# 回帰分析の理論と応用
## ―リッジ回帰を中心に―

2003MM117 山田樹里
指導教員: 木村美善

## 1　はじめに

3,4年次のゼミで回帰分析の多重共線性という問題を扱った.回帰分析では,通常説明変数は直交していないが,多くの場合そのことは重大な問題とはならず,分析に大きな影響は無い.しかし,説明変数間に強い線形関係が存在すると多重共線性の問題が発生し,回帰分析の結果は信頼性を欠くものになる.このような問題をもつ困難なデータを検出し,解決するための手法を研究したいと考えた.

本研究では,多重共線性の検出方法を紹介し,多重共線性を含むデータを用いて,最小2乗法とリッジ回帰と主成分回帰とを比較し,考察する.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　線形回帰モデル

従属変数$y_1, \cdots, y_i$,説明変数$x_1, \cdots, x_p$についての$n$個の観測値が与えられた場合,回帰式は次のように表せる.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i \qquad i = 1, 2, \cdots, n \quad (1)$$

ただし,$\beta_0, \beta_1, \cdots \beta_k$は回帰係数,$\epsilon_i$は誤差項である.$k = 1$のときを「単回帰モデル」といい,$k \geq 2$のときを「重回帰モデル」という.

次の誤差項に関する仮定のうち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.

1. $E(\epsilon_i) = 0$　(不偏性)
2. $V(\epsilon_i) = \sigma^2$　(等分散性)
3. $Cov(\epsilon_i, \epsilon_j) = 0$　(無相関性)
4. $\epsilon \sim N(0, \sigma^2 I)$　(正規性)

## 3　最小2乗法

線形回帰モデル$Y = X\beta + \epsilon$において実測値$y_i$と予測値$\hat{y_i}$の差を残差といい,この残差の2乗和(SSE)を最小とする考え方を最小2乗法という.

SSEは$\beta$ の2次関数になるので,$\beta$ で偏微分したものを0とすれことにより,最小2乗(OLS)推定量は次のように得られる.

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2)$$

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.([4]参照)

## 4　多重共線性

### 4.1　多重共線性(Multicollinearity)とは

回帰分析において説明変数間に強い相関関係が存在するとき,その回帰分析の結果は信頼性に欠けるものとなることがある.説明変数同士の非直交性が極端な場合,その状態を「多重共線性が存在する」という.

### 4.2　多重共線性の検出

多重共線性の検出について,ここでは２つの方法を紹介する.まず,固有値による検出について述べる.説明変数間の相関行列を$V(x)$とし,その固有値を$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$とする.固有値が0に近いとき,多重共線性が存在することがわかる.また,Condition Numberを

$$t = \sqrt{\frac{\lambda_1}{\lambda_p}} \quad (3)$$

とする. $t$が大きいと,強い多重共線性の証拠となる.

もうひとつの検出法は,分散拡大要因(VIF)を用いるものである.$R_i$を$i$番目の説明変数を他のすべての説明変数に回帰したときの重相関係数$R_i$とするとき,第$i$番目の説明変数のVIFは,

$$VIF(i) = 1/(1 - {R_i}^2) \quad (4)$$

で与えられる.一般的に,(4)式が10以上のとき深刻な多重共線性が存在すると考えられる.([1],[4]参照)

## 5　リッジ回帰

### 5.1　リッジ回帰(ORR)推定量

Hoer and Kennard(1970)は,パラメータ$k > 0$を取り入れることによって回帰推定量の安定性を高めるために次のリッジ回帰(Ordinary Ridge Regression)推定量

$$\hat{\beta_k} = (X'X + kI)^{-1}X'Y \quad (5)$$

を提案している.特に,$k = 0$のとき$\hat{\beta_k}$はOLS推定量に等しい.

### 5.2　一般化リッジ回帰(GRR)推定量

(5)式のORR推定量を一般化した便利なものとして,GRR推定量がある.対称正値定符号行列

$$K = Udiag(k_1, \cdots, k_p)U' \quad (6)$$

を考える.ただし,$X'X = Udiag(\lambda_1, \cdots, \lambda_p)U'$で,$U$は直交行列である.この$K$を用いた推定量

$$\hat{\beta}_K = (X'X + K)^{-1}X'Y \quad (7)$$

が,一般化リッジ回帰(Generalize 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できる.

$k$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案されているが,実際には,一般化リッジ回帰においてはリッジトレースからトレースの値が動かなくなったときの値を$k$の値として採用する方法が一般的に用いられている.このような$k$の決定にはある程度恣意性が入ることはやむを得ない.([2],[5]参照)

## 6 主成分回帰

主成分回帰とは,説明変数の主成分のうち対応する固有値が極端に小さいものを落として残りの主成分を用いて回帰を行い,得られた係数を元の説明変数の係数に戻す方法である.主成分の数を決定する方法はScree plotを示すのがよい.

### 6.1 主成分回帰(PCR)モデル

説明変数の固有値$\lambda_1, \cdots, \lambda_p$に対する正規化された固有ベクトルを$V_1,\cdots,V_p$とすると,回帰モデルは次のように表すことができる.ただし,$E(\epsilon)=0$,$E(\epsilon\epsilon')=I$で,$Y$と$X$は,$X'X$と$X'Y$が相関係数の行列となるように標準化されている.

$$\begin{aligned} Y &= X\beta+\epsilon \qquad (8)\\ &= XVV'\beta+\epsilon \\ &= C\alpha+\epsilon \qquad (9)\end{aligned}$$

ただし,

$$C=XV, \alpha=V'\beta \qquad (10)$$

行列$C$は,$p$個の列$C_1, \cdots, C_p$からなり,$C_1, \cdots, C_p$は説明変数$X_1, \cdots, X_p$の線形関数で,主成分と呼ばれる.また,$C_1, \cdots, C_p$は直交で,${C_j}'C_j=\lambda_j$と${C_i}'C_j=0 (i\neq j)$を満たす.さらに,(7)式より次の式が得られる.

$$C_i=\sum_{i=1}^{p} V_{ij}X_i \qquad (11)$$

$\lambda_i=0$のとき,これは多重共線性の原因となっている線形関係を正確に表現するものである.([1]参照)

## 7 実行例

### 7.1 データ

データは,1975年 Moter Trend Magazineから引用したもので,30種類の車について集められたものである.それぞれの車の燃費が,他の11個の測定値によって与えられている. このデータの固有値は,$\lambda_1=7.703$, $\lambda_2=1.403$, $\lambda_3=0.773$, $\lambda_4=0.577$, $\lambda_5=0.211$, $\lambda_6=0.142$, $\lambda_7=0.085$, $\lambda_8=0.050$, $\lambda_9=0.033$, $\lambda_{10}=0.008$, $\lambda_{11}=0.003$ である.また,Condition Number は$t=50.67$である. VIFの値は,**表1**に示した.

### 7.2 分析結果

4.2で説明したCondition Number$t$やVIFの値からこのデータは多重共線性の存在することがわかった.そこで,OLSとORRとPCRを用いて分析を行った結果を示す.分析方法は[3]を参考にした.ORRについては,リッジトレースを観察して,$k=0.15$を使って回帰分析を行った.またPCRについては,固有値0.01以下の第10,11主成分を落とし,第９主成分までを用いた.それぞれの回帰分析で得た回帰係数は**表1**に示した.

表 1: 分析結果

| 変数 | VIF | OLS ($k=0$) | ORR ($k=0.15$) | PCR |
|---|---|---|---|---|
| $x_1$ | 128.83 | −1.429 | −0.801 | −0.609 |
| $x_2$ | 43.92 | −0.532 | −0.179 | −0.331 |
| $x_3$ | 160.44 | 1.626 | 0.656 | −0.273 |
| $x_4$ | 2.06 | 0.057 | 0.084 | 0.146 |
| $x_5$ | 7.78 | 0.512 | 0.394 | 0.316 |
| $x_6$ | 5.33 | 0.051 | 0.000 | −0.070 |
| $x_7$ | 11.74 | −0.335 | −0.258 | −0.310 |
| $x_8$ | 20.59 | 0.628 | 0.548 | 0.433 |
| $x_9$ | 9.42 | −0.352 | −0.300 | −0.320 |
| $x_{10}$ | 85.68 | −0.806 | −0.732 | −0.405 |
| $x_{11}$ | 5.14 | 0.035 | 0.022 | 0.038 |

## 8 おわりに

データの固有値やVIFを調べることによって,多重共線性が存在すると考えられることが確認できた.ORRとPCRを研究し,多重共線性が存在する場合の回帰分析方法として,ORRとOCRは,OLSよりも安定していることがわかった.しかしながら,実用上問題となる点があることもわかった.ORRにおいては,kの値を決定することに恣意性があり,PCRでは,どこまで主成分を落とすかはっきりとした基準がない.さらに,OLS推定量,ORR推定量,PCR推定量の結果を比較すると,結果が互いに異なり,どれを用いたらよいのか迷う.これらの推定法を実際に適用する場合にまだまだ解明すべき多くの問題があると考えられる.しかしながら,データに極度の多重共線性の疑いがある場合には,OLS推定値の他に少なくともう１つの推定値を得ることは,試みる価値があるといえる.

## 参考文献


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

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

[3] Jalian J. Faraway: Linear Models with R, CHAPMAN& HALL/CRC,(2005).

[4] 佐和隆光：回帰分析, 朝倉書店 (2002).

[5] 武山嵩弘.澤田謹志：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(2005).