回帰分析の理論とその応用
− リッジ回帰を中心にして −

2000MM092 戸開 秀聡
指導教員　木村 美善

1　はじめに

推測統計学上の諸問題と 3,4 年次でのゼミの内容から「リッジ回帰」に興味を持ち研究課題にしようと考えた. 興味をもった理由は $\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'y}$ という回帰推定量が提案され, 多重共線性問題の回避が出来るにも関かわらず, それほど重要視されていないのが現実であり, その理由を知りたかったからである. 本研究の目的は,「リッジ回帰」の現状を理解するとともに, その問題点を整理し, 最小 2 乗 (OLS) 推定量とリッジ回帰 (RR) 推定量をシミュレーションにより比較・考察することである.

簡単のためモデル式などにある変数はベクトル ($n\times1$, 小文字) と行列 ($n \times p$, 大文字) で表示するものとする.

なお, 本論文では参考文献 [1],[2],[3],[4] を参照した.

2　回帰モデルのまとめ

2.1　線形 (正規) 回帰モデル

モデル　$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$
仮定 1　$E[\boldsymbol{\epsilon}] = \boldsymbol{0}$
仮定 2　$V[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}$
仮定 3　$\mathrm{rank}\boldsymbol{X} = p$
仮定 4　$\boldsymbol{\epsilon}$〜$N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$

このモデルと仮定 1〜3 までが与えられている状態を「線形回帰モデル」と呼び, 仮定 4 を付け加えたものを「線形正規回帰モデル」と呼ぶ. 仮定 4 のもとで,

$$\boldsymbol{y}\text{〜}N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}) \tag{1}$$

であり, $\boldsymbol{y}$ の密度関数は

$$f(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right] \tag{2}$$

である.

また, このモデルの $\boldsymbol{X}$ は与えられたものとしているが,$\boldsymbol{X}, \boldsymbol{y}$ を確率変数とみなす考え方もある. これについてはここでは触れないことにする.

2.2　「線形」の意味

$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ が線形というのは, パラメータ $\boldsymbol{\beta}$ に関して「線形」という意味であることに注意する.

3　リッジ回帰

3.1　リッジ回帰 (RR,ORR) 推定量

Hoerl and Kennard(1970) の RR 回帰推定量は, それまでに最良とされていた OLS 推定量に手を加えたものとして提案されたものである.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \tag{3}$$
$$\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'y} \tag{4}$$
$$\hat{\boldsymbol{\gamma}}_k = (\boldsymbol{\Lambda}^{-1} + \boldsymbol{K})\boldsymbol{Z'y} \tag{5}$$
$$\tilde{\boldsymbol{\beta}}_k = \boldsymbol{P}(\boldsymbol{\Lambda} + \boldsymbol{K})^{-1}\boldsymbol{Z'y} \tag{6}$$

ここで,$\boldsymbol{PP'} = \boldsymbol{I}$　,$\boldsymbol{Z} = \boldsymbol{XP}$ である. 現在, リッジ回帰はデータ $\boldsymbol{X}$ に多重共線性がある場合に回帰係数の推定量を得る方法として用いられている.

3.2　多重共線性

多重共線性とは簡単に言えば,「変数同士の相関が互いに強い」ということになる. 例えば, 次のような相関行列を考えると,$\boldsymbol{x}_2$ と $\boldsymbol{x}_3$ に共線関係が存在する.

$$\begin{pmatrix} 1.000 & -0.369 & -0.455 \\ -0.369 & 1.000 & 0.985 \\ -0.455 & 0.985 & 1.000 \end{pmatrix}$$

変数同士で一次関係が成り立ち,$\boldsymbol{x}_1 = a\boldsymbol{x}_2 + b\boldsymbol{x}_3$ のような独立ではなく従属の関係にある場合や $\boldsymbol{x}_1 \propto \boldsymbol{x}_2$ と極限的な従属関係にある場合には $\mathrm{rank}\boldsymbol{X} = p-1$ と次元が減少することになる. 次元の減少は多変量解析をする際に問題となってしまう.

このように多重共線性がデータに確認された場合,「多重共線性問題」とみなし" 回避", または" 解決" しておかなければならない.

3.3　多重共線性の" 回避" と" 解決"

Hoerl and Kennard(1970) の論文では多重共線性を" 回避" するために RR 推定量を提案しているわけではない. データの積率行列 $\boldsymbol{X'X}$ の" 固有値からなる対角行列 $\boldsymbol{\Lambda}$" に注目し, その固有値の最小の値がかなり小さい場合に $(\boldsymbol{X'X})^{-1}$ を求めるのが困難になる. それで極小の変化をパラメータ k によって与え, 逆行列を求めやすいようにしている. 結果的に共線関係を" 回避" したと言えることになる.

リッジ回帰が主流とならない理由は" 解決" としての方法「変数選択」によってモデル式を考えることが現在

の統計学上では良いとされているからであるように思える.

$$C_p = ||\boldsymbol{y} - \hat{\boldsymbol{\beta}}||^2 + (2p - n)\sigma^2 \quad (7)$$

$$AIC = -2\log f(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + 2p \quad (8)$$

ここに示した 2 つの式は「C_p 基準」と「AIC 基準」と呼ばれているが, 変数選択の基準として用いられ, 非常に洗練された式であると言っても過言ではない. これら以外に「前進選択法」,「後退消去法」,「逐次法」が代表的である. 前進選択法と後退消去法は互いに逆のプロセスにより変数選択を行う. 逐次法は前進選択法を改良した方法である. 現在, これらの方法によって変数選択が行われているが, 様々な問題点がある. 逐次法は膨大な計算量を必要とする. また, 各変数選択法により, 選択される変数の組が異なってくる. いかに効率良く, 納得のいく変数選択を行っていくかは, 現在も重要な研究課題となっている.

4 OLS と RR の比較

OLS 推定量 $\hat{\boldsymbol{\beta}}$ と RR 推定量 $\hat{\boldsymbol{\beta}}_k$ が用いられる場合は次の通りである.

表 4.1OLS と RR の用いられる状態

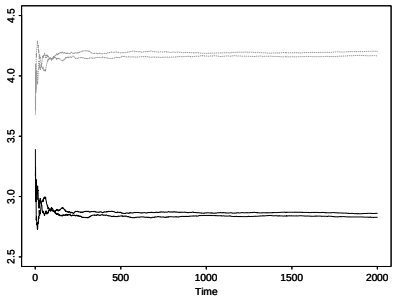
ケース	多重共線性	$\boldsymbol{\epsilon}\sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$	推定量
1	なし	従う	$\hat{\boldsymbol{\beta}}$
2	なし	従わない	
3	あり	従う	$\hat{\boldsymbol{\beta}}_k$
4	あり	従わない	

表 4.1 に示すように, 場合によって推定量を使い分ける必要がある, 誤差項が正規分布に従っていない場合 (2 と 4) に OLS と RR のどちらが良いのかについて考えることにする. 例として

$$\boldsymbol{\epsilon}\sim(1-\delta)N(0,1) + \frac{\delta}{2}N(2,0.1) + \frac{\delta}{2}N(-2,0.1) \quad (9)$$

という混合正規分布 $(0 < \delta < 1)$ に従う場合のシミュレーションを行った. $\boldsymbol{X}(100 \times 4)$ のデータを 2 つ (多重共線あり, なし) 用意し, そのデータと誤差の一次結合 $\boldsymbol{y} = \boldsymbol{x}_1 + 2\boldsymbol{x}_2 + 3\boldsymbol{x}_3 + 4\boldsymbol{x}_4 + \boldsymbol{\epsilon}$ で真の係数を決めると同時に,$\boldsymbol{y}$ を用意し,OLS と RR をそれぞれ試し, 計 4 種類行った. 試行回数 10000 回とし, 誤差項のみ毎回正規乱数により生成している. 図 4.1 は RR と OLS の毎回の結果を平均したものである.

図 4.1.OLS と RR のシミュレーション結果



多重共線性がある場合 ($\boldsymbol{x}_3$ と $\boldsymbol{x}_4$ が共線関係) に RR と OLS による係数推定値が図 4.1 である. ここに記載したものは 10000 回中の 2000 回分である.1000 回を越えてから係数の変化はみられず, 収束した. この結果から RR による推定の方が真の値に近くなり, 収束が速いことがわかる. $\boldsymbol{\epsilon}$ の分布をここでは両側に重みをつけた状態のみ示したが, 片側につけた場合についても同様の結果となった.

5 おわりに

本論文の目標の一つとしていたリッジ回帰推定量を改良して, 不偏推定量を導出しようとしたが, 導出の途中で矛盾が生じてしまう結果となったので本論文には記載していない. しかしながら, 当初は特に考えていなかった仮定が正規分布に従わない状態に対してシミュレーションをした結果, リッジ回帰を用いることで, OLS よりうまく収束することを示すことが出来た. また, リッジ回帰による推定にはバイアスが生じることが参考文献 [2] に示されている. パラメータ k が小さければバイアスの値は小さくなり,RR の方が OLS 推定値より安定していることを確認できた.

参考文献

[1] Anderson,T.W.: Introduction to Multivariate Statistical Analysis, Second Edition.John Wiley & Sons, Inc., New York, 1984.

[2] Hoerl,A.E. and Kennard,R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics,42,55-67, 1970.

[3] Kibria,B.M.G.: Perfomance of some new ridge regression estimators. Communication in Statistics-Simulation and Computation Vol.32, No.2, 419-435, 2003.

[4] 佐和隆光: 回帰分析, 朝倉書店 2000.