線形回帰におけるロバスト推定量の研究

M2005MM013 金子　元紀
指導教員　木村　美善

1　はじめに

線形回帰モデルの中心的な課題の一つは回帰係数の推定であるが, 一般的に用いられる最小２乗法による推定量は正規分布が仮定される場合にはすべての不偏推定量の中で最良である. この仮定を満たした『理想のモデル』からわずかなずれが生じるとき, そのずれに過剰に反応し, 結論を大きく変えるならば, 統計分析の方法として好ましくない. こうした危険を回避し, 標準的仮定からのずれや外れ値に対して影響が小さいロバスト (頑健) な推定量を用いることが望ましい. 回帰モデルの外れ値は不適切なモデル, 誤差項の不均一分散, 突発的原因あるいは正規分布より長い尾をもつ分布から発生しうるが, ロバスト推定の必要性が最も高まるのは最後のケースである. 現実の諸問題において標準的仮定はせいぜい近似的に成り立つ程度であるから, 実用的見地からすれば推定量のロバストネス (頑健性) は重要な問題である. 本研究は, ロバスト線形回帰モデルの理論と手法について整理するとともに, 観測値データへの適用を通してロバスト線形回帰手法の有効性と問題点を探っていくものである.

2　線形回帰モデル

2.1　線形回帰モデル

回帰パラメータを β, 目的変数を y，p 個の説明変数を x_1,$\cdots$,x_p, 誤差を ε とし，n 個の観測値に対して，次のモデルを考える.

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

ここで ϵ_i は互いに独立で正規分布 $N(0, \sigma^2)$ に従うと仮定する.

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. y_i の予測値は

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

で与えられる. また実測値 y_i の残差を $e_i = y_i - \hat{y}_i$ とする.

3　ロバストネス

3.1　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 ([3] 参照). 回帰推定量 T の G における影響関数とは

$$IF(x; T, G) = \lim_{t \to 0} \frac{T((1-t)G + t\delta_x) - T(G)}{t} \quad (3)$$

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. 影響関数から得られるロバストネスの尺度として次の３つが挙げられる.

3.1.1　gross-error sensitivity

$$\gamma^*(T, G) = \sup_x |IF(x; T, G)| \quad (4)$$

これは微小な汚染によって T が受ける最大の影響を表す. $IF(x; T, G)$ が存在するならば x に関して上限を取ることが出来る. γ*が小さいほどよい.

3.1.2　local-shift sensitivity

T の G における local-shift sensitivity は

$$\lambda^*(T, G) = \sup_{x \neq y} \frac{|IF(x; T, G) - IF(y; T, G)|}{|x - y|} \quad (5)$$

これは x から y への変化に対する影響関数の最大平均変化率を表す. ただし $|x - y|$ で割っているので無限大になっても影響関数の変化がさほどでないことがある.

3.1.3　rejection point

$$\rho^*(T, G) = \inf\{r > 0 | IF(x; T, G) = 0, |x| > r\} \quad (6)$$

r が存在しない時は $\rho^*(T, G) = \infty$ とする. ρ^* より大きな x に対して影響関数は 0 をとるため x は T に影響を与えないことになる. これは x(外れ値) が完全に除去されることを意味する.

3.2　漸近効率 (asymptotic efficient)

T に対して漸近正規性

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

が成り立つとき, 漸近分散 $V(T, G)$ は IF によって

$$V(T, G) = \int IF(x; T, G)^2 dG(x) \quad (7)$$

と表される. 分布 F における T_n の漸近効率をフィッシャー情報量 $J(F)$ で表すと

$$e = \frac{\frac{1}{J(F)}}{V(T, F)} = \frac{1}{V(T, F)J(F)} \quad (8)$$

となり，0≤ e ≤ 1 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).

3.3 偏り (bias)

n 個の標本点を

$$Z = \{(x_{11}, ..., x_{k1}, y_1), ..., (x_{1n}, ..., x_{kn}, y_n)\} \quad (9)$$

とし,Z から得られる β の推定量を $\hat{\beta(Z)}$ とする. n 個の観測点の中から m 個の観測点を任意の値にとりかえることによって得られる標本を Z' とする. この m 個の汚染された値, すなわち外れ値による推定量の変化の最大の大きさを $bias(m; \hat{\beta(Z)})$ と書くと

$$bias(m; \hat{\beta(Z)}) = \sup_{z'} \parallel \hat{\beta(Z')} - \hat{\beta(Z)} \parallel \quad (10)$$

と表せる.

3.4 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 ([3] 参照). 推定量 T の変化の最大の大きさ $bias(m; T, Z)$ が無限であるとき, これを推定量の破綻 (breakdown) という. 有限標本 Z での推定量 T の破綻点は

$$\varepsilon_n{}^* = \min\{m/n; bias(m; T, Z) = \infty\} \quad (11)$$

として定義される. $0 \leq \varepsilon_n{}^* \leq 1/2$ であり, 高い破綻点が望ましい.

4 ロバスト推定量

4.1 M 推定量

線形回帰モデルにおける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

$$\sum_{i=1}^{n} \psi\Big(\frac{y_i - x_i'\theta}{\sigma}\Big) x_i = 0 \quad (12)$$

を解くことによって与えられる ([4] 参照). ここで ρ は微分可能かつ 0 のまわりで対称な凸関数とする.

•Huber の ψ

$$\psi(u) = \begin{cases} -c, & u < -c \\ u, & |u| \leq c \\ c, & u > c \end{cases} \quad (13)$$

•Tukey の ψ (双加重 (biweight))

$$\psi(x) = \begin{cases} u[1 - (\frac{u}{c})^2]^2, & |u| \leq c \\ 0, & u > c \end{cases} \quad (14)$$

この他に Andrews の ψ 関数や Cauchy の ψ などがある. 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.

4.2 GM 推定量

M 推定量の影響関数は, 誤差項に関しては有界であったが, 説明変数に関しては有界ではなかった. 影響関数が誤差項と説明変数の両方に対して有界であることが望ましい. 作用点となる外れ値 x_i の影響を有界にするために, 重み関数 w を用いた GM 推定量が考案された ([5] 参照). Mallows(1975) が提案した GM 推定量は

$$\sum_{i=1}^{n} w(x_i)\psi((y_i - x_i'T_n)/\sigma)x_i = 0 \quad (15)$$

を満たすものであるが、GM 推定量 T_n については $\epsilon_n(T_n, X) \leq 1/(p+1)$ なので p が無限大となるとき破綻点は 0 となる.

4.3 LMS 推定量

Rousseeuw(1984) はさらに高い破綻点を得るためにＬＭＳ推定量を提案した. これは

$$med_i e_i^2(\hat{\theta}_{LMS}) = \min_{\theta} med_i e_i^2(\theta) \quad (16)$$

により定義され,Hampel(1975) のアイデアに基づいたものである. LMS 推定量の破綻点は $([n/2] - p + 2)/n$ であり,$n \to \infty$ のとき 1/2 となる.

4.4 LTS 推定量

Rousseeuw(1985) は, 上の LMS 推定量に手を加えた LTS 推定量を提案した. 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.

$$\sum_{i=1}^{h} (e_i(\hat{\theta}_{LTS}))^2_{i:n} = \min_{\theta} \sum_{i=1}^{h} (e_i(\theta))^2_{i:n} \quad (17)$$

有限標本破綻点は LMS 推定量と同じであり, 収束のオーダーも $n^{-1/2}$ であるが目的関数の計算ステップ数が LMS 推定量に比べて多い.

4.5 S 推定量

M 推定量の柔軟性と漸近的性質の良さを保持しながらも高い破綻点をもち, LMS や LTS よりも高い漸近効率をもつことを狙ったのが Rousseeuw and Yohai (1984) による S 推定量である ([6] 参照).

$$s(\hat{\theta}_s) = \min_{\theta} s(\theta) \quad (18)$$

を満たすものとして定義される. ここで $s(\theta)$ は

$$\frac{1}{n} \sum_{i=1}^{n} \rho(e_i(\theta)/s(\theta)) = b \quad (19)$$

を満たすものであり,ρ は $(-\infty, \infty)$ 上の有界な関数で有界, 原点対象,$(0, \infty)$ 上で非減少, $\rho(0) = 0$ であり,k はある定数, すなわち尺度 $s(\theta)$ を推定した後, この $s(\theta)$ を最小にする θ を推定量とするものである.

4.6 2 段階 S 推定量

Rousseeuw and Yohai (1984) は 2 段階 S 推定を導入している ([6] 参照). Tukey の双加重を例にとる. 第 1 段階で調整定数を 50 %の破綻点を与える 1.548 に設定し, 次の式を繰り返し加重最小 2 乗法 (IRLS) で $\hat{\theta}$ を求める.

$$\frac{1}{n} \sum_{i=1}^{n} \psi\Big(\frac{y_i - x'\hat{\theta}}{\hat{\sigma}}\Big) x_i = 0 \quad (20)$$

次の式を $\hat{\sigma}^2$ を求めるための繰り返し型で表す.

$$[\hat{\sigma}^{m+1}]^2 = \frac{1}{(n-k-1)E_\Phi(\rho)} \sum_{i=1}^{n} \rho(r_i^{(m)})[\hat{\sigma}^{(m)}]^2 \quad (21)$$

$$r_i^{(m)} = \frac{Y_i - x_i'\hat{\theta}^{(m)}}{\hat{\sigma}^{(m)}} \quad (22)$$

第 2 段階では $\hat{\sigma}$ を第 1 段階の収束値で固定し、調整定数を高い有効性を与える値, 95 %の漸近効率ならば c=4.685 に設定する. この第 2 段階は $\hat{\sigma}$ を高い頑健性が得られる第 1 段階の値で固定したもとでの M 推定に他ならない. この 2 段階 S 推定は本質的に MM 推定 (Yohai,1987) と同じである.

4.7 τ 推定量

Yohai and Zamar (1988) によって導入された τ 推定量は次の式の解として与えられる ([7] 参照).

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} \psi(\frac{y_i - x_i'\theta}{\sigma})x_i = 0 \\ \frac{1}{n}\sum_{i=1}^{n} \rho_1(\frac{y_i - x_i'\theta}{\sigma}) = b_1 \end{cases} \quad (23)$$

ここで ρ_1,ρ_2 は M 推定量と同様, 微分可能かつ,0 まわりで対称的な凸関数とする.

$$W_n = \frac{\sum_{i=1}^{n}[2\rho_2(r_i) - \psi_2(r_i)r_i]}{\sum_{i=1}^{n}\psi_1(r_i)r_i} \quad (24)$$

が得られ, さらに ρ_2 は

$$2\rho_2(u) - \psi_2(u)u \geq 0 \quad (25)$$

を満たすとすれば $W_n \geq 0$ であり、θ の τ 推定量は

$$\psi(u) = W_n\psi_1(u) + \psi_2(u) \quad (26)$$

を ψ 関数としてもつ M 推定量と考えることができる. よって ρ_1 に高い破綻点を与える関数, ρ_2 に高い効率を与える関数を選ぶことで高い破綻点と有効性をもつ推定量を得ることができる.

4.8 GS 推定量

S 推定量よりも高い漸近効率を求めて Croux et al.(1994) は GS(generalized S) 推定量を考案した ([2] 参照). これは S 推定量と同様に尺度の M 推定量を最小にする $\hat{\theta}$

$$s_n(\hat{\theta}_{GS}) = \min_\theta s_n(\theta) \quad (27)$$

として定義する. ここで $s_n(\theta)$ はある定数 k に対して

$$\binom{n}{2}^{-1} \sum_{i<j} \rho\Big(\frac{e_i - e_j}{s_n(\theta)}\Big) = k_{n,p} \quad (28)$$

を満たすものである.

4.9 LQD 推定量

Croux,Rousseeuw and Hossjer(1994) は GS 推定量の特別な場合として, LQD(Least quartile difference) 推定量を提案した. LQD 推定量は

$$Q_n(\hat{\theta}_{LQD}) = \min_\theta Q_n(\theta) \quad (29)$$

によって定義される. ここで e_i は i 番目の残差であり,Q_n は以下のような尺度推定量である.

$$Q_n^{(\theta)} = \{|e_i - e_j|; i < j\}_{\binom{h_p}{2}:\binom{n}{2}} \quad (30)$$

ここで $h_p = [(n+p+1)/2]$ とおくと Q_n は集合 $\{|e_i - e_j|; i < j\}$ の $\binom{n}{2}$ 個の要素を小さい順に並べた場合の $\binom{h}{2}$ 番目の要素の値となる.

GS 推定量において ρ 関数と $k_{n,p}$ を

$$\rho(x) = I(|x| \geq 1) \quad (31)$$

$$k_{n,p} = \left[\binom{n}{2} - \binom{h_p}{2} + 1\right] \Big/ \binom{n}{2} \quad (32)$$

と置くと,$s_n(\theta) = Q_n(\theta)$ となり,LQD 推定量は GS 推定量の特別な場合であることがわかる.

5 ロバスト推定法の適用例

5.1 重回帰における各手法の比較

用いるデータは 1973 年 5 月 1 日から 31 日までの空気の質に関するデータである. Y(空気の密度) について X_1(観測番号),X_2(日光の周波数),X_3(平均風速),X_4(温度) の 4 つの説明変数をもとに関係を探るものである. 最小 2 乗法を使用したときの回帰式は次のようになった.

$$\hat{Y}_1 = -72.48 + 0.399X_1 - 0.112X_2 - 2.045X_3 + 1.742X_4$$

寄与率は 0.48 と低く, あてはまりの悪い結果となった. よって残差分析から外れ値を把握するため LS のスチューデント化残差の大きな 3,21,23,30 を除いた. 除いたデータに再び最小 2 乗法を用いる.

$$\hat{Y}_2 = -35.39 - 0.031X_1 - 0.022X_2 - 1.153X_3 + 1.156X_4$$

しかし, この回帰式の寄与率は 0.46 となり改善はみられない. そこで τ と GS のスチューデント化残差を用いて外れ値を把握することを試みる. 回帰式は次のようになっている.

$$\hat{Y}_\tau = -62.83 - 0.075X_1 + 0.002X_2 - 0.27X_3 + 1.31X_4$$
$$\hat{Y}_{GS} = -98.78 - 0.27X_1 + 0.004X_2 + 0.18X_3 + 1.83X_4$$

τ と GS のスチューデント化残差の値が共通して特に大きいところは 1,3,22,24,30 である. この 5 つが本当に回帰式の当てはめの支障となっているのかを確認するため, この 5 つを除き再び最小 2 乗法を用いる.

$$\hat{Y}_3 = -90.95 - 0.150X_1 + 0.007X_2 + 0.381X_3 + 1.637X_4$$

となり, 寄与率は約 7 割の値を算出し回帰式がデータの大部分を表していることを示している.

6 シミュレーション

R 上で推定量がどこまで信頼出来るのかをシミュレーション実験を通じて確認する. 説明変数 X は 0 から 10 の一様乱数を用いて生成し, 目的変数 $Y1$ は $3\times X+1+\epsilon_1$ で生成する. ϵ_1 は平均 0, 分散 1 の正規乱数である. 目的変数 $Y2$ を作るために $3\times X+1+\epsilon_2$ を作る. ϵ_2 は平均 8, 分散 1 の正規乱数である. $Y2$ を汚染されたデータと仮定する. 汚染されたデータと汚染されていないデータを合わせたものをつくる. LS,M,LMS,LTS,S,MM, τ ,GS の 8 つの推定量で切片と傾きを算出し, 傾きと切片が共に, 汚染がないデータでの,LS の 95% 信頼区間内であるときカウント 1 を加える. 標本数 n に対する汚染の割合 m は,0,10,20,30,40,49% とし, 計算回数は 1 万回とする. 結果はそれぞれ % で表示している.

表 1　n=20

m	LS	M	LMS	LTS	S	MM	τ	GS
0	96	95	64	67	77	95	92	82
10	39	87	66	69	81	94	85	85
20	10	61	66	69	84	85	81	90
30	0.1	0.1	68	73	87	57	80	89
40	0	0	69	78	81	0.2	52	71
49	0	0	61	76	0.4	0	0.2	35

表 2　n=100

m	LS	M	LMS	LTS	S	MM	τ	GS
0	93	92	43	41	62	92	92	69
10	0.1	73	46	43	69	91	80	78
20	0	21	48	47	73	88	76	82
30	0	0	51	53	79	65	78	83
40	0	0	50	62	81	7	36	79
49	0	0	40	79	0.4	0	0	40

表 3　n=1000

m	LS	M	LMS	LTS	S	MM	τ	GS
0	87	86	22	22	52	86	85	60
10	0	4	24	26	58	83	71	68
20	0	0	24	31	63	81	67	73
30	0	0	24	39	67	40	67	76
40	0	0	24	49	71	0	15	69
49	0	0	16	64	0	0	0	47

6.1 考察

傾向として, 標準的仮定をほぼ満たしていると予測されるデータのときに有効なロバスト推定量は M 推定量,MM 推定量, τ 推定量であることが確認できる. これらの推定量は漸近効率が 95% を達成する推定量である. また, 汚染の割合が多少あると予測されるデータの場合,MM 推定量と GS 推定量が有効的であると考えられる. そして, 汚染の割合がかなり高いと予測されるデータの場合,GS 推定量と LTS 推定量が有効なロバスト推定量であると今回の実験から考えることができる.

7 おわりに

これまで, ロバスト推定量は M,GM 推定量が一般的とされ,S 推定量などは, その高い頑健性のあまり M 推定量などに比べ使われることが少なかった. 本研究では, これまで計算機上で簡単に使用できなかった GS 推定量の使用を可能にしただけでなく, 既存のロバスト推定量との比較を行い, τ 推定量と GS 推定量が今後, 一般的な推定量となる可能性を示すこともできた. また, ロバスト回帰の理論上での有用性は勿論のこと, 実データを用いた解析においてもその有用性を示すことが出来たのではないかと思う. 日本語によるロバスト手法をまとめた文献が極端に少ないなかで本研究が多少とも貢献できたら幸いに思う.

謝辞

本論文を作成するにあたり, 熱心にご指導していただいた木村美善教授, 安藤雅和先生, その他協力して頂いたすべての方に深く感謝致します.

参考文献

[1] 安藤雅和 (1996). 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.

[2] Croux,C.,Rousseeuw,P.J. and Hossjer,O.(1994). Generalized S-estimators,Journal of the American Statistical Association,Vol.89,1271-1281.

[3] Hampel,F.R.,Ronchetti,E.M.,Rousseeuw,P.J. and Stahel,W.A.(1986). Robust Statistics:The Approach Based on Influence Functions,Wiley,New York.

[4] Huber,P.J.(1984). Finite sample breakdown of M- and P-estimators,The Annals of Statistics,Vol.12,119-126.

[5] 蓑谷千凰彦 (1992). 計量経済学における頑健推定, 多賀出版.

[6] 田中辰雄, 中妻照雄 (2006). 計量経済学のフロンティア, 慶應義塾大学出版会.

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