τ 推定量に基づくロバスト・リッジ回帰の研究

M2012MM045 塚原一翔

指導教員：木村美善

1 はじめに

回帰モデルにおいて, 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, 外れ値や多重共線性が存在する場合にはその値に対して敏感であり, 推定量の正確性を失ってしまうことが知られている．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, 多重共線性の問題に対しては, リッジ回帰が広く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, 外れ値に対して有効に対処できず, 外れ値の影響を受けやすいという欠点がある．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．このような場合には, ロバスト回帰とリッジ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で, 外れ値と多重共線性という 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，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, 様々なロバスト推定量に基づくロバスト・リッジ回帰推定量, 特に τ 推定量に基づくロバスト・リッジ回帰推定量に着目し, その有効性について調べることである.

2 回帰分析

2.1 線形回帰モデル

目的変数を y, p 個の説明変数を $x_1, x_2, \cdots, x_p$, 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

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \cdots, n \quad (1)$$

を考える. ここで, ε は近似によって生じる誤差を表す．このモデルを行列で表記すると

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

となる．ここで

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

である.

2.2 最小 2 乗法

最小 2 乗法は, (1) 式のモデルにおける残差平方和 (RSS)

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

を最小にするような $\boldsymbol{\beta}$ を求める手法であり, 回帰分析手法の中で最も基本的かつ最も広く用いられているものである．(2) 式は $\boldsymbol{\beta}$ の 2 次関数になっており, これを $\boldsymbol{\beta}$ で偏微分したものを 0 とすることにより, 最小 2 乗推定量は

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

となる.

モデルにおける誤差が等分散性, 不偏性, 無相関性の仮定を満たすとき, 最小 2 乗法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り, さらに正規分布が仮定されている場合には最良不偏推定量となる．しかし, 実際のデータは, 外れ値や多重共線性が存在したりするなど, これらの仮定をすべて満たすような場合は稀である.

3 リッジ回帰推定量

3.1 多重共線性とは

重回帰分析において説明変数の間に強い相関関係が存在する場合, これらの説明変数の間には多重共線性があるという．この場合, 回帰分析により得られる結果に悪い影響が出ることがある．具体的には, 同時に用いる説明変数の増減により回帰式の係数が大きく変化したり, 決定係数が高い一方で t 値が低く, 有効な推定結果が得られなかったり, 通常考えられる符号と異なる結果が得られる, などの症状が生じる.

3.2 リッジ回帰推定量

$X'X$ の固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p+1}$, 回帰係数 $\boldsymbol{\beta}$ の最小 2 乗推定量 (LS 推定量)$\hat{\boldsymbol{\beta}}^{LS}$ は不偏推定量であるので, その平均 2 乗誤差は

$$\begin{aligned} MSE[\hat{\boldsymbol{\beta}}] &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\ &= \sigma^2 \sum_{i=1}^{p+1} \lambda_i^{-1} \end{aligned} \quad (4)$$

となる．データに多重共線性が存在する場合に, $X'X$ の固有値 λ には 0 に極めて近いものが存在するため, (4) の LS 推定量の平均 2 乗誤差は大きく発散してしまう可能性がある．Hoerl and Kennard(1970) はこのようなときでも平均 2 乗誤差を小さく抑えるための手法としてリッジ回帰推定量を提案した．リッジ回帰推定量はモデルにリッジ・パラメータとよばれる定数 $k \geq 0$ を取り入れることで LS 推定量 $\hat{\boldsymbol{\beta}}^{LS}$ を縮小させたものであり

$$\hat{\boldsymbol{\beta}}^{LS}_{(k)} = (X'X + kI)^{-1}X'Y \quad (5)$$

により定義される．特に $k=0$ のとき，$\hat{\boldsymbol{\beta}}_{(k)}^{LS}$ は最小 2 乗推定量に等しくなる. しかし, リッジ回帰推定量は不偏推定量ではない．したがって k の増加に伴い, 偏りも大きくなってしまう.

4　ロバスト回帰推定量

4.1　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である．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

$$\hat{\boldsymbol{\beta}}^{M}=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{n}\rho(r_i(\boldsymbol{\beta})),$$

$$r_i(\boldsymbol{\beta})=y_i-(\beta_0+\beta_i x_{i1}+\cdots+\beta_p x_{ip}) \tag{6}$$

と定義される．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．また, (6) 式からもわかるように, $\rho(r_i)=r_i^2$ とすると, これは最小 2 乗推定量に等しいことがわかる.

4.2　LMS 推定量

LMS 推定量は, Rousseeuw(1984) により提案されたロバスト回帰推定量であり,

$$\hat{\beta}^{LMS}=\arg\min_{\beta}\ med\{r_1(\boldsymbol{\beta})^2,\cdots,r_n(\boldsymbol{\beta})^2\} \tag{7}$$

として定義される.

4.3　LTS 推定量

LTS 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

$$\hat{\boldsymbol{\beta}}^{LTS}=arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{m}r_{(i)}^2(\boldsymbol{\beta}) \tag{8}$$

として定義される．ここで $r_{(1)}^2(\boldsymbol{\beta})\leq r_{(2)}^2(\boldsymbol{\beta})\leq\cdots\leq r_{(n)}^2(\boldsymbol{\beta})$ である．破綻点は $([n/2]\text{-}p+2)/n$ であり, $n\to 0$ のとき 1/2 となる. LTS 推定量は y 方向だけでなく 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.

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, 高い破綻点をもち, LMS 推定量や LTS 推定量よりも高い漸近効率をもつことを狙ったのが Rousseeuw と Yohai(1984) による S 推定量である. S 推定量は

$$(\hat{\boldsymbol{\beta}}^{s})=\arg\min_{\boldsymbol{\beta}}s(\boldsymbol{\beta}) \tag{9}$$

を満たすものとして定義される．ここで $s(\boldsymbol{\beta})$ は

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\frac{r_i(\boldsymbol{\beta})}{s(\boldsymbol{\beta})}\ )=b,\quad 0\leq b\leq 1 \tag{10}$$

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 はある定数である．すなわち尺度 $s(\boldsymbol{\beta})$ を推定した後, この $s(\boldsymbol{\beta})$ を最小にする $\boldsymbol{\beta}$ を推定量とするものである.

4.5　τ 推定量

τ 推定量は Yohai と Zamar(1988) により提案されたロバスト回帰推定量であり,

$$\hat{\boldsymbol{\beta}}^{\tau}=\arg\min_{\beta\in R^p}\tau(\boldsymbol{\beta}) \tag{11}$$

によって定義される. ここで,$\tau(\boldsymbol{\beta})$ は

$$\tau^2(\boldsymbol{\beta})=s^2(\boldsymbol{\beta})\frac{1}{nb_2}\sum_{i=1}^{n}\rho_2\left(\frac{r_i(\boldsymbol{\beta})}{s(\boldsymbol{\beta})}\right) \tag{12}$$

であり, 尺度 $s(\boldsymbol{\beta})$ は (10) により定義されるものである. ρ_2 は ρ と同じ条件を満たす関数である.

5　多重共線性の存在する人工データ作成法

金・田中 (1993) による多重共線性の存在するデータの作成手順は次の通りである.

5.1　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 $\boldsymbol{V}_{p\times p}$ を作る:
   (a) 線形独立な p 次元ベクトル $\{\boldsymbol{e}_i\}_1^p$ を生成する.
   (b) $\{\boldsymbol{e}_i\}_1^p$ 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 $\{\boldsymbol{v}_i\}_1^p$ に変換し, それを直交行列 $\boldsymbol{V}$ にする.
3. 対角行列 $\boldsymbol{D}_{p\times p}$ 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 $\boldsymbol{D}_{p\times p}$ を作る.
4. 行列 $\boldsymbol{U}_{n\times p}$ を作る:
   (a) $N(0,\boldsymbol{I})$ に従う p 変量正規乱数 $\{\boldsymbol{y}_i\}_1^n$ を発生する.
   (b) $\{\boldsymbol{y}_i\}_1^n$ の平均ベクトル $\bar{\boldsymbol{y}}$ と分散行列 $\boldsymbol{S}$ を計算する.
   (c) $\boldsymbol{S}$ のスペクトル分解 $\boldsymbol{S}$=$\boldsymbol{QGQ}'$ を行う.
   (d) 各 $\boldsymbol{y}_i$ を次のように変換する.
   $$\boldsymbol{z}_i=\boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{Q}'(\boldsymbol{y}_i-\bar{\boldsymbol{y}}),\quad i=1,2,\cdots,n \tag{13}$$
   (e) 各 $\boldsymbol{z}_i'$ を行とする $\boldsymbol{U}_{n\times p}$ を作る.
5. データ $\boldsymbol{X}_{n\times p}$ を作る:
   上で求めた行列 $\boldsymbol{V}$,$\boldsymbol{D}$,$\boldsymbol{U}$ を用いて, 行列の積 $\boldsymbol{UDV}'=\boldsymbol{X}$ を計算し, 人工データを作る.

6 シミュレーション

ここでは多重共線性のデータを用いて, 次の 3 通りの場合についてシミュレーションを行い, ロバスト・リッジ回帰の有効性を調べる.

1. 外れ値がない場合
2. X 方向に外れ値がある場合
3. X と y 方向の両方に外れ値がある場合

6.1 シミュレーションの手順

線形回帰モデル

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon, \quad \varepsilon \sim N(0,1) \tag{14}$$

をモデルとして考え, 回帰係数の真値を $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ とする.

- N(0,1) に従い, 多重共線性をもつ 20 組のデータ

$$(x_{i1}, x_{i2}, x_{i3}, x_{i4}), \qquad i = 1, 2, \cdots, 20$$

を金・田中 (1993) の方法により作成し,

$$y_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + \varepsilon_i$$

とする.

- 20 組のデータ

$$(y_i, x_{i1}, x_{i2}, x_{i3}, x_{i4}), \qquad i = 1, 2, \cdots, 20$$

にモデル (14) を当てはめ, 回帰係数 $\boldsymbol{\beta}$=(β_1, β_2, β_3, β_4)$'$ の推定量 $\hat{\boldsymbol{\beta}}$ を求め, $(\hat{\boldsymbol{\beta}} - \mathbf{1})'(\hat{\boldsymbol{\beta}} - \mathbf{1})$ を計算する.

- X 方向への外れ値として

$$x_5 \sim (1-\eta)N(0,1) + \eta N(8,3)$$

を考え (η =0.2), $\tilde{y}_i$(i=1, 2,⋯,20) を次のように定義する.

$$\tilde{y}_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + \tilde{x}_{i5} + \varepsilon_i.$$

ここで,x_{i5} が外れ値でないとき $\tilde{x_{i5}}$=x_{i5}, x_{i5} が外れ値のとき $\tilde{x}_{i5}$=0 とする. 20 組のデータ

$$(\tilde{y}_i, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}), \qquad i = 1, 2, \cdots, 20$$

にモデル

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon, \varepsilon \sim N(0,1) \tag{15}$$

を当てはめ, 回帰係数 $\boldsymbol{\beta}$=(β_1, β_2, β_3, β_4, β_5)$'$ の推定量 $\hat{\boldsymbol{\beta}}$ を求め, $(\hat{\boldsymbol{\beta}} - \mathbf{1})'(\hat{\boldsymbol{\beta}} - \mathbf{1})$ を計算する.

- y 方向への外れ値として ε^* を

$$\varepsilon^* \sim (1-\eta)N(0,1) + \eta N(8,3)$$

を考え ($\eta = 0.2$), y_i^*(i=1, 2, ⋯, 20) を次のように定義する.

$$y_i^* = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + \tilde{x}_{i5} + \varepsilon_i^* \tag{16}$$

20 組のデータ

$$(y_i^*, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}), \qquad i = 1, 2, \cdots, 20$$

にモデル (15) を当てはめ, 回帰係数 $\boldsymbol{\beta}$=(β_1, β_2, β_3, β_4, β_5)$'$ の推定量 $\hat{\boldsymbol{\beta}}$ を求め, $(\hat{\boldsymbol{\beta}} - \mathbf{1})'(\hat{\boldsymbol{\beta}} - \mathbf{1})$ を計算する.

- この一連の作業を 30 回繰り返し, 第 i 回目で得られる $\hat{\boldsymbol{\beta}}$ を

$$\hat{\boldsymbol{\beta}}_j, \quad j = 1, 2, ..., 30$$

として

$$\widehat{MSE}(\hat{\boldsymbol{\beta}}) = \frac{1}{30} \sum_{j=1}^{30} \{(\hat{\boldsymbol{\beta}}_j - \mathbf{1})'(\hat{\boldsymbol{\beta}}_j - \mathbf{1})\}$$

を求める.

- LS 推定量, LMS 推定量, S 推定量, τ 推定量について k=0, k=0.01, k=0.05 の 3 つの場合を求める.

6.2 シミュレーション結果

シミュレーションによる $MSE(\hat{\beta})$ を下の表に示す. 値が小さいほど精度良く推定できていることになる.

表 1 k=0 の場合のシミュレーション結果

	外れ値なし	X 方向	両方
LS 推定	31.59	114.86	6.63
LMS 推定	287.20	80.02	6.49
S 推定	531.90	38.50	6.35
τ 推定	47.27	49.12	5.70

表 2 k=0.01 の場合のシミュレーション結果

	外れ値なし	X 方向	両方
LS 推定	3.83	4.90	3.56
LMS 推定	7.32	7.84	3.81
S 推定	6.15	6.18	3.63
τ 推定	5.44	7.83	3.78

表 3 k=0.05 の場合のシミュレーション結果

	外れ値なし	X 方向	両方
LS 推定	3.43	4.54	3.56
LMS 推定	5.02	5.93	3.74
S 推定	4.42	4.94	3.59
τ 推定	4.48	6.03	3.66

6.3 外れ値なしの場合の結果と考察

分析結果より, 外れ値なしの場合は k の値に関わらず,> LS 推定が一番精度が良い結果となった. 特に, k=0 のときに, 他の推定量との差が非常に大きくなっている. k の値が大きくなるにつれて他の推定量との差はあまりなく

なっている．また, 図 1 のように LS 推定では各変数の安定する k の値が非常に小さくなっていることが多かった．これより, 外れ値が存在しない場合には, ロバスト・リッジ回帰推定量よりリッジ回帰推定量のほうが適しているということがわかった．

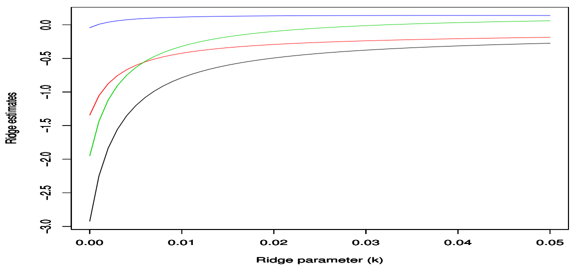

図 1　LS 推定のリッジ・トレース図

6.4　X 方向に外れ値がある場合の結果と考察

X 方向へ外れ値を入れた場合は, 外れ値なしの場合とは逆にロバスト・リッジ推定量のほうが精度が良かった．ロバスト・リッジ推定量の中でも特に S 推定量が一番良かったが, τ 推定量も同じくらいの精度であった．k=0 のときは 3 つのロバスト・リッジ推定量が LS 推定量より精度が良いが, k=0.01, k=0.05 のときは LS 推定量の各変数の推定量が急速に収束しており, LS 推定量のほうがわずかに精度が良くなっている．

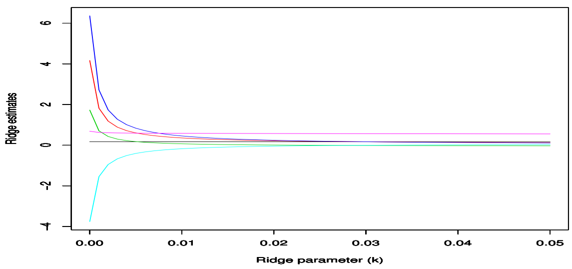

図 2　S 推定のリッジ・トレース図

6.5　X と Y 方向の両方に外れ値がある場合の結果と考察

X 方向と y 方向の両方に外れ値を入れた場合は, τ 推定, S 推定, LMS 推定, LS 推定の順番で精度が良かった．外れ値なしの場合や X 方向に外れ値がある場合とは逆に, ロバスト・リッジ回帰推定量のほうがリッジ回帰推定量より精度の良い結果となった．わずかな差ではあるが, ロバスト・リッジ回帰推定量の中でも特に τ 推定量が一番精度の良い結果となっている．どの分析手法でも, 各変数は推定量が 0 のあたりに収束しているが, X 方向の外れ値の変数のみ, 図の上のほうにある直線のように 0 ではなく 1 に収束していた．また, この変数は k=0 のときの値からあまり変化がなかった．これより, X 方向と Y 方向の両方に外れ値がある場合は, ロバスト・リッジ回帰推定量, 特に τ 推定量が一番良いということがわかった．

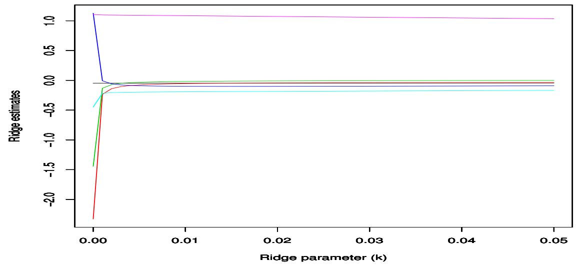

図 3　τ 推定のリッジ・トレース図

6.6　各結果を踏まえたまとめ

外れ値なしの場合と X 方向へ外れ値がある場合は LS 推定量, X と y 方向の両方に外れ値がある場合は τ 推定量が一番良いということがわかった．各ロバスト・リッジ推定は, X と y 方向の両方に外れ値があると, LS 推定より精度が良くなっている．これより, 外れ値が両方向へあるような複雑なデータに対しては, ロバスト・リッジ回帰が有効であり, ロバスト・リッジ推定量でも特に τ 推定量が, 多重共線性と外れ値が混在しているデータに対して安定した分析ができることがわかった．k の値が大きくなるにつれて各変数は 0 付近に収束しており, k の値が 0.01 前後で安定することが多かった．

7　おわりに

外れ値と多重共線性が混在するデータに対して, τ 推定量に基づくロバスト・リッジ回帰はそれらに影響されないことがわかった．今回は $N(0,1)$ に従う独立な説明変数を用いた回帰式に, 多重共線性と外れ値の存在する説明変数を用いた回帰式を当てはめたが, k の値をさらに細かく設定して分析したり, 説明変数をさらに増やすなど研究の余地はまだまだあると感じる．また, 試行回数が少なかったり, τ 推定量以外の他の推定量との比較は行っていないため, 他の推定量に対する τ 推定量の優位性についての分析など, τ 推定を含めたロバスト・リッジ回帰分析のさらなる研究が必要である．

参考文献

[1] 阿部智成, 暮石一樹, 木村美善 (2013)．ロバスト・リッジ回帰推定量について，ACADEMIA Information Sciences and Engineering Vol.13.

[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] Matias,Salibian.Barrera,Gert,Willems,and Ruben,Zamar.(2008),The Fast-τ Estimator for Regression,Journal of Computational and Graphical Statistics,659-682

[4] 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.