社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

2値データの分割表における逐次変数選択法

大野　　学†　　垂水　共之††

† 岡山大学大学院自然科学研究科　〒 700-8530 岡山県岡山市津島中 3–1–1
†† 岡山大学アドミッションセンター　〒 700-8530 岡山県岡山市津島中 3–1–1
E-mail: †ohno@ems.okayama-u.ac.jp , ††tarumi@ems.okayama-u.ac.jp

あらまし　本稿は, [4], [5] で提案された分割表データの変数選択に関して、2 値データの場合の問題を扱うものである．目的変数に対して，より良い説明変数の組合せを選択する問題について，セルの条件付確率モデルの AIC を利用した逐次変数選択法の 1 つである変数増減法を [4], [5] は提案した．データのカテゴリー数が 3 個以上の場合には、その方法で良好に動作する．本稿では，2 値データで選択されるべき説明変数の次数が低い場合，変数増減法では高い精度で変数を選択できないことを示し，その解決法として，ブートスラップ法を用いることを提案する．ブートストラップ法によって選択する変数の信頼性を評価し，信頼性の低い無駄な変数を選択しないようにする．この方法の有効性を数値結果によって示す.

キーワード　AIC, 逐次選択法, 2 値データ, ブートストラップ

Stepwise Variable Selection Method for Contingency Tables on Binary Data

Manabu OHNO† and Tomoyuki TARUMI††

† Graduate School of Natrual Science and Technology, Okayama University　Tsushimanaka 3–1–1, Okayama-shi, 700–8530 Japan
†† Addmission Center, Okayama University　Tsushimanaka 3–1–1, Okayama-shi, 700-8530 Japan
E-mail: †ohno@ems.okayama-u.ac.jp , ††tarumi@ems.okayama-u.ac.jp

Abstract　This paper focused a problem in case of binay data about the variable selection method of contingency table proposed in [4], [5]. [4], [5] proposed a stepwise variable selection method which used AIC of conditional probability model for searching best explanatory variable. We show a proplem that the method cannot select a correct variable with high precision, when correct explanatory variable is low dimension on binary data. For the problem, this paper suggests a method that use bootstrap method. We evaluate reliability of a variables to select using bootstrap method and do not select a variableset that is low waste of reliability. We show the effectiveness of our method by numerical experiment results.

Key words　AIC, Stepwise Method, Binary Data, Bootstrap

1. はじめに

近年のインターネットの技術革新に伴い，消費者調査から社会調査といった広い分野の調査がネットを介して行われている．そのため，インターネットの関連技術によって安価で高速に大量のサンプルを回収することにが可能になった．また，従来の紙を使った調査では，さまざまなコストの問題で不可能だった調査を可能にしている．Web 調査以外の分野では，流通業に導入された POS システムによって，顧客がどの商品を買ったかという購買履歴データが大量に安価に蓄積され，企業の意思決定や商品戦略に役立てられている．これらのデータには，カテゴリカルデータが多く含まれており，その分析方法の 1 つとして，対数線形モデルが挙げられる．また，購買履歴データなどの商品を買った・買わなかったという 2 値データを扱った分析法としては，アソシエーション分析がよく用いられている．

カテゴリカルデータは，分割表の形で分析される場合が多い．分割表のセルの確率モデルのモデル分析を [4], [5] が提案した．その確率モデルに関して AIC を導出し，比較することによって有効な分割表を選択する方法を提案している．特に，分割表において目的変数をより有効に説明する説明変数の組合せを探

表 1　分割表の例 (3 元分割表)

		X_0	
X_1	X_2	1	2
1	1	16	22
	2	34	44
2	1	78	32
	2	33	54
total	1	94	54
	2	67	98

す問題に焦点を当てており，その目的において適合度検定や対数線形モデルに対しての優位性を述べている．

その説明変数の選択法として, [4], [5] は，逐次変数選択法の 1 つである変数増減法の適用を提案している．変数増減法は，目的変数に対して有効な説明変数の組合せを選択する方法として大変有効な方法である．しかしながら，2 値データの場合に，さまざまな問題が起こることがわかった．その問題点と解決法を次章以降で述べる．

2.　AIC による逐次変数選択法

2.1　条件つき確率モデルの AIC

分割表の確率モデルの AIC は, [4], [5] より，次のように定式化される．データは，$p+1$ 個の変数 $X_0, \cdots, X_p$ と，n 個のサンプルによって構成されており，各々の変数は，$C_0, \cdots, C_p$ 個のカテゴリー数をもつとする．X_0 を目的変数とし，それ以外を説明変数とする．変数選択の目的は，説明変数の集合 $I = \{X_1, \cdots, X_p\}$ $k \leqq p$ に対してその任意の部分集合 $J = \{X_{j_1}, \cdots, X_{j_k}\}$ が最も有効な説明変数の集合であるかを識別する問題になる．この目的において，考える確率モデルは，J のもとでの X_0 の条件付き確率

$$p(x_0 \mid x_1, \cdots, x_i) = p(x_0 \mid x_{j_1}, \cdots, x_{j_k}) \quad k \leqq i \tag{1}$$

として考えることができる．尤度比検定統計量との対応を考慮すると，その AIC は

$$\begin{aligned} \mathbf{AIC}(X_0; J) \;=\; & (-2) \sum_{x_0} \sum_{x_{j_1}} \cdots \sum_{x_{j_k}} n(x_0, x_{j_1}, \cdots, x_{j_k}) \\ & \times \log \frac{n \cdot n(x_0, x_{j_1}, \cdots, x_{j_k})}{n(x_0) n(x_{j_1}, \cdots, x_{j_k})} + 2(C_0 - 1) \\ & \times (C_{j_1} \cdots C_{j_k} - 1) \end{aligned} \tag{2}$$

で与えれる．$n(x_0, x_{j_1}, \cdots, x_{j_k})$ は，変数 $X_0, X_{j_1}, \cdots, X_{j_k}$ のとる値 $(x_0, x_{j_1}, \cdots, x_{j_k})$ に関する同時観測度数を表す．(2) は，一般的な尤度ではなく条件つき対数尤度に基づいて導かれた AIC である．その導出については，付録で示している．

AIC の例として，表 1 の分割表が与えられた場合を考える．考えられる説明変数の組合せと対応する AIC は，表 2 のようになる．変数選択の観点からみれば，AIC の値が最も低い変数を説明変数として選択するので，この場合，X_1X_2 が選択される．

2.2　AIC を用いた逐次変数選択法

(2) の AIC を変数選択の規準に用いると，変数増減法の動作は次のようになる．変数増減法は，変数増加と変数減少の 2 つのステップを繰り返す．変数増加ステップでは，加えた変数の組合せが，最も低い AIC をもてばその変数を実際に加え，変数減少ステップでは，削除した変数の組合せが，最も低い AIC をもてば，その変数を実際に削除する．変数増減法は，最初，説明変数が 3 変数になるまで変数増加ステップ行い，その後，変数減少ステップ，次に変数増加ステップを行い，変数が更新されなくなるまでそれらのステップを繰り返す．

表 2　表 1 の分割表の AIC

k	J	AIC
0	$\emptyset$	0
1	X_1	-3.136
1	X_2	-14.546
2	X_1X_2	-20.937

変数増減法の他に変数減増法の適用が考えられるが，変数の個数が多いとき，高次の分割表を考えなければならない．そのため，0 セルが多く出現し，分割表の変数選択の問題への適用を困難にする場合がある．その意味で逐次的に変数を増加させていく変数選択法のうち，変数増加法の欠点を補った変数増減法が適切であるといえる．

3.　2 値データにおける問題

上述した目的変数を有効に説明する説明変数の選択問題について, [4], [5] で提案するように (2) の AIC を用いた変数増減法を適用する方法が有効である．このような選択問題に対して，総当り法を適用すると明らかに変数が多い場合に破綻する．逐次的な方法は，最適の説明変数の組合せを探すことできない可能性を含むが，それとトレードオフの関係で圧倒的に計算コストが低い．現実的な範囲において，有効な変数を選択する方法として十分に機能するため，一般的によく利用されている．しかし，その変数増減法では，2 値型データを含めた変数のカテゴリー数が低い場合に，選択精度が低くなる傾向があることがシミュレーションによって確認した．その実験方法と結果を以下に示す．

選択精度をシミュレーションによって確認するため，以下のように人工データの生成する．複雑さを避けるために，すべて同じカテゴリー数 z をもつとする．すなわち，$C_1 = \cdots = C_p = z$. 目的変数を X_0，k 次の説明変数を $X_1, \cdots, X_k$ とする．いま，z 値データにおける $X_0, \cdots, X_k$ は, $X_0, \cdots, X_k$ の z^{k+1} 個のとりうるすべての実現値を全事象 Ω としたときに，要素 $E_i \in \Omega$ が確率 p_i $(i = 1, \cdots, z^{k+1})$ に従って生起するように生成する．ただし，p_i は，$p_1 + \cdots + p_{z^{k+1}} = 1$ の制約のもとで一様分布 $U[0, 1]$ に従う乱数によってあたえる．例えば，$z = 2$, $k = 1$ のとき $\Omega = \{E_1, E_2, E_3, E_4\}$ となり，X_0, X_1 の 2 次元確率分布は表 3 のようになる．残りの変数 $X_k, \cdots, X_p$ は，互いに独立な 0.5 のベルヌーイ乱数とする．データは，15 変数，4000 サンプルとして，10000 回のシミュレーションを行う．そのときの変数の選択率を表 4 に示す．

表 4 より，明らかに 2 値データ ($z = 2$) の $k = 1, 2, 3$ では，

表 3　$z = 2, k = 1$ の確率分布の例

event	X_0	X_1	probability
E_1	1	1	p_1
E_2	1	2	p_2
E_3	2	1	p_3
E_4	2	2	p_4
total			$p_1 + p_2 + p_3 + p_4$=1.0

表 4　変数増減法の選択率 (%)

	z		
k	2	3	4
1	6.59	68.36	99.24
2	18.52	99.25	100.00
3	47.73	99.98	99.98
4	83.72	99.78	87.95

表 5　2 値データにおける変数の選択割合 (%)

	k		
selected variables	1	2	3
(1) $\{X_1, \cdots, X_k\}$	6.59	18.52	47.73
(2) $\emptyset$	0.24	0.04	0.00
(3) including $\{X_1, \cdots, X_k\}$	86.87	78.41	51.59
(4) not including $\{X_1, \cdots, X_k\}$	6.30	3.03	6.8
total	100.00	100.00	100.00

$z > 2$ に比べて変数の選択率が極端に低いことがわかる．逆に，2 値データ以外のデータに対して，変数増減法は非常に強力な変数選択法であることがわかる．次に，2 値データ ($k = 1, 2, 3$) における選択した変数の構造に注目する．表 4 の 2 値データで選択した変数について，表 5 は，次の 4 つの変数の選択割合を表す．

（1） 正しい変数 $\{X_1, \cdots, X_k\}$
（2） 説明変数として何も選択されていない
（3） $\{X_1, \cdots, X_k\}$ が含まれている変数 (超集合)
（4） $\{X_1, \cdots, X_k\}$ が含まれていない変数

表 5 より，各 k において，正しい変数 $\{X_1, \cdots, X_k\}$ を含んだ変数が，正しい変数 $\{X_1, \cdots, X_k\}$ よりも多く選択されている．これは，変数増減法が適切な変数の個数で終了せずに余分な変数を多く加えていることを意味する．

以上の結果より，2 値データで $k \leqq 3$ の場合，無駄な変数く取り込んでしまうことがわかった．これは，カテゴリー数が少ない，特に 2 値データのような 2 カテゴリーの場合，目的変数に対してまったく独立に生成した変数であっても，確率的に有効な説明変数になってしまうためである．それは，たまたま有効な変数になっただけであり，そのデータの発生の背景を考慮すれば，選択されるべき変数ではない．よって，2 値データについて変数選択を行う場合，2 値データの不安定さを考慮し，変数増減法に余計な変数を選択しないよう仕組みを導入する必要がある．その仕組みにブートストラップ法を導入することを提案する．

4. ブートストラップ法を用いた変数増減法

[1] によって提案されたブートストラップ法はリサンプリング法の 1 つである [2]．ブートストラップ法はさまざまな分野で応用されており，パラメーターの信頼区間の構成，系統樹推定の信頼性評価にも利用されている．また，バギング (bagging; bootstrap aggregating) もブートストラップ法に基づいた方法であり，不安定な解を安定させ精度の向上に機能する方法として利用されている．

前述の実験結果を受け，変数増加法において変数の次元が低いときの増加ステップで無駄な変数を取り込まないようにするために，増加ステップだけにブートストラップ法を適用する．変数増減法に限らず AIC を用いた変数選択法は，各ステップの AIC の比較で，より低い AIC をもつ変数を選択する．

$\boldsymbol{X}$ を $n \times p$ のデータ行列とする．変数増減法において，データ $\boldsymbol{X}$ から計算される AIC の各比較によって選択される変数は，データによって AIC が変化するので $\boldsymbol{X}$ の関数となり，AIC がデータから計算する統計量であること考慮して $\hat{s}(\boldsymbol{X})$ と書く．ブートストラップ法を適用した変数増減法は，$s(\boldsymbol{X})$ が信頼できる変数なのか，すなわち，AIC の大小関係が信頼できるものかどうかを測るために，$\boldsymbol{X}$ からの B 回のブートストラップ複製 $\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_b^*$ から，$s(X_b^*) = \hat{s}(\boldsymbol{X})$ となる回数

$$L = \#\left\{\hat{s}(\boldsymbol{X}) = s(\boldsymbol{X}_b^*),\ b = 1, \cdots, B\right\} \tag{3}$$

を数えて，

$$\frac{L}{B} \geq \alpha$$

となる $\hat{s}(\boldsymbol{X})$ を選択する．$\frac{L}{B} < \alpha$ であれば $s(\boldsymbol{X})$ に関する AIC の関係は信頼できないので選択しないものとする．ただし，X_b^* のサンプルサイズは n とする．本提案方法では，オリジナルデータにおいて加えた変数がより低い AIC を持つ場合に，ブートストラップ確率によってその変数を加えるかどうかの判断を行なう．これは，変数を増加させる観点からは，保守的な方法であるが，ブートストラップの計算コストを大幅に減らすことができる．

このように，ブートストラップ法を変数増加法に用いることによって，不安定な 2 値データに対してその信頼性を評価することにより，無駄な変数をできるだけ選択しないようにする．この方法の有効性をシミュレーションによって確認する．

5. 比 較 実 験

ブートストラップ法による選択精度の改善をシミュレーションによって確認する．人工データの生成方法は，3 章で用いた方法と同じであり，データサイズは，10 変数, 10000 サンプルである．ブートストラップ反復回数 $B = 1000, 200$ の各々のもとで，300 回のシミュレーションを行う．α は，一般的に用いられている有意水準の 5% に基づいて，$\alpha = 0.95$ とした．ブートストラップを行わない通常の変数増減法 (Normal と記す) とブートストラップを適用した方法 (Bootstrap と記す) の変数の選択率を表 6 に示す．

表 6　変数の選択率の比較 (%)

$B = 200$						
	$z = 2$		$z = 3$		$z = 4$	
k	Normal	Bootstrap	Normal	Bootstrap	Normal	Bootstrap
1	29.0	93.3	85.6	96.0	100.0	100.0
2	47.3	94.3	100.0	100.0	100.0	100.0
3	81.6	98.6	100.0	100.0	100.0	100.0
4	95.0	97.3	100.0	100.0	100.0	100.0
$B = 1000$						
	$z = 2$		$z = 3$		$z = 4$	
k	Normal	Bootstrap	Normal	Bootstrap	Normal	Bootstrap
1	28.6	89.0	85.3	98.6	100.0	100.0
2	50.3	97.0	99.6	99.6	100.0	100.0
3	77.0	97.0	100.0	100.0	100.0	100.0
4	93.6	98.3	100.0	99.6	100.0	99.6

表 6 より，問題となっていた 2 値データの変数の選択率が改善されているのが確認できる．また，2 値データ以外の $z > 2$ において，ブートストラップを適用しても，もともと高い選択率に悪影響を及ぼしていないことがわかる．$B = 200$ のブートストラップ複製でも，$B = 1000$ と同等の選択率を向上させる能力があることがわかる．

6. 実データの適用例

実際のデータに対しての提案手法の適用例を示す．用いるデータは，表 7 の 1008 人に対する洗剤の選好データ [3] である．銘柄を目的変数 X_0 として，このときの可能な全て説明変数における (2) の AIC の値は，表 8 のようになる． AIC の値

表 7　洗剤選好データ

			硬度 (X_3)		
使用経験 (X_1)	水温 (X_2)	好み (X_0)	硬水	中位	軟水
なし	低温	銘柄 X	68	66	63
		銘柄 M	42	50	53
	高温	銘柄 X	42	33	29
		銘柄 M	30	23	27
あり	低温	銘柄 X	37	47	57
		銘柄 M	52	55	49
	高温	銘柄 X	24	23	19
		銘柄 M	43	47	29

表 8　説明変数の AIC

explanatory variable-set	AIC
X_1X_2	-21.16
X_1	-18.58
X_1X_3	-16.19
$X_1X_2X_3$	-10.82
X_2	-2.36
X_3	3.60
X_2X_3	5.17

より，最小の値をもつ説明変数の組合せとして X_1X_2 が選択される必要があり，変数増減法を用いると X_1X_2 が選択される．一方，提案方法を用いた 1000 回シミュレーションでは，X_1 が 100% の選択率で選択された．ただし，このときのブートストラップ複製は 1000 回である．提案手法で X_1 が選ばれるのは，X_1X_2 と X_1 の AIC の値がそれぞれ-21.16 と-18.58 であり，その差が小さいためにブートストラップ複製によって明確な大小関係が成り立つとは言えないという結論に至るためである．このことは，X_1 の AIC の値と分割表の可視化という観点からも，より自然な結果を与えるといえる．

7. ま と め

本稿では，2 値データにおける変数増減法の選択率低下の問題を指摘し，その改善方法として，変数増減法の増加ステップにブートストラップ法の適用を提案した．シミュレーションを用いた実験によってブートストラップ法が，その問題に対して有効に機能することがわかった．

ブートストラップ法を用いても 2 値データ以外のデータに関して，選択率を低下させる影響がないためにユニバーサルな方法であるといえる．また，少なくとも 200 回ほどのブートストラップ複製でも十分に効果があり，そのため，計算コストもあまり問題にならないといえる．

文　献

[1] B. Efron, "Bootstrap methods: Another look at the jackknife", Ann. Statist, no.7, pp.1-26, 1979.

[2] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall/CRC, New York, 1993.

[3] P.N. Ries and H. Smith, "The use of chi-square for preference testing in multidimensional problem", Chemical Engineering Progress, vol.59 pp.39-42, 1969.

[4] Y. Sakamoto, Categorical Data Analysis, Kluwer Academic Publishers, Dordrecht, 1991.

[5] 坂元慶行, カテゴリカルデータのモデル分析, 共立出版, 東京, 1985.

付　録

ここでは, [4] [5] より，確率モデル

$$p(x_0, x_1, \cdots, x_k) = p(x_0|x_1, \cdots, x_k)$$

の AIC の導出方法を示す．いま，データには $p+1$ 個の変数 X_j $(j = 0, 1 \cdots, p)$ があり，変数 X_j は値 $1, \cdots, C_j$ をとる．k 個の変数からなる任意の変数集合 $c = \{X_0, \cdots, X_k\}$ $(k \leqq p+1)$ が値 $x_0, \cdots, x_k$ をとる確率を $p(x_0, \cdots, x_k)$ とし，対応する観測度数を $n(x_0, \cdots, x_k)$ とすると，

$$\sum_{x_0=1}^{C_0} \cdots \sum_{x_k=1}^{C_k} p(x_0, \cdots, x_k) = 1 \qquad \text{(A·1)}$$

$$\sum_{x_0=1}^{C_0} \cdots \sum_{x_k=1}^{C_k} n(x_0, \cdots, x_k) = n \qquad \text{(A·2)}$$

である．ただし，n はサンプルサイズである．母集団が十分大きいとすると，確率 $p(x_0, \cdots, x_k)$ の下での観測度数の集合 $n(x_0, \cdots, x_k)$ が得られる確率は，多項分布

$$M(n(x_0,\cdots,x_k)|p(x_0,\cdots,x_k))$$
$$= \frac{n!}{\prod_{x_0=1}^{C_1}\cdots\prod_{x_k=1}^{C_k} n(x_0,\cdots,x_k)!}$$
$$\times \prod_{x_0=1}^{C_0}\cdots\prod_{x_k=1}^{C_k} p(x_0,\cdots,x_k)^{n(x_0,\cdots,x_k)}$$

で与えられる．いま，$n(x_0,\cdots,x_k)$ が与えられたもとでの $p(x_0,\cdots,x_k)$ の対数尤度関数 $l(p(x_0,\cdots,x_k)|n(x_0,\cdots,x_k))$ は，$p(x_0,\cdots,x_k)$ に無関係な定数項を無視すれば

$$l(p(x_0,\cdots,x_k)|n(x_0,\cdots,x_k))$$
$$= \sum_{x_0=1}^{C_0}\cdots\sum_{x_k=1}^{C_k} n(x_0,\cdots,x_k)\log p(x_0,\cdots,x_k) \qquad \text{(A·3)}$$

で与えられる．一般に成り立つ関係

$$p(x_0,x_2,\cdots,x_k) = p(x_0|x_1\cdots,x_k)p(x_1,\cdots,x_0) \quad \text{(A·4)}$$

を，(A·3) に代入すると，その右辺は，

$$\sum_{x_0=1}^{C_0}\cdots\sum_{x_k=1}^{C_1} n(x_0,\cdots,x_k)\log p(x_0|x_1,\cdots,x_k) +$$
$$\sum_{x_1=1}^{C_1}\cdots\sum_{x_k=1}^{C_k} n(x_1,\cdots,x_k)\log p(x_1,\cdots,x_k) \qquad \text{(A·5)}$$

となる．この第 2 項を無視し，第 1 項だけを用いて，新たに

$$l(p(x_0|x_1,\cdots,x_k))$$
$$= \sum_{x_0=1}^{C_0}\cdots\sum_{x_k=1}^{C_k} n(x_0,\cdots,x_k)\log p(x_0|x_1,\cdots,x_k) \quad \text{(A·6)}$$

を考える．この形で定義される対数尤度を条件付き対数尤度とよぶ．いま，確率モデル

$$p(x_0,x_1,\cdots,x_k) = p(x_0|x_1,\cdots,x_k)$$

を考えると，この AIC は制約条件

$$\sum_{x_0=1}^{C_0} p(x_0|x_1,\cdots,x_k) = 1 \quad x_i = 1,\cdots,C_i \ (i=1,\cdots,k) \qquad \text{(A·7)}$$

と (A·6) における最尤推定量を考慮すると

$$(-2)\sum_{x_0=1}^{C_0}\sum_{x_1=1}^{C_1}\cdots\sum_{x_k=1}^{C_k} n(x_0,x_1,\cdots,x_k)$$
$$\times\log\frac{n\cdot n(x_0,x_1,\cdots,x_k)}{n(x_0)n(x_1,\cdots,x_k)} + 2(C_0-1)(C_1\cdots C_k-1)$$

となる．