Multivariantna analiza nominalnih podatkov

BLAŽ VODOPIVEC

Logistični modeli in logistična regresija

Problemi, s katerimi se ukvarjajo družboslovni raziskovalci, so kompleksni. V vsakem nastopa vrsta dejavnikov , ki so medsebojno prepleteni in vplivajo drug na drugega. Multivariantne metode analize podatkov so v mnogočem omogočile raziskovalcem vpogled v probleme, v katerih bi se izgubili, če bi jih proučevali zgolj s preprostimi univariantnimi in bivariantnimi metodami. Na žalost pa je bila do pred kratkim uporaba multivariantnih metod statistične analize, z nekaj izjemami, omejena na podatke, izmerjene vsaj na intervalnem nivoju. Večina multivariantnih metod tako ali drugače izhaja iz matrike Pearsonovih produkt-moment korelacijskih koeficientov oziroma kovarianc. Postopki ocenjevanja parametrov pa temeljijo na predpostavki o multivariantni normalni distribuciji variabel, oziroma vsaj rezidualov (napake).

V mnogih raziskavah je del ali večina zbranih podatkov na nominalnem merskem nivoju. To je do pred kratkim za raziskovalca, ki je hotel analizirati kompleksne odnose med več spremenljivkami, predstavljalo hudo neprijetno situacijo. Če se je hotel izogniti pregledovanju množice kontingenčnih tabel, ki so rezultat hkratnega križanja več spremenljivk, je bila njegova edina možnost, da se je s takimi ali drugačnimi triki "prisleparil" skozi multivariantne tehnike, namenjene intervalnim spremenljivkam. Vendar je pri tem vzel v zakup napake pri ocenah parametrov in težave pri interpretaciji outputa analize.

V zadnjih dvajsetih letih pa so statistiki razvili vrsto novih metod, namenjenih multivariantni analizi nominalnih podatkov. Uporaba nekaterih od teh metod, ki analizo nominalnih podatkov postavljajo tako rekoč ob bok analizi podatkov višjih merskih nivojev, je v zadnjem času doživela v družboslovnem raziskovanju pravi razmah. Zato jih velja na kratko predstaviti.

Tokrat bodo predstavljene metode za analizo vplivov med manifestnimi spremenljivkami, ki imajo med klasičnimi multivariantnimi metodami vzporednico v multipli regresiji. Prihodnjič bodo predstavljene metode za analizo latentnih spremenljivk, oziroma vpliva le-teh na manifestne, ki imajo svojo vzporednico v faktorski analizi oziroma v merskem modelu analize kovariančnih struktur (glej npr. Vodopivec 1988).

Metode za analizo vplivov med manifestnimi nominalnimi spremenljivkami lahko v grobem razdelimo takole: za analizo vpliva nominalnih neodvisnih spremenljivk na

nominalno odvisno uporabljamo *logistične* modele. Za analizo vpliva intervalnih neodvisnih spremenljivk na dihotomno odvisno spremenljivko pa ponavadi uporabljamo *logistično regresijo* ali *probit regresijo.*

Logistični modeli

Vpliv nominalne spremenljivke na drugo analiziramo s pomočjo kontingenčne tabele. S hi-kvadrat testom preverimo predpostavko o neodvisnosti spremenljivk. S pomočjo raznih koeficientov (koeficient kontingence, fi, tau, gama itd.) in s pomočjo pregledovanja vsebine celic tabele dobimo vpogled v moč zveze med spremenljivkama. Če pa nas zanima simultan vpliv več nominalnih neodvisnih spremenljivk na nominalno odvisno, se situacija zaplete. Vsaka nadaljnja spremenljivka pomeni dodatno dimenzijo v kontingenčni tabeli in pri več kot treh dimenzijah postanejo tabele praviloma popolnoma nepregledne.

Vrh tega so vsi prej omenjeni statistični testi in koeficienti prirejeni za dvodimenzionalne tabele. Zato raziskovalec nima nobenega statističnega indikatorja,ki bi mu povedal, koliko dodatne neodvisne spremenljivke, ki jih je vključil v analizo, prispevajo k pojasnjevanju razpršitve odvisne spremenljivke. Tudi če tako večdimenzionalno tabelo sploščimo v dvodimenzionalno (z odvisno spremenljivko na eni dimenziji in z vsem kombinacijami vrednosti neodvisnih spremenljivk na drugi dimenziji), nam prej omenjeni indikatorji ustrezno ne pokažejo skupnega vpliva vseh spremenljivk (npr. analogno koeficientu multiple korelacije), prav tako pa ne moremo oceniti prispevka vsake posamezne spremenljivke (npr. analogno posameznim regresijskim koeficientom).

Logistični modeli so podzvrst log-linearne analize. Log-linearna analiza na razmeroma enostaven način razreši probleme analize večdimenzionalnih kontingenčnih tabel. Vprašanje, ki ga skušamo razrešiti z log-linearno analizo, je, koliko posamezne spremenljivke in njihove interakcije prispevajo k razporeditvi frekvenc v celicah take tabele, oziroma k verjetnostim, da se bo posamezna enota (npr. respondent) znašla v določeni celici take tabele. Splošna log-linearna analiza ne pozna delitve na odvisne in neodvisne spremenljivke. Odvisna spremenljivka je v tem primeru ravno celična verjetnost.

Denimo, da analiziramo i X k X j tabelo. P_{ijk} naj bo verjetnost, da ima posameznik i-ti atribut na prvi spremenljivki, j-ti atribut na drugi spremenljivki in k-ti na tretji. Če predpostavljamo, da med spremenljivkami ni interakcij, velja $P_{ijk} = P_i * P_j * P_k$. Ker pa so ravno interakcije tisto, kar nas pri analizi večdimenzionalnih kontingenčnih tabel zanima, pomeni, da je

$$P_{ijk} = P_i * P_j * P_k * P_{i*j} * P_{i*k} * P_{i*j*k}.$$

Log-linearna analiza s pomočjo take ali drugačne logaritmične transformacije pretvori multiplikativni model na desni strani zgornje enačbe v linearni aditivni:

$$f(P_{ijk}) = f(P_i) + f(P_j) + f(P_k) + f(P_{i*j}) + ... + f(P_{i*j*k})$$

Nato s pomočjo enega od statističnih algoritmov (ponavadi je to metoda maksimalne zanesljivosti - *maximum likelihood*) z analizo enačb za vse celice tabele (P_{ijk}, P_{i-1jk}...)

oceni vrednost parametrov $f(P_i)$, $f(P_{i-1})$, ... $f(P_j)$, $f(P_{j-1})$, ... $f(P_{i*j*k})$, $f(P_{i-1*j*k})$... S pomočjo antilogaritmične transformacije lahko potem izračunamo originalne parametre modela. Osnovna enota pri log-linearni analizi torej niso posamezniki (primeri), ampak celice kontingenčne tabele.

Logistični modeli so kategorija log-linearnih modelov, pri katerih raziskovalec eno od spremenljivk vnaprej določi za odvisno in ugotavlja vpliv ostalih spremenljivk na njeno razpršitev. Logistična analiza da oceno parametrov, ki povedo, v kakšni meri posamezne prediktorske spremenljivke in njihove interakcije vplivajo na verjetnost, da bo posameznik prišel v določeno kategorijo odvisne spremenljivke. Ti parametri so analogni regresijskim koeficientom v regresijski enačbi, z dvema razlikama. Prvič, model ni aditiven, ampak multiplikativen. In drugič, enačba ne napoveduje vrednosti odvisne spremenljivke, temveč verjetnost, da bo posameznikov odgovor v določeni kategoriji odvisne spremenljivke. Denimo, da je posameznikov odgovor v i-ti kategoriji prve neodvisne spremenljivke in j-ti kategoriji druge neodvisne spremenljivke. Razmerje med verjetnostjo, da bo njegov odgovor v k-ti kategoriji odvisne spremenljivke in verjetnostjo, da bo v referenčni kategoriji, kaže enačba (1)

$$P_{ijk}/P_{ijR} = B_k * B_{i*k} * B_{j*k} * B_{i*j*k}, \qquad (1)$$

kjer so z B označeni parametri logističnega modela.

Izraz modeliranje se uporablja zato, ker raziskovalec hkrati z računanjem parametrov preizkuša, ali določen načrt analize vplivov (model), ki vključuje določene glavne in interakcijske učinke, v zadostni meri pojasnjuje razpršitev odvisne spremenljivke v tabeli (torej, ali napovedana razpršitev statistično pomembno odstopa od dejanske). Za testiranje modela se uporabljajo razne variante hi-kvadrat testa. Poleg tega so raziskovalcu na voljo še sumarni indikatorji (npr. koeficient entropije, koeficient koncentracije), ki, analogno koeficientu multiple korelacije, povedo, kolikšen delež razpršitve odvisne spremenljivke pojasnujeje specificirani model.

Ponazorimo uporabo logistične analize na primeru. Kot vse multivariantne metode, tudi logistična analiza pokaže pravo vrednost šele pri kompleksnejših problemih. Vendar nam bo za vpeljavo v to metodo bolj prav prišel preprost primer.

Za primer 1 bomo uporabili nekaj podatkov iz raziskave Slovensko javno mnenje 1988. Zanima nas, kako nekatere socio-demografske značilnosti respondenta vplivajo na njegov odgovor, da je za obrambo domovine pripravljen žrtvovati tudi življenje. Neodvisni spremenljivki sta respondentov spol in kmečko oz. nekmečko poreklo, ki se kaže v odgovoru, da del ali vsi dohodki respondentove družine izhajajo iz kmetijske dejavnosti. Izhodiščni podatki za analizo so prikazani v tabeli 1.

Tabela 1: Izhodiščni podatki za primer 1

spol	kmet	žrtvoval bi življenje	ne bi žrtvoval življenja
M	DA	80	188
M	NE	275	435
Ž	DA	48	168
Ž	NE	199	682

Bivariantna hi-kvadrat testa sta pokazala, da hipoteza o neodvisnosti odvisne in neodvisnih spremenljivk drži v primeru porekla in ne drži v primeru spola. Zaradi statistično pomembne interkorelacije prediktorjev in njunega morebitnega interakcijskega delovanja določimo izoliran vpliv vsakega posebej in skupni vpliv s pomočjo logistične analize. Zaenkrat predpostavimo samo model glavnih vplivov. Rezultati analize so naslednji:

	Ocena parametrov modela	z
konstanta	0.38	-16.2
spol/moški	1.41	7.0
kmet/da	0.88	-2.1

Hi-kvadrat = 2.54
SS = 1
p = .12
Koeficient entropije = .02
Koeficient koncentracije = .03

Logistična analiza je pokazala, da je tudi vpliv porekla na pripravljenost žrtvovati življenje za obrambo domovine statistično pomemben ($z > 1.96$). Pri bivariantni analizi se ta vpliv ni pokazal, ker je v kategoriji s kmečkim poreklom manj žensk. Interakcijski vpliv spola in porekla na pripravljenost žrtvovati življenje za obrambo domovine statistično ni pomembno, saj že sam model glavnih učinkov v zadostni meri pojasnjuje razpršitev odvisne spremenljivke v tabeli (hi-kvadrat test statistično ni pomemben). Neodvisni spremenljivki skupaj pojasnjujeta 2-3% razpršitve odvisne.

Verjetnost, da bo posameznik moškega spola s kmečkim poreklom pripravljen žrtvovati življenje za obrambo domovine, dobimo z enačbo 0.38 * 1.41 * 0.88. Ta verjetnost je torej 0.47 : 1 ali 32%. Verjetnost, da bo to pripravljen storiti posameznik moškega spola nekmečkega porekla, dobimo z enačbo 0.38 * 1.41 * (1/0.88). Ta verjetnost je torej 0.61 : 1 ali 38%.

Interkorelacija prediktorjev in interakcijsko učinkovanje prediktorjev

Kadar proučujemo učinek več neodvisnih spremenljivk na odvisno, se bivariantne metode slabo obnesejo, saj učinka nove neodvisne spremenljivke ne moremo preprosto prišteti učinku prejšnje oz. prejšnjih. Če dva prediktorja med seboj korelirata, je njun skupni vpliv na odvisno spremenljivko ponavadi manjši (lahko je tudi večji) od seštevka posamezno ugotovljenih vplivov. Po drugi strani lahko dva prediktorja v kombinaciji delujeta drugače kot vsak posebej, se pravi, na odvisno spremenljivko vplivata tudi interakcijsko. Čeprav je povedano raziskovalcem dobro znano, pa raziskovalec, ki razpolaga samo z bivariantnimi metodami, omenjene fenomene računsko in kognitivno težko obvlada. Zato raziskovalci nanje večinoma raje pozabijo. Iz izkušenj vemo, da se večina takšnih ali drugačnih raziskav, ki proučujejo vpliv različnih dejavnikov na nek pojav, ustavi pri naštevanju vplivov posameznih dejavnikov oziroma statistično

pomembnih korelacijskih ali kontingenčnih zvez. Tako ravnanje je skoraj pravilo, če so podatki v raziskavi merjeni na nominalnem nivoju. Logistična analiza nam pomaga te fenomene analizirati. Poglejmo si hipotetičen primer (primer 2).

Denimo, da je določen študij v predpisanem času končala polovica študentov. Zanima nas, kako na njihovo uspešnost vplivata spol in vrsta srednje šole, ki so jo končali. Primerjajmo hipotetične rezultate A, B in C iz tabele 2. Če te podatke analiziramo s pomočjo dvo-dimenzionalnih kontingenčnih tabel (kar bi najverjetneje storil hipotetični raziskovalec), bi v vseh treh primerih dobili enake rezultate, ki so prikazani v tabeli 3.

Tabela 2: Izhodiščni podatki za analizo vpliva spola in srednje šole na študijsko uspešnost (primer 2).

PRIMER		A		B		C	
USPEH		ne	da	ne	da	ne	da
spol	šola						
Ž	družbos.	35	65	60	90	40	60
Ž	naravos.	50	50	25	25	45	55
M	družbos.	50	50	25	25	45	55
M	naravos.	65	35	90	60	70	30

Tabela 3: Rezultati bivariantne analize vpliva spola in srednje šole na študijsko uspešnost (primer 2).

PRIMER A, B IN C

		USPEH	
		ne	da
SPOL	Ž	85	115
	M	115	85

Hi-kvadrat = 9
SS = 1
p = 0.003

		USPEH	
		ne	da
SRED. ŠOLA	družb.	85	115
	narav.	115	85

Hi-kvadrat = 9
SS = 1
p = 0.003

Rezultati logistične analize, prikazani v tabeli 4, pa so v primerih A, B in C zelo različni. Glavni učinki v primeru B so šibkejši kot v primeru A, saj prediktorja statistično pomembno korelirata. V primeru C pa je za razliko od primerov A in B pomemben tudi interakcijski vpliv obeh prediktorjev.

Tabela 4: Rezultati logistične analize vpliva spola in srednje šole na študijsko uspešnost (primer 2).

UČINEK	PRIMER		
	A	B	C
konstanta	1.00	1.00	1.01
spol	1.35	1.22	1.36
s. šola	1.35	1.22	1.35
spol * s. šola	1.00	1.00	0.81
koeficient koncentracije	.044	.030	0.054

Logistična regresija

Logistična regresija se uporablja za proučevanje vpliva intervalnih neodvisnih spremenljivk na diskretno, ponavadi dihotomno odvisno spremenljivko. Čeprav včasih naletimo na poročila o raziskavah, kjer avtorji v takem primeru uporabljajo kar običajno multiplo regresijo (metoda najmanjših kvadratov), ta postopek ni ustrezen iz več razlogov:

1. Ocenjevanje parametrov regresijske enačbe po metodi najmanjših kvadratov predpostavlja normalno distribucijo rezidualov. Ker ima distribucija rezidualov v primeru dihotomne odvisne spremenljivke ponavadi U obliko, so ocene parametrov enačbe izkrivljene, ravno tako koeficient multiple korelacije.

2. Kršena je tudi predpostavka o homogenosti varianc odvisne spremenljivke pri različnih kombinacijah vrednosti neodvisnih. Pri dihotomnih spremenljivkah je varianca neposredno odvisna od srednje vrednosti ($sigma^2 = p(1-p)$). Heterogenost varianc izkrivi ocene parametrov.

3. Problematična je interpretacija rezultatov (enačbe). Če odvisna spremenljivka lahko zavzame samo vrednosti 0 in 1 in če regresijska enačba napove rezultat posameznika nekje med tema vrednostima, je težko reči, kaj to pomeni, še posebej, če imamo v mislih prej omenjeno povezavo med proporcem in varianco. Še bolj je interpretacija problematična, če je napovedani rezultat večji od 1 ali manjši od 0.

Oglejmo si na primeru, kako je mogoče omenjene probleme rešiti. Denimo, da nas zanima, kako starost (x1) in število ur vožnje (x2) vplivajo na uspeh oz. neuspeh na vozniškem izpitu. Kandidat bo opravil izpit, če bo njegova vozniška spretnost presegla določen nivo zi. Če bo y' iz enačbe (2)

$$y_i' = b_0 + b_1x_{1i} + b_2x_{2i} (+ ...) + e \text{ oz. } y_i' = \mathbf{bx_i} + e \qquad (2)$$

večji od zi, bo izid uspeh (a_1), drugače pa neuspeh (a_0). Da bi rešili prej naštete težave (vrednost izida omejena na a_0 in a_1, heterogenost varianc, distribucija rezidualov), je potrebno ustrezno transformirati izraz bxi in ga probabilistično povezati z verjetnostjo odgovorov a_0 in a_1:

$$P(y_i = a_1) = P(z_i < \mathbf{bx_i}) = F(\mathbf{bx_i}) \qquad (3)$$

Verjetnost, da bo posameznik opravil izpit, je torej enaka verjetnosti, da je njegova vozniška spretnost presegla določen (naključen) prag zi. Ta verjetnost pa je odvisna od

njegove starosti in absolviranih ur vožnje ter regresijskih koeficientov, ki izražajo vpliv teh dejavnikov na pridobivanje spretnosti. Transformacijska funkcija F je lahko funkcija normalne distribucije, ki za določeno absciso vrne ustrezno površino (verjetnost). V tem primeru imamo opraviti z *probit* regresijo. Manj računskih problemov, a zelo podobne rezultate daje logistična transformacija $F(t) = (1 + e^{-t})^{-1}$. V tem primeru imamo opraviti z *logit* regresijo. Probabilistična interpretacija napovedanega rezultata yi' kot $P(y_i = a_1)$ in $P(y_i = a_0)$ je vsekakor edina smiselna, če imamo opraviti s kategorialno odvisno spremenljivko.

S pomočjo logistične analize lahko tudi simultano ugotavljamo vpliv intervalnih in nominalnih neodvisnih spremenljivk na dihotomno odvisno, vendar moramo v tem primeru kontinuirane spremenljivke spremeniti v diskretne, se pravi, moramo jih razvrstiti v razrede. Oglejmo si tak primer.

Ugotoviti skušamo, kateri dejavniki vplivajo na mnenje, ali naj bo obrambni minister civilist ali general. Vprašanje je bilo zastavljeno v raziskavi Slovensko javno mnenje 1988. Bivariantne analize so pokazale, da je ta odgovor v statistično pomembni zvezi z vrsto stališč in ocen o armadi, pa tudi z nekaterimi socio-demografskimi spremenljivkami. Vendar je multivariantna logistična analiza pokazala, da nam kot statistično pomembni prediktorji ostanejo samo tri spremenljivke. Drugi prediktorji so zaradi visokih interkorelacij izpadli iz modela. Tri relevantne spremenljivke so bile:

1. Pripravljenost žrtvovati svoje življenje za obrambo domovine.
2. Ocena, da imamo glede na mednarodne razmere preveč vojaštva.
3. Sumarna ocena o pripravljenosti in usposobljenosti armade in o odnosih v njej. Ta je vsebovala ocene o vojakih, starešinah, organizaciji, odnosih, opremljenosti, zavzetosti itd. Sumarna ocena je bila standardizirana na 7-stopenjski lestvici, kjer 1 pomeni zelo slabo in 7 odlično (intervalna spremenljivka).

Rezultate logistična analize kaže tabela 5. Približno 2/3 anketirancev je bilo mnenja, naj bo minister civilist. Na to mnenje predvsem vpliva respondentova ocena usposobljenosti armade: bolje, kot jo respondent ocenjuje, večja je verjetnost, da bo izbral ministra - generala. Tudi pripravljenost žrtvovati življenje za obrambo domovine vpliva na preferenco ministra - generala. Po drugi strani pa mnenje, da imamo preveč vojaštva, vpliva na preferenco ministra - civilista.

Tabela 5: Rezultati logistične analize vplivov na mnenje, ali naj bo minister za obrambo civilist ali general.

prediktor		minister naj bo general		civilist
konstanta		0.15	:	1
žrtvoval življenje	da	1.31	:	1
	ne	0.76	:	1
imamo preveč vojaštva	da	0.68	:	1
	neodločen	1.14	:	1
	ne	1.29	:	1
ocena armade	(n=1 slabo n=7 dobro)	$e^{n*.24}$	:	1

VIRI

Bock, R.D., "Multivariate statistical methods in behavioral research", New York: McGraw-Hill, 1975.
Haberman, S.J., "Analysis of qualitative data", New York: Academic press, 1979.
Toš, Niko, Klinar, Peter, Markič, Boštjan, Mlinar, Zdravko, "Slovensko javno mnenje 88", Ljubljana: RI FSPN, 1988.
Vodopivec, Blaž, "Epistemološki vidiki analize kovariančnih struktur", *Anthropos*, str. 4-6, 1988.