ロバスト・リッジ回帰推定量について

阿 部 智 成[1]　暮 石 一 樹[2]　木 村 美 善[3]

概要

回帰モデルにおいて多重共線性と外れ値が混在する場合には, 最小 2 乗推定量やこれに基づく従来のリッジ回帰推定量ではうまく機能せず, ロバスト推定量に基づくリッジ回帰推定量が望ましいことを Silvapulle (1991) と武山・木村（2009）がシミュレーションにより示した. 前者はこのロバスト推定量として M 推定量を用いて考察しており, 後者は LMS, LTS, GS および 最深回帰推定量を用いている. 本論文では, ロバスト推定量として, M, LMS, LTS に加えて新たに S, MM, τ 推定量を取り上げ, これらのロバスト推定量に基づくリッジ回帰推定量の有効性についてシミュレーションにより明らかにする.

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最良線形不偏となり, さらに正規分布の下では最良不偏となる望ましい推定量である. しかし, この最小 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. しかし, この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 乗推定量に基づくリッジ回帰推定量は好ましくない.

Silvapulle (1991) は多重共線性と外れ値が混在する場合に, 最小 2 乗推定量の代わりに M 推定量を用いた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, このリッジ回帰推定量は M 推定量に基づいていることから, 誤差の外れ値に対しては有効であるが, 説明変数の外れ値に対しては依然として対応できていない. 説明変数の外れ値にも対応するためには, 説明変数に対してもロバストな推定量に基づくリッジ回帰推定量を用いることが必要と考えられる. 武山・木村 (2008) は, 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, LTS, GS (generalized S estimator) や 最深回帰推定量 (deepest regression estimator) といったロバスト推定量に

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

基づく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.

本論文では, 多重共線性と外れ値が混在するデータに対して, M, LMS, L 推定量に加えて新たに S, MM および τ 推定量に基づくリッジ回帰推定量を提案し, その有効性についてシミュレーションにより明らかにする.

2　線形回帰モデルと最小 2 乗推定量

応答変数 $\mathbf{y}$ と p 個の説明変数 $x_1, x_2, \cdots, x_p$ に関する n 個の観測値 $y_i, x_{i1}, x_{i2}, \cdots, x_{ip}$, $i = 1, \cdots, n$ 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

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

として定義される. LS 推定量は, 誤差ベクトル $\boldsymbol{\varepsilon}$ が $E[\boldsymbol{\varepsilon}] = \mathbf{0}$, $V[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}_n$ 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 $N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ 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. そして, 実際のデータ解析において, 標準的仮定は近似的にしか満たされないことが多い.

3　多重共線性とリッジ回帰推定量

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

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

と表される. TMSE はの真の回帰ベクトルからの推定量の平均的な離れの大きさ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. 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, (4) 式による $\hat{\boldsymbol{\beta}}$のTMSE 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, (3) 式の LS 推定量 $\hat{\boldsymbol{\beta}}$ を縮小することによって回帰推定値の安定化を図るリッジ (RID) 回帰推定量

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

を提案した. この $\hat{\boldsymbol{\beta}}(k)$ は, $k = 0$ のとき, $\hat{\boldsymbol{\beta}}$ に等しい. リッジ回帰推定量 $\hat{\boldsymbol{\beta}}(k)$ の TMSE は

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

である (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[$\hat{\boldsymbol{\beta}}(k)$] <TMSE[$\hat{\boldsymbol{\beta}}$] を満たす $k > 0$ が存在することを示した. TMSE[$\hat{\boldsymbol{\beta}}(k)$] 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.

4 ロバスト回帰推定量

本論文で用いるロバスト推定量は次の通りである.

- **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

$$\hat{\boldsymbol{\beta}}^M = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})), \quad r_i(\boldsymbol{\beta}) = y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{7}$$

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. (7) 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

$$\hat{\boldsymbol{\beta}}^{LMS} = \arg\min_{\boldsymbol{\beta}} \mathrm{med}\{r_1^2(\boldsymbol{\beta}), \cdots, r_n^2(\boldsymbol{\beta})\} \tag{8}$$

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 方向のみでなく $\boldsymbol{X}$ 方向に対してもロバストであるが, 漸近効率は高くない.

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

$$\hat{\boldsymbol{\beta}}^{LTS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{m} r_{(i)}^2(\boldsymbol{\beta}) \tag{9}$$

として定義される．ここで $r_{(1)}^2(\boldsymbol{\beta}) \le r_{(2)}^2(\boldsymbol{\beta}) \le \cdots \le r_{(n)}^2(\boldsymbol{\beta})$. 破綻点は $([n/2]-p+2)/n$ であり, $n \to 0$ のとき $1/2$ となる. LTS 推定量は y 方向のみでなく $\boldsymbol{X}$ 方向に対してもロバストであるが, 漸近効率は高くない.

- **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．

- **MM 推定量** $\hat{\boldsymbol{\beta}}^{MM}$: MM 推定量は Yohai (1987) により提案されたものであり

$$\hat{\boldsymbol{\beta}}^{MM} = \arg\min_{\boldsymbol{\beta}} \eta_n(\boldsymbol{\beta}) \tag{12}$$

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

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

であり, s_n は (11) により定義される $s_n(\boldsymbol{\beta})$ の最小値で, ρ_2 は ρ_1 と同じ条件を満たす関数.

- 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

$$\hat{\boldsymbol{\beta}}^{\tau} = \arg\min_{\boldsymbol{\beta}} \ \tau_n(\boldsymbol{\beta}) \tag{14}$$

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

$$\tau_n^2(\boldsymbol{\beta}) = s_n^2(\boldsymbol{\beta})\frac{1}{n}\sum_{i=1}^{n}\rho_2(\frac{r_i(\boldsymbol{\beta})}{s_n(\boldsymbol{\beta})}) \tag{15}$$

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, そして 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．

5　ロバスト・リッジ回帰推定量

データに目的変数 y 方向の外れ値と説明変数 $\boldsymbol{X}$ の多重共線性とが混在する場合に, Silvapulle (1991) は LS 推定量を用いる通常のリッジ回帰推定量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 方向の外れ値だけでなく $\boldsymbol{X}$ に外れ値と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

$$\hat{\boldsymbol{\beta}}^{rob}(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}^{rob} \tag{16}$$

を提案した．そして $\hat{\boldsymbol{\beta}}^{rob}$ として、$\hat{\boldsymbol{\beta}}^{LMS}$, $\hat{\boldsymbol{\beta}}^{LTS}$, $\hat{\boldsymbol{\beta}}^{GS}$, $\hat{\boldsymbol{\beta}}^{DR}$ を用いたロバスト・リッジ回帰推定量 $\hat{\boldsymbol{\beta}}^{rob}(k)$ を取り上げて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 GS (generalized S) 推定量であり, $\hat{\boldsymbol{\beta}}^{DR}$ は Rousseeuw and Hubert (1999) により提案された最深回帰 (deepest regression) 推定量である．以下の節では, $\hat{\boldsymbol{\beta}}^{LMS}$, $\hat{\boldsymbol{\beta}}^{LTS}$ に加えて, 新たに S 推定量, MM 推定量, τ 推定量に基づくロバスト・リッジ回帰推定量 $\hat{\boldsymbol{\beta}}^{S}(k)$, $\hat{\boldsymbol{\beta}}^{MM}(k)$ および $\hat{\boldsymbol{\beta}}^{\tau}(k)$ の有効性をシミュレーションにより明らかにする．

6　シュミレーション 1

外れ値と多重共線性の両方が混在するデータを作成し, ロバスト・リッジ推定量 $\hat{\boldsymbol{\beta}}^{M}(k)$, $\hat{\boldsymbol{\beta}}^{S}(k)$, $\hat{\boldsymbol{\beta}}^{MM}(k)$ の有効性と精度についてシミュレーションにより調べる．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．

6.1　データの作成

作成するデータは標本数が 100 で, 3 つの説明変数 x_1, x_2, x_3 を正規乱数により生成し，多重共線性として x_1 と x_2 に強い相関を持たせた．重回帰モデルは

$$y = x_1 + x_2 + x_3 + \varepsilon$$

である．次の 3 通りの場合に分けて考察する．

1. 目的変数 y のみが汚染される（外れ値をもつ）場合:

$$\varepsilon \sim (1-\eta)\cdot N(0,0.1)+\eta\cdot t(0) \tag{17}$$

ここで $t(0)$ はコーシー分布であり, η は混合の割合を表す定数である.

2. 説明変数 x_3 が汚染される（外れ値をもつ）場合：

$$x_3 \sim (1-\eta)\cdot N(0,0.1)+\eta\cdot t(0) \tag{18}$$

3. 目的変数 y と説明変数 x_3 の両方が汚染される（外れ値をもつ）場合:
誤差 ε と説明変数 x_3 に (17) と (18) を仮定する.

リッジ回帰推定量に必要となるパラメータ k に関しては固定せず，各回のシミュレーションにおいて, k^* を範囲 $(0,1)$ で $k^*=0.0001$ ごとに増加させ $(\hat{\boldsymbol{\beta}}(k^*)-\mathbf{1})'(\hat{\boldsymbol{\beta}}(k^*)-\mathbf{1})$ を最小にする（TMSE を最小にする）k^* を k として用いる. 有効性を判断するための方法としては, 1000 回のシミュレーションの各回ごとに生成されるそれぞれの回帰推定値が, 汚染も多重共線性もない場合に求められた回帰係数の 95% 信頼区間に何回入るかによって判断する. 説明変数は 3 個あるため, 3 個の回帰推定値のすべてが信頼区間に入った場合を数える.

6.2 実行結果と考察

3 通りのシミュレーションの結果がそれぞれ表 1, 2, 3 である. ここで汚染率は η の値, RID は LS に基づく通常のリッジ回帰推定量である. 表 2 よりわかるように, 汚染がなく共線関係のみのデータ (0%) で最小 2 乗推定量（LS) はすでに 5 割しか信頼区間に入っておらず，1% の汚染により，2 割を切ってしまっている. このことからも最小 2 乗推定量は多重共線性や外れ値に対して非常に弱いことが分かる. リッジ回帰推定量に関しては, 共線関係のみの場合 (0%) では最も高い回数を得ているが，最小 2 乗推定量と同様に 1% の汚染により大きく精度を落とし，5 ～ 10% の汚染でほとんど信頼区間に入らなくなることからも外れ値に対応できていないことがわかる.

ロバスト・リッジ推定量では，M 推定量を用いた場合，表 1 でわかるように目的変数 y の汚染については抵抗力があり頑健であるが, 表 2, 3 から見られるように説明変数の汚染に対しては弱い. 汚染率の増加に伴い，大幅に回数を落としており，5% のところで約 2 割となっている. S 推定量を用いたリッジ回帰推定量は, 汚染率が上がっても信頼区間に入る回数は減少せず, その頑健性の強さが見て取れる. また，MM 推定量を用いたリッジ回帰推定量は全体を通して最も安定しており, 表からもわかるように，汚染率 30% まで，汚染がない状態を除けばすべてで最高値を取っている. しかし，汚染率が約 30% を超えるあたりから S 推定量によるリッジ回帰推定量の回数がMM推定量によるものより高い回数となり、頑健性は強い. どのこのシミュレーション結果からの全体的なまとめとして, 通常のデータにおいては汚染率 30% というのはほとんど考えられないため，多重共線性や外れ値がデータに含まれていると考えられる場合には，MM 推定量を用いたリッジ回帰推定量が適切であり実用性があると考えられる.

表 1: 信頼区間に収まった回数 (汚染：目的変数)

汚染率	LS	RID	M	S	MM
1%	265	369	593	290	606
2%	147	210	607	296	608
3%	91	139	589	286	602
4%	50	95	569	301	597
5%	35	59	560	295	592
10%	5	9	495	307	567
20%	0	3	378	337	516
30%	0	0	246	366	393
40%	0	1	135	400	264

表 2: 信頼区間に収まった回数 (汚染：説明変数)

汚染率	LS	RID	M	S	MM
0%	489	626	613	283	610
1%	183	226	498	242	592
2%	77	90	398	240	565
3%	31	37	323	239	563
4%	7	13	257	231	580
5%	9	10	191	242	553
10%	0	0	31	233	498
20%	0	0	0	271	410
30%	0	0	0	287	287
40%	0	0	0	300	108

7　シミュレーション 2

前節よりも複雑な多重共線性と外れ値を持つデータに対してシミュレーションを行ない, ロバスト・リッジ回帰推定量 $\hat{\boldsymbol{\beta}}^{M}(k)$, $\hat{\boldsymbol{\beta}}^{LMS}(k)$, $\hat{\boldsymbol{\beta}}^{LTS}(k)$, $\hat{\boldsymbol{\beta}}^{S}(k)$, $\hat{\boldsymbol{\beta}}^{\tau}(k)$ の有効性を明らかにする. 多重共線性の存在する人工データの作成は金・田中 (1993) に従う.

$\boldsymbol{X}$ を階数 r の $n \times p$ 行列とする. このとき $\boldsymbol{X}$ は $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ と特異値分解される. ここで $\boldsymbol{D}$ は正の対角要素を持つ $r \times r$ の対角行列, $\boldsymbol{U}$ は $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$ であるような $n \times r$ 行列, $\boldsymbol{V}$ は $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$ であるような $p \times r$ 行列, $\boldsymbol{I}$ は $r \times r$ 単位行列である. この特異値分解の式の行列 $\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V}$ を実際に求めてその積で多重共線性を持つデータ $\boldsymbol{X}$ を作るわけである.

表 3: 信頼区間に収まった回数（汚染：目的・説明変数)

汚染率	LS	RID	M	S	MM
1%	135	167	527	293	595
2%	32	50	419	279	572
3%	20	27	370	288	576
4%	8	8	284	281	563
5%	2	6	259	272	551
10%	0	0	72	252	489
20%	0	0	7	283	442
30%	0	0	0	289	319
40%	0	0	0	302	178

7.1 多重共線性データ作成法（金・田中, 1993)

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・シュミッド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 $\lambda_i = c/(\kappa_j \sum_{i=1}^p \kappa_j^{-1})$ 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 $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{G}\boldsymbol{Q}'$ 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.

$$\boldsymbol{z}_i = \boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{Q}'(\boldsymbol{y}_i - \bar{\boldsymbol{y}}), \quad i = 1, 2, \cdots, n \tag{19}$$

   (5) 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:
   行列 $\boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U}$ を用いて $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ とする.

7.2 シミュレーションの仮定

1. 標本数 $n = 50$ とし, 5 つの説明変数 x_1, x_2, x_3, x_4, x_5 を用いる.
2. x_1, x_2, x_3, x_4 は金・田中（2003) の方法により作成され, 複雑な多重共線性を持っている.
3. 重回帰モデル：

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + \varepsilon, \quad \varepsilon \sim N(0, 1) \tag{20}$$

4. x_5 の汚染： x_1, x_2, x_3, x_4 の多重共線性を維持したまま, x_5 に外れ値を入れる.

$$x_5 \sim (1 - \eta)N(0, 1) + \eta N(0, 9) \tag{21}$$

5. y の汚染： y に外れ値を入れる.

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + \tilde{\varepsilon}, \quad \tilde{\varepsilon} \sim N(0, 9) \tag{22}$$

6. $\eta = 0.2$ とする.

7.3 y 方向のみに外れ値が存在する場合

それぞれのロバスト推定量に基づくリッジ回帰推定量による係数推定値と k の値 (k の決定方法は, それぞれのリッジトレースから視覚的に判断したものである) を表 4 に, S 推定量と τ 推定量に基づくリッジ回帰推定量によるリッジ・トレースを図 1 と図 2 に示す.

表 4: 係数推定値: y 方向の外れ値

推定量	k	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
M	0.0010	0.408	0.147	0.332	1.101	0.234
LMS	0.0010	0.271	0.159	0.300	0.718	-0.094
LTS	0.0010	0.159	0.039	0.283	0.674	0.024
S	0.0005	0.735	0.312	0.209	0.662	0.155
τ	0.0008	0.916	0.744	0.249	0.683	0.274

表 4 より, 全ての係数推定値が概ね 1 に近い値を取っており, うまく推定できている. このことから, y 方向の外れ値に対しては全てのロバスト推定量に基づくリッジ回帰推定量が有効であるといえる.

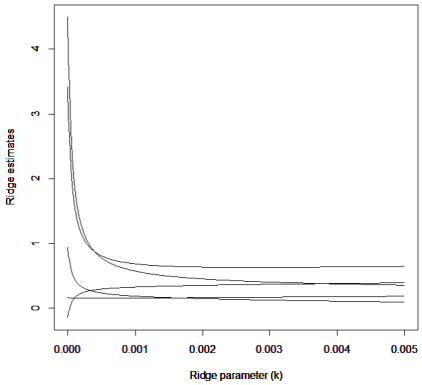

図 1: S・リッジ

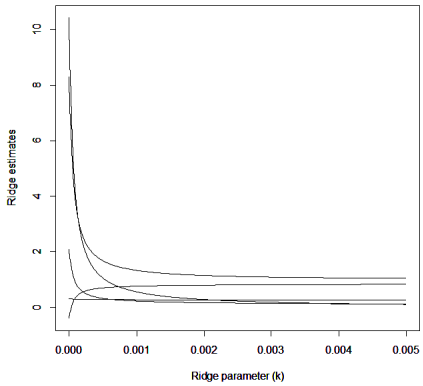

図 2: τ・リッジ

7.4 x 方向のみの外れ値が存在する場合

それぞれのロバスト推定量に基づくロバスト・リッジ回帰推定量による係数推定値と k の値を表 5, S 推定量と τ 推定量に基づくリッジ推定量によるリッジ・トレースを図 3 と図 4 に示す.

表 5: 係数推定値: x 方向の外れ値

推定量	k	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
M	0.0010	-0.198	-0.652	0.233	0.781	0.004
LMS	0.0010	1.009	0.035	0.263	0.780	0.548
LTS	0.0010	1.076	0.076	0.003	0.775	0.578
S	0.0015	1.193	0.379	0.145	0.912	0.489
τ	0.0008	0.708	0.362	0.336	1.500	0.526

表 5 より, M 推定量に基づくリッジ回帰推定量の値は $\hat{\beta}_4$ 以外, 真の係数値からずれてしまっていることがわかる. これは M 推定量が x 方向の外れ値に影響されたためと考えられる. これは M 推定量に基づくリッジ回帰推定量が M 推定量の特性を引き継ぎ, x 方向の外れ値に対しては有効でないことを示している.

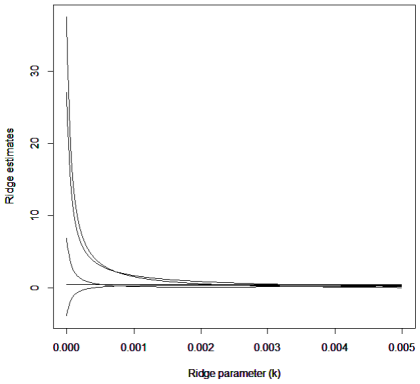

図 3: S・リッジ

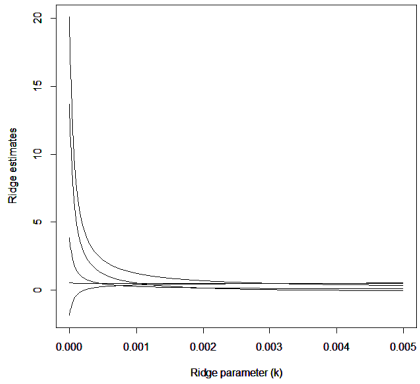

図 4: τ・リッジ

7.5 x, y 方向両方に外れ値が存在する場合

それぞれのロバスト推定量に基づくリッジ回帰推定量による結果を表 6, S 推定量と τ 推定量に基づくリッジ回帰推定量によるリッジ・トレースを図 5 と図 6 に示す.
表 6 より, M 推定量に基づくリッジ回帰推定値 $\hat{\beta}_2$, $\hat{\beta}_5$ が負の値を取り, 真の係数推定値から大きくずれてしまっていることがわかる. これは M 推定量が x 方向の外れ値に影響されたためと考えられる. また τ 推定量を除くすべての推定量の x_5 の係数推定値 $\hat{\beta}_5$ が負の値となっており 真の係数から離れている. そして, τ 推定量は x_3 の係数推定値以外のすべてで最も良い推定値を与えている.

表 6: 係数推定値: x, y 両方向の外れ値

推定量	k	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
M	0.0006	0.141	-0.383	0.223	1.956	-2.304
LMS	0.0010	0.327	-0.104	0.285	0.760	-0.411
LTS	0.0010	0.365	-0.081	0.258	0.694	-0.390
S	0.0007	0.742	0.232	0.235	0.804	-0.367
τ	0.0005	1.043	0.288	0.263	1.009	1.187

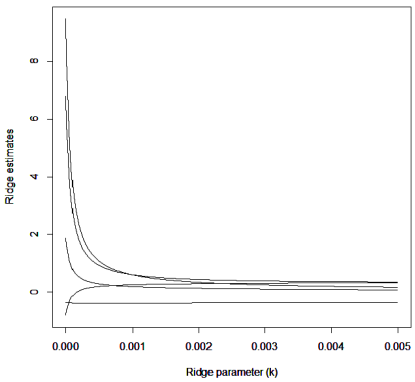

図 5: S・リッジ

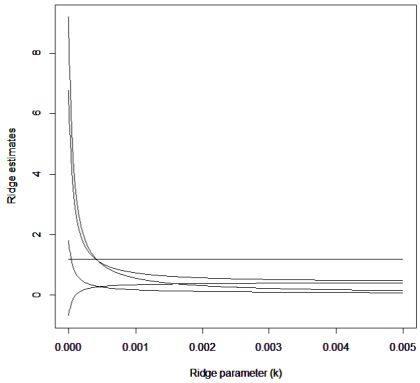

図 6: τ・リッジ

8　おわりに

本論文では, 多重共線性だけでなく外れ値も含むデータに対しては, 従来の最小 2 乗推定量に基づく通常のリッジ回帰推定量ではうまく対処できないこと, そして この場合にはロバスト推定量に基づくリッジ回帰推定量が有効に機能することをシミュレーションにより示した. リッジ回帰のためのロバスト推定量としては, これまでにすでに用いられている M, LMS, LTS に加えて新たに S, MM, τ 推定量を用いた. シミュレーションの結果によると, これらのロバスト・リッジ推定量のうちでは MM 推定量と τ 推定量に基づくリッジ回帰推定量が優れている. また, ロバスト・リッジ回帰推定量にはそれに用いたロバスト推定量の性質が強く反映することも確認した. この意味においても, M 推定量に基づくリッジ回帰推定量は多重共線性と y の外れ値にはうまく機能するが, 説明変数の外れ値には対処できないことに注意すべきである. MM 推定量は R で利用できるが, τ 推定量はまだ R に実装されていない. 室梅秀平氏（南山大学数理情報研究科）には, 本論文の τ に関する部分の計算等に必要な自作プログラムを提供していただき感謝いたします.

参考文献

[1] Beaton, A. E. and Tukey, J. W (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Tecnometrics*, **16**, 147-185.

[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. (1970a). 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] 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.

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

[16] 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.