回帰分析法のロバストネスの理論とその応用に関する研究

2001MM075 佐久間　彩
指導教員　木村　美善

1　はじめに

多変量解析法の中でも，多くの方法論の考え方の基礎になる回帰分析に焦点を当てて研究する．回帰分析において，データが正規分布に従っている場合では，推定値を求める方法として最小二乗法 (least squares (LS) method) が使われてきた．しかし，正規分布に従っていることはまれであり，外れ値に影響されやすいという欠点があり,LS 法は分析方法としてよくない場合がある．そこで，外れ値に影響されにくい特徴を持つロバスト回帰分析を知り，興味を持った．本研究の目標は，ロバスト回帰の理論を理解し，データを使って LS 法とロバスト回帰分析手法の違いを見ることである．

2　線形回帰モデル

2.1　モデルの定式化

n 個の観測値がある場合，回帰式は

$$y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + e_i\,, i = 1, \cdots, n \qquad (1)$$

y_i は応答変数，x_{i1}，$\cdots$，x_{ip} は説明変数，e_i は誤差項である．ここで，説明変数が 1 個の時は単回帰モデル，2 個以上の時は重回帰モデルと呼ぶ．e_i は平均 0，未知の標準偏差 σ の正規分布に従うと仮定される．

2.1.1　最小二乗推定量 (LS)

$$\sum_{i=1}^{n} r_i(\hat{\theta}_{LS})^2 = \min_{\hat{\theta}} \sum_{i=1}^{n} r_i(\hat{\theta})^2 \qquad (2)$$

最小二乗法は残差の二乗和を最小にする．

3　ロバスト回帰

3.1　Breakdown point(破綻点)

T を回帰推定量とし，$T(Z) = \hat{\theta}$ と表す．元のデータの中の m 個を，任意の値 (かなり悪い外れ値を考慮に入れる) に置き換えた時のデータを Z' とし，あらゆる可能な Z' を考える．汚染によって生じる最大バイアスは,

$$bias(m; T, Z) = \sup_{z'} ||T(Z') - T(Z)|| \qquad (3)$$

有限標本 Z での推定量 T の破綻点は

$$\varepsilon_n^* = min\left\{\frac{m}{n}; bias(m; T, Z) is\ infinite\right\} \qquad (4)$$

3.2　LMS 推定量 (Least median of squares estimator)

$$\underset{i}{median}\ r_i(\hat{\theta}_{LMS})^2 = \min_{\hat{\theta}}\ \underset{i}{median}\ r_i(\hat{\theta})^2 \qquad (5)$$

これは 1984 年に Rousseeuw によって提案された．LS 推定量の和の部分を中央値に置き換えることで非常にロバストである．LMS 推定量は 50% 破綻点 (50% は破綻点の最大値) をもつ ([3] を参照).

3.3　LTS 推定量 (Least trimmed squares estimator)

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

Rousseeuw によって,LMS 推定量とともに提案されたのが LTS 推定量である．(6) は LS に似ていて，残差平方和の大きなものを使わないことが, 唯一の違いである．それによって，外れ値の影響を受けず，データの主要部分によく当てはまる．

3.4　LQS 推定量 (Least quantile of squares estimator)

$$\hat{\theta}_{LQS}(Z) = \underset{\theta}{argmin}(r^2(\theta))_{h:n} = \underset{\theta}{argmin}\,|r(\theta)|_{h:n} \qquad (7)$$

通常では最適解 $h_{opt} = [(n+p+1)/2]$ と定義する．LQS 推定量は LMS 推定量を一般化したものであり，$h = [n/2] + 1$ にすると，LMS 推定量になる．[ ] はガウス記号で，[n/2] は n/2 を超えない最大の整数である ([4] を参照).

4　決定係数

今回,LMS 推定量で使う決定係数は, 文献 [3] より，回帰式に切片を含む場合は

$$R^2 = 1 - \left(\frac{med|r_i|}{mad(y_i)}\right)^2 \qquad (8)$$

$$mad(y_i) = 1.4826\ \underset{i}{med}\{|y_i - \underset{j}{med}\ y_j|\} \qquad (9)$$

また [4] より LMS，LTS，LQS 推定量の決定係数は

$$R^2_{LQ(T)S} := 1 - \frac{s^2_{LQ(T)S}(X, y)}{s^2_{LQ(T)S}(1, y)} \qquad (10)$$

ただし，

$$s_{LQS}(X, y) = c_{h,n}|r(\hat{\theta}(Z))|_{h:n} \qquad (11)$$

$$s_{LTS}(X, y) = d_{h,n}\sqrt{\frac{1}{h}\sum_{i\,=\,1}^{h} (r^2(\hat{\theta}_{LTS}(Z)))_{i:n}} \qquad (12)$$

$c_{h,n}$と $d_{h,n}$ は定数である．

5　重回帰分析

文献 [1] よりたばこの消費量のデータを扱う．このデータは 51 個の観測値をもつ．Income(x_1), Price(x_2), Sales (y) の 3 つの変数を使って解析する．回帰式は

$$\begin{aligned}
\text{LS}: \hat{y} &= 0.022x_1 - 3.018x_2 + 153.338 \\
\text{LMS}: \hat{y} &= 0.018x_1 - 1.31\ x_2 + 93.317 \\
\text{LTS}: \hat{y} &= 0.016x_1 - 1.45\ x_2 + 106.605 \\
\text{LQS}: \hat{y} &= 0.018x_1 - 1.3\ \ x_2 + 95.133
\end{aligned}$$

図は, 横軸が予測値で, 縦軸が標準化残差である.

＜ LS ＞

図 1 より残差は 0 のまわりにランダムに分布してるようである．ほとんどのデータが± 2.5 の間に収まっているが, 30 番が外れ値である．決定係数も 0.250 と低く回帰式の当てはまりは 25% で良くない．R_a^2=0.219.

＜ LMS ＞

図 2 より LMS の残差は 0 のまわりにランダムに分布しており，± 2.5 の間に収まっているデータもある．一方，30，29，34，9，12，18，45，38，8，20，49，10 番の 12 個が外れ値である. 式 (8) の決定係数は 0.983, 式 (10) は 0.69．2 つ決定係数を計算したが, データの散らばり方から当てはまりは 69% というのが妥当である.

＜ LTS ＞

図 3 より LTS は外れ値 10 個である．LMS より 2 個少ない．LMS と比べると 10，49 番がない．決定係数は 0.719 ので当てはまりは 72% である.

＜ LQS ＞

LQS の回帰式や**図 4** より残差プロットは LMS に近い．LQS は外れ値 12 個で LMS の時の番号と同じであった．決定係数は 0.674 であるので当てはまりは, 67% である.

6　おわりに

たばこの例で，LS 推定量では外れ値が 1 個でも存在すれば, 推定量が影響を受けることや, LMS, LTS, LQS 推定量では外れ値が半分以下であったので影響を受けないことが分かった.

参考文献

[1] Chatterjee, S., Hadi, A. S. and Price, B. ：Regression Analysis by Example, Wiley, New York (2000).

[2] 藤木美江：回帰分析とその応用に関する研究，南山大学経営学部情報管理学科卒業論文 (2001).

[3] Rousseeuw, P. J. and Leroy, A. M. ：Robust Regression and Outlier Detection, Wiley, New York (1987).

[4] Rousseeuw, P. and Hubert, M. ：
Recent developments in PROGRESS,
`http://win-www.uia.ac.be/u/statis`.

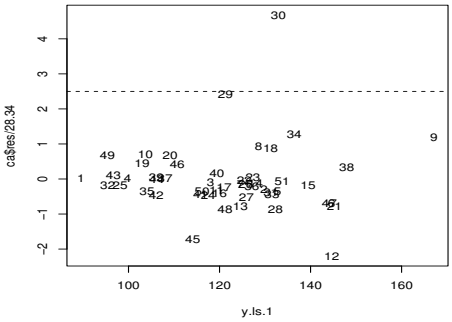

図 1　LS 推定量の残差プロット

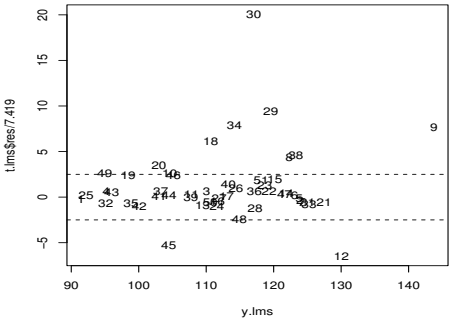

図 2　LMS 推定量の残差プロット

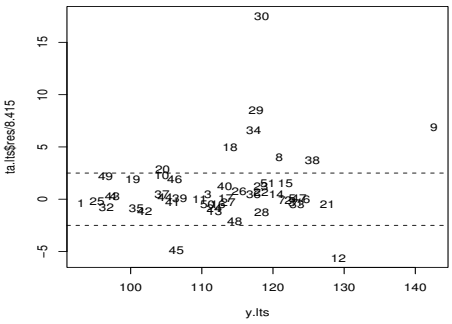

図 3　LTS 法の残差プロット

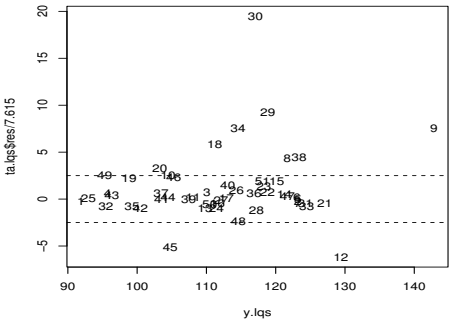

図 4　LQS 法の残差プロット