学位論文要旨

An unbiased C_p type criterion for ANOVA model with a tree order restriction
(Tree order restriction が課せられた ANOVA モデルにおける不偏な C_p タイプの規準量)

稲津　佑

いくつかの群間の平均構造の差異をモデル化する手法のひとつに ANOVA モデルがある. ANOVA モデルは実データ解析の分野でも頻繁に用いられる重要なモデルである. また, 解析者はいくつもの候補となるモデルを想定することができ, その中から最適なモデルを選ぶこと, すなわちモデル選択も重要な問題である. これまでに, Akaike [1] による AIC 規準や Mallows [3] による C_p 規準等の様々なモデル選択規準が提案されてきた. AIC 規準は -2 倍の最大対数尤度に, C_p 規準は残差二乗和をフルモデルにおける分散の推定量で割ったものに, 罰則項として 2 倍のパラメータ数を加えたシンプルな定義であり, リスクに対する漸近不偏性をもつ規準量である. また, Fujikoshi and Satoh [2] により導出された修正 C_p 規準, MC_p はリスクに対する不偏性を持つ規準量である. しかしながら, C_p および MC_p のリスクに対する (漸近) 不偏性は特定の正則条件下で成立するものであり, パラメータに対する順序制約等がある場合は, (漸近) 不偏性は必ずしも成立しない.

本研究では, パラメータに順序制約が課せられた ANOVA モデルにおける, リスクに対する不偏性を持った C_p 規準を導出する. 以下の設定を考える. Y_{ij} を第 i 群における第 j 番目の個体の観測値とする. ただし, $1 \leq i \leq k^*$ とし, 各 i に対し, $j = 1, \ldots, N_i$ とする. また, $k^* \geq 2$ とする. 更に, $N = N_1 + \cdots + N_{k^*}$, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})'$, $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_{k^*}')'$ および $\boldsymbol{N} = (N_1, \ldots, N_{k^*})'$ と記号を定義する.

次に, $Y_{11}, \ldots, Y_{k^*N_{k^*}}$ は互いに独立な確率変数であり, Y_{ij} は

$$Y_{ij} \sim N(\mu_{i,*}, \sigma_*^2), \tag{1}$$

に従うとする. ここで, $\mu_{i,*}$ および σ_*^2 は未知の真値であり, $\mu_{i,*} \in \mathbb{R}$, $\sigma_*^2 > 0$ を満たすとする. 式 (1) で与えられるモデルを真のモデルとする.

次に, 候補モデルを定義する. $Q_1, \ldots, Q_k$ を空でない互いに素な集合とし, $Q_1 \cup \cdots \cup Q_k = \{1, 2, \ldots, k^*\}$ を満たすとする. ただし, $2 \leq k \leq k^*$ とする. ここで, $Y_{11}, \ldots, Y_{k^*N_{k^*}}$ は互いに独立な確率変数で

$$Y_{ij} \sim N(\mu_i, \sigma^2), \tag{2}$$

に従うとする. ただし, $\mu_1, \ldots, \mu_{k^*}$ と $\sigma^2 (> 0)$ は未知パラメータである. ここで, $\mu_1, \ldots, \mu_{k^*}$ に対し,

$$^\forall s \in \{1, \ldots, k\},\ ^\forall u_1, u_2 \in Q_s, \quad \mu_{u_1} = \mu_{u_2}, \tag{3}$$

および

$$^\forall t \in \{2, \ldots, k\},\ ^\forall \nu \in Q_t, \quad \mu_q \leq \mu_\nu, \tag{4}$$

を仮定する. ただし, $q \in Q_1$ である. このとき, モデル (2) に (3) および (4) を仮定したモデルを候補モデルとする. 特に, 順序制約 (4) は Tree Ordering (TO), もしくは tree order restriction と呼ばれる. あるいは, $\boldsymbol{X}_s = (\boldsymbol{Y}'_{q_1^{(s)}}, \ldots, \boldsymbol{Y}'_{q_v^{(s)}})'$, $\mu_{q_1^{(s)}} = \cdots = \mu_{q_v^{(s)}} \equiv \theta_s$, $n_s = N_{q_1^{(s)}} + \cdots + N_{q_v^{(s)}}$ とおけば, 候補モデルは以下のモデル

$$X_{st} \sim N(\theta_s, \sigma^2), \quad t = 1, \ldots, n_s,$$

のパラメータに

$$\theta_1 \leq \theta_2, \ldots, \theta_1 \leq \theta_k,$$

を課したものと書き換えることもできる. ただし, 各 s に対し, $q_1^{(s)}, \ldots, q_v^{(s)}$ は Q_s の元であり, v は Q_s の要素数である.

これらの設定の下, 予測平均二乗誤差に基づくリスクに対する, 不偏な C_p 規準として TOC_p を導出した.

定理 1. 以下で与えられる TOC_p はリスクに対する一様最小分散不偏推定量である:

$$\mathrm{TO}C_p := (N - k^* - 2)\frac{\hat{\sigma}^2}{\bar{\sigma}^2} + 2(\hat{m} + 1).$$

ただし,

$$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}, \quad \bar{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{k^*}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2,$$

$$\bar{X}_s = \frac{1}{n_s}\sum_{t=1}^{n_s} X_{st}, \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{s=1}^{k}\sum_{t=1}^{n_s}(X_{st} - \hat{\theta}_s)^2, \quad \hat{m} = 1 + \sum_{a=2}^{k} 1_{\{\hat{\theta}_1 < \hat{\theta}_a\}}$$

であり, $\hat{\theta}_1, \ldots, \hat{\theta}_k$ は $\theta_1, \ldots, \theta_k$ の最尤推定量である.

参考文献

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, In 2nd International Symposium on Information Theory (eds. B. N. Petrov & F. Csáki), (1973), 267–281, Akadémiai Kiadó, Budapest.

[2] Y. Fujikoshi and K. Satoh, Modified AIC and C_p in multivariate linear regression, Biometrika, **84** (1997), 707–716.

[3] C. L. Mallows, Some comments on C_p, Technometrics, **15** (1973), 661–675.