研究ノート

モデル選択基準とその正規線形モデルへの適用

松　尾　精　彦

要　約

経済データ分析において，正規線形回帰モデルを想定し，その枠内でモデルを特定しようとする場合を考える．この研究ノートで焦点を当てる問題は，核となる説明変数（外生変数，独立変数とも言う）は分かっているが，それに付け加える説明変数群の候補が2つあり，そのどちら(あるいは両方)をモデルに付け加えるべきかを決定するというものである．

この問題に対し，Non-Nested モデル検定や逐次変数選択法といった，モデル選択アプローチがあるが，これらはいずれも得られたデータに対するモデルの適合度に基づくものである．それに対し，ここで述べるモデル選択基準は，得られたデータをもとに予測を行う際の最適性に基づくものであり，より実践的な意味を持つ．

ここでは，Non-nested モデル検定，逐次変数選択法，そしてモデル選択基準の違いを述べた後，AIC (Akaike Information Criterion) や Mallows の C_p, そして Schwarz の SC といったモデル選択基準について議論する．

キーワード：Model selection; Forecasting; AIC; Mallows' C_p; Schwarz's SC.
経済学文献季報分類番号：16-10

1　紹介

経済データ分析の目的の一つに，2つの説明変数群のどちらか(あるいは両方)をモデルに付け加えるべきか決定しようとするものがある．例えば，秋岡 (2002) では，沖縄電力の民営化効果の有無について議論している．しかし，民営化以前・以後に対応するダミー変数は，技術革新とよく似た効果を示していて，どちらを採用すべきかの問題がある．また，会計の分野では，株式の収益率を説明するのに，会計数値を用いるか，キャッシュ・フローを用いるかの問題がある (百合草, 2001)．これら2つの場面で考えなくてはならないのは，よく似た効果を与える2つの説明変数群のうち，どちらを採用すべきかという問題である．可能性としては，「どちらも無い」，「どちらか一方が効果がある」，「両方効果がある」の4通りが考えられる．

何らかの意味で適切なモデルを選択するためのアプローチとして，逐次変数選択法や Non-nested モデル検定，そしてここで述べるモデル選択基準の3つがあるが，上のような問題に

*この研究は平成13年度関西大学学部共同研究費によって行った研究の一部である．本研究ノートを作成するにあたり，秋岡弘紀助教授，松本茂助教授（関西大学経済学部）そして 太田浩司氏（武蔵大学経済学部）には，数々の有益な助言を受けた．記して感謝の意を表する次第である．

たいしては，モデルを予測に用いる際の最適性にもとづくモデル選択基準を採用することが適切である．逐次変数選択法は，説明変数と応答変数との関連がよく分かっていない状態で，探索的に変数を選択するためのアプローチであり，基本的には変数増加法・変数減少法に基づいたアルゴリズムが提案されている．一方，Non-nested モデル検定では，正規線形回帰モデル v.s. ガンマ線形回帰モデル，正規線形回帰モデル v.s. 正規非線形モデルのように，どちらか一方が真のモデルを含んでいると仮定し，それがどちらのモデル (群) かを決定しようとするものである．もちろん，モデルが互いに Non-nested なら利用できるので，「どちらか一方に効果がある.」という場面では適用可能であるが，その基準は逐次変数選択法と同じく得られたデータをより良く説明するモデルを見つけるためのものなのである．

次の節で示すように，モデルはデータ数に依存して選ばれる．つまり，データが少なければそれだけ単純なモデルが選ばれるということである．単純なモデルが選ばれるとき，母数推定量には必然的にバイアスが生じる．そのため，個々の母数推定値よりはむしろ，モデル全体としてのパフォーマンスに意味があると言える．推定されたモデルのパフォーマンスをどのように測ればよいかとなると，その基準を，推定されたモデルを用いた“予測”に求めるのは極めて自然なことといえる．

2 節では，上述の問題を定式化し，なぜモデル選択基準が効果的であるかを議論する．その上で，3 節では，モデル選択基準を紹介しその性質について説明を行う．4 節では，3 節で紹介したモデル選択基準について総合的に論じる．5 節では，本文の展開に必要となる事柄を付け加える．

2 問題の定式化

先に述べたように，この研究ノートでは正規線形回帰モデルの枠内でモデル選択問題を考えることにする．つまり $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ を n 次元応答変量（被説明変量，内生変量あるいは従属変量とも言う）ベクトルとするとき，各 X_i $(i = 1, 2, \cdots, n)$ は独立に正規分布 $N(\eta_i, \sigma^2)$ に従っているとする．Z_O, Z_A, Z_B をそれぞれ $n \times k$, $n \times p$, $n \times q$ 説明変数行列とし，Z_O はモデルが必ず含む説明変数からなり，Z_A, Z_B はどちらか片方あるいは両方がモデルに含まれる可能性のある説明変数からなるものとする．すると次の 4 つのモデル（仮説）が考えられる．

$$M_O : \boldsymbol{\eta} = Z_O\boldsymbol{\omega}, \tag{1}$$

$$M_A : \boldsymbol{\eta} = Z_O\boldsymbol{\omega} + Z_A\boldsymbol{\alpha}, \tag{2}$$

$$M_B : \boldsymbol{\eta} = Z_O\boldsymbol{\omega} + Z_B\boldsymbol{\beta}, \tag{3}$$

$$M_{AB} : \boldsymbol{\eta} = Z_O\boldsymbol{\omega} + Z_A\boldsymbol{\alpha} + Z_B\boldsymbol{\beta}, \tag{4}$$

ここで，$\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ はそれぞれ $k \times 1$, $p \times 1$, $q \times 1$ 未知母数ベクトルとする．これら 4 つのモデルのうち，どれが一番妥当かを決定する問題を考える．

この問題は，統計的検定の繰り返しにより解決されるとは限らない．例えば，$H_o : M_O$, $H_1 : M_A$ では $H_1 : M_A$，$H_o : M_O$, $H_1 : M_B$ では $H_1 : M_B$，$H_o : M_O$, $H_1 : M_{AB}$ では $H_1 : M_{AB}$ が採択されるが，$H_o : M_A$, $H_1 : M_{AB}$ では $H_o : M_A$，$H_o : M_B$, $H_1 : M_{AB}$ では $H_o : M_B$ が採択されるかもしれない．このような場合，3 つのモデル M_A，M_B, M_{AB} が候補として残り，どのモデルが最も適切かの情報は得られない．

M_A と M_B のどちらが良好かを決める場合には，モデル間に包含関係がない (Non-nested) ので，普通の検定では優劣を決められない．このような場合には，J 検定，Cox 検定，あるいは，Vuong 検定などの Non-nested 検定が考案されている．統計的検定を繰り返す立場では，M_{AB} は除外され，M_A と M_B のどちらが良いかという検定に持ち込まれる．しかし，ここではモデル選択基準を用いるアプローチについて考えたい．モデル選択基準にもとづくアプローチは，推定されたモデルを用いて予測する際の，モデル・パフォーマンスの良し悪しに基づくものであり，以下の理由により現実的なものと言える．

現実的に見て，真のモデルは M_{AB} であろう．なぜなら $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ の少なくとも一方が厳密に $\mathbf{0}$ であることは考え難いからである．それ故，データ数 (n) が十分大きくなればどちらも有意になる筈である．いま，直交射影行列を

$$\begin{aligned} P_O =& Z_O(Z_O^T Z_O)^{-1} Z_O, \\ P_A =& Z_A(Z_A^T Z_A)^{-1} Z_A, \\ P_B =& Z_B(Z_B^T Z_B)^{-1} Z_B, \end{aligned} \tag{5}$$

とおく．真のモデル M_{AB} を推測に用いる際，$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ の有意性を F 検定するときの分子はそれぞれ，

$$\boldsymbol{X}^T P_A(I-P_B)(I-P_O)\boldsymbol{X},\ \boldsymbol{X}^T P_B(I-P_A)(I-P_O)\boldsymbol{X} \tag{6}$$

であり，自由度が p, q, 非心度が

$$\frac{\boldsymbol{\alpha}^T Z_A^T(I-P_B)(I-P_O)Z_A\boldsymbol{\alpha}}{\sigma^2},\quad \frac{\boldsymbol{\beta}^T Z_B^T(I-P_A)(I-P_O)Z_B\boldsymbol{\beta}}{\sigma^2} \tag{7}$$

のカイ自乗分布の σ^2 倍になる．上の非心度は，データ数 n が増えれば増えるほど大きくなるので，それだけ有意になりやすくなる．

しかし得られたデータの範囲内で推測を行うことが統計解析の目的であるため，統計解析の結果として，真のモデル M_{AB} ではなく，M_A (あるいは M_B) が選ばれる可能性が考えられる．ここで議論しているのは Z_A と Z_B の説明変数は相関が大きく，しかも母数の符号が同じであることが想定される場合である．いま仮に M_A が選ばれたとき，M_A を推測に用いたときの母数推定値 $\tilde{\boldsymbol{\beta}}$ には Z_B を除いたことによる bias が入ってしまう．つまり，$Z_A^* = (I-P_O)Z_A$, $Z_B^* = (I-P_O)Z_B$, $\boldsymbol{X}^* = (I-P_O)\boldsymbol{X}$ とするとき，

$$\tilde{\boldsymbol{\beta}} = ({Z_A^*}^T Z_A^*)^{-1} {Z_A^*}^T \boldsymbol{X}^*$$

の期待値は，

$$\begin{aligned} E(\tilde{\boldsymbol{\beta}}) =& ({Z_A^*}^T Z_A^*)^{-1} {Z_A^*}^* E(\boldsymbol{X}^*) \\ =& ({Z_A^*}^T Z_A^*)^{-1} {Z_A^*}^T (Z_A^*\boldsymbol{\alpha} + Z_B^*\boldsymbol{\beta}) \\ =& \boldsymbol{\alpha} + ({Z_A^*}^T Z_A^*)^{-1} {Z_A^*}^T Z_B^*\boldsymbol{\beta} \end{aligned} \tag{8}$$

となる．$({Z_A^*}^T Z_A^*)^{-1} {Z_A^*}^T Z_B^*\boldsymbol{\beta}$ は $Z_B^*\boldsymbol{\beta}$ を Z_A^* に回帰させたときの母数推定値であるから，Z_B^* が Z_A^* との相関が強くければ，そのバイアスはかなり大きくなる．Z_A^* と Z_B^* の説明変数が張る空間が互いに直交していれば，互いの推定量に影響を及ぼさないが，相関が高いとき（多重共線性が疑われるとき）には，一方を含むか否かが他方の推定量に大きな影響を及ぼすことが分かる．

このように，仮に M_A が選ばれたとしても，それは Z_B の効果を検出するだけの十分なデータが得られなかったと考えるのが自然である．選ばれたモデルは，母数を推定するものと言うよりは，観測が得られるメカニズムをより上手く説明するものと考えるのが妥当であろう．言い換えれば，個々の母数推定値に興味を持つよりもむしろ，選ばれたモデル全体としてのパフォーマンスに関心を持つべきである．それ故，予測の際の最適性をもとにモデル選択を行うことは，データを用いて予測を行う際には特に重要となる．

3 モデル選択基準

前節で述べたように，データ数が増えれば増えるほど，説明変数は有意になりやすくなる．つまり，データ数が増えるほど詳細な分析が可能になるのである．しかしながら，実際の場面ではデータ数は限られているのが普通であり，何らかの目的のために最適なモデルを選ぶという立場が合理的であろう．この節では，予測の最適性に基づく基準である，モデル選択基準を紹介する．

3.1 Mallows の C_p

Mallows (1963) の C_p は次のように導出される．n 次元確率変数 $\boldsymbol{X}$ が $n \times p$ 説明変数行列 Z と $n \times q$ 説明変数行列 Z_ω に対して，

$$\boldsymbol{X} = \boldsymbol{\eta} + \boldsymbol{\epsilon} = Z\boldsymbol{\beta} + Z_\omega \boldsymbol{\beta}_\omega + \boldsymbol{\epsilon}, \quad \mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad \mathrm{V}(\boldsymbol{\epsilon}) = \sigma^2 I \tag{9}$$

と表されているとする．ここで $\boldsymbol{\beta}, \boldsymbol{\beta}_\omega$ はそれぞれ $p \times 1, q \times 1$ 未知母数ベクトルとする．いま $\mathrm{E}(\boldsymbol{X}) = Z\boldsymbol{\beta}$ と仮定したときの $\boldsymbol{\beta}$ の推定量 $\tilde{\boldsymbol{\beta}}$ は，

$$\tilde{\boldsymbol{\beta}} = (Z^T Z)^{-1} Z^T \boldsymbol{X} \tag{10}$$

となる．このときのモデルの予測誤差として，scaled sum of squared error である

$$K = \frac{1}{\sigma^2} ||Z\tilde{\boldsymbol{\beta}} - \boldsymbol{\eta}||^2 \tag{11}$$

を採用する．$P = Z(Z^T Z)^{-1} Z^T$ とおくとき，

$$\begin{aligned} K &= \frac{1}{\sigma^2} ||Z(Z^T Z)^{-1} Z^T \boldsymbol{X} - (Z\boldsymbol{\beta} + Z_\omega \boldsymbol{\beta}_\omega)||^2 \\ &= \frac{1}{\sigma^2} || - (I - P) Z_\omega \boldsymbol{\beta}_\omega + P\boldsymbol{\epsilon}||^2 \\ &= \frac{1}{\sigma^2} \{ \boldsymbol{\beta}_\omega^T Z_\omega^T (I - P) Z_\omega \boldsymbol{\beta}_\omega + \boldsymbol{\epsilon}^T P \boldsymbol{\epsilon} \} \end{aligned} \tag{12}$$

となり，

$$\mathrm{E}(K) = \frac{1}{\sigma^2} \boldsymbol{\beta}_\omega^T Z_\omega^T (I - P) Z_\omega \boldsymbol{\beta}_\omega + p \tag{13}$$

が成立する．

一方，

$$\begin{aligned}\mathrm{RSS} &= ||\boldsymbol{X} - P\boldsymbol{X}||^2 = \boldsymbol{e}^T\boldsymbol{e} \\ &= ||(I-P)(Z\boldsymbol{\beta} + Z_\omega\boldsymbol{\beta}_\omega + \boldsymbol{\epsilon})||^2 \\ &= ||(I-P)(Z_\omega\boldsymbol{\beta}_\omega + \boldsymbol{\epsilon})||^2 \\ &= \boldsymbol{\beta}_\omega^T Z_\omega^T (I-P) Z_\omega \boldsymbol{\beta}_\omega + \boldsymbol{\beta}_\omega^T Z_\omega^T (I-P)\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T(I-P)\boldsymbol{\epsilon}\end{aligned} \tag{14}$$

が成立し，

$$\mathrm{E}(RSS) = \boldsymbol{\beta}_\omega^T Z_\omega^T (I-P) Z_\omega \boldsymbol{\beta}_\omega + (n-p)\sigma^2 \tag{15}$$

を得る．

(13) と (15) より $\boldsymbol{\beta}_\omega^T Z_\omega^T (I-P) Z_\omega \boldsymbol{\beta}_\omega$ を消去して，

$$\mathrm{E}(K) = \frac{\mathrm{E}(\mathrm{RSS})}{\sigma^2} - n + 2p = \frac{\mathrm{E}(\boldsymbol{e}^T\boldsymbol{e})}{\sigma^2} - n + 2p \tag{16}$$

となる．σ^2 は未知なので，適当な推定値 $\hat{\sigma}^2$ で置き換えた，

$$C_p = \frac{\mathrm{RSS}}{\hat{\sigma}^2} - n + 2p = \frac{\boldsymbol{e}^T\boldsymbol{e}}{\hat{\sigma}^2} - n + 2p \tag{17}$$

が用いられる．この基準を小さくするものが望ましいモデルとなる．

3.2 Schwarz の SC

この基準は，ベイズ理論から導出されるもので，サンプルサイズ n を無限大にする操作により事前分布に依存しない基準を求めることができる．Schwarz (1978) は非常に簡潔に書かれているので，ここでは解説を加えて詳細に説明する．パラメータ $\boldsymbol{\theta}$ が与えられたときの，確率変数 X の条件付分布として指数型分布族を想定する．つまり条件付分布の密度関数が $\mathcal{R}^1$ 上のルベーグ測度にたいし，

$$f(x,\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T\boldsymbol{y}(x) - b(\boldsymbol{\theta})\right)p(x), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{18}$$

の形を持つとする．なお，分布の自然母数 (natural parameter) $\boldsymbol{\theta}$ とその十分統計量 $\boldsymbol{y}(x)$ は k 次元ベクトルであり，自然母数空間 $\boldsymbol{\Theta}$ は k 次元ユークリッド空間 $\mathcal{R}^k$ の凸部分集合である[1]．

m_j $(j = 1, 2, \cdots, l)$ を $\mathcal{R}^k$ の k_j 次元線形部分空間とするとき，競合するモデルが $m_j \cap \boldsymbol{\Theta}$, $j = 1, 2, \cdots, l$ と表されるとしよう．α_j を j 番目のモデルが真である確率，$\mu_j(\boldsymbol{\theta})$ を j 番目のモデルが真のときの $m_j \cap \boldsymbol{\Theta}$ 上の事前分布の密度関数とするとき，$\boldsymbol{\theta}$ の事前分布は，$\sum_{j=1}^{k} \alpha_j \mu_j(\boldsymbol{\theta})$ により表される．各 $\mu_j(\boldsymbol{\theta})$ が $m_i \cap \boldsymbol{\Theta}$, $j = 1, 2, \cdots, k$ 上有界で，しかも局所的に 0 から一定以上離れているとする．いま Loss function を，m_j を真のモデル，$\delta(x_1, x_2, \cdots, x_n) = \delta(\boldsymbol{x})$ を推定されたモデルとするとき，

$$L(j, \delta(\boldsymbol{x}))) = \begin{cases} 0 & \text{if } \delta(\boldsymbol{x}) = j \\ 1 & \text{otherwise} \end{cases} \tag{19}$$

[1]指数型分布族の性質については，例えば，稲垣 (2003) §.14 を参照されたい

と定義しよう.

以上の設定のもとで，$\boldsymbol{\theta}$ の事後分布は,

$$\frac{\exp\Big(\sum_{i=1}^{n}\big(\boldsymbol{\theta}^T\boldsymbol{y}(x_i)-b(\boldsymbol{\theta})\big)\Big)\sum_{j=1}^{l}\alpha_j\mu_j(\boldsymbol{\theta})}{\int_{\Theta}\exp\Big(\sum_{i=1}^{n}\big(\boldsymbol{\theta}^T\boldsymbol{y}(x_i)-b(\boldsymbol{\theta})\big)\Big)\sum_{j=1}^{l}\alpha_j\mu_j(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{20}$$

となる. 上式の分母はパラメータ $\boldsymbol{\theta}$ と無関係であり, 先に仮定した μ_j の直交性より，α_j^*, μ_j^* をそれぞれ事後確率, 事後確率分布とすると,

$$\alpha_j^*\mu_j^*(\boldsymbol{\theta})=C(\boldsymbol{x})\exp\Big(\sum_{i=1}^{n}\boldsymbol{\theta}^T\boldsymbol{y}(x_i)-nb(\boldsymbol{\theta})\Big)\alpha_j\mu_j(\boldsymbol{\theta}),\quad j=1,2,\cdots,l \tag{21}$$

と表される. この両辺を積分することにより,

$$\int_{m_j\cap\Theta}\alpha_j^*\mu_j^*(\boldsymbol{\theta})d\boldsymbol{\theta}=\alpha_j^*=C(\boldsymbol{x})\int_{m_j\cap\Theta}\exp\Big(\sum_{i=1}^{n}\boldsymbol{\theta}^T\boldsymbol{y}(x_i)-nb(\boldsymbol{\theta})\Big)\alpha_j\mu_j(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{22}$$

となる. これを最大化するモデル m_j がベイズ推定量となる. $C(\boldsymbol{x})$ は j について共通だから, $\bar{\boldsymbol{y}}=\dfrac{\sum_{i=1}^{n}\boldsymbol{y}(x_i)}{n}$ として,

$$SC(m_j;\bar{\boldsymbol{y}},n)=\log\int_{m_j\cap\Theta}\alpha_j\exp\Big(n\big(\boldsymbol{\theta}^T\bar{\boldsymbol{y}}-b(\boldsymbol{\theta})\big)\Big)\mu_j(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{23}$$

をモデル選択基準とするのが Schwarz (1978) の考え方である. このままでは, SC が α_j, μ_j に依存するので, $\bar{\boldsymbol{y}}$, m_j を一定のまま, サンプルサイズ n を無限大にすることにより事前分布への依存しない形にしたのが,

$$SC(m_j;\bar{\boldsymbol{y}},n)=n\sup_{\boldsymbol{\theta}\in m_j\cap\Theta}\big(\boldsymbol{\theta}^T\bar{\boldsymbol{y}}-b(\boldsymbol{\theta})\big)-k_j\log n+R \tag{24}$$

である[2]. ここで, $R=R(m_j;\bar{\boldsymbol{y}},n)$ は $O(1)$ であり n が大きいとき無視される部分を表す.

この基準を求める際の近似計算において, $m_j\cap\Theta$ が真のモデル $\boldsymbol{\theta}_*$ を含まなくてもよいことは注目に値する. この分節の最後に, 正規線形モデルが上述の指数型分布族の形をしていることを示し, 正規線形モデルで使われる SC の式を与える. 平均 μ 分散 σ^2 の正規分布 $N(\mu,\sigma^2)$ の密度関数は,

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \tag{25}$$

より対数尤度は,

$$l(\mu,\sigma^2;x)=\frac{\mu}{\sigma^2}x+\big(-\frac{1}{2\sigma^2}\big)x^2-\big(\frac{\mu^2}{2\sigma^2}+\log\sigma\big) \tag{26}$$

という形を持つ. x_i を $X_i\sim N(\mu_i,\sigma^2)$ の観測値とし, 各 X_i は独立で, $\boldsymbol{z_i}$ を i 番目の観測に伴う p 次元説明変数ベクトルとするとき $\mu_i=\boldsymbol{\beta}^T\boldsymbol{z}_i$ と表されるものとする. このとき尤度は,

$$l_n(\boldsymbol{\beta},\sigma^2;\boldsymbol{x},Z)=\sum_{i=1}^{n}l(\mu_i,\sigma^2;x_i,\boldsymbol{z}_i) \tag{27}$$

$$=\Big(\frac{\boldsymbol{\beta}}{\sigma^2}\Big)^T\sum_{i=1}^{n}x_i\boldsymbol{z}_i+\Big(-\frac{1}{2\sigma^2}\Big)\sum_{i=1}^{n}x_i^2-\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(\boldsymbol{\beta}^T\boldsymbol{z}_i\big)^2-n\log\sigma\Big) \tag{28}$$

[2](23) から (24) を求める証明の概略については, 付録 5.3 を参照されたい.

と書け，

$$\begin{cases} \theta_j = \frac{\beta_j}{\sigma^2}, & j = 1, 2, \cdots, p \\ \theta_{p+1} = -\frac{1}{2\sigma^2} & \\ y_j(\boldsymbol{x}) = \sum_{i=1}^n x_i z_{ij}, & j = 1, 2, \cdots, p \\ y_{p+1}(\boldsymbol{x}) = \sum_{i=1}^n x_i^2 & \end{cases} \tag{29}$$

とすることにより，(18) の分布形を持つ．ここで，z_{ij} は $\boldsymbol{z}_i$ の第 j 成分を表すものとする．正規線形モデルの場合 $\boldsymbol{\beta}$ の最尤推定値は最小自乗推定値と一致し，残差ベクトルを $\boldsymbol{e} = (I - P)\boldsymbol{x}$ とするとき，

$$\hat{\boldsymbol{\beta}} = (Z^T Z)^{-1} Z^T \boldsymbol{x}, \quad \hat{\sigma^2} = \frac{\boldsymbol{x}^T (I - Z(Z^T Z)^{-1} Z^T)\boldsymbol{x}}{n} = \frac{\boldsymbol{e}^T \boldsymbol{e}}{n}. \tag{30}$$

この推定値を (24) に代入すると，

$$SC(p, \boldsymbol{y}, n) = -\frac{n}{2} \log 2\pi \frac{\boldsymbol{e}^T \boldsymbol{e}}{n} - n - \frac{p}{2} \log n \tag{31}$$

となる．この式から必要な部分を取り除き (-2) 倍した，

$$SC(p, \boldsymbol{y}, n) = n \log \frac{\boldsymbol{e}^T \boldsymbol{e}}{n} + p \log n \tag{32}$$

が一般に利用されている (Greene, 2000, p.306)．この基準を小さくするモデルが好ましい．

Schwarz の SC は，分布形 (18) が一見制約的に見えるのだが，これは正準リンク (cannonical link) 関数を持つ一般化線形モデルが共通して持つ形であり，正規線形モデルの他にもポアソン分布を想定した対数線形モデルや二項分布を想定したロジスティックモデルといった有用なモデルに対して適用できる．

3.3　赤池の AIC

様々な場面で，モデル選択基準として用いられるようになってきた AIC (Akaike Information Criterion) について説明する．AIC はモデルの関数形を与えさえすれば計算可能であり，

$$AIC = -2 \times (\text{ モデルの最大対数尤度 }) + 2 \times (\text{ モデルのパラメータ数 }) \tag{33}$$

により与えられる．この基準を小さくするモデル好ましいことになる．この節では，AIC の大まかな導出法を示し[3]，その利用法について述べる．

AIC は真の分布とモデルにより推定される分布との距離を，Kullback-Leibler (K-L) 情報量を用いて測定するものである．$g(x)$ を真の分布，$f(x)$ をモデルから推定された分布とするとき，モデルに関する真の分布の K-L 情報量は，

$$I(g; f) = \int_{-\infty}^{\infty} \log\{\frac{g(y)}{f(y)}\}\, g(y) dy \tag{34}$$

[3]坂元他 (1982) を参考にした．今後の説明では，分布族の support が一定であるとか，積分と微分の交換可能性とかいった正則条件は全て成り立つものとしている．

により定義される[4]. いま, データ $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ を確率変数 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$ の実現値とする. X_i は互いに独立に同じ真の分布 $g(\cdot)$ に従うとする. これに対しモデルは

$$\text{Model}(p):\ \big\{f(\cdot|\boldsymbol{\theta})\ ;\ \boldsymbol{\theta}=(\theta_1,\cdots,\theta_p)\in\Theta_p\big\} \tag{35}$$

とし, このモデルは真の分布を含んでいる, つまり, $\boldsymbol{\theta}^*\in\Theta_p$ が存在して $g(\cdot)=f(\cdot|\boldsymbol{\theta}^*)$ であるとする. Model(p) をデータに当てはめるとき, 対数尤度は

$$l_n(\boldsymbol{\theta})=\sum_{i=1}^{n}\log f(x_i|\boldsymbol{\theta}) \tag{36}$$

と書ける.

$\hat{\boldsymbol{\theta}}_p$ を最尤推定値とするとき, Model(p) の良さの基準として,

$$I\big(g(\cdot);f(\cdot|\hat{\boldsymbol{\theta}}_p)\big)=\int_{-\infty}^{\infty}\log\Big\{\frac{g(y)}{f(y|\hat{\boldsymbol{\theta}}_p)}\Big\}\ g(y)dy \tag{37}$$

$$=\int_{-\infty}^{\infty}\log\big\{g(y)\big\}\ g(y)dy-\int_{-\infty}^{\infty}\log\big\{f(y|\hat{\boldsymbol{\theta}}_p)\big\}\ g(y)dy \tag{38}$$

を採用する[5]. つまり, 推定されたモデルと真のモデルの距離をモデル選択の基準とするのである. (38) の1項目はモデルの選び方に関係なく一定なので除外し[6], 2項目を n 倍した

$$n\int_{-\infty}^{\infty}\log\big\{f(y|\hat{\boldsymbol{\theta}}_p)\big\}\ g(y)dy \tag{39}$$

を考える. これが大きい程よいモデルであると言える. 真の分布 $g(\cdot)$ は未知なので, 上式を推定することにしよう.

$$l_n^*(\boldsymbol{\theta}):=n\ \mathrm{E}_Y\big\{\log f(Y|\boldsymbol{\theta})\big\}=n\ \int_{-\infty}^{\infty}\log f(y|\boldsymbol{\theta})\ g(y)dy \tag{40}$$

と定義すると, (39) は $l_n^*(\hat{\boldsymbol{\theta}}_p)$ と表される. $\hat{\boldsymbol{\theta}}_p$ は確率変数なので $l_n^*(\hat{\boldsymbol{\theta}}_p)$ の期待値

$$l_n^*(p):=\mathrm{E}_{\boldsymbol{X}}\big\{l_n^*(\hat{\boldsymbol{\theta}}_p)\big\}=\int l_n^*(\hat{\boldsymbol{\theta}}_p)\prod_{i=1}^{n}g(x_i)d\boldsymbol{x} \tag{41}$$

を考える. $l_n^*(\hat{\boldsymbol{\theta}}_p)$ を真の値 $\boldsymbol{\theta}^*$ のまわりでティラー展開して,

$$\begin{aligned}l_n^*(\hat{\boldsymbol{\theta}}_p)\doteq l_n^*(\boldsymbol{\theta}^*)&+n(\hat{\boldsymbol{\theta}}_p-\boldsymbol{\theta}^*)^T\mathrm{E}_Y\Big\{\frac{\partial\log f(Y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big\}_{\boldsymbol{\theta}^*}\\&+\frac{1}{2}n(\hat{\boldsymbol{\theta}}_p-\boldsymbol{\theta}^*)^T\mathrm{E}_Y\Big\{\frac{\partial^2\log f(Y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big\}_{\boldsymbol{\theta}^*}(\hat{\boldsymbol{\theta}}_p-\boldsymbol{\theta}^*)\end{aligned} \tag{42}$$

という近似式を得る. 右辺第2項は, $\mathrm{E}_Y\big\{\log f(Y|\boldsymbol{\theta})\big\}$ が $\boldsymbol{\theta}^*$ で最大値を取るため 0 になる. $J_*:=-\mathrm{E}_Y\Big\{\dfrac{\partial^2\log f(Y|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big\}_{\boldsymbol{\theta}^*}$ とおくとき, 尤度理論より漸近的に

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_p-\boldsymbol{\theta}^*)\sim\mathrm{N}(0,{J_*}^{-1}) \tag{43}$$

---

[4] $g\neq f$ ならば $I(g;f)>0$ であり, $I(g;f)=0$ if and only if $f=g$ であることが容易に示される

[5] 上の式で x ではなく y を用いた理由は, データの関数である $\hat{\boldsymbol{\theta}}_p$ とは独立であることを明確にするためである.

[6] AIC は比例尺度ではなく, 間隔尺度である理由はここにある.

が成り立つので[7],

$$n(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}^*)^T J_*(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}^*) \sim \chi^2(p) \tag{44}$$

が近似的に成り立つ. このことから, (42) の両辺の期待値を取ることにより,

$$l_n^*(p) \doteq l_n^*(\boldsymbol{\theta}^*) - \frac{p}{2} \tag{45}$$

を得る.

次に, $l_n(\boldsymbol{\theta}^*) = \sum_{i=1}^n \log f(x_i|\boldsymbol{\theta}^*)$ を $\hat{\boldsymbol{\theta}}_p$ のまわりでティラー展開すると近似式,

$$l_n(\boldsymbol{\theta}^*) \doteq l_n(\hat{\boldsymbol{\theta}}_p) + (\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_p)^T \Big\{\frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big\}_{\hat{\boldsymbol{\theta}}_p} + \frac{1}{2}(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_p)^T \Big\{\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big\}_{\hat{\boldsymbol{\theta}}_p} (\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_p) \tag{46}$$

が得られる. $l_n(\boldsymbol{\theta})$ は $\hat{\boldsymbol{\theta}}_p$ で最大値を達成するので, 上式第 2 項は 0 である. また, $n \to \infty$ のとき, $\hat{\boldsymbol{\theta}}_p \to \boldsymbol{\theta}^*$ *a.s.* が成り立つため,

$$-(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_p)^T \Big\{\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big\}_{\hat{\boldsymbol{\theta}}_p} (\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_p) \sim \chi^2(p) \tag{47}$$

が近似的に成り立つ. そこで, (46) の両辺の期待値をとることにより近似的に,

$$l_n^*(\boldsymbol{\theta}^*) \doteq \mathrm{E}_{\boldsymbol{X}}[l_n(\hat{\boldsymbol{\theta}}_p)] - \frac{p}{2} \tag{48}$$

を得る. (45) に (48) を代入することにより,

$$l_n^*(p) \doteq \mathrm{E}_{\boldsymbol{X}}[l_n(\hat{\boldsymbol{\theta}}_p)] - p \tag{49}$$

という近似式が得られる. $\mathrm{E}_{\boldsymbol{X}}[l_n(\hat{\boldsymbol{\theta}}_p)]$ をその推定値 $l_n(\hat{\boldsymbol{\theta}}_p)$ で置き換え, (-2) 倍した,

$$(-2)l_n(\hat{\boldsymbol{\theta}}_p) + 2p \tag{50}$$

を赤池の情報量基準 (AIC) と呼ぶのである.

以上の導出は, モデルが真の分布を含む, つまり $g(\cdot) = f(\cdot|\boldsymbol{\theta}^*)$ の場合に限り有効である. AIC はモデルが真の分布を含まなくても, AIC は (50) により与えられる. このとき暗黙に, “データ数 n が大きくなるにしたがいパラメータ数 p もそれに応じて大きくなり, モデルの中で真の分布にいくらでも近い分布が存在する” という仮定をおいているのである[8]. 言い換えれば, AIC は真の分布を含む（あるいはモデルの中に, 真の分布をかなり良く近似する分布が存在する）いくつかのモデルの中で最良のものを見つけるための基準と言える.

AIC をその導出法に基づき厳密に適用するとなると, かなり制約的となり, 実質的には従来の尤度比検定とほとんど変わらないものとなる[9]. 赤池 (1976) 自身, “AIC の利用に際しては何等の数表も主観的な議論も必要としなかった” ことを特長に挙げているように, 従来の尤度比検定に代わる簡便法であり, 理論的な厳密性よりも道具としての汎用性から提唱されたもの言える. このことは, 坂元他 (1982) を見ても明らかである. そこでは, 尤度比検定が可能な場面での AIC の利用について述べている. AIC を道具として割り切るとき, $F(x_t|\mathbf{z}_t;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}$

[7]例えば, 稲垣 (2003) を参照されたい.

[8]詳しくは稲垣他 (1977), 竹内 (1976) を参照されたい

[9]稲垣他 (1977) は, 尤度比検定の枠組みから AIC および C_p を捉え, 3 者が漸近的に同等であることを厳密に示している.

と $G(x_t|\mathbf{z}_t;\gamma),\gamma\in\mathbf{\Gamma}$ の2つのモデルのうち，どちらが真の分布に近いかを判定することも可能になる．仮にどちらか一方が漸近的に真のモデルを含まないとしても，そのモデルの最大尤度の部分が小さくなるため，モデルの候補から自然に脱落するであろうというものである．

正規線形モデルでの変数選択の場面では，前の基準と同じ設定で，

$$AIC = n\log\frac{\boldsymbol{e}^T\boldsymbol{e}}{n} + 2p \tag{51}$$

となる．この基準を小さくするモデルが好ましい．

3.4 Adjusted R^2 ; $\bar{R}^2$

この基準は，これまでに紹介してきたものとは違い，何らかの最適性から導出されたものではない．とは言え，これまでの基準が持っている性質を共有している．つまり，モデルのパラメータ数に応じたペナルティが与えられているという意味で，モデル選択基準の一つとして扱われるのが一般的である．

Mallows' C_p と同じ設定で，モデル $\mathrm{E}(\boldsymbol{X}) = Z\boldsymbol{\beta}$ をあてはめた時の自由度調整済決定係数 (Adjusted R^2; $\bar{R}^2$) は，$P = Z(Z^TZ)^{-1}Z^T$ として，

$$\bar{R}^2 = 1 - \frac{\boldsymbol{x}^T(I-P)\boldsymbol{x}/(n-p)}{\boldsymbol{x}^T(I-P_{\mathbf{1}})\boldsymbol{x}/(n-1)} = 1 - \frac{\boldsymbol{e}^T\boldsymbol{e}/(n-p)}{\sum_{i=1}^n(x_i-\bar{x})^2/(n-1)} = 1 - \frac{n-1}{n-p}R^2 \tag{52}$$

となる．ここで，Z は $n\times p$ 説明変数行列，$P_{\mathbf{1}} = \mathbf{1}(\mathbf{1}^T\mathbf{1})\mathbf{1}^T = (1/n)_{ij}$ である．モデルの適合度として，

$$R^2 = 1 - \frac{\boldsymbol{x}^T(I-P)\boldsymbol{x}}{\boldsymbol{x}^T(I-P_{\mathbf{1}})\boldsymbol{x}} = 1 - \frac{\boldsymbol{e}^T\boldsymbol{e}}{\sum_{i=1}^n(x_i-\bar{x})^2} \tag{53}$$

を用いると，説明変数を増やせば必ず適合度は上がるので，決定係数 R^2 のままではモデル選択の基準とはなりえないのである．

4 結び

まず，前節で紹介したモデル選択基準を，正規線形モデルの枠組みの中に限定して比較を行おう．得られたデータ $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ が $\boldsymbol{X} \sim \mathrm{N}(Z\boldsymbol{\beta}, \sigma^2 I)$ の観測値であると仮定したとき，各基準は次の形に表される．

$$C_p = \frac{\boldsymbol{e}^T\boldsymbol{e}}{\hat{\sigma}^2} + 2p \tag{54}$$

$$SC = n\log\frac{\boldsymbol{e}^T\boldsymbol{e}}{n} + p\log n \tag{55}$$

$$AIC = n\log\frac{\boldsymbol{e}^T\boldsymbol{e}}{n} + 2p \tag{56}$$

$$\bar{R}^2 = 1 - \frac{\boldsymbol{e}^T\boldsymbol{e}/(n-p)}{\sum_{i=1}^n(x_i-\bar{x})^2/(n-1)} \tag{57}$$

上で，Z は $n\times p$ 説明変数行列，I は n 次単位行列，$\boldsymbol{e} = (I - Z(Z^TZ)^{-1}Z^T)\boldsymbol{x}$ は残差を表すものとする．C_p, SC, AIC の場合は値が小さいほうが望ましく，$\bar{R}^2$ は大きいほうが望まし

い．どの基準を採用するにせよ，包含関係にある (Nested) 2 つのモデルを比較する際は，検定を行うことと同値である．

AIC を例にとろう．帰無仮説 H_0 が 対立仮説 H_1 の制約になっているときを考える．H_0 のパラメータ数を p としその場合の残差を $\boldsymbol{e}_0$, H_1 のパラメータ数を $p+q$ としその場合の残差を $\boldsymbol{e}_1$ とすれば，$n\log\frac{\boldsymbol{e}_0^T\boldsymbol{e}_0}{n}+2p$ と $n\log\frac{\boldsymbol{e}_1^T\boldsymbol{e}_1}{n}+2p+2q$ との差，つまり，

$$n\log\frac{\boldsymbol{e}_0^T\boldsymbol{e}_0}{n}-n\log\frac{\boldsymbol{e}_1^T\boldsymbol{e}_1}{n}-2q \tag{58}$$

が正ならば H_0 を，負ならば H_1 を採択することを意味する．それ故，尤度比を用いて棄却限界値を $2q$ とする検定と全く同じになる．他の基準の場合も同様で，統計量と棄却限界値が変化するだけである．このようにモデル選択基準とは，仮説検定を行う際，自動的に棄却限界値を与える方式と捉えることも可能である．

導出法から見るとき，$\bar{R}^2$ は何ら最適性を持たないため，やや採用し難い．C_p と SC は仮定されたモデルと真のモデルが違っていることが許容されるため，モデル同士が包含関係を持たなくても利用可能である．これは，C_p と SC が AIC よりも適用範囲が狭いことに起因している．もちろん，竹内 (1976) が感想として述べた，“真のモデルを漸近的に含まないモデルは，最大尤度の部分が小さくなるためモデル候補から勝手に脱落する” という考えを採用すれば，AIC もまた包含関係を持たない Non-nested なモデルの比較にも用いることが可能になる．

稲垣他 (1978) が厳密に示したように，AIC と C_p は共に，漸近的に尤度比検定と同等である．また，SC も AIC とは棄却限界値の異なる尤度比検定として解釈できる．一般に尤度比検定は漸近的な最適性を持つことが示されている．SC は，パラメータ数によるペナルティー部分が，AIC の $2p$ に対し $p\log n$ であり，n が大きければ $2p$ よりはるかに大きくなる．その結果，AIC よりも母数節約的になりすぎる．

次に，サンプル数がそれほど大きくない時にどれを選ぶかということを考えよう．AIC はデータ数 n を無限大にしたときに得られるもので n がそれほど大きくない場面での有効性に問題があること，AIC の導出法は一般的な尤度理論に基づくものであることから，σ^2 の推定値をどのように与えるかの任意性があるものの，正規線形モデルでの変数選択では C_p の利用が最適であると結論づけてもよいだろう．

正規線形モデルの枠内での，モデル選択基準としての優位性は持たないものの，AIC は尤度を指定しさえすれば適用可能であるため，その適用範囲は他の基準を圧倒して最も広い．尤度を指定することを制約的と見ることも出来るが，指定しなければほとんどの推測が不可能になることを考えれば，それほどでもないと言えよう．何度か述べたように，道具として割り切ったときの AIC は，正規線形モデルに限らず，ありとあらゆるモデル間の比較に用いることのできる便利な基準 (道具) であることを最後に強調しておきたい．

5　付録

5.1　逆行列の公式

$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ を正方行列の分割とし，A, D は対角行列とする．このとき

$$|A|\neq 0,\ |D-CA^{-1}B|\neq 0$$

ならば，次の逆行列の公式が成り立つ．

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1}+A^{-1}BW^{-1}CA^{-1} & -A^{-1}BW^{-1} \\ -W^{-1}CA^{-1} & W^{-1} \end{bmatrix}, \text{where } W = D - CA^{-1}B. \quad (59)$$

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} V^{-1} & -V^{-1}BD^{-1} \\ -D^{-1}CV^{-1} & D^{-1}+D^{-1}BV^{-1}CD^{-1} \end{bmatrix}, \text{where } V = A - BD^{-1}C. \quad (60)$$

5.2 正規線形モデルでの推測

前の分節で述べた逆行列の公式を用いて，説明変数を追加する効果について述べる．$\boldsymbol{X}$ を n 次元正規分布に従う確率変数とし，Z_O, Z_1 をそれぞれ $n \times k$, $n \times p$ 説明変数行列，$\boldsymbol{\beta}_O$, $\boldsymbol{\beta}_1$ を k, p 次元母数ベクトルとするとき，

$$\boldsymbol{X} \sim N(Z_O\boldsymbol{\beta}_O + Z_1\boldsymbol{\beta}_1, \sigma I) \quad (61)$$

が成り立つとしよう．ただし，Z_O, Z_1 の合わせて $k+p$ 個の説明変数ベクトルは一次独立とする．

ここで，$H_O : \boldsymbol{\beta}_1 = \mathbf{0}$, $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$ を検定しよう．検定統計量は，$H_O : \boldsymbol{\beta}_1 = \mathbf{0}$ を仮定したとき残差平方和，

$$\boldsymbol{X}^T(I - Z_O(Z_O^TZ_O)^{-1}Z_O)\boldsymbol{X} \quad (62)$$

と $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$ を仮定したときの残差平方和，

$$\boldsymbol{X}^T(I - [Z_OZ_1]([Z_OZ_1]^T[Z_OZ_1])^{-1}[Z_OZ_1])\boldsymbol{X} \quad (63)$$

の差を用いる．式 (63) を，公式 (59) を用いて展開しよう．

$$([Z_OZ_1]^T[Z_OZ_1])^{-1} = \begin{bmatrix} Z_O^TZ_O & Z_O^TZ_1 \\ Z_1^TZ_O & Z_1^TZ_1 \end{bmatrix}^{-1}$$

は，$W = Z_1^T(I - Z_O(Z_O^TZ_O)^{-1}Z_O^T)Z_1$ と置くとき，

$$\begin{bmatrix} (Z_O^TZ_O)^{-1}+(Z_O^TZ_O)^{-1}Z_O^TZ_1W^{-1}Z_1^TZ_O(Z_O^TZ_O)^{-1} & -(Z_O^TZ_O)^{-1}Z_O^TZ_1W^{-1} \\ -W^{-1}Z_1^TZ_O(Z_O^TZ_O)^{-1} & W^{-1} \end{bmatrix} \quad (64)$$

なので，H_1 を仮定したときの残差平方和 (63) は，

$$\boldsymbol{X}^T(I - P_O - (I - P_O)Z_1W^{-1}Z_1^T(I - P_O))\boldsymbol{X} \quad (65)$$

となる．よって，残差平方和の差は，

$$\boldsymbol{X}^T((I - P_O)Z_1W^{-1}Z_1^T(I - P_O))\boldsymbol{X}, \quad (66)$$

いま，$Z_1^* = (I - P_O)Z_1$, $\boldsymbol{X}^* = (I - P_O)\boldsymbol{X}$ とするとき，$W = Z_1^{*T}Z_1^*$ であることから，

$$\boldsymbol{X}^T((I - P_O)Z_1W^{-1}Z_1^T(I - P_O))\boldsymbol{X} = \boldsymbol{X}^{*T}Z_1^*(Z_1^{*T}Z_1^*)^{-1}Z_1^{*T}\boldsymbol{X}^*. \quad (67)$$

よって, (66) は, $\boldsymbol{X}^*$ を Z_1^* に回帰させたときの回帰による変動と解釈できる. $Z_1^*(Z_1^{*T}Z_1^*)^{-1}Z_1^{*T}$ はベキ等行列であるので, $\boldsymbol{X}^{*T}Z_1^*(Z_1^{*T}Z_1^*)^{-1}Z_1^{*T}\boldsymbol{X}^*$ は非心度 $\boldsymbol{\eta}^{*T}Z_1^*(Z_1^{*T}Z_1^*)^{-1}Z_1^{*T}\boldsymbol{\eta}^*$, 自由度 p のカイ自乗分布に従う. ここで, $\boldsymbol{\eta}^* = (I - P_O)\boldsymbol{\eta}$ を表す.

H_O の仮定のもとでは, $\boldsymbol{\eta}^* = (I - P_O)\boldsymbol{\eta} = (I - P_O)Z_O\boldsymbol{\beta}_O = \boldsymbol{0}$ であり, $\boldsymbol{X}^{*T}(I - Z_1^*(Z_1^{*T}Z_1^*)^{-1}Z_1^{*T})\boldsymbol{X}^*$ は自由度 $n-k-p$ のカイ自乗分布に従うことより,

$$\frac{\boldsymbol{X}^{*T}Z_1^*(Z_1^{*T}Z_1^*)^{-1}Z_1^{*T}\boldsymbol{X}^*}{\boldsymbol{X}^{*T}(I - Z_1^*(Z_1^{*T}Z_1^*)^{-1}Z_1^{*T})\boldsymbol{X}^*} \sim F(p, n-k-p) \tag{68}$$

を用いて検定が行われる.

検定の場合に限らず, $\boldsymbol{\beta}_1$ についての推測では, $\boldsymbol{X}$, Z_1 の代わりに $\boldsymbol{X}^* = (I - P_O)\boldsymbol{X}$, $Z_1^* = (I - P_O)Z_1$ を用いれば良いことが, 上の計算と同様にして導かれることに注意されたい. 特に,

$$\hat{\boldsymbol{\beta}}_1 = (Z_1^{*T}Z_1^*)^{-1}Z_1^{*T}\boldsymbol{X}^* \tag{69}$$

が成り立つ.

5.3 SC の導出法

この分節では, (24) がどのように導かれるかを解説する. $A = \sup_{\boldsymbol{\theta}\in\Theta}(\boldsymbol{\theta}^T\boldsymbol{y} - b(\boldsymbol{\theta}))$ とし, 最大を達成する $\boldsymbol{\theta}$ 値を $\boldsymbol{\theta}_0$ で表す. いま, $0 < \rho < e^A$ を満たす ρ を適当に定め, $\boldsymbol{\theta}_0$ の近傍

$$W_\rho = \{\boldsymbol{\theta} : \exp\{\boldsymbol{\theta}^T\boldsymbol{y} - b(\boldsymbol{\theta})\} \geq \rho\} \tag{70}$$

を考える. 指数型分布族の性質より $b(\boldsymbol{\theta})$ は C^∞ 級 (無限回連続的微分可能関数) であり, $\dfrac{\partial^2 b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ は正定値行列である. それ故 W_ρ は, $\boldsymbol{\theta}_0$ をその内点として含む, 有界凸閉集合であることが分かる. $\boldsymbol{\theta} \in W_\rho$ に対し, $\boldsymbol{\theta}^T\boldsymbol{y} - b(\boldsymbol{\theta})$ を $\boldsymbol{\theta}_0$ の周りで展開すると,

$$\begin{aligned}\big(\boldsymbol{\theta}^T\boldsymbol{y} - b(\boldsymbol{\theta})\big) &= \big(\boldsymbol{\theta}_0^T\boldsymbol{y} - b(\boldsymbol{\theta}_0)\big) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\Big\{\frac{\partial^2 b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big\}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= A - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\Big\{\frac{\partial^2 b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big\}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\end{aligned} \tag{71}$$

が成り立つ. ここで, $\boldsymbol{\theta}_1$ は $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0 = \alpha(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, $0 < \alpha < 1$ を満たすものである. W_ρ 上 $\dfrac{\partial^2 b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ の固有値の最大値を λ_1, 最小値を λ_2 とすれば, $\boldsymbol{\theta} \in W_\rho$ に対し

$$A - \lambda_1||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 < \big(\boldsymbol{\theta}^T\boldsymbol{y} - b(\boldsymbol{\theta})\big) < A - \lambda_2||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 \tag{72}$$

が成り立つ. これより,

$$\int_{W_\rho}\exp\big\{n\big(\boldsymbol{\theta}^T\boldsymbol{y} - b(\boldsymbol{\theta})\big)\big\}\mu(\boldsymbol{\theta})d\boldsymbol{\theta} < \int_{W_\rho}\exp\big\{n(A - \lambda_2||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2)\big\}\mu(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{73}$$

を得る. L_n ノルムの性質として, V を確率変数とするとき一般に $\lim_{n\to\infty}(E[V^n])^{1/n} = \sup V$ が成り立つので, n を十分大きくとれば,

$$\int_{\Theta}\exp\big\{n\big(\boldsymbol{\theta}^T\boldsymbol{y} - b(\boldsymbol{\theta})\big)\big\}\mu(\boldsymbol{\theta})d\boldsymbol{\theta} < \int_{\Theta}\exp\big\{n(A - \lambda_2||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2)\big\}\mu(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{74}$$

が成り立つ. $\mu(\boldsymbol{\theta})$ ルベーグ測度とするとき[10], つまり $\mu(\boldsymbol{\theta}) \equiv 1$ とするとき,

$$\exp\{SC(\Theta; \boldsymbol{y}, n)\} < \exp\{nA\}\left(\frac{\pi}{n\lambda_2}\right)^{(k/2)}, \tag{75}$$

よって,

$$SC(\Theta; \boldsymbol{y}, n)\} < nA - (k/2)\log n + R_2, \tag{76}$$

ここで R_2 は $O(1)$ の部分, となる. 同様にして,

$$SC(\Theta; \boldsymbol{y}, n)\} > nA - (k/2)\log n + R_1, \tag{77}$$

ここで R_2 は $O(1)$ の部分, も成立するので, 最終的に (24) を得る. 以上の説明は, Θ についてのものだが, $m_j \cap \Theta$ の場合も同様に証明できる. なぜなら, モデル m_j は $\mathcal{R}^k$ 内の k_j 次元線形部分空間なので, パラメータを適当に一次変換してやれば,

$$\left\{ \exp\left(\boldsymbol{\theta}^T \boldsymbol{y}(x) - b(\boldsymbol{\theta})\right) p(x),\ \boldsymbol{\theta} \in m_j \cap \Theta \right\}$$

は, k_j 次元ベクトル $\boldsymbol{\theta}'$, $\boldsymbol{y}'(x)$ と, k_j 次元ユークリッド空間内の凸部分集合 Θ_j が導き出されて,

$$\left\{ \exp(\boldsymbol{\theta}'^T \boldsymbol{y}'(x) - b(\boldsymbol{\theta}')\ ) p'(x),\ \boldsymbol{\theta}' \in \Theta_j \right\}, \tag{78}$$

という形に書き表されるからである.

参考文献

[1] Akaike, H. (1972) Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, 267-281.

[2] 赤池 弘次 (1976) 情報量基準 AIC とは何か, 数理科学, No.153, 5-11.

[3] 秋岡 弘紀 (2002) 電気事業の完全民営化に関する一考察, 関西大学経済論集, 第 52 巻, 393-434.

[4] 稲垣 宣生, 尾形 良彦 (1978) 尤度統計量の弱収束とその応用, 数学, 第 30 巻, 193-206.

[5] 稲垣 宣生 (2003) 数理統計学, 改訂版, 裳華房.

[6] 坂元 慶行, 石黒 真木夫, 北川 源四郎 (1982) 情報量統計学, 共立出版.

[7] 佐和 隆光 (1979) 回帰分析, 朝倉書店.

[8] 竹内 啓 (1976) 情報統計量の分布とモデルの適切さの基準, 数理科学, No.153, 12-18.

[9] 百合草 裕康 (2001) キャッシュ・フロー会計の有用性, 中央経済社.

[10] Greene, W. H. (2000) *Econometric Analysis*, 4th edition. Prentice Hall.

[10] どんな事前分布を仮定しても, データ数 n を大きくすれば結果は変わらない.

[11] Gourieroux, C. and Monfort, A. (1994) Testing Non-nested Hypotheses, *Handbook of Econometrics*, Vol.4, 2585-2633.

[12] Mallows, C. L. (1973) Some Comments on C_p, *Technometrics*, Vol.15, 661-675.

[13] Schwarz, G. (1978) Estimating the Dimension of a Model, *The Annals of Statistics*, Vol.6, 461-464.