最深回帰推定量とそのRによる実用化

大見　俊司[1]　　安藤　雅和[2]　　木村　美善[3]

1　はじめに

線形回帰において, Rousseeuw and Hubert (1999) は regression depth という新しい概念を導入し, これを最大にする最深回帰推定量を提案した. regression depth は当てはめた超平面に対するデータ（または確率分布）のバランスの良し悪しの程度をはかるものであり, 最深回帰推定量は 1 次元の場合にはメディアンと一致する. 最深回帰推定量はメディアンを多次元に一般化した優れたロバスト推定量であり, そのロバストネスと諸性質は Rousseeuw and Hubert (1999), Aelst and Rousseeuw (2000), Aelst, Rousseeuw, Hubert and Struyf (2002) などにおいて研究されている. よく知られているように最小 2 乗推定量は誤差の標準的仮定のもとでは, 線形不偏推定量の中で最良であり, さらに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が, 標準的仮定からの「ずれ」に対しては敏感であり, たった一つの外れ値によって大きな影響を受けてしまう. したがって, 標準的仮定からの「ずれ」や外れ値が生じる可能性のある場合には, これらの「ずれ」や外れ値に対して影響を受けにくく良さの損失の少ないロバスト推定量を用いることが望ましい. 現実の多くの問題においては, 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, ロバストネスの問題は本質的で重要な問題である.

最深回帰推定量以外に, これまで様々なロバスト回帰推定量が提案されている. 主なものだけでも, M 推定量（Huber, 1973）, GM 推定量（Hampel et al., 1986), LMS 推定量 (Rousseeuw, 1984), LTS 推定量 (Rousseeuw, 1984), MM 推定量（Yohai, 1987), S 推定量 (Rousseeuw and Yohai, 1984), τ 推定量 (Yohai and Zamar, 1988), GS 推定量 (Croux, Rousseeuw and Hossjer, 1994) などがある. これらのロバスト回帰推定量の統計解析ソフト R による利用は年々増え続けており, 今後ますます拡大していくと思われる. 最深回帰推定量については, regression depth に基づいた新しい推定量であることもあり, 統計研究者の間でもまだ十分に理解されておらず, R で利用できるようになっていない. 我が国ではロバスト統計分野の研究者が少なく, 最近のロバスト推定法に対する理解とその応用は非常に遅れている. 本論文では次の 3 点を目的とする.

1) 最深回帰推定量とその基本的性質を紹介する.
2) 最深回帰推定量を R で利用できるようにする.
3) 最深回帰推定量と他の回帰推定量を R を用いて比較する.

本論文の構成は次の通りである. 第 2 節では, regression depth の定義と性質を述べる. 第 3 節では, 最深回帰推定量の定義とその基本的性質について考察する. 第 4 節では, Aelst et

[1] 南山大学数理情報研究科　E-mail: m05mm020@nanzan-u.ac.jp
[2] 日本学術振興会特別研究員　E-mail: andomasa@econ.nagoya-cu.ac.jp
[3] 南山大学数理情報学部　E-mail: kimura@ms.nanzan-u.ac.jp

al. (2002) による Fortran で書かれた最深回帰推定量のプログラム（MEDSWEEP[4]）を R 用に書き直し, このプログラムの評価を行なう. 第 5 節では, R を用いて最深回帰推定量と LAD（least absolute values estimator）, LMS, LTS, S を有限標本相対効率によりシミュレーション比較する. 第 6 節では最深回帰推定量と他の回帰推定量の相違を視覚的に見るために単回帰の例を取り上げる.

2 Regression depth

線形回帰モデル

$$y = (\boldsymbol{x}', 1)\boldsymbol{\theta} + \varepsilon \tag{1}$$

を考える. ここで, $\boldsymbol{x} = (x_1, \cdots, x_{p-1})'$ は $p-1$ 次元確率ベクトル, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$ は p 次元回帰係数ベクトル, ε は確率誤差, y は応答変数とする. n 個の観測値データを $\boldsymbol{Z}_n = \{\boldsymbol{z}_i = (x_{i1}, \cdots, x_{i,p-1}, y_i); i = 1, \cdots, n\} \subset \mathbb{R}^p$ とし, $\boldsymbol{Z}_n$ に回帰式 (1)（この回帰式を $\boldsymbol{\theta}$ と表す）を当てはめたときの残差を $r_i(\boldsymbol{\theta}) = y_i - (\boldsymbol{x}', 1)\boldsymbol{\theta} = y_i - \theta_1 x_{i1} - \cdots - \theta_{p-1} x_{i,p-1} - \theta_p$ とする. データ $\boldsymbol{Z}_n$ に対する regression depth を定義するためにまず $nonfit$ (不適合) を次のように定義する.

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すべての $\boldsymbol{x}_i$ 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すべての $\boldsymbol{x}_i$ に対して $r_i(\boldsymbol{\theta}) < 0$ ならば $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$ は $\boldsymbol{Z}_n$ に対して $nonfit$ と呼ばれる.

これを用いて, $\boldsymbol{Z}_n$ に対する regression depth を次のように定義する.

定義 2　データ $\boldsymbol{Z}_n \subset \mathbb{R}^p$ に対する $\boldsymbol{\theta}$ の $rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n)$ 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る. すなわち

$$rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n) = \min_{\mathbf{u}, v} \{\sharp(r_i(\boldsymbol{\theta}) \geq 0 \text{ かつ } \boldsymbol{x}_i'\boldsymbol{u} < v) + \sharp(r_i(\boldsymbol{\theta}) \leq 0 \text{ かつ } \boldsymbol{x}_i'\boldsymbol{u} > v)\} \tag{2}$$

ここで, すべての $(\boldsymbol{x}_i', y_i) \in \boldsymbol{Z}_n$ に対して, 最小は $\boldsymbol{x}_i'\boldsymbol{u} \neq v$ を満たすすべての単位ベクトル $\boldsymbol{u} = (u_1, \cdots, u_{p-1})' \in \mathbb{R}^{p-1}$ と, $v \in \mathbb{R}$ でとられるものとする. $\sharp$ は要素数を表す.

この定義により, データ $\boldsymbol{Z}_n \subset \mathbb{R}^p$ 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 $\boldsymbol{\theta}$ を傾けるとき, 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.

regression depth の概念を理解しやすくするために, 単回帰（p=2）の場合を考えてみよう. この場合には定義 1 と 2 は次のようになる.

定義 $1'$　どの x_i とも一致しない実数 $v_{\boldsymbol{\theta}} = v$ が存在し, 次の 1) または 2) が成り立つとき, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ を $\boldsymbol{Z}_n$ に対して $nonfit$ という.

[4] MEDSWEEP　http://www.agoras.ua.ac.be/

1)　$r_i(\boldsymbol{\theta}) < 0, \forall x_i < v$ かつ $r_i(\boldsymbol{\theta}) > 0, \forall x_i > v$
2)　$r_i(\boldsymbol{\theta}) > 0, \forall x_i < v$ かつ $r_i(\boldsymbol{\theta}) < 0, \forall x_i > v$

定義 $2'$　データ集合 $\boldsymbol{Z}_n \subset \mathbb{R}^2$ に対する $\boldsymbol{\theta} = (\theta_1, \theta_2)$ の $rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n)$ 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る.

図 1 と図 2 でそれぞれ $nonfit$ と $rdepth$ の例を示す.

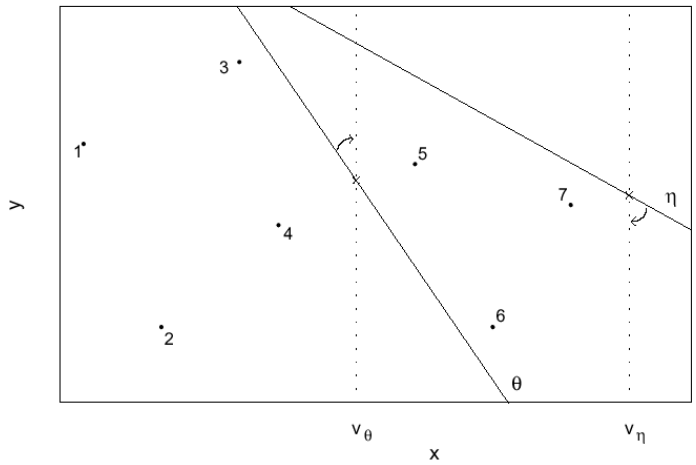

図 1　$nonfit$ となる θ と η

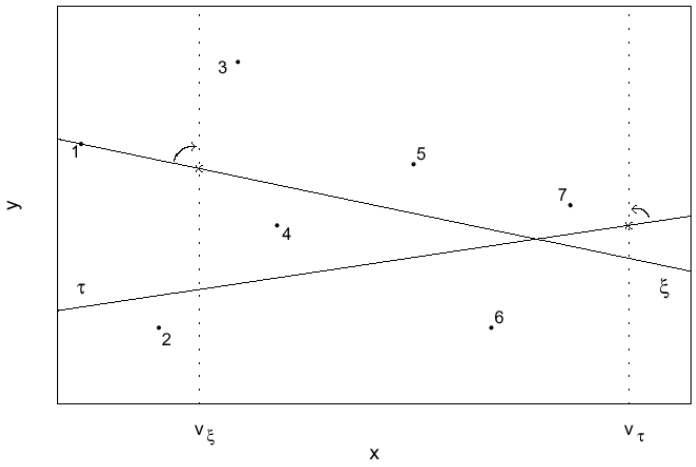

図 2　$rdepth = 2$ をもつ τ と $rdepth = 3$ をもつ ξ の例

図 1 では直線 θ, η に対応する v を v_θ, v_η とした. v_θ, v_η 上の $\times$ 印を中心にして直線 θ, η を垂直になるまで傾けると 1 つの観測値も通らずに回転することができるので直線 θ, η は $nonfit$ である. 図 2 では直線 τ, ξ に対応する v を v_τ, v_ξ とし, v_τ, v_ξ 上の $\times$ 印を中心にして直線 τ, ξ を垂直になるまで傾けると直線 τ は 2 つの観測値とぶつかり, その観測値を取り除かなくてはならないので $rdepth$ は 2 となる. 直線 ξ は 3 つの観測値とぶつかるので $rdepth$ は 3 となる. $rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n)$ については次の定理が成り立つ.

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

$$k \leq rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n) \leq \left[\frac{n+k}{2}\right]. \tag{3}$$

よって, $k = n$ のとき $rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n) = n$ 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 H 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

$$\begin{aligned} &rdepth(\boldsymbol{\theta}, H) \\ &= \min_{\boldsymbol{u}, v}\left\{H(y - (\boldsymbol{x}', 1)\boldsymbol{\theta} > 0 \text{ かつ } \boldsymbol{x}'\boldsymbol{u} < v) + H(y - (\boldsymbol{x}', 1)\boldsymbol{\theta} < 0 \text{ かつ } \boldsymbol{x}'\boldsymbol{u} > v)\right\} \end{aligned} \tag{4}$$

によって定義される. ここで H は確率変数 $(\boldsymbol{x}', y)$ の分布であり, 最小は $H(\boldsymbol{x}'\boldsymbol{u} = v) = 0$ を満たすすべての単位ベクトル $\boldsymbol{u} = (u_1, \cdots, u_{p-1})' \in \mathbb{R}^{p-1}$ と $v \in \mathbb{R}$ でとられるものとする.

この $rdepth(\boldsymbol{\theta}, H)$ は v を中心に垂直になるまで超平面 $\boldsymbol{\theta}$ を傾けるとき, 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しても同等である. $rdepth(\boldsymbol{\theta}, H)$ について次の 2 つの定理が成り立つ.

定理 2　$\boldsymbol{Z}_n$ が密度関数をもつ分布 H からの標本のとき

$$\frac{rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n)}{n} \xrightarrow[n\to\infty]{a.s.} rdepth(\boldsymbol{\theta}, H). \tag{5}$$

ここで $a.s.$ は概収束を表す.

定理 3

a. $(\boldsymbol{x}_i', y_i)$ 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

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n) \leq \left[\frac{n+p}{2}\right]. \tag{6}$$

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

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) \leq \frac{1}{2}. \tag{7}$$

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

$$med[y|\boldsymbol{x}] = \tilde{\theta}_1 x_1 + \cdots + \tilde{\theta}_{p-1} x_{p-1} + \tilde{\theta}_p \tag{8}$$

を満たすならば

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) = rdepth(\tilde{\boldsymbol{\theta}}, H) = \frac{1}{2}. \tag{9}$$

3 最深回帰推定量

3.1 定義と Fisher-consistency

定義 4 データ $\boldsymbol{Z}_n$ に対する最深回帰推定量 $DR(\boldsymbol{Z}_n)$ は $rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n)$ を最大にする $\boldsymbol{\theta}$ として定義する. すなわち

$$DR(\boldsymbol{Z}_n) = \arg\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n). \tag{10}$$

最深回帰推定量 $DR(\boldsymbol{Z}_n)$ は分布の仮定を必要とせず, 回帰共変性, 尺度共変性, アフィン共変性を満たす推定量である. また, $\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, \boldsymbol{Z}_n)$ を与える $\boldsymbol{\theta}$ が複数ある場合はそれら $\boldsymbol{\theta}$ の平均をもって推定量とする. $\boldsymbol{Z}_n$ が 1 変量データである場合, 任意の $\theta \in R$ に対して, $\max_{\theta} rdepth(\theta, \boldsymbol{Z}_n) = \max_{\theta} \min(\sharp\{y_i \leq \theta\}, \sharp\{y_i \geq \theta\})$ であるので, $DR(\boldsymbol{Z}_n)$ は y_i のメディアンになっていることがわかる. このように, 最深回帰推定量は 1 変量の場合のメディアンを線形回帰へと一般化させたものである.

定義 5 p 次元確率変数 $(\boldsymbol{x}', y)$ の分布 H に対する最深回帰推定量 $DR(H)$ は $rdepth(\boldsymbol{\theta}, H)$ を最大にする $\boldsymbol{\theta}$ として定義する. すなわち

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

ここで分布 H は狭義に正の密度関数を持ち

$$med_H(y|\boldsymbol{x}) = (\boldsymbol{x}', 1)\tilde{\boldsymbol{\theta}} \tag{12}$$

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) によるものであるが, 関数型がパラメトリックであり, 誤差分布がノンパラメトリックであるような大きなあるセミパラメトリックモデル $\mathcal{H}$ に H が属するとき, 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. ここで H_nは $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$ の経験分布関数を表す.

次に最深回帰推定量のロバストネスを測る. 推定量のロバストネスをはかる指標として, 有限標本破綻点, 影響関数, 感度関数および相対効率を考える.

3.2 有限標本破綻点

n 個の観測値からなるデータ $\boldsymbol{Z}_n = \left\{(\boldsymbol{x}_1', y_1), \cdots, (\boldsymbol{x}_n', y_n)\right\}$ に対する推定量 T_n の有限追加型破綻点 $\varepsilon_n^*(T_n, \boldsymbol{Z}_n)$ は, $\boldsymbol{Z}_n$ に m 個の観測値を加えて, 推定量 T_n を破綻させるために必要な m の最小値に対する $\frac{m}{n+m}$ として定義する.

定義 6

$$\varepsilon_n^*(T_n, \boldsymbol{Z}_n) = \min\left\{\frac{m}{n+m}; \sup_{\boldsymbol{Z}_{n+m}} \parallel T_{n+m}(\boldsymbol{Z}_{n+m}) - T_n(\boldsymbol{Z}_n) \parallel = \infty\right\}.$$

この有限標本追加型破綻点から一般的に扱われている有限標本破綻点を得ることは Zuo(2001) により与えられている. $\varepsilon_n^*(T_n, \boldsymbol{Z}_n)$ について次の 2 つの結果が得られる.

定理 5 $\boldsymbol{Z}_n$ において $\boldsymbol{x}_i$ が general position にあるならば

$$\varepsilon_n^*(DR, \boldsymbol{Z}_n) \geq \frac{1}{n}\left(\left\lceil \frac{n}{p+1} \right\rceil - p + 1\right) \underset{n\to\infty}{\longrightarrow} \frac{1}{p+1}. \tag{13}$$

最深回帰推定量 DR の ε_n^* はもとのデータ $\boldsymbol{Z}_n$ がそれ自身異常なとき $\frac{1}{p+1}$ になる.

定理 6 $\boldsymbol{Z}_n$ 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 (12) を満たすならば

$$\varepsilon_n^*(DR, \boldsymbol{Z}_n) \underset{n\to\infty}{\overset{a.s.}{\longrightarrow}} \frac{1}{3}. \tag{14}$$

3.3 影響関数

分布 H における推定量 T の影響関数 $IF(\boldsymbol{z}, T, H)$ は $\boldsymbol{z} = (\boldsymbol{x}', y)$ に小さい確率が加わることによる T への影響を測るものである. $\Delta_{\boldsymbol{z}}$ によって $\boldsymbol{z}$ で確率 1 をもつ確率分布を表し, $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_{\boldsymbol{z}}$ と書くとき, 影響関数は次のように定義される.

$$\begin{aligned} IF(\boldsymbol{z}, T, H) &= \lim_{\varepsilon\to 0} \frac{T((1-\varepsilon)H + \varepsilon\Delta_z) - T(H)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{T(H_\varepsilon) - T(H)}{\varepsilon} \\ &= \frac{\partial}{\partial\varepsilon} T(H_\varepsilon)|_{\varepsilon=0} \end{aligned} \tag{15}$$

最深回帰推定量 $DR = (DR_1, DR_2)$ は, 回帰共変, 尺度共変, アフィン共変であるので平均 $\boldsymbol{\mu}$, 分散共分散行列 $\boldsymbol{\Sigma}$ の楕円型分布 $H = H_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ における影響関数は $H = H_{\boldsymbol{0},\boldsymbol{I}}$ におけるそれから求められる. ここで DR_1, DR_2 はそれぞれ傾き, 切片の推定量を表す. H が 2 変量

標準正規分布 $N_2(\mathbf{0}, \boldsymbol{I})$ のとき, 次の結果が得られる.

定理 7　$H = N_2(\mathbf{0}, \boldsymbol{I})$ における最深推定量の影響関数は

$$IF((x,y), DR_1, H) = \frac{sgn(x)sgn(y)}{2\phi(0)} \left(\frac{I(\phi(x) \geq \phi(0)/3)}{4\phi(x)} + \frac{I(\phi(x) < \phi(0)/3)}{\phi(0) + \phi(x)} \right) \tag{16}$$

$$IF((x,y), DR_2, H) = \frac{sgn(y)}{2\phi(0)} \left(\frac{I(|x| \leq \Phi^{-1}(\frac{2}{3}))}{\Phi(|x|)} + \frac{I(|x| > \Phi^{-1}(\frac{2}{3}))}{2(2\Phi(|x|) - 1)} \right) \tag{17}$$

である. ただし, Φ は 1 変量標準正規分布 $N(0,1)$ の分布関数であり, ϕ はその密度関数である.

図 3 と図 4 は定理 7 の $H = N_2(\mathbf{0}, \boldsymbol{I})$ における最深回帰推定量 DR の傾きと切片の影響関数のグラフである. これらの 2 つの図は Aelst and Rousseeuw (2000) の FIG.4 を引用した. 図を見るとわかるように DR の傾きと切片の影響関数はともに有界である. これは Hampel の用語を使うと, DR は B-robust であることを意味する. T が B-robust とは T の H における gross-error sensitivity

$$\sup_{\boldsymbol{z}} | \, IF(\boldsymbol{z}; T, H) \, |$$

が有限であることをいう. gross-error-sensitivity は微小な汚染によって T が受ける最大の影響を表す.

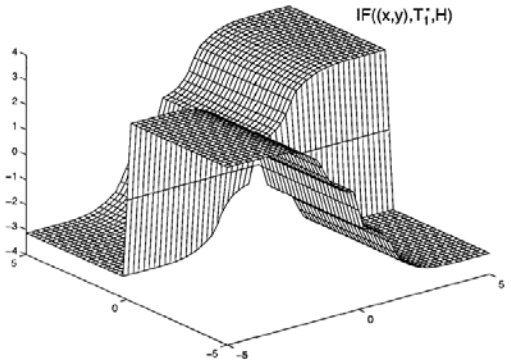

図 3　最深回帰推定量の傾きの影響関数

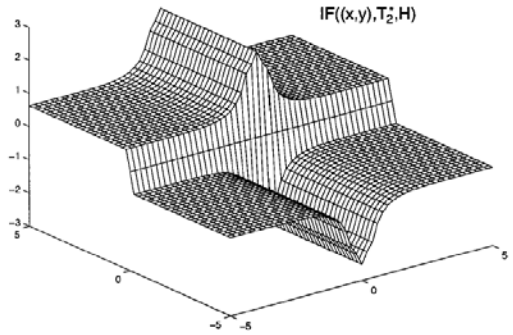

図 4　最深回帰推定量の切片の影響関数

3.4　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を計算する. 任意の推定量 T_n に対する感度関数は標本 $\boldsymbol{Z}_n = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n\}$ 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,

$$SF_n(\boldsymbol{z}, T, \boldsymbol{Z}_n) = n\big(T_{n+1}(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \boldsymbol{z}) - T_n(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)\big). \tag{18}$$

感度関数は実際の標本 $\boldsymbol{Z}_n$ に強く依存するので置換型標本 $\boldsymbol{Z}_n(\pi) = \{(x_i^s, x_{\pi(i)}^s); i = 1, \cdots, n\}$ 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 $N_2(\mathbf{0}, I)$ に近い分布をしており, 周辺分布はメディアン 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

$$APSF_n(\boldsymbol{z}) = \frac{1}{B} \sum_B SF_n(\boldsymbol{z}, T, \boldsymbol{Z}_n(\pi)) \tag{19}$$

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 で計算し, 傾きと切片の $APSF$ 図を描いたものが図 5 と図 6 である, これらは Aelst and Rousseeuw (2000) による図よりもさらによく図 3, 図 4 と似ており, 影響関数に基づく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.

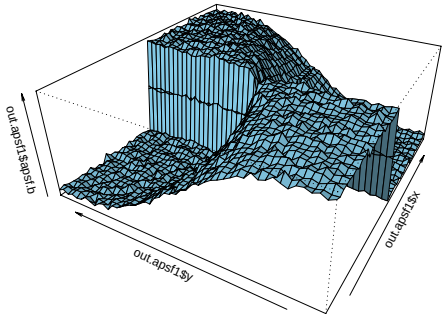

図 5　シミュレーションによる最深回帰推定量の傾きの $APSF$ 図

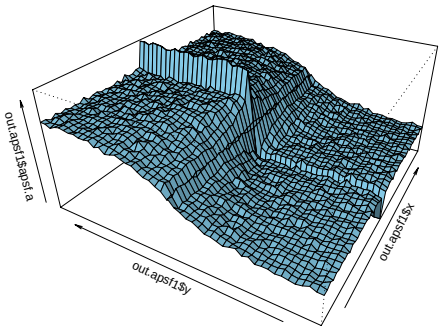

図 6　シミュレーションによる最深回帰推定量の切片の $APSF$ 図

3.5　相対効率

最深回帰推定量は漸近正規性をもたないが, 正規分布からわずかに異なる正規分布に近い極限分布をもつことが He and Portnoy(1998) によって証明されている. そこで最深回帰推定量の最小 2 乗回帰推定量に対する相対効率として, 正規分布の下での 2 つの推定量の分散比を考えることにする. そして, 標本数 n に対して, シミュレーションにより, 分散比の近似値を求めた結果が表 1 である. 相対効率はデータ数が増加してもほとんど変わらないことがわかる.

4　プログラム

Aelst et al.(2002) によって Fortran で書かれた最深回帰推定量の近似プログラム MEDSWEEP がこれまで使用されてきた. 我々はこれを統計解析システム R で使用で

表 1　10000 回のシミュレーションに基づいた最深回帰推定量と LS の相対効率

n	切片	傾き
20	55.97	39.97
50	61.64	40.00
100	61.67	41.71
500	62.26	40.51
1000	64.01	39.68
10000	63.99	40.71

きるように S 言語に書き換えた. 次にこのプログラムの性能を検証する.

4.1　プログラムの性能評価

与えられた n, p に対して正規分布と自由度 2 の t 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 アルゴリズムから最深回帰推定量 $(\hat{\theta}_1^{(j)}, \cdots, \hat{\theta}_p^{(j)})$ を求めて傾きの平均 2 乗誤差

$$MSE(\hat{\theta}_1, \cdots, \hat{\theta}_{p-1}) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{p-1} \sum_{i=1}^{p-1} (\hat{\theta}_i^{(j)} - \theta_i)^2 \tag{20}$$

を計算する. ここで真値は $\theta_i = 0; i = 1, \cdots, p-1$ である. 切片の MSE は $\frac{1}{m} \sum_{j=1}^{m} (\hat{\theta}_p^{(j)})^2$ である. 正規分布と自由度 2 の t 分布の乱数からそれぞれの 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 2, 3 に載せる. データ数が増加すると MSE が減少していることが見てとれる. 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 点を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り, 特に自由度 2 の t 分布における MSE が非常に大きいのは外れ値の影響を受けているからだと思われる. データ数が少ないときには良い推定が出来ない場合があることがわかる. また, 書き換えたプログラムの n と p に対する平均計算時間を表 4 に載せる. $n = 1000, p = 10$ のとき, 20 秒ほどかかるが実際の分析においては問題ないので, このプログラムは実行可能であろう.

5　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, 各誤差分布のメディアンを 0 と仮定するだけでよい. これらはかなり弱い条件である. 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, このモデルは誤差分布が歪んでいたり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認めない. LMS, LTS, S などの推定量の目的は最頻値を捜すことである. それはこれらの推定量が大部分のデータを含む集中した線形雲を

表 2 正規分布における切片と傾きの $MSE(\times 10^{-3})$

n	MSE	p 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

表 3 自由度 2 の t 分布における切片と傾きの $MSE(\times 10^{-3})$

n	MSE	p 2	3	5	10
20	切片	141.12	193.73	404.93	6.91e+15
	傾き	123.51	156.90	335.12	6.99e+14
50	切片	45.28	49.59	57.56	99.92
	傾き	33.77	34.54	37.15	56.35
100	切片	20.72	22.87	25.73	33.16
	傾き	14.21	14.83	15.90	19.64
500	切片	4.02	4.06	4.31	4.63
	傾き	2.32	2.31	2.35	2.52

表 4 各 n, p に対する MEDSWEEP の計算時間 (秒). ただし, 各計算時間は 10000 個のサンプルの平均である.

n	p 2	3	4	5	10
20	0.022	0.081	0.178	0.284	0.761
50	0.073	0.102	0.183	0.271	1.033
100	0.077	0.290	0.632	0.987	2.283
500	0.343	1.306	2.796	4.481	9.628
1000	0.663	2.584	5.557	8.530	19.33

捜すことを意味する. 一方, 最深回帰推定量はデータの線形雲の中心を捜すメディアンタイプの推定方法である.

5.1 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求めて LS の切片と傾きの分散比（相対効率）を計算すると表 6 を得た. また, 表 5 で各回帰推定量の破綻点と有限標本破綻点を載せる. 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非常に高い. それに対して, Deepest は効率と破綻点がともに高く, バランスがとれている.

表 5 破綻点

	LS	LAD	LMS	LTS	S	Deepest
破綻点	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{p+1} \leq \varepsilon_n^*(DR, \boldsymbol{Z}_n) \leq \frac{1}{3}$
有限標本破綻点	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{[\frac{n}{2}]-p+2}{n}$	$\frac{[\frac{n-p}{2}]+1}{n}$	$\frac{[\frac{n}{2}]-p+2}{n}$	$\frac{\lceil \frac{n}{p+1} \rceil - p+1}{n} \leq \varepsilon_n^*(DR, \boldsymbol{Z}_n) \leq \frac{1}{3}$

表 6　LS との有限標本相対効率

	切片の有限標本相対効率					傾きの有限標本相対効率				
n	LAD	LMS	LTS	S	Deepest	LAD	LMS	LTS	S	Deepest
20	67.02	21.53	23.41	36.71	55.97	63.10	19.89	22.78	33.56	39.97
50	63.27	17.04	16.25	30.00	61.64	63.82	18.58	17.84	29.39	40.00
100	63.57	13.89	12.93	29.48	61.67	63.21	16.09	14.29	28.04	41.71
500	62.41	8.85	8.65	28.07	62.26	65.39	10.43	9.43	29.00	40.51
1000	63.14	7.09	8.01	28.47	64.01	64.19	8.67	8.60	28.60	39.68

5.2　単回帰

最深回帰推定量と他の推定量 (LS, LAD, LMS, LTS, S) にどのような相違があるのかを視覚的にみるために単回帰における回帰直線を例として取り上げる．Chatterjee et al. (2000, p177) から「1986 年の広告ページ数と広告収入のデータ」を引用する．広告ページ数を x(百枚), 広告収入を y(百万ドル) とする．

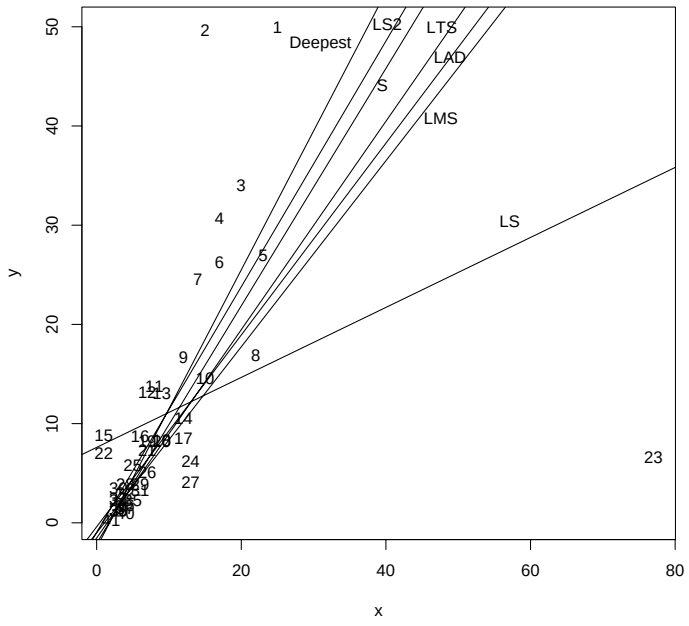

図 7　広告データに対する回帰直線図

LS は 23 番目の観測値の影響を強く受けているが, ロバスト推定量による回帰直線は影響を受けていない．また, LS による外れ値 (1, 2, 23) を除いた LS(LS2) に最深回帰推定量は近く, 他のロバスト推定量による直線と LS2 を挟んで反対側にあることが見てとれる．

6 おわりに

本論文では, regression depth に基づく最深回帰推定量を紹介し, その基本的な性質について考察した. 最深回帰推定量を計算し, その性質を調べるに際しては R を用いた. 最深回帰推定量以外の主要なロバスト推定量は GS 推定量を除きほとんど R で利用できるようになっている. 最深回帰推定量は提案されてからまだ日が浅く, よく理解されていないこともあり, R での利用はこれまで可能ではなかった. 筆者達は Fortran のみで利用可能であった最深回帰推定量のプログラムを R 上で利用できるように書き直すことに思い至った. そして, 書き直した R プログラムの機能と性能をチェックしたうえで, これを使用して本研究を行なった. 最深回帰推定量を R で使用したのは, 著者達の知る限りでは, 本研究が最初である. 我が国ではロバスト統計学分野の研究者が少なく, ロバスト推定の研究は理論と応用の両面で大きく遅れている. その主な理由の一つにロバスト理論が難しく取っ付きにくいことがある. 最深回帰推定量は他のロバスト推定量に劣らず魅力のある推定量であるので, R で利用可能となったことを契機として, もっと身近なものになり, 広く活用されるようになることを願う.

参考文献

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

[2] Aelst, S.V., Rousseeuw, P.J., Hubert, M. and Struyf, A. (2002). The deepest regression method, *J. Multivariate Anal.*, **81**, 138-166.

[3] Bai, Z. and He, X. (1999). Asymptotic distributions of the maximal depth estimators for regression and multibariate location, *Ann. Statist.*, **27**, 1616-1637.

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

[5] Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994). Generalized S-estimators, *J. Amer. Statist. Assoc.*, **89**, 1271-1281.

[6] 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.

[7] 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.

[8] He, X. and Portnoy, S. (1998).Asymptotics of the deepest line, *Applied Statistical Science : Nonparametric Statistics and Related Topics*, Nova Science Publishers, New York, 71-81.

[9] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Statist.*, **1**, 799-821.

[10] Rousseeuw, P.J. (1994). Least median of squares regression, *J. Amer. Statist. Assoc.*, **79**, 871-880.

[11] Rousseeuw, P.J. and Hubert, H. (1999). Regression depth, *J. Amer. Statist. Assoc.*,

94, 388-402.

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

[13] Rousseeuw, P.J., and Yohai, V.J. (1994). Robust regression by means of S-estimators, *Robust and Nonlinear Time Series Analysis*, J.Franke, W.Hardle and R.D.Martin (eds.), Lectures Notes in Statistics, **26**, 256-272, Springer, New York.

[14] Yohai, V.J. (1987). High breakdown-point and high efficiency estimates for regression, *Ann. Statist.*, **15**, 642-656.

[15] Yohai, V.J., and Zamar, R.H. (1988). High breakdown estimates of regression by means of the minimization of an efficient scale, *J. Amer. Statist. Assoc.*, **83**, 406-413.

[16] Zuo, Y. (2001). Some quantitative relationships between two types of finite sample breakdown point, *Statist. Probab. Lett.*, **51**, 369-375.