ロバスト化残差の効用

～誤差分布に関する正規性の検討

野　口　和　也

本論は，線型モデルを用いて最小二乗推定を行なう際の前提条件である誤差分布型の正規性についての検討様式について論じたものである。最初に，分布型が未知な場合の推定方式はロバストでなければならないことを強調し，次には残差分析において残差をロバスト化する方法について述べる。さらに正規性の検討様式として，*exploratory mode* を導入し，最後に若干の数値例を用いてロバスト化残差の効用を示す（**本論文は，昭和57年7月に千葉大学で開催された，第50回日本統計学会における筆者の報告「trimmed-β による残差プロット」に基づくものです**）。

1. ロバスト推定

a）　ロバスト推定の必要性

今，30cm の定規を用いて，机の長さをミリ単位までの精度で測定する問題が生じたとしよう。このとき，ただ一回の測定では心許無いから，何回かくり返し測定してみることにする。測定を n 回行なうことにすれば，測定値 $X_1, X_2, \cdots, X_n$ が得られ算術平均 $\bar{X}=\frac{1}{n}\Sigma X_i$ を机の長さと見なすことができる。

このような行為を統計理論の側に立って説明してみると，まず各々の測定値は次のような構造を持っている。

$$(1.\ 1)\quad X_i=\theta+\epsilon_i,\ i=1,\cdots,n$$

θ は机の真の長さ（未知の値）を表わし，ϵ_i は測定に因る誤差である。算術平均は，この測定誤差の2乗和 $\Sigma\epsilon_i^2$ を最小にするような推定量（**最小二乗推定**

量）である。いうまでもなく，θの推定量には$\bar{X}$だけでなく他にも多くの推定量が存在するが，各々の測定誤差 ϵ_i が，独立に平均0の正規分布に従うと仮定した場合，$\bar{X}$に優るものはない。問題は，果たして測定誤差が平均0の正規分布に従うかどうかである。もし，定規の長さが30cmでなく，実際には30.5 cmあったとすれば平均ゼロとはならないし，あるいは個人的なクセによって平均的に大きめに測定する人もいれば，少なめに測定する傾向を持つ人もいるであろう。しかし，こういった問題には立ち入らず，ここでは測定誤差の正規分布性のみを検討することにしよう。もともと，正規分布[1]は，K. F. Gauss（1777-1855）の天体観測による測定誤差の研究[2]に端を発している。さらにA. Quetelet（ケトレー；1796-1874）は1835年に，人間の身長，体重，胸囲，腕の長さ，etc. の分布が正規分布をしていることを実証した。このことから，他の多くの測定値（観測値）の分布も正規分布をしていると考えられるようになり，20世紀になって始まる推測統計の方法は，多くがこの正規分布の仮定から出発している。

ここで問題なのは，ガウスやケトレーの時代には，いわゆる「ガリレオの方法」が広く実践されており，科学の対象は質量，長さ，速度といった定量的性質だけであり，色，におい，味，嗜好といったような性質は“観念の産物”として科学の対象から外されていた，という事実である。その結果，社会科学や行動科学といったような，この“観念の産物”を直接，あるいは間接的に対象とするような分野では，正規性の仮定の妥当性を疑問視[3]して，もし分布型が正規分布からはずれていても，推定効率がそれほど著しく低下しないような統計手法の開発が必要となった。それが robust statistics と呼ばれるものである[4]。

b） ロバスト推定の方法

パラメータ θ の推定量としての算術平均は2乗和 $\Sigma(X_i-\theta)^2$ を最小にするような最小二乗推定量である。最小二乗推定量は，観測値それ自身の θ からの距離をウェイトとして使用しているため，θ から遠く離れれば離れるほどその

観測値が推定量に及ぼす影響は強まる。そこで，θ からの距離を2乗する代りに，適当な凸関数 ρ を用いて，

(1. 2) $\Sigma\rho(X_i-\theta)\to\min$ あるいは $\Psi(X_i-\theta)\to 0$ $(\Psi=\rho')$

となるように推定量を選べば，中心からの距離が大きくなるにしたがってウェイトを小さくすることができる。問題は，凸関数 $\rho(u)$ の型であるが，$\rho(u)=|u|$ とおけば，$\Sigma|X_i-\theta|\to\min$ を満たすのは $\theta=X\text{med}$ となる（med は中央値）ことから，$\rho(u)$ は u^2 と $|u|$ との間にあるような関数にすればよい。このようなタイプの推定量は Huber(1964) 以来，現在まで様々な型のものが考案されている。これらは最尤推定（most liklihood estimation）の考え方に基づくことから，M-推定量と呼ばてれいる。代表的なものとして，Huber-type と Tukey の biweight と呼ばれるものを挙げておこう。

(1. 3) Huber type M-estimator

$$\rho(u)\begin{cases}=u^2/2 & |u|\leqslant c\\ =c|u|-c^2/2 & |u|>c\end{cases}$$

$$\Psi(u)\begin{cases}=u & |u|\leqslant c\\ =c\cdot sgn(u) & |u|>c\end{cases}$$

(1. 4) Tukey's biweight type M-estimator

$$\rho(u)\begin{cases}=\dfrac{c^2}{6}\left\{1-\left[1-\left(\dfrac{u}{c}\right)^2\right]^3\right\} & |u|\leqslant c\\ =c^2/6 & |u|>c\end{cases}$$

$$\Psi(u)\begin{cases}=u\left\{1-\left(\dfrac{u}{c}\right)^2\right\} & |u|\leqslant c\\ =0 & |u|>c\end{cases}$$

定数 c は5～9とする[5]。これらの M-推定量は，一般の線型モデルの母数推定にも拡張できる。$\Sigma(Y_i-\hat{Y}_i)^2\to\min$ の代りに，$\Sigma\rho(Y_i-\hat{Y}_i)\to\min$ とすればよい。

上述のような凸関数をウェイトとして用いる方法は n が大きいときには良い

が，nが小さいときには，あまり微妙に変化する関数を用いても有効に作用しない。そこで，正規分布に従わない観測値の分布が，一般的にはスソの長い分布であることに注目して，スソに当る部分の観測値に小さい（あるいは0の）ウェイトを与える方法がある。最も簡単なのは観測値のなかで最小のものと最大なものとをα個づつ取り除いた残りを平均することである。このような推定量は trimmed mean と呼ばれ，標本の大きさnの場合の両スソからα個づつ取り除く trimmed mean を $T_n[\alpha]$ で定義する。

(1. 5) trimmed mean: $$T_n[\alpha]=\frac{1}{n-2\alpha}\sum_{i=\alpha+1}^{n-\alpha} X_{(i)}$$

$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ は $X_1, X_2, \cdots, X_n$ の順序統計量である[6]。trimmed mean に良く似たものにウィンザー化平均がある。これは両スソのα個の観測値を $X_{(\alpha+1)}$ および $X_{(n-\alpha)}$ におきかえて平均をとる方法である。

(1. 6) Winsorized mean:

$$W_n[\alpha]=\frac{1}{n}\left\{(\alpha+1)X_{(\alpha+1)}+\sum_{i=\alpha+2}^{n-\alpha-1}X_{(i)}+(\alpha+1)X_{(n-p)}\right\}$$

これら2つの推定量は観測値とウェイトとの線型結合（linear combination）で表わせることから，L-推定量と呼ばれる。

もう一つのロバストな推定量のクラスにR-推定量と呼ばれるものがある。これはパラメータの分布型に関する想定を全く必要としないタイプの方法で，distribution-free，あるいは nonparametric method と呼ばれており，1950年代から研究されている分野である。順位（rank）を用いて統計量を定義する。良く知られているものに，Hodges=Lehmann（1963）によって考案された統計量，

(1, 7) $$HL=\mathrm{med}[(X_i+X_j)/2 \mid 1\leqslant i\leqslant j\leqslant n]$$

がある。これはn個の観測値を2個づつ対にして，${}_nC_2$個の平均を求め，その中央値をとる方法である。また，Tukey（1977）による trimean は簡単であるが有効な推定量である。Tukey は観測データの 1st quartile と 3rd quartile

を lower hinge と upper hinge と呼び，両ヒンジの外側の観測値を wing と称して

$$(1.\ 8)\quad \text{trimean} = \frac{1}{4}\left[\text{lower hinge} + 2(\text{med}) + \text{upper hinge}\right]$$

を定義している[7]。R-推定量には，この他にも非常に多くのものがあり，実際の応用面での使用度も他のクラスの推定量より多い。詳細は Lehmann(1975) や David(1981) のようなテキストを参照されたい。

c） 推定量の選択

Andrews, et al. (1972) は65種のロバスト推定量をとりあげ，人工数値による実験を行なっている。その結果をまとめると，自由度10以上の t 分布ならばほぼ正規分布であると見なしてよく，$\bar{X}$ を用いればよい。また，これより少しスソの長い自由度5～10程度の t 分布ならば，HL や5％ trimmed mean が良い。自由度が2～5の t 分布に対しては10％～25の trimmed mean がよく，n が大きければ（n>20～30), trimean がよい。これ以上スソが長い分布に対しては中央値が良い。

以上のような数値実験の結果は，分布の型と推定量の優劣とを明らかにはしているが，注意すべきは，一般の実務レベルでのデータ解析においては観測値の分布が未知なことである。正規分布であるかどうかを見きわめることはできるが，正規分布でないことが判明しても，それが自由度5の t 分布にあてはまる，とかロジスティック分布にあてはまる，とかいう結論をひきだすことはできない。したがって，優れたロバスト推定量というのは，ある一つの状況において最適な推定量でなく，現実に起こり得るような様々な状況のもとで，最適に近いような性質を維持する推定方式であると考えなければならない。

このような点を考慮して，Stigler (1977) は現実のデータを用いて推定量の比較を行なった。Stigler のデータは，1761年に James Short が観測した金星の運行に関するデータ，1874年と1882年に行なわれたMichelson-Newcomb の光の速度の測定値，Cavendish が1798年におこなった地球の平均密度の測

定値のような，科学史上著名なものばかりである。Stigler の結論は，trimmed mean が最も優れ，それ以外のものは計算コストを上回るものではない，というものであった。さらに，Spjøtoll＝Aastveit(1980) が農業データを，Rocke, et al.(1982) が化学データを，Smith＝Iqlewicz(1982) が人間の推定能力に関するデータ(8) をとりあげている。これらの分析者たちの結論をまとめると，多くの状況下で trimmed mean がよく，データが標準化されている場合は trimean がよい，ということである。

注(1) '正規分布' という名は F. Galton (1821-1911) による。それまでは '誤差分布' と呼ばれていた。定義式も現在のものとはやや異なる。

(2) K. F. Gauss(1821～1826)に誤差分布および最小二乗法の展開と応用が見られる。

(3) 今日では，心理学的な測定値はまず正規分布でないと考えられている。また，Taylor＝Hudson(1972) は社会科学の領域における観測値は定量的な性質のものと "観念の産物" とが相互関連して生成される場合が多く，やはり正規分布と考えられるようなものは殆んどないことを紹介している。

(4) ロバストな手法が熱心に論ぜられるようになったのは1970年以降である。

(5) 中川・小柳 (1981) 167ページ。

(6) $n=5$ の場合の 20% trimmed mean は $\frac{1}{3}(X_{(2)}, X_{(3)}, X_{(4)})$ となり，オリンピックの体操等の競技の採点方式に一致する。このような採点方式は trimmed mean が統計学で論じられるより以前から行なわれていた。

(7) Tukey(1977) 46p. Tukey はまた，wing をとり除いたデータの平均を midmean と呼んでいる。

(8) 小包の重さ，時間の経過，ボールの直径，壁についた印の高さ，1ページ当りの単語数，角度，etc. を20人に当てさせたデータである。

2. ロバスト化残差

線型回帰モデル $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ におけるロバスト化残差 (robustified residual) $\boldsymbol{e}^*$ は，$\boldsymbol{\beta}$ のロバスト推定量 $\boldsymbol{b}^*$ を用いて次のように定義される。

(2. 1) $\boldsymbol{e}^*=\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}^*$

これは，手順としては，$\boldsymbol{\beta}$ の推定をまず通常の推定方式（例えば最小二乗法）を用いて推定量 $\boldsymbol{b}$ を求め，残差 $\boldsymbol{e}(=\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})$ の分布の中心にかんしてロバス

トな手法を使うことによってβの推定量を再定義することである。注意すべきは，こうして得られたロバスト化残差は最小二乗残差が満たしている制約を満足していないことである。さらに，最小二乗残差を studentization する場合のようなスケーリングの問題も残る。しかし，本論は回帰パラメータβのロバスト推定を目的とするものではなく，ロバスト化残差によって誤差分布の型を検討することに主眼を置いているので，「通常の残差と同じ制約をそれが満たさないという事実は，多分重要でない[1]」と考えておこう。

前節で述べた理由から，以下ではロバスト推定として trimmed mean を取り上げ，それによって得られるロバスト化残差の効用について考察しよう。一般に，残差プロットと呼ばれるものは，visual display によって残差の構造を検討し，線型モデルに与えられている制約条件の妥当性を吟味するための方法である。制約条件には，大別するとモデルの型の specification と誤差分布の型の想定があるが，ロバスト化残差の効用が発揮されるのは後者である。誤差分布に対して正規性の仮定をおいて最小二乗推定を行なった場合，“真の”誤差分布が正規でなければ，一変数の分布を$\bar{X}$で推定するのと同様の問題が起こる。それと同時に，最小二乗残差もまた，“真の”分布のスソの長さにひきずられて，本来正規分布でないものを正規分布らしく見せかけるような行動をとる。

k-trimmed mean → k-trimmed β の拡張を考えてみよう。これは前者における1と0のウェイトを，ベクトル$\mathbf{1}$と$\mathbf{0}$におきかえるだけでよい。すなわち，手順は次のようになる（$k=\alpha/n$）。

(2. 2) ①全ての観測値ベクトル（y_i, $\boldsymbol{x}_i$）にウェイト$\mathbf{1}$を与え，最小二乗残差e_iを求める。

②n個のe_iについて順序統計量$e_{(i)}$を求めて，$i \leqslant nk$および$i \geqslant n-nk+1$に対応する観測値ベクトル（$y_{(i)}$, $\boldsymbol{x}_{(i)}$）にウェイト$\mathbf{0}$を与え，再び最小二乗残差を計算することにより$e(k)$を得る。

$e(k)$は，k-trimmed β を通じて導出される残差系列であり，$k=0$のとき最

小二乗残差に一致する。

注(1) Gnanadesikan(1977) 邦訳 244 ページ。

3. 正規性の検定様式

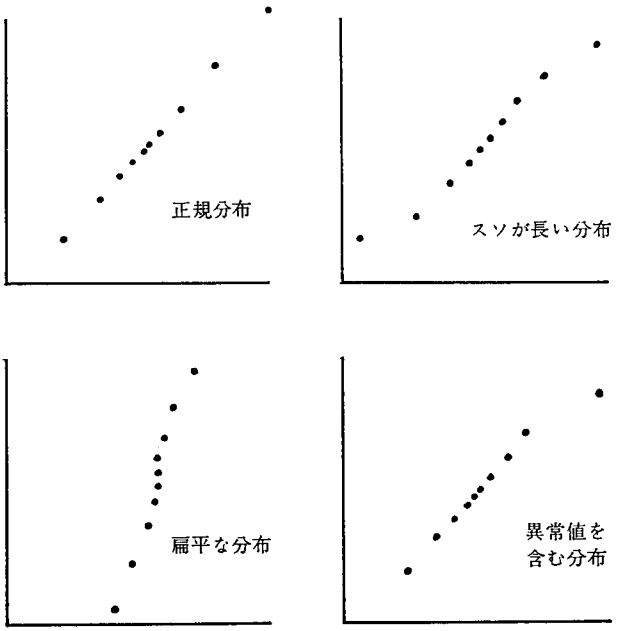
a) *half normal plotting*

確率プロットと呼ばれる一連の手法の中に，半正規プロットと呼ばれるものがある。これは線型モデルにおける残差系列を検討することにより誤差分布型の非正規性を検出するために広く用いられている。その形式は，

(3. 1) $Z_i=F^{-1}(p_i)$

ただし，$p_i=(i-\alpha)/(n-2\alpha+1)$　　$(0\leqq\alpha<1)$

で定義される Z_i をタテ軸にとり（$F(p)$ は正規分布の累積分布関数），ヨコ軸には $x_{(i)}$ をとる[1]。p_i は $[0,1]$ の区間を動き，α は（n があまり大きくないときには）0.5が使われることが多い[2]。

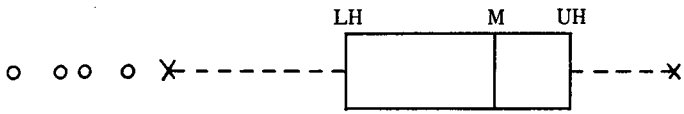


<図　1>

半正規プロットは computer display で行なえば便利であるが，市販の確率紙を用いても簡単にできる。確率紙とはタテ軸に p_i を目盛りとして表わしているような用紙で，特に「正規確率紙」は点 $(X_{(i)}, p_i)$ をプロットすれば，X_i の分布の正規性を検討することができる。正規確率紙上では X_i が $N(0, \sigma^2)$ に従う場合，点 $(X_{(i)}, p_i)$ の plotting は原点を通る傾き σ の直線のまわりに集積する。<図1>は半正規プロットの代表的な形状を表わしたものである[3]。

b） *Box-and-Whisker*

Tukey (1977) が提案している *Box-and-Whisker*[4] は，分布の主要特性を容易に把握せしめるような visual display である。<図2>において，LH=lower hinge, UH=upper hinge, M=median である。LH から UH までの距

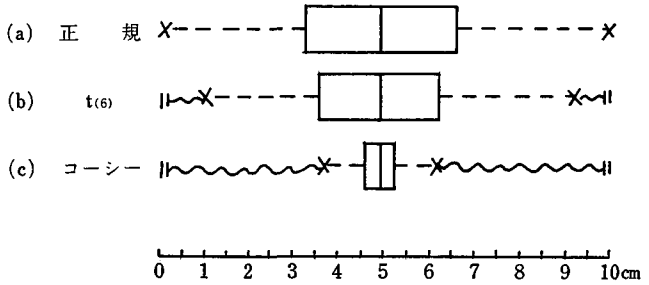


<図 2>

離を MS（=midspread[5]）と呼び，この範囲（*Box* の部分）に全データ点の50%が含まれる。ヒゲ（*Whisker*）の部分は点線で示され，×印は LH-MS および UH+MS を示す。ただし，$X_{(1)}$>LH-MS のときは×は $X_{(1)}$ を，$X_{(n)}$<UH+MS のときは×は $X_{(n)}$ を示す。<図2>の場合，データは中央値より大きい範囲に密で，小さい部分ではかなりばらついており，右に歪んだ分布であることがわかる。×印の外側は wing を示す。

標準正規分布においては M=0, LH=−0.6745, UH=0.6745, MS=1.349 となることを利用すれば，×—× 間の距離を設定することによって *Box-and-Whisker* の正規分布パターンを知ることができる。<図3>は ×—× の作図スケールを10cm として描いたものである (a)。この範囲に全観測値の95%が含まれる。同じく全観測値の95%が ‖—‖ の間におさまるようにして描いたものが (b) と (c) である。(b) は自由度6の t 分布を，(c)はコーシー分布を表わしている。〰〰で示された部分は wing である。コーシー分布の場合，全体の

の50％観測値が中央値のまわりに凝集し，他の50％が残りの広い部分に散在している状態が良くわかる。



<図　3>

c） *confirmatory mode*

半正規プロットや *Box-and-Whisker* のような visual display は，研究者にとって，与えられたデータの分布型に関して未知の状態から，その分布の特徴的な形状を探索（exploration）するのに不可欠である。このようなデータ解析の方式は *exploratory mode* と呼ばれている。これに対して，予めデータの分布型＝正規と仮説をおき，データより得られる検定統計量によって仮説検定を行なうことも考えられる。このような仮説の検定や母数の推定は *confirmatory mode* と呼ばれる。*confirmatory mode* では，いくつかの要約統計量を用いて，データから計算された値が偶然に生ずる確率が求められ，それに基づいて推論が行なわれるわけであるが，ここで問題となるのは要約統計量の選定である。もし算術平均と標準偏差というような要約統計量を<図２>のような分布をもつデータに用いてしまうと，検定や推定の結果はひどく惨めなものになる。このような事態をさけるためには，*exploratory mode* の解析を前もって施しておくべきであり，その結果を考態して要約統計量の選定を行なわなければならない。このことは正規性の検定についても言えることである。正規性の検定はまず *exploratory made* でデータの分布型を検討し，もし正規分布と見なしてよいかどうか微妙な状態であれば *confirmatry made* によればよい。

柴田（1981）は正規性の検定方式として，χ^2 適合度検定，Neyman のスムーズ・テスト，標本の歪度・尖度を用いる検定，Geary 検定，Shapiro-Wilk の検定をとりあげ，統計量の漸近分布やシミュレーションの結果をまとめている[6]。それによると，分布の非対称性については標本歪度，スソの長さに対しては標本尖度ないしは Geary 検定が良い。

歪度（skewness）は分布の非対称性の度合と方向を，尖度（kurtosis）は分布のスソの長さを表わす統計量で，平均値のまわりの r 次のモーメントを μ_r としたとき

$$(3.\ 2)\quad b_1=\frac{\mu_3}{\mu_2^{3/2}},\quad b_2=\frac{\mu_4}{\mu_2^{2}}-3$$

で定義される[7]。b_1 は分布が対称の場合にゼロとなり，左[右]に偏った分布では $b_1>0$ [$b_1<0$] となる。b_2 は正規分布の場合ゼロとなり（$b_2=0$ だからといってその分布が正規分布であるわけではない），尖っている（＝スソが長い）分布の場合 $b_2>0$，扁平な分布の場合 $b_2<0$ となる。

Geary 検定は，スソの長さに敏感な統計量として

$$(3.\ 3)\quad G=\frac{\Sigma|X_i-\bar{X}|}{\sqrt{n\Sigma(X_i-\bar{X})^2}},\quad G^*=\frac{\Sigma|X_i-X\mathrm{med}|}{\sqrt{n\Sigma(X_i-\bar{X})^2}}$$

を用いるものである。スソの長い分布では G は平均的に小さくなる。G^* は $\bar{X}$ を中央値におきかえることにより，ロバスト化したものである[8]。

注(1) 脇本・後藤・松原（1979），119ページ。尚，本論では市販の確率紙に合わせて座標を設定したが，最近の欧米の文献では，$(Z_i,\ X_{(i)})$ のようにとるのが普通である。

(2) α の決め方については，柴田（1981）239-242ページを参照されたい。

(3) 半正規プロットの拡張概念として，Gnanadesikan(1977) は多変量正規分布を想定した‘ガンマプロット’を考案している。邦訳235ページ。

(4) Tukey (1977), 39p. McNel (1977), 6p. McNeil は単に‘Boxplot’と呼んでいる。

(5) MS は一般に四分位偏差と呼ばれているが，ここでは McNeil の用語に従った（Tukey は H-spread と呼んでいる）。UH, LH も同様である。

(6) 柴田（1981），第7章

(7) 定義からすると b_1, b_2 は分布の位置に依存するように見えるが，b_1, b_2 ともに分布の位置・尺度に対して不変である。柴田 (1981) 228-229ページ。このふたつの統計量によって正規性の検定を行うためには，仮説 $H_0: b_1=0$ and $b_2=0$ に対して，有意水準$=\alpha/2$ として，b_1 および b_2 の各々に対して検定すればよい。また，$n \gg 0$ のとき，b_1 および b_2 は漸近的に正規分布に従う。詳細は柴田 (1981) 228-235ページ。

(8) $n \gg 0$ のとき，G は漸近的に $N\left(\sqrt{2/\pi}, \frac{1}{n}\left(1-\frac{2}{\pi}\right)\right)$ に従う。

4. 数値例

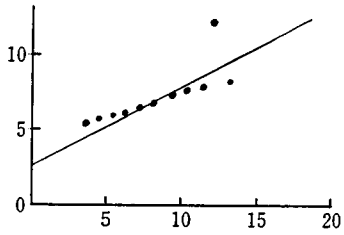
最後に数値例を示そう。Anscombe (1973) は残差分析の重要性を説くために4つの数値例をつくり出した。その中のEX-Cは単純回帰における 'outlier' の混入の影響を示したものである。$\hat{y}=3.0+0.5x$ と最小二乗推定されるような11組のデータ (x_i, y_i) が<表1>に与えられている。このとき<図4>の

<表1>

** ANSCOMBE—EX.C

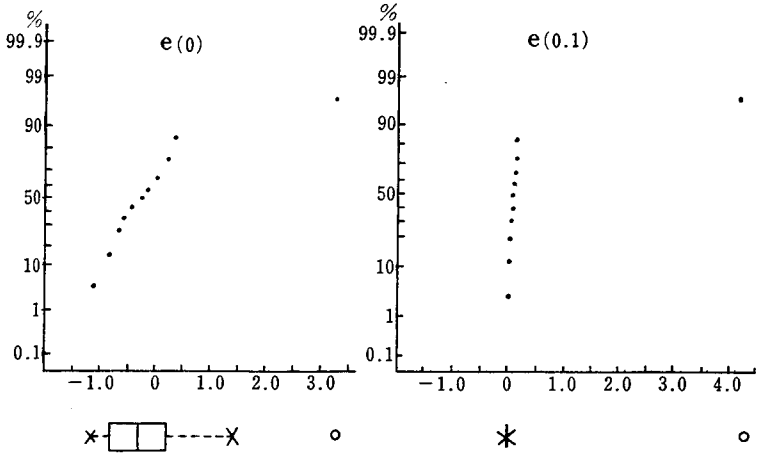
N	X	Y
1	10.00	7.46
2	8.00	6.77
3	13.00	12.74
4	9.00	7.11
5	11.00	7.81
6	14.00	8.84
7	6.00	6.08
8	4.00	5.39
9	12.00	8.15
10	7.00	6.42
11	5.00	5.73

A=3.00245
B=0.49973
COR=0.81629



<図4>

X-Y plot からも outlier の存在は明らかであり，No. 3 の観測値の存在ゆえに $\hat{b}$ が over-estimate されている。この EX-C について $e(0)$ および $e(0.1)$ を visual display したのが<図5>である。$e(0)$ のプロットを見ると No. 3 の観測値を除いた残りのデータはよく正規分布をしているようである。*Box-and-Whisker* からは分布の中心50%は正規とみなせるが，残りの50%が分布の右側に集中しているように読みとれる。しかし，一つとび離れた○印がoutlierの存在を警告するので，注意深い分析者であればそれをとり除いて，残りの10



<図 5>

組のデータについての検討を再び始めるであろう。*confirmatory mode* で残差系列を調べてみると，$b_1=2.04$* (* は正規性の仮説に対して5％棄却域に入ることを指す。以下も同じ)，$b_2=-3.57$*，$G=0.64$* で，左に偏り過ぎると同時にスソが長すぎるここがわかる。そこで e_3 を除く10組について同様の統計量を求めてみると，$b_1=-0.18$，$b_2=-0.01$*，$G=0.86$ となり，ほぼ対称ではあるが，正規分布にしては扁平すぎることがわかる[1]。これに対して，$e(0.1)$ のディスプレイは，一見して以上のようなことを伝達する。この場合，*Box-and-Whisker* は一ヶ所に2つの×印と3本の縦線が重なってしまって良く描けない。これは明らかに，一様分布とただひとつの飛びはなれた値との合成を示すものである[2]。

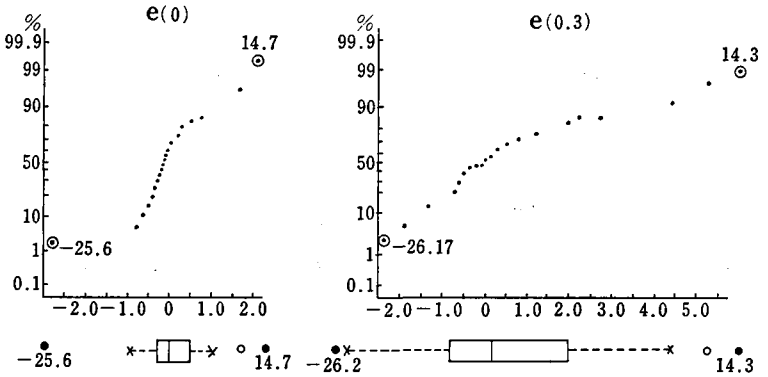
<表2>

I	RND	N	X	Y
1	14.711	1	6.32	20.80
2	0.070	2	0.24	3.12
3	0.066	3	1.48	3.74
4	0.079	4	0.97	3.49
5	−0.226	5	6.72	6.06
6	0.636	6	5.67	6.40
7	1.223	7	7.55	7.93
8	0.390	8	3.48	5.06
9	−0.411	9	2.28	3.62
10	0.312	10	5.64	6.06
11	−0.117	11	4.57	5.10
12	2.482	12	8.93	9.88
13	0.285	13	3.07	4.75
14	5.276	14	3.45	9.93
15	−0.942	15	5.17	4.57
16	−25.798	16	5.83	−19.95
17	4.005	17	3.11	8.49
18	−1.752	18	1.20	1.78
19	−0.594	29	0.22	2.55
20	1.613	20	1.28	5.19

A=3.03394
B=0.49128
COR=0.17588

この Anscombe の例にならって，ロバスト化残差の効用を示すような例を作ってみよう。誤差項としてコーシー分布に従うような乱数（<表2>の RND）を発生させ[3]，X_i,

$i=1, \cdots, 20$ に (0, 10) の一様乱数を与える。誤差分布の平均を $\overline{\mathrm{RND}}$ とし，$Y_i = (3-\overline{\mathrm{RND}}) + 0.5X_i + \mathrm{RND}_i$ として生成したのが<表2>の X, Y である。このデータに対して最小二乗推定を施してみると，$\hat{Y}=3.03+0.49X$ を得る。ところが，X, Y の相関係数が0.18と極端に低いことからもわかるように，回帰線の当てはまりは非常に悪い。そこで $e(0)$ をディスプレイしてみると，<図6>のようになる。これを見ると，2,3の観測値によって残りがひっぱられ，本来正規分布であったものがスソの長い分布に変形されてしまっているかのような印象を受ける。そこで，これらの回帰に対して害を及ぼすような観測値を排除してしまうことにしよう。No. 1 と No. 16 の観測値を除けば最小二乗推定 $\hat{Y}_i=3.09+0.65\,X_i$ を得，相関関数は 0.71 となる。さらに，No. 14 の観測値も除けば，$\hat{Y}=2.80+0.66X_i$ を得，X, Y の相関係数も 0.81 となり，あてはめの状態は大分良くなる。そこで，これら3つの観測値を 'outlier' と見なし，誤差項分布の正規性の仮定を妥当と認定し最終的な回帰式 $\hat{Y} = 2.80 + 0.66X_i$ を結論としてしまう可能性がある。



<図 6>

このような“錯誤”は，X と Y との相関係数が低い場合に起きやすい。今，誤差分布が F で，X と Y との間の相関係数はゼロであると仮定しよう。その場合，残差は X と Y との線型関係によってもたらされるから，中心極限定理

により，誤差分布 F の分散が有限であるかぎり，残差の分布は正規分布に近づく。このことを明らかにするために，次のような数値実験を行なってみよう。X_i に一様乱数を与え，$Y_i=kr_i+(1-k)r_j$, $(i\neq j)$ とする。このようにして n 組の X, Y を生成する。$k=0$ の場合，X, Y は独立であるから，最小二乗残差 $e_i=Y_i-a-bX_i$ の分布は中心極限定理により，尖ったものとなるはずである。<表3>は $n=(20, 35, 50)$, $k=(0, 0.3, 0.8)$ の9種の設定のもとに，各々100回づつ最小二乗残差の分布の b_1 と b_2 を求め，さらにそれを100でわったものである。$k=0$ のときには b_2 がかなり大きく，相当にスソの長い分布をしていることがわかる。逆に $k=0.8$ の場合は n が大きくなるにつれてスソが短かくなり，中心極限定理は働いていない。しかし $k=0.3$ 程度では，（本来，中心極限定理は働かないはずであるが）やはり残差の分布はスソが長い。したがって，<表2>のような低い相関をもつデータについては，最小二乗残差 $e(0)$ の visual display は，本来残差に含まれている非正規性の証拠を洗い流してしまう可能性を孕んでいるのである。このような場合，ロバスト化残差のディスプレイは，そのような‘落とし穴’への危険をある程度回避する能力を持っている。<図6>の $e(0.3)$ のディスプレイは両端にある観測値が他と連繋していることを知らしめるのに十分な役割を発揮している。

<表3>

n \ k	0	0.3	0.8
20	$b_1=$ 2.41 $b_2=$ 6.51	$b_1=$ 1.52 $b_2=$ 2.64	$b_1=$ 0.50 $b_2=-0.73$
35	$b_1=$ 2.41 $b_2=$ 6.78	$b_1=$ 1.66 $b_2=$ 2.69	$b_1=$ 0.46 $b_2=-0.75$
50	$b_1=$ 2.46 $b_2=$ 7.05	$b_1=$ 1.70 $b_2=$ 2.80	$b_1=$ 0.49 $b_2=-0.81$

注(1) Gについては，スソの長すぎる点に関してのみ有意点の数表が作成されているので，スソが短すぎることの検定には役立たない。

(2) 一様分布の場合，$E(b_1)=0$, $E(b_2)=-1.2$ であるから，confirmatory mode でも e_3 を除く残りのデータは一様分布をすると結論できる。この一様分布の正体は，Anscombe が $Y=4.0+0.346X$ に従うようにデータを作り出した際の四捨五入による誤差であろう。$e(0.1)$ は No. 3 を除けば，全て $[-0.004, 0.005]$ の範囲にある。

(3) RND 系列は，$b_1=-2.16$, $b_2=8.17$ である。

文 献

[1] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972) *Robust Estimates of Location—Survey and Advances.* Princeton Univ. Press.

[2] Andrews, D. F. (1979) *The Robustness of Residual Displays* (a contribution to ROBUSTNESS IN STATISTICS, edited by Launer, R. L. & Wilkinson, G. N. Academic Press)

[3] Anscombe, F. J. (1973) *Graphs in Statistical analysis.* The American Statistician 27-1.

[4] Atkinson, A. C. (1982) *Regression Diagnostics, Transformations and Constructed Variables* (with Discussion). JRSS (B) 44-1

[5] David, H. A. (1981) *Order Statistics,* 2nd ed. Wiley

[6] Gauss, K. F. (1821-26) *Theoria Combinations Observationum: Erroribus Minimis Obnoxiae* (飛田武幸・石川耕春訳「誤差論」1981, 紀伊国屋書店)

[7] Gnanadesikan, R. (1977) *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley (丘本正・磯貝恭史訳「統計的多変量データ解析」1979, 日科技連出版社)

[8] Hodges, J. L. & Lehmann, E. L. (1963) *Estimates of Location based on rank tests.* Annals of Mathematical Statistics, 34.

[9] Huber, P. J. (1964) *Robust estimation of a location paramater.* Annals of Mathematical Statistics. 35.

[10] Huber, P. J. (1981) *Robust Statistics.* Wiley

[11] Lehmann, E. L. (1975) *Nonparametrics : Statistical Methods Based on Ranks.* Holden Day (鍋谷清治・刈屋武昭・三浦良造訳「ノンパラメトリックス——順位にもとづく統計的方法」1978, 森北出版)

[12] McNeil, D. R. (1977) *Interactive Data Analysis, A practical primer.* Wiley

[13] 中川徹・小柳義夫 (1982)「最小二乗法による実験データ解析」東大出版会

[14] Rocke, D. M., Downs, G. W. & Rocke, A. J. (1982) *Are Robust estimators Really Necessary?* TECHNOMETRICS 24-2

[15] 柴田義貞 (1981)「正規分布 ～ 特性と応用」東大出版会

[16] Smith, T. M., & Iglewicz, B. (1982) *An Effective Classroom Technique for Comparison of Robust Estimators.* The American Statistician 36-3 (Part 1)

[17] Spjøtvoll, E. & Aastveit, A. H. (1980) *Comparison of Robust Estimators Based on Data From Field Experiments.* Scandinavian Journal of Statistics 7.

[18] Stigler, S. M. (1977) *Do Robust Estimators Work With Real Data?* Annals of Statistics 6.

[19] Taylor, C. L., & Hudson, M. C. (1972) *World Handbook of Political and Social Indicators.* Yale Univ. Press

[20] Tukey, J. W(1977) *Exploratory Data Analysis* Addison-Wesley

[21] 脇本和昌・後藤昌司・松原義弘 (1979)「多変量グラフ解析法」朝倉書店

1982. 9. 29 脱稿

(後期課程第3年度生・統計学　保田順三郎研究指導)