ロバスト回帰の理論とその応用
– Regression Depth を中心に –

2001MM064 大見　俊司
指導教員　木村　美善

1　はじめに

回帰分析では最小 2 乗法による推定量が通常よく用いられているが, この回帰係数の最小 2 乗推定量はよく知られているように線形回帰の標準的仮定のもとでは線形不偏推定量の中で最良のものであり, 正規分布が仮定される場合には全ての不偏推定量の中で最良である. しかし, 最小 2 乗推定量は標準的仮定からのずれに対して敏感であり, 一つの外れ値によっても大きな影響を受けてしまう. こうした危険を回避し, 安全性と信頼性を確保するためには標準仮定からのずれや外れ値に対して影響が小さく, よ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.

本論文の目的は Regression Depth の理論を理解することと, 最小 2 乗法とロバスト回帰推定法の比較することである.

2　回帰分析

y と $x_1, x_2, \cdots, x_p$ の関係は線形重回帰モデル

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

により定式化される.p=1 の時を単回帰モデルという. 回帰モデルは次の 3 つの仮定を満たすと仮定する.
仮定 1　E(ϵ_i)=0 , 仮定 2　V(ϵ_i) = σ^2
仮定 3　説明変数 $x_{j1}, \cdots, x_{jn}$が互いに独立である ($j = 1, 2 \cdots, p$).

3　ロバスト回帰

1. 仮定された分布のもとで、かなり高い効率がある。
2. 仮定された分布から少しずれた分布のもとで効率が少ししか下がらない。
3. 仮定された分布からかなりずれた分布のもとでも極端に効率が下がらず破局に至ることはない。

このような 3 つの性質をもつ統計量はロバストであるといわれる。

3.1　いろいろなロバスト推定量

これまでに様々なロバスト推定量が提案されており, 特筆すべきものだけでも,M 推定量,GM 推定量,repeated median 推定量, LMS 推定量,LTS 推定量, MM 推定量, τ 推定量,S 推定量 GS 推定量などがある. その中で研究を行った LMS 推定量と LTS 推定量を取り上げる.

3.1.1　LMS 推定量 $\hat{\theta}_{LMS}$

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

により定義される. Breakdown point は 50 %である.

3.1.2　LTS 推定量 $\hat{\theta}_{LTS}$

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

として定義される. $(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$ のように残差の 2 乗を小さいほうから並び替える.LTS は最小 2 乗法 (LS) に似ているが、LS との違いは大きい方の残差平方が使われていないことである. この推定量の Breakdown point は h がだいたい $\frac{n}{2}$ のときに,50 %に達する ([4] 参照).

4　Regression Depth

Regression Depth はロバスト回帰の推定法として新しく提案されたばかりの概念であり, ロバスト回帰の分野の研究において今後注目されるであろう推定量である ([1],[3] 参照). 適合の深さを導入するために, 最初に不適合を定義する. 単回帰においてデータ集合 $Z_n = \{(x_i, y_i); i = 1, \cdots, n\} \subset R^2$ に対して $y = bx + a$ を当てはめる。b を傾き,a を切片とし,$\theta = (b, a)$ で表す。θ に対する Z_n の残差を $r_i = r_i(\theta) = y_i - bx_i - a$ とする.

定義 1 どの x_i とも一致しない実数 $v_\theta = v$ が存在し次のことが成り立つとき, $\theta = (b, a)$ は Z_n に対して不適合と言う.

$r_i(\theta) < 0, \forall x_i < v$ かつ $r_i(\theta) > 0, \forall x_i > v$
or
$r_i(\theta) > 0, \forall x_i < v$ かつ $r_i(\theta) < 0, \forall x_i > v$

定義 2 データ集合 $Z_n \subset R^2$に対する$\theta = (b, a)$ の $rdepth(\theta, Z_n)$ はθ を不適合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.

$rdepth(\theta, Z_n)$ を計算するためにまず観測値を $x_1 \leq x_2 \leq \cdots \leq x_n$に並び替える.$x_i = x_j$ $(i \neq j)$ の場合は y の値の小さいほうが前にくる. 次に

$$rdepth(\theta, Z_n) = \min_{1 \leq i \leq n} (\min\{S^+(x_i) + G^-(x_i), G^+(x_i) + S^-(x_i)\})$$

を使って rdepth を計算する事ができる. ここで

$$S^+(v) = \sharp\{j; x_j \leq v \quad and \quad r_j \geq 0\}$$
$$G^-(v) = \sharp\{j; x_j > v \quad and \quad r_j \leq 0\}$$
$$G^+(v) = \sharp\{j; x_j \geq v \quad and \quad r_j \geq 0\}$$
$$S^-(v) = \sharp\{j; x_j < v \quad and \quad r_j \leq 0\}.$$

この結果 Regression Depth の直線はデータの奥深くに入り込むことになる.

4.1 Catline の説明

Catline とは Regression Depth の概念をもとにした単回帰分析の方法である．まずデータ集合 $Z_n = \{(x_i, y_i); i = 1, \cdots, n\}$ を $x_1 \leq x_2 \leq \cdots \leq x_n$ 順に並び替える．そしてデータ集合 Z_n に対して, 次のように L,M,R,3 つのグループに分ける．

n=3m のとき $\sharp L = m, \sharp M = m, \sharp R = m$.

n=3m+1 のとき $\sharp L = m, \sharp M = m+1, \sharp R = m$

n=3m+2 のとき $\sharp L = m+1, \sharp M = m, \sharp R = m+1$

定義 4 Catline $\theta_{cat} = (b_{cat}, a_{cat})$ は $L \cup M$ と $M \cup R$ を同時に 2 等分する直線 $y = b_{cat}x + a_{cat}$ である．

直線 $y = b_{cat}x + a_{cat}$ によって定義された 2 つの開半空間が [n/2] 点より多くを含まないならば N 点からなる集合を直線 $y = b_{cat}x + a_{cat}$ は 2 等分する. もし N が奇数ならば直線は少なくとも 1 点を通る．

([r] は r を超えない最大の整数を表す.)

4.2 Catline のアルゴリズム

(b,a) が Catline になるために $r_{LM}(b,a)$ と $r_{MR}(b,a)$ によって $L \cup M$ および $M \cup R$ の中の残差の集合を表す. すなわち $r_{LM}(b,a) = \{y_i - bx_i - a; i \in L \cup M\}, r_{MR}(b,a) = \{y_i - bx_i - a; i \in M \cup R\}$. 2k 個の値 $t_1 \leq t_2 \leq, \cdots, \leq t_{2k}$ のメディアンは間隔 $[t_k, t_{k+1}]$ である．

定理 7 任意のデータ集合 $Z_n \subset R^2$ に対して

$$(b,a) \text{ は } Catline$$
$$\Updownarrow$$
$$med\ r_{LM}(b,a) \cap med\ r_{MR}(b,a) \neq \emptyset \qquad (1)$$
$$\Updownarrow$$
$$a \in med\ r_{LM}(b,0) \cap med\ r_{MR}(b,0). \qquad (2)$$

n が 3 の倍数でない時,$\sharp(L \cup M)$ と$\sharp(M \cup R)$ は奇数となる. 従って (2) の両方のメディアンは 1 つとなる．この場合 Catline は

$$med\ r_{LM}(b_{cat},0) \cap med\ r_{MR}(b_{cat},0) \qquad (3)$$
$$a_{cat} = med\ r_{LM}(b,0) \qquad (4)$$

で特徴付けることができる．

n が 3 の倍数のとき Catline の傾きを得るために f(b)=0 を解けばよい.

$$f(b) = med\ r_{LM}(b,0) - med\ r_{MR}(b,0)$$

その後切片を (2),(4) から求めることができる.

5 最小２乗法とロバスト回帰法の比較

[2] の P177 の 1986 年の広告枚数 P(百枚) と広告収入 R(百万ドル) のデータを使う.

右上の表は, 各手法で推定した回帰直線の式である. プロット図を見ると極端な値, すなわち 23 番目の観測値がある. 最小 2 乗法ではこの観測値を通常の値と見ているので,23 番目の観測値のほうに回帰直線が傾いている. そのため他の観測値が外れ値になっている．それに対して Catline,LMS,LTS で求めた直線は 23 番目の観測値の影響を受けずにいる.LMS,LTS で求めた回帰直線は似ているが Catline で求めた回帰直線はすこし違っている. なぜこのようにロバスト回帰直線に違いがあるか考えると,LMS は残差の中央値を最小にするのでそれ以上の残差に影響しない．LTS は今回計算に使った R では $h = [\frac{n}{2}] + 1$ なので残差を小さいほうから並べて $h = [\frac{n}{2}] + 1$ 個しか使わないのでそれ以上の残差に影響しない．2 つの方法とも Breakdown point は約 50% で LMS と LTS が似た回帰直線を引くことがわかる．Catline は Breakdown point が約 33% であり LMS,LTS よりも多くの観測値から直線を引いているので違いが生じたと考えられる．

表 1　広告収入データの回帰直線の比較

手法	回帰直線
最小 2 乗法 (LS)	$\hat{y_i} = 0.3527x_i + 7.6041$
1,2,23 番目の観測値を抜いた最小 2 乗法 (LS2)	$\hat{y_i} = 1.2380x_i - 0.9619$
catline	$\hat{y_i} = 1.400663x_i - 2.504642$
LMS	$\hat{y_i} = 0.9375x_i - 1.0062$
LTS	$\hat{y_i} = 1.050x_i - 1.529$

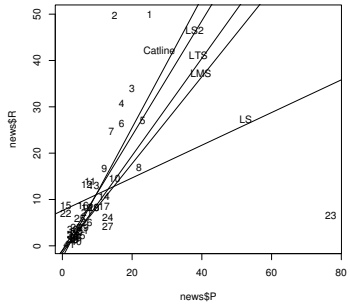

図 1　広告収入データのプロット図

参考文献

[1] 藤木美江：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 (2003)

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

[3] Hubert,M.and Rousseeuw,P.J：The Catline for Deep Regression, Journal of Multivariate Analysis ,66, 270-296(1998).

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