回帰分析の理論と応用
-最深回帰推定量を中心に-

M2006MM028 澤田　謹志
指導教員　木村　美善

1　はじめに

回帰分析では, 一般に最小 2 乗法による推定量が用いられることが多い．この推定量は, 標準的仮定のもとでは線形不偏推定量の中で最良である．さらに, 正規分布の仮定のもとでは, すべての不偏推定量の中で最良になる．しかし, 最小 2 乗推定量は線形回帰の標準的仮定からのずれに対して敏感であり, 外れ値による影響を受けやすい．こうした問題点を解決する方法として, これまでにさまざまなロバスト推定量が提案されてきた．その中のひとつに Regression Depth の概念に基づく最深回帰推定量がある．本論文の目的は, 線形回帰モデルにおける回帰係数の推定問題に対して, Rousseeuw and Hubert [6] により提案された最深回帰推定量の特徴と有効性について, 理論とシミュレーションの両面から明らかにすることである．

2　線形回帰

2.1　回帰モデル

n 個の観測値が与えられたとき, 目的変数を y, p 個の説明変数を $x_1, \cdots, x_p$ とすると, 回帰式は次のように表すことができる．

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

ただし $\beta_0, \cdots, \beta_p$ を回帰係数, ϵ_i を誤差とする．

2.2　最小 2 乗法 (Least squared method : LS)

最小 2 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

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

としたとき, 実測値 y_i との残差は $r_i(\hat{\beta}) = y_i - \hat{y}_i$ と表される．

3　Regression(回帰) Depth

3.1　重回帰の場合

まず単回帰と同様に重回帰における不適合 $(nonfit)$ を定義する．データ集合 $Z_n = \{(x_{i1}, \cdots, x_{i,p-1}, y_i); i = 1, \cdots, n\} \subset R^p$ に対し, y_i を

$$\theta_1 x_{i1} + \cdots + \theta_{p-1} x_{i,p-1} + \theta_p = (x_i^t, 1)\theta$$

すなわち R^p におけるアフィン超平面であてはめる．ここで $\theta = (\theta_1, \cdots, \theta_p)^t \in R^p$, $x_i = (x_{i1}, \cdots, x_{i,p-1})^t \in R^{p-1}$ とする．

定義 1　どの x_i も属さないアフィン超平面 V が存在し, 次の 1 かつ 2 が成り立つとき,　$\theta = (\theta_1, \cdots, \theta_p)$ は Z_n に対して不適合 $(nonfit)$ という．

1.　$r_i(\theta) > 0$, V の一方の開半空間.
2.　$r_i(\theta) < 0$, V のもう一方の開半空間.

図 3 は, 3 次元データ集合における不適合の例である．

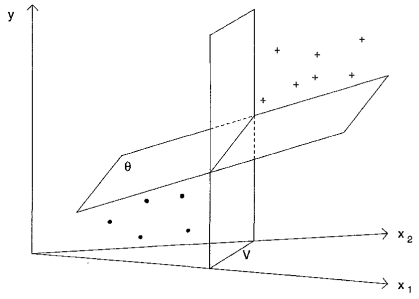

図 1　$nonfit$　$\theta \in R^3$ の例

ここで x 空間は, $y = 0$ によって与えられた水平面とみなすことができる．また, アフィン超平面 V は残差が正になる観測値と負になる観測値を分離する．

定義 2　データ集合 $Z_n \subset R^p$ に対する $\theta \in R^p$ の $rdepth(\theta, Z_n)$ は次のように定義される．

$$rdepth(\theta, Z_n)$$
$$= \min_{u,v} \{\sharp(r_i(\theta) \geq 0 \text{ かつ } x_i^t u < v) + \sharp(r_i(\theta) \leq 0 \text{ かつ } x_i^t u > v)\}$$

これは, 超平面 θ を垂直になるまで傾けるとき, 通る必要がある観測値の最少数に等しいということができる．

定義 2 によって k 個の点を通るすべての推定量は, 少なくとも $rdepth = k$ $(0 \leq k \leq n)$ をもつ．

定理 1　(exact fit property)　θ 上にある観測値の数が k $(0 \leq k \leq n)$ ならば, そのとき

$$k \leq rdepth(\theta, Z_n) \leq \left[\frac{n+k}{2}\right]$$

ここで, $[\lambda]$ は λ 以下の最大の整数を表す．

また $k = n$ のとき, $rdepth(\theta, Z_n) = n$ となることに注意する.

次に確率分布に対する $rdepth$ を定義する.

定義 3 R^p 上の分布 H に対する θ の $rdepth(\theta, H)$ は

$$rdepth(\theta, H) = \min_{u,v}\{H(y - (x^t, 1)\theta > 0 \text{ かつ } x^t u < v) + H(y - (x^t, 1)\theta < 0 \text{ かつ } x^t u > v)\}$$

によって定義される. ここで H は確率変数 (x^t, y) の分布であり, 最小値は全ての単位ベクトル $u = (u_1, \cdots, u_{p-1})^t \in R^{p-1}$ と, $H(x^t u = v) = 0$ を満たす全ての $v \in R$ でとられるものとする. これは, $rdepth(\theta, H)$ が, v を中心に垂直になるまで超平面 θ を任意の方向に傾けるとき, 通る必要がある確率の最小値として定義される.

定理 2 Z_n が密度関数をもつ分布 H からの標本とするとき

$$\frac{rdepth(\theta, Z_n)}{n} \xrightarrow[n\to\infty]{a.s.} rdepth(\theta, H).$$

3.2 最大 $rdepth$

定理 3

1. (x_i^t, y_i) が general position にある, すなわちどの $p-1$ 次元アフィン部分空間にも p 点以上の観測値がないならば

$$\max_\theta rdepth(\theta, Z_n) \le \left[\frac{n+p}{2}\right].$$

2. 密度関数をもつ R^p 上の分布 H のに対して

$$\max_\theta rdepth(\theta, H) \le \frac{1}{2}.$$

3. H が密度関数をもち, ある $\tilde{\theta} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_p)^t \in R^p$ に対して

$$med\,[y \mid x] = \tilde{\theta}_1 x + \cdots + \tilde{\theta}_{p-1} x_{p-1} + \tilde{\theta}_p$$

を満たすならば

$$\max_\theta rdepth(\theta, H) = rdepth(\tilde{\theta}, H) = \frac{1}{2}.$$

4 最深回帰推定量

データ集合 Z_n に対する p 次元における最深回帰推定量 $\mathrm{DR}(Z_n)$ は, $rdepth(\theta, Z_n)$ を最大にする θ と定義する. すなわち

$$\mathrm{DR}(Z_n) = \arg\max_\theta rdepth(\theta, Z_n).$$

ただし, $rdepth(\theta, Z_n)$ を最大にする θ は一つとは限らない. 同じ最大値をとる θ が複数ある場合は, それらの平均をとる. また, 1 変量データのとき, 任意の $\theta \in R$ に対して

$$rdepth(\,\theta\,, Z_n) = \min\,(\sharp\,\{y_i \le \theta\},\, \sharp\,\{y_i \ge \theta\}).$$

したがって, $\mathrm{DR}(Z_n)$ は y_i の中央値になる.
最深回帰推定量は定義に分布の仮定を必要とせず, 回帰共変, 尺度共変, アフィン共変推定量である.

5 最深回帰推定量のロバストネス

5.1 有限標本破綻点 (breakdown point)

任意のデータ集合 Z_n に対して回帰推定量を T_n とすると

$$T_n(Z_n) = \hat{\theta}$$

と表される. これは T_n を用いて Z_n から回帰係数ベクトルを求めることを意味する.

推定量が外れ値に対してどれくらい強いか（外れ値に対するロバストネス）を評価するために, 破綻点と呼ばれる指標が用いられる ([4] 参照).

n 個のデータ集合 Z_n に, 汚染された m 個の値を加えたものを Z_{n+m} とする. これにより生じるバイアス (偏り) の最大値が無限大になるとき, m 個の外れ値は T_n に対して大きな影響をもち, 推定量 T_n は破綻する. この値を追加破綻点と定義する:

$$\epsilon_n^*(T_n, Z_n) = \min\left\{\frac{m}{n+m};\ \sup_{Z_{n+m}} \parallel T_{n+m}(Z_{n+m}) - T_n(Z_n) \parallel = \infty\right\}$$

最深回帰推定量の破綻点は常に正であるが, 元のデータ自身に異常がある場合, $\frac{1}{p+1}$ という低い値をとる.

定理 4 密度関数をもつ分布 $H \subset R^p$ $(p \ge 2)$ からの標本を $Z_n = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ とする. $med\,(y \mid x) = (x^t, 1)\tilde{\theta}$ を満たすような $\tilde{\theta} \in R^p$ が存在すると仮定すると,

$$\epsilon_n^*(\mathrm{DR}, Z_n) \xrightarrow[n\to\infty]{a.s.} \frac{1}{3}.$$

となる ([2] 参照).

5.2 影響関数

分布 H における推定量 T の影響関数は, $z = (x^t, y)$ に小さい確率を加えることで T への影響を測るものである ([4] 参照).
Δ_z によって z で確率 1 をもつ確率分布を表し, $H_\epsilon = (1-\epsilon)H + \epsilon\Delta_z$ と書くとき, 影響関数は

$$\begin{aligned} IF(z, T, H) &= \lim_{\epsilon\to 0} \frac{T((1-\epsilon)H + \epsilon\Delta_z) - T(H)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{T(H_\epsilon) - T(H)}{\epsilon} \\ &= \frac{\partial}{\partial\epsilon} T(H_\epsilon)|_{\epsilon=0} \end{aligned}$$

により定義される.

p=2 次元において, 最深回帰推定量 $T_1^* = (T_1^*, T_2^*)^t$ の影響関数を得る. ここで T_1^* は傾き, T_2^* は切片である.

定理 5　2 変量標準正規分布 $H = N_2(0, I)$ に対して

$$IF((x,y), T_1^*, H) = \frac{sgn(x)sgn(y)}{2\phi(0)}\Big(\frac{I(\phi(x)) \geq \phi(0)/3)}{4\phi(x)} + \frac{I(\phi(x) < \phi(0)/3}{\phi(0)+\phi(x)}\Big),$$

$$IF((x,y), T_2^*, H) = \frac{sgn(y)}{2\phi(0)}\Big(\frac{I(|x| \leq \Phi^{-1}(\frac{2}{3}))}{\Phi(|x|)} + \frac{I(|x| > \Phi^{-1}(\frac{2}{3}))}{2(2\Phi(|x|)-1)}\Big).$$

ここで, Φ は 1 変量正規分布 $N(0,1)$ の分布関数であり, ϕ はその密度関数である.

5.3　感度関数

影響関数は母集団分布で定義されているので, その有限標本版の影響関数と比較する. 任意の推定量 T_n に対する感度関数は, 標本 $Z_n = \{z_1, \cdots, z_n\}$ に一つの観測値 $z = (x, y)$ を加えることによる影響を測る. すなわち,

$$SF_n(z, T, Z_n) = n(T_{n+1}(z_1, \cdots, z_n, z) - T_n(z_1, \cdots, z_n))$$

感度関数は実際の標本 Z_n に強く依存するが, 置換型標本 $Z_n(\pi) = \{(x_i^s, x_{\pi(i)}^s); i = 1, \cdots, n\}$ を用いることでこの影響を軽減することができる. ここで $x_i^s = \Phi^{-1}(\frac{i}{n+1})$, π は $\{1, \cdots, n\}$ 上の置換を表す. この変換された標本は無作為標本の場合よりも母集団分布 $N_2(0, I)$ に近い分布をし, 周辺分布は中央値 0 で対称になるという利点をもつ. 特定の置換 π の効果は順列の集合上で感度関数を平均化した

$$APSF_n(z) = \frac{1}{n!}\sum_{\pi} SF_n(z, T, Z_n(\pi))$$

によって調整される ([1] 参照).

6　ロバスト推定量

6.1　LMS 推定量

LMS 推定量 $\hat{\theta}_{LMS}$ は, Hampel (1975) の考えに基づいて Rousseeuw (1984) によって提案され

$$\hat{\theta}_{LMS} = \arg\min_{\theta} med_i r_i^2(\theta)$$

で定義される. 残差の 2 乗の中央値を最小にするこの推定量は, x 方向の外れ値と同様に y 方向の外れ値に関してロバストである.
破綻点は $\frac{[n/2]-p+2}{n}$ で, $n \to \infty$ のとき $\frac{1}{2}$ となる.

6.2　LTS 推定量

LTS 推定量 $\hat{\theta}_{LTS}$ は, LMS 推定量に手を加えたものとして Rousseeuw (1984) によって提案され

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

で定義される. 残差の 2 乗値を小さい順に並べた順序統計量の h 番目までの和を最小にする. 最小 2 乗法によく似ているが, 大きな残差が和に含まれないため外れ値の影響を回避することができる.
破綻点は $\frac{[(n-p)/2]+1}{n}$ で, h が $\frac{n}{2}$ に近いところで $\frac{1}{2}$ となる.

6.3　相対効率

線形回帰の標準的仮定を満たしているデータを用い, LS と他の推定量との相対効率を求める. 正規分布から m=10000 となるようなサンプル $Z_n = \{(x_i, y_i);\ i = 1, \cdots, n\} \subset R^2,\ j = 1, \cdots, m$ を生成し, 各回帰推定量の傾きと切片の分散を計算した. それらに対し LS の傾き, 切片それぞれの分散比をとったものを相対効率とする.

表 1　LS との有限標本相対効率

n		LMS	LTS	DR
20	傾き	19.4%	22.1%	37.8%
	切片	21.6%	23.2%	55.2%
40	傾き	17.9%	18.5%	37.4%
	切片	18.3%	17.5%	61.2%
60	傾き	17.4%	15.8%	39.3%
	切片	16.4%	15.4%	62.5%
80	傾き	16.7%	14.5%	39.5%
	切片	15.1%	14.1%	63.4%
100	傾き	16.1%	15.1%	39.5%
	切片	14.1%	13.2%	61.3%
500	傾き	11.1%	9.5%	41.3%
	切片	8.7%	8.6%	64.5%

表を見ると, LMS, LTS に比べ, DR の効率は高い値を示しているだけでなく, データ数に関わらずその値を維持していることがわかる.

7　シミュレーション

Regression Depth を用いた手法の性能を評価するために, シミュレーションを通して推定量の平均二乗誤差を比較する ([3][7] 参照). モデルを $y = \theta_1 x_1 + \cdots + \theta_{p-1} x_{p-1} + \theta_p + \epsilon$, $(\theta_1, \cdots, \theta_p) = \mathbf{0}$ とし, 外れ値の割合を 5%, 30% と変化させた. データ数は 20, 40, 60, 80, 100 と設定し, 繰り返し回数は 10000 回とした. 比較する推定量には, LS 推定量のほかに 2 つのロバスト推定量 (LMS 推定量, LTS 推定量) を用いる.
ここで平均二乗誤差 (MSE) は

$$\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

として計算する. なお最深回帰推定量には近似プログラムである MEDSWEEP を使用する ([5] 参照).
p=3, 誤差分布を正規分布 $N_1(0,1)$ とし, シミュレーションを行った結果を以下に示す.

表 2　MSE の比較 (正規分布 $N_2(0,25)$：5 %)

n		LS	LMS	LTS	DR
20	傾き	1.996	0.339	0.286	0.199
	切片	1.811	0.266	0.229	0.126
40	傾き	0.916	0.154	0.142	0.077
	切片	0.830	0.142	0.134	0.050
60	傾き	0.533	0.102	0.100	0.048
	切片	0.562	0.101	0.099	0.032
80	傾き	0.434	0.079	0.083	0.036
	切片	0.407	0.083	0.079	0.024
100	傾き	0.345	0.064	0.068	0.028
	切片	0.324	0.069	0.071	0.019

表 3　MSE の比較 (正規分布 $N_2(0,25)$：30 %)

n		LS	LMS	LTS	DR
20	傾き	11.820	0.367	0.231	1.371
	切片	10.440	0.294	0.183	0.681
40	傾き	5.135	0.142	0.124	0.179
	切片	4.760	0.131	0.106	0.102
60	傾き	3.358	0.094	0.085	0.101
	切片	3.202	0.091	0.081	0.063
80	傾き	2.556	0.070	0.066	0.072
	切片	2.408	0.072	0.063	0.047
100	傾き	1.979	0.056	0.054	0.052
	切片	1.885	0.058	0.050	0.034

表 4　MSE の比較 (t 分布：5 %)

n		LS	LMS	LTS	DR
20	傾き	0.081	0.352	0.268	0.174
	切片	0.070	0.282	0.225	0.112
40	傾き	0.051	0.151	0.146	0.071
	切片	0.076	0.149	0.138	0.047
60	傾き	0.023	0.104	0.102	0.044
	切片	0.024	0.106	0.103	0.029
80	傾き	0.020	0.076	0.083	0.032
	切片	0.021	0.082	0.083	0.021
100	傾き	0.017	0.065	0.070	0.026
	切片	0.017	0.068	0.072	0.017

表 5　MSE の比較 (t 分布：30 %)

n		LS	LMS	LTS	DR
20	傾き	0.231	0.365	0.297	0.198
	切片	0.222	0.287	0.225	0.116
40	傾き	0.104	0.152	0.149	0.078
	切片	0.085	0.138	0.133	0.050
60	傾き	0.063	0.102	0.105	0.050
	切片	0.060	0.100	0.096	0.032
80	傾き	0.053	0.077	0.082	0.035
	切片	0.054	0.076	0.083	0.023
100	傾き	0.042	0.065	0.069	0.028
	切片	0.045	0.063	0.067	0.017

7.1　考察

正規分布 $N_2(0,25)$ を外れ値として 5% 含む場合を見ると, データ数の大きさに関わらず DR の MSE が一番小さな値をとっている.

外れ値が全体の 30% になると, LS は他の推定量に比べ非常に大きな値をとってしまっている. DR の MSE は, n=20 のときにやや大きな値をとってはいるものの, データ数が増えるにつれて MSE は小さくなり n=100 のときに最小値をとっている. LMS, LTS については, 外れ値が 5% のときよりも 30% のときのほうが MSE は小さくなった.

次に自由度 2 の t 分布を外れ値として含む場合について考察する. 外れ値が 5% のときはデータ数の大きさに関わらず, 4 つの推定量のなかで LS の MSE が最小となっている. そして次に小さな MSE をとっているのは DR である.

外れ値を 30% 含むときをみると, 5% のときに比べて全体的に大きな値をとっている LS に対し, 外れ値の割合に関係なく安定した値を示した DR の MSE が最小となっている.

8　おわりに

本研究では, Regression Depth の概念を基に, 最深回帰推定量の定義や性質についてまとめ, ロバストネスを考察した. これらから最深回帰推定量のもつ高い破綻点と, 効率の低下が少ないことが確認できた. また, 最深回帰推定量とその他の推定量との比較をデータ解析とシミュレーションを通して行い, 分析結果から最深回帰推定量が実用的かつ有効であることを示すことができた. 最深回帰推定量による推定を行う際に MEDSWEEP という近似プログラムを用いたが, 今後の課題として, 近似ではなく真の係数値を推定し, より多くの実データを用いた解析と合わせて評価を行いたい.

参考文献

[1] Aelst,S.V. and Rousseeuw,P.J.(2000). Robustness of deepest regression, *Journal of Multivariate Analysis*, **73**, 82-106.

[2] Aelst,S.V., Rousseeuw,P.J., Hubert,M.and Struyf, A.(2002).The Deepest Regression Method,*Journal of Multivariate Analysis*, **81**,138-166.

[3] Fujiki,M.(2005). A Study on Theory of Regression Depth and Its Application, 大阪大学博士論文.

[4] 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.

[5] 大見俊司 (2006). 最深回帰推定量の研究, 南山大学数理情報研究科修士論文.

[6] Rousseeuw,P.J. and Hubert,H.(1999). Regression Depth, *Journal of the American Statistical Association*, **94**,388-402.

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