ロバスト回帰の研究

2002MM036 神谷　美紀　　2002MM058 水戸　藍　　2002MM095 竹内　愛希代
指導教員　木村　美善

1　はじめに

回帰分析において通常用いられる最小 2 乗法は外れ値がある場合には, それによって大きな影響を受け, 良くない方法になってしまうことが知られている. この欠点を克服するのがロバスト回帰法である. 本研究では, ロバスト回帰法の理論を学習し, 実際のデータを使ってロバスト推定量の良さと特徴を最小 2 乗推定量と比較したり, 回帰診断を行うなかで示していくことを目的とする.

2　回帰分析

2.1　線形回帰モデル

従属変数 y と p 個の説明変数 $(x_1, x_2, \cdots, x_p)$ に関する n 個の観測値がデータ (数値) として与えられたとする. $(x_1, x_2, \cdots, x_p)$ から y の値を予測するとき,$(x_1, x_2, \cdots, x_p)$ と y の関係を表すための一つの数式モデルを以下に設定する.

$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \varepsilon_i \quad i = 1, 2, \cdots, n \tag{1}$$

この数式モデルは, 線形重回帰モデル (a liner multiple regression model) と呼ばれ,$\theta_0, \cdots, \theta_p$ は偏回帰係数と呼ばれる未知のパラメータ, ε_i は誤差項で,y_i の変動のうち $(x_{i1}, x_{i2}, \cdots, x_{ip})$ では説明しきれない部分の予測誤差を表している. 特に $p = 1$ のときを単回帰モデルと言い, このとき θ_0 は直線の切片,θ_1 は傾きを表す.

2.2　誤差項の性質

誤差 ε_i は次のような性質をもつと仮定する.

1. [不偏性]　$E(\varepsilon_i) = 0 \quad (i = 1, 2, \cdots, n)$
2. [等分散性]　$V(\varepsilon_i) = \sigma^2 (i = 1, \cdots, n)$
3. [無相関性]　誤差 ε_i は互いに無相関である.
   $Cov(\varepsilon_i, \varepsilon_j) = 0$
   すなわち,ε_i は平均 0, 分散 σ^2 を持ち, 互いに影響を受けずに発生すると想定する.
   推定や検定を行うときは次の正規分布の仮定を追加する.
4. [正規性]　ε_i は互いに独立に正規分布 $N(0, \sigma^2)$ に従う.

予測値と残差をそれぞれ以下のように表す. 残差とは, 実測値と予測値との差である.

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \cdots + \hat{\theta}_p x_{ip} \tag{2}$$

$$r_i(\hat{\theta}) = y_i - \hat{y}_i \tag{3}$$

2.3　最小 2 乗法 (Least square method : LS)

最小 2 乗法は残差平方和が最小になるように推定値 $\left(\hat{\theta}_0 \cdots \hat{\theta}_p\right)$ を定める方法である. よって, 最小 2 乗推定量 $\hat{\theta}_{LS}$ は

$$r_i^2(\hat{\theta}_{LS}) = \min_{\hat{(\theta)}} \sum_{i=1}^{n} r_i^2\left(\hat{\theta}\right) \tag{4}$$

と表すことが出来る.

最小 2 乗法は理解する事が簡単で, 正規分布が仮定されている場合すべての不偏推定量の中で最良であるのでよく使われるのだが, 実際のデータは大体標準的な仮定からずれていたり, 外れ値が存在したりする. 最小 2 乗法は仮定からのずれや外れ値にとても敏感で影響を受けやすく, 外れ値を含んだデータであっても, 推定値からの誤差がそれほど大きくならないので外れ値を検出するのが難しい.

3　ロバスト回帰

3.1　ロバスト回帰とは

通常外れ値は最小 2 乗法による残差を用いて行われる. しかし, 外れ値が $leverage\ point$ の場合これはうまく行かない. $leverage\ point$ が存在する場合 LS 直線はデータに上手に当てはまらない. 単回帰の場合は実際に 2 次元平面上で外れ値を見ることが出来るので, それを取り除いてもう一度解析し直せばよい. しかし重回帰の場合 $leverage\ point$ を見つけることはプロット図が高い次元であるため困難であろう.

このことから外れ値の検出は最小 2 乗法ではうまくいかないことが多い. そしてこの問題を克服するためにロバスト回帰という方法が考案された. ロバスト回帰は外れ値の影響を受けないように工夫した推定量であり, データの多数部分への当てはめを考える.

3.2　漸近効率 (asymptotic efficient)

T に対して漸近性

$$L_G(\sqrt{n}(T_n - T(G))) \Longrightarrow N(0, V(T, G)) \tag{5}$$

が成り立つとしよう. ここで $L_G(S)$ は G のもとでの統計量 S の分布を表わす.$T(G), V(T, G)$ はそれぞれ T_n の漸近値 (asymptotic value), 漸近分散 (asymptotic variance) といわれる.

また $V(T, G)$ は影響関数 IF によって

$$V(T, G) = \int IF(x; T, G)^2 dG(x) \tag{6}$$

と表わされる.

分布 F における Fisher 情報量は,

$$J(F) = \int \left(\frac{f'(x)}{f(x)}\right)^2 dF(x) \tag{7}$$

である. ただし $f = F'$ である. この場合
クラメール．ラオの不等式

$$V(T,F) \geq \frac{1}{J(F)} \tag{8}$$

が成り立ち, 等号は

$$IF(x;T,F) = -J(F)^{-1}(\frac{f'(x)}{f(x)}) \tag{9}$$

の時かつこのときに限り成り立つ,F における T_n の漸近効率は

$$\begin{aligned}\varepsilon &= \frac{\frac{1}{J(F)}}{V(T,F)} \\ &= \frac{1}{V(T,F)J(F)}\end{aligned} \tag{10}$$

となり,$0 \leq \varepsilon \leq 1$ の間の値をとる.
漸近分散 $V(T,F)$ が小さく $J(F)^{-1}$ に近い程 ε は大きくなることから,ε が 1 に近い T_n ほど望ましい.([9] 参照)

3.3 *Breakdown point*

外れ値は, 最小 2 乗推定量に大きな影響を与える．これに対してデータの中に外れ値が一定の割合含まれていてもそのデータを扱うことが出来る推定量があることがわかった．これを形式化する目的で *breakdown point* の概念が導入された．その古い定義 (Hodges 1967) は一次元に制限されていた．Hample (1971) はこれを一般化したが, その定義は漸近的であり，数学的であった．*breakdown point* の有限標本形を取り扱う方が好ましいので以下これについて述べる．Z は n 個の標本データの点からなる集合を表すとする．

$$Z = \{(x_{11},\cdots,x_{1p},y_1),\cdots,(x_{n1},\cdots,x_{np},y_n)\} \tag{11}$$

T は回帰推定量とし標本 Z から回帰係数のベクトル θ を推定するものである．

$$T(Z) = \hat{\theta}. \tag{12}$$

元のデータの中の m 個を任意の値に置き換えること（これは非常に悪い外れ値を許す）によって得られたデータを Z' とし、汚れた標本 Z' を考える．このような汚染が引き起こす偏り $bias\,(m;T;Z)$ の最大を

$$bias\,(m;T,Z) = sup\|T(Z') - T(Z)\| \tag{13}$$

で表す，もし $bias\,(m;T;Z)$ が無限ならばこれは m 個の外れ値は T において大きな影響を与えることを意味する．これは推定量の"*breaks down*"といえる．それゆえ，標本 Z における推定量 T の *breakdown point* は

$$\epsilon_n^* = min\left\{\frac{m}{n}; bias\,(m;T,Z) = \infty\right\} \tag{14}$$

として定義される．言い換えれば推定量 T が $T(Z)$ から任意に離れた値をとるための最小の汚染の割合である．この理論は確率分布を含まない．最小 2 乗法の場合，1 つの外れ値は T を動かすのに十分である．ゆえにその *breakdown point* は

$$\epsilon_n^*\,(T,Z) = \frac{1}{n} \tag{15}$$

となる．これは標本の大きさ n の増加によって 0 へ近づく．よって最小 2 乗推定量は 0% の *breakdown point* といえる．これは，最小 2 乗法が外れ値に対して極端に敏感であることを示している.([14] 参照)

3.4 様々なロバスト推定量

ロバスト推定量はこれまでいくつか提案されており, L1 推定量 (least absolute values estimator),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,S 推定量などがある.

3.4.1 L1 推定量

ロバスト推定量としての最初のものは最小絶対値推定量 (Least absolute values estimator):$\hat{\theta_{L1}}$ である．これは次のように定義される.

$$\sum_{i=1}^{n}|r_i(\hat{\theta}_{L1})| = minimize_{\hat{\theta}}\sum_{i=1}^{n}|r_i| \tag{16}$$

これは LS 推定量と異なり y_i の外れ値に対しては防御するが,*leverage point*(梃子比または作用点)x_i に対しては無防備であり，x_i は線形式の当てはめに大きな影響を与えてしまう上に LS と同じく *breakdown point* は,$\frac{1}{n}$ である.

3.4.2 M 推定量

M 推定量は L1 推定量よりも y_i の外れ値に関してロバストであり

$$\sum_{i=1}^{n}\psi(\frac{r_i}{\hat{\sigma}})x_i = 0 \tag{17}$$

と定義される．しかしながら,x_i の外れ値の影響によって *breakdown point* は $\frac{1}{n}$ で標本が大きくなると 0 へ近づく.

3.4.3 GM 推定量

leverage point の弱さを克服するために一般化 M 推定量 (GM 推定量) が提案された．GM 推定量は

$$\sum_{i=1}^{n}w(x_i)\psi(\frac{r_i}{\hat{\sigma}})x_i = 0 \tag{18}$$

と定義される．しかし，これも次元の増加と共に *breakdown point* は減少する.([14] 参照)

3.4.4 Repeated median 推定量

中央値に基づいたロバスト回帰は Theil(1950), Brown,Mood(1951),Sen(1968),Andreews(1974) などによって考えられてきた．repeated median は 50% の *breakdown point* を持ち, 多くの他のロバスト手法よりも *breakdown point* が高い．中央値の演算子 M は

$$M\left\{\tilde{\theta}\left(i_1,\cdots,i_p\right)\right\} = med\left\{\tilde{\theta}\left(i_1,\cdots,i_{p-1},j\right)\right\} \quad (19)$$

によって定義される. ここで中央値は $\tilde{\theta}\left(i_1,\cdots,i_{p-1},j\right)$ の j を $\{1,\cdots,n\}-\{i_1,\cdots,i_{p-1}\}$ の上で動かして求められる．M が適用されるごとに変数の数が 1 つずつ減少する．これは

$$\hat{\theta} = M^p\left\{\tilde{\theta}\left(i_1,\cdots,i_p\right)\right\} \quad (20)$$

のように θ の repeated median 推定値を定義することを可能にする.([18] 参照)

3.4.5 LMS 推定量

LMS 推定量 (least median of squares)$\hat{\theta}_{LMS}$ は、Rousseeuw(1984) により, 高い破綻点を得るように Hample(1975) のアイディアを元に提案されたもので

$$med_i r_i^2(\hat{\theta}_{LMS}) = \min_{\hat{\theta}} med_i r_i^2(\hat{\theta}) \quad (21)$$

により定義される．ここで $med_i r_i^2$ は残差の二乗 r_i^2 の中央値である．この推定量は LS や L1 と異なり，y の外れ値と同様に x の外れ値に関してもロバストであり，*breakdown point* は 50% である．

3.4.6 LTS 推定量

LTS 推定量（least trimmed squares estimator）$\hat{\theta}_{LTS}$ は, Rousseeuw(1985) により

$$\sum_{i=1}^{h}(r^2(\hat{\theta}_{LTS}))_{i:n} = \min_{\hat{\theta}}\sum_{i=1}^{h}(r^2(\hat{\theta}))_{i:n} \quad (22)$$

により定義される．$(r^2)_{1:n} \le \cdots \le (r^2)_{n:n}$ のように残差の 2 乗を小さいほうから並び替えた h 番目の和を最小とする $\hat{\theta}$ である．LTS と LS は似ているが，大きい残差が和に含まれないことで外れ値を避ける事ができるので, その影響を受けなくなる．この推定量の *breakdown point* は h が $[\frac{n}{2}]+1$ の時に，50% に達する.([14] 参照)

3.5 実データの分析

アンモニアから硝酸へ酸化する影響力のデータを扱う.([8] 参照) このデータは 21 個の 4 次元の観測値を含む．Stackloss(y) は, 情報率 (x_1), 入り江の冷たい水の温度 (x_2), 酸の濃縮度 (x_3) の変数を使って解析した.

分析結果

<LS>

図 1 は，最小 2 乗法での残差プロットであるが，すべての観測値が -2.5 から 2.5 の間にランダムに分布しており，外れ値はないように思われる．このプロット図から，このデータセットはまったく外れ値を含まないと結論を出してしまいそうである．決定係数も 0.914, 自由度修正決定係数は,0.898 となり, よい適合となっている.

表 1　回帰直線

手法と回帰直線
LS
$\hat{y} = -39.920 + 0.716x_1 + 1.295x_2 - 0.152x_3$
LMS
$\hat{y} = -34.250 + 0.714x_1 + 0.357x_2 + 0.000x_3$
LTS
$\hat{y} = -34.956 + 0.740x_1 + 0.345x_2 - 0.006x_3$

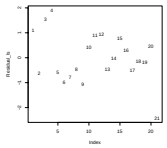

図 1　LS の残差プロット

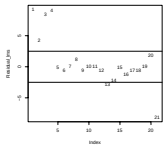

図 2　LMS の残差プロット

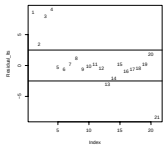

図 3　LTS の残差プロット

<LMS>

図 2 は，LMS 法での残差プロットである．LMS の残差は 0 のまわりにランダムに分布しており，-2.5 から 2.5 の帯に収まっているデータもある．しかし，いくつかの観測値が -2.5 から 2.5 の帯から離れたところにあり，外れ値の存在を表していることもわかる．観測値 $1,2,3,4,13,21$ が外れ値であることは明らかである.

<LTS>

図 3 は，LTS 法での残差プロットである．LTS の残差は, LMS の残差プロットとほぼ等しい．この図からも，観測値 $1,2,3,4,13,21$ は外れ値となるのは明らかである．このデータでは，LMS 法と LTS 法でほとんど差はなかった.

これより観測値 $1,2,3,4,13,21$ が外れ値であることは明らかである．この例で，ロバスト回帰分析法での残差プロット図と最小 2 乗法の残差プロット図は大きく違い，最小 2 乗法の残差プロット図はあまり信用できないことがわかる．ここで，LMS 法と LTS 法両方で明らかになっ

た顕著な外れ値 $1,2,3,4,13,21$ を除いて推定してみると, 回帰直線は以下のようになる.

$$\hat{y} = -34.058 + 0.757x_1 + 0.453x_2 - 0.052x_3 \qquad (23)$$

このときの決定係数は 0.962, 自由度修正決定係数は 0.952 となり, 当てはまりは良くなった.
今回, 私たちは観測値 (1,2,3,4,13,21) の 6 つを外れ値として検出することに成功した. この結果, 最小 2 乗法では外れ値が一つでも存在すれば影響を受けてしまうことや, 外れ値が外れ値を覆い隠してしまうこと, ロバスト回帰分析法の LMS 推定量, LTS 推定量は, 外れ値の影響を受けず, 外れ値の存在を明らかにすることがわかった.

3.6 ロバスト回帰の特徴

- ロバスト回帰の目的は, 仮定からのずれに対する防衛と安全をはかることである.
- データの少数派 (悪いデータ) を無視またはウェイトを小さくして, 多数派 (よいデータ) に当てはめる.
- 外れ値に影響されにくい手法.
- 悪いデータが 50% に近づくと良いデータとの区別がつかなくなる. この場合には,2 つの部分グループに分けるのが望ましい.

4 回帰診断

4.1 回帰診断とは

外れ値を含むデータを扱う別の方法として, 回帰診断がある. ロバスト回帰は外れ値のウエイトを下げることによって, 外れ値の影響を受けないようにした. それとは逆の方法で, 回帰診断は回帰モデル式の妥当性や, 誤差に関する仮定の正当性をチェックすることで, 仮定からのずれや外れ値を回避する方法である.

4.2 残差プロットと正規 Q-Q プロット, 残差の検定

独立同一正規誤差と言う標準的仮定の是非を簡単に確認できるのが, 残差プロットと正規 Q-Q プロットである. 正規 Q-Q プロットは点がほぼ直線状に並べば, 与えられたデータが正規分布に近い分布をしていることになる.
ダブリュ統計量 (Shapiro and Wilk's W-statistic) はさらに残差を分析する方法で, 正規性の仮説 H_0:the residuals are normal (この残差は正規である) に対する検定は,

$$W = \frac{1}{s^2}\left\{\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{n,m}(x_{(m)} - x_{(i)})\right\}^2, \ m = n - i + 1 \qquad (24)$$

として表される. ここで,s^2 は全体の平方和で,$\left[\frac{n}{2}\right]$ は $\frac{n}{2}$ の整数部分を指し,$a_{n,m}$ は定数で, 別に数表が用意されている.$x_{(i)}$ は i 番目の順序統計量を表わす.

4.3 影響力の大きなデータのチェック

4.3.1 ハット行列 (Hat Matrix)

通常のモデルは $E(\varepsilon) = \mathbf{0}$ で $cov\,(\varepsilon) = \sigma^2\boldsymbol{I}$ の仮定の下で, $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{\varepsilon}$ として与えられる. ここで $\boldsymbol{y}$ は $n\times 1$, $\boldsymbol{X}$ はランクが $k+1 < n$ の $n\times(k+1)$ 行列,$\boldsymbol{\theta}$ は $(k+1)\times 1$ である.$\boldsymbol{X}$ は確率変数ではないことを仮定する.
最小 2 乗推定量 $\hat{\boldsymbol{\theta}} = \boldsymbol{X(X'X)^{-1}X'y}$ を使って予測値 $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$ のベクトルは,

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X(X'X)^{-1}X'y} \\ &= \boldsymbol{Hy}\end{aligned} \qquad (25)$$

$\boldsymbol{H} = \boldsymbol{X(X'X)^{-1}X'}$ で,$\boldsymbol{H}$ は $\boldsymbol{y}$ から $\hat{\boldsymbol{y}}$ へ変換するので, ハット行列又は射影行列と呼ばれる. ハット行列 $\boldsymbol{H}$ は対称でベキ等である.

4.3.2 梃子比 (Leverage)

行列 $\boldsymbol{H}$ の第 (i,i) 要素 h_{ii} とおくと, 個々の残差の推定値の分散は

$$Var(\hat{\varepsilon}) = \sigma^2(1 - h_{ii}) \qquad (26)$$

となる. これより,h_{ii} が 1 に近ければ $Var(\hat{\varepsilon})$ は本来の分散値 σ^2 よりも小さくなり,$\hat{y}$ は y に近づくことが予想される. 値 h_{ii} を梃子比と呼ぶ.
h_{ii} の平均的な値は $(k+1)/n$ と見積ることができ, 実践的見地からは以下のように判断される.
$0.2 < h_{ii} \leq 0.5 \longrightarrow$ 対応データは危険
$0.5 < h_{ii} \longrightarrow$ 対応データは解析からは除外する
ただし梃子比が大きいことだけで, 対応する y_i が外れ値であると判断するのは困難である.

4.4 残差のばらつきの一様性のチェック

小さい残差を加えた h_{ii} の大きな値の追加の検証は

$$1/n \leq h_{ii} + \frac{\hat{\varepsilon}_i^2}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}} \leq 1 \qquad (27)$$

によって与えられる. 同じ分散を持つように残差を調整することが好ましい. 変数の比を調整する 2 つの共通の方法がある.
変数の比を調整する初めの方法は,$Var(\hat{\varepsilon}) = \sigma^2(1 - h_{ii})$ を使って平均 0 分散 1 の標準化残差 $\hat{\varepsilon}/\sigma\sqrt{1 - h_{ii}}$ を得る.σ と s を交換することによって, 内的にスチューデント化された残差 (internally studentized residual)

$$r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1 - hii}} \qquad (28)$$

を得る. ここで,$s^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}\hat{\boldsymbol{\theta}})^2$
2 つ目の方法は,i 番目の観測値を削除して得られる σ の推定値

$$t_i = \frac{\hat{\varepsilon}_i}{s_{(i)}\sqrt{1 - hii}} \qquad (29)$$

を使う. ここで,$s_{(i)}$ は $(y_i, \boldsymbol{x_i'})$ を省いた後, 残った n-1 個の観測値で計算したものである. これをスチューデント化残差 (studentized residual) または, 外的にスチューデント化された残差 (externally dtudentized residuals) と呼ぶ. スチューデント化残差は母集団分散が全て 1 の量になるはずなので, 残差そのものを見るより, 異常を発見し

やすい. スチューデント化残差をプロットすることによって, 絶対値が 2 よりも大きかったり, 一定の増加傾向等のパターンが見られれば, 誤差の分散が一定という仮定が疑わしくなる.

4.5 個々のデータの影響度チェック

1 つのデータが推定された回帰モデルパラメータに大きく影響を与えることがある.i 番目の観測値 $(y_i, \boldsymbol{x_i'})$ を削除して回帰分析を行うことによってえられた推定値

$$\hat{\boldsymbol{\theta}}_{(i)} = (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{y}_{(i)} \tag{30}$$

と全データを用いた推定値 $\hat{\boldsymbol{\theta}}$ との違いが i 番目のデータの回帰推定値への影響の大きさを示す. Cook の距離 (Cook's distance) によって,$\hat{\boldsymbol{\theta}}_{(i)}$ と $\hat{\boldsymbol{\theta}}$ を比較することができる. Cook の距離は

$$D_i = \frac{(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})}{(p+1)s^2} \tag{31}$$

として定義される. これは

$$\begin{aligned} D_i &= \frac{(\boldsymbol{X}\hat{\boldsymbol{\theta}}_{(i)} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{X}\hat{\boldsymbol{\theta}}_{(i)} - \boldsymbol{X}\hat{\boldsymbol{\theta}})}{(p+1)/s^2} \\ &= \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})'(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{(p+1)s^2} \end{aligned} \tag{32}$$

として書き直すことができる. このことから,D_i は $\hat{\boldsymbol{y}}_{(i)}$ と $\hat{\boldsymbol{y}}$ の間の通常のユークリッド距離に比例している. ゆえに D_i が大きいと観測値 $(y_i, \boldsymbol{x_i})$ は $\hat{\boldsymbol{\theta}}$ と $\hat{\boldsymbol{y}}$ 両方に対して大きな影響力を持っている.([13] 参照)

4.5.1 回帰診断の分析

データはロバスト回帰と同様,stackloss データを使用.

分析結果のそれぞれのプロット図

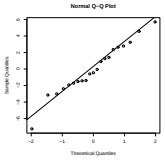

図 4　正規 1Q-Q プロット

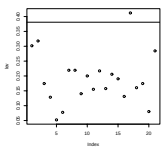

図 5　梃子比 1 プロット (h_{ii})

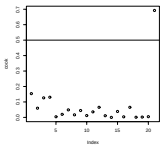

図 6　Cook1 の距離プロット (D_i)

誤差が全体的に大きめで Q-Q プロットを見ると, 正規性はかなり疑わしいように見える. よって回帰診断を続けた. 回帰診断をした結果, $h_{11} = .306 > 0.2$ で, 他の $\hat{\varepsilon}_i, r_i, t_i, D_i$ も異常には大きくないのだが若干大きい値なので, 外れ値の可能性があるかもしれない. 観測値 4 は $\hat{\varepsilon}_4 = 1.882, r_4 = 2.052, t_4 = 2.211$ とそれぞれ大きな値なので外れ値である可能性がある. 観測値 21 は $\hat{\varepsilon}_{21} = -2.638, r_{21} = -3.330, t_{21} = -3.352, D_{21} = 6.920e-01$ と, それぞれ基準値を大幅に上回っているのでこれは外れ値と言えるだろう. 以上のことから, 観測値 1,4,21 が外れ値である可能性がある.

観測値 1,4,21 を除いて再び回帰診断を行う.

分析結果のそれぞれのプロット図

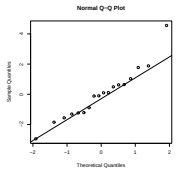

図 7　正規 2Q-Q プロット

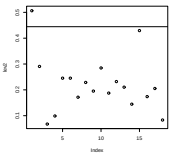

図 8　梃子比 2 プロット (h_{ii})

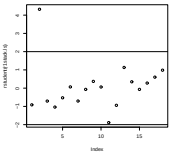

図 9　スチューデント化された 2 残差プロット (r_i)

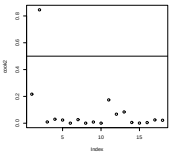

図 10　Cook2 の距離プロット (D_i)

Q-Q プロットより, 誤差の正規性があるとは全く言えない. よって, 回帰診断を続ける. 観測値 3 は $\hat{\varepsilon}_3 = 2.873, r_3 = 4.321, t_3 = 4.115, D_3 = .846$ で, 梃子比 h_{33} 以外の値が異常値なので, これは明らかに外れ値であると言える. よって観測値 3 を外れ値として分析を進める.
観測値 1,3,4,21 を除いて再び回帰診断を行う.

分析結果

W 検定

W=0.980　p-value=0.955

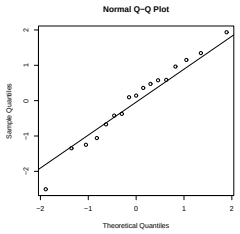

図 11　正規 Q-Q プロット

誤差は W 検定から, 正規性を満たしていると言える. このときの Q-Q プロットもわりと直線に当てはまっているように見える. よって誤差が正規分布を仮定できる場合には, 最小 2 乗法を適用するのが最良である. 決定係数もとても高く回帰式へのあてはまりが非常によいと言える. よって最適なモデルは, 観測値 1,3,4,21 を削除したもので, 回帰式は

$$\hat{y} = -37.652 + 0.798x_1 + 0.578x_2 - 0.067x_3$$

で, このときの決定係数は 0.975 でよい当てはまりといえる. 回帰診断では 4 つの外れ値を検出することができた.

5　おわりに

ロバスト回帰と回帰診断

- ロバスト回帰と回帰診断は同じ問題を反対側から考察し, どちらも重要で互いに相補的である.
- ロバスト回帰だけでは不十分であるし, 回帰診断だけでも不十分である.
- 外れ値の分析は回帰診断であり, ロバスト回帰ではない. しかし, ロバスト手法は外れ値に対して敏感ではない (抵抗力のある) ので, 外れ値の検出には力を発揮する.
- ロバスト回帰での手法は仮定のモデルの下で効率が低い (良くないことがある) ので, 注意する必要がある.
- データ分析においては, ロバスト回帰と回帰診断を個々に議論するのではなく, 総合的に判断する精神が大切である.
- 機械的な統計手法の適用は, 誤った結論をもたらすことも多い.
- データの持つ情報だけでなく, データの外の知識や情報も有効に活用すべきである.
- グラフなどを用いて, 視覚的な考察をすることは非常に有用である.

参考文献

[1] 安藤雅和. 木村美善:線形回帰モデルにおける S 推定量のバイアス:南山経営研究 第 11 巻 第 3 号 (1997).

[2] 浅井麻貴:回帰分析手法のロバストネスとその応用に関する研究, 南山大学数理情報学部数理科学科卒業論文 (2004).

[3] C.,D.and F.,W.S.:Fitting Equations to Data,Wiley New York(1980).

[4] Chatterjee,S., Hadi,A.and Price,B.:Regression Analysis By Example Third edition,INC(2000).

[5] Chatterjee,S.and Price,B.:回帰分析の実際, 新曜社 (1980).

[6] Faraway,J.J.:Linear Models with R, A CRC Press Company(2004).

[7] 藤木美江:回帰分析とその応用に関する研究, 南山大学経営学部情報管理学科卒業論文 (2001).

[8] Huber,P.J.:Between Robustness and Diagnostics, Directions in Robust Statistics and Diagnostics, Part I (Werner Stahel and Sanford Weisberg,Eds.)121-130,Springer-Verlag(1991).

[9] 金子元紀:ロバスト線形回帰, 南山大学数理情報学部数理科学科卒業論文 (2004).

[10] 中澤港:R による統計解析の基礎,Pearson Education Japan(2003).

[11] 間瀬茂・金藤浩司:工学のためのデータサイエンス入門, 数理工学社 (2004).

[12] 大見俊司:ロバスト回帰の理論とその応用－ Regression Depth を中心に－, 南山大学数理情報学部数理科学科卒業論文 (2004).

[13] Rencher,A.C.:Linear Models in Statistics,John Wiley&Sons,Inc(2000).

[14] Rousseew,P.J. and Leroy,A.M.:Robust Regression and Outlier Detection,Wiley,NewYork(1986).

[15] 佐久間彩:回帰分析法のロバストネスの理論とその応用に関する研究, 南山大学数理情報学部数理科学科卒業論文 (2004).

[16] 佐和隆光:回帰分析, 朝倉書店 (2002).

[17] 白旗慎吾:統計解析の入門, 共立出版 (2002).

[18] Siegel,A.F.: Robust regression using repeated medians,Biometrics,341-350(1982).

[19] 田中豊. 脇本和昌:多変量統計解析法, 現代数学社 (1998).

[20] 立松和明:ロバスト回帰におけるロバスト統計手法とその応用, 南山大学大学院修士 (経営学) 論文 (2002).