回帰分析の理論とその応用

2007MI002 阿部智成
指導教員：木村美善

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, 望ましい推定量である. しかし, 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が知られている. 多重共線性の問題に対して, 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一つである. しかし, このリッジ回帰は, 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という欠点がある. 外れ値が存在するとき, 外れ値に対する影響を受けにくいロバスト回帰を用いることが望ましい.
本研究の目的は, リッジ回帰とロバスト回帰の理論を理解し, 実際にデータを用いて, それらの比較考察を行うことである.

2　線形回帰モデル

2.1　最小 2 乗推定量

目的変数 y_i の $n \times 1$ ベクトルを y, 定数項と説明変数 $x_{i1}, x_{i2}, \cdots, x_{ip}$ の $n \times (p+1)$ ベクトルを X, 回帰係数 $\beta_0, \beta_1, \cdots, \beta_p$ の $(p+1) \times 1$ ベクトルを β, 誤差項 ε_i の $n \times 1$ のベクトルを ε とし, 線形回帰モデル

$$y = X\beta + \varepsilon \tag{1}$$

を考える. 残差平方和 (RSS) は

$$RSS[\beta] = \| \varepsilon \| = (y - X\beta)'(y - X\beta)$$

により定義される. RSS は

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2}$$

のとき最小になる. これが式 (1) における $\beta = (\beta_0, \beta_1, \cdots, \beta_p)'$ の最小 2 乗 (LS) 推定量である.LS 推定量は,ε が $E[\varepsilon] = 0$,$V[\varepsilon] = \sigma^2 I$($I$ は単位行列) を満たすとき最良線形不偏推定量であり, さらに正規分布に従うときには最良不偏推定量となる. しかし, 実際のデータはこのような標準的仮定からずれていたり, 外れ値や多重共線性が存在したりする. このような場合 LS 推定量は不安定になる.([5] 参照)

3　リッジ回帰推定量

3.1　リッジ回帰とは

この手法は, リッジ・パラメータと呼ばれる $k > 0$ を取り入れることによって回帰係数の安定化を図るものである. リッジ回帰推定量は偏りをもつ推定量であるが, 適切な k を選ぶことによって最小 2 乗推定量よりも小さい平均 2 乗誤差を持つようにすることができる.$X'X$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_{p+1}$ とする. このとき, 回帰係数 β の LS 推定量 $\hat{\beta}$ の平均 2 乗誤差 (MSE) は

$$MSE[\hat{\beta}] = E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] = \sigma^2 \sum_{i=1}^{p+1} \lambda_i^{-1} \tag{3}$$

と表される. ただし, X,Y は標準化されているものとする. ここで, データに多重共線性の問題があるとき, 固有値 λ にはきわめて 0 に近いものが存在するため,(3) 式で与えられる LS 推定量の MSE は大きく発散する可能性がある. そこで, リッジパラメータと呼ばれる定数 $k > 0$ を導入し,LS 推定量 $\hat{\beta}$ を縮小することによって回帰係数の安定化をはかる. この推定量は

$$\hat{\beta}_k = (X'X + kI)^{-1}X'y \tag{4}$$

である.$\hat{\beta}_k$ は $k > 0$ のときバイアスを伴うため不偏推定量とならないが, 説明変数間に多重共線性の問題があるとき,$\hat{\beta}$ よりも小さい MSE を与える k が存在する. 当然,$k = 0$ のとき LS 推定量と一致する.([3] 参照)

4　ロバスト回帰推定量

4.1　ロバスト回帰とは

ロバスト回帰法とは真の分布が指定した分布とずれがあっても効率がそれほど減少しない回帰分析法である. 外れ値がある場合,LS 直線は外れ値に大きく影響を受け, データにうまくあてはまらない. 単回帰の場合には実際に 2 次元平面上で外れ値を見ることができるので, それを取り除いてもう一度解析しなおせばよいが, 重回帰の場合それが困難である. このことから外れ値の検出は最小 2 乗法ではうまくいかないことが多い. そして, この問題を克服するためにロバスト回帰という方法が考案された.([2] 参照)

4.2　LMS 推定量と LTS 推定量

LMS 推定量 $\hat{\theta}_{LMS}$ は,Rousseeuw(1984) により, 高い破綻点を得るように Hample(1975) のアイディアに基づき提案されたもので

$$med_i \ \ r_i^2(\hat{\theta}_{LMS}) = \min_{\theta} med_i \ \ r_i^2(\theta) \tag{5}$$

により定義される. ここで $med_i \ \ r_i^2$ は残差の 2 乗 r_i^2 の中央値である. この推定量は,y の外れ値と同様に x の外れ値に関しても強く, 破綻点は 50 % である.LTS 推定量 $\hat{\theta}_{LTS}$ は,Rousseeuw(1985) により

$$\sum_{i=1}^{h} (r^2(\hat{\theta}_{LTS}))_{i:n} = \min_{\theta} \sum_{i=1}^{h} (r^2(\theta))_{i:n} \tag{6}$$

と定義される.$(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:m}$ は残差の 2 乗を小さいほうから並び替えたものであり, 小さいほうから h 番目までの和を最小とする θ である.LTS は LS と似ているが, 大きい残差が和に含まれないことで外れ値を避けることができるのでその影響を受けにくくなる. この推定量の破綻点は h が $[n/2]+1$ のときに,50 % に達する.([4] 参照)

5 実行例

5.1 リッジ回帰推定量

多重共線性が存在する自作のデータを使ってみていく. データは独立に正規分布に従う乱数を用いて多重共線性が存在するように生成したものである. このデータを用いて,LS 推定量とリッジ回帰推定量を比較する. このデータの相関係数を見ただけでは 1 に近い値はなく, 共線性があるか判断がつかないが, 固有値を見ると,0.0026 と非常に 0 に近いものがあるすなわち, このデータには, 多重共線性の疑いがあることがわかる.
このデータをリッジ回帰分析すると,$k = 0.09$ の時最も MSE が小さくなり当てはまりが良くなる. この結果を LS 推定量と比較してみる.SSE は, リッジ回帰の方が少し大きいが,MSE はリッジ回帰のほうが小さく, よく当てはまっているといえる. これより, 多重共線性の存在するデータにおいては,LS 推定量よりリッジ回帰推定量の方が良い当てはまりを示していることがわかる.

5.2 LS,LTS,LMS 推定量の比較

外れ値の存在する自作のデータに対する LS,LMS,LTS で回帰推定量は以下の表のようになる.
回帰係数の LS 推定量と比べて LMS と LTS による推定量の方がよく当てはまっているのがわかる. このことから, 重回帰分析においても LMS,LTS 推定量は, 外れ値の影響を受けにくいことが確かめられた.

表 1 回帰係数

	x1	x2	x3
ls	3.300	3.028	3.308
lms	1.003	2.010	2.995
lts	1.003	2.004	2.999

6 ロバストリッジ回帰推定量

実際のデータには多重共線性に加えてさらに外れ値が存在するような場合がある. この問題を解決するために考えられた手法としてロバスト・リッジ回帰法がある.

6.1 M 推定量に基づくロバスト・リッジ回帰

M 推定量に基づくロバスト・リッジ回帰推定量は,LS 推定量 $\hat{\beta}$ を β の M 推定量 $\hat{\beta}^M$ で置き換えたリッジ推定量であり, 次のように定義される.

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

これは, 第 1 段階で加重最小 2 乗法による解を求め, 第 2 段階でリッジ回帰を適応するものである. こうして, リッジ回帰と M 推定量などのロバスト回帰を組み合わせることにより, 多重共線性の問題に加え, 外れ値が存在する場合にもほとんど性能を損なわずに分析を行うことが可能になる. このことを実際に統計解析ソフト R を用いて示したのが次の例である.
データは独立に正規分布に従う乱数を用いて生成し, 多重共線性と外れ値を混在させている. これを M 推定量に基づくロバスト・リッジ回帰推定量は表 2 のようになる.

表 2 分析結果

k	x1	x2	x3	x4	x5	SSE
0.0	−0.750	−0.700	−1.992	−2.234	6.556	83.603
0.1	−0.752	−0.702	−1.990	−2.231	6.556	83.598
0.2	−0.754	−0.704	−1.987	−2.227	6.555	83.595
0.3	−0.755	−0.706	−1.985	−2.224	6.555	83.595
0.4	−0.758	−0.708	−1.983	−2.220	6.555	83.596

6.2 考察

結果を見たところあまり当てはまりも良くないし,SSE も $k = 0.5$ 以上になると増加してしまう. これは M 推定量が y 軸方向の外れ値に対してのみ頑健であるという性質が維持されているからと考えられる. この結果から, 他のロバスト回帰推定量に基づくロバスト・リッジ回帰推定量についてもはロバスト推定量の特性をそのまま受け継ぐのではないかと推測される.([6] 参照)

7 おわりに

本研究では, リッジ回帰とロバスト回帰は実行例を用いてその有効性を示した. しかし, ロバスト・リッジ回帰の理論に少ししか取り組むことができなかった.M 推定量以外のロバスト回帰推定量を縮小たものを比較したかったが, 時間が足りなかった. 大学院に進んでからも引き続きこの研究を進め, リッジ回帰とロバスト回帰の発展に貢献したい.

参考文献

[1] Faraway,J. J.：*LinearModelswithR*. Chapman and Hall, 2003.

[2] Grob,J.G:*LinearRegression*.Springer,1991.

[3] Hoerl,A.E.and Kennard,R.W.:Ridge Regression Applications to Nonortogoual. *Technometrics*, 1974.

[4] Rousseeuw,P.J.and Leroy,A. M.:*RobustRegressionandOutlierDetection*. John Wiley and Sons, 1986.

[5] 佐和 隆光：回帰分析. 朝倉書店, 1979.

[6] Sibapulle,M.J.:Robust ridge regression based on M-estimator. *Austral.J.Statist*, 1991.