回帰分析の理論とその応用
– リッジ回帰を中心に –

2002MM084 澤田 謹志　　2002MM097 武山 嵩弘
指導教員　木村 美善

1　はじめに

通常, 回帰モデルにおいて, 説明変数は互いに独立であることが仮定されている. しかし, 説明変数間に強い線形関係が存在すれば多重共線性の問題が生じ, 回帰分析の結果は不安定なものとなってしまう. この問題を解決するためには様々な方法があるが, その中でも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ひとつである.

本論文の目的は, リッジ回帰の理論を理解し, 最小２乗推定量, ロバスト推定量, リッジ回帰推定量の比較・考察を行うことである. また, 新しい試みとして, データに多重共線性の問題と外れ値が混在するような場合についても分析できるロバスト・リッジ回帰に, 種々のロバスト推定量を適用して分析を行った.

モデルを簡略化するため, 第 3 節以降の式中における変数は $n \times 1$ のベクトルと $n \times p$ の行列で表記する.

2　回帰モデル

2.1　モデルの定式化

n 個の観測値が与えられた場合, 目的変数を y_i, 説明変数を x_{ki} とすると, 回帰式は

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, \cdots, n \tag{1}$$

と表される. ただし,$\beta_0, \beta_1, \cdots, \beta_k$ は回帰係数,ε_i は誤差項を示す.

2.2　回帰モデルの種類

(1) 式のモデルにおいて, 説明変数が唯一つである場合（$k = 1$ のとき）を「単回帰モデル」, 複数ある場合（$k \geq 2$ のとき）を「重回帰モデル」と呼ぶ.

また, 次の仮定を考える.

1. $E(\varepsilon_i) = 0$（不偏性）
2. $V(\varepsilon_i) = \sigma^2$（等分散性）
3. $cov(\varepsilon_i, \varepsilon_j) \neq 0$（無相関性）
4. $\varepsilon \sim N(0, \sigma^2 I)$（正規性）

上に示した４つの仮定のうち,1〜3 を満たす回帰モデルを「線形回帰モデル」, 1〜4 全ての仮定を満たすモデルを「線形正規回帰モデル」と呼ぶ. また, 線形回帰モデル以外のモデルを総じて「非線形回帰モデル」と呼ぶ.

3　最小２乗法

3.1　残差２乗和 (SSE)

線形回帰モデル $Y = X\beta + \varepsilon$ において, 実測値 y_i と予測値 $\hat{y}_i$ の差を残差といい, その値を e とすると, このノルムの平方は次のように表される.

$$\begin{aligned} \|e\|^2 &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - 2\beta' X'Y + \beta'(X'X)\beta \end{aligned} \tag{2}$$

これを残差２乗和 (SSE) と呼ぶ.

3.2　最小２乗 (OLS) 推定量

最小２乗法とは SSE の値を最小にする考え方である.SSE は β の 2 次関数になっているので, これをを最小にするため,(2) 式を β で偏微分したものを 0 とすることにより,OLS 推定量は次のように得られる.

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

この推定量は最も基本的で, かつ最も広く用いられている.

3.3　ガウス・マルコフの定理

(1) 式の線形回帰モデルにおける β_i の任意の線形不偏推定量を $\hat{b}_i$ とすると

$$V(\hat{\beta}_i) \leq V(\hat{b}_i) \tag{4}$$

が成り立つ. さらに $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$,$\hat{b}' = (\hat{b}_0, \hat{b}_1, \cdots, \hat{b}_p)$ より, これを一般化すれば

$$V(\hat{\beta}) \leq V(\hat{b}) \tag{5}$$

となる. ここで,(5) 式は $V(\hat{b}) - V(\hat{\beta})$ が非負値定符号行列であることを表し,OLS 推定量の分散は他のどの線形不偏推定量の分散よりも小さい, すなわち最良線形不偏推定量 (BLUE) であることが示される. これは, ガウス・マルコフの定理として知られている.

4　多重共線性

4.1　多重共線性とは

回帰分析において, 説明変数は直交していないのが普通だが, その非直交性が極端でなければ, 分析を行う上で大きな問題にはならない. しかし, 非直交性が極端であり, データの説明変数間に強い相関関係があるときにそのまま回帰分析を行ってしまうと, その結果は不明確で意味をなさないものとなってしまうことがある.

特に, 変数どうしの相関が ± 1 の場合には,OLS における正規方程式の解そのものを求めることさえできない. し

かし, 相関が完全に ±1 でなくても ±1 に近づけば近づくほど $|X'X|$ が 0 に近くなるため, やはり正規方程式の解は非常に不安定なものになってしまう. このような事例を「多重共線性の問題」という.

4.2 多重共線性の検出

多重共線性を検出する方法のひとつに, 分散拡大要因 (VIF) を用いたものがある. 第 i 番目の説明変数の係数に対する VIF は

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

で計算される. ここで,R_i^2 は第 i 番目の説明変数を他の説明変数に回帰したときの重相関係数の 2 乗値を表す.$R_i^2 = 0$ のとき $VIF(i) = 1$ となり R_i^2 の値が 1 に近づくにつれ,$VIF(i)$ の値は発散していく. ゆえに, 全ての説明変数に対して VIF を調べることによって, どの変数が多重共線性の原因となっているのかを特定することができる. 一般的な基準として,VIF が 10 を超えるような変数は他の変数と共線関係にあるとされている.

また,$(VIF(i))^{-1}$ を「トレランス」と呼ぶことがある. トレランスは VIF の逆数になっているため, その値が 0.1 以下となるときに多重共線性が示唆される. (文献 [1],[5] 参照)

5 リッジ回帰

5.1 平均２乗誤差 (MSE)

説明変数間の分散共分散行列を $V(x)$ とし, その固有値を $\lambda_1 \geq \lambda_2 \geq \cdots\cdots \geq \lambda_p$ とする. 偏回帰係数 β の OLS 推定量を $\hat{\beta}$ とすると

$$\begin{aligned} E(\|\hat{\beta} - \beta\|^2) &= E(\|(X'X)^{-1}X'(Y - \eta)\|^2) \\ &= \sigma^2 \cdot \mathrm{tr}V(x)^{-1} \\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} \end{aligned} \tag{7}$$

が成り立つ. この値を平均２乗誤差 (MSE) という. (ただし,$E(Y) = \eta$ であり,σ^2 は誤差分散である)

多重共線性が存在するとき,$V(x)$ の固有値には極めて 0 に近いものがあり, (7) 式の MSE は非常に大きなものになってしまう恐れがある.

5.2 リッジ回帰 (ORR) 推定量

文献 [3] の Hoerl and Kennard(1970) は, 平均２乗誤差を小さくするような推定量を求めるために, モデルにリッジ・パラメータと呼ばれる正定数 k を取り入れることにより回帰係数の安定性を高める手法である「リッジ回帰」(Ordinary Ridge Regression) を提案した. (この手法については, 文献 [2] に詳しい) その推定量は以下のように表される.

$$\hat{\beta}_k = (X'X + kI)^{-1}X'Y \tag{8}$$

特に $k = 0$ のとき,$\hat{\beta}_k$ は OLS 推定量に等しい.

しかし, 回帰式からも分かるように,ORR 推定量は不偏推定量ではない. したがって k が増加していくにつれてバイアス (偏り) もまた大きくなり,SSE の値も増加していく. それでも多重共線性のあるデータでは, ほとんどの場合において SSE の増加量よりも MSE の減少量の方が大きくなるため,OLS よりも ORR を用いた方が良い推定量を得られる可能性が高いと言える.

図 1 に, この例として, 多重共線性の問題があることで有名な Longley データ (1947 年から 1962 年までの 16 年間に観測された, 米国経済に関するデータ) を分析した際の SSE,MSE を示す.

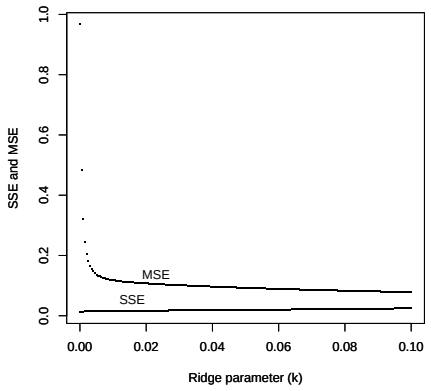

図 1 パラメータ k に対する SSE,MSE ($k \in [0, 0.1]$) : Longley データ

5.3 ORR 推定量の分散とバイアス

ORR における MSE の値は次のようになる.

$$\begin{aligned} E(\|\hat{\beta}_k - \beta\|^2) &= E[(\hat{\beta}_k - \beta)'(\hat{\beta}_k - \beta)] \\ &= \sigma^2 \sum_{1}^{p} \lambda_i/(\lambda_i + k)^2 + k^2\beta'(X'X + kI)^{-2}\beta \\ &= \gamma_1(k) + \gamma_2(k) \end{aligned} \tag{9}$$

ここで, 分割した 2 番目の要素 $\gamma_2(k)$ は,$Z\beta$ から β の差の平方であり, $k = 0$ のときには $Z = I$ となるから,$\gamma_2(k) = 0$ となる. このように $\gamma_2(k)$ はバイアスの 2 乗を示しており, パラメータ k の値について単調増加する. また,1 番目の要素 $\lambda_1(k)$ は分散の総和を示しており, パラメータ k の値について単調減少する.

6 一般化リッジ回帰

6.1 一般化リッジ回帰 (GRR) 推定量

(8) 式の ORR 推定量を一般化したものとして,GRR 推定量がある.$X'X$ の固有値を対角要素として持つような行列を $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ とする. このとき $P'X'XP = \Lambda$ となるような直交行列 P が存在すると仮定すると, $PP' = I$ より, 線形回帰モデルは

$$\begin{aligned} Y &= XPP'\beta + \varepsilon \\ &= Z\alpha + \varepsilon \end{aligned} \tag{10}$$

と表される. ただし,$Z = XP$,$\alpha = P'\beta$ である. ここでさらに $K = diag(k_1, k_2, \ldots, k_p)$ とすると,(8) 式を一般化し

たものとして, 以下のような推定量が得られる.

$$\begin{aligned}\hat{\alpha}_k &= (Z'Z+K)^{-1}Z'Y \\ &= (P'X'XP+K)^{-1}Z'Y \\ &= (\Lambda+K)^{-1}Z'Y \end{aligned} \tag{11}$$

これが,GRR(Generalized Ridge Regression) 推定量である. この推定量を使って, 最適なパラメータ k の値を数式により求めることが可能になる.

6.2 パラメータ k の決定方法

最適なパラメータ k の値の決定方法は, リッジ回帰に関する研究論文で様々なものが提示されている.（これについては, 文献 [4] に詳しい）

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

$$\hat{k}_{LW} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\Lambda\hat{\beta}} \tag{13}$$

$$\hat{k}_{AM} = \frac{1}{p}\sum_{i=1}^{p}\frac{p\hat{\sigma}^2}{\hat{\alpha}_i} \tag{14}$$

$$\hat{k}_{MED} = Median\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_i}\right) \tag{15}$$

k の決定方法には, 上に示した方法以外にも様々なものがあるが, 実際の分析では, 視覚によりリッジ・トレースが安定状態に達したと認められる時点の k を選ぶことが多いようである.

6.3 リッジ回帰の利用法

リッジ回帰は通常, データに共線性があり OLS で分析を行うことが困難なときに, 安定した係数を求める際に用いられるが, パラメータ k の値の変化に対する推定量の変化を見ることにより, 多重共線性の存在を発見したりすることも可能である. また, 変数選択法による分析を行う際に, リッジ・トレースの様子から変数の選定を行うことも出来る.

7 ロバスト回帰

7.1 M 推定量

M 推定量は Huber(1964) によって提案された, ロバスト推定量の中でも最も一般的なものであり, 次のような関数を最小にする.

$$\sum_{i=1}^{n}\rho(e_i) = \sum_{i=1}^{n}\rho(y_i - x_i'b) \tag{16}$$

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,(16) 式からもわかるように,$\rho(e_i) = e_i^2$ とすると, これは OLS 推定量に等しい.

7.2 LMS 推定量

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

$$\mathrm{MED}(e_1^2, \cdots, e_n^2) \tag{17}$$

を最小にするように求められる.

7.3 LTS 推定量

LTS(Least Trimmed Squares) 推定量は,Rousseeuw (1984) によって提案された手法であり, 残差平方を小さいものから順に並び替えたときの m 番目までの和

$$\sum_{i=1}^{m} e_{(i)}^2 \tag{18}$$

を最小にするような回帰係数 b を決定する.

8 ロバスト・リッジ回帰

8.1 外れ値がある場合の ORR 推定量

従来のリッジ回帰は, 重回帰分析において多重共線性が生じた際でも, 変数選択を行わずに安定した係数を推定することが可能であった. しかし, データに外れ値がある場合にはその影響を強く受けてしまい, 分析が困難となってしまう.

8.2 ロバスト・リッジ回帰 (RRR) 推定量

この問題を解決するために, 文献 [6] の Silvapulle(1991) が提案した手法が「ロバスト・リッジ回帰」である. その推定量は以下のように表される.

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

(19) 式からわかるように,RRR 推定量は ORR 推定量における $\hat{\beta}$ をロバスト推定量 $\hat{\beta}^{rob}$ に置き換えたものであり, 仮定された分布から「ズレ」が生じた場合にも, それほど性能を損なわずに分析することができる.

9 実行例

9.1 多変量データ

統計ソフト R を用いて, 正規分布に従うデータを作成した. ただし, データに多重共線性を持たすため, 変数 x_2 と x_3 は共線関係にあるようにした.

```
> x1 <- rnorm(20,m=10,s=1)
> x2 <- rnorm(20,m=20,s=1)
> x3 <- 5*x2+rnorm(20)
> x <- cbind(x1,x2,x3)
> y <- 3*x1+5*x2+x3+rnorm(20)
```

このデータをベースとして, 次のように 3 通りに外れ値を与えたデータについて ORR,RRR を適用する.

1. 説明変数 x_1 に外れ値を与えた場合
2. 目的変数 y に外れ値を与えた場合
3. 多重共線性の問題がある説明変数 x_3 に外れ値を与えた場合

9.2 分析結果

表 1 に, このデータを分析した結果を示す. ただし,k の値はリッジ・トレースの様子から推定している. リッジトレースとは, 横軸にリッジ・パラメータ k をとり, 縦軸にそれに対する推定量の軌跡をプロットしたものである. ここではその例として,LTS 推定量を用いた RRR のリッジ・トレースを示した. (図 2)

表 1　回帰分析の結果

x_1 に外れ値のあるデータ				
	ORR	M(Huber)	LMS	LTS
変数	$k=0.07$	$k=0.05$	$k=0.02$	$k=0.02$
x_1	-0.039	-0.052	3.203	3.239
x_2	5.848	6.070	5.104	5.039
x_3	0.969	1.042	1.018	1.034
y に外れ値のあるデータ				
	ORR	M(Huber)	LMS	LTS
変数	$k=0.10$	$k=0.11$	$k=0.13$	$k=0.13$
x_1	-19.078	3.304	3.077	3.084
x_2	3.913	5.168	5.135	5.123
x_3	1.666	0.866	0.818	0.810

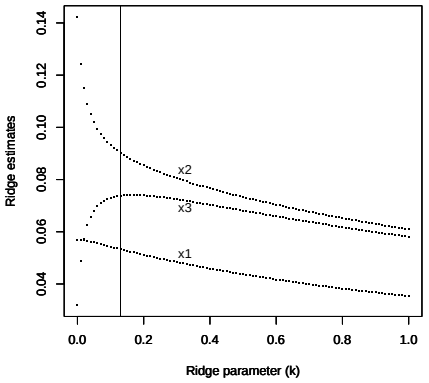

図 2　LTS 推定量を用いた RRR のリッジ・トレース ($k \in [0, 1.0]$) : y に外れ値のあるデータ

9.3 考察

R によって作成したデータから, 真の回帰直線は

$$y = 3x_1 + 5x_2 + x_3 \tag{20}$$

となることがわかっている. 以下,9.1 節に示した 3 通りのデータを分析した際の各推定量の精度を検証する.

≪説明変数 x_1 に外れ値がある場合≫

ORR 推定量と M 推定量を用いた RRR 推定量には大きな差異はなく,x_1 の係数推定値が外れ値の影響を強く受け, マイナスの値となってしまっている. これは,M 推定量がモデルの誤差項についてはロバストであるが, 説明変数項に対してはロバストではないという性質を持っているためであると推測できる. 一方,LMS 推定量,LTS 推定量は共に外れ値の影響をほとんど受けていない. パラメータ k の値は共に 0.02 と決定され, 真の係数値に非常に近い値を示した.

≪目的変数 y に外れ値がある場合≫

x_1 に外れ値がある場合と同様,LMS 推定量と LTS 推定量を用いた際のリッジ・トレースにより決定された最適な k の値は共に 0.13 となり, そのときの RRR 推定量も互いに似た値を示した. また, 真の係数値にも非常に近くなっている. M 推定量を用いたときの RRR 推定量を見ても,x_1 の係数推定値が若干高くなっているが x_3 については LMS,LTS を用いた際よりも真の係数値に近い値を示しており, かなり良い結果が得られた.

≪説明変数 x_3 に外れ値を与えた場合≫

1 つの変数に多重共線性と外れ値の問題が混在するようなこのケースでは,RRR を用いても $0 \leq k \leq 1$ におけるリッジ・トレースが終始不安定な状態を示しており, 最適な k の値を決定することが困難となってしまった. そのため, 表 1 にも分析結果を載せていない. また, 各 RRR について $0 \leq k \leq 1$ の範囲で k の値を徐々に増加させ, そのときの推定量の変化を観察してみても,ORR よりは多少良い結果が得られた程度で, 真の係数値に極めて近くなることはなかった.

10 おわりに

外れ値と多重共線性の問題が 1 つの説明変数に混在してしまうと, ロバスト・リッジ回帰を用いても良い結果が得られにくいということが判明したが, データにこの 2 種類の問題が示唆されていても,1 つの変数に同時に存在していない限りは最小 2 乗法や従来のリッジ回帰, ロバスト回帰を用いるよりもうまく分析できる可能性が高いということが確認できた.

参考文献

[1] Chatterjee,S.,Hadi,A.S. and Price,B.: Regression Analysis By Example, *John Wiley & Sons*,2000

[2] Grrob,J.: Linear Regression, *Springer*,2003

[3] Hoerl,A.E. and Kennard,R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics 12*, 55-67,1970

[4] Kibria,B.M.G.: Performance of Some New Ridge Regression Estimators, *Communications in Statistics—Teory and Methods 32*, 419-435, 2003

[5] 佐和隆光: 回帰分析, 朝倉書房, 2000

[6] Silvapulle: Robust Ridge Regression Based on an M-Estimator, *Australian Journal of Statistics 33*, 319-333, 1991