説明変数空間における観測値の影響力評価

竹　内　秀　一

Assessment of Influence of Observations in the Space Spanned by Explanatory Variables

Hidekazu TAKEUCHI

Several influence measures have been proposed to assess the influence of observations in linear regression. Some such influence measures employ leverages which are the diagonal elements of the hat matrix composed of explanatory variables. In this paper, two new influence measures related to existing leverages are derived. These new influence measures are improved by their use of the eigenvalues (or singular values) and eigenvectors of a matrix based on the space spanned by the explanatory variables. The properties of the new influence measures are demonstrated through the analysis of real and artificial data sets.

1　はじめに

回帰分析における診断統計量（influence measure）に基づく観測値の影響力評価の事例が，Cook and Weisberg[3]，Chatterjee and Hadi[1] それに Weisberg[8] などの研究をはじめ数多く取り上げられている。最近では，データサイエンスの視点からビッグデータへの対応を検討する場合に，1つのアプローチとして回帰分析の応用手法である説明変数（explanatory variable）空間の次元縮小などを適用することが，Cook and Forzani[2] の研究で試みられている。

回帰診断（regression diagnosis）における多くの診断統計量は，説明変数に基づくハット行列（hat matrix）を通して観測値の影響力を評価している。本論文では，診断統計量の主要な構成要素であるハット行列の対角成分，すなわち「てこ比（leverage）」の性質を説

明変数行列に基づく固有値（特異値）および固有ベクトルを利用して多次元的に検討する（もう一つの主要な構成要素である「残差」についての研究は竹内・近河・篠崎[7]などを参照）。これまでにも，Cook and Weisberg[3]において，同様の先行研究はあるが，説明変数行列に基づく固有値および固有ベクトルを利用してデータ分析に適用されることはあまりなかった。本論文において，実際のデータ分析においても利用しやすくなるようにてこ比を修正することにより，新たな指標の適用方法を提案する。

本論文の構成は以下のとおりである。第2節では線形回帰モデルおよび各種の基本的な統計量を与える。第3節において，説明変数行列に基づく固有値（特異値）および固有ベクトルを利用して新たな指標の提案をする。第4節において実データおよび人工データに対して新たな指標を適用することにより，てこ比の特徴を再確認する。第5節は全体のまとめと今後の課題である。

2　定義

本論文では，竹内[6]と同様に，以下の一般的な線形回帰モデルを考える。

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

ただし，$\boldsymbol{y}$ 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，$\boldsymbol{\varepsilon}$ は $n\times 1$ の誤差ベクトルであり，その期待値は $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_n$ で，分散共分散行列は $\mathrm{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$ である。このとき，$\mathbf{0}_n$ は $n\times 1$ の成分がすべて0の列ベクトルであり，σ^2 は未知分散であり，$\mathbf{I}_n$ は $n\times n$ の単位行列であり，$n > p \geq 2$ とする。

また，$\boldsymbol{\beta}$ の最小2乗推定量は $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ となる。ただし，「$'$」は行列やベクトルの転置を表す。ここで，$\boldsymbol{y}$ の予測値ベクトルを $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ とし，この $\boldsymbol{y}$ の係数部分に相当する行列を $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ と定義する。このとき，$\mathbf{H}$ は説明変数行列から構成される予測行列でありハット行列と呼ばれ，その第 (i, j) 成分を $h_{ij} = \boldsymbol{x}_i(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_j'$ と表す。ただし，$\boldsymbol{x}_i$ は説明変数行列

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

の第 i 番目の行ベクトルである。特に，$\mathbf{H}$ の第 i 対角成分 $h_{ii} = \boldsymbol{x}_i(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i'$ を，第 i 番目の観測値に対するてこ比（基本的な性質については竹内[5]を参照）という。

3　てこ比

本節では，まずてこ比に関する先行研究を紹介する。つぎに，説明変数行列に基づく固有値（特異値）および固有ベクトルを利用して，てこ比に関する新たな指標を提案する。

3.1　てこ比の別表現

説明変数行列 $\mathbf{X}$ の第 i 番目の行ベクトル $\boldsymbol{x}_i$ について，$\tilde{\boldsymbol{x}}_i^*=\boldsymbol{x}_i-\bar{\boldsymbol{x}}$ と中心化し，$\tilde{\boldsymbol{x}}_i^*$ の成分から，その第 1 成分を除去（$\boldsymbol{x}_i$ の第 1 成分は定数項 1 のため平均も 1 になり自明であるのでこれを別に分離）して $1\times(p-1)$ の行ベクトルを $\tilde{\boldsymbol{x}}_i$ とする。ここで，$\bar{\boldsymbol{x}}$ は説明変数空間の中心，つまり平均ベクトル

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$$

である。Cook and Weisberg[3] が提案しているように，定数項および中心化した説明変数ベクトル $\tilde{\boldsymbol{x}}_i$ により，てこ比は

$$h_{ii}=\frac{1}{n}+\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'\sum_{k=1}^{p-1}\frac{\cos^2\Theta_{ik}}{\lambda_k}=\frac{1}{n}+\|\tilde{\boldsymbol{x}}_i\|^2\sum_{k=1}^{p-1}\frac{\cos^2\Theta_{ik}}{\lambda_k} \tag{3.1}$$

と分解することができる。ただし，Θ_{ik} は $\tilde{\boldsymbol{x}}_i$ および中心化された説明変数行列

$$\tilde{\mathbf{X}}=\begin{pmatrix}\tilde{\boldsymbol{x}}_1\\ \tilde{\boldsymbol{x}}_2\\ \vdots\\ \tilde{\boldsymbol{x}}_n\end{pmatrix}$$

に基づく，$\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ の第 k 固有値 ($\lambda_k>0$) に対する固有ベクトルとのなす角であり，$\|\boldsymbol{a}\|$ はベクトル $\boldsymbol{a}$ のノルムを表す。(3.1)式から，Cook and Weisberg[3] は，てこ比が大きくなる（説明変数空間における観測値の影響力が大きくなる）場合として，以下の 2 つのことを掲げている。

i）$\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'$（あるいは $\|\tilde{\boldsymbol{x}}_i\|^2$）の値が大きくなる場合，つまり，$\tilde{\boldsymbol{x}}_i$ が中心（平均 $\bar{\boldsymbol{x}}$ より第 1 成分の定数項に対応する部分を除いたもの）から離れている場合

ii）$\sum_{k=1}^{p-1}\frac{\cos^2\Theta_{ik}}{\lambda_k}$ の値が大きくなる場合，つまり，$\tilde{\boldsymbol{x}}_i$ が小さな固有値に対する固有ベクトル方向を向いている場合

これらの性質は，説明変数空間における観測値の影響力評価を行ううえで重要な視点になるが，実際のデータ分析において上記の性質に基づいた検討はあまり行われていない。実際のデータ分析では，各観測値のてこ比の相対的な大小比較あるいは一定値（たとえば，てこ比の平均 p/n の 2 倍程度が目安）を超過しているといったことなどの検討に留まっている

ことが多い。これは，(3.1)式のように元のデータを固有ベクトル空間に変換された多次元空間における観測値の影響力評価を検討することになるため，データ数 (n) や説明変数の数 (p) が多くなると，データ分析を進めていくうえで観測値の特性以外にも数学的な問題を含めて考慮することが必要となり煩雑さが増すものと考えられる。

そこで，てこ比の表現について見直し，データ分析に適用しやすくなるように工夫する。Cook and Weisberg[3] の考え方を一般化すると，ハット行列の第 (i, j) 成分は

$$h_{ij} = \|\boldsymbol{x}_i\| \cdot \|\boldsymbol{x}_j\| \sum_{k=1}^{p} \frac{\cos\theta_{ik} \cdot \cos\theta_{jk}}{\delta_k} \tag{3.2}$$

と表現することができる(付録 A を参照)。ここで，θ_{ik} は説明変数行列 $\mathbf{X}$ に基づく，$\mathbf{X}'\mathbf{X}$ の第 k 固有値 (δ_k) に対する固有ベクトル $\boldsymbol{u}_k$（定義式は付録 A を参照）と説明変数ベクトル $\boldsymbol{x}_i$ のなす角である（固有値の大小関係などについては 3.2 節を参照)。(3.2)式は (3.1)式と異なり，観測値の影響力評価を直接行うことができるように敢えて中心化をしていない。これにより，データ分析において観測値の影響力評価が (3.1)式よりは容易になる。(3.2)式から，ハット行列の対角成分であるてこ比の別表現は，単純に

$$h_{ii} = \|\boldsymbol{x}_i\|^2 \sum_{k=1}^{p} \frac{\cos^2\theta_{ik}}{\delta_k} \tag{3.3}$$

となる。

3.2　説明変数行列の変換

前項の説明変数行列に基づく固有値および固有ベクトルから得られる (3.3)式によるてこ比の別表現を見直し，データ分析に適用しやすくなるように修正することを考える。

説明変数行列 $\mathbf{X}$ を以下のように特異値分解する。

$$\mathbf{X} = \mathbf{L}\mathbf{G}^{\frac{1}{2}}\mathbf{U}'$$

ただし，$\mathbf{G}$ は $p \times p$ の対角行列であり，その第 j 対角成分が $\mathbf{X}'\mathbf{X}$ の第 j 固有値 $\delta_j(>0)$ である（固有値は大きいものから付番する，つまり $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_p$ とする)。その第 j 固有値 δ_j (平方根 $\sqrt{\delta_j}$ が特異値)に対応する固有ベクトルを第 j 列にもつ行列が $\mathbf{U}$ であり，$p \times p$ の正方行列になり $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}_p$ を満たす。また，$\mathbf{L}$ は $n \times p$ の行列であり

$$\mathbf{L} = \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \\ \vdots \\ \boldsymbol{\ell}_n \end{pmatrix}$$

とする。ここで，$\mathbf{L}'\mathbf{L} = \mathbf{I}_p$ を満たし，第 i 番目の行ベクトルは $\boldsymbol{\ell}_i = (\ell_{i1}\ \ell_{i2} \cdots \ell_{ip})$ という成分をもつ。

以上から，ハット行列は

$$\mathbf{H} = \mathbf{L}\mathbf{L}' \tag{3.4}$$

と表される(付録 B を参照)。(3.4)式の第 (i, j) 成分は $h_{ij}=\boldsymbol{\ell}_i\boldsymbol{\ell}_j'$ となる。特に，ハット行列の対角成分であるてこ比は

$$h_{ii} = \boldsymbol{\ell}_i\boldsymbol{\ell}_i' = \sum_{k=1}^{p} \ell_{ik}^2 \tag{3.5}$$

となる。

説明変数空間における新たな指標として，(3.5)式における各次元の影響力を特異値で重み付けすることにより

$$q_{ii} = \frac{1}{\sum_{k=1}^{p} \sqrt{\delta_k}} \sum_{k=1}^{p} \sqrt{\delta_k}\,\ell_{ik}^2 \tag{3.6}$$

を定義する。同様に，(3.5)式における各次元の影響力を固有値で重み付けすることにより

$$q_{ii}^* = \frac{1}{\sum_{k=1}^{p} \delta_k} \sum_{k=1}^{p} \delta_k \ell_{ik}^2 \tag{3.7}$$

も定義する。次節において，これら 2 つの新指標をデータ分析に適用した事例を示す。

なお，上記の (3.6)式および(3.7)式は，以下のような行列の成分として扱うこともできる。(3.6)式の q_{ii} は，以下のような行列

$$\mathbf{Q} = \frac{1}{\mathrm{trace}(\mathbf{G}^{\frac{1}{2}})}\mathbf{L}\mathbf{G}^{\frac{1}{2}}\mathbf{L}' \tag{3.8}$$

を定義すると，この第 i 対角成分になる。ただし，$\mathrm{trace}(\mathbf{A})$ は正方行列 $\mathbf{A}$ の対角成分の和(合計)を表す。同様に，(3.7)式の q_{ii}^* は，以下のような行列

$$\mathbf{Q}^* = \frac{1}{\mathrm{trace}(\mathbf{G})}\mathbf{L}\mathbf{G}\mathbf{L}' \tag{3.9}$$

を定義すると，この第 i 対角成分になる。

4　説明変数空間における観測値の影響力評価

てこ比を修正した新たな指標である (3.6)式および (3.7)式を基に，説明変数空間における観測値の影響力評価を行う。4.1 節において実データの事例を，4.2 節において人工データに基づく事例をそれぞれ示し，てこ比の特徴についても言及する。

4.1　配達時間データ

回帰診断においてよく利用されるデータ分析例の一つとして，Montgomery and Peck［4］に掲げられている「配達時間データ（Delivery Time Data)」を取り上げる。このデータは，ある清涼飲料水会社が，自動販売機への最適配達ルートを分析するために収集したものである。特に，この会社は，そのルートドライバーが自動販売機への配達（配送）に要する時間を予測することに興味をもっている。目的変数（本論文ではデータを省略）は，配達に要する時間 (y) であり，これに影響を与えている重要な要因（説明変数）は，X_1 が自動販売機に補充された清涼飲料水のケース数（個）であり，X_2 がルートドライバーの歩いた距離（フィート）である。X_1（横軸）および X_2（縦軸）の散布図を図 4.1（No. 15 および No. 23 が誤差の範囲で重なって打点されている）に示す。

本論文では，説明変数空間における個々の観測値の影響力を検討するので，ハット行列の対角成分であるてこ比 h_{ii} と (3.5)式における各次元の成分 $\ell_{i1}^2, \ell_{i2}^2, \ell_{i3}^2$，それに新指標である q_{ii} および q_{ii}^* を表 4.1 にまとめた。なお，計算結果は R によるものであり，特異値（括弧内が固有値）は，$\sqrt{\delta_1}=2{,}593.9\,(\delta_1=6{,}728{,}369.4)$，$\sqrt{\delta_2}=19.74\,(\delta_2=389.72)$，それに $\sqrt{\delta_3}=2.970$ $(\delta_3=8.819)$ となる。

表 4.1 の結果から，説明変数空間における個々の観測値の影響力評価をすると，てこ比の値だけに注目した場合は，目安となる $2p/n=0.24$ を超える No. 9 $(X_1=30, X_2=1460)$ および No. 22 $(X_1=26, X_2=810)$ が影響力の大きい観測値として検出される。図 4.1 の散布図か

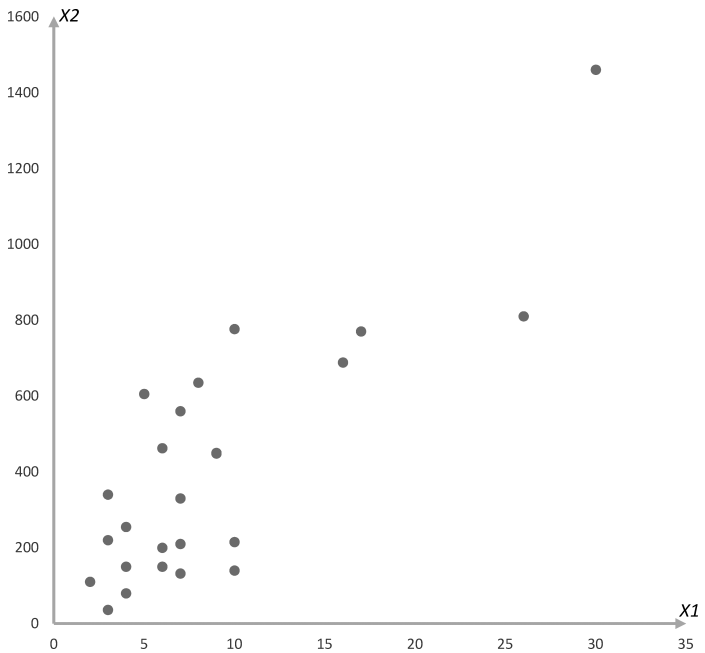

図 4.1　配達時間データの散布図

表 4.1　配達時間データの結果

No.	h_{ii}	ℓ_{i1}^2	ℓ_{i2}^2	ℓ_{i3}^2	q_{ii}	q_{ii}^*
1	0.10180	0.04661	0.04397	0.01122	0.04655	0.04661
2	0.07070	0.00719	0.00469	0.05882	0.00723	0.00719
3	0.09873	0.01718	0.03601	0.04554	0.01735	0.01718
4	0.08537	0.00095	0.01528	0.06914	0.00114	0.00095
5	0.07501	0.00335	0.02374	0.04793	0.00355	0.00335
6	0.04287	0.01619	0.00052	0.02616	0.01609	0.01619
7	0.08180	0.00180	0.00006	0.07994	0.00187	0.00180
8	0.06373	0.00656	0.02075	0.03642	0.00670	0.00656
9	0.49829	0.31694	0.00200	0.17936	0.31441	0.31692
10	0.19630	0.05440	0.12700	0.01490	0.05490	0.05440
11	0.08613	0.07039	0.01355	0.00220	0.06988	0.07038
12	0.11366	0.00688	0.08456	0.02221	0.00748	0.00688
13	0.06112	0.00967	0.00284	0.04862	0.00966	0.00967
14	0.07824	0.03173	0.02600	0.02052	0.03167	0.03173
15	0.04111	0.02984	0.00002	0.01125	0.02960	0.02984
16	0.16594	0.08951	0.07626	0.00017	0.08931	0.08951
17	0.05943	0.00595	0.01072	0.04276	0.00603	0.00595
18	0.09626	0.00259	0.04967	0.04400	0.00300	0.00260
19	0.09645	0.00019	0.01380	0.08246	0.00039	0.00019
20	0.10168	0.08816	0.00708	0.00644	0.08746	0.08816
21	0.16528	0.00292	0.13439	0.02797	0.00394	0.00293
22	0.39158	0.09760	0.24915	0.04483	0.09868	0.09761
23	0.04126	0.03011	0.00001	0.01114	0.02986	0.03011
24	0.12061	0.05994	0.05516	0.00551	0.05984	0.05994
25	0.06664	0.00335	0.00279	0.06050	0.00341	0.00335

らも，右上方向に離れている 2 つのデータがこれらであることがわかる。

これを特異値分解あるいは固有値分解した結果から見直すと，No. 9 は第 1 次元 ℓ_{i1}^2 の値が大きく，No. 22 は第 2 次元 ℓ_{i2}^2 の値が大きいことがわかる。ただし，新指標である q_{ii} あるいは q_{ii}^* の値から考えると，No. 9 の値が他のデータに比較して著しく大きく，説明変数空間における影響力の大きいデータとみなせるが，No. 22 は 2 番目に大きいが No. 16（$X_1=10$, $X_2=776$）や No. 20（$X_1=17$, $X_2=770$）と大きな違いはない。これは，第 1 特異値（固有値）と第 2 特異値（固有値）の差異が大きいために，第 2 特異値（固有値）において影響力の大きい No. 22 が過少に評価された結果であると考えられる。

こうした点が，従来から提案されている (3.1)式や (3.3)式のような固有ベクトル空間まで細かく分解して検討しなくても，ある程度解明できるものといえる。

4.2　人工データ

この人工データは，説明変数空間における個々の観測値の影響力評価をイメージしやすくするために単純化して構成したものである。表 4.2 において示されているように，No. 1〜10 までのデータは $X_1^2+X_2^2=10^2$ に基づく円周上の点であり，No. 11〜15 までのデータは $X_1^2+X_2^2=5^2$ に基づく円周上の点であり，No. 16〜20 までのデータは $X_1^2+X_2^2=7.5^2$ に基づく円周上の点である。具体的に，データは図 4.2 の散布図のように配置される。

本論文では，説明変数空間における個々の観測値の影響力を検討するので，4.1 節と同様にハット行列の対角成分であるてこ比とその各成分，それに 2 つの新指標を表 4.3 にまとめ

表 4.2　人工データ

No.	X_1	X_2	No.	X_1	X_2
1	10	0.00	11	1	4.90
2	8	−6.00	12	3	4.00
3	6	8.00	13	5	0.00
4	4	−9.17	14	−2	4.58
5	2	9.80	15	−4	3.00
6	−2	−9.80	16	−1	−7.43
7	−4	9.17	17	−3	−6.87
8	−6	−8.00	18	−5	−5.59
9	−8	6.00	19	2	−7.23
10	−10	0.00	20	4	−6.34

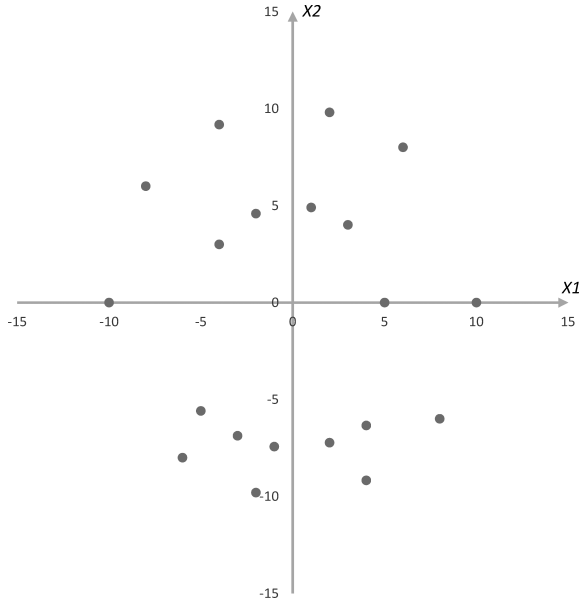

図 4.2　人工データの散布図

表 4.3　人工データの結果

No.	h_{ii}	ℓ_{i1}^2	ℓ_{i2}^2	ℓ_{i3}^2	q_{ii}	q_{ii}^*
1	0.23369	0.00064	0.18132	0.05173	0.07864	0.07081
2	0.19407	0.05043	0.10382	0.03982	0.07148	0.07081
3	0.21374	0.06598	0.07841	0.06934	0.07134	0.07081
4	0.15824	0.10410	0.02019	0.03396	0.06427	0.07087
5	0.19416	0.10756	0.01327	0.07333	0.06627	0.07084
6	0.15427	0.10847	0.01331	0.03248	0.06358	0.07084
7	0.19462	0.10321	0.02025	0.07116	0.06673	0.07087
8	0.18045	0.06669	0.07852	0.03523	0.06910	0.07081
9	0.21699	0.04981	0.10394	0.06324	0.07304	0.07081
10	0.23205	0.00057	0.18149	0.04999	0.07854	0.07081
11	0.09167	0.02678	0.00331	0.06158	0.01986	0.01824
12	0.09574	0.01641	0.01959	0.05974	0.02107	0.01823
13	0.09677	0.00017	0.04531	0.05129	0.02263	0.01823
14	0.09128	0.02564	0.00507	0.06057	0.01992	0.01821
15	0.09526	0.01238	0.02600	0.05688	0.02141	0.01823
16	0.10395	0.06307	0.00429	0.03659	0.03693	0.04011
17	0.11122	0.05160	0.02219	0.03743	0.03845	0.04010
18	0.12451	0.03193	0.05301	0.03958	0.04116	0.04014
19	0.10447	0.06335	0.00395	0.03716	0.03698	0.04015
20	0.11286	0.05122	0.02274	0.03890	0.03859	0.04010

た。なお，4.1 節と同様に，計算結果も R によるものであり，特異値（括弧内が固有値）は，$\sqrt{\delta_1}=29.30(\delta_1=858.3), \sqrt{\delta_2}=23.42(\delta_2=548.4)$，それに $\sqrt{\delta_3}=4.43(\delta_3=19.7)$ となる。

表 4.3 のてこ比の数値から，説明変数空間の周辺（この例の場合は図 4.2 における外側の円周上のデータ No. 1〜10）の影響力が，内部（この例の場合は図 4.2 における内側の 2 つの円周上のデータ No. 11〜20）よりも大きくなることがわかる。表 4.3 の結果から，説明変数空間における個々の観測値の影響力評価をすると，てこ比の値だけに注目した場合は，一般的な目安である $2p/n=0.300$ を超える観測値はないが，次善の目安となる $1.5p/n=0.225$ を超える No. 1 および No. 10 が影響力の大きい観測値として検出される。図 4.2 の散布図からも，X_1 軸（横軸）方向に最も離れている両端の 2 つのデータがこれらであることがわかる。

これを特異値分解あるいは固有値分解した結果から見直すと，No. 1 および No. 10 は第 2 次元 ℓ_{i2}^2 の値が（0.1 を超えて）大きく，第 1 次元 ℓ_{i1}^2 の値が（0.1 を超えて）大きいデータは X_2 軸（縦軸）方向に最も離れている 4 つのデータ No. 4〜7 であることがわかる。ただし，新指標である q_{ii} の値から考えると，外側の円周上のデータ No. 1〜10 の値が内側の円周上のデータに比較して大きいことがわかる。また，q_{ii}^* の値からも，同様のことがわかる。な

お，この人工データは，第１特異値（固有値）と第２特異値（固有値）の差異があまり大きくないために，１次元と２次元についての重要度における差があまりない。

以上から，説明変数空間においては，空間の周辺部ほど影響力が大きい観測値として検出されやすいことが判明した。4.1 節の実データから，説明変数空間における外れ値に相当するデータが影響力の大きい観測値とみなされやすい印象を受けるが，厳密には説明変数空間の周辺部であり，かつデータの密集度が低い場合に影響力の大きい観測値と判断されやすいことがわかる。

5　まとめと今後の課題

本論文では，回帰診断における主要な構成要素であるてこ比を，説明変数に関わる固有値（特異値）および固有ベクトルを利用して修正することにより新たな指標を提案した。この新たな指標の特徴をデータ分析事例を通して，Cook and Weisberg[3] が提案する(3.1)式あるいは本論文で示した (3.3)式を利用して固有ベクトル方向まで細かく分解して検討しなくても，代替的な方法として (3.6)式や(3.7)式の新指標を利用すれば観測値の影響力を十分に評価できることを示した。

また，従来の Cook and Weisberg[3] が提案するてこ比の表現を一般化し，(3.2)式のようにハット行列の非対角成分を含めることも試みた。新指標については (3.8)式や(3.9)式のように行列形式への拡張も可能であるので，この非対角成分の利用方法については，複数個の観測値の影響力評価への拡張を含め今後の検討課題としたい。

付録 A：(3.2)式の導出

(3.2)式の導出過程を示す。

$$
\begin{aligned}
h_{ij} &= \boldsymbol{x}_i(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_j' = \boldsymbol{x}_i'(\mathbf{U}\mathbf{G}^{\frac{1}{2}}\mathbf{L}'\mathbf{L}\mathbf{G}^{\frac{1}{2}}\mathbf{U}')^{-1}\boldsymbol{x}_j' \\
&= \boldsymbol{x}_i'(\mathbf{U}\mathbf{G}\mathbf{U}')^{-1}\boldsymbol{x}_j' = \boldsymbol{x}_i'\mathbf{U}\mathbf{G}^{-1}\mathbf{U}'\boldsymbol{x}_j' \\
&= \sum_{k=1}^{p}\frac{\boldsymbol{x}_i\boldsymbol{u}_k\boldsymbol{u}_k'\boldsymbol{x}_j'}{\delta_k} \\
&= \|\boldsymbol{x}_i\|\cdot\|\boldsymbol{x}_j\|\sum_{k=1}^{p}\frac{1}{\delta_k}\cdot\frac{\boldsymbol{x}_i\boldsymbol{u}_k}{\|\boldsymbol{x}_i\|\cdot\|\boldsymbol{u}_k\|}\cdot\frac{\boldsymbol{x}_j\boldsymbol{u}_k}{\|\boldsymbol{x}_j\|\cdot\|\boldsymbol{u}_k\|} \\
&= \|\boldsymbol{x}_i\|\cdot\|\boldsymbol{x}_j\|\sum_{k=1}^{p}\frac{\cos\theta_{ik}\cdot\cos\theta_{jk}}{\delta_k}
\end{aligned}
$$

と表現することができる。ここで，θ_{ik} は固有ベクトル $\boldsymbol{u}_k$ と説明変数ベクトル $\boldsymbol{x}_i$ のなす角である。また，適宜，$\mathbf{L}'\mathbf{L}=\mathbf{I}_p$ および $\mathbf{U}'\mathbf{U}=\mathbf{U}\mathbf{U}'=\mathbf{I}_p$ を利用した。なお，$\mathbf{U}$ は $p\times p$ の行列

であり，列ベクトルで表現すると

$$\mathbf{U} = (\boldsymbol{u}_1 \ \boldsymbol{u}_2 \ \cdots \ \boldsymbol{u}_p)$$

となり，$\mathbf{U}'\mathbf{U}=\mathbf{I}_p$ から $k=1, 2, \cdots, p$ について $\boldsymbol{u}_k'\boldsymbol{u}_k=\|\boldsymbol{u}_k\|^2=1$（あるいは $\|\boldsymbol{u}_k\|=1$）となる。

よって，$\boldsymbol{x}_i\mathbf{U}$ の第 k 列（第 k 成分）は，内積 $\boldsymbol{x}_i\boldsymbol{u}_k$ であり

$$\boldsymbol{x}_i\boldsymbol{u}_k = \|\boldsymbol{x}_i\|\cdot\|\boldsymbol{u}_k\|\cos\theta_{ik} \quad \text{あるいは} \quad \cos\theta_{ik} = \frac{\boldsymbol{x}_i\boldsymbol{u}_k}{\|\boldsymbol{x}_i\|\cdot\|\boldsymbol{u}_k\|}$$

となる。内積 $\boldsymbol{u}_k'\boldsymbol{x}_j'=\boldsymbol{x}_j\boldsymbol{u}_k$ についても同様に表現することができる。

付録 B：(3. 4)式の導出

説明変数行列 $\mathbf{X}$ を特異値分解することにより

$$\begin{aligned}\mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{L}\mathbf{G}^{\frac{1}{2}}\mathbf{U}'(\mathbf{U}\mathbf{G}^{\frac{1}{2}}\mathbf{L}'\mathbf{L}\mathbf{G}^{\frac{1}{2}}\mathbf{U}')^{-1}\mathbf{U}\mathbf{G}^{\frac{1}{2}}\mathbf{L}' \\ &= \mathbf{L}\mathbf{G}^{\frac{1}{2}}\mathbf{U}'(\mathbf{U}\mathbf{G}\mathbf{U}')^{-1}\mathbf{U}\mathbf{G}^{\frac{1}{2}}\mathbf{L}' \\ &= \mathbf{L}\mathbf{G}^{\frac{1}{2}}\mathbf{U}'\mathbf{U}\mathbf{G}^{-1}\mathbf{U}'\mathbf{U}\mathbf{G}^{\frac{1}{2}}\mathbf{L}' \\ &= \mathbf{L}\mathbf{L}'\end{aligned}$$

と表される。

参 考 文 献

［1］ Chatterjee, S. and Hadi, A. S. (1988), *Sensitivity Analysis in Linear Regression*, New York: Wiley.

［2］ Cook, R. D. and Forzani, L. (2018), Big data and partial least-squares prediction, *The Canadian Journal of Statistics*, **46**, 62–78.

［3］ Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.

［4］ Montgomery, D. C. and Peck, E. A. (1992), *Introduction to Linear Regression Analysis*, Second Edition, New York: Wiley.

［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（2018），新たな予測行列に基づく診断統計量，人文自然科学論集，**142** 号，3–20.

［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.

［8］ Weisberg, S. (2014), *Applied Linear Regression*, Fourth Edition, New York: Wiley.