AIC の多重比較法によるモデル集合の研究

M2006MM032 棚橋 昌也
指導教員　松田 眞一

1　はじめに

重回帰分析では、普通、モデル選択として変数選択を行う。その変数選択において AIC を用いる場合、AIC が最小となるように変数の組み合わせを考える。このように、AIC が最も小さくなるようにモデル選択を行うことを AIC 最小化法という。

AIC 最小化法は、重回帰分析における変数選択に限らず、様々なモデル選択の場面で用いられる。AIC は、確かに、モデル選択に対して一つの有効な解を与える。しかし、AIC は本来、統計量であるので、AIC 最小化法によって一意に決定したモデルが、必ずしも適切な予測を与えるとは限らない。したがって、サンプリングエラーを考慮して、AIC 最小モデルに対して、有意に差が認められないモデルに関しては、AIC 最小モデルと同様に検討すべきである (下平 [8]、[9] 参照)。そこで本研究では、棚橋 [10] から引き続いて、AIC 最小モデルに対して、有意差が認められないモデルを集めたモデル集合を導くことを考える。

2　2 つのモデルの有意差検定について

データセットの型を**表 1** に示す。ただし、これらは列方向に基準化してあるものとする。

表 1　データセットの型

No.	説明変数 (r 変数)				目的変数
1	x_{11}	x_{12}	$\cdots$	x_{1r}	y_1
2	x_{21}	x_{22}	$\cdots$	x_{2r}	y_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ir}	y_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
N	x_{N1}	x_{N2}	$\cdots$	x_{Nr}	y_N

$\boldsymbol{x}_i = (x_{i1},\ x_{i2},\ \cdots,\ x_{ir})'$

フルモデル $\Omega = \{1,\ 2,\ \cdots,\ r\}$ に対して、モデル式を

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_r x_{ir} + \epsilon_i \tag{1}$$
$$= \boldsymbol{\theta}\boldsymbol{x}_i + \epsilon_i \tag{2}$$

と定義する。ただし、$\boldsymbol{\theta} = (\beta_1,\ \beta_2,\ \cdots,\ \beta_r)$ をフルモデル Ω に対するパラメータベクトルとする。

今、N 個の r 次元確率ベクトル $\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ \cdots,\ \boldsymbol{X}_N$ について、それらが独立に同一の分布に従い、密度関数が $f(\boldsymbol{x}_n|\boldsymbol{\theta})$ $(n = 1,\ 2,\ \cdots,\ N)$ であるとき、その尤度関数は

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{n=1}^{N} f(\boldsymbol{x}_n|\boldsymbol{\theta}) \tag{3}$$

となる。したがって、その対数尤度関数は、

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log f(\boldsymbol{x}_n|\boldsymbol{\theta}_\omega) \tag{4}$$

と表される。ただし、$\boldsymbol{x} = (\boldsymbol{x}_1',\ \boldsymbol{x}_2',\ \cdots,\ \boldsymbol{x}_N')'$ である。

ここで、$\boldsymbol{\theta}_\omega = \{(\beta_1,\ \beta_2,\ \cdots,\ \beta_r) \mid \beta_i = 0,\ i \notin \omega\}$ をモデル $\omega(\subseteq \Omega)$ に対するパラメータベクトルとすると、モデル $\omega_0,\ \omega_1$ において、$\omega_0 \subset \omega_1$ または $\omega_0 \supset \omega_1$ であるとき尤度比検定を、そうでないとき Linhart の AIC 有意差検定を行う。

2.1　Linhart の AIC 有意差検定

AIC(赤池情報量規準) とは、モデル選択において規準とされる情報量の 1 つである。モデル $\omega(\subseteq \Omega)$ に対する AIC は、次のように定義される。

$$\mathrm{AIC}_\omega = -2(l(\hat{\boldsymbol{\theta}}_\omega) - \dim(\boldsymbol{\theta}_\omega)) \tag{5}$$

ただし、$\hat{\boldsymbol{\theta}}_\omega$ はパラメータベクトル $\boldsymbol{\theta}_\omega$ の最尤推定量、$\dim(\boldsymbol{\theta}_\omega)$ は $\boldsymbol{\theta}_\omega$ の自由度を表す。一般に、AIC が小さい方が良いモデルとされる。

Linhart によって、2 つのモデルの AIC における有意差を検定する方法が提案されている。これは 2 つのモデルが階層関係になければ、それらの AIC の差は、漸近的に平均 0 の正規分布によって比較的良く近似されることを利用したものである。

モデル $\omega_0,\ \omega_1(\omega_0, \not\subset \omega_1)$ の AIC を $\mathrm{AIC}_0,\ \mathrm{AIC}_1$ としたとき、これらの差 $\Delta\mathrm{AIC} = \mathrm{AIC}_0 - \mathrm{AIC}_1$ が有意に 0 から離れていなければ、これらのモデルは同等であるとする。

まず、AIC の差の分散を推定する。2 つのモデルの最大対数尤度を $l_0 = l(\hat{\boldsymbol{\theta}}_0),\ l_1 = l(\hat{\boldsymbol{\theta}}_1)$ とする。ただし、$\hat{\boldsymbol{\theta}}_0,\ \hat{\boldsymbol{\theta}}_1$ は、$\omega_0,\ \omega_1$ における尤度推定量である。

ここで、

$$h(\boldsymbol{x}) = \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_0) - \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_1) \tag{6}$$

とおくと、(4) より、2 つのモデルの最大対数尤度の差 Δl は

$$\Delta l = l_0 - l_1 = \sum_{i=1}^{N} h(\boldsymbol{x}_i) \tag{7}$$

と表され、その分散は

$$\hat{\mathrm{var}}(\Delta l) = \frac{n}{n-1} \sum_{i=1}^{N} \left\{ h(\boldsymbol{x}_i) - \frac{\sum_{j=1}^{N} h(\boldsymbol{x}_j)}{n} \right\}^2 \tag{8}$$

である。したがって、AIC の差の分散は

$$\hat{\mathrm{var}}(\Delta\mathrm{AIC}) = 4 \times \hat{\mathrm{var}}(\Delta l) \tag{9}$$

となる。

ここでは、AIC の差を推定した標準偏差で割った統計量

$$z = \frac{\Delta\mathrm{AIC}}{\sqrt{\hat{\mathrm{var}}(\Delta\mathrm{AIC})}} \tag{10}$$

が標準正規分布に従うとして検定を行う (下平ら [7] 参照)。

2.2 尤度比検定

モデル間に階層関係がある場合には、尤度比検定が知られている。今、モデル ω_0, ω_1 に、$\omega_0 \supset \omega_1$ という階層関係があるとすると、一般に $l_0 > l_1$ である。これらの最大対数尤度の差 $\Delta l = l_0 - l_1$ が有意に 0 から離れていなければ、これらのモデルに有意差はないとする。

尤度比検定では、$2\Delta l$ は漸近的に自由度 $n = \dim(\hat{\boldsymbol{\theta}}_0) - \dim(\hat{\boldsymbol{\theta}}_1)$ のカイ 2 乗分布に従うとして、検定を行う (稲垣 [4] 参照)。

3 多重比較法について

r 個の説明変数において、すべての組合せを考えた 2^r 個のモデルについて、AIC 最小モデルを ω_0、その他のモデルを ω_k $(k = 1,\ 2,\ \cdots,\ 2^r - 1)$ とする。また、それぞれのモデルにおけるパラメータの真値を $\boldsymbol{\theta}_0{}^*,\ \boldsymbol{\theta}_1{}^*,\ \cdots,\ \boldsymbol{\theta}_{2^r-1}{}^*$ とする。本研究では、ω_0 と ω_k $(k = 1,\ 2,\ \cdots,\ n)$ に有意差があるかを調べる。したがって、次のような帰無仮説族を考える。

$$\left\{ \begin{array}{ccccc} H_{01} & : & \boldsymbol{\theta}_0{}^* & = & \boldsymbol{\theta}_1{}^* \\ H_{02} & : & \boldsymbol{\theta}_0{}^* & = & \boldsymbol{\theta}_2{}^* \\ \vdots & & & \vdots & \\ H_{02^r-1} & : & \boldsymbol{\theta}_0{}^* & = & \boldsymbol{\theta}_{2^r-1}{}^* \end{array} \right. \tag{11}$$

今、帰無仮説の個数が 31 個、それぞれの帰無仮説における有意水準が 0.05 であるとすると、少なくとも 1 つの検定において誤りが起こる確率は

$$1 - 0.95^{31} = 0.796\cdots$$

より、約 80% となり、これはもはや誤りが起こる確率として認められる範疇を大きく超えてしまっている。そこで、もし検定全体として誤りが起こる確率をある一定の値に抑えたい場合は、何らかの補正が必要となる。

本研究では、その補正の方法として、Holm の方法と BH 法を用い、これらを比較する。

3.1 Holm の方法

Holm の方法とは FWER(Family-Wise Error Rate) を制御する多重比較法で、Bonferroni の不等式に基づく多重比較法 (Bonferroni の方法) に閉検定手順の考え方を組み合わせて改良したものである。

FWER とはすべての帰無仮説のうち、すくなくとも 1 つの帰無仮説が誤って棄却される確率である。

以下に、Holm の方法の手順を簡潔に示す (Holm[2]、永田・吉田 [5] 参照)。

Holm の方法の手順

1. 帰無仮説族に含まれる帰無仮説の個数 k を求める。
2. 有意水準 α を定める。
3. 各帰無仮説における検定統計量を計算し、p 値 P_i $(i = 1,\ 2,\ \cdots,\ k)$ を求める。
4. $P_1,\ P_2,\ \cdots,\ P_k$ を昇順に並べ替えたものを $P^{(1)},\ P^{(2)},\ \cdots,\ P^{(k)}$ とする。
5. $i = 1$ とする。
6. $P^{(i)} > \dfrac{\alpha}{k - i + 1}$ となるなら、帰無仮説 $H^{(i)},\ H^{(i+1)},\ \cdots,\ H^{(k)}$ をすべて保留して検定作業を終了する。そうでなければ、帰無仮説 $H^{(i)}$ を棄却して次へ進む。
7. $i = k$ なら手順を終了する。$i < k$ なら、$i = i + 1$ として 6 から繰り返す。

3.2 BH 法

BH 法とは、FDR(False Discovery Rate) を制御する多重比較法である。BH 法は、堀内・松田 [3] において、仮説間にどのような関連があっても、安定して FDR を保つことができるということが示されている。

FDR とは、棄却されたすべての帰無仮説のうち、真の帰無仮説であるにもかかわらず誤って棄却されたものの割合である。m 個の帰無仮説の検定結果についてまとめたものを**表 2** に示す。

表 2　m 個の帰無仮説の検定結果の分割表

	検定		
	採択	棄却	計
真の帰無仮説	U	V	m_0
偽の帰無仮説	T	S	$m - m_0$
計	$m - R$	R	m

FDR を制御する多重比較法では、$Q = \dfrac{V}{R}$(ただし、$R = 0$ のときは、$Q = 0$ とする) を制御することを考える。しかし、Q は決して知ることのできない確率変数であるので、実際には、その期待値 $Q_e = E(Q)$ を制御することを考える。

以下に、BH 法の手順を簡潔に示す (Benjamini and Hochberg[1]、堀内・松田 [3] 参照)。

BH 法の手順

4. まで Holm の方法と同様。

5. $i = k$ とする。
6. $P^{(i)} \leq \dfrac{i}{m} q*$ となるなら、帰無仮説 $H^{(1)},\ H^{(2)},\ \cdots,\ H^{(k)}$ をすべて棄却する。そうでなければ、帰無仮説 $H^{(i)}$ を保留して次へ進む。
7. $i = 1$ なら手順を終了する。$i > 1$ なら、$i = i - 1$ として 6 から繰り返す。

BH 法を適用することは、Holm の方法に比べて、検出力が上昇することが期待される。

4 モデル集合作成の手順

以下に、本研究におけるモデル集合作成の手順を示す。

モデル集合作成の手順

1. 説明変数の数が r 個のデータにおいて、説明変数の全ての組合せを考えた 2^r 個のモデルを用意する。
2. 全てのモデルの AIC を計算する。
3. AIC を最小とするモデルを M_0、それ以外を $M_1,\ M_2,\ \cdots,\ M_{2^r-1}$ とする。
4. M_0 に対して、$M_i\ (i = 1,\ 2,\ \cdots,\ 2^r - 1)$ が階層関係にあるかどうか調べる。
5. 階層関係にあれば尤度比検定、階層関係になければ Linhart の AIC 有意差検定を行い、p 値 $P_1,\ P_2,\ \cdots,\ P_{2^r-1}$ を求める。
6. Holm の方法、または BH 法を用いて有意水準 α で検定し、モデル集合を作成する。

ただし、5. において、2 つの検定の整合性をとるために、AIC 有意差検定における p 値は 2 倍する。

5 シミュレーション

5.1 AIC 最小化法の信頼性

(2) のような線形回帰モデルにおいて、$\beta_1,\ \cdots,\ \beta_r$ を最小値 0.1、最大値 1.0 の一様乱数で発生させる。このとき、そのいくつかは 0 に置き換える。さらに ϵ_i を平均 0、標準偏差 0.1 の正規乱数で、$x_{i1},\ \cdots,\ x_{ir}\ (i = 1,\ \cdots,\ N)$ を平均 0、標準偏差 1 の正規乱数で発生させ、シミュレーションデータを作成する。

シミュレーション回数 1000、説明変数の数 $r = 2, 5, 8$、データ数 $N = 10,\ 20,\ 50,\ 100,\ 200,\ 500,\ 1000$ とする。AIC 最小モデルが真のモデルと一致した数を**図 1** [*1] に示す。

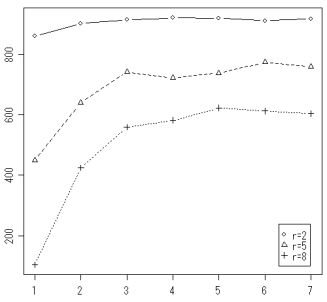

図 1　AIC 最小モデルが真のモデルと一致した数

図 1 から、モデルの候補の数が多くなるにしたがって、AIC 最小化法によって一意にモデルを決定することが、必ずしも正しくない場合が多くなることがわかる。したがって、モデル集合を導くことが必要となる。

5.2 モデル集合の傾向

ここでは、モデル集合がどのような傾向をもって決定されているかを観察する。

真のモデルにおいて、0 に置き換えられた説明変数の数を i とする。$r = 5$、$N = 100$ に対して、$i = 0,\ 1,\ 2,\ 3,\ 4$ のときの、Holm の方法によるモデル集合の大きさの分布を**図 2** に示す。シミュレーション回数は、それぞれ 1000 とする。

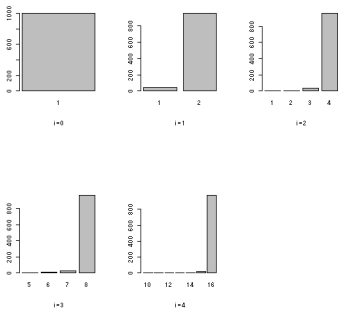

図 2　Holm の方法によるモデル集合の大きさ $(i = 0,\ 1,\ 2,\ 3,\ 4)$

図 2 からわかるとおり、それぞれ 2^i 部分に大きく山ができている。つまり、モデル集合は、真のモデルの構造に依存して、その大きさが決定されているということがわかる。

いくつかのモデル集合を観察すると、真のモデル M^*、モデル集合に含まれるモデル $M_i\ (i = 1,\ 2,\ \cdots)$ について、すべての M_i において、$M^* \subseteq M_i$ となっていることがわかった。したがって、0 に置き換えられた説明変数のすべての組合せ 2^i 個だけ、モデル集合にモデルが含まれるということがわかる。

その他のほとんどの場合においても同様である。したがって、モデル集合に含まれるすべてのモデルの積集合をとることによって、真のモデルの構造を知ることができると考えられる。

5.3 積集合による真のモデルの予測精度

ここでは、$r = 5$、$N = 100$ に対して、ϵ_i の標準偏差が $\sigma = 0.1,\ 1.0$ のときの、前述した積集合による真のモデルの予測の精度を観察する。シミュレーション回数は、それぞれ 1000 回とする。ただし、積集合によって予測された真のモデルは、必ずしもそのモデル集合に含まれているとは限らないことに注意する。結果を**図 3**$(\sigma = 0.1)$、**図 4**$(\sigma = 1.0)$ [*1] に示す。

図 3、**図 4** からわかるとおり、σ が比較的小さい場合には、積集合による予測において、AIC を大きく上回

[*1] 1: $N = 10$, 2: $N = 20$, 3: $N = 50$, 4: $N = 100$, 5: $N = 200$, 6: $N = 500$, 7: $N = 1000$

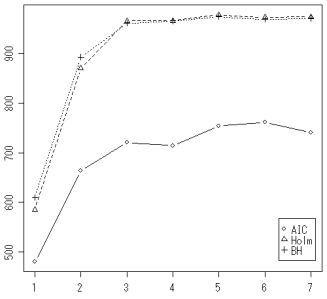

図 3　積集合による真のモデルの予測の精度 ($\sigma = 0.1$)

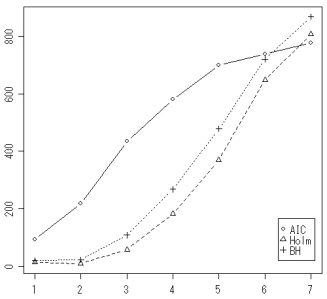

図 4　積集合による真のモデルの予測の精度 ($\sigma = 1.0$)

る結果が得られる。しかし、σ が大きくなると、単純な積集合による予測では、ある程度のデータ数を採らないと、十分な結果を得ることができない。これは、誤差項が大きくなると、総じて、モデル集合が大きくとられることに帰着する。

6　解析例

下平 [8] と同様に、新生児の体重データ (佐和 [6]) を用いて解析を行う。説明変数の数は $r = 4$、データ数は $N = 15$ である。$2^4 = 16$ 個のモデル候補について考える。解析結果について、p 値が上位 8 個のモデルを**表 3** に示す。

有意水準を 0.05 とすると、Holm の方法、BH 法ともに上位 2 個のモデルが選択された。積集合の考え方では、2 番目のモデル {1, 2, 3} が適切な予測を与えるモデルであると考えられる。しかし、説明変数の数に対してデータ数が十分であるとは言えないので、真に効いている説明変数を取りこぼさないという意味でも、1 番目のフルモデルを選択することが無難であるかもしれない。

下平 [8] では、有意水準 20% で、モデル集合に {1, 2, 3, 4}、{1, 2, 3}、{1, 3, 4}、{1, 3}、{1, 2}、{1, 4} 、{1, 2, 4}、{1} の 8 個のモデルが選ばれた。そのうち、TIC の立場から、{1} だけを選ぶのも一つの手であると述べている。

表 3　新生児の体重データにおけるモデル候補の AIC と p 値 (上位 8 個)

	モデル	AIC	p 値
1	{1, 2, 3, 4}	155.8572	—
2	{1, 2, 3}	157.3471	0.06174
3	{1, 3, 4}	160.4402	0.01030
4	{1, 3}	163.0421	0.00373
5	{1, 2}	163.9788	0.00233
6	{1, 4}	164.6006	0.00171
7	{1, 2, 4}	163.9263	0.00151
8	{1}	165.7155	0.00121

7　おわりに

本研究では、多重性を考慮するにあたって、BH 法を用いることによって、従来の FWER を制御する多重比較法に比べて検出力を上げることができた。

シミュレーションを重ねた結果、本研究で導いたモデル集合は、ある傾向をもって決定され、その傾向から真のモデルを予測すること可能であることがわかった。しかし、実用上の場面では、多重共線性がある場合や誤差項が大きい場合、また比較するモデルの候補が多い場合などにさらなる検討の余地がある。

参考文献

[1] Benjamini, Y and Hochberg, Y：Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society series B, No.57, pp.289-300 (1995).

[2] Holm, S：A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics, Vol.6, pp.65-70 (1979).

[3] 堀内賢太郎・松田眞一：FDR を制御する多重比較法の性能評価, 南山大学紀要『アカデミア』 数理情報編, Vol.6, pp.17-30 (2006.3).

[4] 稲垣宣生：数理統計学, 裳華房 (1990).

[5] 永田靖・吉田道弘：統計的多重比較法の基礎, サイエンティスト社 (1997).

[6] 佐和隆光：回帰分析, 朝倉書店 (1979).

[7] 下平英寿・伊藤秀一・久保川達也・竹内啓：統計科学のフロンティア 3 モデル選択, 岩波書店 (2004).

[8] 下平英寿：モデルの信頼集合と地図によるモデル探索, 統計数理, Vol.41, pp.131-147 (1993).

[9] 下平英寿：モデル選択理論の新展開, 統計数理, Vol.47, pp.3-27 (1999).

[10] 棚橋昌也：AIC による回帰分析のモデル候補選択の研究, 南山大学 数理情報学部 数理科学科 卒業論文 (2006.3).