: contingency tables, log-linear models, categorical data, bootstrap.