Cveto Trampuž

PRIMERJAVA ANALIZE VEČRAZSEŽNIH TABEL Z RAZLIČNIMI MODELI REGRESIJSKE ANALIZE DIHOTOMNIH SPREMENLJIVK

POVZETEK. Namen tega dela je prikazati osnove razlik, ki lahko nastanejo pri interpretaciji rezultatov analize istih podatkov z logit modelom in klasičnim regresijskem modelom, kjer so vse, neodvisne in odvisna spremenljivka, dihotomne. Za maksimalne modele sta podani: (1) enostavna izpeljava za izračun parametrov in (2) ugotovitev, da je interpretacija rezultatov, dobljenih na osnovi navedenih dveh modelov lahko različna, če so razmerja (p_j/q_j) frekvenc vrednosti odvisne spremenljivke precej različne od 0.5.

ABSTRACT. This paper is concerned with the comparison of logit analysis with classical linear regression analysis in which the dependent variable and all the explanatory variables are dichotomous. The main aim of this work is to define (1) how the parameters for saturated models can be calculated and (2) to show that the two models can yield different interpretations of results based on the same data sets when the rations (p_j/q_j) of the dependent variable highly differ from 0.5

1. UVOD

Predpostavimo, da so za N enot zbrane vrednosti večih dihotomnih spremenljivk. Problem, ki ga želimo rešiti je naslednji: **Raziskati moramo, če se ugotovitve o odnosih med eno odvisno in večimi neodvisnimi spremenljivkami lahko razlikujejo z uporabo klasične regresijske analize oziroma logit analize.**
Primerjava je izvedena samo na osnovi podobnih obrazcev za izračun parametrov obeh modelov. Z upoštevanjem porazdelitev spremenljivk, je seveda možno dobiti natančnejše sklepe. Rezultati so podani na primeru štirih spremenljivk za maksimalne modele regresijske analize.

Označimo odvisno spremenljivko z Y in neodvisne spremenljivke z X1, X2 in X3. Vsaka od navedenih spremenljivk naj ima vrednosti 0 in 1. Sestavimo še tako imenovano skupno neodvisno spremenljivko Z, ki ima toliko vrednosti, kolikor je možnih kombinacij vrednosti vseh neodvisnih spremenljivk. V našem primeru ima spremenljivka Z 8 vrednosti (v splošnem primeru pa r vrednosti, kjer je $r = 2^m$, m = število neodvisnih spremenljivk). Označili jih bomo kar z Z_j, $(j=1,2,\ldots,r)$, pri čemer vrstni red v splošnem ni pomemben. Večrazsežnostno tabelo lahko prikažemo tako v dveh razsežnostih. V taki dvorazsežnostni tabeli so ohranjene vse informacije, ki jih imajo zbrani podatki. Frekvence so osnovni podatki za analizo odnosov med navedenimi štirimi spremenljivkami s pomočjo klasične regresijske analize in logit analize dihotomnih spremenljivk.
V tabeli (T1) bomo samo zaradi lažjega pisanja indeksov vpeljali za robne frekvence oznake: v_i in s_j, $i=0,1$ in $j=1,\ldots,8$.

Tabela T1. Prikaz štirirazsežne tabele v dveh razsežnostih.

				X1 X2 X3					
K-->	000	100	010	001	110	101	011	111	
Y	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	
0	n_{01}	n_{02}	n_{03}	n_{04}	n_{05}	n_{06}	n_{07}	n_{08}	v_0
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	n_{17}	n_{18}	v_1
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	N

Oznake imajo naslednji pomen:

n_{ij} je frekvenca celice ij, (i = 0,1; j = 1,2,...,r),

$$v_i = \sum_j n_{ij},$$

$$s_j = \sum_i n_{ij},$$

$$N = \sum_{ij}\sum n_{ij} = \sum_i v_i = \sum_j s_j.$$

Iz druge vrstice v glavi tabele (T1) je razvidno, kako so na osnovi kombinacij vrednosti spremenljivk X1, X2 in X3 tvorjene celice tabele . Tako pomeni na primer frekvenca n_{06} število enot, za katere velja: X1 = 1, X2 = 0, X3 = 1 in Y = 0. Pri n_{16} so vrednosti spremenljivk X1, X2 in X3 enake kot pri n_{06}, le Y = 1. S pomočjo indeksa j pri Z_j bomo tudi nadomestili in s tem poenostavili pisanje kombinacije indeksov spremenljivk X1, X2 in X3 v kasnejših izračunih oziroma formulah. Tako bo na primer indeks j = 6 pomenil indekse celic (stolpca) X1 = 1, X2 = 0 in X3 = 1 pri Y = 0 in pri Y = 1.

2. PREGLED TABELE (T1).

Raziščimo predpostavko, da je spremenljivka Y neodvisna od posameznih spremenljivk X1, X2 in X3 in od kakršne koli interakcije med njimi. Za rešitev te naloge bomo uporabili test x^2 (hi-kvadrat). Ker ima spremenljivka Y le dve vrednosti ima izraz za izračun statistike x^2 naslednjo obliko, ki jo navajamo brez izpeljave:

$$x^2 = \frac{1}{pq}\sum_j s_j(p_j - p)^2, \qquad (1)$$

kjer imajo oznake naslednji pomen:

$$p = \frac{v_1}{N} \qquad p_j = \frac{n_{1j}}{s_j}, \qquad q = \frac{v_0}{N} = 1 - p.$$

Če velja za vsak j

$$\frac{n_{1j}}{n_{0j}} \equiv \frac{p_j}{q_j} = \frac{v_1}{v_0} = k, \qquad (2)$$

kjer je $q_j = 1 - p_j$ in k poljubna pozitivna konstanta, je vrednost x^2 v izrazu (1) enaka nič, kar pomeni, da je

spremenljivka Y neodvisna od neodvisnih spremenljivk in njihovih interakcij. Če pa so posamezni členi vsote dovolj različni od nič, lahko razmišljamo, da je spremenljivka Y odvisna od neodvisnih spremenljivk oziroma njihovih medsebojnih interakcij. Pri tem nam lahko pomagajo ocene standardiziranih oziroma popravljenih razlik med empiričnimi in teoretičnimi frekvencami (Haberman).

3. KLASIČNI REGRESIJSKI MODEL.

Klasični maksimalen regresijski model s pomočjo katerega bomo analizirali odvisnost spremenljivke Y od spremenljivk X1, X2, X3 in vseh njihovih možnih produktov, ima naslednjo obliko:

$$Y = b_0 + b_1 X1 + b_2 X2 + b_3 X3 + b_4 X4 + b_5 X5 + b_6 X6 + b_7 X7 + e, \qquad (3)$$

kjer imajo oznake naslednji pomen:

Y, X1, X2, X3 so dane spremenljivke;
vrednosti spremenljivk X4, X5, X6 in X7 so njihove interakcije in jih dobimo tako, da pri vsaki enoti pomnožimo vrednosti spremenljivk X1, X2 in X3 po naslednjih formulah:

$$X4=X1*X2, \quad X5=X1*X3, \quad X6=X2*X3 \text{ in } \quad X7=X1*X2*X3 \qquad (4)$$

Vrednosti posameznih spremenljivk bomo označevali z malimi črkami in enim ali večimi indeksi. Tako bo na primer oznaka x_{ij} pomenila vrednost spremenljivke j pri enoti i. Če spremenljivka nima indeksa, pomeni indeks i njeno vrednost pri enoti z indeksom i (na primer: y_i).

b_j so regresijski koeficienti, (j=0,1,...,7),

e je člen napake.

Predpostavimo, da spremenljivke X_j niso med seboj linearno odvisne in prikažimo izračun ocen (označili jih bomo z B_j) regresijskih koeficientov b_j iz (3) po metodi najmanjših kvadratov v znani matrični obliki:

$$B = (X^T X)^{-1} X^T Y. \qquad (5)$$

Zgradbi matrike X in vektorja Y sta znani, zato ju ne opisujemo. V matriki $X^T X$, ki je razsežnosti rxr je samo 2r (= 16) elementov med seboj različnih. Enaki so ustreznim vsotam robnih frekvenc s_j v tabeli (T1). Zato je možno poiskati analitično njeno inverzno matriko in s tem eksplicitno zvezo med koeficienti B_j in frekvencami oziroma razmerji frekvenc in robnih frekvenc tabele (T1). Da bi to pokazali, vpeljimo najprej naslednje vektorje :

$$y1 = \begin{matrix} n_{11} \\ n_{12} \\ n_{13} \\ n_{14} \\ n_{15} \\ n_{16} \\ n_{17} \\ n_{18} \end{matrix} \quad s = \begin{matrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \end{matrix} \quad B = \begin{matrix} B_0 \\ B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \\ B_6 \\ B_7 \end{matrix} \quad \hat{e} = \begin{matrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \\ \hat{e}_4 \\ \hat{e}_5 \\ \hat{e}_6 \\ \hat{e}_7 \\ \hat{e}_8 \end{matrix} \quad p = \begin{matrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \end{matrix}$$

Elementi vektorjev y1 in s so podani v tabeli (T1).
Elementi vektorja B so izračunani po formuli (5).
Elemente vektorja $\hat{e}$ bomo pojasnili kasneje (str.+4).
Elementi vektorja p so podani v formuli (6).

Vpeljimo še naslednje matrike:

$$A = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{matrix} \qquad S = \begin{matrix} s_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & s_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & s_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & s_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & s_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & s_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & s_7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & s_8 \end{matrix}$$

$$W = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \end{matrix} \qquad V = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{matrix}$$

Matrika A ima r vrstic (spomnimo se, da je r tudi vsota binomskih koeficientov pri potenci m). Je spodnja trikotna in simetrična glede na 'nepravo diagonalo' ($a_{ij} = a_{r-j+1,r-i+1}$). Enice so v celotnem prvem stolpcu in celotni r-ti vrstici. Enice in ničle v stolpcih z indeksi 2, 3 in 4, lahko dobimo tudi tako, da prenesemo v njih kombinacije ničel in enic iz druge vrstice (označili smo jo z K ->) v tabeli (T1), zapovrstjo, v vrstice matrike A. Sicer pa postavljamo enice v matriko A na osnovi kombinacij k (k = 0,1,...,m) elementov izmed m elementov. S pomočjo binomskih koeficientov tudi določimo pozitivne in negativne enice v matriki V in diagonalne vrednosti v matriki W. Ker je postopek razviden že iz primera, ga ne bomo opisovali.

Iz danih matrik A in V in W izpeljemo nato vse ostale matrike in nekatere vektorje, ki jih potrebujemo, po naslednjih obrazcih:

$A^{-1} = VAV$ in ima za naš primer naslednjo obliko:

$$A^{-1} = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1. \end{matrix}$$

$$p = S^{-1}y1, \qquad (6)$$

$$X^TX = A^TSA,$$

$$(X^TX)^{-1} = A^{-1}S^{-1}(A^T)^{-1},$$

$$y1 = SAB,$$

$$X^TY = A^Ty1,$$

zveza med koeficienti B in deleži p je naslednja:

$$B = A^{-1}p = A^{-1}S^{-1}y1 \quad \text{in}$$

$$p = AB.$$

Za zgled navedimo vektorja B in p ter matriki X^TX in $(X^TX)^{-1}$ za naš konkreten primer štirih spremenljivk eksplicitno (podobno dobimo lahko eksplicitne izraze za kompleksnejše primere odnosov med večimi spremenljivkami):

Vektor B je izražen s frekvencami tabele (T1) takole:

$$B = \begin{matrix}
\frac{n11}{s1} \\
-\frac{n11}{s1}+\frac{n12}{s2} \\
-\frac{n11}{s1}+\frac{n13}{s3} \\
-\frac{n11}{s1}+\frac{n14}{s4} \\
\frac{n11}{s1}-\frac{n12}{s2}-\frac{n13}{s3}+\frac{n15}{s5} \\
\frac{n11}{s1}-\frac{n12}{s2}-\frac{n14}{s4}+\frac{n16}{s6} \\
\frac{n11}{s1}-\frac{n13}{s3}-\frac{n14}{s4}+\frac{n17}{s7} \\
-\frac{n11}{s1}+\frac{n12}{s2}+\frac{n13}{s3}+\frac{n14}{s4}-\frac{n15}{s5}-\frac{n16}{s6}-\frac{n17}{s7}+\frac{n18}{s8}.
\end{matrix}$$

Vektor p izražen s koeficienti B:

$$
p = \begin{bmatrix}
B_0 \\
B_0 + B_1 \\
B_0 + B_2 \\
B_0 + B_3 \\
B_0 + B_4 \\
B_0 + B_1 + B_2 + B_4 \\
B_0 + B_1 + B_3 + B_5 \\
B_0 + B_2 + B_3 + B_6 \\
B_0 + B_1 + B_2 + B_3 + B_4 + B_5 + B_6 + B_7
\end{bmatrix}
$$

Matrika X^TX izražena robnimi frekvencami s_j tabele (T1):

N	s2+s5+s6+s8	s3+s5+s7+s8	s4+s6+s7+s8	s5+s8	s6+s8	s7+s8	s8
s2+s5+s6+s8	s2+s5+s6+s8	s5+s8	s6+s8	s5+s8	s6+s8	s8	s8
s3+s5+s7+s8	s5+s8	s3+s5+s7+s8	s7+s8	s5+s8	s8	s7+s8	s8
s4+s6+s7+s8	s6+s8	s7+s8	s4+s6+s7+s8	s8	s6+s8	s7+s8	s8
s5+s8	s5+s8	s5+s8	s8	s5+s8	s8	s8	s8
s6+s8	s6+s8	s8	s6+s8	s8	s6+s8	s8	s8
s7+s8	s8	s7+s8	s7+s8	s8	s8	s7+s8	s8
s8	s8	s8	s8	s8	s8	s8	s8

Matrika $(X^TX)^{-1}$:

$\frac{1}{s_1}$	$-\frac{1}{s_1}$	$-\frac{1}{s_1}$	$-\frac{1}{s_1}$	$\frac{1}{s_1}$	$\frac{1}{s_1}$	$\frac{1}{s_1}$	$-\frac{1}{s_1}$
$-\frac{1}{s_1}$	$\frac{1}{s_1}+\frac{1}{s_2}$	$\frac{1}{s_1}$	$\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_2}$	$-\frac{1}{s_1}-\frac{1}{s_2}$	$-\frac{1}{s_1}$	$\frac{1}{s_1}+\frac{1}{s_2}$
$-\frac{1}{s_1}$	$\frac{1}{s_1}$	$\frac{1}{s_1}+\frac{1}{s_3}$	$\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_3}$	$-\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_3}$	$\frac{1}{s_1}+\frac{1}{s_3}$
$-\frac{1}{s_1}$	$\frac{1}{s_1}$	$\frac{1}{s_1}$	$\frac{1}{s_1}+\frac{1}{s_4}$	$-\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_4}$	$-\frac{1}{s_1}-\frac{1}{s_4}$	$\frac{1}{s_1}+\frac{1}{s_4}$
$\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_2}$	$-\frac{1}{s_1}-\frac{1}{s_3}$	$-\frac{1}{s_1}$	$\frac{1}{s_1}+\frac{1}{s_2}+\frac{1}{s_3}+\frac{1}{s_5}$	$\frac{1}{s_1}+\frac{1}{s_2}$	$\frac{1}{s_1}+\frac{1}{s_3}$	$\frac{1}{s_1}+\frac{1}{s_2}+\frac{1}{s_3}+\frac{1}{s_5}+\cdots$
$\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_2}$	$-\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_4}$	$\frac{1}{s_1}+\frac{1}{s_2}$	$\frac{1}{s_1}+\frac{1}{s_2}+\frac{1}{s_4}+\frac{1}{s_6}$	$\frac{1}{s_1}+\frac{1}{s_4}$	$\frac{1}{s_1}+\frac{1}{s_2}+\frac{1}{s_4}+\frac{1}{s_6}+\cdots$
$\frac{1}{s_1}$	$-\frac{1}{s_1}$	$-\frac{1}{s_1}-\frac{1}{s_3}$	$-\frac{1}{s_1}-\frac{1}{s_4}$	$\frac{1}{s_1}+\frac{1}{s_3}$	$\frac{1}{s_1}+\frac{1}{s_4}$	$\frac{1}{s_1}+\frac{1}{s_3}+\frac{1}{s_4}+\frac{1}{s_7}$	$\frac{1}{s_1}+\frac{1}{s_3}+\frac{1}{s_4}+\frac{1}{s_8}+\cdots$
$-\frac{1}{s_1}$	$\frac{1}{s_1}+\frac{1}{s_2}$	$\frac{1}{s_1}+\frac{1}{s_3}$	$\frac{1}{s_1}+\frac{1}{s_4}$	$-\frac{1}{s_1}-\frac{1}{s_2}-\frac{1}{s_3}-\frac{1}{s_5}$	$-\frac{1}{s_1}-\frac{1}{s_2}-\frac{1}{s_4}-\frac{1}{s_6}$	$-\frac{1}{s_1}-\frac{1}{s_3}-\frac{1}{s_4}-\frac{1}{s_7}$	$\sum_{j=1}^{8}\frac{1}{s_j}$

Elementi matrike $(X^TX)^{-1}$ so tako kot elementi matrike X^TX funkcije samo robnih frekvenc s_j tabele (T1).
Kvadratni koreni diagonalnih elementov matrike $(X^TX)^{-1}$ so del ocene za izračun standardne napake za parametre B, zato bomo navedli formulo za njihov izračun še posebej. Označili jih bomo z d_i in jih razvrstili v vektor d:

$$d = (As^{-1})^{\frac{1}{2}}.$$

Izračunajmo še vektor $\hat{e}$.

Ocene za člene napake e v (3) imajo le šestnajst različnih vrednosti:

$-p_j$, ki se ponavljajo n_{0j} krat in

$1 - p_j$, ki se ponavljajo n_{1j} krat.

Njihova vsota je enaka nič za vsak j posebej.
Zato lahko upeljemo vektor ê takole:

$$\hat{e}_j^2 = n_{0j} p_j^2 + n_{1j}(1-p_j)^2$$

$$= \frac{n_{0j} n_{1j}}{s_j} .$$

Ocena za standardno napako v regresijskem modelu je potem:

$$s_e^2 = \frac{1}{N-m-1} \sum_j \frac{n_{0j} n_{1j}}{s_j}$$

$$= \frac{N}{N-m-1} pq\left(1 - \frac{x^2}{N}\right),$$

kjer je x^2 izračunan po formuli (1).

Brez izpeljave navedimo še, da velja med statistiko x^2 in statistiko R^2 (R je multipli korelacijski koeficient) pri modelu (3) naslednja zveza:

$$R^2 = \frac{x^2}{N}.$$

* * *

Za primerjavo rezultatov, dobljenih z regresijskim modelom (3) in z logit modelom, vpeljimo naslednje transformacije vrednosti spremenljivk Y, X1, X2 in X3:

$$Y = \tfrac{1}{2}(Y'+1), \quad Xj = \tfrac{1}{2}(Xj'+1) \quad (j=1,2,3).$$

S tako transformiranimi vrednostmi spremenljivk izračunajmo po metodi najmanjših kvadratov koeficiente regresijskega modela:

$$Y' = B_0' + B_1' X1' + B_2' X2' + B_3' X3' + B_4' X4' + B_5' X5' + B_6' X6' + B_7' X7' + e' \qquad (7)$$

S pomočjo koeficientov B_j' ocenjujemo količine $2p_j-1$ oziroma p_j-q_j namesto p_j kot v modelu (3).

Ugotovimo, kakšne so zveze med koeficienti B_j' in B_j ter frekvencami v tabeli (T1).

Najpreje vpeljimo matriko G po formuli:

$$G = A^T W^{-1} A^{-1}. \tag{8}$$

Matrika G izgleda v našem primeru takole:

$$G = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 \\ -1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix}.$$

Med vektorji B' in B ter p veljajo naslednje zveze:

$$B' = 2Gp, \tag{9}$$

$$P = \tfrac{1}{2}G^{-1}B',$$

$$B' = 2A^T W^{-1} B = 2GAB \quad \text{in}$$

$$B = \tfrac{1}{2}WA^{-T}B'$$

Pri B_0' moramo še odšteti enico, zaradi transformacije Y v Y'.

Inverzno matriko matrike G lahko izračunamo po obrazcu:

$$G^{-1} = 2^m G^T.$$

Vsote stolpcev in vrstic matrike G so enake 0, razen prve vrstice in zadnjega stolpca, ko sta vsoti enaki 2^m.

4. LOGLINEARNI IN LOGIT MODELI

Če predpostavljamo, da so vse spremenljivke enakopravne (ne predpostavljamo, da je ena izmed njih odvisna in druge neodvisne) ima maksimalen loglinearni model, ki je osnova za možne hierarhične modele (v našem primeru smo dihotomne spremenljivke X1, X2 in X3 zaradi lažjega pisanja indeksov preimenovali zapovrstjo v A, B in C; indeksi i, j, k in l se nanašajo na spremenljivke A, B, C in Y in imajo vrednosti 0 in 1) naslednjo obliko:

$$\begin{aligned}\log(m_{ijkl}) &= L_0 \\ &+ L_i^A + L_j^B + L_k^C + L_l^Y \\ &+ L_{ij}^{AB} + L_{ik}^{AC} + L_{il}^{AY} + L_{jk}^{BC} + L_{jl}^{BY} + L_{kl}^{CY} \\ &+ L_{ijk}^{ABC} + L_{ijl}^{ABY} + L_{ikl}^{ACY} + L_{jkl}^{BCY} \\ &+ L_{ijkl}^{ABCY}.\end{aligned} \tag{10}$$

Pri navedenem modelu veljajo naslednje relacije:

$$\begin{aligned}&\sum_i L_i^A = \sum_j L_j^B = \sum_k L_k^C = \sum_l L_l^Y \\ &= \sum_i L_{ij}^{AB} = \sum_j L_{ij}^{AB} = \ldots \\ &= \sum_i L_{ijk}^{ABC} = \sum_j L_{ijk}^{ABC} = \sum_k L_{ijk}^{ABC} = \ldots \\ &= \sum_i L_{ijkl}^{ABCY} = \sum_j L_{ijkl}^{ABCY} = \sum_k L_{ijkl}^{ABCY} = \sum_l L_{ijkl}^{ABCY} = 0.\end{aligned} \tag{11}$$

Oznake imajo naslednji pomen:

m_{ijkl} so teoretične frekvence;

L_0 konstanta (logaritem geometrijske sredine vseh frekvenc);

$L_i^A, L_j^B, L_k^C, L_l^Y$ vpliv posameznih spremenljivk A, B, C in Y;

$L_{ij}^{AB}, L_{ik}^{AC}, L_{il}^{AY}, L_{jk}^{BC}, L_{jl}^{BY}, L_{kl}^{CY}$ vpliv interakcij 2. reda;

$L_{ijk}^{ABC}, L_{ijl}^{ABY}, L_{ikl}^{ACY}, L_{jkl}^{BCY}$ vpliv interakcij 3. reda;

L_{ijkl}^{ABCY} vpliv interakcij 4. reda.

Za naš model so ocene za teoretične frekvence podane z enačbo:

$$m_{ijkl} = n_{ijkl}, \quad (i,j,k,l = 0,1)$$

kjer so n_{ijkl} dejanske empirične frekvence štirirazsežne tabele. Ker nas na tem mestu izračuni parametrov L za model (10) ne zanimajo, jih ne bomo posebej navajali.

Če predpostavljamo, da je spremenljivka Y odvisna, lahko določimo logit model takole:

$$\frac{1}{2}\log\left(\frac{m_{ijkl}}{m_{ijk0}}\right) = L_l^Y + L_{il}^{AY} + L_{jl}^{BY} + L_{kl}^{CY} + L_{ijl}^{ABY} + L_{ikl}^{ACY} + L_{jkl}^{BCY} + L_{ijkl}^{ABCY} \qquad (12)$$

Pri navedenem modelu veljajo za parametre L podobne relacije kot pri (11) (Haberman).

Izračunajmo parametre L modela (12) z uporabo matrike G, ki smo jo vpeljali v formuli (8).
Zaradi vsot parametrov L, opisanih v (11) nam zadostuje, da izračunamo po en koeficient za vsako spremenljivko in po en za vsako možno interakcijo.

Najprej vpeljimo vektorje:

$$L = \begin{matrix} L_l^Y \\ L_{il}^{AY} \\ L_{jl}^{BY} \\ L_{kl}^{CY} \\ L_{ijl}^{ABY} \\ L_{ikl}^{ACY} \\ L_{jkl}^{BCY} \\ L_{ijkl}^{ABCY} \end{matrix} = \begin{matrix} L_0 \\ L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \\ L_6 \\ L_7 \end{matrix} \qquad \mu = \begin{matrix} \mu 1 \\ \mu 2 \\ \mu 3 \\ \mu 4 \\ \mu 5 \\ \mu 6 \\ \mu 7 \\ \mu 8, \end{matrix}$$

kjer je $\mu_j = \frac{1}{2}\log\left(\frac{n_{1j}}{n_{0j}}\right)$, n_{ij} so frekvence tabele (T1).

parametri L se nanašajo na indeks 1 tako pri spremenljivki Y, kot tudi pri vseh spremenljivkah A, B in C oziroma X1, X2 in X3 in njihovih medsebojnih interakcijah.

Vektor L je dan z izrazom:

$$L = G\mu.$$

Izraz je po obliki podoben izrazu B' = 2Gp izvedenem v (9), ki smo ga razvili pri klasičnem regresijskem modelu (7).

Matrika G je v obeh primerih ista!

Ena od začetnih možnosti za raziskovanje razlik pri interpretaciji rezultatov dobljenih (seveda z istimi podatki) po obeh modelih je, da ugotovimo analitično ujemanje parametrov L in B', ki imajo podoben pomen za praktično uporabo oziroma za interpretacijo.

Spomnimo se količin p_j, ki smo jih vpeljali v izrazih (6) in količine $q_j = 1 - p_j$.

$$p_j = \frac{n_{1j}}{s_j} \quad (j=1,2,\ldots,r).$$

Posamezen člen v vsoti pri izrazu L = Gμ je

$$\tfrac{1}{2}\log\left(\frac{p_j}{q_j}\right) = -\tfrac{1}{2}\log\left(\frac{q_j}{p_j}\right), \qquad (13)$$

pri vsotah pri B' = 2Gp v (9) pa je

$$2p_j - 1 = p_j - q_j. \qquad (14)$$

Poglejmo, kako se razlikujeta izraza (13) in (14) pri različnih vrednostih za p_j.

Vpeljimo v izraz (13) za p_j naslednjo transformacijo

$p_j = \frac{1}{2}(1 + t_j)$ in dobimo

$$\tfrac{1}{2}\log\left(\frac{1 + t_j}{1 - t_j}\right) \qquad (-1 < t_j < 1). \qquad (15)$$

Logaritem (15) razvijemo v Taylorjevo vrsto okoli točke $t_j = 0$ in dobimo:

$$\tfrac{1}{2}\log\left(\frac{1 + t_j}{1 - t_j}\right) = t_j + t_j^3/3 + t_j^5/5 + \ldots, \text{ kjer je}$$

$$2p_j - 1 = t_j.$$

Razlika (13) - (14) pri posameznem členu je torej

$$t_j^3/3 + t_j^5/5 + \ldots$$

Celotna razlika pa je GT, kjer je T vektor s komponentami:

$$t_j = t_j^3/3 + t_j^5/5 + \ldots$$

Za toliko so koeficienti L različni od koeficientov B'.

Oglejmo si tabelo razlik za različne vrednosti p:

p	$-\frac{1}{2}\log(\frac{1-p}{p}) - (2p-1)$
0.50	0.0000
0.55	0.0003
0.60	0.0027
0.65	0.0095
0.70	0.0236
0.75	0.0493
0.80	0.0931
0.85	0.1673
0.90	0.2986
0.95	0.5722
0.975	0.8818
0.995	1.6566
0.9995	2.8012

Vidimo, da se večje razlike začenjajo šele za p > 0.75.
Predpostavimo, da nas relativna napaka ne zanima, saj so odločilne za interpretacijo le absolutne velikosti koeficientov.

5. NUMERIČNI PRIMERI.

Za ilustracijo so prikazani trije izmišljeni primeri. Podane so tabele s statistiko x^2, koeficienti B_j' in njihova standardna napaka SE_B, ter koeficienti L_j s standardno napako SE_L. Pri maksimalnem modelu sta obe standardni napaki neodvisni od indeksa j Predznaki pri koeficientih L_j so urejeni tako, da se ujemajo s predznaki pri koeficientih B_j'.
Kriterij za ocenjevanje razlik med koeficienti v tem članku ni izpeljan. Vsaj za grobo oceno razlik pa bi bilo morda možno po

običajnih postopkih (ob predpostavki, do so B_j' in L_j normalno porazdeljeni) uporabiti standardne napake. V prikazanih zgledih bi razlike pri interpretaciji pomena koeficientov B'_j in L_j lahko nastale, s tveganjem približno 5%, če veljajo naslednje relacije:

$$|B_j'| > 1.96SE_{B_j'} \text{ in } |L_j| < 1.96SE_{L_j} \text{ ali}$$

$$|B_j'| < 1.96SE_{B_j'} \text{ in } |L_j| > 1.96SE_{L_j}.$$

Korektnost tega kriterija pa bi bilo potrebno še preveriti.

V prvem numeričnem primeru so razmerja n_{1j}/n_{0j} približno enaka. Koeficienti B'_j in L_j se ujemajo na 3 decimalna mesta in ni možna različna interpretacija njihovega pomena.
V drugem primeru so ostale vse robne frekvence enake kot v prvem primeru. Spremenjene so le frekvence, ki se nanašajo na spremenljivko X3. Koeficienti B'_j in L_j so sicer različni, vendar je interpretacija njihovega pomena še vedno enaka.
V tretjem primeru pa bi po nakazanem kriteriju lahko razmišljali o razlikah pri B_2' in L_2 ter B_5' in L_5.

Primer 1. Razmerja n_{1j}/n_{0j} so približno enaka.

				X1 X2 X3					
	000	100	010	001	110	101	011	111	
Y									
0	17	22	31	13	25	40	45	39	232
1	18	18	34	12	25	45	50	40	238
	35	40	65	25	50	85	95	79	474
p_j	51.4	45.0	52.3	48.0	50.0	52.9	52.6	50.6	51.1

$x^2 = 0.966 \quad x^2/N = 0.00204$

	B_j'	L_j
0	0.007	0.008
1	−0.014	−0.014
2	0.021	0.021
3	0.014	0.014
4	−0.007	−0.007
5	0.029	0.029
6	−0.009	−0.009
7	−0.028	−0.027

R^2 = 0.00204, $SE_{B'}$ = 0.051, SE_L = 0.050

Primer 2. Razmerja n_{1j}/n_{0j} so spremenjena samo v okviru spremenljivke X3.

		X1 X2 X3							
		000	100	010	001	110	101	011	111
Y									
	0	17	22	60	13	25	40	16	39	232
	1	18	18	5	12	25	45	79	40	238
		35	40	65	25	50	85	95	79	474
p_j		51.4	45.0	7.7	48.0	50.0	52.9	83.2	50.6	51.1

x^2 = 88.92 x^2/N = 0.1876

	B_j'	L_j
0	−0.028	−0.056
1	0.021	0.050
2	−0.015	−0.044
3	0.202	0.261
4	0.028	0.057
5	−0.159	−0.219
6	0.179	0.238
7	−0.216	−0.274

R^2 = 0.1876, $SE_{B'}$ = 0.046, SE_L = 0.056

Primer 3. Razmerja n_{1j}/n_{0j} so spremenjena slučajno.

				X1 X2 X3					
	000	100	010	001	110	101	011	111	
Y									
0	21	55	84	13	25	40	16	39	293
1	18	18	5	12	44	90	81	145	413
	39	73	89	25	69	130	97	184	706
p_j	46.2	24.7	5.6	48.0	63.8	69.2	83.5	78.8	58.5

$x^2 = 203.7 \quad x^2/N = 0.2885$

	B_j'	L_j
0	0.049	0.013
1	0.133	0.183
2	0.109	0.078
3	0.348	0.441
4	0.134	0.191
5	−0.050	−0.110
6	0.116	0.194
7	−0.264	−0.338

$R^2 = 0.2886, \quad SE_{B'} = 0.037, \quad SE_L = 0.053.$

LITERATURA:

Leo A. Goodman. Analyzing Qualitative/Categorical Data. Abt Books,1978.
Shelby J. Haberman. Analysis of Qualitative Data. Volume 1, 2, fAcademic press, New York, 1978–79.
Konstantin Momirović. Uvod u analizu nominalnih variabli. Volume 2, Metodološke sveske, FSPN, Ljubljana, 1988.
Annette J. Dobson. Introduction to Statistical Modelling. Champman and Hall, London, 1983.
G. Nigel Gilbert. Modelling Society. An introduction to Loglinear Analyisis for Social Researchers. George Allen & Unwin, London 1981.
Grahm J. G. Upton. The Analysis of Cross–tabulated Data. John Wiley & Sons, New York, 1970.