ロバスト・リッジ回帰の研究

M2011MM002 阿部智成

指導教員：木村美善

1　はじめに

線形回帰モデルにおいて，通常よく用いられる最小 2 乗推定量は標準的仮定の下では， 望ましい推定量である．しかし，最小 2 乗推定量は多重共線性や外れ値が存在する場合には不安定になり，その良さを失ってしまうことが知られている．多重共線性の問題に対して，リッジ回帰は実際に幅広い分野の研究で用いられている手法の一つである．しかし，このリッジ回帰は，外れ値に有効に対処できるようになっておらず，その影響を受けやすいという欠点がある．外れ値が存在するとき，外れ値に対する影響を受けにくいロバスト回帰を用いることが望ましい．また， 実際の分析に用いられるデータには外れ値と多重共線性が混在している場合が多くある．そこで，リッジ回帰とロバスト回帰を組み合わせたロバスト・リッジ回帰という手法を用いることにより， この 2 つの問題に対して同時に対処することが可能となる．本研究の目的は，とロバスト・リッジ回帰の理論と推定量が持つ性質を整理し，実際のデータなどのデータを分析した際の結果を通して， それらの有効性と問題点について検証していくことである．

2　線形回帰モデル

2.1　最小 2 乗推定量

目的変数 y_i の $n \times 1$ ベクトルを y，定数項と説明変数 $x_{i1}, x_{i2}, \cdots, x_{ip}$ の $n \times (p+1)$ 行列を X，回帰係数 $\beta_0, \beta_1, \cdots, \beta_p$ の $(p+1) \times 1$ ベクトルを $\boldsymbol{\beta}$，誤差項 ε_i の $n \times 1$ ベクトルを $\boldsymbol{\varepsilon}$ とし，線形回帰モデル

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

を考える. 残差平方和 (RSS) は

$$RSS[\boldsymbol{\beta}] = \| \boldsymbol{\varepsilon} \|^2 = (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})$$

により定義される.RSS は

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2}$$

のとき最小になる. これが式 (1) における $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ の最小 2 乗 (LS) 推定量である．LS 推定量は，$\boldsymbol{\varepsilon}$ が $E[\boldsymbol{\varepsilon}] = 0$, $V[\boldsymbol{\varepsilon}] = \sigma^2 I$ (I は単位行列) を満たすとき最良線形不偏推定量であり，さらに正規分布に従うときには最良不偏推定量となる．しかし，実際のデータはこのような標準的仮定からずれていたり，外れ値や多重共線性が存在したりする. このような場合 LS 推定量は不安定になる.([4] 参照)

3　リッジ回帰推定量

3.1　リッジ回帰とは

この手法は, リッジ・パラメータとよばれる $k > 0$ を取り入れることによって回帰係数の安定化を図るものである. リッジ回帰推定量は偏りをもつ推定量であるが，適切な k を選ぶことによって最小 2 乗推定量よりも小さい平均 2 乗誤差を持つようにすることができる．$X'X$ の固有値を $\lambda_1 \geq \cdots, \lambda_{p+1}$ とする. このとき，回帰係数 β の LS 推定量 $\hat{\beta}$ の平均 2 乗誤差 (MSE) は

$$MSE[\hat{\beta}] = E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] = \sigma^2 \sum_{i=1}^{p+1} \lambda_i^{-1} \tag{3}$$

と表される. ただし，X, Y は標準化されているものとする. ここで，データに多重共線性の問題があるとき，固有値 λ にはきわめて 0 に近いものが存在するため，式 (3) で与えられる LS 推定量の MSE は大きく発散する可能性がある. そこで，リッジパラメータと呼ばれる定数 $k > 0$ を導入し，LS 推定量 $\hat{\beta}$ を縮小することによって回帰係数の安定化をはかる. この推定量は

$$\hat{\beta}_k = (X'X + kI)^{-1}X'y \tag{4}$$

である．$\hat{\beta}_k$ は $k > 0$ のときバイアスを伴うため不偏推定量とならないが，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 ([1] 参照). 当然 $k = 0$ のとき LS 推定量と一致する．

4　ロバスト回帰推定量

4.1　ロバスト回帰とは

ロバスト回帰法とは真の分布が指定した分布とずれがあっても効率がそれほど減少しない回帰分析法である. 単回帰の場合には実際に 2 次元平面上で外れ値をみることができるので，それを取り除いてもう一度解析しなおせばよいが，重回帰の場合それが困難である. このことから外れ値の検出は LS 法ではうまくいかないことが多い. そして，この問題を克服するためにロバスト回帰という方法が考案された.

4.2　LMS 推定量と LTS 推定量

LMS 推定量 $\hat{\beta}_{LMS}$ は, $Rousseeuw$ (1984) により，高い破綻点を得るように $Hampel$ (1975) のアイディアに基づき提案されたもので

$$med_i r_i^2(\hat{\beta}_{LMS}) = \min_{\beta} med_i r_i^2(\beta) \tag{5}$$

により定義される. ここで $med_i r_i^2$ は残差の 2 乗である r_i^2 の中央値である. この推定量は，y の外れ値と同様に x の

外れ値に関しても強く，破綻点は 50 %である.LTS 推定量 $\hat{\beta}_{LTS}$ は，$Rousseeuw$(1985) により

$$\sum_{i=1}^{h}(r^2(\hat{\beta}_{LTS}))_{i:n} = \min_{\beta}\sum_{i=1}^{h}(r^2(\beta))_{i:n} \tag{6}$$

と定義される．$(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$ は残差の 2 乗を小さいほうから並び替えたものであり，小さいほうから h 番目までの和を最小とする β である．LTS は LS と似ているが，大きい残差が和に含まれないことで外れ値を避けることができるのでその影響を受けにくくなる. この推定量の破綻点は h が $[n/2]+1$ のときに，50 %に達する.([3] 参照)

4.3 M 推定量

線形回帰モデルにおける M 推定量は，ψ を実軸上の実数値関数 ρ の導関数としたとき

$$\sum_{i=1}^{n}\psi(\frac{y_i - x_i'\beta}{\sigma})x_i = 0 \tag{7}$$

を解くことによって与えられる. ここで ρ は微分可能かつ 0 の周りで対称な凸関数とする.

- $Huber$ の ψ

$$c_{k,l} = \begin{cases} -c, & u < -c \\ u, & |u| \leq c \\ c, & u > c \end{cases} \tag{8}$$

- $Tukey$ の ψ

$$c_{k,l} = \begin{cases} u[1-(\frac{u}{c})^2]^2, & |u| \leq c \\ 0, & u > c \end{cases} \tag{9}$$

この他に $Andrews$ の ψ 関数や $Cauchy$ の ψ などがある. この M 推定量は y 方向に関してはロバストであるが，x に関してはロバストではなく，次元が増えると破綻点が低くなる.

4.4 S 推定量

M 推定量の柔軟性と漸近的性質の良さを保持しながらも高い破綻点を持ち，LMS や LTS よりも高い漸近効率をもつことを狙ったのが $Rousseeuw\ and\ Yohai$(1984) による S 推定量である．S 推定量は

$$s(\hat{\beta}_s) = \min_{\beta} s(\beta) \tag{10}$$

を満たすものとして定義される. ここで $s(\beta)$ は

$$\frac{1}{n}\sum_{i=1}^{n}\rho(r_i(\beta)/s(\beta)) = b \tag{11}$$

を満たすものである．ρ は $(-\infty,\infty)$ 上の有界な関数であり，原点対称,$(0,\infty)$ 上で非減少，$\rho(0)=0$，b はある定数. すなわち尺度 $s(\beta)$ を推定した後，この $s(\beta)$ を最小にする β を推定量とするものである.

4.5 τ 推定量

τ 推定量は以下のように定義される.

$$\hat{\beta}_n = arg\min_{\beta\in R^p}\tau_n(\beta) \tag{12}$$

この $\tau_n(\beta)$ は以下で与えられる

$$\tau_n^2(\beta) = s_n^2(\beta)\frac{1}{nb_2}\sum_{i=1}^{n}\rho_2(\frac{y_i-\beta' x_i}{s_n(\beta)}) \tag{13}$$

ここで，$s_n(\beta)$ は以下の式を解くことにより得られる.

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1(\frac{y_i-\beta' x_i}{s_n(\beta)}) = b_1 \tag{14}$$

ρ_1,ρ_2 は M 推定量と同様に，微分可能かつ，0 の周りで対称な凸関数とする．ρ_2 が

$$2\rho_2(u) - \psi_2(u)u \geq 0 \tag{15}$$

を満たすとすれば，$W_n \geq 0$ であり，β の τ 推定量は

$$\psi(u) = W_n\psi_1(u) + \psi_2(u) \tag{16}$$

を ψ 関数として持つ M 推定量と考えることができる. ここで,

$$W_n = \frac{\sum_{i=1}^{n}[2\rho_2(r_i) - \psi_2(r_i)r_i]}{\sum_{i=1}^{n}\psi_1(r_i)r_i}. \tag{17}$$

よって ρ_1 に高い破綻点を与える関数，ρ_2 に高い効率を与える関数を選ぶことで，高い破綻点と高い効率を同時に持つ推定量を得ることができる.

5 ロバストリッジ回帰推定量

実際のデータには多重共線性に加えてさらに外れ値が存在するような場合がある. この問題を解決するために考えられた手法として $silvapulle$ (1991) によって提案されたロバスト・リッジ回帰法が挙げられる. これは従来のリッジ回帰で縮小されている LS 推定量の代わりに M 推定量を縮小させたものであり，多重共線性と外れ値の問題どちらにも対応できる.

5.1 M 推定量に基づくロバストリッジ回帰

LS 推定量を縮小することによって得られるリッジ回帰推定量

$$\hat{\beta}_k = (X'X + kI)^{-1}X'Y \tag{18}$$

は, LS 推定量 $\hat{\beta}$ に左から行列 $(X'X+kI)^{-1}X'X$ を掛けた

$$\hat{\beta}_k = (X'X + kI)^{-1}X'\hat{\beta} \tag{19}$$

と表現でき, $\hat{\beta}$ を 0 ベクトルの方向に縮小させるものである．$Silvapulle$ (1991) は，LS 推定量 $\hat{\beta}$ を β の M 推定量 $\hat{\beta}_M$ で置き換えたロバスト・リッジ回帰を提案した. その推定量は次のように定義される.

$$\hat{\beta}_k^{rob} = (X'X + kI)^{-1}X'X\hat{\beta}^M \tag{20}$$

こうして, リッジ回帰と M 推定量のロバスト回帰を組み合わせることにより，多重共線性の問題に加え，外れ値が存在する場合にもほとんど性能を損なわずに分析を行うことが可能になる. 式 (20) の $\hat{\beta}_M$ を他のロバスト推定量で置き換えることにより，M 推定量以外のロバスト推定量とリッジ回帰を組み合わせることができる.

6 実行例 (リッジ回帰推定量)

longley データを用いて実行例を示す. 説明変数は x_1 : $GNP.deflator$, x_2 : GNP, x_3 : $Unemployed$, x_4 : $Armed.Forces$, x_5 : $Population$ であり，応答変数は y : $Employed$ である. ここでは，回帰モデルとして

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \qquad (21)$$

を考える. このモデルの妥当性を確認するため，標準化残差をプロットしたもの見ると，残差は 0 の周りでほぼランダムに分布していることがわかるため，モデルの設定には問題がなさそうである.

6.1 結果と考察

まず LS 推定量 $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_5)$ を求める. 回帰分析の結果を表 1 に示す.

表 1 LS 回帰分析の結果:longley データ

変数	$\hat{\beta}$	標準偏差	t 値
x_1	-0.149	0.132	-0.366
x_2	2.038	0.032	2.269
x_3	-0.107	0.004	-0.921
x_4	-0.111	0.003	-1.975
x_5	-0.799	0.330	-1.222
	$n = 16$	$R^2 = 0.987$	

決定係数だけを見れば, 0.987 と高いので一見良い推定ができたように思えるが，x_1, x_5 に対する係数推定値がマイナスの値となった. これは $Employed$ が $GNP.deflator$, $Population$ に反比例していることを意味するが，通常は比例すると考えられる. そこで, 相関係数に着目すると，やはり x_1, x_5 には強い相関関係があることから，多重共線性の存在が示唆される. したがって，x_2, x_3, x_4 の係数についても多少の修正がなされるべき可能性が高い. 次に，このデータにリッジ回帰を適用する. 図 1 に, リッジ・パラメータ k の値を 0 から 0.1 まで増加させていったときのリッジ・トレースを示した.

図 1 のリッジ・トレースから，強い相関関係にあった x_1, x_2, x_5 の係数に対する推定値がパラメータ k の値の微小な変化に大きな影響を受けている様子が窺える. LS 推定量でマイナスの値を示していた x_1, x_5 は, $k = 0.08$ の前後でほぼ安定状態に入り，その符号もプラスへと転じている. 以上のことから，リッジ回帰推定量は分析前の予想にも合致しており，良い推定が行えているといえるだろう.

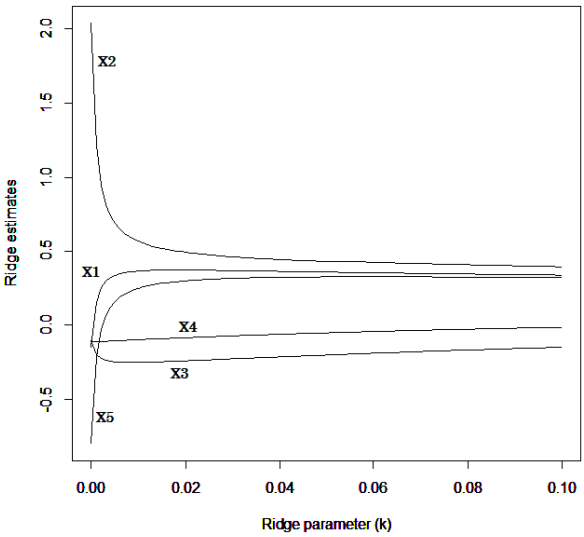

図 1 標準化残差:*longley* データ

7 実行例 (ロバスト回帰推定量)

ここでは，LS, M, LMS, LTS, S 推定量の比較を行う. 扱うデータは自作データとする. データの作成法は，3 つの説明変数を持ち，それぞれ独立した変数とする. 応答変数は真の係数 $\beta = (\beta_1, \beta_2, \beta_3)'$ が全て 1 となるように設定した. y 方向に外れ値を与える場合は，正規分布に従う誤差項の平均・分散を大きくした.x 方向の外れ値は, 説明変数それぞれ p : $(p = 1, 2, 3, 4, 5)$ 割の観測値のデータを $N(5, 5)$ と置き換えた.

7.1 結果と考察

1. y 方向のみの外れ値が存在するとき, LS 推定量による係数推定値は p の値にかかわらず，全ての係数推定値が真の係数とは大きくずれてしまった. また，$p = 3$ までの場合において M 推定量による係数推定値が真の係数に最も近かった. $p = 4$ 以上になると，M 推定量による係数推定値も不安定になる結果となった. よって，$p = 3$ までは LS 推定量以外は良い推定ができ，y 方向のみの外れ値が 30 %以下の場合は M 推定量による推定が優れていると推察できる.
2. x 方向のみの外れ値が存在するとき, y 方向のみの外れ値が存在するときと同様に LS 推定量による係数推定値は p の値にかかわらず，全ての係数推定値も真の係数とは大きくずれてしまった. また，M 推定量は $p = 2$ 以上のときに LS 推定量に近い推定値となり，x 方向の外れ値にロバストでないという欠点が大きく出る結果となった.
3. x, y 方向両方に外れ値が存在するとき, LS, M 推定量は p の値にかかわらず，全ての係数推定値が大きくずれてしまった.LMS, LTS, S 推定量は推定値に大きな差はなく，LMS 推定量が最も真の係数に近い値となった.

以上より，それぞれの推定量の特性が出る結果となった．本例のようにデータ数が少ない場合には LTS, LMS, S 推定量はどれも同様に有効であると推察される．

8　多重共線性の存在するデータ作成法

多重共線性の存在するデータの作成手順は金・田中 (1993) に従った．

8.1　手順

1. 変数の数 (p) と標本の大きさ (n) を固定する．
2. 直交行列 $V_{p\times p}$ を作る:
   (a) 線形独立な p 次元ベクトル $\{e_i\}_1^p$ を生成する．
   (b) $\{e_i\}_1^p$ をグラム・シュミッドの直交化法を用いて，ベクトルのノルムが 1 であるような正規直交ベクトル $\{v_i\}_1^p$ に変換し，それを直交行列 V にする．
3. 対角行列 $D_{p\times p}$ を作る:
   (a) condition index $\kappa_1, \kappa_2, \cdots, \kappa_p$ と分散の和 $c(=\sum_{j=1}^p \lambda_j)$ を指定する．指定された condition index と分散の和 c に基づき，固有値 $\lambda_i = c/(\kappa_i \sum_{j=1}^p \kappa_j^{-1})$ を計算する．
   (b) 求めた各 $\lambda_i^{1/2}$ を対角要素にする対角行列 $D_{p\times p}$ を作る．
4. 行列 $U_{n\times p}$ を作る:
   (a) $N(0, I)$ に従う p 変量正規乱数 $\{y_i\}_1^n$ を発生する．
   (b) $\{y_i\}_1^n$ の平均ベクトル $\bar{y}$ と分散行列 S を計算する．
   (c) S のスペクトル分解 $S = QGQ'$ を行う．
   (d) 各 y_i を次のように変換する．

$$z_i = G^{-\frac{1}{2}} Q'(y_i - \bar{y}), \quad i = 1, 2, \cdots, n \quad (22)$$

   (e) 各 z_i' を行とする $U_{n\times p}$(以下この U を U_E とする) を作る．
5. データ $X_{n\times p}$ を作る:
   上で求めた行列 V, D, U を用いて，それら 3 つの行列の積 $UDV' = X$ を計算して人工データを作る．

9　実行例 (ロバスト・リッジ回帰推定量)

ここで扱う推定量は，それぞれリッジ回帰推定量,M,LMS,LTS,S,τ 推定量に基づくロバスト・リッジ回帰推定量とする．ここでは，上で挙げたデータ作成法を用いてデータを作成し，このデータを用いてロバスト・リッジ回帰推定量の有効性を示していく．

9.1　実行例の仮定

- 標本数を 50 とし，5 つの説明変数 $(x_1, x_2, x_3, x_4, x_5)$ を用いた $(x_1, x_2, x_3, x_4, x_5 \sim N(0, 1))$.
- 目的変数は以下のように設定する．

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon.$$

y 方向に外れ値を入れる場合はこの ε が以下のような混合分布に従う．

$$\varepsilon \sim (1-\eta) \cdot N(0, 1) + \eta \cdot N(0, 9)$$

- x 方向の外れ値は (x_1, x_2, x_3, x_4) の相関を維持したまま $x_5 \sim (1-\eta)N(0, 1) + \eta N(0, 9)$ とし，混入させる．
- $\eta = 0.2$ とする．

9.2　結果と考察

以下に S 推定量と τ 推定量に基づくロバスト・リッジ回帰推定量を適用したリッジ・トレースを示す．これを見て

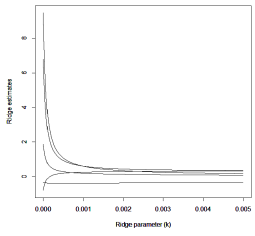

図 2　S・リッジ

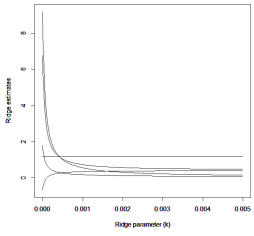

図 3　τ・リッジ

わかるように，k の微小な値の変化に対してリッジ・トレースが大きく変化していることがわかる．また，$k = 0.02$ のところで安定し，係数推定値も真の値に近づいていることがわかる．これは，S 推定量と τ 推定量に基づくロバスト・リッジ回帰推定量が多重共線性と外れ値が混在するデータに対して有効であることを示している．

10　おわりに

本研究では, リッジ回帰, ロバスト回帰，ロバスト・リッジ回帰の実行例を用いてその有効性を示した．しかし，このことは本研究で扱ったデータに対しては有効であるということであり，一般的に有効性を示せたわけではない．また，複雑な多重共線性が存在するデータを作成することに成功し，τ 推定量の有効性を示した．しかし，τ 推定量は複雑であり一般的に τ 推定量の有効性を示すには更に研究が必要である．

参考文献

[1] Hoerl,A.E. and Kennard, R.W(1974). Ridge Regression: Applications to Nonortogoual, Technometrics, 12, 69-82.

[2] 金 鉉彬・田中 豊 (1993).”多重共線性を持つ人工データの作成法の一提案”, 日本計算機統計学.

[3] Rousseeuw,P.J.and Leroy,A. M. (1986) Robust Regression and Outlier Detection, John Wiley and Sons.

[4] Sibapulle,M.J(1991). Robust ridge regression based on M-estimator. Austral.J.Statist, 33, 319-333.

[5] 武山 嵩弘 (2008). ロバスト・リッジ回帰推定量の研究, 南山大学数理情報研究科修士論文.