回帰分析の理論とその応用
— 外れ値とロバスト回帰 —

2009SE009 安藤周平
指導教員：木村美善

1　はじめに

分析を行う際に集められたデータはしばしば外れ値と呼ばれる観測値を含んでいる．このようなデータを用いた場合，回帰分析で通常用いられる最小二乗法は外れ値によって非常に良くない影響を受けてしまう．ロバストなアプローチは，データに外れ値が含まれる場合やデータが仮定された分布に近似的にしか従わないときにも信頼出来るパラメータの推定値を生み出す方法である．本研究の目的は最小二乗回帰推定量とロバスト回帰推定量を比較し，外れ値の検出や外れ値が存在する場合の推定にロバスト推定法が有効な方法であることを示すことである．また，データ解析には統計解析ソフト「R」を使用する．

2　回帰モデル

2.1　線形回帰モデル

目的変数 y と p 個の説明変数 $x_1,\cdots,x_p$ に関する n 個の観測値 $y_i, x_{i1},\cdots,x_{ip},\ i=1,\cdots,n$ が与えられているとして，次の線形重回帰モデルを考える．

$$y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}+\varepsilon_i,\ i=1,\cdots,n \quad (1)$$

ここで $\beta_0,\beta_1,\cdots,\beta_p$ は回帰係数を表し，ε_i は誤差を表す．このモデルを行列で表記すると

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \quad (2)$$

となる．また，誤差 ε_i は次の性質を持つと仮定する．

1. $E(\varepsilon_i)=0 \quad (i=1,2,\cdots,n)$ [不偏性]
2. $var(\varepsilon_i)=\sigma^2 \quad (i=1,2,\cdots,n)$ [等分散性]
3. $cov(\varepsilon_i,\varepsilon_j)=0 \quad (i\neq j)$ [無相関性]

これらは誤差 ε_i の平均が 0 であり，分散 σ^2 を持ち，$\varepsilon_i\neq\varepsilon_j$ のとき互いに影響を受けないことを示している．([2] 参照)

2.2　最小二乗法

最小二乗法とは残差平方和が最小になるように回帰係数の推定値を定める方法である．回帰係数 $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_p)'$ の推定量を $\hat{\boldsymbol{\beta}}=(\hat\beta_0,\hat\beta_1,\cdots,\hat\beta_p)'$，目的変数 $\boldsymbol{y}=(y_1,\cdots,y_n)'$ の予測値を $\hat{\boldsymbol{y}}=(\hat y_1,\cdots,\hat y_n)'$ とするとき，

$$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}} \quad (3)$$

と表すことができ，残差を

$$r_i(\hat\beta)=y_i-\hat y_i,\ i=1,\cdots,n, \quad (4)$$

と表す．最小二乗推定量 (LS 推定量)$\hat\beta_{LS}$ は

$$\sum_{i=1}^{n}r_i(\hat\beta_{LS})^2=\min_{\beta}\sum_{i=1}^{n}r_i(\beta)^2 \quad (5)$$

により定義される．$\hat\beta_{LS}$ は正規分布からの「ずれ」や「外れ値」に対して敏感であり，大きな影響を受けて不安定であることが知られている．そして，外れ値からのこの影響のため外れ値を検出することは難しくなることがある．

3　ロバスト法

3.1　ロバスト法とは

ロバスト法はデータが外れ値を全く含まないならば最小二乗法と近似的に同じ結果を与えるが，外れ値が含まれているならば典型的なデータのみに適用された場合の最小二乗法と近似的に同じ結果を与えることができる．すなわち，データ志向の特徴づけとしてロバスト法はデータの大部分にうまく当てはまるというものである．その結果として，高次元多変量の場合においても，外れ値を検出するための非常に信頼できる方法を与える．([1] 参照)

3.2　LMS 推定量

LMS（Least Median of Squares）推定量 $\hat\beta_{LMS}$ は，Rousseeuw(1984) により提案されたロバスト回帰推定量であり，残差平方の中央値を最小にする

$$\hat\beta_{LMS}=\arg\min_{\beta}med(r_1(\beta)^2,\cdots,r_n(\beta)^2) \quad (6)$$

として定義される．LMS 推定量はわかりやすく，破綻点 (breakdown point) は $([n/2]-p+2)/n$ であり，$n\to\infty$ のときに最大の破綻点である 0.5 を持つ．外れ値に対して強くて影響されにくいことから外れ値の検出に威力を発揮する．外れ値が存在しない場合には LS 推定量と LMS 推定量の差は大きくならないが，差が大きくなる場合には外れ値の存在と LS 推定量の使用に注意すべきである．

3.3　LTS 推定量

LTS（Least Trimmed Squares）推定量 $\hat\beta_{LTS}$ は，Rousseeuw(1984) により提案されたものであり，残差平方を $r^2_{(1)}\le r^2_{(2)}\le\cdots\le r^2_{(n)}$ のように昇順に並び替えた h 番目までの和を最小にする

$$\hat\beta_{LTS}=\arg\min_{\beta}\sum_{i=1}^{h}r_{(i)}(\beta)^2 \quad (7)$$

として定義される．残差平方和の中に大きい残差が含まれないことにより，外れ値を避ける事ができるため影響

を受けなくなる．また，LTS 推定量は h 番目以降を無視した OLS 推定量であるという見方もできる．破綻点は $([(n-p)/2]+1)/n$ である．$n \to \infty$ のとき，最大破綻点である 0.5 となる．([3]，[4] 参照)

3.4 MM 推定量

MM 推定量は 2 つの損失関数 ρ_0, ρ_1 に基づいており，それぞれは破綻点と効率に関するものである．MM 推定量 $\hat{\beta}_n$ は

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1'\left(\frac{y_i - x_i'\hat{\beta}_n}{\hat{\sigma}_n}\right)x_i = 0 \tag{8}$$

を満たし，$\hat{\sigma}_n$ は等式

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i - x_i'\hat{\beta}}{\hat{\sigma}_n(\beta)}\right) = b \tag{9}$$

を満たす最小のものとして定義される．ここで $0 < b < 1$ は指定された定数.

4 実データの分析

4.1 データについて

エコカー減税対象車の値段に影響するデータ (2011 年度) を扱う．このデータは車会社であるトヨタ，ホンダ，ダイハツ，マツダ，日産，スバル，スズキのホームページから一部抜粋したものである．目的変数を y：値段 [万]，説明変数を x_1：燃費 [km/L]，x_2：乗車定員 [人数]，x_3：最高出力 [PS] として 19 個の観測値を持つデータを分析する．

4.2 分析と考察

表 1　回帰直線

手法	回帰直線
LS	$\hat{y} = -185.332 + 4.397x_1 + 14.675x_2 + 2.00x_3$
LMS	$\hat{y} = 162.349 + 3.844x_1 - 62.235x_2 + 2.136x_3$
LTS	$\hat{y} = 138.226 + 4.108x_1 - 58.352x_2 + 2.128x_3$
MM	$\hat{y} = 46.2798 + 2.839x_1 - 27.681x_2 + 2.058x_3$
LS*	$\hat{y} = 42.658 + 2.945x_1 - 27.613x_2 + 2.073x_3$

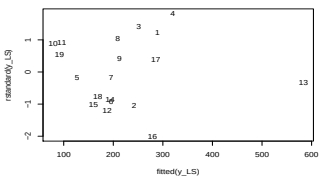

図 1　LS の残差プロット

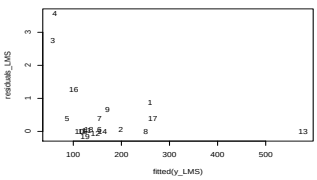

図 2　LMS の残差プロット

表 1 は LS，LMS，LTS，MM 推定量による回帰直線である．LS 推定量によって得られた寄与率 R^2 は 0.896 であり，自由度調整済寄与率 R^{*2} は 0.8752 となり，よく適合している．図 1 は最小二乗法による残差プロットである．図 1 から観測値の 16 番が外れ値であるように思われる．16 番を取り除いて分析をすると，R^2 は 0.9237 であり，R^{*2} は 0.9074 という結果になった．外れ値だと思われた観測値を除いたため R^2，R^{*2} は上がっている．しかし，うまく外れ値を除けたかどうか判断できない．図 2 は LMS 推定量，図 3 は LTS 推定量，図 4 は MM 推定量による残差を標準偏差で割った値をプロットした図である．LMS，LTS，MM は同じような結果を示しており，図 2，3，4 から観測値の 3, 4 番が外れ値であることは明らかである．これらの外れ値を除いて得た最小二乗推定量による回帰直線が表 1 の LS*である．このとき R^2 は 0.9612 であり，R^{*2} は 0.9522 という結果になり，これは外れ値を除く前と比べて，より適合している．これらのことは，ロバスト推定は外れ値を持つデータに対して有効な方法であることを示している．また，いずれのロバスト推定量も 3，4 が外れ値という結果であったが，次に大きな影響を与えている番号が LMS，LTS では 16 番であり，MM では 1 番で異なっている．3，4 番に加え 16 番を除いた場合の R^2 は 0.963，R^{*2} は 0.9537 であるのに対して，1 番を除いた場合の R^2 は 0.974，R^{*2} は 0.9675 であることから MM 推定量の方が LMS，LTS 推定量よりも良い回帰式を与えている．

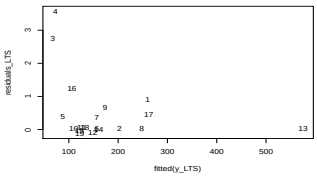

図 3　LTS の残差プロット

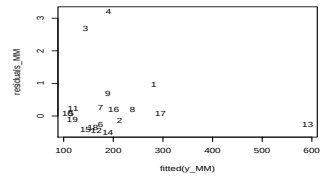

図 4　MM の残差プロット

5 おわりに

ロバスト推定法の理論の理解については，M，LMS，LTS，MM 推定量について学習し，そして，それらを用いることで最小二乗推定量による問題点を論じ，ロバスト推定法の有効性を示すことができたと思う．

参考文献

[1] Maronna R. A. , Martin R. D. and Yohai V. J. : *Robust Statistics : Theory and Methods*, Wiley, 2011.

[2] Rencher A. C. and Schaalje G. B. : *Linear Models in Statistics*, John Wiley & Sons, 2008.

[3] Rousseeuw P. J. and Leroy A. M. : *Robust Regression and Outlier Detection*, John Wiley & Sons, 1987.

[4] 武山嵩弘, 木村美善: ロバストリッジ回帰推定量とそのシミュレーション評価, 南山大学紀要「アカデミア」数理情報編, 第 8 巻, 35-46, 2008.