Ⅳ-186　マローズのC_p-統計量を基礎にした予測誤差基準について

建設省　正員　○桐越　信
北海道大学　正員　五十嵐日出夫
北海道大学　正員　山形耕一

1. はじめに

重回帰モデルは土木計画の種々の分野で予測モデルとして数多く用いられている。重回帰モデルの構築にあたって直面する最も重要でかつ困難な問題は，モデル構築時にモデルに取り入れる説明変数をどのように選択するべきかという問題である。モデル構築以前に，説明変数があらかじめ決定されていることはほとんどない。一般には，モデル構築時に説明変数のいくつかの候補群が，モデル構築者の経験や勘などによって主観的に用意され，そのなかから試行錯誤の後に，なんらかの客観的基準により最も望ましいと思われる説明変数の組合せが選択される。

マローズのC_p-統計量は説明変数選択の客観的基準のひとつであり，モデル構築時の偏りを明示的に定式化のなかにとり入れているところにその特徴がある。本研究は，このマローズのC_p-統計量を基礎にして，さらにモデルに取り入れられた説明変数の予測時における予測誤差をも考慮した予測誤差基準について考察するものである。

2. マローズのC_p-統計量[1)]

マローズのC_p-統計量では，モデル構築時の基本構造式を

$$y_i=\beta_0+\sum_{k=1}^{p}\beta_k x_{ik}+\xi_i+u_i \qquad (1)$$

と設定する。(1)式でξ_iは想定されたモデルの未知の偏りを表わし，u_iは正規分布$N(0,\eta^2)$に従うと仮定されている。ここで，ξ_iについて，

$$\sum_{i=1}^{n}x_{ik}\xi_i=0 \quad ,k=1,2,\cdots,p \qquad (2)$$

$$\sum_{i=1}^{n}\xi_i=0 \qquad (3)$$

を仮定する。(2)，(3)式を仮定すると偏回帰係数β_0，β_kの最小2乗推定量b_0，b_kはやはり不偏推定量となる。説明変数が$(x_{i1},x_{i2},\cdots,x_{ip})$をとるときの予測対象$y_i^*$を(1)式に従って(4)式で定義する。

$$y_i^*=\beta_0+\sum_{k=1}^{p}\beta_k x_{ik}+\xi_i+u_i^* \qquad (4)$$

(4)式において，u_i^*は(1)式のu_iと独立に，かつすべて互いに独立に正規分布$N(0,\eta^2)$に従うものとする。このとき予測値$\hat{y}_i$は，

$$\hat{y}_i=b_0+\sum_{k=1}^{p}b_k x_{ik} \qquad (5)$$

で与えられるので，

$$C_p=\sum_{i=1}^{n}(y_i^*-\hat{y}_i)^2 \qquad (6)$$

とおくと，

$$E[C_p]=(n+p+1)\eta^2+\sum_{i=1}^{n}\xi_i^2 \qquad (7)$$

となる。

一方，この場合，残差平方和の期待値$E[RSS]$は，

$$E[RSS]=(n-p-1)\eta^2+\sum_{i=1}^{n}\xi_i^2 \qquad (8)$$

となるので，(7)，(8)式より，

$$E[C_p]=E[RSS]+2(p+1)\eta^2 \qquad (9)$$

となる。ここで，η^2について不偏推定量$\hat{\eta}^2$が求められるなら，

$$C_p=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2(p+1)\hat{\eta}^2 \qquad (10)$$

と書きかえて，(10)式を(9)式の推定量とすることができる。$\hat{\eta}^2$については，一般には，

$$\min\hat{\eta}_p^2=\min\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-p-1)\right] \qquad (11)$$

が採用されている。

(10)式はモデルが偏りをもつことを前提にしたうえで，そのモデルの望ましさを評価する基準であると考えることができる。(10)式で，第1項は説明変数の数の増加に伴って小さくなるが，第2項は大きくなるので，マローズのC_p-統計量はQSS，PSS，AICなどと同じようにモデルの簡略さと適合性とのトレード・オフの関係を表わしていると考えることができる。QSS[2)]，PSS[3)]，AIC[4)]はそれぞれ(12)，(13)，(14)式で表現される。

$$QSS=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2(p+1)\hat{\sigma}^2$$
$$=(n+p+1)\hat{\sigma}^2$$
$$=\frac{(n+p+1)}{(n-p-1)}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad (12)$$

$$PSS=\sum_{i=1}^{n}(y_i-\hat{\hat{y}}_i)^2 \qquad (13)$$

ここで，$\hat{\hat{y}}_i=b_{0i}+\sum_{k=1}^{p}b_{ki}x_{ik}$

$$AIC=n\ln\hat{\hat{\sigma}}^2+2(p+1) \qquad (14)$$

(12)，(13)，(14)式では，モデル構築時の基本構造式を，

$$y_i=\beta_0+\sum_{k=1}^{p}\beta_k x_{ik}+\varepsilon_i \qquad (15)$$

と設定し，ε_iは正規分布$N(0,\sigma^2)$に従うと仮定されている．(12)式のQSSにおける$\hat{\sigma}^2$はσ^2の不偏推定量である．(13)式のPSSにおけるb_{0i}, b_{ki}はi組目のデータ$(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$を除いた$(n-1)$組のデータより求める偏回帰係数である．(14)式のAICにおける$\hat{\hat{\sigma}}^2$はσ^2の最尤推定量である．したがって$\hat{\hat{\sigma}}^2$は不偏推定量ではなく偏推定量であり，漸近的不偏推定量である．

3. C_p-統計量を基礎にした予測誤差基準

(4)式の予測対象y_i^0の予測にあたって，モデルに取り入れられた説明変数の予測時における予測誤差も考慮して，それをΔx_{ik}[5),6)]とすると，予測値は

$$\hat{y}_i^{0*}=b_0+\sum_{k=1}^{p}b_k(x_{ik}+\Delta x_{ik}) \qquad (16)$$

となる．したがって，予測誤差v_i^{0*}は

$$\begin{aligned}v_i^{0*}&=y_i^0-\hat{y}_i^{0*}\\&=\beta_0+\sum_{k=1}^{p}\beta_k x_{ik}+\varepsilon_i+u_i^0\\&\quad-\{b_0+\sum_{k=1}^{p}b_k(x_{ik}+\Delta x_{ik})\}\\&=(\beta_0-b_0)+\sum_{k=1}^{p}(\beta_k-b_k)x_{ik}\\&\quad+\varepsilon_i+u_i^0-\sum_{k=1}^{p}b_k\Delta x_{ik}\end{aligned} \qquad (17)$$

となる．ここで，Δx_{ik}の期待値$E[\Delta x_{ik}]$について

$$E[\Delta x_{ik}]=0, \quad i=1,2,\cdots,n, \quad k=1,2,\cdots,p \qquad (18)$$

と仮定すると，

$$\begin{aligned}E[\sum_{i=1}^{n}v_i^{0*2}]&=E[\sum_{i=1}^{n}(y_i^0-\hat{y}_i^{0*})^2]\\&=(n+p+1)\eta^2+\sum_{i=1}^{n}\varepsilon_i^2\\&\quad+\sum_{i=1}^{n}V[\sum_{k=1}^{p}b_k\Delta x_{ik}]\\&=E[RSS]+2(p+1)\eta^2\\&\quad+\sum_{i=1}^{n}V[\sum_{k=1}^{p}b_k\Delta x_{ik}]\end{aligned} \qquad (19)$$

となる．ここで，Δx_{ik}と$\Delta x_{ik'}$とが互いに独立であるとすると，(19)式は

$$E[\sum_{i=1}^{n}v_i^{0*2}]=E[RSS]+2(p+1)\eta^2+\sum_{i=1}^{n}\sum_{k=1}^{p}V[\Delta x_{ik}]\{(E[b_k])^2+V[b_k]\} \qquad (20)$$

となる．

実際には，(20)式を

$$K=\sum_{i=1}^{n}(y_i-\hat{\hat{y}}_i)^2+2(p+1)\hat{\eta}^2+\sum_{i=1}^{n}\sum_{k=1}^{p}V[\Delta x_{ik}]\{b_k^2+\hat{V}(b_k)\} \qquad (21)$$

として推定する．ここで(21)式のKを予測誤差基準とよぶことにする．

(21)式において，説明変数の数の増加に伴って，第1項は小さくなるが，第2項は大きくなるので，第1項+第2項はモデルの簡略さと適合性とのトレード・オフの関係を表わしている．このことは，第1項+第2項が(10)式のC_pと同じものであることからも明らかである．ところが(21)式では，説明変数の予測時における予測誤差のために，説明変数の数の増加に伴って第3項も大きくなる．このことは，(21)式のKが，説明変数の数の増加に対して(10)式のC_pよりもさらに大きなペナルティをつけることを示している．

4. おわりに

本研究では，重回帰モデル構築における説明変数選択のひとつの客観的基準であるマローズのC_p-統計量をとりあげ，これにさらにモデルに取り入れられた説明変数の予測時における予測誤差をも明示的に考慮した予測誤差基準について考察を行なった．説明変数の予測誤差をも考慮している点で予測誤差基準は予測モデルを評価するのにふさわしい評価方法といえるが，予測誤差基準によればC_p-統計量によるよりもさらに一層予測モデルとしての重回帰モデルの簡略化の方向が示唆される．

〈参考文献〉

1) 竹内啓；現象と行動のなかの統計数理，新曜社，pp.74~85，1980年

2) 奥野忠一，芳賀敏郎；続多変量解析法，日科技連出版社，P.73，1976年

3) 芳賀敏郎，竹内啓，奥野忠一；重回帰分析における変数選択の新しい規準，品質，Vol.6，No.2，pp.35~40，1976年

4) 佐和隆光；回帰分析，朝倉書店，pp150~153，1979年

5) 山形耕一，桐越信；交通需要予測の予測精度について，第1回土木計画学研究発表会講演集，pp.52~57，1979年

6) 桐越信，他；予測誤差基準による発生重回帰モデルの評価方法に関する研究，第3回年講集，pp79~80