回帰における多重共線性と外れ値に関する研究

M2011MM044 暮石一樹
指導教員：木村美善

1　はじめに

線形回帰分析において推定量の信頼性は非常に重要であり，代表的な最小二乗法は正規分布が仮定されるなど理想のモデルの下では精度の高い推定量を得ることが可能である．しかし，以下の 2 つの状況においては，推定量の信頼性は崩れてしまう．
1. 外れ値の存在やモデル分布からのずれが存在する場合．
2. 説明変数間に強い線形関係が存在する場合である．
そして，それぞれの問題の解決法としてロバスト回帰分析法とリッジ回帰分析法がある．本研究の目的は，この 2 つの問題点を同時に解決するためのロバスト・リッジ回帰分析法について考察するとともに，シミュレーションを通し，その有効性や特徴を明らかにしていくことである．

2　回帰分析

2.1　モデルの定式化

目的変数 y_i からなる $n \times 1$ ベクトルを $\mathbf{y}$, 説明変数 x_{i0} , $x_{i1}, \cdots, x_{ip}$ からなる $n \times (p+1)$ 行列を X，回帰係数 β_0，β_1，$\cdots$，β_p からなる $(p+1) \times 1$ ベクトルを β，誤差項 ϵ_i からなる $n \times 1$ ベクトルを ϵ とする時，回帰モデルは

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \qquad \mathbf{E}(\epsilon_\mathbf{i}) = \mathbf{0}, \qquad \mathbf{V}(\epsilon_\mathbf{i}) = \sigma^\mathbf{2}\mathbf{I} \quad (1)$$

と表すことができる．最小 2 乗法 (Ordinary Least Square) とは残差平方和

$$\|\epsilon\|^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \quad (2)$$

を最小にするような β を求める方法であり，得られる最小二乗推定量 (OLS 推定量) は

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

で与えられる．

2.2　平均 2 乗誤差

平均 2 乗誤差 (Mean Square Error) とは推定量が真の母数に平均的にどれだけ近いかを表すものである．$n \times p$ の説明変数行列 X，$n \times 1$ の目的変数ベクトル $\mathbf{y}$ が与えられた時，$X'X$ の固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \lambda_{min} > 0$ とする．その時，偏回帰係数 β の OLS 推定量を $\hat{\beta}$ とすると MSE は以下のようになる．

$$MSE[\hat{\beta}] = E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] \quad (4)$$

$$= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} \quad (5)$$

ただし，σ^2 は誤差分散とする．

3　多重共線性

3.1　多重共線性

重回帰分析において，そこに含まれる説明変数間に強い線形関係が存在する場合，OLS を用いた係数推定値はデータの少しの変化や変数選択に対して敏感に反応してしまうため不安定になり，信頼性が失われて回帰モデルからそれぞれの変数が及ぼす影響を予測することが困難になる．これは，変数間の線形関係が強い場合，$|\ X'X\ | \approx 0$ となってしまい，正規方程式の解が安定しないことが原因である．

3.2　多重共線性の検出法

一つ目の検出法として分散拡大要因 (Variance Inflation Factor) がある．これは OLS 推定量と真の値との誤差に対して，多重共線性が及ぼす影響を調べるものである．第 i 番目の説明変数の係数に対する VIF は以下のようになる．

$$VIF(i) = (\frac{1}{1 - R_i^2}) \quad (6)$$

ただし，R_i^2 は第 i 番目の説明変数を式中の他の説明変数に回帰した時の重相関係数の 2 乗値である．この VIF の値が 10 を超えるものは他の変数との共線関係があるとされ，多重共線性が示唆される．
二つ目の方法として，固有値による検出法がある．説明変数間の相関行列を $\mathrm{V}(x)$，その固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ とする時，0 に極めて近い λ_i があるならば多重共線性が示唆される．例として，式 (5) の場合においては λ が極めて 0 に近い値を取る場合，MSE が大きく発散してしまうことが分かる．

4　リッジ回帰

4.1　リッジ回帰

リッジ回帰 (Ordinary Ridge Regression) は OLS 推定量よりも小さな MSE をもつ推定量を得るための方法である．リッジ回帰推定量 (ORR 推定量) はリッジ・パラメータと呼ばれる定数 $k \geq 0$ を取り入れることで OLS 推定量の不安定さを解決することができ，

$$\hat{\beta_k} = (X'X + kI)^{-1}X'\mathbf{y} \quad (7)$$

により定義される．$k = 0$ の時は，$\hat{\beta_k}$ は OLS 推定量に等しく，ORR 推定量の MSE は，

$$\begin{aligned} MSE[\hat{\beta_k}] &= E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2\beta'(X'X + kI)^{-2}\beta \\ &= \gamma_1(k) + \gamma_2(k) \end{aligned} \quad (8)$$

となる．
また，$\gamma_1(k)$ は分散の総和であり，単調減少し，$\gamma_2(k)$ はバイアスを示し，単調増加していく．

4.2 一般化リッジ回帰

一般化リッジ回帰 (Generalized Ridge Regression) とはORR を一般化したものであり，GRR 推定量がある．これを用いることで最適なパラメータ k の値を計算により求めることができる．Λ=diag(λ_1，λ_2，$\cdots$，λ_p) は $Z'Z$ の固有値からなる行列である．$P'X'XP$ =Λ となるような直交行列 P が存在すると仮定すると，線形回帰モデルは

$$\mathbf{y} = \mathbf{Z}\alpha + \epsilon \tag{9}$$

となる．ただし，$Z = XP$，α=$P'\beta$ とする．これにより一般化推定量を得ることができる．

$$\hat{\alpha}_k = (\Lambda + kI)^{-1}Z'Y \tag{10}$$

4.3 パラメータ k の決定方法

パラメータ k の決定方法は多く存在するため，ここでは最も実用的と言われているリッジ・トレースと理論による方法をいくつかあげる．リッジ・トレースとは縦軸に標準化した係数推定値，横軸にパラメータ k をとり，その軌跡をプロットし，トレースが安定した位置の k の値を採用するというものである．一方，理論による方法では代表的な方法として，Hoerl and Kennard(1980) と Hoerl,Kennard and Baldwin(1975) があり，以下に挙げる．

$$\hat{k}_{HK} = (\frac{p\hat{\sigma}^2}{\hat{\alpha}^2_{max}}) \tag{11}$$

$$\hat{k}_{HKB} = (\frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}) \tag{12}$$

5 ロバスト推定量

ロバストとは頑健という意味であり，一定の仮定を基に得られる統計的手法が，仮定の一部が保障されない場合において，どれほどの影響を受けるのかを研究するにあたり，Box(1953) により提案された概念である．最も代表的な例は最小二乗推定量であり，標準的な仮定からのずれや外れ値が存在する場合は大きな影響を受けてしまう．ロバスト推定量はそのような状況下でも影響を軽減し，より正確な推定を行うことができる．また，Huber(1981) では，望ましい統計量の性質として 3 つ挙げている．
1. 仮定された分布の下で高い効率を持つ．
2. 仮定された分布から幾分のずれが生じた場合においても，高い効率を維持することができる．
3. 仮定された分布から大きくずれた分布の下でも大きな効率の減少を防ぐことができ，破綻することがない．
Huber はこのような性質をもつ統計量をロバスト推定量と定義している．そして，多くの文献でロバストネスを測るための尺度が提案されており，その代表的な例である破綻点について以下で述べていく．

5.1 有限標本破綻点

大域的なロバストネスを測るための尺度として破綻点がある．Z を n 個のオリジナルデータとする．

$$Z = \{(x_1, y_1), \cdots, (x_n, y_n)\} \tag{13}$$

とし，Z から得られる β の推定量を $\hat{\beta}(Z)$ とする．次に汚染されたデータ Z' を得るためにオリジナルデータから m 個の任意の値（外れ値を含ませるため，値の範囲は制限しないものとする）を置き換え，あらゆる可能な汚染データ Z' を考える．汚染によって生じるバイアスの最大バイアスは

$$MB(m; \hat{\beta}, Z) = \sup_{Z'} \parallel \hat{\beta}(Z') - \hat{\beta}(Z) \parallel \tag{14}$$

と定義され，$MB(m; \hat{\beta}, Z)$ が無限になる場合は，置き換えられた m 個の値が $\hat{\beta}$ に大きな影響を与えていることを意味する．これを推定量の破綻 (Breakdown) という．そして，標本 Z における推定量 β の有限破綻点は以下のように定義される．

$$\epsilon^*_n(\beta, Z) = \min\{\frac{m}{n}; MB(m, \hat{\beta}, Z) = \infty\} \tag{15}$$

ただし，ϵ^* は $0 \leq \epsilon *_n \leq 1/2$ であり，高い破綻点が望ましい．最小二乗推定量の場合では，

$$\epsilon^*_n(\hat{\beta}, Z) = \frac{1}{n} \tag{16}$$

となり，たった一つの外れ値に対しても推定量 β は大きく影響を受ける．また，(16) 式から標本数 n が増加するにつれ，0 に収束していくため，最小二乗推定量の漸近的破綻点は 0 となる．
他にもロバストネスを測るための手段として影響関数や漸近効率などが存在する．

5.2 M 推定量

M 推定量は Huber(1964) により提唱された最も一般的なものであり，

$$\hat{\beta}^M = arg,\ \min_{\beta} \sum_{i=1}^{n} \rho(r_i) \qquad r_i = y_i - x_i'\beta \tag{17}$$

により求めることができる．ρ は微分可能な偶関数であり，さまざまな提案がされている．
・Huber

$$\rho_H(r) = \begin{cases} \frac{1}{2}r^2, & \mid r \mid \leq a \\ a \mid r \mid -\frac{1}{2}a^2, & \mid r \mid > a \end{cases} \tag{18}$$

・Bisquare

$$\rho_B(r) = \begin{cases} \frac{a^2}{6}\{1 - [1 - (\frac{r}{a})^2]^3\}, & \mid r \mid \leq a \\ \frac{a^2}{6}, & \mid r \mid > a \end{cases} \tag{19}$$

この他にも Andrews，Tukey により提案されたものが存在する．M 推定量は y 方向に対してはロバストであるが，x 方向に対してはロバストではない．また，$\rho(r_i) = r_i^2$ とすることで，最小二乗推定量を得ることができる．

5.3 LMS 推定量

残差平方の中央値を最小にする推定量

$$\hat{\beta}^{LMS} = arg \ \min_{\beta} med(r_i^2) \tag{20}$$

を LMS(Least Median of Squares) 推定量といい，Rousseeuw(1984) により定義された．有限標本破綻点は $([\frac{n}{2}]-p+2)/n$ であり，$n \to \infty$ のとき，50% となる．しかし，効率が悪く (収束速度が $n^{-1/3}$)，漸近的に正規分布に従わない．

5.4 LTS 推定量

LMS(Least Trimmed Squares) 推定量の収束の遅さを解決するために Rousseeuw(1984) によって提案され，LTS 推定量は

$$\hat{\beta}^{LTS} = arg \ \min_{\beta} \sum_{i=1}^{h} (r_{(i)}^2) \tag{21}$$

により定義される．この推定量は残差平方を昇順に並び替えた順序統計量の h 番目までの和を最小にする推定量である．$h = n$ のとき，最小二乗推定量と等しい．大きな残差を加えないことで外れ値を避けている．有限標本破綻点は $([\frac{n-p}{2}]+1)/n$ であり，h が $\frac{n}{2}$ に近い位置で $\frac{1}{2}$ となる．

5.5 S 推定量

残差のばらつきを最小にすることで高い効率，高い破綻点を得ること目的として，Rousseuw and Yohai(1984) により提案された S 推定量は

$$\hat{\beta}^{S} = arg \ \min_{\beta} s(r_i) \tag{22}$$

により定義される．ここでの s は尺度推定量であり，

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\frac{r_i}{s}) = b \tag{23}$$

を満たすものである．ただし，ρ は次の条件を満たしている関数であると仮定する．(1) ρ は対称，連続微分可能である．(2) ρ は $[0, c]$ で単調増加であり，$[c, \infty]$ で一定となる $c > 0$ が存在する．また，調整定数 c に関しては小さく設定すれば高い破綻点を得ることができるが，漸近効率が低下してしまうため，S 推定量では破綻点と漸近効率にトレードオフの関係がある．

5.6 MM 推定量

Yohai(1987) により提案され，M 推定量では y 軸方向のみロバストであった性質を改善し，MM 推定量では x 軸に対してもロバストの性質を持つことが可能になった．その方法として，二つの異なった ρ_0,ρ_1 を使用する．ρ_0 により，高い破綻点，ρ_1 により高い効率を得るためである．初期推定量 $\hat{\beta}_0$ は S 推定量を用いることが多い．そして，残差 $r_i(\hat{\beta}_0)$ のロバスト尺度 $\hat{\sigma}$ を計算する．ここでは M 推定量が用いられることが多く，その際に高い効率を持てるように調整関数を設定する．

$$\hat{\beta}^{MM} = arg \ \min_{\beta} \sum_{i}^{n} \rho_1(\frac{r_i}{\hat{\sigma}}) \tag{24}$$

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0(\frac{r_i}{\hat{\sigma}}) = 0.5 \tag{25}$$

その後は反復再加重法 (IRWLS) を用いることで得ることができる．このように，第一段階で高い破綻点，第二段階で高い効率を保障するようになっている．

5.7 ロバスト・リッジ回帰

上記で述べた通常のリッジ回帰では多重共線性による影響はリッジ・パラメータ k を加えることで軽減できるものの，外れ値や分布のズレに対しては最小二乗法と同様に大きな影響を受けてしまう．そのため，2 つの問題 (多重共線性と外れ値) を同時に解決するための方法として，Silvapulle(1991) は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 (Robust Ridge Regresion) を提案した．これは従来のリッジ推定量

$$\hat{\beta_k} = (X'X + kI)^{-1} X'\mathbf{y} \tag{26}$$

が最小二乗推定量を $\hat{\beta}$ を用いて，

$$\hat{\beta}_k = (X'X + kI)^{-1} XX\hat{\beta} \tag{27}$$

と書くことができるでの，最小二乗推定量の $\hat{\beta}$ をロバスト推定量 $\hat{\beta}^M$ に置き換えることで

$$\hat{\beta}_k = (X'X + kI)^{-1} XX\hat{\beta}^M \tag{28}$$

として得られる．これにより，従来のリッジ回帰では対応することができなかった外れ値や分布のズレに対しても性能を損なうことなく分析することが可能になる．そして，Silvapulle(1991) では (28) 式を一般化した

$$\hat{\alpha}_k^M = (\Lambda + kI)^{-1} C'C\hat{\alpha}^M \quad \hat{\alpha}^M = P\hat{\beta}^M \tag{29}$$

も提案しており，これをリッジ型M推定量 (Ridge-type M-estimatior) と呼んでいる．ただし，$C = XP$ とする．

6 シュミレーション

外れ値と多重共線性の両方を持たせるために，誤差項と説明変数を汚染させたデータを用いて，ロバストな M, MM, S 推定量を使用したロバスト・リッジ推定量の有効性や精度について検証していく．なお，シミュレーションの計算にはオープン・ソースの統計解析ソフト R を使用する．

6.1 データ

作成するデータは標本数が 100 個，3 つの説明変数 X_1 , X_2, X_3 を正規乱数により生成し，多重共線性を発生させるため，説明変数 X_1, X_2 に強い相関関係を持たせた．目的変数 $\mathbf{y}$ は $\mathbf{y} = \mathbf{X_1} + \mathbf{X_2} + \mathbf{X_3} + \epsilon$ とし，$\epsilon \sim N(0, 1)$

とする．ここで，ある定数 η の割合で正規分布にコーシー分布を加えることで，混合分布を作成していく．目的変数を汚染する場合は $\epsilon \sim (1-\eta)\cdot N(0,0.1)+\eta\cdot t(0)$ とし，説明変数を汚染する場合は，目的変数を先に作成した後に，$X_{3i} \sim (1-\eta)\cdot N(0,0.1)+\eta\cdot t(0)$ と置き換えていく．また，リッジ回帰分析を行う際に必要となるパラメータ k に関しては，今回のシミュレーションでは固定せず，範囲を $(0,1)$ とし $k=0.0001$ ごとに増加させながら MSE をベースとしながら，最小の MSE の値をとったときのパラメータ k を使用するとした．

$$MSE=\frac{1}{p}(\hat{\beta}-1)^2 \tag{30}$$

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内に回帰係数の 95% 信頼区間にそれぞれの推定値が何回入るかによって判断していく．信頼区間は共線関係も汚染もない状態で求めたものとし，すべての係数が信頼区間に入った場合のみカウントするとする．

7　考察

共線関係のみのデータの時点で最小 2 乗推定量はすでに 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ることができたが，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てしまい，5% の汚染までで，ほとんど信頼区間に入らないことからも，外れ値に対応できない事が分かる．一方のロバスト・リッジ推定量では，M 推定量を用いた場合，汚染率の増加に従い，大幅に数を落としており，5% の地点で約 2 割となっていることからも，説明変数への外れ値に対応できていないことが分かる．一方，S 推定量は信頼区間に入る数は徐々に増えていく傾向にあり，汚染率に基づく最大バイアスの推移を調べたところ，約 40% のあたりまでほとんど真の係数値に近い値を取っていたことを考えると，その頑健性の強さを示すことができた．また，MM 推定量に関しては，全体を通して最も安定した結果を出せた推定量であり，表からも分かるように，汚染率 20% まで，汚染がない状態を除けばすべてで一番の値を取っている．しかし，最大バイアスの推移を見てみると，汚染率が約 30% の地点から推定量に影響が出始めていたため，S 推定量よりも頑健性の観点においては劣っていた．
全体的な総括として，通常のデータにおいて，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，ロバスト MM 推定量を用いたロバスト・リッジ推定量が最適で，実用性があると考えられる．

8　おわりに

シミュレーションでは目的変数，説明変数，両方の変数への汚染を行い，それぞれの推定量の有効性や精度を確かめることができた．しかし，本研究ではロバスト M，MM，S 推定量のみの使用だったため，他のロバスト推定量である τ や最深回推定量なども同時に検証することができれば，ロバスト・リッジ推定量のさらなる有用性を示すことができたと思う．また，ロバスト・リッジ推定量を求めるにあたり，必要となるリッジパラメータ k や MSE を計算によって導出することができれば，ロバスト・リッジ推定量のさらなる実用性が示すことができるため，その点まで研究を進められなかったことは残念である．ただ，日本ではロバスト手法の研究はあまりされておらず，日本語による文献はとても少ない．そのため，本研究が多少なりとも貢献することができれば幸いに思う．

表 1　信頼区間に収まった回数 (説明変数)

汚染率	LS	RIG	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

表 2　信頼区間に収まった回数 (目的・説明変数)

汚染率	LS	RIG	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

参考文献

[1] Askin, R. G. and Montgomery, D. C. (1980). Augmented Robust Estimators, Technometrics 22

[2] 金子元紀. (2007). 線形回帰におけるロバスト推定量の研究，南山大学大学院数理情報研究科修士論文

[3] Maronna, R. A. and Martin, R. D. and Yohai, V. J. (2006). Robust Statistics Theory and Methods, John Wiley & Sons, Ltd

[4] Silvapulle, M. J. (1975). Robust Ridge Regression based on an M-estimator, Austral. J, Statist, 33(3), 319-333

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

[6] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estiomates for regression. The Annals of Statictics 15, 642-656