回帰分析の理論とその応用に関する研究
ーリッジ回帰と多重共線性ー

2009SE067 今枝建史郎
指導教員：木村美善

1　はじめに

卒業研究を始めるまではデータ分析の勉強が中心であったため分析の方法や分析の多様な手法は理解することができたが，統計学に関する数学的理論については学ぶ機会が少なかった．つまり、結果が得られる仮定，仕組みに関しては理解できていない所が多かったと言える．この背景から統計学についてより理解していきたいと思ったことと，理解を深めることで得られた数値結果からデータをより上手に活用したいと思い，この研究課題を選んだ．本研究の目的は多重共線性とリッジ回帰の理論を理解することである．また，データ解析は統計解析ソフト「R」を使用する．

2　線形回帰モデル

n 個の観測値が与えられた場合，目的変数を y, 説明変数を x_j とすると，回帰式は

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

と表される．ただし，$\beta_0, \beta_1, \cdots, \beta_p$ は回帰係数 ε_i は誤差項を示す．また目的変数 y_i の $n \times 1$ ベクトルを$\boldsymbol{Y}$，定数項と説明変数 $x_{1i}, x_{2i}, \cdots, x_{pi}$ の $n \times (p+1)$ ベクトルを $\boldsymbol{X}$，回帰係数 $\beta_0, \beta_1, \cdots, \beta_p$ の $(p+1) \times 1$ ベクトルを $\boldsymbol{\beta}$，誤差項 ε_i の $n \times 1$ のベクトルを $\boldsymbol{\varepsilon}$ とすると

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (2)$$

のように行列で書くことができる．([2] 参照)

3　最小 2 乗 (OLS) 推定量

(2) 式における $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ の最小 2 乗推定量は

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

であり，このとき，残差平方和は最小の値になる．最小 2 乗推定量は最良線形不偏推定量であり，さらに正規分布である場合は最良不偏推定量となる．([7] 参照)

4　リッジ回帰

4.1　リッジ回帰 (ORR) 推定量

それぞれの説明変数の間に多重共線性が存在する場合，最小 2 乗推定量 $\hat{\boldsymbol{\beta}}$ を縮小し安定化を図るため $\boldsymbol{X}'\boldsymbol{X}$ の対角要素にリッジ・パラメータと呼ばれる正定数 k を加え，推定量の平均 2 乗誤差が小さくなるように

$$\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad (3)$$

とする方法をリッジ回帰という．そして，(3) 式をリッジ推定量という．([5] 参照)

4.2　リッジ回帰における SSE と MSE

最小 2 乗推定量とリッジ回帰推定量を比べるときに用いられる指標の中に残差平方和 SSE と平均 2 乗誤差 MSE の 2 つがある．リッジ回帰推定量の SSE と MSE を求める際は最小 2 乗推定量のときとは異なる計算式になる．リッジ回帰推定量のときのそれぞれの計算式は以下の通りとなる．

$$SSE(\hat{\boldsymbol{\beta}}_k) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}) \quad (4)$$

$$MSE(\hat{\boldsymbol{\beta}}_k) = \sigma^2 \sum_{j=1}^{p} \lambda_j (\lambda_j + k)^{-2} + k^2 \boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-2}\boldsymbol{\beta} \quad (5)$$

ここで (5) 式の右辺の第一項は $\hat{\boldsymbol{\beta}}_k$ の成分の分散の和（総分散），第二項は偏りの 2 乗を表す．正定数 k の値に対して，右辺の第一項は単調減少し，第二項は単調増加することがわかる．([1], [3] 参照)

4.3　実行例

この節では，統計ソフト「R」の組み込みデータである longley データをリッジ回帰で分析する．

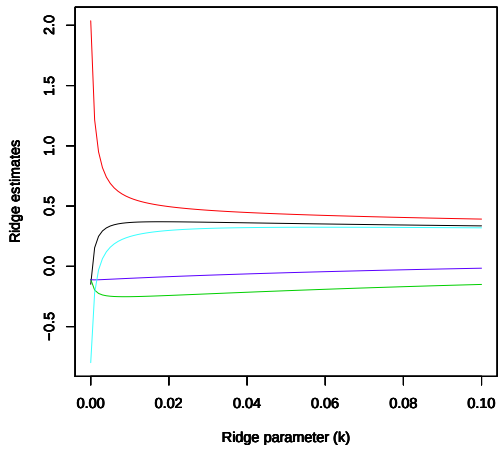

図 1　longley データのリッジ・トレース

この分析結果よりそれぞれの変数の係数は $k = 0.01$ でほぼ安定状態に入ることが分かった．よって，最小 2 乗

推定量である $k=0$ の結果と $k=0.01$ の結果を比較していく．$\tilde{\boldsymbol{\beta}}$ は標準化した場合の回帰係数ベクトルであり $\hat{\boldsymbol{\beta}}_i = \tilde{\boldsymbol{\beta}}_i(s_y/s_{x_i})$ となっている．SSE，MSE の値は $\tilde{\boldsymbol{\beta}}$ を用いて計算されている．

表 1　longley データの回帰分析結果

	OLS ($k=0$)		ORR ($k=0.01$)	
変数	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}$	$\tilde{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}$
Intercept	0	92.4613	0	35.1646
x_1	−0.1489	−0.0485	0.3639	0.1184
x_2	2.0378	0.0720	0.5685	0.0201
x_3	−0.1075	−0.0040	−0.2516	−0.0095
x_4	−0.1111	−0.0056	−0.0993	−0.0050
x_5	−0.7992	−0.4035	0.2471	0.1248
SSE	0.1893		0.2426	
MSE	19.3152		1.3662	

表 1 の結果より，$k=0.01$ のモデルは

$$y = 35.1646 + 0.1184x_1 + 0.0201x_2 - 0.0095x_3 - 0.0050x_4 + 0.1248x_5 \quad (6)$$

となる．$k=0$ と $k=0.01$ のときの $\tilde{\boldsymbol{\beta}}$ の SSE と MSE について見ると，$k=0$ の SSE は 0.1893 なのに対し，$k=0.01$ のときは 0.2426 と値が大きくなっている．SSE とは本来モデルの誤差を表すものなので MSE と同じく小さい方が望ましいが，この結果からは SSE が大きくなってしまっている．しかし，その増加量は微小であり影響力の小さいものであると考えることができる．次に MSE について考える．MSE は $k=0$ のとき 19.3152 であるのに対し，$k=0.01$ のときは 1.3662 まで大きく減少している．この大きな減少は推定量に大きく影響を与えるものである．以上より，係数推定値，SSE，MSE の値を総合的に考察すると，データに多重共線性が存在する場合は，最小 2 乗法よりリッジ回帰を用いて分析を行う方が良い結果が得られることがわかる．([4], [6] 参照)

5　ダミー変数を含むデータのリッジ回帰

5.1　多変量データ

longley データは最小 2 乗法よりリッジ回帰の方が優れた推定量を求めることができた。しかし、リッジ回帰は必ずしも最小 2 乗法より優れた推定量を求めることができるわけではない。ここではその一例を見ていく。統計ソフト R を用いて，独自に正規分布に従うデータを作成した．ただし，変数 x_2 と x_3 には多重共線性が現れるよう共線関係をもたせてある．本章では乱数を用いて作成したダミー変数を持つデータを分析し，ダミー変数をデータに含んでいる場合にリッジ回帰で満足する推定量を導くことができるかを考察していく．

表 2　OLS との比較

	x_1	x_2	x_3	x_4
真の係数値	1	2	3	3
k=0	1.0100	1.2237	3.1489	2.5801
k=0.02	1.1868	7.1269	1.9221	3.0198

5.2　実行例

結果を見ると $k=0.02$ のときの値は最小 2 乗推定量で求めたものと比べ真の係数値に近づいた係数値もあるが，ほとんどの係数が大きく離れてしまっている．特に多重共線性を持っている変数 x_2 の係数値は他の変数の係数値と比べ特異に大きく変化している．これはダミー変数である x_4 のような変数がないときには見られない結果であり，MSE は $k=0$ のときとは 0.1589 から 0.0135 に小さくなっているが，係数値が大きく異なることを考えると満足に分析することができていないと言える．また，別の乱数を用いて行なっても，結果は同じように多重共線性を持つ変数一つが特異に大きくなり同じような結果が得られた．以上の結果より，説明変数にダミー変数などの質的変数を含むときにリッジ回帰を用いる場合は注意が必要である．

6　おわりに

数学的観点から見る統計学の研究は大変勉強になったと感じる．回帰モデル，特にリッジ回帰についての勉強は参考にする資料のほとんどが海外の文献であったため英語を訳しながらの研究となり時間がかかったが，これにより，より理解を深めることができたと思う．それに加え，リッジ回帰分析，SSE と MSE を導くプログラムを書いたこともリッジ回帰に対し理解を深めることに繋がった．全体的に見て予測していた以上に研究することができ，自分では納得し満足している．

参考文献

[1] Grob,J. : *Linear Regression*, Springer, 2003.

[2] Rencher A.C. and Schaalje G.B: *Linear Models in Statistics*, John Wiley & Sons,Inc, 2008.

[3] S. チャンジー・B. プライス（佐和隆光・加納悟 訳）: 回帰分析の実際，朝倉書房，2011.

[4] Shewhart A.C. and WILKS S.S. : *Regression Analysis by Example*, John Wiley & Sons,Inc, 2006.

[5] 佐和隆光：回帰分析，朝倉書房，2011.

[6] 武山嵩弘：回帰分析の理論とその応用－リッジ回帰を中心に－，南山大学数理情報学部数理科学科卒業論文，2006.

[7] 武山嵩弘・木村美善：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，南山大学紀要『アカデミア』数理情報編，第 8 巻， pp. 35-46，2008.