ロバスト回帰の研究
ーMM 推定量と τ 推定量を中心にー

2012SE244 田川 聖也

指導教員 木村 美善

1 はじめに

最小二乗法で推定した回帰係数は正規分布からのずれや外れ値に影響されるため，ずれや外れ値がある場合はロバスト回帰が望ましい．ロバスト回帰推定量はこれまで多く提案されており，その中にロバスト (頑健) 性と漸近効率の両立に成功した MM 推定量と τ 推定量がある．本研究の目的はロバストネスに関する理解を深め，この 2 つの推定量の違いを明確化することである．

2 線形回帰モデル

2.1 モデルの定式化

応答変数を y，p 個の説明変数を x_1，x_2，$\cdots$，x_p，誤差を ε とする．n 個の観測値が与えられたとき重回帰モデルは

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

と表すことができる．ここで $\beta_0, \beta_1, \cdots, \beta_p$ は回帰係数.

2.2 最小二乗推定量 (Least squares estimator)

$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ の最小二乗推定量 (LS 推定量) は (1) 式における残差平方和を最小にする $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ である．y_i の残差を $r_i(\hat{\boldsymbol{\beta}}) = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})$ とする．

3 ロバスト回帰推定量

3.1 破綻点 (breakdown point)

回帰推定量を $\boldsymbol{T}$ としたとき, 元データ $\boldsymbol{Z}$ から回帰係数を求めると $\boldsymbol{T}(\boldsymbol{Z}) = \hat{\boldsymbol{\beta}}$ となる．元の n 個のデータのうち m 個を任意の値 (非常に悪い外れ値を含む) に置き換えたデータを $\boldsymbol{Z'}$ とするとき，その汚染によって引き起こされる最大バイアスは

$$bias(m; \boldsymbol{T}, \boldsymbol{Z}) = \sup_{Z'} \parallel \boldsymbol{T}(\boldsymbol{Z'}) - \boldsymbol{T}(\boldsymbol{Z}) \parallel \tag{2}$$

である．$\boldsymbol{Z}$ における有限標本破綻点は

$$\varepsilon_n^*(\boldsymbol{T}, \boldsymbol{Z}) = \min\{m/n; bias(m; \boldsymbol{T}, \boldsymbol{Z}) = \infty\} \tag{3}$$

と定義される．$bias(m; \boldsymbol{T}, \boldsymbol{Z})$ が無限であるとき, 推定量 $\boldsymbol{T}$ は破綻することを意味する ([2] 参照).

3.2 漸近効率

推定量 $\hat{\boldsymbol{\beta}}$ の漸近効率は LS 推定量の漸近分散との分散比で与えられ，正規分布の下では

$$\mathrm{eff}(\hat{\boldsymbol{\beta}}) = \frac{(E_{\Phi}(\psi'(r_i/\sigma)))^2}{E_{\Phi}(\psi(r_i/\sigma)^2)} \tag{4}$$

となる．(4) 式の値が 1 に近ければその推定量は漸近効率が高いといえる．正規分布の下で LS 推定量の漸近効率は 1 として考えられる．

3.3 S 推定量の定義

S 推定量は Rousseuw and Yohai (1984) により提案されたもので，頑健性の高い推定量である．損失関数 ρ は $(-\infty, \infty)$ 上の有界関数であり，Tukey の biweight 損失関数は

$$\rho(t) = \begin{cases} 3(t/c)^2 - 3(t/c)^4 + (t/c)^6 & |\frac{t}{c}| \leq 1 \\ 1 & |\frac{t}{c}| > 1 \end{cases} \tag{5}$$

によって定義される．ただし $\rho(t)' = \psi(t)$ である．$\rho_0(t)$=$\rho(t/c_0)$ とする．標本が n 個のとき残差を $\boldsymbol{r}(\hat{\boldsymbol{\beta}}) = (r_1(\hat{\boldsymbol{\beta}}), r_2(\hat{\boldsymbol{\beta}}), \cdots, r_n(\hat{\boldsymbol{\beta}}))^T$ とし

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{r_i}{s_n(\boldsymbol{r})}\right) = b, \quad 0 \leq b \leq 1 \tag{6}$$

を満たす尺度の推定値を $s_n(\boldsymbol{r})$ とする．$s_n(\boldsymbol{r}(\hat{\boldsymbol{\beta}}))$ を最小にする $\hat{\boldsymbol{\beta}}$ が S 推定量である．調整定数 c_0 の値によって推定量の破綻点や効率が変化する．r が標準正規分布 $N(0, 1)$ に従うとき漸近的に

$$E_{\Phi}(\rho_0(r/s)) = b \tag{7}$$

が成り立つ．(6) 式を満たす尺度の推定値 s_n は反復重み付き最小二乗法 (IRLS : Iteratively Reweighted LS) によって求めることができる．S 推定量は $n \to \infty$ のとき, $b \fallingdotseq \varepsilon_n^*$ となるので $b = 0.5$ となるように c_0 を変動させる．Tukey の biweight の場合 $c_0 = 1.547$ となる ([1]，[4] 参照).

3.4 MM 推定量

MM 推定量は Yohai (1987) により提案されたもので (6) 式の s_n を用いる．$\rho_1(t)$=$\rho(t/c_1)$ とし

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{\boldsymbol{s_n}}\right) \tag{8}$$

を考える．この $S(\boldsymbol{\beta})$ を最小にする $\hat{\boldsymbol{\beta}}$ が MM 推定量である．すなわち $S(\boldsymbol{\beta})$ を $\boldsymbol{\beta}$ で微分した方程式

$$\sum_{i=1}^{n} \psi_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right)\boldsymbol{x_i} = 0 \tag{9}$$

の解 $\hat{\boldsymbol{\beta}}$ が MM 推定量となる．ここで $\rho_1(t)' = \psi_1(t)$ である．c_0 の値は頑健性を高めるが，c_1 の値は漸近効率を高めるように選ばれる．このように 2 段階で高い頑健性と高い漸近効率を保つようにする．Tukey の biweight の場合 $c_1 = 4.685$ のとき漸近効率が 0.95 となる ([4] 参照).

3.5 τ 推定量

高い破綻点と高い漸近効率を与える推定量として MM 推定量のほかに Yohai and Zamar (1988) によって導入された τ 推定量がある．τ 推定量は次の $\tau_n(\boldsymbol{\beta})$ を最小にする $\hat{\boldsymbol{\beta}}$ として定義される．

$$\tau_n^2(\boldsymbol{\beta}) = s_n^2(\boldsymbol{\beta})\frac{1}{n}\sum_{i=1}^{n}\rho_2(r_i/s_n(\boldsymbol{\beta})) \tag{10}$$

ただし s_n は (6) 式の尺度の推定値であり，$\rho_2(t)$=$\rho(t/c_2)$．$\frac{d\tau_n^2(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = 0$，$\frac{ds_n(\boldsymbol{\beta})}{d\boldsymbol{\beta}}$ を解くと

$$\psi_n^*(r_i/s_n(\boldsymbol{\beta})) = W_n\psi_0(r_i/s_n(\boldsymbol{\beta})) + \psi_2(r_i/s_n(\boldsymbol{\beta})) \tag{11}$$

$$W_n = \frac{\sum_{i=1}^{n}[2\rho_2(r_i/s_n(\boldsymbol{\beta})) - \psi_2(r_i/s_n(\boldsymbol{\beta}))r_i/s_n(\boldsymbol{\beta})]}{\sum_{i=1}^{n}\psi_0(r_i/s_n(\boldsymbol{\beta}))r_i/s_n(\boldsymbol{\beta})} \tag{12}$$

が得られる．ただし $\rho_2(t)' = \psi_2(t)$ である．したがって τ 推定量は 2 段階で構成した (9) 式のような M 推定量と考えることができる ([3] 参照)．

4 分析結果

ここでは自作した S 推定量のプログラムを用いる．調整定数を変動させて $c_0 = 1.547$, 2.142, 2.621 とし，LS 推定量も加えて単回帰分析を行う．データは $n = 13$ で 1931 年から 1943 年 1 月の Libby(x) と Newgate(y) での水流の記録を用いる ([2] 参照)．

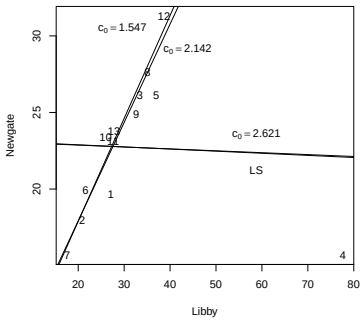

図 1: 調整定数を変化させた S 推定量と LS 推定量の比較

$c_0 = 1.547$ は b が 0.5 となる値であり，外れ値に影響されないが $c_0 = 2.621$ になると 4 番の外れ値に影響され，回帰式が LS 推定量と似た結果になる．

表 1: 調整定数を変化させた $E_\Phi(\rho_0(r/s)) = b$

c_0	1.547	1.984	2.253	2.561	3.421	4.092	4.685
b	0.500	0.401	0.350	0.300	0.200	0.150	0.120

$n \to \infty$ のとき，$b \fallingdotseq \varepsilon_n^*$ となるので表 1 より b の値は c_0 の値が高くなるにつれて低くなることが分かる．

5 MM 推定量と τ 推定量の違い

MM 推定量は $c_1 = 4.685$ の ψ_1，τ 推定量は $c_0 = 1.547$, $c_2 = 6.081$ とした (11) 式の ψ_n^* を漸近的にみた ψ^* を使うと表より漸近効率が 0.95 となる．また表 3 より τ 推定量は c_0 に対して c_2 の値を高くし効率を高めようとすると W が小さくなり c_0 の影響を減らしていることが分かる．

表 2: MM の漸近効率

c_1	eff
1.547	0.287
1.984	0.458
2.253	0.560
2.561	0.661
3.421	0.847
4.092	0.917
4.685	0.949

表 3: τ の漸近効率

c_0	c_2	eff	W
1.547	1.547	0.287	1.569
1.547	2.253	0.429	0.777
1.547	3.421	0.704	0.242
1.547	4.685	0.878	0.081
1.547	5.246	0.917	0.053
1.547	5.813	0.942	0.036
1.547	6.081	0.951	0.031

IRLS を用いて第 4 節のデータの尺度の推定値 s_n を求めるとき s_n が収束する様子を考察する (τ 推定値 [1] 参照)．横軸は k(IRLS での繰り返し数)，縦軸は k 回毎の尺度の値 s である．図 2 は $c_0 = 1.547$ として (6) 式を解く場合のものである．MM 推定量はこの収束値 s_n が 2 段階で変動しない．図 3 は初期値を s_n，$c_0 = 1.547$，$c_2 = 6.081$ として τ 推定値を求める場合のものである．図 3 より τ 推定値は s_n が 2 段階で変動することが分かる．また τ 推定値は s_n を求める手順を 2 回行ったものであるといえる．

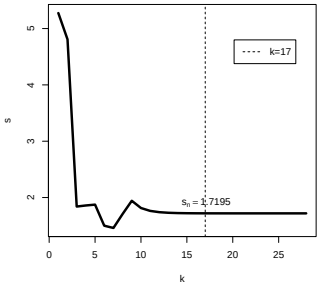

図 2: 尺度推定値 s_n

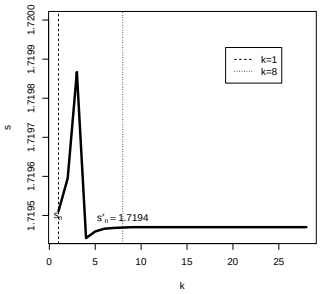

図 3: 尺度推定値 (τ の場合)

6 おわりに

理解を深めるため，S 推定量や尺度を求めるプログラムを実装し，それを応用することで MM 推定量，τ 推定量の違いをより明確に把握することができた．

参考文献

[1] 室梅秀平 : ブートストラップ法とそのロバスト推定への応用，南山大学大学院数理情報研究科修士論文 (2012)．

[2] Rousseeuw P. J. and Leroy A. M. : Robust Regression and Outlier Detection, John Wiley & Sons(1987).

[3] Yohai, V. J. and Zamar, R. H. : High breakdown-point estimates of regression by means of minimization of an efficient scale, Journal of the American Statistical Association, Vol. 83, 406-413(1988).

[4] Yohai, V. J. : High breakdown-point and high efficiency robust estimates for regression, The Annals of Statistics, Vol. 15, 642-656(1987).