多重共線条件下における複数の回帰モデル選択

川崎　能典　モデリング研究系　准教授

概要: スパース正則化法は近年活発に研究され，ゲノムデータ等の高次元データ解析に適用されているが，背後のメカニズムからかけ離れた結論が導かれ，再現性にも欠けるという現象も多々報告されている．高次元データになるほど多重共線性は避けられない問題である．ここでは多重共線条件下での回帰分析を念頭に，モデル選択のように一つのモデルに結論を絞るのではなく，結果の解釈可能性を担保することを目的に，複数の競合的モデルを手元に残す方法を提案する．[本研究は，統計数理研究所共同研究課題(25-共研-1018)に基づく，植木優夫氏(東北大学メディカル・メガバンク機構)との共同研究である．]

1. 標準化更新度に基づくモデル探索

n次元の応答変数ベクトルyと大きさ$n \times p$のデザイン行列$X = (x_1, \ldots, x_p)$を得たもとでの線形重回帰$y = X\beta + \epsilon$を考える．本研究では，飽和モデルとの隔たりからあてはまりの良さ(Goodness-of-fit; GOF)に関する基準を置き，その一方で各変数を取り込んだときのあてはまりの改善度として，標準化更新度と呼ぶ基準を提案する．回帰モデルの添字集合をCと書く．モデルCは以下を満たすときあてはまりが良いと定義．

$$\mathrm{GOF}_C = \frac{||y - X_C\hat{\beta}_C||^2 - ||y - X\hat{\beta}||^2}{||y - X\hat{\beta}||^2} \leq z_{1-\alpha}$$

ここで閾値zは，Cが真の回帰関数と一致するときに，以下を満たすように決める．(実際にはF分布の分位点となる．)

$$P(\forall C : \mathrm{GOF}_C \leq z_{1-\alpha}) > 1 - \alpha$$

ここでαは有意水準ではなくチューニングパラメータである．αを小さくするとGOF判定は緩くなる．シミュレーションの結果，$\alpha = 1/n$が適切と判断した．

一方，変数の取り込みの可否は，モデルCに変数kを加えたときの改善度を，標準化更新度(Standardized Update; SU)により判定する．

$$\mathrm{SU}_{k,C} = \frac{||y - X_C\hat{\beta}_C||^2 - ||y - X_{C\cup\{k\}}\hat{\beta}_{C\cup\{k\}}||^2}{||y - \bar{y}1||^2 - ||y - X\hat{\beta}||^2}, k \notin C$$

最小モデルからみた飽和モデルの改善度を1としたとき，変数kを取り入れることによるあてはまりの改善度がSUで，0から1の値をとる．閾値sに対しSU $> s$であれば変数kを採用するが，ここではシミュレーションの結果に基づき0.1を採用した．これにより，最小モデルからみた飽和モデルの改善度に照らして，変数kの貢献度の占める割合が10% 以上であることを要請している．

GOFとSUを使い，以下の手続きでモデルを探索する．

1. 各変数をひとつだけ含んだp個の一変量モデルからスタート(並列的探索)
2. GOF基準を満たしたモデルにはお墨付きを与えて終了
3. 満たさないモデルについて，それ以外の変数を各ステップでひとつずつ取り込み，GOF基準を満たすまで深掘り
4. 取り込みの可否はSUにより判断(取り込める変数がなければ良いモデルはなかったとして終了)

詳細は[1]を参照されたい．

2. シミュレーション

標本数nと説明変数の個数pは$(n, p) = (200, 50)$とし，(j,k)成分が$0.1^{|j-k|}$の行列Σによって，サイズ$n \times p$のデータ行列$Z \sim \mathrm{N}(0, \Sigma)$を発生させる．このとき以下のデータ生成機構をモデル1と呼ぶ．

$$\text{Model 1: } X_1 = Z_1 - Z_2, X_j = Z_j (j = 2, \ldots, p) \quad (1)$$

観測値yは以下のように生成する．

$$y = 2Z_1 + 2Z_2 + 2Z_{25} + \epsilon, \epsilon \sim \mathrm{N}(0, 2I) \quad (2)$$

ZではなくXが観察可能な説明変数になっていることに注意．拾い上げたい変数はモデル1では{1,2,25}．モデル2では，(1)の代わりに以下のようにXを生成する．

$$\text{Model 2: } X_1 = Z_1 + Z_2 + Z_{25}, X_2 = Z_1 - Z_2, X_{j+1} = Z_j (j = 2, \ldots, p)$$

応答変数の生成は(2)と同じ．モデル2は完全に多重共線条件下にあることに注意．拾い上げるべき変数は{1,2,3,26}である．

提案手法も含め，比較するのは以下の5つ．

1. 提案手法, $\alpha = 1/n$, SU $> s = 10\%$
2. 1と同じだが，SU基準をMallows' C_p ($\log n$)で置き換える
3. 飽和重回帰モデルの係数のt検定 ($P < 0.05$，多重性の補正なし)
4. 各変数について一変量回帰した係数のt検定 ($P < 0.05$でボンフェローニ補正)
5. Elastic Net．クロスバリデーションによりL_2パラメータを(0, 0.01, 0.1, 1, 10, 100)から選定．

200回の実験で，偽陽性FP，偽陰性FNと，右隣の括弧内にその平均数を記した．手法5のFPの高さが気になる．特に完全多重共線性下(Model 2)では，常に余計な変数を(平均的に20個以上)選びながら，重要な変数を(平均2個弱)殆ど常に取りこぼしている．

Model 1	1	2	3	4	5
FP	0.00 (0.00)	0.975 (13.305)	0.84 (2.205)	0.085 (0.085)	1.00 (20.605)
FN	0.01 (0.03)	0.005 (0.015)	0.00 (0.00)	1.00 (1.00)	0.00 (0.00)
Model 2	1	2	3	4	5
FP	0.00 (0.00)	1.00 (19.94)	0.86 (2.48)	0.03 (0.04)	1.00 (20.59)
FN	0.03 (0.12)	0.01 (0.01)	1.00 (2.84)	1.00 (1.00)	0.98 (1.89)

3. 応用：前立腺がんデータの解析

応答変数は97人のPSA値の対数(lpsa)．説明変数群は8つの変数からなる．(1)がんの大きさの対数値(lcavol)，(2)前立腺重量の対数値(lweight)，(3)患者の年齢(age)，(4)良性前立腺過形成BPHの対数値(lbph)，(5)精嚢侵襲の有無(svi, 0-1の2値変数)，(6)莢膜侵食の対数値(lcp)，(7)グリーソンスコア(gleason)，(8)グリーソンスコアが4ないし5のパーセンテージ(ppg45)．

提案手法を適用すると，6つのモデルが得られた．{lcavol, lweight}, {lcavol, svi}, {lcavol, lweight,svi}, {lcavol, lweight, lcp}, {lcavol, lweight, gleason}, {lcavol,lweight, pgg45}.

変数の相関行列(% 表示)と，再下段には重回帰分析でのP値を示す．

	lcavol	lweight	age	lbph	svi	lcp	gleason	ppg45
lcavol	—	—	—	—	—	—	—	—
lweight	28	—	—	—	—	—	—	—
age	22	35	—	—	—	—	—	—
lbph	3	44	35	—	—	—	—	—
svi	54	16	12	−9	—	—	—	—
lcp	68	16	13	−1	67	—	—	—
gleason	43	6	27	8	32	51	—	—
ppg45	43	11	28	8	46	63	75	—
lpsa	73	43	17	18	57	55	37	42
Multiple P	$< 10^{-8}$	0.0026	0.058	0.098	0.002	0.24	0.75	0.31

最大のVariance Inflation Factorはlcpの3.1で，深刻な多重共線があるとは結論されない．提案手法が挙げた6個の変数は，lpsaとの相関がある．MCP(Minimax Concave Panalty)またはSCADを用いるとgleason以外が生き残った．Elastic Netでは全ての変数が残ったが，上掲のシミュレーション結果(高FP)を踏まえれば自然である．

参考文献

[1] Ueki, M. and Kawasaki, Y. (2013) Multiple choice from competing regression models under multicollinearity based on standardized update, *Computational Statistics & Data Analysis*, Vol. 63, 31-41.