Primer primerjave analize kontingenčnih tabel z modeli regresijske analize nominalnih spremenljivk

Cveto Trampuž

An Illustrative Comparison Logit Analysis with Dummy Variable Regression Analysis. Two different regression models in which the dependent variable and all the explanatory variables are nominal are presented. The comparison is made at on elementary technical level on the three variables (age, employment, opinion on possible shortening of weekly working hours) from the Public Opinion Survey 89' in Slovenia.

1. Uvod

1.1 Predstavitev podatkov v obliki večrazsežnostnih frekvenčnih tabel in pojem modela analize.

V empiričnih raziskavah, predvsem v družboslovju, moramo velikokrat analizirati odnose med lastnostmi opazovanih enot, ki so opisane z nominalnimi spremenljivkami. Pri tem nastane vprašanje, kakšne in kako zanesljive informacije nam podatki nudijo, ter kako jih lahko dobimo, predstavimo in ocenimo.

Prvi korak je prav gotovo izračun eno in večrazsežnostnih frekvenčnih porazdelitev (v nadaljnjem besedilu jih bomo imenovali kar tabele) za tiste spremenljivke, ki nas zanimajo. Če ne znamo nič drugega, so tako dobljene frekvence, skupaj z vsemi možnimi strukturnimi odstotki, edine informacije za zaključke o odnosih med spremenljivkami oziroma o odnosih med enotami. V tabelah so ohranjene vse informacije, ki jih podatki vsebujejo.

V naslednjem koraku nas zanima, ali je možno objektivneje oceniti zanesljivost rezultatov analiz in s tem naših odločitev. To vprašanje je še posebej pomembno zato, ker so v veliki večini raziskav opazovane enote le vzorec, vzet iz neke populacije. Radi bi vedeli, s kakšno zanesljivostjo lahko sklepamo iz vzorčnih podatkov na razmere v celotni populaciji.

Problem rešujemo tako, da na osnovi poznavanja problematike, ki jo raziskujemo, predpostavimo določene odnose med spremenljivkami in preizkusimo, če dani podatki naše predpostavke dovolj dobro potrjujejo. Zamisel mora biti torej taka, da dopušča tudi možnosti ocenjevanja in posploševanja. Pravimo, da si zamislimo model odnosov med spremenljivkami oziroma enotami. Model mora biti tak, da dopušča ocenitev tako količine informacije, ki je prenesena v model (kadar podatke prikazujemo v bolj strnjeni obliki,

kot so [illegible], lahko izgubimo del informacije, ki jo sicer vsebujejo), kot tudi parametre modela, ki nam v njegovem okviru ponazarjajo predpostavljene odnose.

1.2 Opis izbranih spremenljivk.

Splošna predstavitev analize nominalnih spremenljivk z različnimi regresijskimi modeli bi bila v okviru tega prispevka preobsežna, zato prikazujemo primerjavo le pri analizi odnosov med tremi konkretnimi spremenljivkami z loglinearno analizo in regresijsko analizo nominalnih spremenljivk. Pri tem nam ne gre toliko za vsebinski pomen dobljenih rezultatov, kot za pot, po kateri ugotavljamo medsebojne odnose med nominalnimi spremenljivkami z izbranimi postopki. V bolj obsežnih primerih, ko analiziramo več spremenljivk z več vrednostmi, so postopki podobni, kot bodo opisani v naslednjih poglavjih.

Za prikaz in tudi za razlago primerjav smo izbrali naslednje tri spremenljivke:

Z zaposlenost — vrednosti: (1=da, 0=ne).

S starost — vrednosti: (1=do 30, 2=31-50, 3=nad 50)

Pri starosti smo tvorili še naslednje dihotomne spremenljivke:

$S1$ ima vrednost 1, če je $S = 1$ in 0 v ostalih primerih

$S2$ ima vrednost 1, če je $S = 2$ in 0 v ostalih primerih

$S3$ ima vrednost 1, če je $S = 3$ in 0 v ostalih primerih.

T ali naj se zmanjša število delovnih ur v tednu — vrednosti: (1=ne, 0=ostalo: da+ne vem).

Naš problem je, da ugotovimo ali je mišljenje, da delovnega tedna ni potrebno skrajšati, različno glede na starost in zaposlitev in glede na interakcijo med zaposlitvijo in starostjo (vpliv interakcije spremenljivk S in Z na spremenljivko T si predstavljamo tako, da $S(Z)$ spremeni svoj vpliv na T pri različnih vrednostih spremenljivke $Z(S)$).

2. Nekatere metode za analizo odnosov med nominalnimi spremenljivkami.

Za analizo nominalnih spremenljivkami imamo na razpolago veliko metod. Med najpogosteje uporabljenimi so tabele in statistika χ^2. Če je spremenljivk več in če imajo več vrednosti, postane tak način analize nepregleden in naporen. Kadar odnosi niso enostavni, se namreč težko odločamo, kako potekajo spremembe frekvenc 'po eni spremenljivki vzdolž vrednosti ostalih spremenljivk' in kakšne zaključke lahko naredimo o povezanosti med temi spremenljivkami. Zato je smiselno izbrati še postopke, ki omogočajo splošnejši in celovitejši pregled informacij, ki so v podatkih na razpolago. Splošne ugotovitve lahko dopolnimo še s podrobnostmi, ki jih ugotavljamo z bolj enostavnimi postopki. Pri tem pa naše ugotovitve ne smejo biti različne na osnovi različnih metod analize. Če bi se to zgodilo, bi pomenilo, da metode niso enakopravne za našo raziskavo, da niso enako občutljive glede na dane podatke ali pa, da so naše ugotovitve na meji zanesljivosti, kakršnokoli smo si že zastavili.

V našem zgledu smo preizkusili tri možnosti.

2.1 Večrazsežnostna tabela, prikazana v dveh razsežnostih.

V celu tabele je spremenljivka T, v glavi pa spremenljivka $ZS1S2$, ki predstavlja vse možne kombinacije vrednosti spremenljivk Z, $S1$ in $S2$.

	$ZS1S2$						
T	000	001	010	100	101	110	
0	f_{10}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	$f_{1.}$
1	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	$f_{2.}$
	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	$f_{.5}$	$f_{.6}$	$f_{..}$

Tabela 1:

V zgornji tabeli so podani osnovni podatki za raziskovanje medsebojnih odnosov med navedenimi spremenljivkami s pomočjo nelinearnih regresijskih modelov. Vendar bi radi še preizkusili, če je možno s standardno analizo take tabele dobiti smiselno podobne ugotovitve, kot pri regresijski analizi.

Zanima nas, če sta spremenljivki T in $ZS1S2$ povezani. Nalogo lahko formalno rešujemo takole:

Postavimo ničelno hipotezo H_0: *spremenljivki nista povezani* in alternativno hipotezo H_1: *spremenljivki sta povezani.* Hipotezo H_0 bomo testirali s pomočjo testa χ^2. V primeru, da bomo ničelno hipotezo zavrnili in sprejeli z določenim tveganjem alternativno hipotezo, bomo morali ugotovljeno povezanost še razložiti. Pri tem si bomo pomagali s takoimenovanimi standardiziranimi (r'_{ij}) in popravljenimi reziduali (r''_{ij}), ki so porazdeljeni pri velikem številu enot približno po standardizirani normalni porazdelitvi [Hab78]. Da pa lahko test χ^2 uporabimo, morajo biti teoretične frekvence (F_{ij}) v vseh celicah tabele (1) vsaj 5. Poleg tega je še odprto vprašanje, če je uporaba statistike χ^2 v vseh primerih takih tabel ustrezna.

2.2 Regresijski model, ko so odvisna in vse neodvisne spremenljivke dihotomne.

Model ima naslednjo obliko:

$$T = b_0 - b_1 Z - b_2 S1 - b_3 S2 - b_4 ZS1 - b_5 ZS2 - \varepsilon. \qquad (1)$$

Pomen oznak:

$ZS1$ pomeni interakcijo med Z in $S1$ (produkt Z s $S1$),
$ZS2$ pomeni interakcijo med Z in $S2$ (produkt Z s $S2$),
T je odvisna spremenljivka;
Z, $S1$, $S2$, $ZS1$, $ZS2$ so neodvisne spremenljivke;
b_0 je regresijska konstanta;
b_i, $i = 1,2,3,4,5$ so parametri oziroma koeficienti regresijskega modela
ε je člen napake.

Pomen koeficientov b_p $(p = 0,...,5)$ si lahko razlagamo tudi takole: Vrednost spremenljivke T, pri danih vrednostih neodvisnih spremenljivk, pomeni delež enot, ki pripadajo razredu, ki ga določajo vrednosti neodvisnih spremenljivk in vrednost 1 pri T. Člen napake mora zaradi statističnega ocenjevanja parametrov modela zadoščati posebnim pogojem [Gur86]. Zanima nas, od katerih neodvisnih spremenljivk je T odvisna in kako lahko z danimi vrednostmi neodvisnih spremenljivk ocenimo vrednost spremenljivke T.

Formalno bomo postopali takole:

Najprej postavimo ničelno hipotezo H_0 za celoten model. H_0: $b_j = 0$ za vsak j $(j = 1,...,5)$, ali še enostavnejše H_0: $R = 0$, kjer je R multipli korelacijski koeficient. Če bomo lahko na osnovi testa F ničelno hipotezo zavrnili, bomo lahko z določenim tveganjem sprejeli alternativno hipotezo H_1, namreč, da je odvisna spremenljivka odvisna vsaj od ene izmed neodvisnih spremenljivk, oziroma, da je vsaj en b_j različen od nič ali pa, da je $R > 0$.

Poleg tega nas seveda zanima, kako je odvisna spremenljivka odvisna od posameznih neodvisnih spremenljivk. Zato postavimo še nadaljnje hipoteze za vsako posamezno neodvisno spremenljivko: H_{0j}: $b_j = 0$ in H_1: $b_j \neq 0$ ali $b_j > 0$ ali $b_j < 0$. H_0 testiramo s pomočjo testa F, H_{0j} pa s pomočjo testov T ali F.

V modelu (1) lahko obravnavamo interakcije tako, da imajo spremenljivke $ZS1$ in $ZS2$ vrednost 1, če sta vrednosti Z in $S1(S2)$ enaki, in vrednost 0 v ostalih primerih:

$$\begin{aligned} ZS1 &= (Z)(S1) + (1 - Z)(1 - S1) \\ ZS2 &= (Z)(S2) + (1 - Z)(1 - S2) \end{aligned} \tag{2}$$

Koeficienti b_p $(p = 0,...,5)$ imajo lahko tudi v tem primeru že opisani pomen.

Enake rezultate (razen b_0), kot v zgornjem primeru dobimo, če spremenimo vrednost 0 pri vseh spremenljivkah v vrednost -1. Ocene vrednosti spremenljivke T iz modela (1) pa v tem primeru pomenijo razlike deležev enot med razredoma, ki ju določajo iste vrednosti neodvisnih spremenljivk in vrednosti 1 oziroma -1 spremenljivke T.

2.3 Loglinearni modeli.

Z loglinearnimi modeli analiziramo pri večrazsežnostnih tabelah frekvence in njihova razmerja kot funkcije parametrov, ki na osnovi modela pojasnjujejo karakteristike spremenljivk in njihove medsebojne odnose. Pa tudi obratno je res. Parametri so funkcije teoretičnih frekvenc oziroma njihovih razmerij. Ocenimo jih iz dejanskih (empiričnih) frekvenc. Predpostavimo, da v tabeli nobena empirična frekvenca ni enaka 0. čeprav obstojajo posebne tehnike v okviru loglinearnih modelov, ki bolj ali manj dobro rešujejo tudi ta problem. Model izberemo podobno, kot pri regresijski analizi. Tudi loglinearni model predstavlja naše teoretične predpostavke o odnosih med spremenljivkami. Na osnovi modela izračunamo poleg parametrov lambda (λ) še teoretične frekvence. Če se teoretične frekvence dovolj dobro ujemajo z empiričnimi, lahko sklepamo, da so bile naše predpostavke o povezanosti spremenljivk z določenim tveganjem pravilne. Oceniti skušamo tudi zanesljivost informacij, ki nam jih posredujejo parametri. Opišimo celoten postopek nekoliko bolj podrobno.

Najprej si oglejmo tako imenovani 'hierarhični saturirani' loglinearni model:

$$log(F_{ijk}) = \lambda_0 + \lambda_i^T + \lambda_j^Z + \lambda_k^S + \lambda_{ij}^{TZ} + \lambda_{ik}^{TS} + \lambda_{jk}^{ZS} + \lambda_{ijk}^{TZS} \quad (3)$$

in še tako imenovani 'saturirani' logit model:

$$\frac{1}{2} log(\frac{F_{ijk}}{F_{ijk}}) = \lambda + \lambda_j^Z + \lambda_k^S + \lambda_{jk}^{ZS} \quad (4)$$

Pri obeh modelih imajo indeksi naslednje vrednosti: $i = 0, 1$; $j = 0, 1$; $k = 1, 2, 3$.

Pomen oznak:

F_{ijk} je teoretična frekvenca, izračunana pri pogoju, da model velja.

λ_0 je logaritem geometričnega povprečja dejanskih frekvenc $\{f_{ijk}\}$ vseh celic tabele. Razlagamo si ga podobno, kot konstanto pri regresijski analizi. Parametri λ predstavljajo vplive posameznih spremenljivk ali njihovih medsebojnih interakcij na teoretične frekvence v posameznih celicah tabele.

Razmišljamo podobno, kot pri regresijski analizi. Če je λ enak nič, vpliva, ki ga parameter predstavlja, ni. Če je dovolj različen od 0, pa poenostavljeno pomeni, da je v celici empirična frekvenca prevelika (pri pozitivnih λ) ali premajhna (pri negativnih λ), da bi lahko trdili, da vpliva ni. Pri vplivih interakcij dveh ali večih spremenljivk moramo ugotoviti po lastni presoji, katera izmed njih je za to odgovorna. Pri tem moramo seveda upoštevati, da se vsote parametrov λ po ustreznih indeksih enake 0 [Hab78].

λ_i^T predstavlja vpliv spremenljivke T.

λ_j^Z predstavlja vpliv spremenljivke Z.

λ_k^S predstavlja vpliv spremenljivke S.

λ_{ij}^{TZ} predstavlja vpliv interakcije spremenljivk T in Z.

λ_{ik}^{TS} predstavlja vpliv interakcije spremenljivk T in S.

λ_{jk}^{ZS} predstavlja vpliv interakcije spremenljivk Z in S.

λ_{ijk}^{TZS} predstavlja vpliv interakcije spremenljivk T, Z in S.

λ v modelu logit razlagamo podobno, kot regresijsko konstanto.

Statistične ocene parametrov λ in rezidualov (razlik med teoretičnimi in empiričnimi frekvencami).

Predpostavimo, da nas zanimajo od 0 različni λ. Postavimo ničelno hipotezo H_0: $\lambda = 0$ in alternativno H_1: $\lambda \neq 0$. Kako veliko absolutno vrednost mora imeti λ, da ničelno hipotezo lahko zavrnemo? Vsaj približen odgovor na zastavljeno vprašanje dobimo, če izračunamo njegovo standardizirano vrednost. Označimo jo z λ'. λ' je pri velikih vzorcih porazdeljen približno po standardizirani normalni porazdelitvi. Če je njegova absolutna vrednost večja od 2.0, lahko s tveganjem, manjšim od 0.05, H_0 zavrnemo. Seveda pa ta test ni popolnoma zanesljiv, saj v splošnem ne vemo, kdaj je vzorec dovolj velik.

Druga možnost je, da izračunamo reziduale $r_{ijk} = f_{ijk} - F_{ijk}$. Za hitro oceno rezidualov izračunamo njihove standardizirane (r'_{ijk}) in popravljene vrednosti (r''_{ijk}). Oboji so porazdeljeni približno po standardizirani normalni porazdelitvi, zato jih znamo verjetnostno oceniti. Statistično ocenjevanje parametrov modela (3) s pomočjo testov Pearsonov χ^2 in L.R. (likelihood ratio) χ^2 je opisano v [Hab79].

3. Rezultati primerjav analiz.

Analiza odvisnosti spremenljivke T od S in Z s pomočjo tabele (1) je v našem primeru ekvivalentna analizi rezultatov z regresijsko analizo (1).

To je razvidno iz odnosov med frekvencami v tabeli (1) in koeficienti modela (1). Vpeljimo naslednje oznake:

$$T' = B_0 + B_1 Z + B_2 S1 + B_3 S2 + B_4 ZS1 + B_5 ZS2,$$

kjer so B_p $(p = 0, ..., 5)$ ocene za koeficiente b_p in $p_{i/j} = f_{ij}/f_{.j}$; $(i = 1, 2;\ j = 1, ..., 6)$.

Med koeficienti B in količinami p veljajo naslednje zveze:

$$\begin{aligned}
B_0 &= p_{2/1} & p_{2/1} &= B_0 \\
B_1 &= p_{2/4} - p_{2/1} & p_{2/4} &= B_0 + B_1 \\
B_2 &= p_{2/3} - p_{2/1} & p_{2/3} &= B_0 + B_2 \\
B_3 &= p_{2/2} - p_{2/1} & p_{2/2} &= B_0 + B_3 \\
B_4 &= p_{2/6} - p_{2/3} - p_{2/4} + p_{2/1} & p_{2/6} &= B_0 + B_1 + B_2 + B_4 \\
B_5 &= p_{2/5} - p_{2/2} - p_{2/4} + p_{2/1} & p_{2/5} &= B_0 + B_1 + B_3 + B_5
\end{aligned}$$

Gornje zveze je seveda možno posplošiti za poljubno število spremenljivk za tabele (1) in modele (1).

Zveze med ocenami regresijskih koeficientov B_p in $p_{i/j}$ je možno dobiti na preprost način tudi v primeru, ko so produkti $ZS1$ in $ZS2$ računani po formulah (2), oziroma, ko so vrednosti 0 vseh spremenljivk spremenjene v -1.

Tudi med statistiko χ^2, izračunano iz tabele (1), in multiplim korelacijskim koeficientom R iz (1) velja zveza $R^2 = \frac{\chi^2}{n}$ (n je število enot), kar ni težko dokazati.

S tem je dana zveza tudi med statistiko F pri regresijskih modelih (1) in χ^2 pri tabelah.

Potrebno pa bi bilo še ugotoviti, kakšni so odnosi med pogoji, ki jim morajo zadoščati podatki za uporabo testa χ^2 pri tabelah (1) in testov F pri regresijskih modelih (1). Vprašanje, ali so tabele in ustrezni regresijski modeli v splošnem ekvivalentni, ostaja še odprto.

Rezultati analiz odvisnosti spremenljivke T od spremenljivk Z in S s pomočjo večrazsežnostnih frekvenčnih porazdelitev in z regresijsko analizo nominalnih spremenljivk[1].

Opis spremenljivk za večrazsežnostno tabelo (2) je podan na strani 2. Spremenljivko $ZS1S2$ smo izračunali po formuli $ZS1S2 = 100\,Z + 10\,S1 + S2$. Vrednosti spremenljivke $ZS1S2$ bi lahko določili (šifrirali) tudi drugače, na primer s števili 1,2,...,6. Predlagano formulo smo izbrali le zato, da so vrednosti že kar kombinacije vrednosti spremenljivk Z in S.

Test χ^2 nam s tveganjem $sig\chi^2 = 0.0003$ omogoča sprejetje hipoteze, da sta spremenljivki T in $ZS1S2$ povezani. Reziduali (`Std Res` in `Adj Res`) nam kažejo, kako lahko

[1]Vsi rezultati na tej in naslednjih straneh so bili dobljeni s programi CROSSTABS, REGRESSION in HILOGLINEAR v okviru SPSSPC+ in LOGLINEAR v okviru SPSS-X.

Crosstabulation: T Zmanjsati st. del. ur na teden

By ZS1S2 Zaposlenost-Starost

Count / Exp Val / Row Pct / Col Pct / Tot Pct / Residual / Std Res / Adj Res	ZS1S2 S1=0 IZ=0 S2=0 000	S1=0 IZ=0 S2=1 001	S1=0 IZ=0 S2=0 010	S1=1 IZ=1 S2=0 100	S1=0 IZ=1 S2=1 101	S1=0 IZ=1 S2=0 110	Row Total
0	222	67	42	48	317	162	858
	189.7	53.4	39.5	58.0	342.8	174.5	42.1%
ostalo	25.9%	7.8%	4.9%	5.6%	36.9%	18.9%	
	49.2%	52.8%	44.7%	34.8%	38.9%	39.0%	
	10.9%	3.3%	2.1%	2.4%	15.5%	7.9%	
	32.3	13.6	2.5	-10.0	-25.8	-12.5	
	2.3	1.9	.4	-1.3	-1.4	-.9	
	3.5	2.5	.5	-1.8	-2.4	-1.4	
1	229	60	52	90	499	253	1182
	261.3	73.6	54.5	80.0	472.2	240.5	57.9%
ne	19.4%	5.1%	4.4%	7.6%	42.1%	21.4%	
	50.8%	47.2%	55.3%	65.2%	61.1%	61.0%	
	11.2%	2.9%	2.5%	4.4%	24.4%	12.4%	
	-32.3	-13.6	-2.5	10.0	25.8	12.5	
	-2.0	-1.6	-.3	1.1	1.2	.8	
	-3.5	-2.5	-.5	1.8	2.4	1.4	
Column Total	451 22.1%	127 6.2%	94 4.6%	138 6.8%	815 40.0%	415 20.3%	2040 100.0%

Chi-Square	D.F.	Significance	Min E.F.	Cells with E.F.< 5
23.62992	5	.0003	39.535	None

Tabela 2: Večrazsežnostna frekvenčna tabela.

Dependent Variable.. T Zmanjsati st. del. ur na teden

Multiple R	.10763	Analysis of Variance			
R Square	.01158		DF	Sum of Squares	Mean Square
Adjusted R Square	.00915	Regression	5	5.75846	1.15169
Standard Error	.49151	Residual	2034	491.37683	.24158

F = 4.76730 Signif F = .0002

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
Z	.14441	.04781	.13749	3.020	.0026
S2	-.03532	.04937	-.03567	-.715	.4745
ZS1	-.08797	.07375	-.07173	-1.193	.2331
S1	.04543	.05573	.03982	.815	.4150
ZS2	-.00581	.06697	-.00577	-.087	.9309
(Constant)	.50776	.02314		21.939	.0000

Tabela 3: Regresijski model

povezanost na hitro razložimo. Največje vrednosti so v stolpcih 000, 001 in 101. Tisti, ki niso zaposleni ($Z = 0$) in so v starostni skupini $S3$ ($S1 = 0$ in $S2 = 0$) in delno v $S2$ se manj odločajo, da delovni teden ne bi bil skrajšan. Obratno pa velja za zaposlene ($Z = 1$) predvsem v starostnem razredu $S2 = 1$. Tabelo lahko seveda še natančneje raziščemo po standardnih postopkih.

V tabeli (3) je prikazan le zadnji korak 'forward' analize, ko so v model vključene vse neodvisne spremenljivke.

Iz rezultatov regresijske analize lahko opazimo, da je T odvisna le od Z ($sigT = 0.0026$) in da konstanta, ki pomeni skupino $Z = 0$, $S = 3$ v veliki meri prispeva k večji vrednosti T, kar je v skladu z ugotovitvami pri analizi tabele.

3.1 Na osnovi rezultatov, dobljenih iz regresijske analize (1) in loglinearnih modelov (3) in (4), ugotavljamo v našem konkretnem primeru enake odnose med spremenljivkami T, Z in S.

Zveze med parametri λ pri loglinearnih modelih in parametri B pri regresijskem modelu niso linearne. Zato je njihov prikaz težji in ga tu le nakazujemo brez poglobljene analize.

Za vse ocene parametrov λ v loglinearnih modelih (3) in (4) veljata formuli (5) in (6)

(vrednosti spremenljivk T in Z smo zato, da bi se ujemali z indeksi v formulah, povecali za enoj):

$$\lambda^{TZS}_{i'j'k'} = \frac{1}{tzs}\sum_{i=1}^{t}\sum_{j=1}^{z}\sum_{k=1}^{s}(t\delta_{i'i}-1)(z\delta_{j'j}-1)(s\delta_{k'k}-1)\mu_{ijk} \qquad (5)$$

$$\lambda^{TZS}_{j'k'} = \frac{1}{tzs}\sum_{i=1}^{t}\sum_{j=1}^{z}\sum_{k=1}^{s}(z\delta_{j'j}-1)(s\delta_{k'k}-1)\frac{\mu_{ijk}}{\mu_{ijk}} \qquad (6)$$

Oznake imajo naslednji pomen:

$\lambda^{TZS}_{i'j'k'}$ je oznaka za parametre λ v modelih (3) in (4). Indekse pri λ dobimo tako, da upoštevamo vse smiselne kombinacije (glede na model) indeksov i', j' in k' prvega, drugega in tretjega reda. Ko računamo na primer, $\lambda^{T}_{i'}$ si mislimo, da indeksov j' in k' in konstant z in s v vsoti formule (5) ni.

$$\lambda^{T}_{i'} = \frac{1}{tzs}\sum_{i}\sum_{j}\sum_{k}(t\delta_{i'i}-1)\mu_{ijk}$$

Podobno reduciramo formulo (5) za računanje na primer

$$\lambda^{TS}_{j'k'} = \frac{1}{tzs}\sum_{i}\sum_{j}\sum_{k}(z\delta_{j'j}-1)(s\delta_{k'k}-1)\mu_{ijk}$$

μ_{ijk} so naravni logaritmi empiricnih frekvenc, $\mu_{ijk} = \ln(f_{ijk})$.

Indeksi imajo naslednje vrednosti:

i, i' = $1,\ldots,t$ (v našem primeru je $t = 2$);

j, j' = $1,\ldots,z$ (v našem primeru je $z = 2$);

k, k' = $1,\ldots,s$ (v našem primeru je $s = 2$).

$\delta_{qq'}$ so simboli, ki imajo vrednost 1, če sta indeksa q in q' enaka ($q = q'$) in vrednost 0, če velja $q \neq q'$

Rezultati analize odvisnosti spremenljivke T od spremenljivk Z in S s pomočjo loglinearnih modelov.

Hierarhični popolni (saturirani) loglinearni model.

Rezultati, ki jih dobimo s tem modelom, nam omogočajo ugotoviti, kateri λ v modelu (3) so dovolj različni od 0. Kaj pomenijo vsi rezultati, ki jih izpiše računalniški program (hiloglinear - backward elimination), je možno prebrati v SPSS-X Advanced Statistics Guide, itd. Izmed njih naj omenimo le naslednje ugotovitve:

Končni hierarhični model je na osnovi testa χ^2 določen s $T * Z$ in $Z * S$ in ne s $T * Z * S$ ter $T * S$. Če upoštevamo rezultate logit modela (4), kjer kombinacija $Z * S$ ne nastopa, $T * Z * S$ pa zaradi gornje ugotovitve ne pride v poštev, lahko ugotovimo, da je mnenje, da delovnega tedna ni potrebno skrajšati, odvisno predvsem od zaposlitve. Tisti, ki so zaposleni, razmišljajo, da delovnega tedna ni potrebno skrajšati, kar se ujema z ugotovitvami, ki smo jih dobili na osnovi rezultatov regresijske analize. Seveda pa obstoja velika povezanost med starostjo in zaposlitvijo, kar je trivialna ugotovitev.

Na tem mestu ponovno poudarjamo, da nam v tem članku ne gre za vsebinsko razmišljanje o odnosih med spremenljivkami T, Z in S (zvezo med njimi bi lahko raziskali še z drugimi metodami (ne da bi spremenljivke razvrstili v razrede), temveč le za preizkus možnosti primerjav analiz s pomočjo tabel (1), regresijske analize (1) in loglinearnih modelov (3) in (4).

V nadaljnjem bi bilo potrebno raziskati (a to presega okvir tega članka), kdaj se lahko ugotovitve o odnosih med spremenljivkami razlikujejo z uporabo loglinearnih in regresijskih modelov. [Goo76] navaja, da nastajajo razlike takrat, kadar so pogojne verjetnosti v (1) izven intervala 0.25 – 0.75. V našem primeru ta pogoj ni bil izpolnjen.

		STAROST											
		do 30 let			od 31 do 50 let			nad 50 let			SKUPAJ		
		ZAPOSLENI			ZAPOSLENI			ZAPOSLENI			ZAPOSLENI		
		da	ne	$\sum$	da	ne	$\sum$	da	ne	$\sum$	da	ne	$\sum$
	f	253	52	305	198	60	558	90	229	319	811	341	1182
ne	l	-0.057	0.057	0.026	0.025	-0.025	-0.052	0.032	-0.032	0.026	0.116	-0.166	0.11
	l'	-1.280	1.280	0.588	0.603	-0.603	-1.283				3.850	-3.850	6.23
da+	f	462	12	201	317	67	381	48	222	270	527	331	858
ne	l	0.057	-0.057	-0.026	-0.025	0.025	0.052	-0.032	0.032	-0.026	-0.116	0.116	-0.11
veto	l'	1.280	-1.280	-0.588	-0.603	0.603	1.283				-3.850	3.850	-1.59
	f	115	91	509	815	127	942	138	451	589	1368	672	2040
$\sum$	l	0.386	-0.386	-0.237	0.571	-0.571	0.251	-0.957	0.957	-0.011	0.343	-0.343	
	l'	8.628	-8.628	-5.298	11.01	-11.01	6.153	-22.61	22.61		11.11	-11.11	

Tabela 4: Rezultati analize modela (3).

Pomen oznak v tabeli (4):

f empirična frekvenca,

l parameter lambda pri modelu (3),

l' standardizirani lambda.

Ustrezni parametri l pri modelu (4) so enaki parametrom v modelu (3).

Literatura

[Agr84] Alan Agresti. *Analysis of Ordinal Categorical Data.* John Wiley & Sons, New York, 1984.

[BFH75] Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Massachusetts, London, 1975.

[Fie77] Stephen E. Fienberg. *The Analysis of Cross-Classified Categorical Data.* MIT Press, Massachusetts, London, 1977.

[Gil81] G. Nigel Gilbert. *Modelling Society. An introduction to Loglinear Analysis for Social Researches.* George Allen & Unwin, London, 1981.

[Goo76] Leo A. Goodman. The relationship between modified and usual multiple-regression approaches to the analysis of dichotomous variablesy. *Sociological Methodology*, 83-110, 1976.

[Goo78] Leo A. Goodman. *Analyzing Qualitative/Categorical Data. Log-linear Models and Latent Structure Analysis.* Abt Books, 1978.

[Gur86] Damodar Gurajati. *Basic Econometrics.* International student edition. McGraw-Hill Co, Singapore, 1986.

[Hab78] Shelby J. Haberman. *Analysis of Qualitative Data.* Volume 1, Academic press, New York, 1978.

[Hab79] Shelby J. Haberman. *Analysis of Qualitative Data.* Volume 2, Academic press, New York, 1979.

[KB80] David Knoke and J. Peter Burke. *Log-linear Models.* Sage University Paper 20, Sage Publications, London, 1980.

[Mom88] Konstantin Momirović. *Uvod u analizu nominalnih variabli.* Volume 2 of *Metodološke sveske*, FSPN, Ljubljana, 1988.

[Nor88] J. Marija Norusis. *SPSS-X Advanced Statistics Guide.* SPSS Inc.Chicago, Cambridge,Massachusetts, 1988.

[Upt70] Graham J. G. Upton. *The Analysis of Cress-tabulated Data.* John Wiley & Sons, New York, 1970.