回帰分析手法のロバストネスとその応用に関する研究

2001MM004 浅井　麻貴　　2001MM045 前田　温子
指導教員　木村　美善

1　はじめに

多変量解析法には回帰分析, 主成分分析, 判別分析, 因子分析, クラスター分析, 数量化理論等がある．中でも最も基本的であり，理論がしっかりしており，最も広く実用されている回帰分析を 3、4 年次でのゼミの学習主題として理論と応用を学んだ．その中心的課題の一つは回帰係数の推定であるが，最小二乗法による推定量が通常用いられている．この回帰係数の最小二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，最良であるが，最小二乗推定量は標準的仮定からのずれに対して敏感であり，わずか一つの外れ値によっても大きな影響を受けてしまう．したがって，これらのずれや，外れ値の存在が想定される場合には，重大な間違いを引き起こすことにもなりかねない．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，標準的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（頑健な）推定量を用いることが望ましい．現実にはモデルの一部または全部が厳密に満たされることは少なく，せいぜい近似的に成り立つ程度のものである．それゆえ，現実問題への統計的手法の適用にあたってロバストネスの問題は避けて通ることのできない重要なものである．本研究の目的は，このような統計的手法のロバストネスについて理論と応用の両面から考察することである．([1], [3] 参照)
第 2 章：前田温子，第 3 章：浅井麻貴，第 4 章：共同

2　重回帰分析

一つの目的変数 y と p 個の説明変数 x_1,x_2,・・・,x_p について n 個のデータが与えられたとする．単回帰分析と同様に x_1,x_2,・・・,x_p から y の値を予測するとき，x_1,x_2,・・・,x_p と y の関係を示す一つの数式モデルを設定しなければならない．

$$y_i = a_0 + a_1x_{1i} + a_2x_{2i} + ・・・ + a_px_{pi} + e_i, \quad (1)$$
$$i = 1,・・・,n$$

この式を線形重回帰モデル（linear multiple regression model) と呼ぶ．単回帰の場合 (x,y) 平面上の n 個の点の集まりに直線を当てはめたが，重回帰の場合には (x_1,・・・,x_p,y) の (p+1) 次元空間での n 個の集まりに対して p 次元超平面

$$y = a_0 + a_1x_1 + a_2x_2 + ・・・ + a_px_p \qquad (2)$$

を当てはめ, それによって説明変数の値 x_{1i},x_{2i},・・・,x_{pi} から目的変数の y_i を予測する。

$$\sum_{i=1}^{n} e_i^2 = \{y_i-(a_0+a_1x_{1i}+a_2x_{2i}+・・・+a_px_{pi})\}^2 \quad (3)$$

を最小にする a_1,a_2,・・・,a_p の値を $\hat{a}_1$,$\hat{a}_2$,・・・,$\hat{a}_p$ と書く．ここで，誤差項 e は，説明変数 x によって説明されない y の変動部分を表す．誤差 e について，次の仮定をおく．

1. 不偏性：e_i の期待値はゼロ ($E(e_i)$=0)
2. 等分散性：e_i の分散は一定 ($V(e_i)$=σ^2)
3. 無相関性：e_1,・・・,e_n は互いに無相関

([6] 参照)

3　ロバスト手法の種類について

線形回帰モデル

$$y_i = x_i'\theta + u_i \qquad i = 1,・・・,n$$

を考える．ここで x_i=(x_{1i},・・・,x_{pi})′ は p 次元確率ベクトル，θ=(θ_1,・・・,θ_p)′ は p 次元回帰母数ベクトル，u_i は確率誤差，y_i は従属変数とする．標本 (x_1,y_1),・・・,(x_n,y_n) に基づき，回帰母数 θ を推定することが目的である．残差を $r_i(\theta)$=y_i − $x_i'\theta$ とするとき、最小二乗推定量 (least squares estimator)$\hat{\theta}_{LS}$ は

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

を満たすものである．$\hat{\theta}_{LS}$ は u_i に関する標準的仮定のもとでは良い推定量であるが，外れ値に対して敏感に反応し，ロバストではない．わずか一つの外れ値であっても結果に大きな影響を与えてしまう．

ロバスト推定量としても最初のものは最小絶対値推定量 (least absolute values estimator)$\hat{\theta}_{L1}$ であり，これは

$$\sum_{i=1}^{n} \mid r_i(\hat{\theta}_{L1}) \mid = min_\theta \sum_{i=1}^{n} \mid r_i(\theta) \mid$$

で定義される．この $\hat{\theta}_{L1}$ は y_i の外れ値に対しては防御するが，線形式の当てはめに大きな影響を与える x_i すなわち作用点 (leverage point) に対して無防備であり，$\hat{\theta}_{LS}$ と同じく $\epsilon_n^*(\hat{\theta}_{L1},X)$=$\frac{1}{n}$ である．

Rousseeuw(1984) はさらに高い破綻点を得るために LMS 推定量 (least median of squares estimator)$\hat{\theta}_{LMS}$ を提案した．これは

$$median(r_1^2(\hat{\theta}_{LMS}),・・・,r_n^2(\hat{\theta}_{LMS}))$$

$$= min_\theta median(r_1^2(\hat{\theta}), \cdots, r_n^2(\hat{\theta}))$$

により定義され，Hampel(1975) のアイデアに基づいたものである．$\hat{\theta}_{LMS}$ は $\epsilon_n^*(\hat{\theta}_{LMS},X)=([\frac{1}{n}] - p+2)/n$ であり，$n\to\infty$ のとき $\epsilon_n^*(\hat{\theta}_{LMS},X)\to\frac{1}{2}$ で，収束のオーダーが $n^{-1/3}$ と遅い．

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

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

によって定義した．ここで h=$[\frac{n}{2}]$+1 であり，$r(\theta)_{i:n}^2$ は $r_1^2(\theta), \cdots, r_n^2(\theta)$ の第 i 番目順序統計量である．$\hat{\theta}_{LTS}$ の有限標本破綻点は $\hat{\theta}_{LMS}$ と同じであり，収束のオーダーが $n^{-1/2}$ である．([2], [4] 参照)

3.1 hについて

h について，
一般化された h の式：$h = [\frac{n}{2}] + [\frac{(p+1)}{2}]$
ここで，$p \geq 1$（p:次元）なので，$h \geq [\frac{n}{2}] + 1$ となる．
以上のことをもとに，h のいくつかの値に対して LTS 推定量の値を求めてみる．

表 1　10 ヶ月の血圧測定 (Rosner(1977))

1	2	3	4	5	6	7	8	9	10
40	75	80	83	86	88	90	92	93	95

- h=$[\frac{n}{2}]$ のとき
  $h = 5$(上記の h の条件には当てはまらず，例外として)
  $Minimize_{\hat{\theta}} \sum_{i=1}^{5} (r^2)_{i;n}$：$\hat{\theta}$=91.6
- h=$[\frac{n}{2}] + 1$ のとき
  $h = 6$
  $Minimize_{\hat{\theta}} \sum_{i=1}^{6} (r^2)_{i;n}$：$\hat{\theta}$=90.7
- h=$[\frac{n}{2}] + 2$ のとき
  $h = 7$
  $Minimize_{\hat{\theta}} \sum_{i=1}^{7} (r^2)_{i;n}$：$\hat{\theta}$=89.6

ここで，LTS 推定量を求めるに当たって，一つの基準となる LMS 推定量 ($[\frac{n}{2}] + 1$) を求めると，

- $[\frac{n}{2}] + 1$=6
  T=$\frac{86+95}{2}$：$\hat{\theta}$=90.5

よって，LTS 推定量は 90.7 の h=$[\frac{n}{2}] + 1$ の時であることが分かる．
この結果から，h が $[\frac{n}{2}]$+1 のときの LTS 推定量の値 90.7 が LMS 推定量の値 90.5 に最も近い値をとることが見てとれる．別のデータにおいても同様の結果を得た．
よって，$h = [\frac{n}{2}] + 1$ と言える．([4], [5] 参照)

4　重回帰データ

4.1　塩分濃度のデータ

ここで用いるデータは Ruppert and Carroll(1980) で取り上げられたもので，ノースカロライナ州のパムリコ湾の塩分濃度を他の 3 つの変数をもとに関係を探るものである．
i=観測番号，x_1=2 週間分の貯まった塩分濃度，x_2=季節，x_3=海峡への川の流出量，y=塩分濃度

表 2　Salinity Data

(i)	(x_1)	(x_2)	(x_3)	(y)
1	8.2	4	23.005	7.6
2	7.6	5	23.873	7.7
3	4.6	0	26.417	4.3
4	4.3	1	24.868	5.9
5	5.9	2	29.895	5.0
6	5.0	3	24.200	6.5
7	6.5	4	23.215	8.3
8	8.3	5	21.862	8.2
9	10.1	0	22.274	13.2
10	13.2	1	23.830	12.6
11	12.6	2	25.144	10.4
12	10.4	3	22.430	10.8
13	10.8	4	21.785	13.1
14	13.1	5	22.380	12.3
15	13.3	0	23.927	10.4
16	10.4	1	33.443	10.5
17	10.5	2	24.859	7.7
18	7.7	3	22.686	9.5
19	10.0	0	21.789	12.0
20	12.0	1	22.041	12.6
21	12.1	4	21.033	13.6
22	13.6	5	21.005	14.1
23	15.0	0	25.865	13.5
24	13.5	1	26.290	11.5
25	11.5	2	22.932	12.0
26	12.0	3	21.313	13.0
27	13.0	4	20.769	14.1
28	14.1	5	21.393	15.1

表 3　回帰分析（最小二乗法）

変数	係数	標準誤差	t 値	p 値
(intercept)	9.59026	3.12509	3.069	0.00527
x_1	0.77711	0.08622	9.013	3.5×10^{-9}
x_2	− 0.02551	0.16108	− 0.158	0.87548
x_3	− 0.29504	0.10680	− 2.762	0.01083
R^2=0.8264				

表 4　回帰分析（LMS）

変数	係数
(intercept)	40.8877
x_1	0.3312
x_2	− 0.1377
x_3	− 1.4678
R^2=0.976	

表 5　回帰分析（LTS）

変数	係数
(intercept)	36.9751
x_1	0.3847
x_2	− 0.1157
x_3	− 1.3187
R^2=0.960	

考察

データの表を見ると観測値 5 と 16 は他の観測値と比べ，とても激しい流出量だということがわかる．よって観測値 5 と 16 は外れ値になるだろうと推定できる．また，正規性の検定として，Shapiro-Wilk 検定を行ったところ，有意確率が 0.158 ＞有意水準 0.05 で正規分布であるという仮説は棄却されない．すなわち，この検定では正規分布でないということはいえない．しかし，Q-Q プロット図から，直線近くにデータが載っているとは言えず，とても正規性に従っているとは言えない．よって，このデータには Shapiro-Wilk 検定は対応できておらず，正規分布に従っていないことが言える．LS 法を使って求めた回帰式は

$$\hat{y} = 0.777x_1 - 0.026x_2 - 0.295x_3 + 9.59$$

で与えられ，最小二乗法での残差のプロット図をみると，すべての観測値が− 2.5 から 2.5 の間にあり，外れ値はないように思われる．次に，データの表から外れ値と思われていた観測値 5 と 16 に注目してみた．ここで Carroll と Ruppert(1985) は，観測値 3 と 16 が外れ値と思われる観測値 5 を覆い隠していると考えた．実際，観測値 3 と 16 のデータを取り除いてもう一度最小二乗法を行ってみたら，観測値 5（図 4 では 4 になる．) は最初の最小二乗法での残差プロット図より，離れたところに移動したが数値的には外れ値とは言いがたい．そこで，私たちは最小二乗法での残差のプロット図 (図 1) を見て，− 2.0 から 2.0 より離れたところに位置する観測値 16 と 17 を除いてもう一度最小二乗法を行ったところ，観測値 3 と 16 のデータを取り除いた場合の時と同様，数値的には外れ値とは言いがたかった．従って，最小二乗法では，外れ値を見つけにくい．

一方，ロバスト回帰分析の中の LMS，LTS 法はこの外れ値に影響されない．LMS 法を使って求めた回帰式は

$$\hat{y} = 0.331x_1 - 0.138x_2 - 1.468x_3 + 40.888$$

LTS は

$$\hat{y} = 0.385x_1 - 0.116x_2 - 1.319x_3 + 36.975$$

LMS，LTS 法での残差プロット図をみると，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示していることがわかる．観測値 5 と 16 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二乗法の残差プロット図は大きく違い最小二乗法の残差プロット図はあまり信頼できないことがわかる．また，最小二乗法では外れ値の影響を覆い隠してしまうがロバスト回帰分析法では外れ値の存在を明らかにするということがわかった．([5] 参照)

Shapiro-Wilk 検定 (W = 0.96, p-value = 0.158)

図 1　QQ プロットと残差プロット (LS)

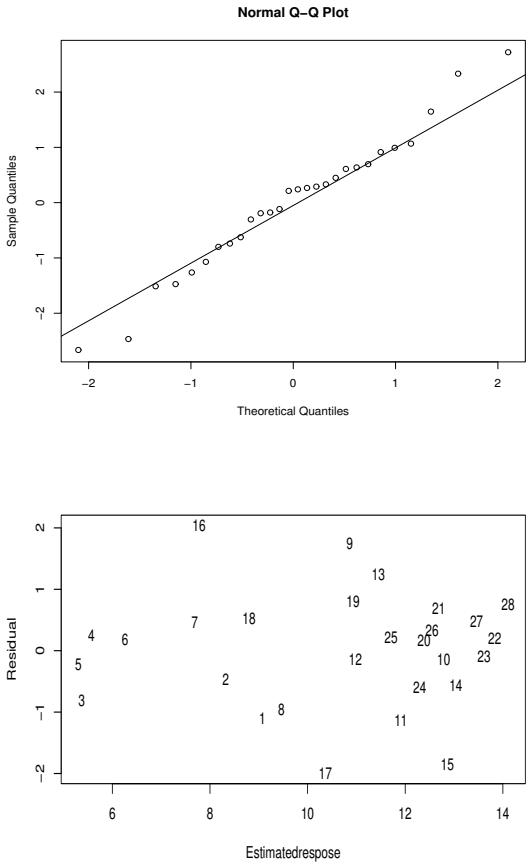


図 2　残差プロット (LMS)

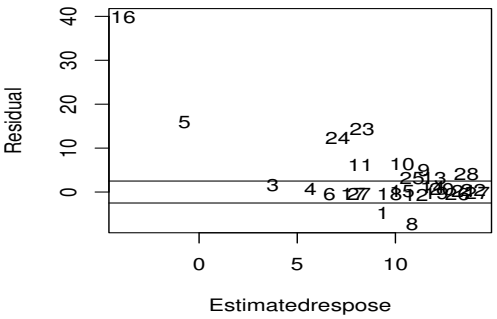


図 3　残差プロット (LTS)

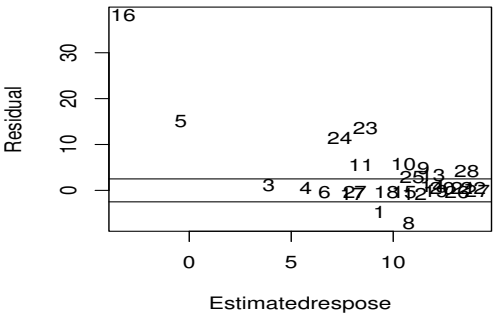
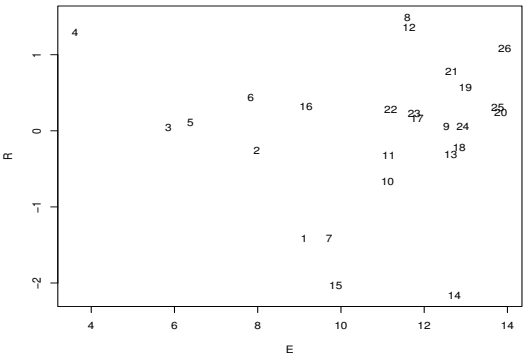
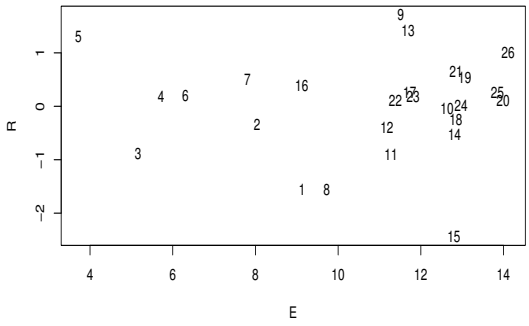


図 4　LS:観測値 3 と 16 を除いた場合の残差プロット



観測値 3 と 16 を除いた場合の寄与率：0.8754
観測値 16 と 17 を除いた場合の寄与率：0.9096

図 5　LS:観測値 16 と 17 を除いた場合の残差プロット



5　おわりに

統計学の理論を学んでいく際に，線形代数、微積分学などの数学の知識がとても重要になってくることがわかった．そして，一つ一つ，定理を証明して理解することの難しさを学んだ．一番苦労したことは，ロバスト回帰とは何かを示せるような分析結果が出せるデータを探すことで，結局，参考文献のデータを使った．インターネットなどで,「ロバスト」について調べてみたが，日本のサイトでは取り上げられている研究が少なく，海外のホームページを見ることになった．また，日本語で書かれた文献もなく今回の研究で英語の論文を読んで理解することの難しさを痛感した．

本論文を作成するにあたり，熱心にご指導いただきました，木村美善教授，安藤雅和先生，その他協力して頂いたすべての方に深く感謝します．

参考文献

[1] 安藤雅和：線形回帰モデルにおけるロバスト推定量の研究，南山大学経営学研究科修士論文 (1996).

[2] 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，Vol. 11，No. 3，568-570(1997).

[3] Chatterjee，S. and Price，B: Regression Analysis by Example，Wiley, New York(1977).

[4] Rousseeuw，P. J：Least median of square regression，Journal of American Statistical Association，Vol. 79，871-880(1984).

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

[6] 田中　豊・脇本和昌：多変量統計解析法，現代数学社 (1983).