AICによる回帰分析のモデル候補選択の研究

2002MM098 棚橋 昌也
指導教員　松田 眞一

1　はじめに

一般的にはAIC最小モデルが真のモデルを近似したものと考えることができるが、AICはデータから算出する統計量であるためサンプリングによるばらつきが生じてしまう。そのため、モデル間のAIC値の差が十分に大きくなければ、どのモデルが本来選択されるされるべき良いモデルなのか判断しかねる。

そこで本研究ではまず、白旗 [5]、下平 [4]、小西・北川 [3] を用いて、従来の尤度原理によるモデル選択を学ぶと共に、その尤度原理だけでは何が不十分であるのかを考え、さらにその尤度原理とAICの関連について述べる。またAICを用いた変数選択をするにあたって、AIC最小モデルを真のモデルを近似するものとして挙げるだけにとどまらず、それに続いていくモデルとして、AIC最小モデルを基準に上位いくつかのモデルについて、それらが検定によってAIC最小モデルと同等であると判断された場合にはそれらを良いモデルの候補として挙げていくことを考える。上記の検定について、その具体的方法を下平 [4] を用いて学ぶ。またそれらの検定の際には多重比較を想定しているので、その多重性を考慮する際の補正の仕方について永田・吉田 [6] を用いて考える。

これらのことを踏まえた上で、統計処理ソフト“R”上でのプログラムの作成を行う。プログラムの作成にあたっては、大村・山原 [2] よりAIC値を求める関数、井上・桑山 [1] よりAIC値を基準とした変数増減法を行う関数をそれぞれ引用する。

2　AICによるモデル選択

これまでの最尤法や、尤度比検定などの尤度原理を用いたモデル選択では、それぞれ問題点を抱えている。最尤法はに、線形回帰モデルにおいて適用するとパラメータをすべて含んだ最大モデルが選択されてしまうという問題点があり、尤度比検定はモデル間に包含関係がなければ適用することができない。そこで、これらを改善したAICによるモデル選択が考え出された。

モデル M_k のAICは

$$\mathrm{AIC}_k = -2 \times (l_k(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}; \boldsymbol{X}) - \dim \boldsymbol{\theta}_{\boldsymbol{k}})$$

と定義され、モデル選択の1つの基準として考えられる。AICは全体が -2 倍されているので、AICの値がより小さいモデルがあてはまりの良いモデルであるということがいえる。つまり、モデルの候補 $M_1, \cdots, M_k$ に対して、$\mathrm{AIC}_1, \cdots, \mathrm{AIC}_\mathrm{k}$ を計算し、AIC_k を最小にする k として $\hat{k}$ を定義する。このようにしてAIC最小モデルを選ぶのが、モデル選択、すなわち回帰分析における変数選択にAICを用いた方法である。(白旗 [5]、下平 [4]、小西・北川 [3] 参照)

3　AIC最小モデルの各モデル候補との検定方法

3.1　AICの差の有意性検定

2つのモデル M_k, $M_{k'}$ を比較した場合にそのAICの値の差 ΔAIC が微小であればどちらのモデルが良いかわからないというような検定について考える。2つのモデル M_k, $M_{k'}$ におけるAICの値をそれぞれ AIC_k, $\mathrm{AIC}_{k'}$ とし、そのAICの値の差 ΔAIC について考える。M_k と $M_{k'}$ が包含関係になければ、この ΔAIC が正規分布によって比較的良く近似できることを利用した検定が提案されている。

2つのモデル M_k, $M_{k'}$ において、各データ t における尤度をそれぞれ $p_k(\boldsymbol{x_t}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$, $p_{k'}(\boldsymbol{x_t}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k'}})$ とした場合に、それぞれの対数をとった対数尤度の差 $\Delta l(\boldsymbol{x_t})$ は

$$\Delta l(\boldsymbol{x_t}) = \log p_k(\boldsymbol{x_t}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) - \log p_{k'}(\boldsymbol{x_t}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k'}})$$

とすることができる。ここで、モデル M_k と $M_{k'}$ の最大対数尤度の差を

$$\Delta l(\boldsymbol{X}) = \sum_{t=1}^{n} \Delta l(\boldsymbol{x_t})$$

と表現すると、その分散は

$$\hat{\mathrm{var}}(\Delta l(\boldsymbol{X})) = \frac{n}{n-1} \sum_{t=1}^{n} \left\{ \Delta l(\boldsymbol{x_t}) - \frac{1}{n} \sum_{t'=1}^{n} \Delta l(\boldsymbol{x_{t'}}) \right\}^2$$

と推定することができる。AICは対数尤度からパラメータ数をひいたものを -2 倍したものであり、分散ではさらに2乗をとっていることからAICの差の分散 $\hat{\mathrm{var}}(\Delta\mathrm{AIC}(\boldsymbol{X}))$ は対数尤度差の分散 $\hat{\mathrm{var}}(\Delta l(\boldsymbol{X}))$ を4倍したもの、つまり、

$$\hat{\mathrm{var}}(\Delta\mathrm{AIC}(\boldsymbol{X})) = 4 \times \hat{\mathrm{var}}(\Delta l(\boldsymbol{X}))$$

となる。

この検定では、AICの差を標準偏差で割った統計量

$$z = \frac{\Delta\mathrm{AIC}(\boldsymbol{X})}{\sqrt{\hat{\mathrm{var}}(\Delta\mathrm{AIC}(\boldsymbol{X}))}}$$

が分散1の正規分布に従うと近似して検定を行う。(下平 [4] 参照)

3.2 対数尤度を利用した検定

本研究では、上記の検定の方法のほかに対数尤度を利用した検定を提案する。

2 つのモデル M_k, $M_{k'}$ について、それぞれの各データにおける対数尤度を $\log p_k(\boldsymbol{x_t}; \hat{\boldsymbol{\theta}_k})$, $\log p_{k'}(\boldsymbol{x_t}; \hat{\boldsymbol{\theta}_{k'}})$ とし、これら対数尤度の和がそれぞれのモデルにおける最大対数尤度となるような一元配置データについて考える。このようなデータから 2 つのモデル M_k, $M_{k'}$ に差があると言えるかどうかの検定を行う。それぞれのモデルのデータが $\mathrm{N}(\mu_k,\ {\sigma_{kk'}}^2)$, $\mathrm{N}(\mu_{k'},\ {\sigma_{kk'}}^2)$ に従うとし、両モデル間に差があるかどうかを t 検定によって判断する。

4 多重性について

本研究では上記の検定において、複数回これらを繰り返すことが想定されるので多重性について考慮する必要がある。それは各検定における誤りの程度 5 %、すなわち有意水準を 5 %と設定し、検定を 3 回繰り返さなければならないとした場合にこの検定を 3 回繰り返すという一つの検定として考えると、全体としては 3 つの検定のうち少なくとも一つの検定において誤って棄却されてしまう確率が $1 - 0.95^3 = 0.141625$ となりおよそ 14 %となってしまうためである。

そこでその代表的な方法として、ボンフェローニ (Bonferroni) の方法がある。これは、k 個の事象に対して $E_i (i = 1,\ 2,\ \cdots,\ k)$ とした場合、

$$P\left(\bigcup_{i=1}^{k} E_i\right) \leq \sum_{i=1}^{k} P(E_i)$$

というボンフェローニの不等式に基づいて多重性を考慮するという方法である。本研究では、このボンフェローニの方法を発展させたホルム (Holm) の方法を用いる。

ホルムの方法とはボンフェローニの方法のステップダウン版であり、各検定の有意水準に対する補正の仕方に改良が施されている。

まずそれぞれの統計量に対する p 値を求め、それらを昇順に並べる。昇順に並べた p 値を $P_1,\ P_2,\ \cdots,\ P_k$ と表し、これらの p 値に対する有意水準を

$$\alpha_1 = \frac{\alpha}{k},\ \alpha_2 = \frac{\alpha}{k-1},\ \cdots,\ \alpha_k = \frac{\alpha}{1} = \alpha$$

と設定する。ここから p 値が小さい順に検定を行う。P_i 値が α_i よりも小さい場合にはその検定における帰無仮説を棄却ししていくわけだが、P_i が α_i よりも大きくなった場合にはその検定に対する帰無仮説を保留し、まだ検定を行っていないすべての帰無仮説も同時に保留する。ホルムの方法は、順に検定を行っていけば残りの検定の数が減っていくということに着目して、それぞれの検定における有意水準に補正を行っている。それによって、ホルムの方法はボンフェローニの方法よりも一つ一つの検定において棄却がしやすくなっており、より検出力の高い手法となっている。(永田・吉田 [6] 参照)

5 プログラム

5.1 概要

プログラムは 3 節で述べた 2 つの検定方法に対して作成し、両プログラムともホルムの方法によって多重性を考慮した。

5.2 両プログラムにおける出力結果とその比較

データには 2005 年度日本女子プロゴルフ年間獲得賞金ランキング (LPGA[7] 引用) より、説明変数を獲得賞金、目的変数を試合数、ラウンド数、優勝回数、パーオン率、パーセーブ率、パーブレーク率、リカバリー率、平均ストローク数、平均パット数、バーディ数、イーグル数とするサンプル数 98 のものを用いた。

その結果、それぞれ AIC 最小モデルを含めて、3.1 節を基にしたプログラムではモデルの候補が 110、3.2 節を基にしたプログラムでは 143 挙げられた。両プログラムとも候補として挙げられたモデルがかなり多く、一つの原因として、ホルムの方法による補正が厳しすぎたということが言える。また両プログラムとも AIC 最小モデルと包含関係にあるモデルが比較的上位に位置している。

6 おわりに

本研究において二つのプログラムを作成したわけであるが、両プログラムとも多重比較の方法が十分であったとは言えず、今後の研究でより的確な多重比較法を検討する必要がある。さらにそれぞれの検定方法において、前提としている正規分布による近似であるなどしっかりと根拠を示すことができなかった。そういった意味で、満足のいく結果であったとは言えないが、今後の研究でプログラムの改善に取り組んでいき、満足のいくものに仕上げていきたい。

参考文献

[1] 井上勤・桑山智裕：S-plus における回帰分析の変数選択関数の作成, 南山大学経営学部情報管理学科卒業論文，2001.

[2] 大村学・山原強志：S-plus における回帰分析の変数選択の研究, 南山大学経営学部情報管理学科卒業論文, 2000.

[3] 小西貞則・北川源四郎：シリーズ「予測と発見の科学」2 情報量基準，朝倉書店，2004.

[4] 下平英寿：統計科学のフロンティア 3 モデル選択 第 1 部 情報量規準によるモデル選択とその信頼性評価，岩波書店，2004.

[5] 白旗慎吾：統計解析入門，共立出版株式会社，1992.

[6] 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.

[7] LPGA 日本女子プロゴルフ協会：http://www.lpga.or.jp/