τ-リッジ回帰推定量のシミュレーション評価

塚原一翔[1]　木村美善[2]

概要

Silvapulle (1991）は, 線形回帰モデルにおいて多重共線性と目的変数の外れ値が混在する場合には, 通常の最小 2 乗推定量やこれに基づくリッジ回帰推定量（LS-リッジ回帰推定量）では対処できず, M 推定量に基づくリッジ回帰推定量（M-リッジ回帰推定量) を用いるのが望ましいことをシミュレーションにより示した. また, 武山・木村 (2009) と阿部・暮石・木村 (2013) は, 目的変数の外れ値に加えて説明変数に外れ値がある場合に, 様々なロバスト推定量に基づくリッジ回帰推定量（ロバスト・リッジ回帰推定量）を提案し, シミュレーションによりその性能を評価した. そして, ロバスト・リッジ回帰推定量はそれに用いるロバスト推定量の性質を受け継ぎ, 多重共線性と外れ値が混在する場合にはロバスト・リッジ回帰推定量が有効であること, M-リッジ回帰推定量は説明変数の外れ値にはうまく機能しないことを明らかにした. それらのシミュレーション結果は, また, τ-推定量に基づくリッジ回帰推定量（τ-リッジ回帰推定量) がロバスト・リッジ回帰推定量のうちでもバランスよく優れた性質を持つものであることを示唆している. 本論文では, 他のリッジ回帰推定量とのシミュレーション比較によって, τ-リッジ回帰推定量の有効性をさらに解明する.

1　はじめに

線形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が. 多重共線性や外れ値が存在する場合には不安定になり, その良さが失われてしまうことはよく知られている. 説明変数間に強い線形関係が存在するという多重共線性の問題に対して, Hoerl and Kennard (1970a,1970b) は最小 2 乗回帰推定量の安定化をはかるため, パラメータ $k > 0$ を持つリッジ回帰推定量（LS-リッジ回帰推定量）を提案し, その特徴と有効性を明らかにした. リッジ回帰推定量は偏りを持つ推定量であるが, 適切な k を選ぶことにより最小 2 乗推定量よりも小さい平均 2 乗誤差を与えることが可能である（Groβ, 2003, Theorem 3.8）. しかし, この LS-リッジ回帰推定量は最小 2 乗推定量を縮小して作られているため, 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. したがって, 多重共線性と外れ値が同時に生じる場合には, 最小 2 乗推定量に基づく LS-リッジ回帰推定量は好ましくない.

Silvapulle (1991) は多重共線性と目的変数 y に外れ値が混在する場合に, 最小 2 乗推定量ではなく M 推定量を用いたリッジ回帰推定量 (M-リッジ回帰推定量) を提案し, その有効性をシミュレーションにより示した. しかし, 武山・木村 (2008) は この M-リッジ回帰推定量は, 目的変数（誤差）の外れ値に対しては有効であるが, 説明変数 X の外れ値に対し

[1]南山大学数理情報研究科
[2]南山大学情報理工学部　E-mail: kimura@ms.nanzan-u.ac.jp

ては依然として対応できないことをシミュレーションにより明らかにした. . そして, このような多重共線性と外れ値が混在する場合に, M 推定量のみでなく, LMS 推定量, LTS 推定量, GS 推定量や 最深回帰推定量などのロバスト推定量に基づくリッジ回帰推定量（ロバスト・リッジ回帰推定量）を提案し, その有効性をシミュレーションにより明らかにした. また, 阿部・暮石・木村 (2013) は多重共線性があり, 目的変数と説明変数の両方に外れ値があるデータに対して, 様々なロバスト推定量（M, LMS, LTS, S, MM, τ) に基づくリッジ回帰推定量を適用し, シミュレーションによりその性能を評価した. これらのシミュレーション結果は, ロバスト・リッジ回帰推定量がそれに用いるロバスト推定量の性質を受け継ぎ, 多重共線性と外れ値が混在する場合にはロバスト・リッジ回帰推定量が有効であること, とりわけ τ 推定量に基づく回帰推定量（τ-リッジ回帰推定量) がバランスよく優れた性質を持っていることを明らかにした.

本論文では, 多重共線性と外れ値が混在するデータを作成し, このデータを用いて他の推定量（LS, M, LMS, S) に基づくリッジ回帰推定量とのシミュレーション比較をすることにより, τ-リッジ回帰推定量の有効性について考察する.

2　線形回帰モデルとリッジ回帰推定量

目的変数 y と p 個の説明変数 $x_1, x_2, \cdots, x_p$ に関する n 組の観測値 $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, 2, \cdots, n$ が与えられているとし, 線形回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

を考える. ここで, $\beta_0, \beta_1, \cdots, \beta_p$ は回帰係数, ε_i は誤差を表す. このモデルを行列で表記すると

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

となる. ただし

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

である. このとき $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ の最小 2 乗（LS）推定量は, (2) のモデルにおける残差平方和 $(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})$ を最小とするような推定量

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

として定義される. LS 推定量は, 誤差ベクトル $\boldsymbol{\varepsilon}$ が $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$, $V[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}_n$ 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 $N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ に従うときには最良不偏推定量となる. しかし, こうした標準的仮定からの「ずれ」があったり, 外れ値や多重共線性が存在したりする場合には, LS 推定量はその「良さ」を失ってしまうことが知られている.

LS 推定量 $\hat{\boldsymbol{\beta}}^{LS}$ の 総平均 2 乗誤差（TMSE) は $\hat{\boldsymbol{\beta}}^{LS}$ が不偏推定量であるから $\boldsymbol{X'X}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_{p+1} \geq 0$ とすると

$$\mathrm{TMSE}[\hat{\boldsymbol{\beta}}^{LS}] = E[(\hat{\boldsymbol{\beta}}^{LS} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^{LS} - \boldsymbol{\beta})] = \sigma^2 \sum_{i=1}^{p+1} \frac{1}{\lambda_i} \tag{4}$$

となる. TMSE はの真の回帰ベクトル $\boldsymbol{\beta}$ からの推定量 $\hat{\boldsymbol{\beta}}^{LS}$ の平均的なずれの大きさを表すものであり, 可能な限り小さいことが望ましい. しかし, データに多重共線性があるとき, 固有値 λ には極めて 0 に近いものが存在するため, TMSE[$\hat{\boldsymbol{\beta}}^{LS}$] は大きくなってしまう.

Hoerl and Kennard (1970a) 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 LS 推定量 $\hat{\boldsymbol{\beta}}^{LS}$ を縮小することによって推定の安定化を図るリッジ回帰推定量 (LS-リッジ回帰推定量)

$$\hat{\boldsymbol{\beta}}^{LS}(k) = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'y} = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'X}\hat{\boldsymbol{\beta}}^{LS} \tag{5}$$

を提案した. ここで, $\hat{\boldsymbol{\beta}}^{LS}(0) = \hat{\boldsymbol{\beta}}^{LS}$ であることに注意する. このとき

$$\mathrm{TMSE}[\hat{\boldsymbol{\beta}}^{LS}(k)] \quad = \quad \sigma^2 \sum_{i=1}^{p+1} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \boldsymbol{\beta}'(\boldsymbol{X'X} + k\boldsymbol{I})^{-2}\boldsymbol{\beta} \tag{6}$$

が成り立つ (Chatterjee, Hadi and Price, 2006). 右辺の第 1 項は総分散で k に関して単調減少であり, 第 2 項は偏りの 2 乗で k に関して単調増加する. Hoerl and Kennard (1970a) は, TMSE[$\hat{\boldsymbol{\beta}}^{LS}(k)$] <TMSE[$\hat{\boldsymbol{\beta}}^{LS}$] を満たす $k > 0$ が存在することを示した. TMSE[$\hat{\boldsymbol{\beta}}^{LS}(k)$] を小さくする k の決定方法としては, 数式で与えられるものとリッジ・トレースによって視覚的に決めるものの 2 種類がある. 前者の k の計算式としては, これまでに様々なものが提案されており, それらのシミュレーションによる比較研究が Kibria (2003) により行われているが, どれも決め手がない状況である. 後者のリッジ・トレースとは横軸にパラメータ k, 縦軸に各回帰係数の推定値を取り, プロットしてできるグラフである. Hoerl and Kennard (1970a) はこのリッジ・トレースが安定する k の値が望ましいものであり, この k を採用するのがよいと主張しているが, その後, 今日に至るまで 多くの研究者たちが, このリッジ・トレースを用いる方法を実用的な方法として好ましいと評価している. リッジ回帰の全体的な解説は Groβ (2003) が詳しい.

3　ロバスト・リッジ回帰推定量

説明変数 $\boldsymbol{X}$ の多重共線性と目的変数 y の外れ値とが混在するデータに対して, Silvapulle (1991) は LS 推定量 $\hat{\boldsymbol{\beta}}^{LS}$ を用いる通常の LS-リッジ回帰推定量 $\hat{\boldsymbol{\beta}}^{LS}(k)$ は外れ値から大きな影響を受けるため好ましいものでなく, その代わりに M 推定量を用いた M-リッジ回帰推定量が有効であることをシミュレーションにより示した. しかし, y の外れ値だけでなく $\boldsymbol{X}$ に外れ値と多重共線性が同時に含まれるデータに対しては, その有効性は損なわれてし

まうことを武山・木村 (2008) は示し, この場合に適切で望ましい推定量として, 説明変数の外れ値にも対応できるロバスト回帰推定量 $\hat{\boldsymbol{\beta}}^{rob}$ に基づくロバスト・リッジ回帰推定量

$$\hat{\boldsymbol{\beta}}^{rob}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}^{rob} \tag{7}$$

を提案した. そして, $\hat{\boldsymbol{\beta}}^{rob}$ として, M 推定量 $\hat{\boldsymbol{\beta}}^{M}$, LMS 推定量 $\hat{\boldsymbol{\beta}}^{LMS}$, LTS 推定量 $\hat{\boldsymbol{\beta}}^{LTS}$, GS 推定量 $\hat{\boldsymbol{\beta}}^{GS}$, 最深回帰推定量 $\hat{\boldsymbol{\beta}}^{DR}$ を用いたロバスト・リッジ回帰推定量を LS-リッジ回帰推定量 $\hat{\boldsymbol{\beta}}^{LS}(k)$ とシミュレーション比較し, その有効性を明らかにした. ここで, $\hat{\boldsymbol{\beta}}^{GS}$ は Croux, Rousseeuw and Hossjer (1994) により提案された推定量であり, $\hat{\boldsymbol{\beta}}^{DR}$ は Rousseeuw and Hubert (1999) により提案された推定量である. $\hat{\boldsymbol{\beta}}^{M}(k)$ が Silvapulle (1991) の提案したリッジ回帰推定量である. また, 阿部・暮石・木村 (2013) は多重共線性と外れ値が混在するデータに対して, $\hat{\boldsymbol{\beta}}^{LS}$ に加えて　$\hat{\boldsymbol{\beta}}^{M}$, $\hat{\boldsymbol{\beta}}^{LMS}$, $\hat{\boldsymbol{\beta}}^{LTS}$, S 推定量 $\hat{\boldsymbol{\beta}}^{S}$, MM 推定量 $\hat{\boldsymbol{\beta}}^{MM}$ および τ 推定量 $\hat{\boldsymbol{\beta}}^{\tau}$ を用いたリッジ回帰推定量をシミュレーション比較した. そして, シミュレーションの結果として, とりわけ τ-リッジ回帰推定量 $\hat{\boldsymbol{\beta}}^{\tau}(k)$ が様々な状況でバランスよく優れた性質を持っていると指摘している. 本論文の第 4 節では, $\hat{\boldsymbol{\beta}}^{\tau}(k)$ の有効性について $\hat{\boldsymbol{\beta}}^{LS}(k)$, $\hat{\boldsymbol{\beta}}^{M}(k)$, $\hat{\boldsymbol{\beta}}^{LMS}(k)$ および $\hat{\boldsymbol{\beta}}^{S}(k)$ とのシミュレーション比較により調べる. このシミュレーションで用いられるロバスト推定量の定義は次の通りである.

- **M 推定量** $\hat{\boldsymbol{\beta}}^{M}$: 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

$$\hat{\boldsymbol{\beta}}^{M} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})), \quad r_i(\boldsymbol{\beta}) = y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \tag{8}$$

として定義される. 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) によるものと Tukey による biweight (Beaton and Tukey, 1974 参照) がよく知られている. (8) 式からもわかるように, $\rho(t) = t^2$ とすると, これは LS 推定量に等しい.

- **LMS 推定量** $\hat{\boldsymbol{\beta}}^{LMS}$: 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

$$\hat{\boldsymbol{\beta}}^{LMS} = \arg\min_{\boldsymbol{\beta}} \mathrm{med}\{r_1^2(\boldsymbol{\beta}), \cdots, r_n^2(\boldsymbol{\beta})\} \tag{9}$$

として定義される. 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 のみでなく $\boldsymbol{X}$ の外れ値に対してもロバストであるが, 漸近効率は高くない.

- **S 推定量** $\hat{\boldsymbol{\beta}}^{S}$: S 推定量は Rousseeuw and Yohai (1984) によって提案されたもので,

$$\hat{\boldsymbol{\beta}}^{S} = \arg\min_{\boldsymbol{\beta}} \ s_n(\boldsymbol{\beta}) \tag{10}$$

により定義される. ここで $s_n(\boldsymbol{\beta})$ は

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\boldsymbol{\beta})}\right) = b, \qquad 0 \le b \le 1 \tag{11}$$

を満たすものである．ρ_1 は $(-\infty, \infty)$ 上の有界関数であり, 原点対称, 連続微分可能, $\rho_1(0) = 0$ かつ ある定数 $c > 0$ に対して $[0, c]$ 上で狭義単調増大, $[c, \infty]$ 上で定数である．

- τ **推定量** $\hat{\boldsymbol{\beta}}^{\tau}$: τ 推定量は. Yohai and Zamar (1988) により提案されたもので

$$\hat{\boldsymbol{\beta}}^{\tau} = \arg\min_{\boldsymbol{\beta}} \ \tau_n(\boldsymbol{\beta}) \tag{12}$$

により定義される．ここで $\tau_n(\boldsymbol{\beta})$ は

$$\tau_n^2(\boldsymbol{\beta}) = s_n^2(\boldsymbol{\beta}) \frac{1}{n} \sum_{i=1}^{n} \rho_2\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\boldsymbol{\beta})}\right) \tag{13}$$

であり, $s_n(\boldsymbol{\beta})$ は (11) により与えられるものである．また, ρ_2 は ρ_1 と同じ条件を満たす関数である．τ 推定量は ρ_1 により高い破綻点を持ち, ρ_2 により高い効率を得るように工夫された推定量である．そして, ρ_1 と ρ_2 をそれぞれ適切に選ぶことにより, 破綻点が最大の 0.5 を持つようにできるし, 正規分布の下での効率を最大の 1 に近づけることもできる. このように τ-推定量は柔軟性のある優れた推定量であるが, 推定値を得るための計算が難しいためか, まだ, 実用面であまり使用されていない状況である (Saribian-Bera, Willems and Zamar, 2008).

3.1 多重共線性を持つデータの作成

説明変数間に多重共線性がある場合には, LS-回帰推定量が不安定になり, 分析結果も不明確になってしまう. さらに, 外れ値も混在する場合には, LS-リッジ回帰推定量は対処できず, ロバスト・リッジ回帰推定量がうまく機能する. このことをシミュレーションで調べるためには, 多重共線性と外れ値が混在するデータを必要とする. 多重共線性を持つデータの作り方はいろいろとあるが, 金・田中（1993) の方法は次の通りである. 阿部・暮石・木村 (2013) のシミュレーションにおいてもこの方法が用いられている.

作成手順

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:
   (1) 線形独立な p 次元ベクトル $\{\boldsymbol{e}_i\}_1^p$ を生成する.
   (2) $\{\boldsymbol{e}_i\}_1^p$ をグラム・シュミットの直交化法を用いて, 各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:
   (1) 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.

(2) 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:

行列 $\boldsymbol{U}$ の作り方としては 3 通り提案されているが, 2 番目の正確な分布データの方法を用いる.

(1) $N(\boldsymbol{0}, \boldsymbol{I})$ に従う p 変量正規乱数 $\{\boldsymbol{y}_i\}_1^n$ を発生する.

(2) $\{\boldsymbol{y}_i\}_1^n$ の平均ベクトル $\bar{\boldsymbol{y}}$ と分散行列 $\boldsymbol{S}$ を計算する.

(3) $\boldsymbol{S}$ のスペクトル分解 $\boldsymbol{S} = \boldsymbol{QGQ'}$ を行う.

(4) 各 $\boldsymbol{y}_i$ を次のように変換する. ただし, $\boldsymbol{G}$ の対角要素 $g_{ii} \leq 0$ のものがあれば, $\boldsymbol{G}^{-1/2}$ の対応する要素を 0 とする.

$$\boldsymbol{z}_i = \boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{Q'}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}), \quad i = 1, 2, \cdots, n \tag{14}$$

(5) 各 $\boldsymbol{z}_i'$ を行とする $\boldsymbol{U}_{n\times p}$ を作る.

(a) データ $\boldsymbol{X}_{n\times p}$ を作る:
行列 $\boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U}$ を用いて $\boldsymbol{X} = \boldsymbol{UDV'}$ とする.

4 シミュレーション

多重共線性のあるデータを金・田中 (1993) の方法で作り, 次の 3 通りの場合について, 5 種類のリッジ回帰推定量 ($\hat{\boldsymbol{\beta}}^{LS}(k)$, $\hat{\boldsymbol{\beta}}^{M}(k)$, $\hat{\boldsymbol{\beta}}^{LMS}(k)$, $\hat{\boldsymbol{\beta}}^{S}(k)$, $\hat{\boldsymbol{\beta}}^{\tau}(k)$) をシミュレーション評価する.

1. 外れ値がない場合（多重共線性のみ）
2. 説明変数 X に外れ値がある場合（多重共線性と X の外れ値の混在)
3. 説明変数 X と目的変数 (多重共線性と X および y の外れ値の混在）

外れ値が入る 2 と 3 の場合には, 多重共線性を保つように、目的変数 y と説明変数 X に外れ値を入れる工夫をし, 多重共線性と外れ値の両方が混在するデータを作成する. 評価の基準としては総平均 2 乗誤差の推定値を用いる. なお, シミュレーションの計算には統計解析ソフト R を使用する.

4.1 シミュレーションの手順

- 外れ値がない場合.

線形回帰モデルとして

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

を考え, 回帰係数の真値を $\beta_0=\beta_1=\beta_2=\beta_3=\beta_4=1$ とする. $x_{i1},\ x_{i2},\ x_{i3},\ x_{i4},\ i=1,2,\cdots,20$ はそれぞれ $N(0,1)$ に従い, 多重共線性をもつように金・田中 (1993) の方法により作成する. そして, $N(0,1)$ に従う $\varepsilon_i,\ i=1,2,\cdots,20$ を用いて

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

とし, 20 組のデータ

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

を作る. このデータに対してモデル (15) を当てはめ, 回帰推定値 $\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3,\beta_4)'$ 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. ただし, $\mathbf{1}=(1,\cdots,1)'$.

- X に外れ値がある場合.
  X に外れ値を入れるために $x_1,\ x_2,\ x_3,\ x_4$ と独立な次の説明変数

$$x_5\sim(1-\eta)N(0,1)+\eta N(8,9)$$

を導入する. そして, $\eta=0.15$ と $\eta=0.25$ に対して, この x_5 のデータ $x_{i5},\ i=1,2,\cdots,20$ を用いて新たなデータ $\tilde{x}_{i5}$ を $\tilde{x}_{i5}=x_{i5}$ (x_{i5} が外れ値でないとき), $\tilde{x}_{i5}=0$ (x_{i5} が外れ値のとき) と定義し

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

とする. このように作った 20 組のデータ

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

に対してモデル

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\varepsilon,\quad \varepsilon\sim N(0,1)\qquad(16)$$

を当てはめ, 回帰係数 $\boldsymbol{\beta}$=$(\beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5)'$ 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 と X に外れ値がある場合.
  y に外れ値を入れるために誤差

$$\varepsilon^*\sim(1-\eta)N(0,1)+\eta N(8,9)$$

を導入し, $\eta=0.15$ と $\eta=0.25$ に対して

$$y_i^*=1+x_{i1}+x_{i2}+x_{i3}+x_{i4}+\tilde{x}_{i5}+\varepsilon_i^*,\quad i=1,2,\cdots,20\qquad(17)$$

とする. このようにして作った 20 組のデータ

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

にモデル (16) を当てはめ, 回帰係数 $\boldsymbol{\beta}$=$(\beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5)'$ 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 j 回目で得られる $\hat{\boldsymbol{\beta}}$ を $\hat{\boldsymbol{\beta}}_j,\ j = 1, 2, \cdots, 30$ として

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

を計算する.

- このシミュレーションを 5 種類のリッジ回帰推定量 $\hat{\boldsymbol{\beta}}^{LS}(k)$, $\hat{\boldsymbol{\beta}}^{M}(k)$, $\hat{\boldsymbol{\beta}}^{LMS}(k)$, $\hat{\boldsymbol{\beta}}^{S}(k)$, $\hat{\boldsymbol{\beta}}^{\tau}(k)$ に対して, それぞれ k=0, k=0.01, k=0.05 の 3 通り行う.

4.2 シミュレーション結果と考察

シミュレーションによる $\widehat{MSE}(\hat{\boldsymbol{\beta}}(k))$ の値を表 1, 表 2, 表 3 に示す. $\widehat{MSE}(\hat{\boldsymbol{\beta}}(k))$ の値が小さいほど推定値は真値に近く, 推定の精度が高いといえる. また, 図 1, 図 2, 図 3 はそれぞれ $\hat{\boldsymbol{\beta}}^{LS}(k)$, $\hat{\boldsymbol{\beta}}^{S}(k)$, $\hat{\boldsymbol{\beta}}^{\tau}(k)$ のリッジ・トレースである.

表 1: $\widehat{MSE}(\hat{\boldsymbol{\beta}}(k)),\quad k = 0$

	外れ値なし	η	X の外れ値あり	y と X の外れ値あり
$\hat{\boldsymbol{\beta}}^{LS}(k)$	31.59	0.15	262.56	23.19
		0.25	330.70	25.36
$\hat{\boldsymbol{\beta}}^{M}(k)$	38.61	0.15	289.33	14.96
		0.25	384.52	22.16
$\hat{\boldsymbol{\beta}}^{LMS}(k)$	287.20	0.15	14.76	4.32
		0.25	20.78	4.72
$\hat{\boldsymbol{\beta}}^{S}(k)$	531.90	0.15	13.16	4.40
		0.25	17.23	4.45
$\hat{\boldsymbol{\beta}}^{\tau}(k)$	47.27	0.15	13.37	4.30
		0.25	12.33	4.38

4.2.1 外れ値なしの場合

k が増加すると $\widehat{MSE}(\hat{\boldsymbol{\beta}}(k))$ は減少し, すべてのリッジ回帰推定量が多重共線性に対処できている. そして, LS と M が良い. $k = 0$ のときは LS が最も良いが, $k = 0.01, 0.05$ のときは M が最も良い. τ は LS と M と比べてもそれほど悪くない. LMS と S は $k = 0$ のとき極端に悪い. 全体的にみると M が LS と同じくらい良いが, τ も悪くない. LMS と S は良くない.

表 2:　$\widehat{MSE}(\hat{\boldsymbol{\beta}}(k)),\quad k=0.01$

	外れ値なし	η	X の外れ値あり	y と X の外れ値あり
$\hat{\boldsymbol{\beta}}^{LS}(k)$	3.83	0.15	15.31	4.69
		0.25	16.26	7.00
$\hat{\boldsymbol{\beta}}^{M}(k)$	3.22	0.15	14.11	4.23
		0.25	16.83	6.36
$\hat{\boldsymbol{\beta}}^{LMS}(k)$	7.32	0.15	8.67	4.00
		0.25	11.00	4.31
$\hat{\boldsymbol{\beta}}^{S}(k)$	6.15	0.15	7.47	3.97
		0.25	7.19	3.98
$\hat{\boldsymbol{\beta}}^{\tau}(k)$	5.44	0.15	7.74	3.74
		0.25	7.13	3.94

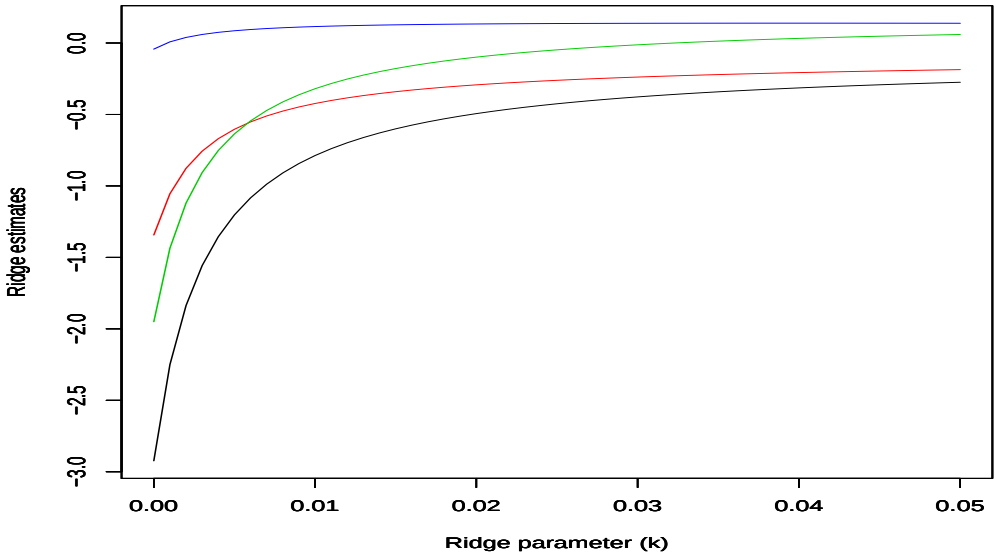

図 1:　$\hat{\boldsymbol{\beta}}^{LS}(k)$ のリッジ・トレース

4.2.2　X に外れ値がある場合

k が増加すると $\widehat{MSE}(\hat{\boldsymbol{\beta}}(k))$ は減少し, すべてのリッジ回帰推定量が多重共線性に対処できている. 外れ値の割合 η が増えると S と τ 以外はすべての k に対して増加する. 外れ値がない場合とは逆に LS と M が悪く, LMS, S, τ が良い. $k=0$ のときは, LS と M は特に悪い. 全体的には S と τ が同じくらい最も良い. M が X の外れ値に対応できないことがわかり, S と τ が X の外れ値に強いこともわかる.

表 3:　$\widehat{MSE}(\hat{\boldsymbol{\beta}}(k)), \quad k = 0.05$

	外れ値なし	η	X の外れ値あり	y と X の外れ値あり
$\hat{\boldsymbol{\beta}}^{LS}(k)$	3.43	0.15	10.07	4.13
		0.25	11.29	4.86
$\hat{\boldsymbol{\beta}}^{M}(k)$	3.07	0.15	9.50	3.90
		0.25	10.99	4.61
$\hat{\boldsymbol{\beta}}^{LMS}(k)$	5.02	0.15	6.00	3.84
		0.25	6.25	4.05
$\hat{\boldsymbol{\beta}}^{S}(k)$	4.42	0.15	5.67	3.88
		0.25	5.06	3.84
$\hat{\boldsymbol{\beta}}^{\tau}(k)$	4.48	0.15	5.94	3.65
		0.25	4.97	3.81

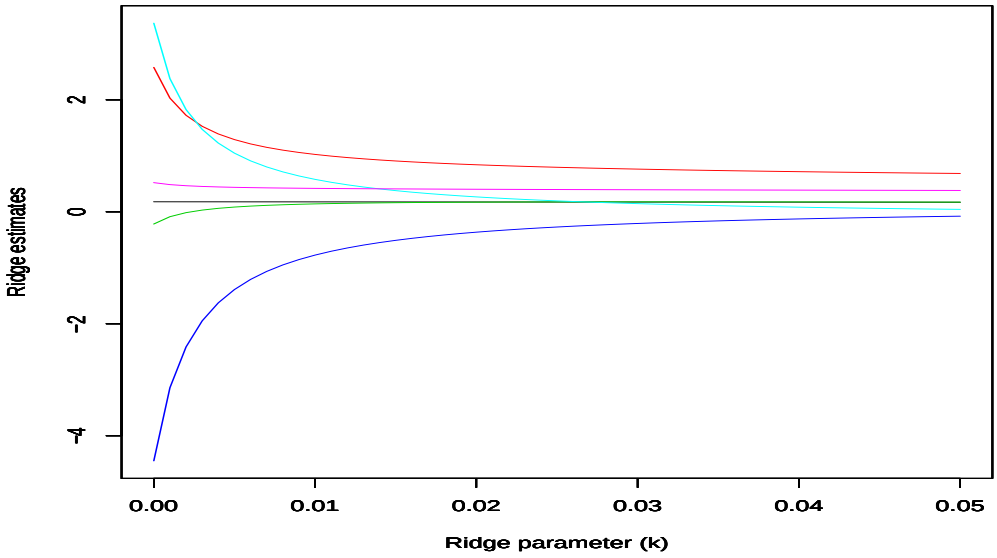

図 2:　$\hat{\boldsymbol{\beta}}^{S}(k)$ のリッジ・トレース

4.2.3　X と y の両方に外れ値がある場合

k が増加すると $\widehat{MSE}(\hat{\boldsymbol{\beta}}(k))$ は減少し, この場合もすべてのリッジ回帰推定量が多重共線性に対処できている. 特に LS と M の減りが大きく外れ値の影響が少なくなっている. η が増加すると, すべての推定量で増える. X に外れ値がある場合と同様に M 以外のロバスト・リッジ推定量が良い. ロバスト・リッジ推定量の中では τ-リッジ推定が k と η の値によらず最も良い.

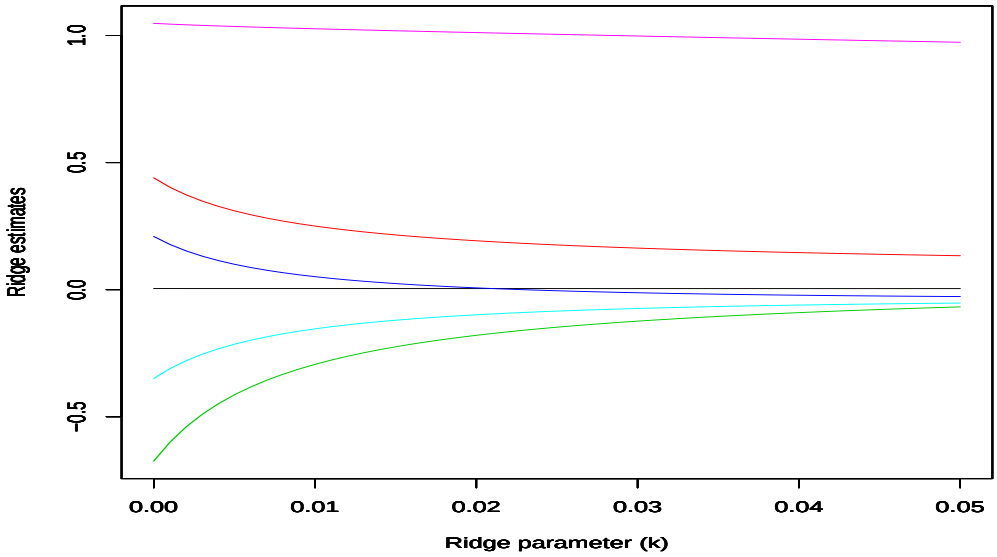

図 3: $\hat{\boldsymbol{\beta}}^{\tau}(k)$ のリッジ・トレース

4.2.4 総合的考察

多重共線性に対しては, すべてのリッジ回帰推定量が対応できている. LS-リッジ回帰推定量は外れ値のない場合は良いが, 外れ値に非常に弱い. M-リッジ回帰推定量は X の外れ値にうまく機能しない. X に外れ値がある場合と, X と y の両方に外れ値がある場合は M-リッジ回帰推定量以外のロバスト・リッジ推定量が良いが, 特に τ 推定量が良い. 外れ値のない場合も悪くないことを考慮すると, 最もバランスがよく優れているのは τ-リッジ回帰推定量であるといえる. また, リッジ回帰推定量にはそれに用いられている推定量の性質が強く反映していることもわかる. k の値が 0.01 前後で安定することが多かったので, k を 0.01 と 0.05 とした.

5　おわりに

本研究では, 多重共線性のあるデータに対してリッジ回帰推定量が有効であること, 多重共線性だけでなく外れ値も同時に含むデータに対しては LS 回帰推定量や LS-リッジ回帰推定量ではうまく対処できず, ロバスト・リッジ回帰推定量が有効であること, さらに, M-リッジ回帰推定量は X の外れ値に対しては機能しないことをシミュレーションにより確認した. また, 阿部・暮石・木村 (2013) はシミュレーションの比較評価の結果からロバスト・リッジ回帰推定量の中で τ-リッジ回帰推定量が優れていると述べており, 本研究では τ-リッジ回帰推定量を中心としたシミュレーション評価を行った. シミュレーションの設定や評価基準は異なるが, 同じように τ-リッジ回帰推定量の優位性が見られる結果を得た. このように τ-リッジ回帰推定量は魅力的であるが, 計算が難しいこともあり, τ-推定量自体がまだ R に実装されていない. 本研究では室梅秀平氏（南山大学数理情報研究科 2012 年度修了) 作成の τ-推定量計算プログラムを利用させていただいた. 室梅氏には感謝したい. 今後の課題としては, τ-リッジ回帰推定量の有効性をさらに明確にするために, 説明変数を増やし, 多重共線性と外れ値の混在の影響をもっと多様な状況のもとで調べるとともに, シミュレーション精度をさらに上げる必要があると思われる.

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とそのシミュレーション評価, 「アカデミア」情報理工学編, **13**, 47-59.

[2] Chatterjee, S., Hadi, A. S. and Price, B. (2006). *Regression Analysis by Example,* Forth Edition, John Wiley & Sons.

[3] Croux, C., Rousseeuw, P. J. and Hössjer, O. (1994). Generalized S-estimators, *Jornal of the American Statistical Association.,* **89**, 1271-1281.

[4] Groβ, J. (2003). *Linear Regression,* Springer.

[5] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics.,* **12**, 55-67.

[6] Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems, *Technometrics.,* **12**, 69-82.

[7] Huber, H. J. (1964). Robust estimation of a location parameter , *The Annals of Statistics,* **35**, 73-101.

[8] Kibria, B. M. G. (2003). Performance of some new ridge regression estimators, *Communications in Statistics—Teory and Methods.,* **32**, 419-435.

[9] 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, 日本計算機統計学会シンポジウム論文集 (8), 26-29.

[10] Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871-880

[11] Rousseeuw, P. J. and Hubert, M. (1999). Regression depth, *Journal of the American Statistical Association,* **94**, 388-402.

[12] Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators, *Robust and Nonlinear Time Series Analysis. LectureNotes in Statistics.,* **26**, eds. J. Franke, W. Härdle, and R. D. Martin, New York, Springer-Verlag, pp. 256-272.

[13] Silvapulle, M. J. (1991). Robust ridge regression based on an M-estimator, *Australian Journal of Statistics,* **33**, 319-333.

[14] Saliban-Barrera, M., Willems, G. and Zamar, R. (2008). The fast-τestimator for regression, *Journal of Computational and Graphical Statistics,* **17**, 659-682.

[15] 武山嵩弘・木村美善. (2008). ロバストリッジ回帰推定量とそのシミュレーション評価, 「アカデミア」数理情報編, **8**, 35-46.

[16] Yohai, V. J. (1987). High breakdown-point and high efficiency estimates for regression , *The Annals of Statistics,* **15**, 642-656.

[17] Yohai, V. J. and Zamar, R. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale , *Journal of the American Statistical Association,* **83**, 406-413.