

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、これにもとづ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わせ $A_i B_j$, ある i, j 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 $\{f_{ij}\}$ で表す。このとき LLM とは f_{ij} の期待値 F_{ij} ($i=1, 2, \dots, r; j=1, 2, \dots, c$) 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。以下モデル (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、

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

$$\therefore \text{て、 } u = \frac{\sum_i \sum_j E_{ij}}{rc},$$

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

とかける。なお、 $\sum_i u_{A(i)} = \sum_j u_{B(j)} = 0$ 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、

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

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)	(b)	(c)	(d)	(e)	(f)
	モデルの構造	対数線形モデル	パラメータ数	d. f.	帰無仮説: Ho	略記法
M_1	$p_{ij} = \frac{1}{rc}, F_{ij} = \frac{n}{rc}$	$\log F_{ij} = u$	1	$rc-1$	$u_{A(i)}=0$ $u_{B(j)}=0$ $u_{AB(ij)}=0$	{ \emptyset } (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} }
M_4	$p_{ij} = p_{i.} \cdot p_{.j},$ $F_{ij} = \frac{f_{i.} \cdot f_{.j}}{n}$	$\log F_{ij} = u + u_{A(i)} + u_{B(j)}$	$1+(r-1) + (c-1) = r+c-1$	$rc-(r+c-1) = (r-1) \times (c-1)$	$u_{AB(ij)}=0$	{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、

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

で与えられるが、モデルによ、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くるが、ここでは Iterative Proportional Fitting を使う。この方式によ、て推定した \hat{F}_{ij} を使、て、

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

あるいは、

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

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、

$$\text{d.f.} = [\text{セルの総数}] - [\text{あてはめるモデルのパラメータ数}]$$

となる。

2. χ_L^2 の分解性とモデルの一般的記述 11 表 1 のモデル

M_i に対する χ_L^2 を改めて $\chi_L^2 \{M_i | \}$ (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、

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

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

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

これをまとめると、

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

となる。 $|C_j|$ は項 j の寄与の程度を示す量であり、これも漸近的に 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。これによって 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、

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

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、

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

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、

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

が成り立つ。これを $I(C_j)$ を項 j の 項別情報量 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、つまり $I(C_j)$ は、モデル $M_1 = \{\phi\}$ が持つ全情報 (これを

χ^2_L 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、ここに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。ここからこの例では項目A, Bの単独効果が大半を占め交互作用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(年齢)	1	2	3	4	5	$f_{i\cdot}$
A(性別)						
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.	$I(C_j)$
{ ϕ }	753.19	9	735.19
A	-348.56	1	-346.56
B	-395.39	4	-387.39
AB	-9.24	4	-1.24

} (-735.19)

$\nu = 3$ 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,

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

あるいは,

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

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 通りのモデルを考へればよい。 $d=2$ では $2^2=4$ 通り、つまり M_1, M_2, M_3, M_4 を知れば M_5 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。

3. モデル選択の方法 χ_L^2 の持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。

手順1° 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。このとき必要となる;

項 j 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, $K = K_I \cup K_0 \cup \{j_0\}$ (j_0 は等確率モデルに対する項, つまり $j_0 = 0$ としてよい).

手順 2° $I(C_0)$, 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。

手順 3° $I(C_j) < 0$ 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 $I(C_{AB}) < 0$ のときは 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 ($M^{(i-1)} \rightarrow M^{(i)}$).

そして n とする項の組み合わせを $K^{(i)}$ とかく。

手順 4° $M^{(i)}$ に採用してない $I(C_j)$ のうち負のもの α を K_I の中から 1 つ選ぶ。それが k であるとする。 $K^{(i)} \cup \{k\}$ 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 にこれを追加して与えるモデルを $M_+^{(i)}$ とする。

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

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

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 ($M=2^d-1$) 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。そこで、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 $(3 \times 2 \times 2 \times 2)$ 次の4元クロス表である。内容は洗剤のブライント・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		Previous nonuser of M	
		High temperature	Low temperature	High temperature	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

(a) 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 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 ($i=1, 2, \dots, 10$) で略記し

を觀察すると、次 $a = k$ 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 a 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) 引き続き現われる M_2, M_3, \dots のうちも M_1 を核としてそれぞれに他の因子項が出入りするが、うちも AIC の改良に多少の大きく寄与しない。つまり M_1 にくらべてそれぞれ情報量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 $M_4 = \{A, B, C, D, BC, BD, AD\}$ が選ばれる。しかし表5の度数表から明らかのように項目 A の水準ごとく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 $\{A, B, C, D\}$ を基準に探索を行う $\Rightarrow k$ 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 a の値が少く AD までモデル内に取り込む $\Rightarrow k$ になる。
- v) Brownの Screening-effect方式による M_6 が選ばれる。彼の方式による有意水準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、としてゐる。

結局、 AIC を利用すると総当たり探索 $\Rightarrow k$ までを客観的に、ほぼ

妥当とみられるモデルを選出することが出来る。

b) 逐次選択方式による場合 次に二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。二をもちに $I(d_j)$ 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0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ま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表9の欄に示してある。判定欄の0印は吟味の結果合格と判断し採用したモデルで他は棄却される。残り2つをくくってステップ5の結果を採用し(このよりは先に現れた候補より53がよいので残り2)二を解とする。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。二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 $M_1 = \{B, C, D, BC, BD\}$ を知り次に二を核として個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 を中心にして二に高次の

交互作用項が出入りしてゐる = と分かる。なおこの例では
逐次選択方式も組み合わせ的方式と同じ解に到達したが手順
に指摘のよう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)

MODEL	DEGREES OF FREEDOM	LIKELIHOOD RATIO CHISQUARE (PROBABILITY)	AIC BASED ON LIKELIHOOD RATIO STATISTIC
M1 B, C, D, BC, BD	18	18.487071 0.4240	-17.512929
M2 B, C, D, BC, BD, CD, BCD	16	15.007456 0.5241	-16.992544
M3 B, C, D, BC, BD, CD	17	17.795417 0.4019	-16.204583
M4 A, B, C, D, AD, BC, BD	14	11.886487 0.6154	-16.113513
M5 A, B, C, D, AD, BC, BD, CD, BCD	12	8.406872 0.7526	-15.593128
M6 B, C, D, BD	19	22.848672 0.2441	-15.151328
M7 A, B, C, D, AD, BC, BD, CD	13	11.194833 0.5945	-14.805167
M8 B, C, D, BC, CD	18	21.595569 0.2504	-14.404431
M9 A, B, C, D, BD, CD	16	17.985591 0.3247	-14.014409
M10 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	中 (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	---	---

効率の差を比較すると、この例では、逐次組み合わせ的方式はCPU時間で総当たり法の約1/20に節約できた。5元の場合には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。このGoodmanの例の他に、

(表8) 項別情報量の表

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

何種かの $4-2$ に同じ

様の実験を行、たかゝりお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、考えられている程、複雑でない、つまりそれほど高次の交互作用

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	df	χ^2	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のアプローチがある[4]。そり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)「苦勞は男に多いか女に多いか?」, A_1 =男, A_2 =女に

B: (性別) B_1 =男, B_2 =女

C: (年齢) 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 =20万以上, D_3 =20万以下, D_4 =町村

これは統教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16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

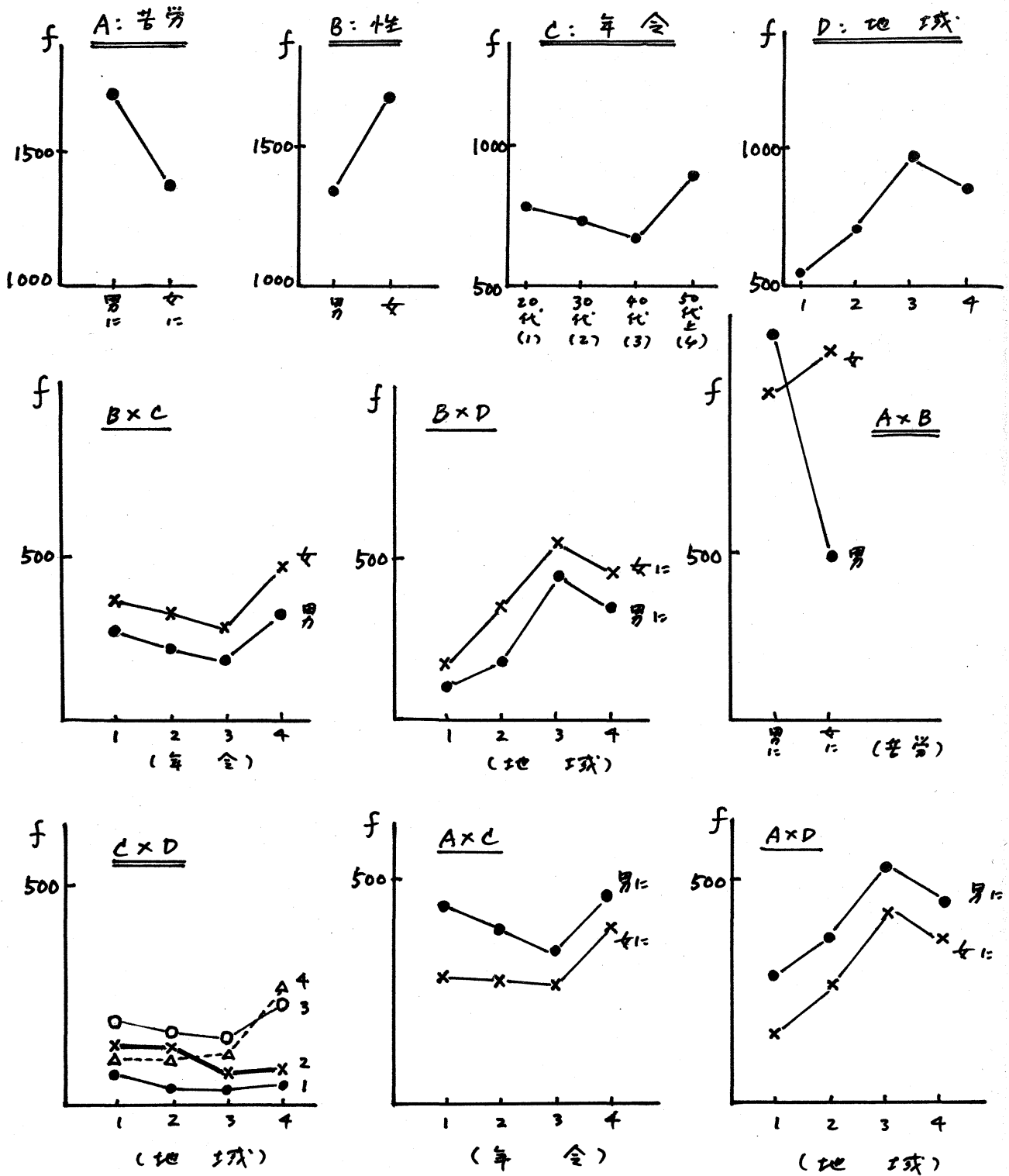
これをみると、AICの大きいではいず、このモデルも大差ない。しかも M_1 を除く残りのモデルに、 M_1 がすべて含まれている。同じデータを我々の逐次選択法により計算すると反復回数

が1回で M_1 に到達する。結局は α クロス表を特徴付ける基本的なモデルは、(性)+(年齢)+(地域)+(賃肉)+(性 \times 賃肉)+(地域 \times 年齢) あるいは $P_1\{\text{性, 賃肉}\} \cdot P_2\{\text{年齢, 地域}\}$ 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「賃肉」、「年齢」と「地域」の間にはそれぞれ交互作用があることが、この α 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 α の度数である。最上段の4図は1因子効果(項目 α の単独効果)であるが、これをみると α の項目も水準間の差が認められる。したがって α の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 $A \times B, C \times D$ の強い交互作用がみられる。さらに $A \times C, A \times D$ あたりには弱く交互作用現象があるがほとんど問題にはならない。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おりよ。

いま α を実用に見地から解釈してみると次の知見がえらわれる。

i) $A \times B$ (「賃肉」 \times 「性別」) の情報から、賃肉に対する回答傾向が男、女によっても全く異なることがわかる。

ii) $C \times D$ (「年齢」 \times 「地域」) の情報から、



(图1) 国民性因子分布

a) 20代では地域差による意見の違いは少ない。

b) 都市プロセス (地域1~3) においては20代→40代の頃に、苦勞が多くと感じてくるが50代以上では逆に減少する。

c) しかし、地域4 (町村) では、年令と苦勞の関係には順序性があり、年令が高まるにつれて苦勞も多くと感じてくる。

つまり50代以上では地域差と傾向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終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 $M_1 \sim M_{10}$ 様に理中化してあることに注目する必要がある)。

さて次に、項元の得た結果とLLMの関連を概述する。項元の結果の一部を示すと表11の(Ⅱ)欄となる。この記法は項元のものを採用した。[4]では表11の7つのモデルを検討し、このうち M_4^* を選んでくる。これは3次元の式を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 $(C_i D_j)$ ($i=1, \dots, 4; j=1, 2$)の $2 \times 4 = 8$ 水準の新しい1項目

を考慮することを相当する。つまり項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を考慮することを同等である。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)~(IV)の各モデルの χ^2 , AIC は完全に一致する。
 そして項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 と書くと項元モデル
 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ゐる(表11の(II)欄の表記を参照)。

(3) 2元は縮約しただけで、それを X, Y とかくと5通りのLLM
 のうちの $\{X, Y\}$ すら必ず独立モデルに相当する。

(表11) LLMと項元モデルの対比

	(I) 対数線形モデル	(II) (I)と同等のモデル	(III) (I), (II)と同等の 項元モデル
M_1^*	A, B, C, D, AB, AC, BC, BD, CD, ABC, BCD	A, (BC), D, A·(BC), (BC)·D	$p(u_1, u_2, u_3) \cdot p(u_2, u_3, u_4) / p(u_2, u_4)$
M_2^*	A, B, C, D, AB, AD, BC, BD, ABD, BCD	A, (BD), C, A·(BD), (BD)·C	$p(u_1, u_2, u_4) \cdot p(u_2, u_3, u_4) / p(u_2, u_4)$
M_3^*	A, B, C, D, AC, AD, BC, BD, CD, ACD, BCD	A, (CD), B, A·(CD), (CD)·B	$p(u_1, u_3, u_4) \cdot p(u_2, u_3, u_4) / p(u_3, u_4)$
M_4^*	A, B, C, D, AB, BC, BD, CP, BCD	A, B, (CD), AB, B·(CD)	$p(u_1, u_2) \cdot p(u_2, u_3, u_4) / p(u_2)$
M_5^*	A, B, C, D, AC, BC, BD, CD, BCD	A, C, (BD), AC, C·(BD)	$p(u_1, u_2) \cdot p(u_2, u_3, u_4) / p(u_3)$
M_6^*	A, B, C, D, AD, BC, BD, CD, BCD	A, D, (BC), AD, D·(BC)	$p(u_1, u_4) \cdot p(u_2, u_3, u_4) / p(u_4)$
M_7^*	A, B, C, D, BC, BD, CD, BCD	A, (BCD)	$p(u_1) \cdot p(u_2, u_3, u_4)$

(注) (III) 欄の添字は, $u_1 = A, u_2 = B, u_3 = C, u_4 = D$ と対応する。

以上からみて項元モデルによる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理の中する。また吟味対象となるモデルは、"の中する独立ある"は条件付独立モデルである。また項目の選択の順序ある"は指定1依存する α で、ある項目を特定化した上で(項元 α のpredictor)モデルの評価を行う。例のデータではA(値内)に対して、残りのB, C, Dの項目がどう働きかけるか、と考える α 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"で対応でない)。つまりAからみて高次の項目を探索する α でC×Dのような情報も理の中するわけではない。

5.まとめ X_L^2 の分解性をAICのそれに置き換え、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1依存する検定方式であるため結果が"ま"に"なるま"があたが、"="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1つのモデルを選ぶことよりも、クロス表の持つ項目間・代表的な関連性のしくみを観察するに主な解析目的がある α で、"="で提案した方式で十分に耐えらると思わゆる。実際にいくつかの適用例でほぼ納得のゆく答をえて"る。また総当り法の結果と、逐次ある"は組み合わせ的方式のそれとが一致する"は最大の利点にある。

今後の問題として、水準 α の与え方、セル内の度数 a の大小、標本数 n などがどのように影響するか調べることも、 χ^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, part B.
- [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 (No. 2),
日本大学理工学部学術講演会論文集, 664-666
- [7] 水野 鋭司, 大隅 昇 (1976): 多重クロス集計 α -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.