論　文

線形回帰分析における尤度距離による影響力評価

竹　内　秀　一

Assessment of Influence based on Likelihood Distance in Linear Regression

Hidekazu TAKEUCHI

Some influence measures based on the case deletion procedure have been proposed in linear regression analysis. Each influence measure assesses the influence of observations from the statistical viewpoint. Likelihood distance is derived to assess the influence based on the log likelihood to estimate unknown parameters such as regression coefficients and a variance of the error distribution in usual linear regression. Although this influence measure has the advantage of the derivation through a mathematical approach, it has some disadvantages of the usage in data analysis. A major disadvantage is that the existing cut-off point based on the log likelihood is independent of the sample size of data since it is derived through an asymptotic confidence region for the unknown parameters. In this paper a new cut-off point is proposed to be dependent on the sample size. Furthermore a theorem shows the condition that the new cut-off point is superior to the existing one.

1　はじめに

線形回帰分析における観測値の影響力評価を考える場合に，その評価規準としてどのような尺度を利用するかにより結果が異なる。これまでは，観測値除去法に基づいて観測値のもつ影響力を評価するための診断統計量（influence measure）として，ノルム化診断統計量（竹内［6］を参照）である Cook の距離（Cook's distance）や行列式型診断統計量（竹内［4］を参照）である一般化分散比（covariance ratio）などを取り上げてきた。本研究では，これ

らとは異質な評価規準として，回帰係数の対数尤度規準に基づく診断統計量である尤度距離（likelihood distance）について検討をする。

尤度距離は，Cook and Weisberg［3］によって導入されたが，データ解析において影響力評価を行う場合に，いくつかの問題点があるためにあまり利用されていないというのが実状である。この尤度距離は確率分布に関わる数少ない診断統計量（その他には，竹内・近河・篠崎［7］などを参照）という特徴がある。けれども，確率的に取り扱いやすい診断統計量であるという利点の反面で，あまり実用上利用されていないという経緯がある。この理由は，安易に確率分布を適用したことにより，影響力の大きさを評価するときに目安となる打切り点（cut-off point または calibration point）が，数学的な自然の流れにより漸近的な形で導出されたため，データ数に依存しないという大きな欠点を抱えてしまったのである。つまり，実際のデータ解析において，データ数に依存しない打切り点では，ほとんど役に立たないのである。そこで本論文では，観測値除去法に基づいて個々の観測値の影響力を測る診断統計量として尤度距離を取り上げ，影響力評価において従来から提案されている確率分布に基づく打切り点を再検討し，新たにデータ数に基づいて調整された（size-adjusted）打切り点を導出する。

本論文の構成は以下のとおりである。2 節では各種の定義を与える。3 節において，尤度距離の新たな打切り点を提案する。4 節では，従来の打切り点と提案する新たな打切り点との比較をし，新たな打切り点が優位になる条件を 1 つの定理として導く。5 節は全体のまとめである。

2　定義

2.1　回帰モデル

ここでは，線形回帰モデルとして，

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

を考える。このとき，$\boldsymbol{y}$ は $n\times 1$ の目的変数ベクトル，$\mathbf{X}$ は $n\times p$ のフルランクの説明変数行列，$\boldsymbol{\beta}$ は $p\times 1$ の回帰係数ベクトル，そして ε は $n\times 1$ の誤差ベクトルであり，正規分布 $\boldsymbol{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ に従うものとする。ただし，$\mathbf{I}_n$ は n 次の単位行列を表す。また，$\boldsymbol{\beta}$ の最小 2 乗推定量は $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ として得られ，σ^2 の不偏推定量は $\hat{\sigma}^2 = \boldsymbol{e}'\boldsymbol{e}/(n-p)$ となる。ただし，「′」は行列あるいはベクトルの転置を表し，$\boldsymbol{e}$ は残差ベクトルであり，$\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{y}$ である。このとき，$\mathbf{H}$ は説明変数行列から構成されるハット行列（hat matrix）$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ であり，その第 i 対角成分 h_{ii} がてこ比である（てこ比の性質については竹内［5］を参照）。ただし，$1/n \le h_{ii} < 1$ とする。さらに，残差ベクトルの第 i 成分 e_i を標準化した $t_i = e_i/(\hat{\sigma}\sqrt{1-h_{ii}})$ を標準化残差（内的スチューデント化残差）と呼び，t_i の定義式において，

$\hat{\sigma}$ の代わりに $\hat{\sigma}_{(i)}$ を用いた $t_i^*=e_i/\{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}\}$ をスチューデント化残差（外的スチューデント化残差）と呼ぶ。ここで，添字の（・）は n 個の観測値の中から除去される観測値の番号を表す。

ところで，$\hat{\sigma}^2$ および $\hat{\sigma}^2_{(i)}$ の関係式は，

$$\hat{\sigma}^2_{(i)}=\frac{n-p-t_i^2}{n-p-1}\,\hat{\sigma}^2$$

であり，また，t_i および t_i^* の関係式は，

$$t_i=t_i^*\sqrt{\frac{n-p}{n-p-1+t_i^{*2}}} \tag{2.1}$$

である。

2.2　尤度距離

Cook and Weisberg［3］は，尤度距離 LD_i をつぎのように定義した。

$$\mathrm{LD}_i=2[L(\hat{\boldsymbol{\beta}})-L(\hat{\boldsymbol{\beta}}_{(i)})] \tag{2.2}$$

ただし，

$$L(\hat{\boldsymbol{\beta}}):\boldsymbol{\beta}\text{ が }\hat{\boldsymbol{\beta}}\text{ のときの対数尤度，}$$

および

$$L(\hat{\boldsymbol{\beta}}_{(i)}):\boldsymbol{\beta}\text{ が }\hat{\boldsymbol{\beta}}_{(i)}\text{ のときの対数尤度}$$

である。

尤度距離 LD_i とは，未知回帰係数ベクトル $\boldsymbol{\beta}$ が $\hat{\boldsymbol{\beta}}$ のとき，つまり，すべての観測値を用いたときの対数尤度と，同様に $\boldsymbol{\beta}$ が $\hat{\boldsymbol{\beta}}_{(i)}$ のとき，つまり，第 i 番目の観測値を除いたときの対数尤度の差を2倍した診断統計量である。この差の大きさから，その除去された観測値の影響力を測定する。したがって，差が大きい，つまり LD_i の値が大きい第 i 番目の観測値を影響力が大きいと判定するのである。

付録1で示されるように，(2.2) 式をてこ比 h_{ii} とスチューデント化残差 t_i^* を使って表すと，以下のようになる。

$$\mathrm{LD}_i=n\log\left(\frac{n}{n-1}\cdot\frac{n-p-1}{n-p-1+t_i^{*2}}\right)+\frac{n-1}{n-p-1}\cdot\frac{t_i^{*2}}{1-h_{ii}}-1 \tag{2.3}$$

通常は（2.3）式を利用して診断統計量を計算する。

2.3　尤度距離の打切り点

尤度距離 LD_i は，未知回帰係数ベクトル $\boldsymbol{\beta}$ に対する漸近的な信頼領域

$$\{\boldsymbol{\beta}:2[L(\hat{\boldsymbol{\beta}})-L(\boldsymbol{\beta})]\le\chi^2_{p+1}(\alpha)\}$$

と関連して導入されている。ここで，$\chi^2_{p+1}(\alpha)$ は有意水準 α（ただし，$0\le\alpha\le1$）のときの自由度 $p+1$ のカイ2乗分布である。

この漸近的な信頼領域に基づく尤度距離 LD_i の導出方法は，Cook and Weisberg［3］に述べられている（付録 2 を参照）。この結果から，統計的類似性により，$\boldsymbol{\beta}$ を $\hat{\boldsymbol{\beta}}_{(i)}$ とみなすことが可能であれば，LD_i の打切り点は $\chi^2_{p+1}(\alpha)$ となるのである。つまり，

$$LD_i > \chi^2_{p+1}(\alpha) \tag{2.4}$$

となる観測値を影響力が大きいと判定するのである。

けれども，尤度距離 LD_i の打切り点は，$\boldsymbol{\beta}$ に対する漸近的な信頼領域に関する統計的類似性から導かれているため，つぎのような問題点がある。

1. 未知回帰係数ベクトル $\boldsymbol{\beta}$ を $\hat{\boldsymbol{\beta}}_{(i)}$ とみなして，統計的類似性を適用している。
2. 打切り点 $\chi^2_{p+1}(\alpha)$ がデータ数 n に依存していない。このため，n がある程度大きくなると，この打切り点によって影響力が大きいと判定される観測値がほとんどなくなってしまう。

これらの問題点を解消するために，n に依存した打切り点，つまり，Belsley, Kuh and Welsch［1］や Chatterjee and Hadi［2］により提唱されている「データ数に基づいて調整された打切り点」を考案する。これは，第 2 点目の問題点に対する回答になっていることはもちろんであるが，確率分布に依存しない打切り点を導くので第 1 点目の問題点に対する回答の一部にも相当すると考えられる。

3　新たな打切り点の提案

尤度距離 LD_i は対数の項を含むので，このままでは，データ数に基づいて調整された打切り点を考案することは難しい。そこで，以下のような式変形を行なった上で，新たな打切り点を導入する。

$$\begin{aligned} LD_i &= n\log\left(\frac{n}{n-1}\cdot\frac{n-p-1}{n-p-1+t_i^{*2}}\right)+\frac{n-1}{n-p-1}\cdot\frac{t_i^{*2}}{1-h_{ii}}-1 \\ &= n\left[\log\left(1+\frac{1}{n-1}\right)+\log\left(1+\frac{-t_i^{*2}}{n-p-1+t_t^{*2}}\right)\right]+\frac{n-1}{n-p-1}\cdot\frac{t_i^{*2}}{1-h_{ii}}-1 \end{aligned}$$

ここで，一次近似として，$\log(1+x)=x$（ただし，$|x|<1$）を適用すると，

$$LD_i^{(1)} = n\left(\frac{1}{n-1}-\frac{t_i^{*2}}{n-p-1+t_i^{*2}}\right)+\frac{n-1}{n-p-1}\cdot\frac{t_i^{*2}}{1-h_{ii}}-1$$

と一次近似式を導出することができる。さらに，$n-1\simeq n$ とみなせば，

$$LD_i^{(1)*} = \frac{nt_i^{*2}}{n-p}\left(\frac{1}{1-h_{ii}}-\frac{n-p}{n-p+t_i^{*2}}\right) \tag{3.1}$$

を得ることができる。

本論文では，(3.1) 式に対して打切り点を考案する。(3.1) 式において，$h_{ii}=p/n$（てこ比がバランスした状態）で $|t_i^*|>2$（外れ値）となる場合を影響力の大きい観測値であるとみな

すことにすると，

$$\mathrm{LD}_i^{(1)*} > 4n\cdot\frac{4n+p(n-p)}{(n-p)^2(n-p+4)} \tag{3.2}$$

となる。つまり，(3.2) 式の右辺がデータ数に基づいて調整された打切り点ということになる。正確には，(3.2) 式の打切り点が LD_i の打切り点とはならないが，一次近似した打切り点であるものとみなす。

4　打切り点の比較

4.1　打切り点の大小比較

(3.2) 式で示される新たな打切り点と (2.4) 式で与えられる従来の打切り点の大小関係を比較する。比較の方法としていくつかあるが，診断統計量と打切り点の大小関係から影響力を評価するので，2 つの打切り点の差を調べることにする。(2.4) 式と (3.2) 式それぞれの右辺について，その差を計算すると

$$\begin{aligned}&(\text{従来の打切り点})-(\text{新たな打切り点})\\ &=\chi^2_{p+1}(\alpha)-4n\cdot\frac{4n+p(n-p)}{(n-p)^2(n-p+4)}\\ &=\frac{\chi^2_{p+1}(\alpha)(n-p)^2(n-p+4)-16n^2-4np(n-p)}{(n-p)^2(n-p+4)}\end{aligned}$$

となる。少なくとも，データ数 n と説明変数の数 p の間には，$n>p\geq 2$ の関係があるので，この式の分母は明らかに正である。よって，分子が正であるか負であるかによって，2 つの打切り点の大小関係が決まる。したがって，今後は分子部分のみに着目して検討するので，

$$f(n,p,\alpha)\equiv\chi^2_{p+1}(\alpha)(n-p)^2(n-p+4)-16n^2-4np(n-p) \tag{4.1}$$

と置き，(4.1) 式の関数 $f(n,p,\alpha)$ の性質を調べることにする。

ここで問題になるのは，カイ 2 乗分布の分布点 $\chi^2_{p+1}(\alpha)$ の取り扱い方である。もちろん，このまま有意水準 α と自由度 $p+1$ の関数として扱うこともできるが，場合分けの組み合わせパターンが増えるだけである。そこで，ひとつの目安として，自由度を $p=2$ の場合，つまり $\chi^2_3(\alpha)$ の場合を選定する。この理由は，カイ 2 乗分布の性質から，ある有意水準に固定して考えると，自由度が増えると分布点の値も増加する。したがって，$\chi^2_3(\alpha)\leq\chi^2_{p+1}(\alpha)$ であるので，従来の打切り点の最小値となり，この値よりも小さければ明らかに影響力が小さい観測値であると判定できることになる。さらに言えば，従来の打切り点の最小値と新しい打切り点を比較することになるので，新しい打切り点に対しては厳しい状況を選定したものと捉えることもできる。

また，有意水準 α をどの程度に設定するのかについても，いくつかの提案がされている。

最も単純な場合は $\alpha=0.50$，つまり 50% 点を適用するというものである。この根拠は影響力が大きいか小さいかの二者択一式の選択をするので，どちらも半々に考えれば $\alpha=0.50$ を選定することになるであろう，という立場である。けれども，この考え方は実用上では極端な結果になるので，通常の仮説検定における有意水準として適用されている $\alpha=0.05$ や $\alpha=0.01$ などを利用することが多い。本研究では，これらの考え方を踏まえて，有意水準として $\alpha=0.10$ を前提にする。この理由は，カイ 2 乗分布の自由度を固定して考えれば，有意水準を厳しくする（α の値を小さくする）と分布点の値が大きくなり，新たな打切り点が比較の上では有利になるので，ある程度有意水準を大きく設定することにより，従来の打切り点の不利な部分を減らすためである。

参考までに，いくつかの有意水準について分布点を示しておくと，有効数字 6 桁では $\chi^2_3(0.50)=2.36597$，$\chi^2_3(0.20)=4.64163$，$\chi^2_3(0.10)=6.25139$，$\chi^2_3(0.05)=7.81473$，それに $\chi^2_3(0.01)=11.3449$ である。有意水準として $\alpha=0.10$ を前提にすれば，$\chi^2_3(0.10)=6.25139$ であるが，式変形を簡略化するために，以下の検討においては少しだけ従来の打切り点が有利になるけれども，$\chi^2_{p+1}(\alpha)=6$ と置き換える。よって，(4.1) 式を

$$f_\alpha(n,p) = 6(n-p)^2(n-p+4)-16n^2-4np(n-p) \tag{4.2}$$

と変形する。

以上のことを前提にすると，以下の定理を導くことができる。

定理：有意水準 $\alpha\leq 0.10$ かつ $n\geq 3p$ (ただし，$p\geq 2$) のとき，従来の打切り点 $\chi^2_{p+1}(\alpha)$ と新たな打切り点 $4n\{4n+p(n-p)\}/\{(n-p)^2(n-p+4)\}$ の大小関係は常に

$$\chi^2_{p+1}(\alpha) > 4n\cdot\frac{4n+p(n-p)}{(n-p)^2(n-p+4)}$$

となる。

証明：次節のとおり。

4.2 定理の証明

まず，(4.2) 式の性質をデータ数 n と説明変数の数 p の関係から検討する。この両者の関係は $n>p\geq 2$ であり，通常は p がそれほど大きな数ではない。よって，n について (4.2) 式の増減を調べる。(4.2) 式を p についての条件付き関数とし，

$$\begin{aligned} f_\alpha(n|p) &= 6(n-p)^2(n-p+4)-16n^2-4np(n-p) \\ &= 6n^3+(8-22p)n^2+(22p-48)pn+6p^2(4-p) \equiv F(n) \end{aligned}$$

と置き換える。関数 $F(n)$ は n についての 3 次関数であり，n^3 の係数は正の値であるので，n の値がある値 ($F(n)=0$ の実数解の最大値) よりも大きくなると，$F(n)>0$ であり，かつ

単調増加関数になる。また，$F(n)=0$ の実数解は1つであり，他の2つの解は虚数解であることがわかる。実数解 n^{**} は繁雑であるので省略するが正の値になる（付録3を参照)。

つぎに，$F(n)$ の1階微分から単調増加になる範囲を探る。

$$F'(n) = 18n^2+2(8-22p)n+(22p-48)p$$

となるので，$F'(n)=0$ となる実数解 $n^*(<n^{**})$ は

$$n^* = \frac{22p-8\pm 2\sqrt{22\left(p+\frac{32}{11}\right)^2-\frac{1872}{11}}}{18}$$

となる。$F(n)$ が単調増加関数になるのは，n^* の2つの解のうちの最大解 n_+^* よりも大きな値の範囲になるので，

$$n_+^* = \frac{22p-8+2\sqrt{22\left(p+\frac{32}{11}\right)^2-\frac{1872}{18}}}{18} \tag{4.3}$$

について，これがどの程度の値であるかを確認する。

(4.3) 式から，

$$\begin{aligned} n_+^* &= \frac{22p-8+2\sqrt{22\left(p+\frac{32}{11}\right)^2-\frac{1872}{11}}}{18} \\ &< \frac{22p-8+2\sqrt{22\left(p+\frac{32}{11}\right)^2}}{18} = \frac{22p-8+2\left(p+\frac{32}{11}\right)\sqrt{22}}{18} \\ &< \frac{22p-8+2\left(p+\frac{32}{11}\right)\sqrt{25}}{18} = \frac{22p-8+10\left(p+\frac{32}{11}\right)}{18} \\ &= \frac{176p+116}{99} < 2p+\frac{3}{2} < 3p \quad (p\geq 2) \end{aligned}$$

となるので，n_+^* は $3p$ よりも小さい値になり，逆に，少なくとも $n\geq 3p$ の範囲では $F(n)$ は必ず単調増加関数であるといえる。

そこで，$F(n)$ における $n=3p$ の場合について，その関数の正負を調べてみると

$$F(3p) = 162p^3+9(8-22p)p^2+3(22p-48)p^2+6p^2(4-p) = 24p^2(p-2) \geq 0 \quad (p\geq 2)$$

となり，非負の値になる。したがって，$n\geq 3p$ の場合は (4.2) 式も非負の値であるといえる。

参考までに，$F(n)$ における $n=\frac{5}{2}p$ の場合について，その関数の正負を調べると

$$F\left(\frac{5}{2}p\right) = \frac{p^2}{4}(21p-184)$$

となり，$p\geq 9$ のときに $F(n)$ は正の値になる。つまり，$2\leq p\leq 8$ の場合には負の値になる。

したがって，新たな打切り点が従来の打切り点よりも必ず小さくなるのは，有意水準 $\alpha=$

0.10 のときに $n \geq 3p$ の場合であることが判明した。さらに，有意水準が $\alpha \leq 0.10$ の場合は，明らかに従来の打切り点が $\alpha = 0.10$ の場合よりも大きくなるので，有意水準の範囲については拡張することが容易に可能である。よって，4.1 節の定理は証明された。

5　まとめ

本論文では，尤度距離に対する新たな打切り点を導出し，従来の打切り点のとの大小比較を行った。従来の打切り点に関する問題点を整理し，それらを改善するように新たな打切り点は考案されている。また，この新たな打切り点が，従来の打切り点よりも，ある条件下で常に小さくなることを定理としてまとめた。

今後の課題としては，複数個の観測値の影響力評価における打切り点の導出，あるいは導出方法を検討する必要がある。また，ノルム化診断統計量や行列式型診断統計量などの従来から存在する一般的な診断統計量との関連性についても，実用上の観点から詳細に検討することが必要であると考えられる。

付録 1：(2. 3) 式の導出

まずはじめに，対数尤度 $L(\hat{\boldsymbol{\beta}})$ および $L(\hat{\boldsymbol{\beta}}_{(i)})$ を求める。σ^2 の最小 2 乗推定量を $\tilde{\sigma}^2 = \boldsymbol{e}'\boldsymbol{e}/n$ とすると，

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}) &= \log f(\boldsymbol{y} : \hat{\boldsymbol{\beta}}, \tilde{\sigma}^2) \\ &= \log\left[\left(\frac{1}{\sqrt{2\pi\tilde{\sigma}^2}}\right)^n \exp\left\{-\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{2\tilde{\sigma}^2}\right\}\right] \\ &= -\frac{n}{2}\log 2\pi\tilde{\sigma}^2 - \frac{\boldsymbol{e}'\boldsymbol{e}}{2\tilde{\sigma}^2} = -\frac{n}{2}\log 2\pi\tilde{\sigma}^2 - \frac{n}{2} \end{aligned}$$

となり，同様に，

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}_{(i)}) &= \log f\left(\boldsymbol{y} : \hat{\boldsymbol{\beta}}_{(i)}, \tilde{\sigma}^2_{(i)}\right) \\ &= \log\left[\left(\frac{1}{\sqrt{2\pi\tilde{\sigma}^2_{(i)}}}\right)^n \exp\left\{-\frac{(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)})'(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)})}{2\tilde{\sigma}^2_{(i)}}\right\}\right] \end{aligned}$$

となる。このとき，

$$\begin{aligned} (\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)})'(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}) &= \{\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})\}'\{\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})\} \\ &= \left\{\boldsymbol{e}+\mathbf{X}\frac{(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i' e_i}{1-h_{ii}}\right\}'\left\{\boldsymbol{e}+\mathbf{X}\frac{(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i' e_i}{1-h_{ii}}\right\} \\ &= \boldsymbol{e}'\boldsymbol{e}+\frac{h_{ii}e_i^2}{(1-h_{ii})^2} = n\tilde{\sigma}^2+\frac{h_{ii}e_i^2}{(1-h_{ii})^2} \end{aligned}$$

であり，他方，

$$(n-1)\,\tilde{\sigma}^2_{(i)} = n\tilde{\sigma}^2 - \frac{e_i^2}{1-h_{ii}}$$

である。よって，

$$(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)})'(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}) = (n-1)\,\tilde{\sigma}^2_{(i)} + \frac{e_i^2}{(1-h_{ii})^2}$$

となるので，

$$L(\hat{\boldsymbol{\beta}}_{(i)}) = -\frac{n}{2}\log 2\pi\tilde{\sigma}^2_{(i)} - \frac{n-1}{2} - \frac{\left(\dfrac{e_i}{1-h_{ii}}\right)^2}{2\,\tilde{\sigma}^2_{(i)}}$$

と導くことができる。したがって，以上のことから，

$$\mathrm{LD}_i = 2[L(\hat{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}_{(i)})] = n\log\frac{\tilde{\sigma}^2_{(i)}}{\tilde{\sigma}^2} + \frac{\left(\dfrac{e_i}{1-h_{ii}}\right)^2}{\tilde{\sigma}^2_{(i)}} - 1$$

となる。このとき，第一項において

$$\frac{\tilde{\sigma}^2_{(i)}}{\tilde{\sigma}^2} = \frac{\dfrac{\boldsymbol{e}'_{(i)}\boldsymbol{e}_{(i)}}{n-1}}{\dfrac{\boldsymbol{e}'\boldsymbol{e}}{n}} = \frac{n}{n-1}\cdot\frac{(n-p-1)\,\hat{\sigma}^2_{(i)}}{(n-p)\,\hat{\sigma}^2} = \frac{n(n-p-1)}{(n-1)(n-p)}\cdot\frac{t_i^2}{t_i^{*2}}$$

であり，さらに（2.1）式から t_i を t_i^* で表せば，

$$\frac{\tilde{\sigma}^2_{(i)}}{\tilde{\sigma}^2} = \frac{n(n-p-1)}{(n-1)(n-p)}\cdot\frac{t_i^{*2}\dfrac{n-p}{n-p-1+t_i^{*2}}}{t_i^{*2}} = \frac{n}{n-1}\cdot\frac{n-p-1}{n-p-1+t_i^{*2}}$$

となる。また，第二項は，

$$\frac{\left(\dfrac{e_i}{1-h_{ii}}\right)^2}{\tilde{\sigma}^2_{(i)}} = \frac{\left(\dfrac{e_i}{1-h_{ii}}\right)^2}{\dfrac{n-p-1}{n-1}\,\hat{\sigma}^2_{(i)}} = \frac{n-1}{n-p-1}\cdot\frac{t_i^{*2}}{1-h_{ii}}$$

であるので，最終的に（2.3）式として

$$\mathrm{LD}_i = n\log\left(\frac{n}{n-1}\cdot\frac{n-p-1}{n-p-1+t_i^{*2}}\right) + \frac{n-1}{n-p-1}\cdot\frac{t_i^{*2}}{1-h_{ii}} - 1$$

を導くことができる。

付録 2：尤度距離の打切り点の導出

2 種類の未知母数を 1 つのベクトルとしてまとめ，

$$\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\beta}\\ \sigma^2\end{pmatrix}$$

とする。このとき，対数尤度関数 $L(\boldsymbol{\theta})$ を $\boldsymbol{\theta}$ の近傍 $\hat{\boldsymbol{\theta}}$ において，2 次の項まで Taylor 展開すると以下のようになる。

$$L(\boldsymbol{\theta}) = L(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' \frac{\partial L}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' \frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$$

ここで，$\hat{\boldsymbol{\theta}}$ を $\boldsymbol{\theta}$ の最小2乗解とすれば，$\frac{\partial L}{\partial \boldsymbol{\theta}} = \mathbf{0}$ となる。

また，$-\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}$ を Fisher の情報行列（information matrix）に置き換えると，

$$\frac{\partial^2 L}{\partial \boldsymbol{\beta}^2} = -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}, \quad \frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial\sigma^2} = \frac{\mathbf{X}'\boldsymbol{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{(\sigma^2)^2},$$

$$\frac{\partial^2 L}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})}{(\sigma^2)^3},$$

であるから，それぞれの期待値が

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \boldsymbol{\beta}^2}\right) = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2}, \quad \mathrm{E}\left(-\frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial\sigma^2}\right) = 0, \quad \mathrm{E}\left(-\frac{\partial^2 L}{\partial (\sigma^2)^2}\right) = \frac{n}{2(\sigma^2)^2}$$

となり，漸近的に

$$-\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} \simeq \begin{pmatrix} \dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \dfrac{n}{2(\sigma^2)^2} \end{pmatrix}$$

となる。したがって，

$$2[L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})] = (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' \begin{pmatrix} \dfrac{\mathbf{X}'\mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \dfrac{n}{2(\sigma^2)^2} \end{pmatrix} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$$

$$= \frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\sigma^2} + \frac{n}{2}\cdot\frac{(\sigma^2-\tilde{\sigma}^2)^2}{(\sigma^2)^2}$$

となる。

他方，

$$\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\sigma^2} \sim \chi^2_p(\alpha)$$

であり，これと独立に

$$\frac{n}{2}\cdot\frac{(\sigma^2-\tilde{\sigma}^2)^2}{(\sigma^2)^2} \sim \chi^2_1(\alpha)$$

であるから，未知分散 σ^2 を所与とするとき，未知回帰係数ベクトル $\boldsymbol{\beta}$ の信頼領域は漸近的に

$$\{\boldsymbol{\beta} : 2[L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta})] \leq \chi^2_{p+1}(\alpha)\}$$

となる。

以上のことから，統計的類似性により，未知回帰係数ベクトル $\boldsymbol{\beta}$ を $\hat{\boldsymbol{\beta}}_{(i)}$ とみなすことができれば，LD_i の打切り点は $\chi^2_{p+1}(\alpha)$ となる。つまり，

$$\mathrm{LD}_i > \chi^2_{p+1}(\alpha)$$

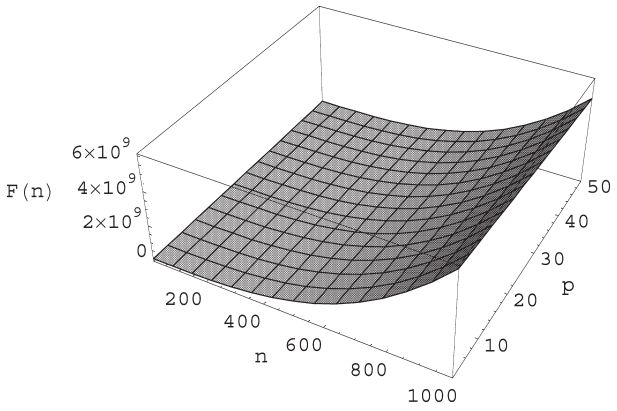

図 1　関数 $F(n)$ のグラフ（$n>0$ の場合）

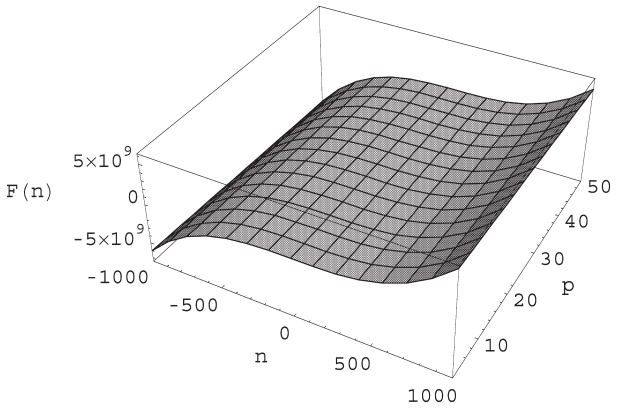

図 2　関数 $F(n)$ のグラフ（$n\leq 0$ を含む場合）

となる場合を影響力の大きい観測値と判定するのである。

付録 3：$F(n)$ の実数解

関数 $F(n)$ のグラフを p を含めて 3 次元グラフとして描くと図 1 のようになる。$F(n)=0$ の解を明確にするために，擬似的に $n\leq 0$ の範囲まで拡張すると図 2 のようになる。図 2 からも $n>0$ となる実数解は 1 つであることが読み取れる。

参 考 文 献

[1] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980), *Regression Diagnostics ; Identifying Influential Data and Sources of Collinearity*, New York : Wiley.

[2] Chatterjee, S. and Hadi, A. S. (1986), Influential observations, high leverage points and

outliers in linear regression, *Statistical Science*, **1**, 379–416.
［3］Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York : Chapman and Hall.
［4］竹内秀一 (1996)，線形回帰における行列式型診断統計量の性質，東京経大学会誌，**199** 号，71–81.
［5］竹内秀一 (1998)，線形回帰におけるてこ比の校正値，人文自然科学論集，**106** 号，97–106.
［6］竹内秀一 (2002)，線形回帰におけるノルム化診断統計量の近似，東京経大学会誌，**231** 号，55–67.
［7］竹内秀一・近河拓也・篠崎信雄 (2000)，複数個の外れ値を検出するときの Cook の距離の検出力，応用統計学，**29**, 83–99.