Keleti András

A loglineáris modell

A 70-es években a statisztika módszertanával foglalkozók érdeklődése középpontjába egy olyan többváltozós elemzési eljárás kidolgozása került, mely három vagy több kategória változó közötti kapcsolat rendszer szerkezetét volt hivatott leírni.

Ha megfigyeléseinket három vagy több kategória változó szerint csoportosítjuk, ezek egy több dimenziós kereszttábla formájában jelennek meg. Az ilyen táblák olyan sajátos elemzési és értelmezési problémákat vetnek fel, melyek a hagyományos több változós elemzési módszerek számára nem hozzáférhetők. Egy kereszttábla két változója közötti kapcsolat az alábbi módon értelmezhető: ha egy megfigyelés az egyik változó bizonyos kategóriájába esik, ez valószínűbbé teszi ugyanennek a megfigyelésnek a másik változó bizonyos kategóriáiba való esését. Az ilyen jellegű kapcsolatokat a két változó közötti interakciónak szokták nevezni. A több dimenziós kereszttáblák elemzésére használt korábbi technikák a tábla különböző két dimenziós szélösszegeit elemezték különkülön, ami azt jelentette, hogy *egyszerre* csak két változót tudtak vizsgálni. Bár ez a megközelítés gyakran enged betekintést a változók közötti kapcsolatba, lényeges korlátai vannak:

a) Összekeveri két változó marginális kapcsolatát[1] a többi változó jelenlétében érvényesülő kapcsolattal.
b) Nem teszi lehetővé a fenti páros kapcsolatok szimultán elemzését.
c) Figyelmen kívül hagyja azt a lehetőséget, hogy a változók között nem csak két, hanem három vagy több irányú interakciók is érvényesülhetnek.

E tanulmányban a keresztosztályozás (cross classification) statisztikai elemzésére nemrégiben kidolgozott módszert szeretnék bemutatni, mely loglineáris modellek segítségével elemzi a több dimenziós kereszttáblákat.[2] Ez a módszer kiküszöböli a fenti hiányosságokat. A modellek alapgondolata a következő: a több dimenziós kereszttábla változói közötti interakciók a becsült esetszámok keresztszorzat hányadosai (cross-product ratio) alapján definiálhatók. Ebből az következik, hogy minden várható esetszám logaritmusa kifejezhető a többi várható esetszám logaritmusai valamilyen lineáris kombinációjaként, úgy hogy a használt súlyok, illetve együtthatók összege nulla. Innen a loglineáris elnevezés.

[1] A két változó közötti interakciót a többi változó mentén összevont marginális táblában vizsgálják.

[2] Bishop, Fienberg, Holland (1975), Cox (1970), Haberman (1974), Lindsey (1973) és Placket (1974) a szóban forgó módszer részletes leírását adják.

Az általános modell három változó esetén

Legyen f_{ijk} egy három dimenziós kereszttábla i-edik sorában, j-edik oszlopában és k-adik rétegében található megfigyelt gyakoriság, és F_{ijk} legyen a modell alapján becsült megfelelő várható gyakoriság, azaz $F_{ijk} = E\{f_{ijk}\}$. A loglineáris modell általános alakja a következő:

$$\log F_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \qquad (1)$$

ahol mint a variancia analízis modellben

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_k u_{3(k)} = 0,$$

$$\sum_i u_{12(ij)} = \sum_j u_{12(ij)} = \sum_i u_{13(ik)} = \sum_k u_{13(ik)} = \sum_j u_{23(jk)} = \sum_k u_{23(jk)} = 0$$

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0. \qquad (2)$$

Ez a modell minden hatást figyelembe vesz, melyek egy háromdimenziós táblában előfordulhatnak:

u — a főátlag, a várható gyakoriságok logaritmusai számtani átlaga;

$u_{1(i)}$, $u_{2(j)}$, $u_{3(k)}$ — főhatások, sor, oszlop és réteg hatások, a megfelelő sor, oszlop és réteg átlagok eltérései a főátlagtól;

$u_{12(ij)}$, $u_{13(ik)}$, $u_{23(jk)}$ — elsőrendű interakciók, a kétirányú kapcsolatok hatását kifejező paraméterek, a megfelelő alacsonyabb rendű paraméterektől való eltéréseket mérik;

$u_{123(ijk)}$ — a három irányú kapcsolat paramétere. (A fenti paraméterek jelentéséről később részletesen szó lesz.)

Nevezzük ezt a modellt telítettnek, mivel minden lehetséges hatást figyelembe vesz, ezért: $f_{ijk} = F_{ijk}$. A kutatót az elemzés során az érdekli, hogy a vizsgált jelenség leírható-e az általánosnál takarékosabb modellel. Más szóval: az összes lehetséges hatás közül azokat akarja kiválasztani, melyek valóban befolyásolják a vizsgált jelenséget. Ezt úgy érheti el, hogy az általános alakú loglineáris modellben bizonyos paramétereket 0-val tesz egyenlővé. Az így kapott modelleket illeszti a megfigyelt adatokhoz, majd valamilyen statisztika segítségével méri az illeszkedés jóságát.

A hipotézisek

Attól függően, hogy mely paramétereket tesszük 0-val egyenlővé, különböző hipotéziseket fogalmazhatunk meg a változók függetlenségéről. A három változó teljes függetlenségét feltételező loglineáris modell:

$$\log F_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)},$$

tehát

$$u_{12(ij)} = u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0 \qquad (3)$$

Ha azt feltételezzük, hogy változóink nem függetlenek egymástól, négyfajta hipotézist fogalmazhatunk meg a három változó közötti kapcsolatok szerkezetéről:

(1) Van egy változónk, mely független a másik kettőtől. Ebben az esetben a három irányú interakcióról és a két irányú interakciók közül kettőről feltételezzük, hogy 0-val egyenlőek. E hipotézis fajtának három lehetséges verziója van, például az egyik:

$$u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0. \tag{4}$$

(2) Két változó független egymástól, ha a harmadik változó értéke adott. E hipotézisnek, mely két változó feltételes függetlenségét fejezi ki, szintén három változata van, például:

$$u_{13(ij)} = u_{123(ijk)} = 0. \tag{5}$$

(3) Páros kapcsolatok a három változó között, úgy hogy mindhárom két változós interakció független a harmadik változó értékétől:

$$u_{123(ijk)} = 0. \tag{6}$$

(4) Ez a hipotézis már a telített modellt (1) hozza vissza, mivel három irányú interakciót is feltételezünk, tehát bármely két változós interakció függ a harmadik változó értékétől.

A becslés

A következőkben a modellek alapján várható gyakoriságok maximum likelihood becsléséről lesz szó. Induljunk ki abból, hogy egy modell várható esetszámai becslésekor csak azokat az információkat vesszük figyelembe, melyeket a modellbe felvett paraméterekhez tartozó megfigyelt szélösszegek tartalmaznak. Eszerint például a változók teljes függetlenségét feltételező modell (3) várható esetszámait az alábbi képlet szerint számolhatjuk:

$$F_{ijk} = \frac{f_{i++} f_{+j+} f_{++k}}{N^2}, \tag{7}$$

ahol a + jelek összeadást jelölnek a megfelelő index mentén és N az összes esetszám.

A változók között egyetlen két irányú interakciót feltételező modell (4) várható gyakoriságainak maximum likelihood becslése a következő:

$$F_{ijk} = \frac{f_{ij+} f_{++k}}{N}. \tag{8}$$

A két változó feltételes függetlenségét feltételező modell (5) várható esetszámait az alábbi képlettel becsüljük:

$$F_{ijk} = \frac{f_{ij+} f_{+jk}}{f_{+j+}}. \tag{9}$$

A (6) $u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} + u_{13(ik)}$ modell várható esetszámai maximum likelihood becslése nem lehetséges a fentiekhez hasonló direkt mó-

don. E modell várható esetszámai a három dimenziós tábla megfigyelt adatainak összevonása útján nyert három két dimenziós $\{f_{ij+}\}$, $\{f_{+jk}\}$, $\{f_{i+k}\}$ marginális tábla függvényei. A várható gyakoriságok a DEMING és STEPHAN (1940) által kidolgozott iterációs illesztési eljárás segítségével becsülhetők.[3] Az iterációs algoritmus bemutatása előtt fogalmazzuk meg a modellek várható esetszámai becslésének általános szabályát:

(1) Minden változóra vonatkozóan keressük meg azt a modellbe felvett legmagasabb rendű paramétert, mely a szóban forgó változót magában foglalja.
(2) Számítsuk ki ezekhez a legmagasabb rendű paraméterekhez tartozó megfigyelt szélösszegeket, például az $\{u_{21(ij)} \mid i = 1, 2, \ldots, I; j = 1, 2, \ldots, J\}$ paraméternek az $\{f_{ij+} \mid i = 1, 2, \ldots, I; j = 1, 2, \ldots, J\}$ szélösszeg felel meg.
(3) A várható esetszámok becsléséhez csak a fenti megfigyelt szélösszegeket használjuk fel.

Az iteráció

Az $u_{123} = 0$ modell várható gyakoriságai becslésével kapcsolatban leszögeztük, hogy az F_{ijk}-k egyedül a $\{f_{ij+}\}$, $\{f_{i+k}\}$, $\{f_{+jk}\}$ szélösszegek függvényei. A maximum likelihood becslés módszerét alkalmazva az F_{ijk}-knak az alábbi egyenlőségeket kell kielégíteniük:

$$\begin{aligned} F_{ij+} &= f_{ij+}, & i &= 1, 2, \ldots, I & j &= 1, 2, \ldots, J \\ F_{i+k} &= f_{i+k}, & i &= 1, 2, \ldots, I & k &= 1, 2, \ldots, K \\ F_{+jk} &= f_{+jk}, & j &= 1, 2, \ldots, J & k &= 1, 2, \ldots, K. \end{aligned} \tag{10}$$

Annak ellenére, hogy a fenti egyenlőségek egyértelműen meghatározzák a modell alapján várható gyakoriságok halmazát, mégsem tudjuk ezeket zárt alakban kifejezni.

Először tegyük az összes F_{ijk}-t 1-gyel egyenlővé: $F_{ijk}^{(0)} = 1$; (a felső index mindig az iteráció megfelelő lépését jelöli), majd az iteráció első ciklusában a további három lépés következik:

$$F_{ijk}^{(1)} = \frac{F_{ijk}^{(0)} f_{ij+}}{F_{ij+}^{(0)}},$$

$$F_{ijk}^{(2)} = \frac{F_{ijk}^{(1)} f_{i+k}}{F_{i+k}^{(1)}},$$

$$F_{ijk}^{(3)} = \frac{F_{ijk}^{(2)} f_{+jk}}{F_{+jk}^{(2)}}.$$

Ezzel befejeződik az iteráció első ciklusa, a ciklusok addig ismétlődnek, míg a várható gyakoriságok változása egyik ciklusról a következőre megfelelően kicsi lesz. Az iteráció első lépésében az 1-esek vektorát vettük induló értékeknek.

[3] A Deming-Stephan algoritmus csak hierarchikus modellekre jó. (Az eljárás kihasználja hogy a peremösszegek teljes elégséges statisztikát adnak.)

Az $u_{123} = 0$ modell illesztéséhez az induló értékek bármely más vektora megfelelő lenne, de BISHOP, FIENBERG és HOLLAND (1975) kimutatták, hogy más induló értékek sem növelik lényegesen a konvergencia sebességét. Az 1-esek mellett az szól, hogy ezek alkalmasak más loglineáris modellek illesztésére is, és ezért megkönnyítik a számítógépes programok készítését.[4]

Az illeszkedés

Egy modell várható gyakoriságainak becslése után a következő lépés annak eldöntése, hogy a modellben megfogalmazott hipotézis elfogadható-e vagy nem. Erre a kérdésre az illeszkedés-vizsgálat adja meg a választ, mely azt teszteli, hogy a várható esetszámok elég jól, szorosan illeszkednek-e a megfigyelt esetszámokhoz. Az illeszkedés „jóságát" az alábbi két statisztika segítségével vizsgáljuk:

$$X^2 = \sum \frac{(f - F)^2}{F} \tag{11}$$

$$G^2 = 2 \sum f \log \left(\frac{f}{F} \right), \tag{12}$$

ahol: f — a megfigyelt esetszám,
F — a modell alapján becsült esetszám.
Ha az illesztett modell hibátlan, és a minta elég nagy, akkor mindkét statisztika megközelítően X^2 eloszlású, az alábbi formula szerinti szabadságfokokkal:

$$s.f. = \# \text{ cellák} - \# \text{ illesztett paraméterek}. \tag{13}$$

Ezt a formulát alkalmazva a tárgyalt öt modell esetében az alábbi szabadságfokokat kapjuk egy $I \times J \times K$ méretű táblában.

1. táblázat

Modell	szükséges szélösszegek	illesztett paraméterek	szabadságfokok
$u + u_1 + u_2 + u_3$	(1) (2) (3)	$[1 + (I-1) + (J-1) + (K-1)]$	$(IJK - I - J - K + 2)$
$u + u_1 + u_2 + u_3 + u_{12}$	(12) (3)	$[1 + (I-1) + (J-1) + + (K-1) + (I-1)(J-1)]$	$[(K-1)(IJ-1)]$
$u + u_1 + u_2 + u_3 + u_{12} + + u_{23}$	(12) (23)	$[1 + (I-1) + (J-1) + + (K-1) + (I-1)(J-1) + + (J-1)(K-1)]$	$[J(I-1)(K-1)]$
$u + u_1 + u_2 + u_3 + u_{12} + + u_{23} + u_{13}$	(12) (23) (13)	$[1 + (I-1) + (J-1) + + (K-1) + (I-1)(J-1) + + (J-1)(K-1) + (I-1) (K-1)]$	$[(I-1)(J-1)(K-1)]$
$u + u_1 + u_2 + u_3 + u_{12} + + u_{23} + u_{13} + u_{123}$	(123)	IJK	0

[4] Ma már a legtöbb hazai számítógép rendelkezik olyan könyvtári programokkal, amelyek alkalmasak az illesztés és a modell-kiválasztás elvégzésére. Ilyen például az MTA SZTAKI „BMDP programok rövid összefoglalása" amelynek BMDP3D ésBMDP3F szakaszai éppen a jelen dolgozat témájával kapcsolatosak.

A X^2 és G^2 statisztikák aszimptotikusan egyenlők. Ez azt jelenti, hogy egyenlők akkor, ha a minta elég nagy (legalább a cellák számának tízszerese), és a null hipotézis igaz.

A paraméterek

Ha kiszámítottuk egy loglineáris modell várható gyakoriságait, és az illeszkedés vizsgálata alapján úgy találtuk, hogy a modellben megfogalmazott hipotézis elfogadható, felmerül az igény, hogy megbecsüljük a modellbe felvett hatások, paraméterek nagyságát, erejét. *A paramétereket a modell várható gyakoriságai felhasználásával becsüljük.* Egy $I \times J \times K$ méretű kereszttábla három változója legyen A, B és C, a modell alapján várható gyakoriságok logaritmusait jelöljük v_{ijk}-val ($\log F_{ijk} = v_{ijk}$). A loglineáris modell lehetséges paraméterei: u_i^A, u_j^B, u_k^C főhatások; u_{ij}^{AB}, u_{jk}^{BC}, u_{ik}^{AC} elsőrendű interakciók; u_{ijk}^{ABC} másodrendű interakció.

Már utaltunk arra, hogy a magasabb rendű u tagok a megfelelő alacsonyabb rendű u tagoktól való eltéréseket mérik, ennek megfelelően:

$$u_i^A = \frac{v_{i++}}{JK} - \frac{v_{+++}}{IJK}$$

$$u_{ij}^{AB} = \frac{v_{ij+}}{K} - \frac{v_{i++}}{JK} - \frac{v_{+j+}}{IK} + \frac{v_{+++}}{IJK} \tag{13}$$

$$u_{ijk}^{ABC} = v_{ijk} - \frac{v_{ij+}}{K} - \frac{v_{i+k}}{J} - \frac{v_{+jk}}{I} + \frac{v_{i++}}{JK} + \frac{v_{+j+}}{IK} + \frac{v_{++k}}{IJ} - \frac{v_{+++}}{IJK}.$$

A (13) formulában a loglineáris modell paramétereit a megfelelő szintű átlagok lineáris kombinációjaként fejeztük ki. Az u_j^B, u_k^C, u_{jk}^{BC}, u_{ik}^{AC} paraméterek is hasonló módon írhatók le.

A variancia analízis ANOVA modelljeivel az analógia nyilvánvaló, ez azonban ne tévessze meg az olvasót. A variancia analízist akkor használjuk, ha független változóknak egy függő változóra való hatását akarjuk megbecsülni. A kereszttáblák elemzésére használt ANOVA-szerű modellek azonban a tábla dimenzióinak megfelelő változók közötti kapcsolat szerkezetét hivatottak leírni. Megkönnyítjük a paraméterek értelmezését, ha bevezetjük a következő korlátozást: $I = J = K = 2$, és a (13) formulában elvégezzük a megfelelő behelyettesítéseket.
Ekkor:

$$2u_i^A = \frac{1}{4} \sum_{j=1}^{2} \sum_{k=1}^{2} (v_{1jk} - v_{2jk}),\ u_1^A = u_i^A;\ u_2^A = -u_i^A. \tag{14}$$

A (14) formula bal oldala nem más mint az F_{1jk}/F_{2jk} esélyek logaritmusainak számtani átlaga.

A tanulmány elején már utaltunk arra, hogy a loglineáris modellben a változók közötti interakciók visszavezethetők az esélyhányadosokra (odds ratio).

Először tehát ezek jelentését kell tisztázni. Egy 2×2 méretű táblában az esély hányados képlete a következő:

$$O = \frac{f_{11}/f_{12}}{f_{21}/f_{22}}.$$

Ez azt fejezi ki, hogy hányszor nagyobb az első sorba tartozó megfigyelés esélye arra, hogy inkább található az első oszlopban mint a másodikban, mint a második sorba tartozó megfigyelés ugyanilyen esélye. Az esélyhányados a publikációkban általában az alábbi keresztszorzat-hányados (cross-product ratio) alakban szokott megjelenni:

$$O = \frac{f_{11}f_{22}}{f_{12}f_{21}}.$$

Egy $I = J = K = 2$ méretű három dimenziós kereszttáblát úgy képzelhetünk el, mint két 2×2-es kereszttáblát, ahol mindkét táblának egy-egy esélyhányados felel meg. Legyenek ezek $O_{AB.1}$ és $O_{AB.2}$, melyeket *feltételes* várható esélyhányadosoknak[5] tekinthetünk, mivel nagyságuk függ a harmadik C változó által felvett értéktől.

$$O_{AB.K} = \frac{F_{11k}F_{22k}}{F_{12k}F_{21k}}. \tag{15}$$

Vezessük be továbbá az A és B változók közötti *parciális* várható esélyhányados fogalmát, $O_{AB.C}$-t, mely $O_{AB.1}$ és $O_{AB.2}$ mértani átlaga, $O_{AB.C} = \sqrt{O_{AB.1}O_{AB.2}}$, tehát nagysága már nem függ C értékétől.

Ha most elvégezzük (13)-ba a megfelelő behelyettesítéseket, az alábbi alakhoz jutunk:

$$u_{11}^{AB} = \frac{1}{8}(v_{111} + v_{221} - v_{121} - v_{211} + v_{112} + v_{222} - v_{122} - v_{212}).$$

Tehát:

$$4u_{11}^{AB} = \frac{\log O_{AB.1} + \log O_{AB.2}}{2} = \log O_{AB.C} \tag{16}$$

$$u_{11}^{AB} = u_{22}^{AB} = u_{ij}^{AB}; \quad u_{12}^{AB} = u_{21}^{AB} = -u_{ij}^{AB}.$$

Tehát a loglineáris modell u_{ij}^{AB} paramétere a megfelelő esélyhányadosok logaritmusai számtani átlagának négyszerese. A másik két elsőrendű interakció is hasonlóképpen írható fel. Például:

$$4u_{jk}^{BC} = \frac{\log O_{BC.1} + \log O_{BC.2}}{2} = \log O_{BC.A}. \tag{17}$$

[5] Azért „várható", mert a loglineáris modell alapján becsült várható gyakoriságokból számoljuk őket.

Ha a (13) formula másodrendű interakció képletébe végezzük el a megfelelő behelyettesítéseket, az alábbi alakhoz jutunk:

$$u_{111}^{ABC} = \frac{1}{8}(v_{111} - v_{121} - v_{211} + v_{221} - v_{112} + v_{122} + v_{212} - v_{222}),$$

melyből további átrendezés után:

$$8u_{111}^{ABC} = \log O_{AB.1} - \log O_{AB.2} = \log \frac{O_{AB.1}}{O_{AB.2}} \tag{18}$$

$$u_{ijk}^{ABC} = u_{111}^{ABC} = u_{221}^{ABC} = u_{212}^{ABC} = u_{122}^{ABC} = -u_{211}^{ABC} = -u_{121}^{ABC} = -u_{112}^{ABC} = -u_{222}^{ABC}.$$

Tehát a loglineáris modell u_{ijk}^{ABC} három irányú interakciót kifejező paramétere arányos az A és B változókhoz tartozó két feltételes esélyhányados hányadosának logaritmusával.

Tekintsük most az $u_{ijk}^{ABC} = 0$ loglineáris modellt, mely szerint a változók közötti kétirányú interakciók függetlenek a harmadik változó értékétől. Ez a modell az esélyhányadosok nyelvén a következőt jelenti:

$$\frac{F_{111}F_{221}}{F_{121}F_{211}} = \frac{F_{112}F_{222}}{F_{122}F_{212}},$$

tehát:

$$4u_{ij}^{AB} = \log O_{AB.1} = \log O_{AB.2} = \log O_{AB.C},$$

és hasonlóképpen:

$$4u_{jk}^{BC} = \log O_{BC.1} = \log O_{BC.2} = \log O_{BC.A}$$
$$4u_{ik}^{AC} = \log O_{AC.1} = \log O_{AC.2} = \log O_{AC.B}.$$

Hierarchikus modellek

Az eddigiekben nem foglalkoztunk a loglineáris modell összes lehetséges változatával. Nem vizsgáltuk például ezt a modellt:

$$F_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{123(ijk)}. \tag{20}$$

Figyelmünket csak olyan modellekre korlátoztuk, melyek magasabb rendű u tagokat csak akkor foglalnak magukban, ha a megfelelő alacsonyabb rendű u tagokat is magukban foglalják. Így az u_{123} paraméter csak akkor lehet a modellben, ha az u_{12}, u_{23} és u_{13} paraméterek is szerepelnek benne. Az így felépített modelleket hierarchikus modelleknek nevezzük. A (20) modell nem hierarchikus, az ilyen modelleknél a fő problémát nem az illesztés jelenti — ennek módszerét Bloomfield (1974) és Haberman (1974) már kidolgozták[6] — hanem az értelmezés. Arról van szó, hogy egy nem hierarchikus modell paramétereit nem tudjuk az előző fejezetben leírt módon értelmezni: a megfelelő alacsonyabb rendű paraméterektől való eltérésekként.

[6] Ez az eljárás nagyon lassú, ezért a számítógépes programok nem számolnak nem hierarchikus modellt.

A modell kiválasztása

A nagyszámú paramétert magukban foglaló bonyolult modellek általában jobban illeszkednek, mint egy egyszerűbb modell, mely a bonyolultabbnak speciális esete. Mégis ha két elfogadhatóan illeszkedő modellünk van, nem a jobban illeszkedőt választjuk akkor, ha a másik rosszabbul illeszkedő, de egyszerűbb, takarékosabb modell várható gyakoriságai és a jobban illeszkedő modell várható gyakoriságai között az eltérés nem szignifikáns. A modellek közötti választás problémája nyilvánvalóan akkor merül fel, ha több elfogadhatóan illeszkedő modellünk van. Mivel az illeszkedés jóságát tesztelő statisztikák statisztikailag nem függetlenek, a modellek illeszkedésének jóságát nem vizsgálhatjuk külön-külön, mint ezt korábban tettük. Szükségünk van tehát egy módszerre, mely segítségünkre lesz a szignifikáns interakciók kiválasztásában. Sajnos a modell kiválasztásnak nincs egy minden célt kielégítő legjobb módszere. BISHOP (1969), BROWN (1976), FIENBERG (1970), GOODMAN (1970, 1971) és KU-KULLBACK (1968) mind más utakon közelítik meg a problémát. Az alábbi technika általános sémát nyújt két modell várható gyakoriságai összehasonlításához, ha az egyik modell a másiknak speciális esete:

$$2 \sum f \log \left(\frac{F_1}{F_2} \right) . \tag{21}$$

Ezt a statisztika arra ad választ, hogy a két modell várható gyakoriságai közötti különbség véletlen vagy szisztematikus. A statisztika X^2 eloszlású és szabadságfoka a két modell szabadságfokainak különbsége.

A modell kiválasztás menetére nézzünk most egy példát (ANDERSEN, 1980). Svéd közlekedési adatokat fogunk elemezni. Baleseti adatokat gyűjtöttek össze 18 héten át 1961-ben és 18 héten át 1962-ben. Mindkét évben 90 km/óra sebesség korlátozást vezettek be bizonyos napokon. A baleseteket feljegyezték mind az autópályákon, mind az egyéb utakon.[7] Így a megfigyelt balesetek három kritérium szerint osztályozhatók: (1) az út típusa melyen a baleset történt (autópályák, egyéb utak), (2) volt-e sebesség korlátozás a baleset napján vagy nem volt, (3) az év melyben a baleset történt. A három dimenziós kereszttábla megfigyelt adatait a 2. táblázat tartalmazza.

A 2. táblázat megfigyelt gyakoriságaihoz nyolc olyan hierarchikus modell illeszthető, melyek mind a három fő hatást tartalmazzák. Az illesztett modelleket H-val jelöljük és a H-kat aszerint különböztetjük meg, hogy mely inter-

[7] Ez a mintavételi eljárás jó példája a Poisson mintavételi modellnek. Keresztosztályozott adatok gyűjtésének három fő módja van:

(1) Poisson modell: egy előre meghatározott időtartam alatt végzik a megfigyelést, anélkül, hogy a minta nagyságát előre meghatároznák.
(2) Multinominális modell: egy előre meghatározott elemszámú mintát veszünk, és ennek elemeit keresztosztályozzuk aszerint, hogy a különböző változók mely kategóriáiba tartoznak.
(3) Product-Multinominális modell: előre meghatározzuk a sor változó minden kategóriájának elemszámát.

Mind a három mintavételi séma ugyanazokat a várható gyakoriságokat és ugyanazokat az illeszkedési statisztikákat eredményezi.

2. táblázat

Év	Sebesség korlátozás	Autópályák	Egyéb utak	Összesen
1961	90 km	8	42	50
	nincs	57	106	163
	összesen	65	148	213
1962	90 km	11	37	48
	nincs	45	69	114
	összesen	56	106	162
Összesen	90 km	19	79	98
	nincs	102	175	277
	összesen	121	254	375

akciókat hagyják figyelmen kívül. A modellek illeszkedését tesztelő G^2 statisztikák eredményeit a 3. táblázat mutatja be.

3. táblázat

Hipotézis	G^2	s. f.
H_{123}	0,19	1
H_{12}	11,36	2
H_{13}	1,34	2
H_{23}	2,44	2
$H_{13,23}$	3,13	3
$H_{12,23}$	13,16	3
$H_{12,13}$	12,05	3
$H_{12,13,23}$	13,85	4

A harmadik tábla alapján nyilvánvaló, hogy a szükséges paraméterek kiválasztása céljából a hipotézisek jóságát az alábbi rendben érdemes tesztelni: $H_{123} \rightarrow H_{13} \rightarrow H_{13,23} \rightarrow H_{12,13,23}$. Az eredményeket a 4. táblázat tartalmazza.

4. táblázat

Okozott variancia	Hipotézis	Teszt	Szabadságfok
másodrendű interakció	H_{123}	0,19	1
elsőrendű interakció (1, 3)	H_{13}	1,15	1
elsőrendű interakció (2, 3)	$H_{13,23}$	1,79	1
elsőrendű interakció (1, 2)	$H_{12,13,23}$	10,72	1

A paraméterek szignifikanciáját (21) alapján teszteljük. Például az (1, 3) interakció hozzájárulását a megfigyelt gyakoriságok varianciájához az alábbi módon mérjük:

$H_{13} - H_{123} = 1{,}34 - 0{,}19 = 1{,}15$, ahol s. f. $= 2 - 1 = 1$ és $H_{13}(u_{123} = u_{13} = 0)$, H_{123}-nak egy speciális esete. A két modell várható gyakoriságai között a

különbség nem szignifikáns, tehát az (1,3) interakció felvétele a modellbe nem indokolt. Hasonló eredményre jutunk ha a (2,3) interakciót vizsgáljuk, ez sem járul hozzá szignifikáns mértékben a megfigyelt gyakoriságok varianciájához. A 4. táblázat azt mutatja, hogy a H_{123}, H_{13} és a $H_{13,23}$ hipotézisek elfogadhatók, tehát az (1,2) másképp u_{12} interakció kivételével minden interakció 0. Esetünkben ez a következőket jelenti: a balesetek megoszlása úttípus szerint egyforma 1961-ben és 1962-ben; a sebesség korlátozás esetén történt balesetek és a sebesség korlátozás hiánya esetén történt balesetek hányadosa ugyanannyi 1962-ben mint 1961-ben; viszont az utak típusa és a sebesség korlátozás léte vagy nem léte között van interakció, tehát a sebesség korlátozásának hatása más az autópályákon és más az egyéb utakon.

A modell kiválasztása során az $u + u_1 + u_2 + u_3 + u_{12}$ modellhez jutottunk, ez az a legegyszerűbb, legtakarékosabb modell, mellyel még kielégítően le tudjuk írni a vizsgált jelenséget. A modell paramétereinek becsült értékei az 5. táblázatban szerepelnek.

5. táblázat

$u_{12(ij)}$	$j = 1$	2
$i = 1$	−0,22	+0,22
2	+0,22	−0,22

	$i = 1$	2
$u_{1(i)}$	−0,49	+0,49
	$j = 1$	2
$u_{2(j)}$	−0,62	+0,62
	$k = 1$	2
$u_{3(k)}$	+0,14	−0,14
$u = 3{,}57$		

Az $u_{12(ij)}$ interakció becsült értékei azt mutatják, hogy a sebesség korlátozásának hatása nagyobb az autópályákon, mint az egyéb utakon. Általánosságban is megállapíthatjuk, hogy egy u paraméter pozitív értéke ($u > 0$) azt jelenti, hogy a megfelelő cella vagy szélösszeg megfigyelt értéke magasabb, mint a modell alapján várható értéke. Ennek megfelelően $u < 0$, ha a megfigyelt érték alacsonyabb, mint a várható. Az $u_{1(i)}$ és $u_{2(j)}$ főhatások becsléseit nehéz értelmezni, mivel nem ismerjük az autópályák és az egyéb utak teljes hosszát, sem a napok számát, melyeken volt, illetve nem volt sebesség korlátozás. Az $u_{3(k)}$ főhatás értelmezése könnyebb, mivel mindkét évben 18 héten át tartott a megfigyelés, és 1961-ben jelentősen több baleset történt, mint 1962-ben, függetlenül a sebesség korlátozásától.

Ha egy gyakorlott szemű statisztikus figyelmesen tanulmányozza a 2. táblázatot, minden bonyolultabb matematikai-statisztikai elemzési eszköz nélkül is ugyanazokra a következtetésekre fog jutni, mint amelyekre mi jutottunk a loglineáris modellek segítségével. Azonban nagyobb méretű három, négy vagy több dimenziós kereszttáblákban az összefüggések elemzése már megkívánja

a loglineáris módszer igénybevételét. E tanulmányban csak a loglineáris modell három változós esetéről volt szó, azonban a különböző hipotézisek értelmezésére, a modellek illesztésére, a paraméterek jelentésére és a modell kiválasztásra vonatkozó minden megállapítás különösebb nehézség nélkül kiterjeszthető a négy vagy több változós esetekre is.

A loglineáris modellekkel kapcsolatban feltétlenül utalnunk kell a mobilitás vizsgálatára. Ez az a terület ugyanis, ahol ilyen modelleket eddig a legeredményesebben alkalmaztak. Az itt felmerült elemzési problémák megoldására tett erőfeszítések jelentős mértékben járultak hozzá a loglineáris módszer fejlődéséhez. A probléma a következő: ha különböző időszakokra, területekre vagy országokra vonatkozó azonos felépítésű kereszttáblákat hasonlítunk össze, azzal a kérdéssel találjuk szembe magunkat, hogy mi okozza a táblázatok között az eltéréseket: a széleloszlások változása-e, vagy a fej illetve oldal rovatban szereplő változók közötti kapcsolat szorosságának változása. A különböző időszakokra vagy régiókra vonatkozó mobilitás táblák esetében a fenti kérdés a következőképpen merül fel: mi okozza a mobilitás megfigyelt változásait az apák és fiaik foglalkozás szerinti megoszlásában, (más szóval: a társadalom foglalkozási struktúrájában) bekövetkező változások, vagy pedig az apa és fia társadalmi helyzete (más szóval: a származás és az elért helyzet) közötti interakció erejének változása. ANDORKA—CSICSMAN—KELETI a Statisztikai Szemle 1981 októberi számában megjelent „A magyar társadalom nyitottságának változásai" című cikkükben a szóban forgó kérdés megválaszolásához az $u_{123} = 0$ loglineáris modellt hívja segítségül.

Ez a modell — ahol 1: az apa foglalkozása; 2: a fiú foglalkozása; 3: a születési kohorsz — azt feltételezi, hogy az apák és fiaik foglalkozás szerinti megoszlása a vizsgált időszakban kohorszról kohorszra változott, és van kapcsolat a származás és az elért helyzet között, de ez a kapcsolat ugyanolyan erejű minden kohorszban. A Statisztikai Szemlében megjelent tanulmány csak a férfiak mobilitásával foglalkozik. Az illeszkedés vizsgálata alapján az $u_{123} = 0$ modellben megfogalmazott nulla hipotézis nem bizonyult elfogadhatónak, ami azt jelenti, hogy szemben a nyugat-európai és az amerikai mobilitás vizsgálatok eredményeivel, Magyarországon az apa-fiú interakció a vizsgált időszakban változott. Időközben az 1973-as vizsgálatban összeírt nők mobilitását is elemeztük a loglineáris módszerrel, a férfiakra és nőkre vonatkozó adatok egy felvételből származnak, az elemzésben ugyanazt a nyolc foglalkozási kategóriát és ugyanazt a négy születési kohorszot használtuk a nők esetében, mint a férfiakéban. Így az eredmények összehasonlíthatók. A 6. táblázatban 6 loglineáris modell illeszkedés vizsgálatának eredményei szerepelnek, külön a férfiakra és külön a nőkre.

6. táblázat

modell		szabadság fokok	G^2	
			férfi	nő
első	1, 2, 3	238	4438,59	4353,38
második	13, 2	217	4144,41	3384,80
harmadik	23, 1	217	3714,36	4010,61
negyedik	12, 3	189	1171,30	1316,19
ötödik	13, 23	196	3420,15	3041,99
hatodik	12, 13, 23	147	236,29	175,34

A társadalom nyitottságának változását érintő szociológiai jelentése a hatodik modellnek van, és ebben a nők és a férfiak mobilitása némileg eltérő képet mutat. A három irányú interakciót 0-nak tételező hipotézist a férfiak esetében el kellett vetnünk, a nők esetében viszont ez a modell elfogadhatóan illeszkedik, tehát az apa-lány interakció a vizsgált időszakban nem változott szignifikáns mértékben. A 7. táblázatban az apa és lánya társadalmi helyzete közötti interakciónak a hatodik modell alapján becsült értékei szerepelnek.

7. táblázat

A (12) loglineáris paraméternek a hatodik modell alapján becsült értékei, osztva a paraméterek standard hibáival

apa foglalkozása	lány foglalkozása							
	1.	2.	3.	4.	5.	6.	7.	8.
1. Vezető	1,308	4,108	0,535	—0,155	0,317	—1,708	—0,684	—2,295
2. Értelmiségi	0,733	10,139	3,362	—0,556	—0,756	—1,087	—1,767	—3,157
3. Egyéb szellemi	0,754	3,974	6,973	—0,613	—1,458	—1,641	—2,195	—2,923
4. Kisiparos	—0,890	—1,884	—2,678	2,741	1,307	—0,182	—1,046	1,124
5. Szakmunkás	—0,418	—1,263	0,778	—0,093	3,383	1,911	0,303	—2,513
6. Betanított m.	—0,882	—2,148	—1,717	—0,412	—0,006	2,803	3,221	4,723
7. Segédmunkás	—0,099	—3,057	—2,212	—0,338	0,579	3,288	3,004	3,096
8. Mezőgazdasági fizikai	—1,390	—3,913	—6,555	0,100	—1,409	2,571	4,495	16,545

A táblázat fődiagonálisában levő értékeket értelmezhetjük az adott rétegre jellemző státusz-öröklési mutatóként. A fődiagonálison kívül eső paraméterek pedig a rétegek közötti társadalmi távolságról adnak információt.

IRODALOM

ANDERSEN (1980): *Discrete statistical models with social science applications.* North-Holland.

BISHOP (1969): Full contingency tables, logits, and split contingency tables. *Biometrics.*

BISHOP—FIENBERG—HOLLAND (1975): *Discrete multivariate analysis: theory and practice.* Cambridge, Mass., The MIT Press.

BLOOMFIELD (1974): Transformations for multivariate binary data. *Biometrics.*

BROWN (1976): Screening effects in multidimensional contingency tables. *Appl. Statist.*

COX (1970): *Analysis of Binary Data.* London, Methuen

LINDSEY (1973): *Inferences from Sociological Survey Data: A Unified Approach.* New York, Elsevier.

PLACKET (1974): *The Analysis of Categorical Data.* London, Griffin.

DARROCH—RATCLIFF (1972): Generalized iterative scaling for loglinear models. *Ann. Math. Statist.*

DEMING—STEPHAN (1940): On a least squares adjustment of sampled frequency table when the expected marginal totals are known.

DEMING—STEPHAN (1940): The sampling procedure of the 1940 population census. *J. Amer. Statist. Assoc.*

FIENBERG (1970): The analysis of multidimensional contingency tables. *Ecology.*

FIENBERG (1977): *The analysis of cross-classified categorical data.* MIT Press.

GOODMAN (1970): The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Amer. Statist. Assoc.*

GOODMAN (1971): The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics.*

GOODMAN (1972): A general model for the analysis of surveys. *Amer. J. Sociol.*

HABERMAN (1974): *The analysis of frequency data.* Univ. of Chicago Press.

KU-KULLBACK (1968): Interactions in multidimensional contingency tables: an information theoretic approach. *J. Res. Nat. Bur. Standards.*