ロバスト線形回帰

2001MM029 金子　元紀
指導教員　木村　美善

1　はじめに

回帰分析を行うときに一般的に用いられる最小二乗法の問題点を克服するためにロバスト (頑健) 回帰が誕生し，発展した．本研究では回帰分析のテキストや論文を読み，実際の分析を通して理論的結果を確認していくことを目的とする．

2　線形回帰

2.1　線形回帰モデル

目的変数 y が，p 個の説明変数 x_1,$\cdots$,x_p,ε を誤差とし，n個の観測値が与えられたとき，それらは次のように表す事が出来る．

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \quad (1)$$

2.2　最小二乗法 (Least squared method:LS)

最小二乗法とは，残差平方和が最小になるように，推定値 $\hat{\beta}_0$,$\cdots$,$\hat{\beta}_p$ を定める方法である．i 番目の予測値を

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi} \quad (2)$$

としたとき，実測値 y_i との残差 (residual) は

$$e_i(\hat{\beta}) = y_i - \hat{y}_i \quad (3)$$

と表せることが出来る．

3　ロバスト回帰

3.1　ロバスト回帰とは

最小 2 乗法は線形回帰の標準的仮定のずれに敏感で，1 つの外れ値によっても大きな影響を受けてしまう．ロバスト回帰とはそれら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．

3.2　ロバストネスの尺度

3.2.1　影響関数 (influence function)

ロバストネスの研究をする上で重要な役割を果たすものが，Hampel(1968) によって導入された影響関数である ([4] 参照)．回帰推定量 T の G における影響関数とは

$$IF(x;T,G) = \lim_{t\to 0} \frac{T((1-t)G + t\delta_x) - T(G)}{t} \quad (4)$$

によって定義される x の実数値関数である．これは T の G における δ_x 方向への方向微分であり，点xでの微小な汚染が T に及ぼす影響の程度を表している．

3.2.2　漸近効率 (asymptotic efficient)

T に対して漸近正規性

$$L_G(\sqrt{n}(T_n - T(G))) \Rightarrow N(0, V(T,G))$$

が成り立つとき, 漸近分散 V(T,G) は IF によって

$$V(T,G) = \int IF(x;T,G)^2 dG(x) \quad (5)$$

と表される．分布 F における T_n の漸近効率は

$$e = \frac{\frac{1}{J(F)}}{V(T,F)} = \frac{1}{V(T,F)J(F)} \quad (6)$$

となり，0≤ e ≤ 1 の間の値をとる．漸近分散 V(T,F) が小さく $J(F)^{-1}$ に近い程 e は大きくなることから，e が 1 に近い程 T_n は望ましい ([1] 参照).

3.2.3　破綻点 (breakdown point)

グローバルな信頼性をはかる尺度として破綻点がある ([5] 参照).

$$T(Z) = \hat{\beta} \quad (7)$$

これは T を用いて Z から回帰係数のベクトルを求めることを意味する．元のデータの中m個を，任意の値（かなり悪い外れ値を考慮にいれる）に置き換えたときのデータを Z' とする．汚染によって生じる偏りの最大は,

$$bias(m;T,Z) = sup_{z'} \|T(Z') - T(Z)\| \quad (8)$$

となる.$bias(m;T,Z)$ が無限であるとき，これを推定量の破綻という．有限標本 z での推定量 T の破綻点は

$$\{\varepsilon_n\}^* = min\{m/n; bias(m;T,Z) = \infty\} \quad (9)$$

となる．$0 \le \varepsilon^* \le 1/2$ であり，高い破綻点が望ましい．

3.3　M 推定量

[5] の論文より，位置母数におけるM推定量は，ψ を実軸上の実数値関数 ρ の導関数としたとき,

$$\sum_{i=1}^{n} \psi(x_i - T_n) = 0 \quad (10)$$

を満たす T_n を β のM推定量であるという ([3] 参照)．線形回帰分析モデルに一般化すると,

$$\sum_{i=1}^{n} \eta(x_i, (y_i - x_i' T_n)/\sigma) x_i = 0 \quad (11)$$

•Huber 推定量

$$\psi_k(r) = r \cdot min(1, k/\mid r \mid) \qquad 0 < k < \infty \quad (12)$$

この M 推定量は誤差項に関してはロバストであるが，x に関してはロバストではなく，次元が増えると破綻点が低くなる．

3.4 LMS 推定量と LTS 推定量

LMS 推定量と LTS 推定量は高い破綻点を目的とした推定量である ([2] 参照).

3.4.1 LMS 推定量

Rousseeuw によって導入されたＬＭＳ推定量は，残差の 2 乗値の中央値を最小にする推定量である．この LMS 推定量は x の外れ値と同様に y の外れ値に関してロバストである．

$$Minimize_{\hat{\beta}} med_i e_i^2(\hat{\beta}) \tag{13}$$

3.4.2 LTS 推定量

Rousseeuw は，上のＬＭＳに手を加えたＬＴＳを提案した．これは $h = [n/2] + 1$ 番目までの残差平方和を最小にするというものである．

$$Minimize_{\hat{\beta}} \sum_{i=1}^{h} (e_i(\hat{\beta}))^2_{i:n} \tag{14}$$

4 単回帰分析・重回帰分析

4.1 単回帰分析

ここで用いるデータは 2003 年に南山大学数理科学科に入学した学生の一部の統計的方法　と統計的方法　の成績である．x = 統計的方法　，y = 統計的方法　とする．

4.1.1 結果・考察

$$\hat{y}_{LS} = -39.2693 + 0.5959x$$
$$\hat{y}_{M:Huber} = 34.5504 + 0.7101x$$
$$\hat{y}_{LMS} = -4.375 + 1.750x$$
$$\hat{y}_{LTS} = -0.1677 + 1.6207x$$

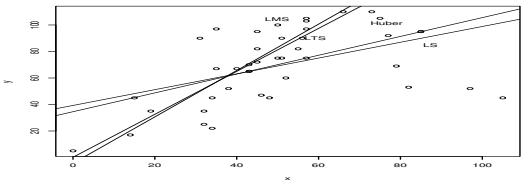

図 1　LS，Huber，LMS，LTS プロット図

LS は外れ値に引っ張られ傾きが浅くなり，M 推定量も x 方向の外れ値に引っ張られている．LMS，LTS はほぼデータの密集しているところを通っている．

4.2 重回帰分析

ここで用いるデータは製造プラントで使用される水の月使用量，生産量, プラントでの月稼働人のデータである．

x_1 = 生産量 (1000 ポンド)，x_2 = プラントでの月稼働人数，$x_3 = y$ = 水の月使用量とする．

4.2.1 結果・考察

$$\hat{y}_{LS} = 4024.7949 + 0.1790x_1 - 16.7534x_2$$
$$\hat{y}_{M:Huber} = 3579.5984 + 0.1254x_1 - 10.8534x_2$$
$$\hat{y}_{LMS} = 1012.9966 - 0.05584x_1 + 14.97592x_2$$
$$\hat{y}_{LTS} = 3442.8997 + 0.1015x_1 - 8.4100x_2$$

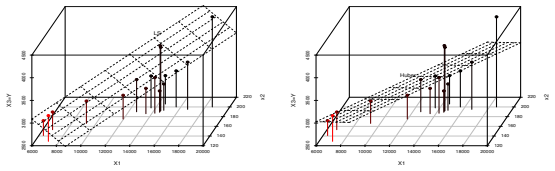

図 2　LS，Huber プロット図

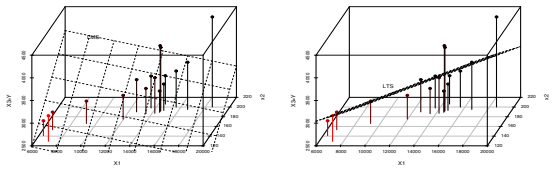

図 3　LMS，LTS プロット図

LS は外れ値に引っ張られているが，ロバスト推定量の方は外れ値に引っ張られていないことが確認出来る．

5 おわりに

今回の研究で特に大きな作業は理論の理解とデータの収集の 2 つだった．特にデータ解析の中でロバスト回帰の特性を示すとき，誰が見ても良さや問題点が明らかになるようなデータを探すのにはとても苦労をした．今回の研究を進めるにつれてロバスト回帰の利点や必要性を更に深く感じるようになった．今後は更に深い基礎知識と幅広い応用の仕方を学んでいきたい．

6 謝辞

本論文を作成するにあたり，多くの助言を賜りご指導していただいた木村美善教授，安藤雅和氏，その他協力して頂いたすべての方に深く感謝いたします．

参考文献

[1] 安藤雅和：線形回帰モデルにおけるロバスト推定量の研究, 南山大学経営学研究科修士論文 (1996).

[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] Huber,P.J.: Finite sample breakdown of M- and P-estimators,The Annals of Statistics,Vol.12,No 1,119-126 (1984).

[4] 木村美善：位置母数のロバスト推定, 南山経営研究, 第 3 巻, 第 1 号 (1988).

[5] Rousseeuw,P.J.and Leroy A.M.:Robust Regression and Outlier Detection, Wiley,New York (1986).