最深回帰推定量の研究

M2005MM020 大見　俊司
指導教員　木村　美善

1　はじめに

本研究では Regression Depth の概念に基づく最深回帰推定量の性質を調べる. また,Fortran でのみ使用可能であった最深回帰推定量の計算用プログラム Medsweep を統計解析システム R で使用できるように書き直し, そのプログラムを使用して, 他の回帰推定量との比較研究を行う.

2　Regression Depth

任意の次元 p に対して n 個のデータを $Z_n = \{\boldsymbol{z}_i = (x_{i1},,\cdots,x_{i,p-1},y_i); i = 1,\cdots,n\} \subset \mathbb{R}^p$ とする. この Z_n に式 $y = \theta_1 x_1 + \cdots + \theta_{p-1}x_{p-1} + \theta_p, \boldsymbol{\theta} = (\theta_1,\cdots,\theta_p)^t \in \mathbb{R}^p$ を当てはめたい. $\boldsymbol{x}_i = (x_{i1},\cdots,x_{i,p-1})^t \in \mathbb{R}^{p-1}$ によって $\boldsymbol{z}_i$ の x の部分を表す. $r_i = r_i(\boldsymbol{\theta}) = y_i - \theta_1 x_{i1} - \cdots - \theta_{p-1}x_{i,p-1} - \theta_p$を $\boldsymbol{z}_i$ の残差とする.

2.1　Regression Depth の定義と性質

定義 1 $\boldsymbol{x}$ 空間上でどの $\boldsymbol{x}_i (i = 1,\cdots,n)$ も属さない超平面 V が存在し,V で分けた開半空間の片方に属する全ての $\boldsymbol{x}_i$ に対して $r_i(\boldsymbol{\theta}) > 0$ であり, もう一方の開半空間に属する全ての $\boldsymbol{x}_i$ に対して $r_i(\boldsymbol{\theta}) < 0$ ならば $\boldsymbol{\theta} = (\theta_1,\cdots,\theta_p)$ は Z_n に対して $nonfit$ と呼ばれる.

定義 2 データ $Z_n \subset R^p$ に対する $\boldsymbol{\theta}$ の $rdepth(\boldsymbol{\theta}, Z_n)$ は $\boldsymbol{\theta}$ を $nonfit$ にするために取り除く必要がある観測値の最小数であり,$rdepth(\boldsymbol{\theta}, Z_n)$ は次のように定義される.

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_{\mathbf{u},v}\{\sharp(r_i(\boldsymbol{\theta}) \geq 0 \text{ かつ } \mathbf{x}_i^t\mathbf{u} < v) + \sharp(r_i(\boldsymbol{\theta}) \leq 0 \text{ かつ } \boldsymbol{x}_i^t\mathbf{u} > v)\} \quad (1)$$

ただし,$(\boldsymbol{x}_i^t, y_i) \in Z_n$ に対して, 最小は $\boldsymbol{x}_i^t\mathbf{u} \neq v$ を満たす全ての単位ベクトル $\mathbf{u} = (u_1,\cdots,u_{p-1})^t \in \mathbb{R}^{p-1}$と ,$v \in \mathbb{R}$ でとられるものとする. この定義により, データ $Z_n \subset \mathbb{R}^p$ に対する $\boldsymbol{\theta} \in \mathbb{R}^p$ の $rdepth$ は v を中心に垂直になるまで超平面 $\boldsymbol{\theta}$ を傾ける時, 通過しなければならない観測値の最少数であるともいえる.

定理 1 (Exact Fit Property) $\boldsymbol{\theta}$ 上にある観測値の数が k $(0 \leq k \leq n)$ ならば, そのとき

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

よって,$k = n$ のとき $rdepth(\boldsymbol{\theta}, Z_n) = n$ となる. ここで $[\lambda]$ は λ 以下の最大の整数である. 次に確率分布に対する $rdepth$ を定義する.

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

$$rdepth(\boldsymbol{\theta}, H) = \min_{\boldsymbol{u},v}\big\{H(y - (\boldsymbol{x}^t, 1)\boldsymbol{\theta} > 0 \text{ かつ } \boldsymbol{x}^t\boldsymbol{u} < v) + H(y - (\boldsymbol{x}^t, 1)\boldsymbol{\theta} < 0 \text{ かつ } \boldsymbol{x}^t\boldsymbol{u} > v)\big\} \quad (3)$$

によって定義される. ここで H は確率変数 $(\boldsymbol{x}^t, y)$ の分布であり, 最小は $H(\boldsymbol{x}^t\boldsymbol{u} = v) = 0$ を満たす全ての単位ベクトル $\mathbf{u} = (u_1,\cdots,u_{p-1})^t \in \mathbb{R}^{p-1}$ と $v \in \mathbb{R}$ でとられるものとする. $rdepth(\boldsymbol{\theta}, H)$ は v を中心に垂直になるまで超平面 $\boldsymbol{\theta}$ を傾ける時, 通過しなければならない部分の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 からの標本のとき

$$\frac{rdepth(\boldsymbol{\theta}, Z_n)}{n} \xrightarrow[n\to\infty]{a.s.} rdepth(\boldsymbol{\theta}, H). \quad (4)$$

2.2　最大 $rdepth$

定理 3

a. $(\boldsymbol{x}_i^t, y_i)$ が general position(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) にあるとき

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) \leq \left[\frac{n+p}{2}\right]. \quad (5)$$

b. 密度関数をもつ $\mathbb{R}^p$ 上の任意の分布 H に対して

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) \leq \frac{1}{2}. \quad (6)$$

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

$$med[y|\boldsymbol{x}] = \tilde{\theta}_1 x_1 + \cdots + \tilde{\theta}_{p-1}x_{p-1} + \tilde{\theta}_p \quad (7)$$

を満たすならば

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) = rdepth(\tilde{\boldsymbol{\theta}}, H) = \frac{1}{2}. \quad (8)$$

次に $rdepth$ の下界を与える. しかし, 証明されているのは 2 次元のときだけで 3 次元以上はまだ証明されておらず, 予想が与えられているのみである (Rousseeuw and Hubert(1999) 参照). 以降はこの予想が正しいものとして話を進める.

推測 1

a. 任意のデータ $Z_n \subset \mathbb{R}^p$ に対して

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) \geq \left\lceil\frac{n}{p+1}\right\rceil. \quad (9)$$

ここで $\lceil\lambda\rceil$ は λ 以上の最小の整数である.
b. 密度関数をもつ $\mathbb{R}^p$ 上の任意の分布 H に対して

$$\max_{\theta} rdepth(\boldsymbol{\theta}, H) \geq \frac{1}{p+1}. \tag{10}$$

3 最深回帰推定量

定義 4 p 次元における最深回帰推定量 $DR(Z_n)$ は $rdepth(\boldsymbol{\theta}, Z_n)$ を最大にする $\boldsymbol{\theta}$ と定義する. すなわち

$$DR(Z_n) = \arg\max_{\boldsymbol{\theta}} \ rdepth(\boldsymbol{\theta}, Z_n). \tag{11}$$

データに対する最深回帰推定量は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. $\max_{\boldsymbol{\theta}} \ rdepth(\boldsymbol{\theta}, Z_n)$ を与える $\boldsymbol{\theta}$ が複数ある場合はそれら $\boldsymbol{\theta}$ の平均を推定量とする.
分布 H に従う p 次元確率変数 $(\boldsymbol{x}^t, y)$ に対して, 最深回帰推定量 $DR(H)$ を次のように定義する.
定義 5

$$DR(H) = \arg\max_{\boldsymbol{\theta}} \ rdepth(\boldsymbol{\theta}, H) \tag{12}$$

ここで分布 H は狭義に正の密度関数を持ち

$$med_H(y|\boldsymbol{x}) = (\boldsymbol{x}^t, 1)\tilde{\boldsymbol{\theta}} \tag{13}$$

を満たす $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ が存在すると仮定する.
このモデルは誤差の分布が非対称であったり, 異なった分散であったりする場合にも有効である. 次の定理は Aelst and Rousseeuw(2000) によるもので, 誤差分布がノンパラメトリックであり,H が大きなセミパラメトリックモデル $\mathcal{H}$ に属するとき, 最深回帰推定量 $DR(H)$ が $\tilde{\boldsymbol{\theta}}$ の Fisher-consistent 推定量であることを示す.
定理 4 (Fisher-consistency) 任意の $H \in \mathcal{H}$ に対して,$DR(H) = \tilde{\boldsymbol{\theta}}$ が成り立つ.
Bai and He(1999) によって示された $\tilde{\boldsymbol{\theta}}$ に対する最深回帰推定量 DR の一致性と定理 4 の Fisher-consistency から $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$ が独立で同一の分布 $H \in \mathcal{H}$ に従うとき,$DR(H_n) = DR_n(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)$ は $DR(H)$ に確率収束する.H_nは $z_1, \cdots, z_n$の経験分布関数を表す.

4 最深回帰推定量のロバストネス

4.1 有限標本破綻点

推測 1 の系 推測 1 が成り立ち,$\boldsymbol{x}_i$ が general position にあるならば

$$\varepsilon_n^*(DR, Z_n) \geq \frac{1}{n}\Big(\Big\lceil\frac{n}{p+1}\Big\rceil - p + 1\Big) \approx \frac{1}{p+1}. \tag{14}$$

ここで $\approx$ は $n \to \infty$ のときの極限値であり,$\varepsilon_n^*(T, Z_n)$ はデータ Z_nに対する推定量 T の有限標本破綻点である. 最深回帰推定量 DR の ε_n^* はもとのデータ Z_n がそれ自身異常なとき $\frac{1}{p+1}$ になる.
定理 5 $Z_n = \Big\{(\boldsymbol{x}_1^t, y_1), \cdots, (\boldsymbol{x}_n^t, y_n)\Big\}$ が狭義に正の密度関数をもつ $\mathbb{R}^p (p \geq 2)$ 上の分布 H からの標本であり,H が (13) を満たすならば

$$\varepsilon_n^*(DR, Z_n) \xrightarrow[n\to\infty]{a.s.} \frac{1}{3}. \tag{15}$$

4.2 影響関数

2 次元の最深回帰推定量 $DR = (DR_1, DR_2)^t$ の影響関数を導く. ただし,DR_1 は傾き DR_2 は切片である. 最深回帰推定量は回帰共変, 尺度共変, アフィン共変なので球形分布 $H = H_{\mathbf{0},I}$ における影響関数を導けばよい.
定理 6 最深回帰推定量の H=$H_{\mathbf{0},I}$ における影響関数は

$$IF((x,y), DR_1, H) = sgn(x)sgn(y) \times \Bigg(\frac{I(G(|x|) \leq 2G(+\infty)/3)}{4[G(+\infty) - G(|x|)]} + \frac{I(G(|x|) \geq 2G(+\infty)/3)}{[2G(+\infty) - G(|x|)]} \Bigg),$$

$$IF((x,y), DR_2, H) = \frac{sgn(y)}{2h_Y(0)} \times \left(\frac{I(H_{X|Y}(|x| \mid 0) \leq \frac{2}{3})}{H_{X|Y}(|x| \mid 0)} + \frac{I(H_{X|Y}(|x| \mid 0) \geq \frac{2}{3})}{2(2H_{X|Y}(|x| \mid 0) - 1)} \right)$$

である. ただし,$G(t) = \int_0^{t^2} g(u)du, h_Y$ は Y の周辺密度関数,$H_{X|Y}$は $Y = 0$ が与えられたもとでの X の条件付累積分布関数である.
図 1 は 2 変量標準正規分布 $H = N_2(\mathbf{0}, I)$ における最深回帰推定量の傾きの影響関数であり, 図 2 は $H = N_2(\mathbf{0}, I)$ における最深回帰推定量の切片の影響関数である.

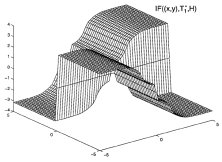

図 1　傾きの影響関数

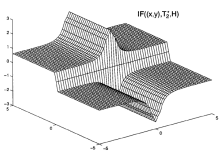

図 2　切片の影響関数

4.3 感度関数

影響関数は母集団分布上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ために, 平均置換型感度関数 (the averaged permutation-stylized sensitivity function) を計算する. 任意の推定量 T_n に対する感度関数は標本 $Z_n = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n\}$ に一つの観測値 $\boldsymbol{z} = (x, y)$ を加えることによる影響を測る. すなわち,

$$SF_n(\boldsymbol{z}, T, Z_n) = n\big(T_{n+1}(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \boldsymbol{z}) - T_n(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)\big). \tag{16}$$

感度関数は実際の標本 Z_n に強く依存するので置換型標本 $Z_n(\pi) = \{(x_i^s, x_{\pi(i)}^s); i = 1, \cdots, n\}$ を使うことでこの影響を軽減する. ここで $x_i^s = \Phi^{-1}(\frac{i}{n+1})$ であり, 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 $N_2(\mathbf{0}, I)$ に近い分布をし, 周辺分布は中央値 0 に関して対称になるという利点をもつ. 特定の置換 π の効果は非復元抽出リサンプリングで感度関数を平均化した

$$APSF_n(\boldsymbol{z}) = \frac{1}{B}\sum_B SF_n(\boldsymbol{z}, T, Z_n(\pi)) \qquad (17)$$

によって和らげられる. ここで,B は非復元抽出リサンプリングの繰り返し回数である. これを用いてデータ数 20, 格子点の数 2500, 繰り返し回数 3000 で計算すると, 傾きと切片の $APSF$ 図はそれぞれ図 1, 図 2 と似ていたので, 影響関数のロバストネスは小標本に対しても有効であるといえるだろう.

4.4 漸近効率

漸近効率とは最小 2 乗推定量が有効推定量であるとき, この漸近分散に対して比較したい統計量の漸近分散を比較したものである. He and Portnoy(1998) によって最深回帰推定量は正規分布からわずかに異なる極限分布を持つことが証明された. よって最深回帰推定量は漸近正規ではないので, 漸近分散による, 漸近効率を測ることができない. しかし, 極限分布が正規分布に似ているのでシミュレーションにより, 近似の漸近効率を求めると傾き 40%, 切片 64% となった.

5 プログラム

Rousseeuw et al.(2002) によって最深回帰推定量の近似プログラム Medsweep が Fortran で書かれた. Fortran は古い言語であるため, 統計解析システム R で使用できるようにプログラムを書き換えた.

5.1 プログラムのパフォーマンス

与えられた n, p に対して正規分布から $m = 10000$ のサンプル $\boldsymbol{Z}^{(j)} = \{(x_{i1}, \cdots, x_{i,p-1}, y_i); i = 1, \cdots, n\}, j = 1, \cdots, m$ を生成する. それらのサンプルに対してそれぞれ Medsweep アルゴリズムの最深回帰推定量 $(\hat{\theta}_1^{(j)}, \cdots, \hat{\theta}_p^{(j)})$ を計算する.

$$MSE(\hat{\theta}_1, \cdots, \hat{\theta}_{p-1}) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{p-1}\sum_{i=1}^{p-1}(\hat{\theta}_i^{(j)} - \theta_i)^2 \qquad (18)$$

ここで真値は $\theta_i = 0; i = 1, \cdots, p$ である. 切片の MSE は $\frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_p^{(j)})^2$ である. それぞれの n と p に対して切片と傾きの平均 2 乗誤差 (MSE) を計算したものを表 1 に載せる. データ数が増加すると MSE が減少し, 次元が増加すると MSE は増加しており, 異常値もないのでこのプログラムに欠陥はないだろう. しかし, データ数が 20 のとき MSE は大きい. これは Medsweep が p 次元に対して p 点通らなければならず, データの影響を強く受けるからである.

表 1　切片と傾きの $MSE(\times 10^{-3})$

		p			
n	MSE	2	3	5	10
20	切片	95.39	113.12	150.06	453.95
	傾き	147.33	166.42	233.64	5095.47
50	切片	33.31	32.68	36.89	44.80
	傾き	53.64	51.57	52.57	59.18
100	切片	16.36	16.66	17.32	17.99
	傾き	25.03	25.28	26.42	27.08
500	切片	3.18	3.23	3.21	3.22
	傾き	4.93	4.92	4.98	4.97
1000	切片	1.57	1.55	1.57	1.61
	傾き	2.53	2.46	2.47	2.49

6 比較

最深回帰推定量 (Deepest) は回帰モデルの誤差分布が互いに独立で, 各誤差分布の中央値を 0 と仮定するだけでよい. これらはとても弱い条件である. 誤差分布が対称であることを仮定する必要がなく, 同一の分布であることを仮定する必要もない. また, このモデルは歪んだ誤差分布や分散が均一でなくてもよい. 他のロバスト回帰推定量は最深回帰推定量よりも多くの制約を必要とし, より制限されたモデルを仮定する. 実際, これらの推定量は歪んだ誤差分布や分散の不均一性を許してはいない.

6.1 単回帰分析

データは Chatterjee et al.(2000)(1986 年の広告枚数と広告収入,p177) から引用する. 説明変数を広告枚数 P(百枚), 目的変数を広告収入 R(百万ドル) とする. ロバスト推定量による回帰直線は 23 番の観測値の影響を受けずにいる. 外れ値 (1,2,23) を抜いた LS(LS2) の Shapiro-Wilk normality test の $p-$ 値 は 0.0706 なので残差の正規性は否定されない. したがって,LS2 と似た直線を引いている S はよい推定をしているだろう. Deepest と Catline はほぼ同じ回帰直線となった. そして, これらの回帰直線は LS2 に似ているが若干傾きが大きい. これは左上の観測値の影響を受けているからである. この広告収入データに対してどの推定量が一番適しているかについては, 外れ値を抜いた LS の結果から残差に正規性があるならば S 推定量が一番適しているかもしれない. しかし, スチューデント化残差図を見ると

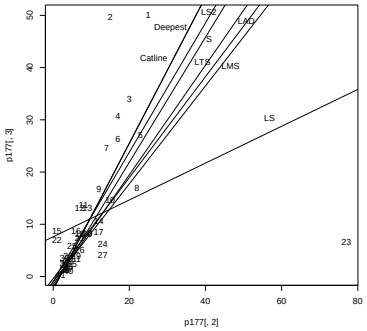

図 3　広告データに対する回帰直線図

等分散性があるとは言い切れないので, 残差に正規性や等分散性の仮定を必要としない Deepest と Catline が適しているだろう.

6.2　相対効率と破綻点の比較

線形回帰の標準的仮定を満たしているデータを用いて LS と他の推定量との相対効率を求める. 与えられた n に対して正規分布から 2 次元の $m = 10000$ のサンプル $\boldsymbol{Z}^{(j)} = \{(x_i, y_i); i = 1, \cdots, n\}, j = 1, \cdots, m$ を生成する. それらのサンプルに対して各回帰推定量の切片と傾きの分散を計算し,LS の切片と傾きの分散との相対効率を測ると表 2 のようになった. LAD は効率は高いが有限標本破綻点は低い. LMS,LTS,S は効率は低いが有限標本破綻点は高い. Deepest は傾きの効率は若干低いが切片の効率は高い. 有限標本破綻点は LMS,LTS,S より低いが LAD よりも外れ値の影響を受けない.

7　終わりに

最深回帰推定量は有限標本破綻点, 影響関数, 相対効率から残差に正規性が認められるデータに対しても, 外れ値のあるデータに対しても良い推定ができるだろう. 他のロバスト回帰推定量との大きな違いは残差の中央値が 0 であれば, それ以外は非対称だろうが分散が違っていようが良い性質をもつことである. 実際に単回帰分析では等分散性のないデータを用いて最深回帰推定量の良さを確認することができた. また, 本研究の課題であった統計解析システム R で最深回帰推定量の近似プログラムを作成することができたので, 今後このプログラムが回帰分析を行う人の役に立てばうれしく思う.

表 2　LS との有限標本相対効率

切片の有限標本相対効率					
n	LAD	LMS	LTS	S	Deepest
20	67.0%	21.5%	23.4%	36.7%	56.0%
50	63.3%	17.0%	16.3%	30.0%	61.6%
100	63.6%	13.9%	12.9%	29.5%	61.7%
500	62.4%	8.9%	8.7%	28.1%	62.3%
1000	63.1%	7.1%	8.0%	28.5%	64.0%
傾きの有限標本相対効率					
n	LAD	LMS	LTS	S	Deepest
20	63.1%	19.9%	22.8%	33.6%	40.0%
50	63.8%	18.6%	17.8%	29.4%	40.0%
100	63.2%	16.1%	14.3%	28.0%	41.7%
500	65.4%	10.4%	9.4%	29.0%	40.5%
1000	64.2%	8.7%	8.6%	28.6%	39.7%

表 3　有限標本破綻点

LAD	LMS	LTS	S
$\frac{1}{n}$	$\frac{[\frac{n}{2}]-p+2}{n}$	$\frac{[\frac{n-p}{2}]+1}{n}$	$\frac{[\frac{n}{2}]-p+2}{n}$
Deepest			
$\frac{\lceil \frac{n}{p+1} \rceil - p+1}{n} \leq \varepsilon_n^*(DR, Z_n) \leq \frac{1}{3}$			

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] Bai, Z. and He, X. (1999).Asymptotic distributions of the maximal depth estimators for regression and multibariate location, Ann. Statist, Vol. 27, No. 5, 1616-1637.

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

[5] 藤木美江 (2003).Regression Depth の理論とその応用に関する研究, 南山大学経営学研究科修士論文.

[6] He, X. and Portnoy, S. (1998).Asymptotics of the deepest line, in “Applied Statistical Science ：Nonparametirc Statistics and Related Topics”, Nova Science Publishers, New York, 71-81.

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

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