誤差分布の非正規性の処理

線形回帰モデルにおける誤差項が，ふつうに仮定されているような正規分布に従わないということは，充分考えられることである．このことについてはすでにこの号の前稿にも述べたし，また小柳氏の論文にも扱われているが，ここではそれについてもう少し立ち入って考えよう．問題を a）非正規性の影響，b）非正規性の原因とパターン，c）正規性の検定，d）ロバスト推定，の四つのテーマにわけて考える．

1. 問題の意味

線形モデルにおいて，誤差項 u_i が互いに独立に平均0，分散 σ^2 の(必ずしも正規分布とは限らない)分布に従うとき，最小2乗推定量 $\hat{\beta}_j$ が母数 β_j の不偏推定量になり，かつその分散が $m^{jj}\sigma^2$ (m^{jj} は説明変数のモーメント行列の逆行列の要素) に等しくなることは，一般的に成り立つ．そうして $\hat{\beta}_j$ の分布は，標本数 n がある程度大きく，また説明変数の値の中で，特定の標本に対応するものだけがとくに大きくなるようなことがないならば，ほぼ正規分布に近い分布に従う．また σ^2 の推定量 $\hat{\sigma}^2$ も不偏推定量であるから，統計量，

$$(\hat{\beta}_j-\beta_j)/\sqrt{m^{jj}}\,\hat{\sigma}$$

の分布は誤差分布が正規分布である場合とほぼ等しくなる．したがって，たとえば β_j に関する信頼区間，

$$\hat{\beta}_j-t_\alpha\sqrt{m^{jj}}\,\hat{\sigma}<\beta_j<\hat{\beta}_j+t_\alpha\sqrt{m^{jj}}\,\hat{\sigma}$$

(t_α は t 分布の両側 100α パーセント点)

が真の値を含む確率はほぼ $1-\alpha$ になる．

したがって正規分布を前提にした推測の方法は，誤差分布が正規分布でない場合にも，少なくとも近似的には妥当な結論を導くといえる．このことを最小2乗法にもとづく推測の方法は validity robustness をもつといいあらわすこともある．

しかしながら，分布が正規分布からいちじるしくかけ離れている場合には，最小2乗推定量の効率はいちじるしく落ちる可能性がある．すなわちそれ以外に最小2乗推定量よりいちじるしく分散の小さい推定量が存在するかもしれない．もし分布の形 $f(u)$ が既知ならば，β_j の推定のために，つぎのような方法を利用することができる．

a）最尤法　$\prod_{i=1}^{n} f(y_i-\beta_0-\beta_1x_{1i}-\cdots-\beta_px_{pi})$

を最大にする $\beta_0,\beta_1,\cdots,\beta_p$ を推定量とする．

b）最良不変推定量(Pitman 推定量)

$$\hat{\beta}_j=\int\cdots\int\beta_j\prod_{i=1}^{n}f(y_i-\sum_{j=1}^{p}\beta_jx_{ji})\prod_{j=1}^{p}d\beta_j \Big/ \int\cdots\int\prod_{i=1}^{n}f(y_i-\sum_{j=1}^{p}\beta_jx_{ji})\prod_{j=1}^{p}d\beta_j$$

n がある程度大きいとき，上記の二つの推定量はいずれも，ほぼ分散 m^{jj}/I の正規分布に従う．ここに I は Fisher 情報量

$$I=\int\frac{|f'(u)|^2}{f(u)}du$$

である．したがって最小2乗推定量の相対効率は n が大きいとき，ほぼ $1/\sigma^2I$ となる．たとえば誤差分布が自由度 ν の t 分布に従うならば，

$$\sigma^2=\nu/(\nu-2)\quad I=(\nu+1)/(\nu+3)$$

であるから，相対効率は，

$$(\nu-2)(\nu+3)/\nu(\nu+1)$$

となる．それゆえ $\nu=5$，10，20のとき，最小2乗推定量の相対効率は，それぞれ80%，94.5%，

98.5%となる．したがって5%程度の効率のロスは容認しうるものとすれば，t 分布で自由度が10以上くらいならば最小2乗推定量を用いることによる情報損失はあまり大きくないといってもよい．逆に誤差分布が自由度5以下の t 分布くらい正規分布から隔っているならば，最小2乗法を用いることによる情報損失が大きくなる可能性がある．

同じことは，区間推定，あるいは仮説検定についてもいうことができる．すなわち誤差分布が正規分布から大きくはなれているならば，t 分布にもとづく区間推定の幅が広すぎたり，t 検定の検出力が低くなったりする可能性がある．その場合の情報の損失は最小2乗推定量の情報損失とほぼ同等と考えてよい．

2. 非正規性の方向

ところで一般に誤差分布が正規分布に従うことを厳密に証明することは不可能であるが，逆に分布の形についてまったく知識がないという場合も稀である．そこで分布がほぼどんな形になるであろうかを，それぞれの場に即して考えてみることが必要である．それには三つの場合がある．a）正規分布でない特定の分布形を想定すべき積極的な理由が存在する場合，b）誤差分布は「理想的」な場合には正規分布と考えられるが現実にはそれからある程度ずれると思われる場合，c）本来の「誤差」以外に大きな攪乱，ないしノイズが（小さい確率で）起こりうる場合．

a）については，離散データの場合，正値データの場合等がある．離散データが正規分布に従わないことは自明であるが，これについてはふつうの回帰分析とは別個に扱うべきである．というのは非正規性の問題よりも，離散分布の母数を，説明変数のどのような関数としてあらわすかということのほうがより重要だからである．たとえば2項分布に従うデータについては，2項確率 p を説明変数の一次関数としてあらわすことは適当でない場合が多い．

この場合ロジットモデル，

$$\log\,[p/(1-p)]=\beta_0+\beta_1x_1+\cdots+\beta_px_p$$

あるいはプロビットモデル，

$$p=\Phi(\beta_0+\beta_1x_1+\cdots+\beta_px_p)$$

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

のほうが適切である．

正値データについては，変動係数がいちじるしく小さい場合以外には，分布の非対称性が明白になるので，一般には変数変換が必要になる．ある種のデータについては対数正規分布に従うと想定できる場合が少なくない．また寿命データについてはワイブル分布を想定できるのがふつうである．これらの問題についてはここではこれ以上ふれない．一般に非正規性が最初から明白である場合には，最小2乗法を適用する際充分注意しなければならない．

b）については二つのことが考えられる．一つは変数変換が不充分であるために，非正規性が完全には除かれない場合，とくに分布の非対称性が残る場合である．もう一つは完全に管理された実験あるいは観測の場合には誤差分布が正規分布に従うと考えられるが，現実には管理が完全ではないためにある程度正規分布からのズレが生じ得る場合である．このときには分布は正規分布よりある程度スソの長い分布になると考えられる．この場合にたとえばCauchy分布ほどスソの長い分布が生ずることはほとんどないといってよい．

c）については，観測機器の故障，操作のあやまり，数字の読み違い，転記・パンチのミス等が考えられる．これらのいわば「まちがいデータ」は現実には意外に多く混入しがちなものである．これをモデル化すれば，観測値誤差分布の密度関数はつぎのようになる．

$$(1-\varepsilon)f(x)+\varepsilon g(x)$$

ここで ε は小さい確率，f は本来の誤差分布の密度関数，g は「まちがい」の分布である．これは

小柳氏も述べているようにHuberの考えたモデルである．

この問題については，このようなモデルをただ想定すればよいというものではなく，むしろそのような「まちがい」を検出することが大切である．そうしていわゆるoutlierを検出するだけでなく，それがどのような原因によって生じたかを可能な限り具体的に追求しなければならない．

3. 非正規性の検出

分布の正規性の検定については，同一分布に従う観測値に関しては，これまで数多くの方式が提案されている．それらの考え方の多くは回帰分析の残差項にも適用できるけれども，一般に残差項は独立同一分布に従うわけではないから，検定統計量の仮説のもとでの分布を正確に求めることはむずかしい．

そこでとにかく一般に残差をプロットして眺めてみることがまず大切である．それによって残差に残っている系統的な偏り，自己相関，異常値，および非正規性などの問題を同時に吟味することができる．ここで一つのプロットによってこれらの多くの偏りを同時に検出することは不可能であると思われるかもしれない．またどのようにしても，たとえば誤差分散の不均一性，異常値の存在，あるいは誤差分布の非正規性などを，明確に区別することは困難である．しかしながら現実にそのような区別をつけることは必ずしも必要ではない．モデルからこれらの偏りは，いずれも最小2乗法にもとづく推測方式の妥当性や効率を損うという点で問題にされるのであって，その間の区別をつけること自体はあまり意味がないことが多い．またこれらの場合を区別することは観念的には考えられても，対象の構造を具体的に理解するうえでは差があまりないという場合もある．

もっとも重要なのは異常値の検出である．そのためには，まず残差 $e_i=y_i-\sum_j \hat{\beta}_j x_{ij}$ $(i=1, 2, \cdots, n)$ を基準化しなければならない．すなわち仮説のもとでは，

$$E(e_i^2)=V(y)-V(\sum \hat{\beta}_j x_{ij})$$
$$=(1-\sum_j\sum_k m^{jk}x_{ij}x_{ik})\sigma^2$$

となるから，これを $c_i\sigma^2$ とあらわせば，e_i を $\sqrt{c_i}$ で割ることによって基準化した残差がえられる．そうしてさらにこの仮説のもとでは，

$$t_i=\frac{\sqrt{n-p-1}\, e_i/\sqrt{c_i}}{\sqrt{\sum e_i^2-e_i^2/c_i}}$$

が自由度 $n-p-1$ の t 分布に従うから，このことを用いて検定を行なうことができる．また全体としては，

$$\max_i |t_i| > t_{\alpha/n}(n-p-1)$$

ただし $t_{\alpha/n}(\nu)$ は自由度 ν の t 分布の両側 $100\alpha/n$ パーセント点，とすれば，全体として異常値の存在を検定することができる．

このような検定方式は分布の非正規性の検出にも，スソの長い分布に対してはかなり高い検出力をもつことが知られている．

そうして「異常値」が検出されたらそれを除いて推定量と残差を再計算し，さらに異常値が残されていないかどうか検定して，異常値がもはや検出するまでつづけるという方式が考えられる．

4. ロバストな推定法

しかしながら最初に異常値がいくつも含まれているとき，最小2乗法を適用すると，推定された式が異常値に影響されて，残差自体が偏ったものになってしまうことがある．同じことは誤差の分布がいちじるしく正規分布からかけ離れている場合にも生ずる．このような場合には係数の推定にロバストな方法を用いなければならない．それにもいろいろな方法が考えられる．小柳氏の論文にも紹介されているように，ρ を適当な関数として，

$$\sum_i \rho(y_i-\sum_j \beta_j x_{ji}) \to \min$$

となるように β_j を決めるのがHuberの方法である．ρ の定め方にもいろいろな考え方があるが，

ここでは $\rho(u)=|u|$ すなわち，

$$\sum_i |y_i-\sum \beta_j x_{ji}| \to \min$$

とする方法を考えよう．この方法は母平均の推定に標本中央値を使うことに対応するから，誤差分布がスソの長い分布であるとき有効である．そうして正規分布のときには，標本数が大きいとき，その効率(推定量の分散の逆数)は最小2乗推定量の $2/\pi=63\%$ となる．

このような方法は最小絶対偏差法とよばれることがある．この方法による推定値を計算するには線形計画法によってつぎの問題を解けばよい．

$$\sum_j \hat{\beta}_j^* x_{ji}+u_i-v_i=y_i \quad i=1,2,\cdots,n$$

$$u_i \geqq 0,\ v_i \geqq 0$$

の条件のもとで $z=-\sum u_i-\sum v_i$ 最大

この問題は変数を $2n+p$ 個含むから，n が大きくなると計算が面倒になるように思われるかもしれない．しかし最小2乗推定量 $\hat{\beta}_j$ をまず求めておいて $\hat{\beta}_j^*=\hat{\beta}_j+\beta_j^+-\beta_j^-$ とおき，

$$\sum \beta_j^+ x_{ji}-\sum \beta_j^- x_{ji}+u_i-v_i=e_i$$

$$\beta_j^+ \geqq 0,\ \beta_j^- \geqq 0,\ u_i \geqq 0,\ v_i \geqq 0$$

のもとで $z=-\sum u_i-\sum v_i$ 最大

という問題を，最初 e_i の符号に応じて u_i または v_i を基底変数としてシンプレックス法を用いて解けば，一般にそれほど多くの計算量を必要としない．

つぎに残差の新しい推定値を，

$$e_i'=y_i-\sum \hat{\beta}_{ij}^* x_j=u_i-v_i$$

とすれば，この値は異常値の存在によって影響を受けることが少ないであろう．そこでこれらの値の中で絶対値の大きいものを標本から除いて，残りの値について改めて最小2乗法によって母数を推定すればよい．

この際 e_i' が大きすぎるか否かをどのような基準によって定めるかという問題が残る．このような方法を適用する場合，σ の正確な推定量を求めることはむずかしいから，厳密な「異常値の検定」はできない．そこで二つの考え方がある．一つは e_i' の絶対値の大きいものから機械的に何個かを，(たとえば標本の10%ずつ)を除くことであり，第2の方法は σ の推定値をなんらかの形で求めて，その一定倍，(たとえば 2σ) 以上のものを除くことである．σ の一つの推定値としては e_i' の値の4分位偏差(大きさの順に $(n+1)/4$ 番目の値と $3(n+1)/4$ 番の値との差) を $3/4 \fallingdotseq 1/1.36$ 倍すればよい．

このような方法はとくに標本の数がある程度大きく，かつデータの中に大きな誤差を含むものが混じっている可能性が高い場合には有効である．ただそのくわしい性質についてはまだよくわかっていないところも多い．

5. 数 値 例

つぎのような数値例を考えよう．

被説明変数 y の7つのデータがつぎのように与えられたとしよう．

78　79　104　114　112　171　154

これに直線回帰モデル

$$y=\alpha+\beta x+u$$

をあてはめる．ただし $x=1, 2, \cdots, 7$ とする．

まず最小2乗法を適用すると，

$$\hat{\beta}=\frac{\sum(x-\bar{x})y}{\sum(x-\bar{x})^2}=\frac{420}{28}=15$$

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}=116-15\times 4=46$$

これから残差 e を計算するとつぎのようになる．

7　−7　3　−2　−19　25　−7

残差分散の推定量は，

$$\hat{\sigma}^2=\sum e^2/5=1146/5=229.2$$

$$\hat{\sigma}=15.14$$

$E(e^2)=c\sigma^2$ とおくと，c の値はつぎのようになる．

15/28　20/28　23/28　24/28　23/28　20/28　15/28

したがって「基準化された残差」$e/\sqrt{c}$ はつぎのような値になる．

0.63　−0.55　0.22　−0.14　−1.38　1.95　−0.63

これを t 値に変換するとつぎのようになる．

0.59 −0.51 0.20 −0.13 −1.57 3.56 −0.59

この中で絶対値の最大のもの3.56は自由度4の t 分布の両側2％点3.747より小さいから，もちろん10/7%点より小さい．したがって水準10%でも有意とならない．

しかしとにかく25，−19という二つの値はある程度他のものより大きいから，ここで絶対偏差法を適用してみる．

$$\alpha^+-\alpha^-+\beta^+x_i-\beta^-x_i+u_i-v_i=e_i \quad i=1,\cdots,7$$

$$\alpha^+,\ \alpha^-,\ \beta^+,\ \beta^-,\ u_i,\ v_i\geqq 0$$

のもとで $z=-\sum u_i-\sum v_i$ 最大とする

e_1 の符号から最初に基底に入る変数は $u_1\ v_2\ u_3\ v_4\ v_5\ u_6\ v_7$ となり，最初のシンプレックス表は**表1**のようになる（ただし0の入るところは省略してある）．

これからふつうのシンプレックス法で計算をつづけると3回の基底の入れかえで解がえられる（**表2，3，4**）．

これから $\hat{\alpha}^*=\hat{\alpha}+9\frac{1}{3}=55.333$

$\hat{\beta}^*=\hat{\beta}-2\frac{1}{3}=12.667$

をえる．また e' の値は，

$0,\ -11\frac{2}{3},\ \frac{2}{3},\ -2,\ -16\frac{2}{3},\ 29\frac{2}{3},\ 0$

となる．これから σ を推定すると，

$\hat{\sigma}'=(\frac{2}{3}-(-11\frac{2}{3}))\times\frac{3}{4}=9.25$

となり，この値に比べると $29\frac{2}{3}$ は大きすぎるといえる．そこでこの値を除いて推定すると，

$\hat{\alpha}'=40 \qquad \hat{\beta}'=12.5$

となる．

この場合，$n=7$ はあまり大きくないから，漸近理論は適用できないと考えられる．したがって精密な確率的論理によって結論を出すことはできないが，異常値が一つふくまれるという判定は十分合理的であるように思われる．

表 1

基底	e	α^+	α^-	β^+	β^-	v_1	u_2	v_3	u_4	u_5	v_6	u_7
u_1	7	1	−1	1	−1	−1						
v_2	7	−1	1	−2	2		−1					
u_3	3	1	−1	3	−3			−1				
v_4	2	−1	1	−4	④				−1			
v_5	19	−1	1	−5	5					−1		
u_6	25	1	−1	6	−6						−1	
v_7	7	−1	1	−7	7							−1
z	−70	1	−1	8	−8	2	2	2	2	2	2	2

表 2

基底	e	α^+	α^-	β^+	v_4	v_1	u_2	v_3	u_4	u_5	v_6	u_7
u_1	7.5	¾	−¾		¼	−1			−¼			
v_2	6	−½	½		−½		−1		½			
u_3	4	¼	−¼		¾			−1	−¾			
β^-	0.5	−¼	¼	−1	¼				−¼			
v_5	16.5	¼	−¼		−5/4				5/4	−1		
u_6	28	−½	½		3/2				−3/2		−1	
u_7	3.5	(¾)	−¾		−7/4				7/4			−1
z	−66	−1	1	0	2	2	2	2	0	2	2	2

表 3

基底	e	α^-	β^+	v_7	v_4	v_1	u_2	v_3	u_4	v_6	u_7
u_1	4			−1	②	−1			−2		1
v_2	8⅓			⅔	−1⅔		−1		1⅔		−⅔
u_3	3⅓			−⅓	1⅓			−1	−1⅓		⅓
β^-	1⅔		−1	⅓	−⅓				⅓		−⅓
v_5	15⅓			−⅓	−⅔				⅔	−1	⅓
u_6	30⅓			⅔	⅓				−⅓		−⅔
α^+	4⅔	−1		1⅓	−2⅓				2⅓		−1⅓
z	−61⅓	0	0	1⅓	−⅓	2	2	2	2⅓	2	⅔

表 4

基底	e	α^-	β^+	v_7	u_1	v_1	u_2	v_3	u_4	v_6	u_7
v_4	2			−½	½	−½			−1		½
v_2	11⅔			−⅙	⅚	−⅚	−1				⅙
u_3	⅔			⅓	−⅔	⅔		−1			−⅓
β^-	2⅓		−1	⅙	⅙	−⅙					−⅙
v_5	16⅔			−⅔	⅓	−⅓					⅔
u_6	29⅔			⅚	−⅙	⅙				−1	−⅚
α^+	9⅓	−1		⅙	1⅙	−1⅙					−⅙
z	−60⅔	0	0	1⅙	⅙	1⅚	0	0	0	0	⅚

5. むすび

位置母数の推定，すなわち，

$$X_i=\theta+u_i \qquad i=1,2,\cdots,n$$

とあらわされるときの θ の推定に関して，ロバストな推定量を求める問題は，この10年ほどの間にすでに述べた Huber をふくめて多くの人々によって精力的に研究された．それによって多くの数値的結果も得られている．推定量が直接回帰分析の場合に拡張できる限り（たとえば，最小2乗推定量は算術平均の，最小絶対偏差法は中央値の直接の拡張になっている），推定量の漸近効率に関する数値的結果は，一般に回帰分析の場合にも直接あてはめることができる．

それらの結果から知られることをやや乱暴にまとめれば，非正規性による推定量の効率損失の問題は，「異常値」の検出に充分注意を払う限りそれほど重大ではないということもできる．むやみに複雑精巧な推定手法などを適用することは，あまり有効ではない．それよりも不適切な説明変数や，回帰式の形のために生ずる「モデルの偏り」のほうが誤まった結論を導き出す危険が大きいのである．

ただし説明変数選択の基準に対する非正規性の影響については，まだ充分調べられていない．とくに「異常値」が新しい説明変数の導入によって解消するという可能性もあるから，この問題にはやや微妙な点がある．しかしながら Mallows の C_p 統計量や Allen の PSS についての議論は，誤差の正規性の仮定とは一応独立になり立つことに注意しておこう．非正規性の問題は，まったく無関心であってよいことではないが，それだけを取り出して過度に注意を向けるのもよくないようなものであるということができよう．