ロバスト回帰の研究

2003MM052 窪田修也
指導教員: 木村美善

1 はじめに

回帰分析を行う際の解析法として最小2乗法がよく知られている. 最小２乗法は外れ値の影響を強く受けるという欠点があり,この欠点を克服するために提案されたのがロバスト回帰法である.本研究では,実際のデータを用いて,最小２乗法で分析し,問題点を明らかにした上でロバスト回帰を適用し,その良さと特徴を考察する.

2 回帰分析

2.1 モデルの記述

目的変数yと，p個の説明変数x_1,x_2,$\cdots$,x_pに関するｎ個の観測値データがある.x_1,x_2,$\cdots$,x_p からyの値を予測するときの関係式として次の１次式を仮定する.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{1}$$

2.2 最小2乗法(LS)

最小2乗法とは,残差の平方和を最小にするような$\hat{\beta_0}$,$\cdots$,$\hat{\beta_p}$を求める方法である. y_iの予測値を

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_{1i} + \hat{\beta_2} x_{2i} + \cdots + \hat{\beta_p} x_{pi} \tag{2}$$

とするときの残差は

$$r_i(\hat{\beta}) = y_i - \hat{y_i} \tag{3}$$

と表される.

3 ロバスト回帰

3.1 ロバスト回帰とは

最小2乗法は線形回帰の標準的仮定のずれに敏感であり,外れ値が存在することによって大きな影響を受けてしまう. ロバスト回帰とはそれらの影響を受けにくく良さの損失が少ない推定法である.(参照[1][3][6])

3.2 ロバストネスの尺度

3.2.1 影響関数（influence function）

モデルF_0における推定量$T = T(F)$の影響関数は観測値xの関数として,次のように定義されるものである.

$$IF(x; F_0, T) = \lim_{t \to \infty} \frac{T[(1-t)F_0 + t\delta_x] - T[F_0]}{t} \tag{4}$$

ただし,δ_xはxで確率１をとる分布を表す. 影響関数の意味は,モデル分布F_0の下での推定量の値$T(F_0)$が1点xに観測値が追加されたとき,どれだけ変化するかを評価するものである.

3.2.2 漸近効率（asymptonic efficient）

Tに対して漸近正規性

$$\mathcal{L}_G(\sqrt{n}(T_n - T(G))) \to N(0, V(T,G))$$

が成り立つとする.漸近分散$V(T,G)$はIFによって

$$V(T,G) = \int IF(x;T,G)^2 dG(x) \tag{5}$$

と表される.分布FにおけるT_{n}の漸近効率は

$$e = \frac{\frac{1}{J(F)}}{V(T,F)} = \frac{1}{V(T,F)J(F)} \tag{6}$$

となり,$0 \leq e \leq 1$の間の値をとる.ただし$J(F)$はFisher情報量である. 漸近分散V(T,F)が小さく$J(F)^{-1}$に近いほどeは大きくなることから,eが１に近いほどT_nは望ましい.

3.2.3 破綻点（breakdown point）

大域的なロバストネスをはかる尺度として破綻点がある.Tをβの推定量とし

$$T(Z) = \hat{\beta} \tag{7}$$

とする.ｎ個からなるデータZの中のｍ個を,任意の値（かなり悪い外れ値を考慮に入れる）に置き換えたときのデータをZ'とする.この汚染によって生じる偏りの最大は

$$bias(m; T, Z) = sup_{z'} \parallel T(Z') - T(Z) \parallel \tag{8}$$

となる.$bias(m;T,Z)$が無限であるとき,これを推定量の破綻という.有限標本Zでの推定量Tの破綻点は

$$\varepsilon_n{}^* = min\{m/n; bias(m;T,Z) = \infty\} \tag{9}$$

と定義される.$0 \leq \varepsilon^* \leq 1/2$ であり,高い破綻点が望ましい.

3.3 様々なロバスト推定量

3.3.1 LMS推定量

Rousseeuwによって導入されたLMS推定量は

$$med_i {r_i}^2(\hat{\beta}_{LMS}) = \min_{\beta} med_i {r_i}^2(\beta) \tag{10}$$

により定義される.ここで$med_i {r_i}^2$は残差の2乗$r_1^2, \cdots, r_{\mathrm{n}}^2$の中央値である.この推定量は$y$の外れ値と同様に$x_1, \cdots, x_n$の外れ値に関してロバストである.

3.3.2 LTS推定量

LTSはRousseeuwによってLMSに手を加えたもので

$$\sum_{i=1}^{h} (r^2(\hat{\beta}_{LTS}))_{i:\mathrm{n}} = \min_{\beta} \sum_{i=1}^{h} (r^2(\beta))_{i:n} \tag{11}$$

により定義される.$(r^2)_{1:\mathrm{n}} \quad \leq \cdots \leq (r^2)_{n:n}$は残差の2乗を小さいほうから並び替えたものである.

4　データによる分析

ここでは参考文献[2]の中古住宅に関するデータを使用することとする.変数は以下の通りである．x_1 = 宅地面積,x_2 = 住宅延べ面積 ,x_3 = 築後経過年数,x_4 = 京都駅からのJR電車時間, x_5 = JR駅前からのバス時間,x_6 = 徒歩時間,y = 中古価格 とする.（参照[2][4][5]）

4.1　結果と考察

まず最小２乗法より回帰式

$$\hat{y}_{LS} = 16.291 + 0.066x_1 + 0.184x_2 - 0.252x_3 - 0.514x_4 - 0.624x_5 - 0.342x_6$$

を得る.

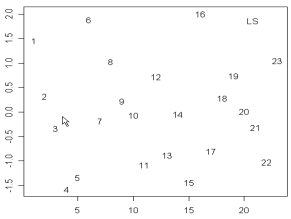

図 1: LS残差プロット

図１は標準化残差をプロットしたものである．LSの残差プロットを見ると外れ値はなく、残差に特に問題は見られない．このとき決定係数は0.9514で,自由度調整済み決定係数は0.9332である.この数値から分析はうまくいってるように見える．次に同じデータを使用し,回帰診断を行なった.図2,3はそれぞれ梃子比とCookの距離によるプロットしたものである.

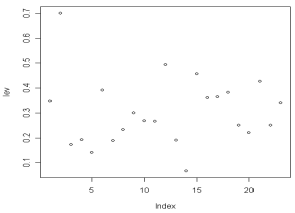

図 2: 梃子比プロット

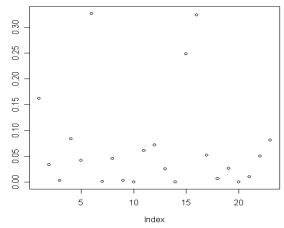

図 3: Cookの距離

回帰係数への影響を与えるCookの距離が0.5を超える目立った値は検出されなかったが, スチューデント化残差より観測番号6,16の２個が外れ値と考えられる値を取っている.そこで,これらの梃子比を調べた結果,いずれも危険と判断する値を取ってはいない．しかし,スチューデント化残差では外れ値として検出されなかった観測番号２の梃子比の値が,非常に大きな値を取っており,外れ値であると考えられる．この観測番号２のデータは他の家の値段に比べて飛びぬけて大きい.その価格59.5（百万円）は, ２番目に高い家の価格29.8の２倍で, 最も安い家の価格5.5の10倍以上あることが理由であると考えられる．以上のことから,観測番号２, ６ ,16が外れ値の可能性がある.

次にロバスト回帰を用いて外れ値の検出を行なった．図４と図５はそれぞれLMS,LTSを用いた場合の残差プロットである.

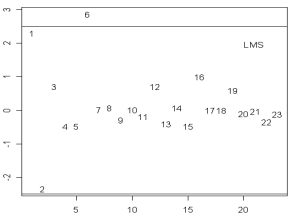

図 4: LMS残差プロット

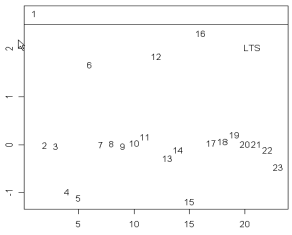

図 5: LTS残差プロット

LMS,LTSのプロットから外れ値の可能性があるものとして観測番号1,2,6,12,16,があげられる.これらの外れ値を除いて分析を行ったところ, 以下のようになった.　決定

表 1: 決定係数の変化

		決定係数	自由度修正決定係数
全データを使用		0.9514	0.9322
外れ値として除いた観測番号	1	0.9574	0.9404
	6	0.9617	0.9464
	1,2,6	0.955	0.9342
	1,6,16	0.985	0.9781
	1,6,12,16	0.9904	0.9855

係数がもっとも高いものは観測番号1,6,12,16を除いたもので決定係数は0.9904,自由度調整済み決定係数は0.9855となり, かなり当てはまりが良くなったと言える.
回帰診断では外れ値として検出されたのは３つだったのに対し,ロバスト回帰では５つの外れ値を検出した.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, どちらも重要であるということがわかる.

5　おわりに

今回の研究でロバスト回帰の誕生した過程を知るとともに,その利便性や必要性をさらに感じることが出来た.また機械的な統計手法の適用は誤った結論をもたらし,誤った認識を持ちかねないということも知り,さまざまな方法や方向から総合的に判断する必要があると感じた.

参考文献

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

[2] 芳賀敏郎・吉澤正: 多変量解析事例集　第１集, 日科技連出版社(1992)

[3] Julian J. Faraway:　Linear Models with R, Chapman& Hall/CRC(2004)

[4] 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）

[5] 間瀬茂・神保雅一・鎌倉稔成・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）

[6] 佐和隆光: 回帰分析, 朝倉書店(1979）