数理解析研究所講究録
345巻 1979年 10-28

10

多重クロス表による社会調査データのモデル解析

統数研　大隅　昇

1. 対数線形モデル　多重クロス表の解析法の1つである，対数線形モデル（Log-linear model：以下LLMと略す）を取りあげ，これにもとずくモデル選択の考え方と適用例を述べる。　多重クロス表とは，複数個の水準（カテゴリー）を持つ項目（アイテム）を多数個（多重）考えて，この項目間の関連を水準ごとの度数情報として分類した表のことである。たとえば2元のクロス表とは，2つの項目A, Bを考え，Aが水準i, Bが水準jとなる水準の組み合せA_iB_j，あるいはセル(i,j)における実現度数をf_{ij}とかく。そしてクロス表全体を$\{f_{ij}\}$で表わす。このときLLMとはf_{ij}の期待値F_{ij} $(i=1,2,\cdots,r;\ j=1,2,\cdots,c)$に対し対数をとって，$E_{ij}=\log F_{ij}$とし$E_{ij}$に関する線形モデルとして度数情報を分析する方法である。このとき若干の変形によりANOVAと同様の表現が可能となる。いまモデル（仮説）として，「A, B間が独立である」という場合を例にとると，

$$E_{ij} = u + u_{A(i)} + u_{B(j)} \qquad (i=1,2,\cdots,r;\ j=1,2,\cdots,c) \qquad (1)$$

ここで, $u = \sum_i \sum_j E_{ij}/rc$,

$$u_{A(i)} = \sum_j E_{ij}/c - u, \quad u_{B(j)} = \sum_i E_{ij}/r - u$$

とかける。なおこのときの制約条件として, $\sum_i u_{A(i)} = \sum_j u_{B(j)} = 0$ がある。式(1)の記法を借りて2元クロス表の情報をさらに一般的にLLMにより表わすと,

$$E_{ij} = u + u_{A(i)} + u_{B(j)} + u_{AB(ij)} \qquad (2)$$

となる。このときの制約条件は, $\sum_i u_{A(i)} = \sum_j u_{B(j)} = \sum_i \sum_j u_{AB(ij)} = 0$ である。(1)はいわゆる「相互独立モデル」であり(2)は項目A,Bの交互作用まで考慮した「飽和モデル」である。2元クロス表の場合には仮説の与え方によりあわせて5通りのLLMがある。次の表1はこれを要約したものである。

(表1) 2元クロス表の場合

	(a) モデルの構造	(b) 対数線形モデル	(c) パラメータ数	(d) $d.f.$	(e) 帰無仮説：Ho	(f) 略記法
M_1	$p_{ij}=\frac{1}{rc}, F_{ij}=\frac{n}{rc}$	$\log F_{ij}=u$	1	$rc-1$	$u_{A(i)}=0$ $u_{B(j)}=0$ $u_{AB(ij)}=0$	{φ} (等確率モデル)
M_2	$p_{ij}=\frac{p_{i.}}{c}, F_{ij}=\frac{f_{i.}}{c}$	$\log F_{ij}=u+u_{A(i)}$	$1+(r-1)$	$r(c-1)$	$u_{B(j)}=0$ $u_{AB(ij)}=0$	{A} (1項目独立モデル)
M_3	$p_{ij}=\frac{p_{.j}}{r}, F_{ij}=\frac{f_{.j}}{r}$	$\log F_{ij}=u+u_{B(j)}$	$1+(c-1)$	$c(r-1)$	$u_{A(i)}=0$, $u_{AB(ij)}=0$	{B} (1項目独立モデル)
M_4	$p_{ij}=p_{i.}p_{.j}$, $F_{ij}=\frac{f_{i.}f_{.j}}{n}$	$\log F_{ij}=u+u_{A(i)}+u_{B(j)}$	$1+(r-1)+(c-1)=r+c-1$	$rc-(r+c-1)=(r-1)\times(c-1)$	$u_{AB(ij)}=0$	{A,B} (相互独立モデル)
M_5	$p_{ij}=\frac{p_{ij}}{n}, F_{ij}=f_{ij}$	$\log F_{ij}=u+u_{A(i)}+u_{B(j)}+u_{AB(ij)}$	$1+(r-1)+(c-1)+(r-1)(c-1)=rc$	0	———	{A,B,AB} (飽和モデル)

このようなモデル化を行うと F_{ij} の（最尤）推定値 $\hat{F}_{ij}$ を求める方法が必要となる。たとえば(1)の場合は，

$$\hat{F}_{ij} = (f_{i\cdot}/n)\cdot(f_{\cdot j}/n)\cdot n = f_{i\cdot} f_{\cdot j}/n \quad (f_{i\cdot}, f_{\cdot j}: \text{周辺度数})$$

で与えられるが，モデルによっては上のように直接 $\hat{F}_{ij}$ を求めることが不可能なケースが出てくる([2]参照)。このため F_{ij} の推定法がいろいろ提案されているが，ここでは Iterative Propotional Fitting を使う。この方式によって推定した $\hat{F}_{ij}$ を使って，

$$\chi_L^2 = 2\sum_i\sum_j f_{ij}\log f_{ij}/\hat{F}_{ij} \quad (\text{尤度比 } \chi^2 \text{統計量}) \qquad (3)$$

あるいは，

$$\chi_G^2 = \sum_i\sum_j (f_{ij} - \hat{F}_{ij})^2/\hat{F}_{ij} \quad (\text{ピアソンの } \chi^2) \qquad (4)$$

によりモデル（仮説）への $\{f_{ij}\}$ の適合性をみる。ここで χ_L^2, χ_G^2 はいずれも仮説の下で漸近的に χ^2 分布に従う。そのときの自由度 d.f. は，

d.f. = [セルの総数] − [あてはめるモデルのパラメータ数]

となる。

2. χ_L^2 の分解性とモデルの一般的記述　いま表1のモデル M_i に対する χ_L^2 を改めて $\chi_L^2\{M_i|(\text{略記法})\}$ で表わす。さらに各項の関連の強さ（寄与度）を C_j ($j = A, B, AB$) で表わす。たとえば項Aは，

$$C_A = \chi_L^2\{M_2|A\} - \chi_L^2\{M_1|\phi\}$$

項 B, AB についてはそれぞれ,

$$C_B = \chi_L^2\{M_3 \mid B\} - \chi_L^2\{M_1 \mid \phi\}$$

$$C_{AB} = \chi_L^2\{M_5 \mid A, B, AB\} - \chi_L^2\{M_4 \mid A, B\}$$

またモデル $M_1 = \{\phi\}$ に対しては $C_0 = \chi_L^2\{M_1 \mid \phi\}$ とおく.

このとき明らかに,

$$\chi_L^2\{M_5 \mid A, B, AB\} = C_0 + C_A + C_B + C_{AB} \ (\equiv 0) \qquad (5)$$

となる. $|C_j|$ は項ごとの寄与の程度を示す量でいずれも漸近的に χ^2 分布に従いその自由度はそれぞれ利用したモデルの間の自由度の差になる. ここにこで χ_L^2 に対して赤池の情報量規準 (AIC) を導入するとモデルの選択が容易にできることは知られている ([5], [8]). モデル M_i に対する AIC は,

$$AIC\{M_i\} = \chi_L^2\{M_i\} - 2(d.f.) \qquad (6)$$

であるから式 (5) にこれを適用してやると, C_j の情報量が考えられる. たとえば, A に対しては,

$$I(C_A) = AIC\{M_2\} - AIC\{M_1\} \qquad (7)$$

となり他の項 B, AB, ϕ についても同様である. したがって (5) から,

$$AIC\{M_5\} = I(C_0) + I(C_A) + I(C_B) + I(C_{AB}) \qquad (8)$$

が成り立つ. この $I(C_j)$ を項 j の <u>項別情報量</u> と名づける. これを利用すると各項のモデルへの寄与の程度を知ることができる. つまり $I(C_j)$ は, モデル $M_1 = \{\phi\}$ が持つ全情報 (このとき

χ_L^2が最大である）を各項がどのように分担したかその寄与の大きさと考えてよい。ここで簡単な例をみる。表2は2元クロス表の例であるが，これに5通りのLLMを適用すると表3の結果を得る。表3をもとにC_jおよび$I(C_j)$を求め整理すると表4となる。これからこの例では項目A, Bの単独効果が大勢を占め交互作用ABの寄与は小さい，つまり独立モデルに近い，ことがわかる。

（表2） 2元クロス表の例（「日本人の国民性」調査から）

B（年令） A（性別）	1	2	3	4	5	$f_{i\cdot}$
1	1096	1051	806	613	645	4211
2	635	706	564	384	380	2669
$f_{\cdot j}$	1731	1757	1370	997	1025	6880

（表3）

モデル	$d.f.$	χ^2_G	χ^2_L	$P\{\chi^2>\chi^2_L\}$	AIC
M_1	9	763.69	753.19	0	735.19
M_2	8	401.15	404.63	0	388.63
M_3	5	354.37	357.80	0	347.80
M_4	4	9.23	9.24	0.055	1.24
M_5	0	0	0	—	—

（表4）

項目：j	C_j	$d.f.$	$I(C_j)$	
$\{\phi\}$	753.19	9	735.19	
A	−348.56	1	−346.56	(−735.19)
B	−395.39	4 (9)	−387.39	
AB	−9.24	4	−1.24	

ところでパラメータ α, β, γ を導入してモデルを $M_{\alpha\beta\gamma}$ とかくと,

$$AIC\{M_{\alpha\beta\gamma}\} = I(C_0) + \alpha \cdot I(C_A) + \beta \cdot I(C_B) + \alpha\beta\gamma I(C_{AB}) \qquad (9)$$

あるいは,

$$\chi_L^2\{M_{\alpha\beta\gamma}\} = C_0 + \alpha \cdot C_A + \beta C_B + \alpha\beta\gamma C_{AB} \qquad (9)'$$

により表1のモデルを一括表現できる. たとえば $\alpha=\beta=\gamma=1$ のとき $M_{111}=M_5$, $\alpha=1, \beta=\gamma=0$ のとき $M_{100}=M_2$ となる. これは項別情報量 $I(C_j)$ を知れば式(9)からすべてのモデルのAICが誘導できることを示している. $I(C_j)$ (または C_j) を求めるためには項目数(クロス表の次元数)を d として 2^d 通りのモデルを考えればよい. $d=2$ では $2^2=4$ 通り, つまり M_1, M_2, M_3, M_5 を知れば M_4 は (9), (9)' から求められる. 一般に d 元のクロス表では 2^d 通りのLLMについて χ_L^2, AIC を求めておけば残りのすべてのモデルの C_j, $I(C_j)$ がえられる. 以上の関係を利用すると次のモデル選択の方式がえられる.

<u>3. モデル選択の方法.</u> χ_L^2 のもつ分解性を利用してモデル選択を行う方式は既に多くの研究者により提案されている. しかしここでは, 前述の項別情報量の性質を利用して客観的に最適モデルの候補を選ぶ1つの方式を提案する. その手順は次の通りである.

<u>手順1°</u> d 元クロス表に対し 2^d 個のLLMを計算する. そして項別情報量 $I(C_j)$ のリストを作る. この時必要となった;

項 k の集合を K とする．さらに1項目（主効果）項の集合を K_0，2項目以上の組み合わせ項（交互作用項）の集合を K_I とする．したがって $K = K_I \cup K_0 \cup \{\delta_0\}$（$\delta_0$ は等確率モデルに対する項．つまり $\delta_0 = 0$ としてよい）．

<u>手順2°</u>　$I(C_0)$，および K_0 内から $I(C_j) < 0$ $(j \neq 0)$ となる項を選んで，これらの項から構成される LLM を初期モデル $M^{(0)}$ とする．

<u>手順3°</u>　$I(C_j) < 0$ となる交互作用項を K_I 内から求めそれを含む LLM を構成するために必要な項目をすべて求める（たとえば $I(C_{AB}) < 0$ のとき A, B, 中の3つの項が必要）．それらの項目が $M^{(0)}$ 内にすでに含まれていれば $M^{(0)}$ にその交互作用項を加え $M^{(0)}$ を $M^{(1)}$ に更新する．以下これを反復（$M^{(i-1)} \to M^{(i)}$）．そしてこのときの項の組み合わせを $K^{(i)}$ とかく．

<u>手順4°</u>　$M^{(i)}$ に採用されていない $I(C_j)$ のうち負のものを K_I の中から1つ選びそれが k であったとする．$K^{(i)} \cup \{k\}$ から構成されるモデルが LLM となるために必要な最少限の項を K_I 内から選ぶ．そして k にこれを追加してえられるモデルを $M_+^{(i)}$ とする．

<u>手順5°</u>
$$\begin{cases} AIC(M_+^{(i)}) \geq AIC(M^{(i)}) \text{ のとき}, & M^{(i)} \to M^{(i+1)}, \\ AIC(M_+^{(i)}) < AIC(M^{(i)}) \text{ のとき}, & M_+^{(i)} \to M^{(i+1)}, \end{cases}$$

とする．$M^{(i+1)}$ を構成する項目 $K^{(i+1)}$ に変化がみられなくな

ったとき収束と判定しそのモデルを解として採用する.

なお, 手順4°でkを1つではなく複数個の項からなる項の部分集合 $K_S \subset K_I$ とすると最小AICを持つモデルが得られる. この場合計算すべきLLMの総数は 2^M ($M=2^d-d-1$) 通りとなるが実際に各種のデータで実験するとこの上限まで到達するケースはまれで, かなり早い段階で解に収束する. ここで, 手順1°～5°に従ってモデル選択を行う方式を逐次選択方式, 手順4°を上述のように変更した方式を組み合わせ的選択方式と仮称する. 次にこの2つの方式によるモデル選択の例をみる. 数値はGoodmanが引用したRies and Smithの例を使う. これは表5のような (3×2×2×2) 次の4元クロス表である. 内容は洗剤のブラインド・テストの結果で, 各項目は表に示す通りである.

(表5) Goodmanの例

Water softness	Brand preference	Previous user of M High temperature	Previous user of M Low temperature	Previous nonuser of M High temperature	Previous nonuser of M Low temperature
Soft	X	19	57	29	63
	M	29	49	27	53
Medium	X	23	47	33	66
	M	47	55	23	50
Hard	X	24	37	42	68
	M	43	52	30	42

A : Water Softness
B : Previous users
C : Temperature
D : Brand Preference

(i) 総当り法による場合　初めに総当り法により4元の場合のモデルの総数167通りについてすべて計算を行う。表6は結果の要約である. χ^2_L にもとづくAICの大きさで上から10個のモデルを挙げてある. これを記号 M_i ($i=1,2,\cdots,10$) で略記し

て観察すると，次のことがわかる．

i) 1項目（1因子項）として意味があるものはB, C, Dである．

ii) 次に2因子項（1次交互作用項）としてBC, BDが寄与が高い．　結局以上の項を含む$M_1=\{B, C, D, BC, BD\}$がデータを説明する中心的なモデルである．

iii) 引き続いて現われる$M_2, M_3, \cdots$のいずれもM_1を核としてそれに他の因子項が出入りするがいずれもAICの改良にそれほど大きく寄与しない．つまりM_1にくらべてそれほど情報を高める項は少ないとみてよい．

iv) Goodman流の逐次選択方式によると$M_4=\{A, B, C, D, BC, BD, AD\}$が選ばれる．しかし表5の度数表から明らかなように項目Aの水準ごとの度数の差はほとんどない．つまりAの寄与は少ない，にもかかわらずAが取り込まれるのは，Goodmanの方式が独立モデル$\{A, B, C, D\}$を基点に探索を行うことに帰因する．したがって単独効果の強いDの影響を受けて交互作用としての効きの少ないADまでモデル内に取り込むことになる．

v) Brownのscreening-effect方式によるとM_6が選ばれる．彼の方式によると有意水準5%でM_1, M_4, M_6, M_{10}の4つのモデルが最終候補として残り，項の少ないモデルが「節約の原理」からみて望ましいからM_6が良い，としている．

結局，AICを利用すると総当り法のときでも客観的に，ほぼ

妥当とみられるモデルを選出することができる。

b) 逐次選択方式による場合　次にここで提案の逐次選択方式による結果をみる。まず $2^4=16$ 通りの LLM について計算を行い表7を得る。これをもとに $I(i,j)$ の表を作り表8を得る。さらに計算をすすめ解を得るまでの途中経過を一覧にしたものが表9である。ステップの欄で○印をつけた番号の部分が逐次選択方式の途中経過であり 1～12 ですべてを行う場合が組み合わせ的方式による場合に相当する。また各ステップで取り込んだ項の組み合わせが次の欄に示してある。判定欄の○印は吟味の結果合格と判断されて採用されたモデルで他は棄却される。残った2つをくらべてステップ5の結果を採用し（というよりは先に現われた候補よりこちらがよいので残った）これを解とする。この解は明らかに総当り法で求めた最適モデル M_1 に相当する。ここで注意を要することは実用上は1つのモデルを選択することが主目的ではない、ということである。むしろ意味のある説明力の高い項から構成される基本的なモデルをまず知り，そのモデルの近傍にあるいくつかの類似のモデルを調べることが要求される。この例ではまず $M_1=$ {B, C, D, BC, BD} を知り次にこれを核として項別情報量の表とにらみあわせながら他のモデルを探索すればよい。総当り法の表6をみても M_1 を中心にしてこれに高次の

交互作用項が出入りしていることがわかる。なおこの例では逐次選択方式も組み合わせ的方式も同じ解に到達したが下欄に指摘のように一般にはそうなるとは限らない。さらに計算

(表6) Goodman の例 (総当り法によるとき)

	MODEL	DEGREES OF FREEDOM	LIKELIHOOD RATIO CHISQUARE	(PROBABILITY)	AIC BASED ON LIKELIHOOD RATIO STATISTIC
M_1	B, C, D, BC, BD	18	18.487071	0.4240	-17.512929
M_2	B, C, D, BC, BD, CD, BCD	16	15.007456	0.5241	-16.992544
M_3	B, C, D, BC, BD, CD	17	17.795417	0.4019	-16.204583
M_4	A, B, C, D, AD, BC, BD	14	11.886487	0.6154	-16.113513
M_5	A, B, C, D, AD, BC, BD, CD, BCD	12	8.406872	0.7526	-15.593128
M_6	B, C, D, BD	19	22.848672	0.2441	-15.151328
M_7	A, B, C, D, AD, BC, BD, CD	13	11.194833	0.5945	-14.805167
M_8	B, C, D, BC, CD	18	21.595569	0.2504	-14.404431
M_9	A, B, C, D, BD, CD	16	17.985591	0.3247	-14.014409
M_{10}	A, B, C, D, AD, BC	15	16.248088	0.3658	-13.751912

(表7) Goodman の例 (16通りの LLM の計算結果)

MODEL SEQ NUM	モデル	DEGREES OF FREEDOM	LIKELIHOOD RATIO CHISQUARE	(PROBABILITY)	AIC BASED ON LIKELIHOOD RATIO STATISTIC
1	φ (等確率モデル)	23	118.626936	0.0000	72.626936
2	A	21	118.125456	0.0000	76.125456
3	B	22	118.563443	0.0000	74.563443
4	C	22	116.705690	0.0000	72.705690
5	D	22	45.414876	0.0024	1.414876
6	A, B, AB	18	117.666664	0.0000	81.666664
7	A, C, AC	18	115.129108	0.0000	79.129108
8	A, D, AD	18	38.814292	0.0030	2.814292
9	B, C, BC	20	96.060731	0.0000	56.060731
10	B, D, BD	20	40.989782	0.0037	0.989782
11	C, D, CD	20	42.240527	0.0026	2.240527
12	A, B, C, AB, AC, BC, ABC	12	88.868779	0.0000	64.868779
13	A, B, D, AB, AD, BD, ABD	12	33.923803	0.0007	9.923803
14	A, C, D, AC, AD, CD, ACD	12	32.889115	0.0010	8.889115
15	B, C, D, BC, BD, CD, BCD	16	15.007456	0.5241	-16.992544
16	A, B, C, D, ···· (飽和モデル)	0	---	---	---

効率の点を比較すると，この例では，逐次あるいは組み合わせ的方式はCPU時間で総当り法の約1/10に節約できた．5元の場合などはLLMの総数は数千通りとなるが，我々の方式では$2^5=32$通りのLLMから項別情報量を求め逐次選択を行えばよいのでかなりの節約となる．

(表8) 項別情報量の表

ADDITIONAL TERM: j		DEGREES OF FREEDOM	EFFECT OF AIC: $I(C_j)$
0	Φ	23	72.6269
1	A	2	3.4985
2	B	1	1.9365
3	C	1	0.0788
4	D	1	-71.2121
12	AB	2	3.6047
13	AC	2	2.9249
14	AD	2	-2.0991
23	BC	1	-18.5815
24	BD	1	-2.3616
34	CD	1	0.7469
123	ABC	2	-1.2201
124	ABD	2	3.9299
134	ACD	2	2.3243
234	BCD	1	-0.2265
1234	ABCD	2	4.0294

この Goodman の例の他に，何種かのケースについて同様の実験を行ったがいずれも適当な答が得られた．とくに，実際の調査データに対して適用すると，考えている程，複雑でない，つまりそれほど高次の交互作用項までを必要としない比較的単純なモデルで説明できるという傾向がみられる．

(表9)

ステップ	モデル	取り入れた項	判定	d.f.	χ^2_L	AIC
①	D	D		22	45.414876	1.414876
②	A, D, AD	AD		18	38.814292	2.814292
③	B, C, D, BC	BC	○	19	22.848672	-15.151328
4	A, B, C, D, AD, BC	AD		15	16.248088	-13.751912
⑤	B, C, D, BC, BD	BD	○	18	18.487071	-17.512929
6	A, B, C, D, AD, BC, BD	AD		14	11.886487	-16.113513
⑦	A, B, C, D, AB, AC, BC, BD, ABC	ABC		10	11.295118	-8.704882
⑧	B, C, D, BC, BD, CD, BCD	BCD		16	15.007456	-16.992544
9	A, B, C, D, AB, AC, AD, BC, BD, ABC	AD, ABC		8	5.196014	-10.803986
10	A, B, C, D, AD, BC, BD, CD, BCD	AD, BCD		12	8.406872	-15.593128
11	A, B, C, D, AB, AC, BC, BD, CD, ABC, BCD	ABC, BCD		8	7.815503	-8.184497
12	A, B, C, D, AB, AC, AD, BC, BD, CD, ABC, BCD	AD, ABC, BCD		6	1.716399	-10.283601
(解)	B, C, D, BC, BD	———		18	18.487071	-17.512929

4. 坂元モデルとの関連　AICを利用したモデル評価の方式として坂元らのアプローチがある[4]. そこに取り上げられた例を用いて坂元モデルとLLMの関連を調べた。データは次の項目，水準からなる(2×2×4×4)の4元クロス表である。

A：(質問)「昔労は男に多いか女に多いか?」A_1=男に，A_2=女に

B：(性別)・B_1=男 ，B_2=女

C：(年令) C_1=20才代，C_2=30才代，C_3=40才代，C_4=50才以上

D：(地域) D_1=6大都市，D_2=20万人以上，D_3=20万人以下，D_4=町村

これは統数研で定期的に実施されている「日本人の国民性」のデータの一部である。[4]に掲載のデータにもとづいて，まず総当り法で167通りのLLMを求めAICの値の小さいほうから10個のモデルを選ぶと表10となる。

(表10)

	モデル	AIC
M_1	A, B, C, D, AB, CD	-56.642
M_2	A, B, C, D, AB, CD, AD	-56.486
M_3	A, B, C, D, AB, CD, AC	-55.733
M_4	A, B, C, D, BD, AB, CD, AD	-55.423
M_5	A, B, C, D, AB, CD, BD	-55.064
M_6	A, B, C, D, AB, CD, AC, AD	-54.805
M_7	A, B, C, D, AB, CD, AC, BD	-54.277
M_8	A, B, C, D, AB, CD, AC, BD, AD	-53.732
M_9	A, B, C, D, AB, CD, BC	-50.900
M_{10}	A, B, C, D, AB, CD, BC, AD	-50.722

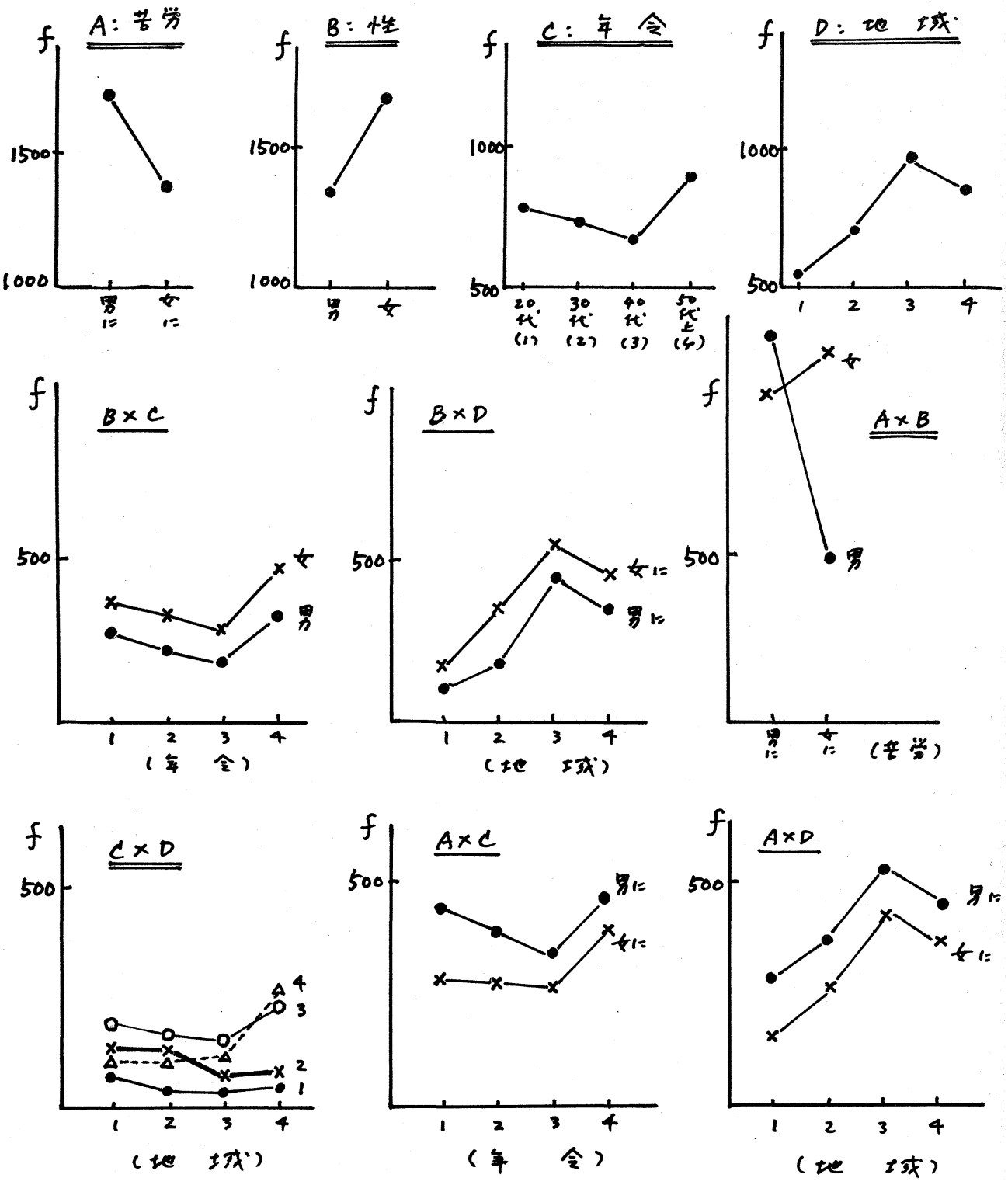
これをみると，AICの大きさではいずれのモデルも大差ない。しかもM_1を除く残りのモデルに，M_1がすべて含まれている。同じデータを我々の逐次選択法により計算すると反復回数

が1回でM_1に到達する．結局このクロス表を特徴づける基本的なモデルは，(性)+(年令)+(地域)+(質問)+(性×質問)+(地域×年令)あるいは$P_r\{$性，質問$\}\cdot P_r\{$年令，地域$\}$と表わせる独立モデルの一種であることが予想される．「性別」と「質問」，「年令」と「地域」の間にそれぞれ交互作用があるといっているのであるから，これを観察するために図を書いてみる[図1]．図中で縦軸は各水準ごとの度数である．最上段の4図は1因子効果（項目ごとの単独効果）であるが，これをみるといずれの項目も水準間の差が認められる．したがってこれら4項目ともモデルに含まれておかしくない．次に，中，下段の2因子項（1次交互作用項）のクロス情報を観察すると確かにA×B，C×Dで強い交互作用がみられる．それに続いてA×C，A×Dあたりがわずかに交絡現象があるがほとんど問題になるほどの差ではない．つまりモデルに積極的に取り入れる必要はないと思われる．こうしてみると，選んだモデル$M_1=\{A,B,C,D,AB,CD\}$はクロス表を説明するに十分な情報を持っているとみてよい．

いまこれを実用の見地から解釈してみると次の知見がえられる．

(i) A×B（「質問」×「性別」）の情報から，質問に対する回答傾向が男，女によって全く異ることがわかる．

(ii) C×D（「年令」×「地域」）の情報から，



(図1) 国民性データの分布

a) 20代では地域差による意見の違いは少い.

b) 都市ブロック(地域Ⅰ3)においては20代→40代の順に,苦労が多いと感じているが50代以上では逆に減少する,

c) しかし,地域4(町村)では,年令と苦労の関係には順序性があり,年令が高まると苦労も多いと感じている.

つまり50代以上では地域差と質問の傾向の間に交絡化現象があることを示している.この例ではA×Bよりもc×Dのやや複雑な構造の意見傾向を抽出できたことに特徴がある.しかも,数多くの項の組み合わせの中からその水準間の絡みをよく反映するような項を客観的に選べたことに意味がある(3因子項以上の組み合わせ項がM_1~M_{10}以内に現われていないことに注目する必要がある).

さて次に,坂元の得た結果とLLMの関連を概述する.坂元の結果の1部を示すと表11の(Ⅱ)欄となる.ここでの記法は坂元のものを用いた.[4]では表11の7つのモデルを検討し,このうちM_4^*を選んでいる.ところでこの方式をLLMで表わすと(Ⅱ)欄になる.この欄で(BC),(CD),(BCD)などの記号は文字の示す項をプールして新しいカテゴリーを作りそれを1項目として扱うことを意味している.たとえばcが4水準,Dが2水準のとき,つまり,$C=(C_1, C_2, C_3, C_4)$, $D=(D_1, D_2)$のとき,(CD)とは(C_iD_j) $(i=1,\cdots 4; j=1,2)$の$2\times4=8$水準の新しい1項目

を考えることに相当する。つまり坂元モデルは多次元クロスの情報をセルの数を同数におさえて3元クロス以下のクロス表に縮約した上でLLMを考えることと同等である。さらに調べると、(II)欄のモデルは実は(I)欄のようなLLMと同等である。実際、(I)～(III)の各モデルの χ^2_L, AICは完全に一致する。そして坂元の提案したモデルは次のように要約できる。

(1) 多次元(4元以上)のクロス情報を3元以下のクロス表に縮約して分析する方式である。

(2) 縮約により得られた3項目をX, Y, Zと書くと坂元モデルでは、16通りのLLMのうち{X, Y, Z, XY, YZ}の形に表記できるモデルに相当している(表11の(III)欄の表現を参照)。

(3) 2元に縮約されたとき、それをX, Yとかくと5通りのLLMのうちの{X, Y}すなわち独立モデルに相当する。

(表11) LLMと坂元モデルの対比

	(I) 対数線形モデル	(II) (I)と同等のモデル	(III) (I), (II)と同等の坂元モデル
M_1^*	A, B, C, D, AB, AC, BC, BD, CD, ABC, BCD	A, (BC), D, A·(BC), (BC)·D	$p(i_1,i_2,i_3)\cdot p(i_2,i_3,i_4)/p(i_2,i_3)$
M_2^*	A, B, C, D, AB, AD, BC, BD, ABD, BCD	A, (BD), C, A·(BD), (BD)·C	$p(i_1,i_2,i_4)\cdot p(i_2,i_3,i_4)/p(i_2,i_4)$
M_3^*	A, B, C, D, AC, AD, BC, BD, CD, ACD, BCD	A, (CD), B, A·(CD), (CD)·B	$p(i_1,i_3,i_4)\cdot p(i_2,i_3,i_4)/p(i_3,i_4)$
M_4^*	A, B, C, D, AB, BC, BD, CD, BCD	A, B, (CD), AB, B·(CD)	$p(i_1,i_2)\cdot p(i_2,i_3,i_4)/p(i_2)$
M_5^*	A, B, C, D, AC, BC, BD, CD, BCD	A, C, (BD), AC, C·(BD)	$p(i_1,i_2)\,p(i_2,i_3,i_4)/p(i_3)$
M_6^*	A, B, C, D, AD, BC, BD, CD, BCD	A, D, (BC), AD, D·(BC)	$p(i_1,i_4)\cdot p(i_2,i_3,i_4)/p(i_4)$
M_7^*	A, B, C, D, BC, BD, CD, BCD	A, (BCD)	$p(i_1)\cdot p(i_2,i_3,i_4)$

(注) (III)欄の添字は、$i_1=A$, $i_2=B$, $i_3=C$, $i_4=D$ と対応する。

以上からみて項元モデルによるとより高次の項を含むモデルとなる。表11にみるように2次の交互作用項まで現われる。また吟味対象となるモデルは、いわゆる独立あるいは条件付独立モデルである。また項目の選択の順序あるいは指定に依存するので、ある項目を特定化した上で（項元のいうprediction）モデルの評価を行う。例のデータではA（質問）に対して、残りのB, C, Dの項目がどう効きかけるか、と考えるので、表の(四)欄の形のモデルをすべて吟味するわけではない（表記法は添字について対称でない）。つまりAからみて寄与の高い項目を探索するのでC×Dのような情報は現われることがない。

<u>まとめ</u>　χ^2_Lの分解性をAICのそれにおきかえ、また項別情報量を考えることで、LLMにおけるモデル選択を容易にする1つの方式を提案した。従来の方式が有意水準設定に依なする検定方式であるため結果があいまいになるきらいがあったが、ここの方式ではかなり客観的に問題を扱うことができる。実用的見地からは1つのモデルを選ぶことよりも、クロス表の持つ項目間の代表的な関連性のしくみを観察することに主な解析目的があるので、ここで提案した方式で十分実用に耐えると思われる。実際にいくつかの適用例でほぼ納得のゆく答をえている。また総当り法の結果と、逐次あるいは組み合わせ的方式のそれとが一致することは最大の利点である。

今後の問題として，水準の与え方，セルの度数の大きさ，標本数などがどのように影響するか調べることや，χ_L^2の近似の程度が結果におよぼす影響を確かめる必要がある。

≪参考文献≫

[1] Brown, M.B. (1976): Screening Effects in Multidimensional Contingency Tables, Appl. Statist. Vol.25, No.1, 37-46.

[2] Fienberg, S.E. (1977): The Analysis of Cross-Classified Categorical Data, MIT Press.

[3] Goodman, L.A. (1971): The Analysis of Multidimensional Contingency Tables, Technometrics, Vol.13, No.1, 33-61.

[4] Sakamoto, Y., Akaike, H. (1978): Analysis of Cross-Classified Data by AIC, Ann. Inst. Statist. Math., vol.30, No.1, partB.

[5] 大隅 昇 (1977): 対数線形モデルによる多次元分割表の評価 日本大学理工学部学術講演会論文集, 891-894

[6] 柳澤幸雄, 大隅 昇 (1978): 同上 (その2), 日本大学理工学部学術講演会論文集, 664-666

[7] 水野欽司, 大隅 昇 (1976): 多重クロス集計の一評価方法について 日本行動計量学会第5回総会

[8] 吉澤 正 (1976): 入学選抜試験における選抜科目の解析, 山梨大学工学部研究報告, vol.27, 107-115.