ロバスト・リッジ回帰推定量の研究

M2006MM031 武山 嵩弘
指導教員　木村 美善

1　はじめに

回帰モデルにおいて, 説明変数間に強い線形関係が存在し多重共線性の問題が示唆されるとき, しばしばリッジ回帰が用いられる. しかし, 最小 2 乗推定量を縮小することにより得られる従来のリッジ回帰推定量は外れ値に対して敏感であることが知られており, このような場合には, 最小 2 乗推定量の代わりにロバスト推定量を縮小することによって得られるロバスト・リッジ回帰推定量を用いた方が望ましいと考えられる.

先行研究では, Silvapulle[4] などにより, 目的変数に外れ値が存在する場合に, 縮小するロバスト推定量として M 推定量を用いたものなどが取り上げられてきたが, これらの推定量は誤差項の外れ値に対してのみ頑健であり, X 方向における外れ値については依然として十分に考慮されていないのが現状である. 本研究では, そのような場合も含めて様々なロバスト推定量を縮小することによって得られるロバスト・リッジ回帰推定量を提案し, 擬似乱数を用いたモンテカルロ・シミュレーションによる各推定量の比較・考察を行う.

2　線形回帰モデルにおける最小 2 乗推定量

2.1　線形回帰モデル

目的変数を Y, p 個の説明変数を $X_1, \cdots, X_p$, 回帰係数を $\beta_0, \beta_1, \cdots, \beta_p$ としたとき, 次のような線形回帰モデル

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \tag{1}$$

を考える. ここで, ε は近似によって生じる誤差を示す.

2.2　最小 2 乗法

最小 2 乗法（OLS：Ordinary Least Square）は, (1) 式のモデルにおける残差平方和

$$\|\varepsilon\|^2 = (Y - X\beta)'(Y - X\beta) \tag{2}$$

を最小にするような係数推定値 $\hat{\beta}$ を求める手法であり, 数ある回帰分析手法のうちでも最も基本的で, かつ最も広く用いられている. (2) 式は β の 2 次関数になっているので, これを β で偏微分したものを 0 とすることにより, OLS 推定量は次のように得られる.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3}$$

モデルにおける残差が不偏性, 等分散性, 無相関性の仮定を満たすとき, ガウス・マルコフの定理により, OLS 推定量はすべての線形不偏推定量のなかで最も小さい分散をもつ最良線形不偏推定量となるが, 実際の分析においてこれらの仮定をすべて満たすようなモデルを立てることが可能となるケースは稀である.

3　リッジ回帰

3.1　多重共線性の問題とその検出

重回帰分析において説明変数が互いに直交していないとき, OLS は係数推定値を求める手法として適当でない. 特に, 非直交性が極端でありデータの説明変数間に強い相関関係があるときには多重共線性の問題が生じ, 正規方程式の解が安定的に定まらなくなってしまう.

多重共線性が引き起こすこのようなリスクを避けるための方法のひとつに分散拡大要因（VIF：Variance Inflation Factor）を用いたものがあり, 第 i 番目の説明変数の係数に対する VIF は次のようにして求められる.

$$VIF(i) = (1 - R_i^2)^{-1} \tag{4}$$

ここで, (4) 式における R_i^2 は第 i 番目の説明変数を他の説明変数に回帰したときの重相関係数の 2 乗値を表す. 一般的な基準として, VIF が 10 を超えるような変数は他の変数と共線関係にあるとされている.

3.2　リッジ回帰推定量

$X'X$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ とすると, OLS 推定量の平均 2 乗誤差（MSE：Mean Squared Error）は

$$\begin{aligned} MSE[\hat{\beta}] &= E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] \\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}, \quad \sigma^2 = \mathrm{var}[\varepsilon] \end{aligned} \tag{5}$$

によって求められる. MSE は真のパラメータまわりの推定量のばらつきを表すものであるから可能な限り小さいことが期待されるが, データに多重共線性の問題があるとき, 固有値 λ には極めて 0 に近いものが存在するため, (5) 式は大きく発散しまう恐れがある.

この問題の解決策として, Hoerl and Kennard[1] は, モデルにリッジ・パラメータとよばれる定数 $k \geq 0$ を取り入れ, (3) 式の OLS 推定量 $\hat{\beta}$ を縮小することによって得られるリッジ回帰（ORR：Ordinary Ridge Regression）推定量を提案した.

$$\hat{\beta}_k = (X'X + kI)^{-1}X'Y \tag{6}$$

$k = 0$ の場合を除いて, $\hat{\beta}_k$ はバイアスを伴うため不偏推定量とはならないが, 説明変数間に多重共線性の問題があるとき, $\hat{\beta}$ よりも小さい MSE を与える k が存在する.

3.3　ORR 推定量における MSE の性質

ORR 推定量における MSE は次のように計算される.

$$\begin{aligned} MSE[\hat{\beta}_k] &= E[(\hat{\beta}_k - \beta)'(\hat{\beta}_k - \beta)] \\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i/(\lambda_i + k)^2 + k^2 \beta'(X'X + kI)^{-2}\beta \\ &= \gamma_1(k) + \gamma_2(k) \end{aligned} \tag{7}$$

ここで, 分割した第 2 項 $\gamma_2(k)$ はバイアスの平方を示し, k に関して単調増加する. また, 第 1 項 $\gamma_1(k)$ は分散の総和を示し, パラメータ k に関して単調減少する.

3.4 一般化リッジ回帰

固有値の対角行列 $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$, $X'X = P'\Lambda P$ となるような直交行列 P, $C = XP'$ を用いて線形回帰モデルを

$$Y = C\alpha + \varepsilon, \quad \alpha = P\beta \tag{8}$$

と書き換えることにより, (6) 式を一般化したものとして, 次式で表される一般化リッジ回帰（GRR：Generalized Ridge Regression）推定量が得られる.

$$\hat{\alpha}_k = (\Lambda + k^* I)^{-1} C'C\hat{\alpha} \tag{9}$$

この推定量を用いて, 最適なパラメータ k の値を数式により求めることが可能になる.

3.5 リッジ・パラメータ k の推定値

パラメータ k の推定値については, 現在もなおリッジ回帰に関する研究論文で様々なものが提案され続けている. 従来の推定値として最もよく用いられているものに,

$$\hat{k}_{HK} = \frac{p\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}} \tag{10}$$

$$\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \tag{11}$$

などが挙げられるが, Kibria[2] は k の新たな推定値として $\hat{\sigma}^2/\hat{\alpha}^2$ の調和平均や中央値を用いたものなどを提案しており, そのなかでもとりわけ, 幾何平均をとることによって得られる推定値

$$\hat{k}_{GM} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^{p} \hat{\alpha}_i^2)^{\frac{1}{p}}} \tag{12}$$

が, シミュレーション結果により従来の推定値に比べ比較的有用であると述べている.

しかし, これらの推定値はいずれも理論的に導かれたものではなく, 計算に手間が掛かるという点からもあまり実用的ではない. 実際の分析では, 横軸にパラメータ k, 縦軸に標準化した係数推定値をとったリッジ・トレースを描き, トレースが安定状態に達した所で視覚により k の値を決定する手法が一般に広く用いられている.

4 ロバスト・リッジ回帰

4.1 リッジ回帰へのロバスト推定量の適用

重回帰分析において ORR を用いることにより多重共線性が生じた際でも変数選択を行わずに安定した係数を推定することが可能となるが, データに外れ値がある場合にはその影響を強く受け, 分析が困難となってしまう.

この問題を解決するための方法として, 第 1 段階で加重最小 2 乗法による解を求め, 第 2 段階でリッジ回帰を適応する手法（Holland, 1973）や, 第 1 段階で先にリッジ・パラメータを決定した後, 第 2 段階で M 推定量を縮小するというアプローチ（Askin and Montgomery, 1980）などにより得られるロバスト・リッジ回帰（RRR：Robust Ridge Regression）が提案されている.

Silvapulle[4] は, 前者に近い手法で ORR における OLS 推定量 $\hat{\beta}$ を M 推定量 $\hat{\beta}^M$ で置き換えた次のような RRR 推定量を提案した.

$$\hat{\beta}_k^M = (X'X + kI)^{-1} X'X\hat{\beta}^M \tag{13}$$

この推定量を用いることにより, 多重共線性の問題に加え, 誤差項の一部が正規分布よりも裾の長い分布に従うような場合についても, それほど性能を損なわずに分析を行うことが可能となる.

4.2 リッジ型 M 推定量

Silvapulle[4] は, 3.4 節における GRR と同様にして, (13) 式を一般化した推定量

$$\hat{\alpha}_k^M = (\Lambda + k^* I)^{-1} C'C\hat{\alpha}^M, \quad \hat{\alpha}^M = P\hat{\beta}^M \tag{14}$$

をリッジ型 M（RM：Ridge-type M）推定量と名付けている. $\Omega = \mathrm{cov}[\hat{\alpha}^M]$ とすると, RM 推定量の MSE は,

$$MSE\left[\hat{\alpha}_k^M\right] = \sum_{i=1}^{p} \mathrm{var}\left[\hat{\alpha}_{ki}^M\right] + \sum_{i=1}^{p} \left\{\mathrm{bias}\left[\hat{\alpha}_{ki}^M\right]\right\}^2$$
$$= \sum_{i=1}^{p} \lambda_i^2 (\lambda_i + k^*)^{-2} \Omega_{ii} + \sum_{i=1}^{p} k^* \alpha_i / (\lambda_i + k^*)^2 \tag{15}$$

によって与えられ, すべての i について $\Omega_{ii} < \sigma^2 \lambda_i^{-1}$ であるとき, すべての正数 $k^* = k$ に対して,

$$MSE[\hat{\alpha}_k^M] < MSE[\hat{\alpha}_k], \quad k > 0 \tag{16}$$

が成り立つ.

4.3 本研究における RRR 推定量

誤差項の外れ値に加え, 説明変数項の外れ値に対しても対処できるよう, (13) 式を次のように一般化した RRR 推定量を提案する.

$$\hat{\beta}_k^{rob} = (X'X + kI)^{-1} X'X\hat{\beta}^{rob} \tag{17}$$

ここで, $\hat{\beta}_k^{rob}$ は種々のロバスト推定量を表す. この推定量を用いて, $\hat{\beta}_k^{rob}$ として LMS 推定量や Deepest などを縮小することにより, 誤差項の外れ値に加え, 説明変数項の外れ値にも対処することが可能となった.

ただし, ロバスト推定量にはそれぞれの長所と短所があるため, 分析対象のデータによって, それに適したものとそうでないものがある. ゆえに, 実際の分析においては, 外れ値の度合いやその方向などを調べることにより, まずデータの特徴を捉え, それに適したロバスト推定量を選ぶことが重要になってくる.

5 縮小するロバスト推定量

5.1 M 推定量

M 推定量は, Huber (1964) によって提案されたロバスト推定量の中でも最も一般的なものであり, 微分可能

な偶関数 ρ を用いて次のように求められる.

$$\hat{\beta}^M = \arg\min_{\beta} \sum_{i=1}^{n} \rho(r_i), \quad r_i = y_i - x_i'\beta \tag{18}$$

関数 ρ はこれまでに様々なものが提案されているが, Huber (1964) によるものが最も一般的である. (18) 式からもわかるように, $\rho(r_i) = r_i^2$ とすると, これは OLS 推定量に等しい.

5.2 LMS 推定量

LMS（Least Median of Squares）推定量は, Hampel (1975) によって提案され, それをさらに Rousseeuw (1984) が発展させたものであり, 残差平方の中央値を最小にすることによって求められる.

$$\hat{\beta}^{LMS} = \arg\min_{\beta} med(r_1^2, \cdots, r_n^2) \tag{19}$$

破綻点は $([n/2] - p + 2)/n$ であり, $n \to 0$ のとき $1/2$ となる. LMS 推定量は Y 方向のみでなく X 方向に対してもロバストであるが, 漸近効率は高くない.

5.3 LTS 推定量

LTS (Least Trimmed Squares) 推定量は, Rousseeuw (1984) によって提案された手法であり, 残差平方を昇順に並び替えた順序統計量の m 番目までの和を最小にすることによって次のように定義される.

$$\hat{\beta}^{LTS} = \arg\min_{\beta} \sum_{i=1}^{m} r_{(i)}^2 \tag{20}$$

(20) 式において, m=n としたときは残差平方すべての和を最小にすることになり, これは OLS 推定量に一致するが, 敢えて m を n 未満の適当な値に留め, 大きな残差については考慮しないようにすることによって, 外れ値が引き起こす汚染の影響を小さく抑えることが可能となる. 破綻点は $m = [n/2] + 1$ のとき最大となり, $1/2$ に達する.

5.4 GS（Generalized S）推定量

Rousseuw and Yohai (1984) によって提案された S 推定量を一般化したものとして, Croux et al. (1994) は GS（Generalized S）推定量を提案した. S 推定では, 尺度の M 推定量である s を残差に基づいて求めていたが, GS 推定量における s は残差の差に基づいて得られる. すなわち, β の GS 推定量 $\hat{\beta}_{GS}$ は

$$\binom{n}{2}^{-1} \sum_{i<j} \rho\left(\frac{r_i - r_j}{s}\right) = k_{n,p} \tag{21}$$

の解として s を定め,

$$\hat{\beta}^{GS} = \arg\min_{\beta} s(r_1, \cdots, r_n) \tag{22}$$

とすることにより求められる. この推定量が導入されたことによって, 漸近効率を犠牲にすることなく高い頑健性をもつことが可能となった. しかしその一方で, 残差の差を見ているため, データ数が大きくなると計算に時間が掛かってしまうという問題点も挙げられる.

5.5 最深回帰推定量

Rousseeuw and Hubert (1999) により, Regression Depth という新しい概念に基づいて提案された手法が最深回帰推定量（Deepest）である. 破綻点は

$$\frac{1}{p+1} \leq \epsilon_n^* \leq \frac{1}{3} \tag{23}$$

でありそれほど高くはないが, OLS に対する相対効率はデータ数によらず非常に高いことが知られている.

p 次元のデータ集合 $Z_n = \{z_i = (x_{i1}, \cdots, x_{i,p-1}, y_i); i = 1, \cdots, n\} \subset \mathbb{R}^p$ に対し, $y = \theta_1 x_1 + \cdots + \theta_{p-1} x_{p-1} + \theta_p$ を当てはめ, $\theta = (\theta_1, \cdots, \theta_p)'$ に対する z_i の残差を $r_i = r_i(\theta)$ とする. このとき, Z_n に対する θ の $rdepth(\theta, Z_n)$ は次のように定義される.

$$rdepth(\theta, Z_n) = \min_{1 \leq i \leq n} \{\min\{S^+(x_i) + G^-(x_i), G^+(x_i) + S^-(x_i)\}\} \tag{24}$$

ただし,

$$S^+(v) = \#\{i; x_i \leq v \text{ and } r_i \geq 0\}$$
$$G^-(v) = \#\{i; x_i > v \text{ and } r_i \leq 0\}$$
$$G^+(v) = \#\{i; x_i \geq v \text{ and } r_i \geq 0\}$$
$$S^-(v) = \#\{i; x_i < v \text{ and } r_i \leq 0\}$$

これにより, 最深回帰推定量 $DR(Z_n)$ は $rdepth(\theta, Z_n)$ を最大にする θ として定義される.

$$DR(Z_n) = \arg\max_{\theta} rdepth(\theta, Z_n) \tag{25}$$

6 シミュレーション

6.1 概要

ORR, Silvapulle[4] が提案した RM 推定量に加え, その他の様々なロバスト推定量を縮小した RRR 推定量を用いて, データにおける共線性や外れ値がそれぞれの推定の精度にどの程度影響を及ぼすか検証する. なお, 計算にはオープン・ソースの統計解析環境『R』を使用し, Deepest に基づく RRR 推定量の計算には, 大見他 [3] により『R』で使用可能となった Medsweep プログラム（Aelst et al., 2002）を用いた.

作成するデータが p 個の説明変数をもち, n 個の観測値からなるものであるとし, $l = [p/2]$ としたとき, 標準正規分布に従う擬似乱数 $z_1, \cdots, z_{p+1}$, 定数 γ を用いて, まず説明変数 $x_1, \cdots, x_p$ を次のように作成する.

$$z_{i1}, \cdots, z_{i,p+1} \sim N(0,1), \quad i = 1, \cdots, n$$
$$\begin{cases} x_{ij} = z_{ij}, & j = 1, \cdots, l \\ x_{ij} = (1-\gamma^2)^{\frac{1}{2}} z_{ij} + \gamma z_{i,p+1}, & j = l+1, \cdots, p \end{cases}$$

すなわち, $x_1, \cdots, x_l$ を独立した変数とし, $x_{l+1}, \cdots, x_p$ には互いに相関関係を持たせてある. 目的変数は

$$y_i = x_{i1} + \cdots + x_{ip} + 1 + \varepsilon_i, \quad \varepsilon \sim N(0, 0.1)$$

とし, 切片項も含めて真の係数が全て 1 となるように設定した. このデータをベースとして, ある定数 ϵ の割合

でコーシー分布を混合することにより, さらに外れ値を加えていく. 誤差項に外れ値を与える場合には

$$y_i = x_{i1} + \cdots + x_{ip} + 1 + E_i,$$
$$E \sim (1-\epsilon)\cdot N(0, 0.1) + \epsilon \cdot t(1)$$

とし, 説明変数項に外れ値を与える場合には, 目的変数のデータを先に作成した後, $1 \leq j \leq l$ を満たす奇数 j に対する x_{ij} のみを次のように置き換える.

$$x_{ij} = E_i, \quad E \sim (1-\epsilon)\cdot N(0,1) + \epsilon \cdot t(1)$$

データ数 n, 説明変数の次元 p, 混合分布の割合 ϵ をそれぞれ変化させ, パラメータ k については敢えて 3.5 節などの推定値によって特定せず, $k = \mathrm{seq}\,(0, n, n/100)$ として一定の範囲内を動かすことにした（$\mathrm{seq}\,(a, b, c)$ は初項 a, 末項 b, 公差 c の等差数列を表す）.

j 回目の推定におけるパラメータ k に対する各 RRR 推定量を $\hat{\theta}(k)^j$ とすると, その MSE は

$$MSE[\hat{\theta}(k)^j] = E[(\hat{\theta}(k)^j - \theta)^2]$$
$$= \frac{1}{p+1}\sum_{i=1}^{p+1}(\hat{\theta}(k)_i^j - 1)^2$$

とすることにより計算されるが, 今回のシミュレーションでは, 推定量の良さを測る基準として 10000 回の推定における $\min MSE[\hat{\theta}(k)^j]$ の平均値を用いている.

6.2 結果

表 1 に, $n = 50, 500$, $p = 3$, $\gamma = 0.99$ のもとで X 方向に外れ値を与えたケースにおける各推定量の MSE を示す. また, 図 1 は $p = 3, \epsilon = 0$ のもとでの n の変化に対する MSE の推移である.

ORR 推定量は, 裾の長い分布をわずかに 1 割混合しただけでも非常に大きな影響を受けてしまい, 外れ値の存在する方向に関わらず簡単に破綻してしまうことがわかる. また, RM 推定量は, 外れ値の存在が誤差項のみであり分布混合の割合がそれほど高くなければ非常に有用な推定量となるが, やはり X 方向の汚染に対しては頑健でない. LMS, LTS 推定量に基づく RRR 推定量は, 誤

表 1　説明変数項に分布を混合, $n = 50, 500$

ϵ	ORR	RM	RRR			
			LMS	LTS	GS	Deepest
$n = 50$						
0	$9.10E-4$	$9.38E-4$	$2.81E-3$	$2.77E-3$	$1.75E-3$	$1.74E-3$
0.1	$1.03E-1$	$4.56E-2$	$2.51E-3$	$2.50E-3$	$1.39E-3$	$2.10E-3$
0.2	$1.72E-1$	$1.27E-1$	$2.40E-3$	$2.35E-3$	$1.17E-3$	$3.36E-3$
0.3	$2.17E-1$	$2.06E-1$	$2.32E-3$	$2.16E-3$	$1.20E-3$	$1.20E-2$
0.4	$2.50E-1$	$2.60E-1$	$3.90E-3$	$3.74E-3$	$3.11E-2$	$7.30E-2$
$n = 500$						
0	$1.60E-4$	$1.68E-4$	$7.84E-4$	$8.55E-4$	$3.69E-4$	$3.12E-4$
0.1	$1.85E-1$	$5.88E-2$	$6.80E-4$	$7.08E-4$	$2.42E-4$	$3.95E-4$
0.2	$2.35E-1$	$1.93E-1$	$6.70E-4$	$6.79E-4$	$2.12E-4$	$7.19E-4$
0.3	$2.56E-1$	$2.56E-1$	$6.78E-4$	$6.55E-4$	$2.08E-4$	$1.56E-3$
0.4	$2.72E-1$	$2.83E-1$	$7.59E-4$	$6.87E-4$	$1.05E-3$	$5.03E-3$

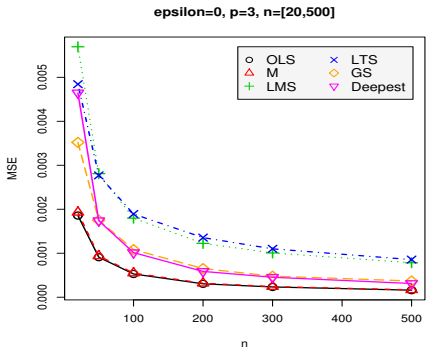

図 1　各 RRR 推定量の漸近的有効性の推移 ($p = 3$)

差項または説明変数項いずれの外れ値に対してもデータ数や次元によらず高い頑健性を示した. しかし, 図 1 からもわかるように, 漸近効率が悪いという欠点も示唆される. GS 推定量に基づく RRR 推定量は, データ数や次元, 外れ値の方向などによらず常に高い頑健性を維持しており, 小標本の分析における漸近的有効性は RM 推定量に次いで高い. これに対し, Deepest に基づく RRR 推定量は, LMS, LTS, GS 推定量のような高い頑健性こそないものの, 低次元で且つデータ数が小さすぎないとき, $\epsilon = 0$ での MSE の値は極めて小さく, 分布混合による汚染がごくわずかであれば, GS 推定量に基づく RRR に優る手法となりうる.

7　おわりに

本研究でとりあげたロバスト・リッジ回帰はリッジ回帰とロバスト回帰の理論を融合したものであり, この組み合わせによりそれぞれの手法がもつ利点を生かし, 有効な分析を行うために回帰モデルが必要とする前提条件を緩和させることが可能となった. 特に GS 推定量に基づく RRR は大標本の分析には計算にかなりの時間が掛かってしまうという問題点があるものの, 様々な側面から総合的に評価すると今回シミュレーションを行ったなかで最もバランスの良い推定量であると言えるだろう.

参考文献

[1] Hoerl,A.E. and Kennard,R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* **12**, 55-67.

[2] Kibria,B.M.G. (2003). Performance of Some New Ridge Regression Estimators, *Communications in Statistics-Simulation and Computation* **32**,419-435.

[3] 大見俊司・安藤雅和・木村美善 (2007). 最深回帰推定量とその R による実用化, 南山大学紀要「アカデミア」数理情報編 **7**, 61-73.

[4] Silvapulle,M.J. (1991). Robust Ridge Regression Based on an M-Estimator, *Australian Journal of Statistics* **33**, 319-333.