# 線形回帰分析における複数個の観測値についての誤差分散の影響力評価

竹　内　秀　一

# Assessment of influence of estimated error variance based on multiple observations in linear regression

Hidekazu TAKEUCHI

In linear regression, the unbiased estimator of error variance has traditionally been addressed in a formalistic manner. This paper proposes a new estimator for error variance, aiming to overcome a problem in regression diagnostics when accounting for multiple observations. The proposed estimator allows for a representation grounded in individual observations, thereby simplifying the interpretation of sets of multiple observations. Moreover, it makes it easier to differentiate between the distinct components of each observation and the combinatorial effects resulting from interactions between observations. As a result, it becomes feasible to clearly discern the impact of individual observations that form the sets of multiple observations from their inherent combined effects. This representation not only paves the way for a more reduced expression but also stands to offer significant utility in data analysis.

## 1　はじめに

回帰分析における診断統計量（influence measure）を利用した観測値の影響力評価の事例が，Cook and Weisberg[2]，Chatterjee and Hadi[1] それに Weisberg[9] などの研究をはじめ数多く取り上げられている。また，複数個の観測値集合に基づいた影響力評価についても，個々の観測値の情報を利用した新たな提案が Nurunnabi, Hadi and Imon[4] や Nur-

unnabi, Nasser and Imon［5］などにより示されている。しかしながら，診断統計量の基本的な構成要素である誤差分散の推定量については，個々の観測値でも複数個の観測値集合の場合でも形式的な取り扱い方をすることが多く，詳細な検討が十分になされていない。

通常の診断統計量は，ハット行列（予測行列）の対角成分，すなわち「てこ比（leverage）」および「残差」（この研究については竹内・近河・篠崎［8］などを参照）を主要な構成要素としている。後者の「残差」については，線形回帰モデルの誤差分散としてデータから推定したものを適用することになるが，形式的に不偏分散を利用することが多い。個々の観測値についての誤差分散の推定量として，形式的な不偏分散を利用することに大きな問題はないが，複数個の観測値集合の場合は，観測値間の相乗効果のようなものがあり，変数間の交互作用効果を考慮するときと同様の対応が必要となる。

そこで，本研究では，複数個の観測値の相乗効果と個々の観測値の単独の効果の関係を調べるために，複数個の観測値集合の誤差分散の不偏推定量を個々の観測値に基づく効果とそれ以外の相乗効果等に分離した表現を提案する。また，相乗効果を省略して簡略化された表現に基づく複数個の観測値集合の誤差分散を利用することでも，実用上は一定程度の影響力評価が行えることを例示する。この際に，複数個の観測値集合の場合に生じるマスク効果（masking effect）が小さい，あるいは無視できる程度の場合であれば，実際のデータ分析においては提案する誤差分散の簡略化された表現で十分に対応できることも示す。

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，一般化した誤差分散の不偏推定量の新たな表現を提案する。第４節では，実データに適用した場合において，誤差分散の推定量の新たな表現およびその簡略化された表現の特徴を確認する。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，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 $n\times q$ のフルランクの説明変数行列，$\boldsymbol{\beta}$ は $q\times 1$ 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，$\sigma^2$ は未知の誤差分散であり，それに $\mathbf{I}_n$ は $n\times n$ の単位行列であり，$n>q\geq 2$ とする。

また，$\boldsymbol{\beta}$ の最小２乗推定量は $\widehat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ となり，$\sigma^2$ の不偏推定量は $\hat{\sigma}^2=\boldsymbol{e}'\boldsymbol{e}/(n-q)$ となる。ただし，「′」は行列やベクトルの転置を表し，$\boldsymbol{e}(=\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})$ は残差ベクトルであ

る。

ここで，$\boldsymbol{y}$ の予測値ベクトルを $\hat{\boldsymbol{y}}=\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ 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（hat matrix）と呼ばれる。

### 2.1　観測値集合の分割

回帰診断の記号法を Chatterjee and Hadi[1] に従って示しておく。行列やベクトルの添字 $I$ は $n$ 個のデータ全体から取り除かれる $m$ 個の観測値の部分集合，つまり，影響力評価の対象となる複数個の観測値集合（集合という意味では 1 個の観測値の場合も含む）を表す。また，添字 $(I)$ はその $m$ 個の観測値集合 $I$ 以外の残りの $n-m$ 個の観測値集合を表す。このとき，一般性を失うことなく，影響力評価の対象となる観測値集合 $I$ はデータあるいは統計量の後半に集中しているものと考える。つまり，$\boldsymbol{y}$，$\mathbf{X}$，$\boldsymbol{e}$，それに $\hat{\boldsymbol{y}}$ は，

$$\boldsymbol{y}=\begin{pmatrix}\boldsymbol{y}_{(I)}\\ \boldsymbol{y}_I\end{pmatrix},\quad \mathbf{X}=\begin{pmatrix}\mathbf{X}_{(I)}\\ \mathbf{X}_I\end{pmatrix},\quad \boldsymbol{e}=\begin{pmatrix}\boldsymbol{e}_{(I)}\\ \boldsymbol{e}_I\end{pmatrix},\quad \hat{\boldsymbol{y}}=\begin{pmatrix}\hat{\boldsymbol{y}}_{(I)}\\ \hat{\boldsymbol{y}}_I\end{pmatrix}$$

というようにそれぞれの部分集合ごとに 2 分割されているものとする。これは影響力評価の対象となる観測値集合 $I$ がデータあるいは統計量の後方にまとまるように，単純に並べ替えただけのことである。ただし，本論文で具体的な観測値集合を扱う場合，たとえば，$m=1$ の場合（個々の観測値の影響力を評価する診断統計量の場合）は添字について，$I$ は $i$ と，$(I)$ は $(i)$ と表示する。

このようにデータ $\boldsymbol{y}$ および $\mathbf{X}$ を 2 つの観測値集合に分けておくと，観測値集合 $I$ の $m$ 個の観測値を取り除いたときの $\boldsymbol{\beta}$ の最小 2 乗推定量は $\hat{\boldsymbol{\beta}}_{(I)}=(\mathbf{X}'_{(I)}\mathbf{X}_{(I)})^{-1}\mathbf{X}'_{(I)}\boldsymbol{y}_{(I)}$ となる。また，$\mathbf{H}$ については

$$\mathbf{H}=\begin{pmatrix}\mathbf{X}_{(I)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(I)} & \mathbf{X}_{(I)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_I\\ \mathbf{X}_I(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(I)} & \mathbf{X}_I(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_I\end{pmatrix}\tag{2.1}$$

と 4 分割される。ただし，(2.1)式における分割行列の第（2,2）成分を $\mathbf{H}_I=\mathbf{X}_I(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_I$ とする。また，$\mathbf{H}$ の第 $(j,k)$ 成分を $h_{jk}=\boldsymbol{x}_j(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}'_k$ と表す。ここで，$\boldsymbol{x}_j$ および $\boldsymbol{x}_k$ はそれぞれ $\mathbf{X}$ の第 $j$ 番目および第 $k$ 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（性質については竹内[6] を参照）という。

以上のように，回帰診断においては，観測値集合 $I$ を対象として影響力評価を直接的に行うか，あるいはそれ以外の観測値集合 $(I)$ を対象として，全データ（すべての観測値）を利用した場合との差異により，間接的に観測値集合 $I$ の影響力評価を行うか，このどちらかになる。ただし，後者については，間接的な場合の診断統計量を式変形することにより，直接的な影響力評価を行う指標に直せる場合が多い。

### 2.2 残差の相関行列

(2.1)式の第（2,2）成分である $\mathbf{H}_I$ の場合と同様に，残差ベクトル $\boldsymbol{e}$ の相関行列（correlation matirix）

$$\mathbf{R} = [\mathrm{diag}(\mathbf{I}_n-\mathbf{H})]^{-\frac{1}{2}}(\mathbf{I}_n-\mathbf{H})[\mathrm{diag}(\mathbf{I}_n-\mathbf{H})]^{-\frac{1}{2}}$$

を 4 分割するときの第（2,2）成分の行列を

$$\mathbf{R}_I=[\mathrm{diag}(\mathbf{I}_m-\mathbf{H}_I)]^{-\frac{1}{2}}(\mathbf{I}_m-\mathbf{H}_I)[\mathrm{diag}(\mathbf{I}_m-\mathbf{H}_I)]^{-\frac{1}{2}} \tag{2.2}$$

と定義する。ただし，diag(**D**) は正方行列 **D** の対角成分のみを取り出し，非対角成分がすべて 0 の行列を表す。

また，**R** あるいは（2.2)式の第 $(j,k)$ 成分は

$$r_{jk} = -\frac{h_{jk}}{\sqrt{(1-h_{jj})(1-h_{kk})}}$$

である（$j=k$ の場合は，明らかに $r_{jj}=1$ である)。なお，(2.2)式において，$j\neq k$ に対しては $\mathbf{R}_I$ が正則行列となるように，$r_{jk}\neq\pm1$ であると仮定する。

最後に，残差ベクトル $\boldsymbol{e}$ を標準化した標準化残差ベクトル

$$\boldsymbol{t} = \frac{1}{\hat{\sigma}}[\mathrm{diag}(\mathbf{I}_n-\mathbf{H})]^{-\frac{1}{2}}\boldsymbol{e}$$

を 2 分割するときの後半（下の部分）のベクトルとして，観測値集合 $I$ に対する残差ベクトル $\boldsymbol{e}_I$ を標準化して

$$\boldsymbol{t}_I = \frac{1}{\hat{\sigma}}[\mathrm{diag}(\mathbf{I}_m-\mathbf{H}_I)]^{-\frac{1}{2}}\boldsymbol{e}_I \tag{2.3}$$

と定義する。また，$\boldsymbol{t}$ あるいは (2.3)式の第 $i$ 成分は

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

である。

## 3 誤差分散の不偏推定量

本節では，観測値集合 $I$ を除去した場合の誤差分散の推定量について，観測値集合 $I$ の具体的な構成要素に基づいた表現を提案する。この構成要素に基づく表現により，観測値集合 $I$ の各成分を除去した場合（$m=1$ の個々の観測値に対する影響力評価の場合）との比較が可能となる。

### 3.1　誤差分散の不偏推定量の定義

観測値集合 $I$ の影響力を調べるために，その観測値集合 $I$ を除去した場合の誤差分散の推定量として

$$\hat{\sigma}^2_{(I)} = \frac{\boldsymbol{e}'_{(I)}\boldsymbol{e}_{(I)}}{n-q-m} = \frac{(n-q)\hat{\sigma}^2 - \boldsymbol{e}'_I(\mathbf{I}-\mathbf{H}_I)^{-1}\boldsymbol{e}_I}{n-q-m} \tag{3.1}$$

を定義する。この (3.1)式を観測値集合 $I$ の残差ベクトルの相関行列である (2.2)式および標準化残差ベクトルである (2.3)式を利用して表すと

$$\hat{\sigma}^2_{(I)} = \frac{\hat{\sigma}^2}{n-q-m}(n-q-\boldsymbol{t}'_I\mathbf{R}_I^{-1}\boldsymbol{t}_I) \tag{3.2}$$

となる。特に，$m=1$ の場合は

$$\hat{\sigma}^2_{(i)} = \frac{n-q-t_i^2}{n-q-1}\hat{\sigma}^2 \tag{3.3}$$

と表す。

### 3.2　誤差分散の不偏推定量の新たな表現

観測値集合 $I$ を具体的な成分で表現する。つまり，$m=1$ の場合は $I=\{i\}$，$m=2$ の場合は $I=\{i,j\}(i<j)$，そして $m=3$ の場合は $I=\{i,j,k\}(i<j<k)$ を例示的に取り上げる。$m\geq 4$ の場合も同様であるが，本論文においては，複数個の観測値集合の基本的な性質を確認するために，よく利用されるケース（$m=1,2,3$）について取り上げ，参考までに $m=4$ の場合についても $m=4$ 以上となることを想定した表現を提示する。

個々の観測値の影響力を評価するための $m=1$ の場合である $I=\{i\}$，つまり $\hat{\sigma}^2_{(i)}$ についてはすでに (3.3)式において表現式を示している。この $m=1$ の場合の表現を利用して，複数個の観測値集合 $m=2$ および $m=3$ の場合を表すことを考える。

まず，$m=2$ の場合である $I=\{i,j\}(i<j)$ について，(3.2)式から具体的には

$$\begin{aligned}\hat{\sigma}^2_{(I)} = \hat{\sigma}^2_{(ij)} &= \frac{\hat{\sigma}^2}{n-q-2}\left[n-q-(t_i\ \ t_j)\begin{pmatrix}1 & r_{ij}\\ r_{ij} & 1\end{pmatrix}^{-1}\begin{pmatrix}t_i\\ t_j\end{pmatrix}\right]\\ &= \frac{\hat{\sigma}^2}{n-q-2}\left[n-q-\frac{t_i^2-2r_{ij}t_it_j+t_j^2}{1-r_{ij}^2}\right]\end{aligned} \tag{3.4}$$

となる。ここで，(3.3)式を変形して

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

および $t_j^2$ のみを置き換えると，(3.4)式は

$$\hat{\sigma}^2_{(ij)} = \frac{\hat{\sigma}^2}{n-q-2}\left[\frac{n-q-1}{\hat{\sigma}^2(1-r_{ij}^2)}\{\hat{\sigma}^2_{(i)}+\hat{\sigma}^2_{(j)}\}+\frac{1}{1-r_{ij}^2}\{2r_{ij}t_it_j-(n-q)(1+r_{ij}^2)\}\right]$$

$$=\frac{n-q-1}{n-q-2}\cdot\frac{\tilde{\sigma}^2_{(i)}+\tilde{\sigma}^2_{(j)}}{1-r_{ij}^2}-\frac{\tilde{\sigma}^2}{n-q-2}\cdot\frac{(n-q)(1+r_{ij}^2)-2r_{ij}t_it_j}{1-r_{ij}^2} \tag{3.5}$$

と表現できる。(3.5)式の第2表現において，第1項は$m=2$の場合を構成する第$i$番目の観測値および第$j$番目の観測値を除去したときのそれぞれの誤差分散の推定量を基にして表すことができ，第2項は第$i$番目の観測値および第$j$番目の観測値の組合せ効果（相乗効果）とみなすことができる。

加えて，(3.5)式において，$r_{ij}=0$とすれば

$$\tilde{\sigma}^{2*}_{(ij)}=\frac{n-q-1}{n-q-2}\{\tilde{\sigma}^2_{(i)}+\tilde{\sigma}^2_{(j)}\}-\frac{n-q}{n-q-2}\tilde{\sigma}^2 \tag{3.6}$$

と簡略化される。さらに加えて，(3.6)式において，$\tilde{\sigma}^2_{(i)}=\tilde{\sigma}^2_{(j)}=\tilde{\sigma}^2$であれば，$\tilde{\sigma}^{2**}_{(ij)}=\tilde{\sigma}^2$となり，$\sigma^2$の不偏推定量に一致する。

つぎに，$m=3$の場合である$I=\{i,j,k\}(i<j<k)$について，(3.2)式から以下のように表現することができる。

$$\begin{aligned}\tilde{\sigma}^2_{(I)}&=\tilde{\sigma}^2_{(ijk)}\\&=\frac{\tilde{\sigma}^2}{n-q-3}\left[n-q-(t_i\ t_j\ t_k)\begin{pmatrix}1&r_{ij}&r_{ik}\\r_{ij}&1&r_{jk}\\r_{ik}&r_{jk}&1\end{pmatrix}^{-1}\begin{pmatrix}t_i\\t_j\\t_k\end{pmatrix}\right]\\&=\frac{\tilde{\sigma}^2}{n-q-3}\left[n-q-\frac{1}{|\mathbf{R}_{ijk}|}\{(1-r_{jk}^2)t_i^2+(1-r_{ik}^2)t_j^2+(1-r_{ij}^2)t_k^2\right.\\&\qquad\left.+2(r_{ik}r_{jk}-r_{ij})t_it_j+2(r_{ij}r_{jk}-r_{ik})t_it_k+2(r_{ij}r_{ik}-r_{jk})t_jt_k\}\right]\end{aligned} \tag{3.7}$$

ただし，行列$\mathbf{R}_I=\mathbf{R}_{ijk}$の行列式$|\mathbf{R}_{ijk}|$は

$$|\mathbf{R}_{ijk}|=1+2r_{ij}r_{ik}r_{jk}-(r_{ij}^2+r_{ik}^2+r_{jk}^2)=(1-r_{ij}^2)+(1-r_{ik}^2)+(1-r_{jk}^2)-2(1-r_{ij}r_{ik}r_{jk})$$

である。ここで，$m=2$の場合と同じく(3.3)式を変形して$t_i^2$，$t_j^2$それに$t_k^2$のみを置き換えると，(3.7)式は

$$\begin{aligned}\tilde{\sigma}^2_{(ijk)}&=\frac{\tilde{\sigma}^2}{n-q-3}\left[\frac{n-q-1}{\tilde{\sigma}^2|\mathbf{R}_{ijk}|}\{(1-r_{jk}^2)\tilde{\sigma}^2_{(i)}+(1-r_{ik}^2)\tilde{\sigma}^2_{(j)}+(1-r_{ij}^2)\tilde{\sigma}^2_{(k)}\}+n-q\right.\\&\qquad-\frac{1}{|\mathbf{R}_{ijk}|}\{(n-q)\{(1-r_{ij}^2)+(1-r_{ik}^2)+(1-r_{jk}^2)\}\\&\qquad\left.+2(r_{ik}r_{jk}-r_{ij})t_it_j+2(r_{ij}r_{jk}-r_{ik})t_it_k+2(r_{ij}r_{ik}-r_{jk})t_jt_k\}\right]\\&=\frac{n-q-1}{n-q-3}\cdot\frac{(1-r_{jk}^2)\tilde{\sigma}^2_{(i)}+(1-r_{ik}^2)\tilde{\sigma}^2_{(j)}+(1-r_{ij}^2)\tilde{\sigma}^2_{(k)}}{|\mathbf{R}_{ijk}|}-\frac{2\tilde{\sigma}^2}{n-q-3}\times\end{aligned}$$

$$\times \frac{(n-q)(1-r_{ij}r_{ik}r_{jk})+(r_{ik}r_{jk}-r_{ij})t_it_j+(r_{ij}r_{jk}-r_{ik})t_it_k+(r_{ij}r_{ik}-r_{jk})t_jt_k}{|\mathbf{R}_{ijk}|} \tag{3.8}$$

と表現できる。(3.8)式の第 2 表現において，第 1 項は $m=3$ の場合を構成する第 $i$ 番目の観測値，第 $j$ 番目の観測値それに第 $k$ 番目の観測値を除去したときのそれぞれの誤差分散の推定量を基にして表すことができ，第 2 項は第 $i$ 番目の観測値，第 $j$ 番目の観測値それに第 $k$ 番目の観測値の組合せ効果とみなすことができる。つまり，$m=2$ と同様の拡張表現ができることとなる。

加えて，(3.8)式において，$r_{ij}=r_{ik}=r_{jk}=0$ とすれば

$$\tilde{\sigma}^{2*}_{(ijk)} = \frac{n-q-1}{n-q-3}\{\tilde{\sigma}^2_{(i)}+\tilde{\sigma}^2_{(j)}+\tilde{\sigma}^2_{(k)}\}-\frac{2(n-q)}{n-q-3}\tilde{\sigma}^2 \tag{3.9}$$

と簡略化される。さらに加えて，(3.9)式において，$\tilde{\sigma}^2_{(i)}=\tilde{\sigma}^2_{(j)}=\tilde{\sigma}^2_{(k)}=\tilde{\sigma}^2$ とすれば，$\tilde{\sigma}^{2**}_{(ijk)}=\tilde{\sigma}^2$ となり，$\sigma^2$ の不偏推定量と一致する。

### 3.3　誤差分散の不偏推定量の一般化表現

前節の（3.5)式および（3.8)式を $m\geq 4$ の場合へと拡張できるように一般化する（通常では，$m$ が観測値の総数 $n$ の半分，つまり $n/2$ を超えることはない)。$m$ 個の観測値集合の場合である $I=\{i_1, i_2, \cdots, i_m\}$ $(i_1<i_2<\cdots<i_m)$ について，(3.2)式および（3.3)式の変形から

$$\begin{aligned}
\tilde{\sigma}^2_{(I)} &= \frac{\tilde{\sigma}^2}{n-q-m}(n-q-\boldsymbol{t}_I'\mathbf{R}_I^{-1}\boldsymbol{t}_I) \\
&= \frac{\tilde{\sigma}^2}{n-q-m}\left[\frac{n-q-1}{\tilde{\sigma}^2|\mathbf{R}_I|}\{a_{11}\tilde{\sigma}^2_{(i_1)}+a_{22}\tilde{\sigma}^2_{(i_2)}+\cdots+a_{mm}\tilde{\sigma}^2_{(i_m)}\}\right. \\
&\qquad -\frac{2}{|\mathbf{R}_I|}(a_{12}t_{i_1}t_{i_2}+a_{13}t_{i_1}t_{i_3}+\cdots+a_{m-1,m}t_{i_{m-1}}t_{i_m}) \\
&\qquad \left.+\frac{n-q}{|\mathbf{R}_I|}(|\mathbf{R}_I|-a_{11}-a_{22}-\cdots-a_{mm})\right] \\
&= \frac{\tilde{\sigma}^2}{n-q-m}\left[\frac{n-q-1}{\tilde{\sigma}^2|\mathbf{R}_I|}\sum_{j=1}^{m}a_{jj}\tilde{\sigma}^2_{(i_j)}-\frac{2}{|\mathbf{R}_I|}\sum_{j=1}^{m-1}\sum_{j<k}^{m}a_{jk}t_{i_j}t_{i_k}+\frac{n-q}{|\mathbf{R}_I|}\left(|\mathbf{R}_I|-\sum_{j=1}^{m}a_{jj}\right)\right] \\
&= \frac{n-q-1}{(n-q-m)|\mathbf{R}_I|}\sum_{j=1}^{m}a_{jj}\tilde{\sigma}^2_{(i_j)} \\
&\qquad -\frac{\tilde{\sigma}^2}{(n-q-m)|\mathbf{R}_I|}\left\{(n-q)\left(\sum_{j=1}^{m}a_{jj}-|\mathbf{R}_I|\right)+2\sum_{j=1}^{m-1}\sum_{j<k}^{m}a_{jk}t_{i_j}t_{i_k}\right\}
\end{aligned} \tag{3.10}$$

となる（詳細な式変形については付録を参照)。ただし，$\mathbf{R}_I$ の逆行列の成分については，表現を簡略化するために

$$\mathbf{R}_I^{-1} = \frac{1}{|\mathbf{R}_I|}\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{12} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{mm} \end{pmatrix}$$

とする。

参考までに，$m=4$ の場合である $I=\{i,j,k,\ell\}=\{i_1,i_2,i_3,i_4\}(i_1<i_2<i_3<i_4)$ について，(3.10)式を具体的な構成成分で表しておく。まず，$\mathbf{R}_I^{-1}$ の各成分を具体的に表示すると以下のようになる。行列式は

$$\begin{aligned}|\mathbf{R}_I| = 1&-2r_{i_1i_2}r_{i_1i_3}r_{i_2i_4}r_{i_3i_4}-2r_{i_1i_2}r_{i_1i_4}r_{i_2i_3}r_{i_3i_4}-2r_{i_1i_3}r_{i_1i_4}r_{i_2i_3}r_{i_2i_4}\\ &+2r_{i_1i_2}r_{i_1i_3}r_{i_2i_3}+2r_{i_1i_2}r_{i_1i_4}r_{i_2i_4}+2r_{i_1i_3}r_{i_1i_4}r_{i_3i_4}+2r_{i_2i_3}r_{i_2i_4}r_{i_3i_4}\\ &-r_{i_1i_2}^2-r_{i_1i_3}^2-r_{i_1i_4}^2-r_{i_2i_3}^2-r_{i_2i_4}^2-r_{i_3i_4}^2+r_{i_1i_2}^2r_{i_3i_4}^2+r_{i_1i_3}^2r_{i_2i_4}^2+r_{i_1i_4}^2r_{i_2i_3}^2\end{aligned}$$

となり，行列式を除く逆行列の各成分は，対角成分が

$$\begin{aligned}a_{11} &= 1+2r_{i_2i_3}r_{i_2i_4}r_{i_3i_4}-(r_{i_2i_3}^2+r_{i_2i_4}^2+r_{i_3i_4}^2)\\ a_{22} &= 1+2r_{i_1i_3}r_{i_1i_4}r_{i_3i_4}-(r_{i_1i_3}^2+r_{i_1i_4}^2+r_{i_3i_4}^2)\\ a_{33} &= 1+2r_{i_1i_2}r_{i_1i_4}r_{i_2i_4}-(r_{i_1i_2}^2+r_{i_1i_4}^2+r_{i_2i_4}^2)\\ a_{44} &= 1+2r_{i_1i_2}r_{i_1i_3}r_{i_2i_3}-(r_{i_1i_2}^2+r_{i_1i_3}^2+r_{i_2i_3}^2)\end{aligned}$$

であり，非対角成分が

$$\begin{aligned}a_{12} &= -r_{i_1i_2}-r_{i_1i_3}r_{i_2i_4}r_{i_3i_4}-r_{i_1i_4}r_{i_2i_3}r_{i_3i_4}+r_{i_1i_3}r_{i_2i_3}+r_{i_1i_4}r_{i_2i_4}+r_{i_1i_2}r_{i_3i_4}^2\\ a_{13} &= -r_{i_1i_3}-r_{i_1i_2}r_{i_2i_4}r_{i_3i_4}-r_{i_1i_4}r_{i_2i_3}r_{i_2i_4}+r_{i_1i_2}r_{i_2i_3}+r_{i_1i_4}r_{i_3i_4}+r_{i_1i_3}r_{i_2i_4}^2\\ a_{14} &= -r_{i_1i_4}-r_{i_1i_2}r_{i_2i_3}r_{i_3i_4}-r_{i_1i_3}r_{i_2i_3}r_{i_2i_4}+r_{i_1i_2}r_{i_2i_4}+r_{i_1i_3}r_{i_3i_4}+r_{i_1i_4}r_{i_2i_3}^2\\ a_{23} &= -r_{i_2i_3}-r_{i_1i_2}r_{i_1i_4}r_{i_3i_4}-r_{i_1i_3}r_{i_1i_4}r_{i_2i_4}+r_{i_1i_2}r_{i_1i_3}+r_{i_2i_4}r_{i_3i_4}+r_{i_1i_4}^2r_{i_2i_3}\\ a_{24} &= -r_{i_2i_4}-r_{i_1i_2}r_{i_1i_3}r_{i_3i_4}-r_{i_1i_3}r_{i_1i_4}r_{i_2i_3}+r_{i_1i_2}r_{i_1i_4}+r_{i_2i_3}r_{i_3i_4}+r_{i_1i_3}^2r_{i_2i_4}\\ a_{34} &= -r_{i_3i_4}-r_{i_1i_2}r_{i_1i_3}r_{i_2i_4}-r_{i_1i_2}r_{i_1i_4}r_{i_2i_3}+r_{i_1i_3}r_{i_1i_4}+r_{i_2i_3}r_{i_2i_4}+r_{i_1i_2}^2r_{i_3i_4}\end{aligned}$$

である。

以上の簡略化した成分を利用すると，

$$\begin{aligned}\tilde{\sigma}_{(I)}^2 &= \tilde{\sigma}_{(i_1,i_2,i_3,i_4)}^2\\ &= \frac{\tilde{\sigma}^2}{n-q-4}\left[n-q-\frac{1}{|\mathbf{R}_I|}(t_{i_1}\ t_{i_2}\ t_{i_3}\ t_{i_4})\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{12} & a_{22} & a_{23} & a_{24} \\ a_{13} & a_{23} & a_{33} & a_{34} \\ a_{14} & a_{24} & a_{34} & a_{44} \end{pmatrix}\begin{pmatrix} t_{i_1} \\ t_{i_2} \\ t_{i_3} \\ t_{i_4} \end{pmatrix}\right]\\ &= \frac{\tilde{\sigma}^2}{n-q-4}\left[n-q-\frac{1}{|\mathbf{R}_I|}\{(a_{11}t_{i_1}^2+a_{22}t_{i_2}^2+a_{33}t_{i_3}^2+a_{44}t_{i_4}^2)\right.\\ &\qquad\left.+2(a_{12}t_{i_1}t_{i_2}+a_{13}t_{i_1}t_{i_3}+a_{14}t_{i_1}t_{i_4}+a_{23}t_{i_2}t_{i_3}+a_{24}t_{i_2}t_{i_4}+a_{34}t_{i_3}t_{i_4})\}\right]\end{aligned} \quad (3.11)$$

と具体的に表すことができる。ここで，$m=2$ および $m=3$ の場合と同じく（3.2)式を変形して $t_{i_1}^2$，$t_{i_2}^2$，$t_{i_3}^2$ それに $t_{i_4}^2$ のみを置き換えると，(3.11)式は

$$\begin{aligned}\hat{\sigma}^2_{(i_1i_2i_3i_4)} = \frac{\hat{\sigma}^2}{n-q-4}\Bigg[ & \frac{n-q-1}{\hat{\sigma}^2|\mathbf{R}_I|}\{a_{11}\hat{\sigma}^2_{(i_1)}+a_{22}\hat{\sigma}^2_{(i_2)}+a_{33}\hat{\sigma}^2_{(i_3)}+a_{44}\hat{\sigma}^2_{(i_4)}\} \\ & -\frac{2}{|\mathbf{R}_I|}(a_{12}t_{i_1}t_{i_2}+a_{13}t_{i_1}t_{i_3}+a_{14}t_{i_1}t_{i_4}+a_{23}t_{i_2}t_{i_3}+a_{24}t_{i_2}t_{i_4}+a_{34}t_{i_3}t_{i_4}) \\ & +\frac{n-q}{|\mathbf{R}_I|}(|\mathbf{R}_I|-a_{11}-a_{22}-a_{33}-a_{44})\Bigg] \\ = \frac{n-q-1}{(n-q-4)|\mathbf{R}_I|} & \{a_{11}\hat{\sigma}^2_{(i_1)}+a_{22}\hat{\sigma}^2_{(i_2)}+a_{33}\hat{\sigma}^2_{(i_3)}+a_{44}\hat{\sigma}^2_{(i_4)}\} \\ -\frac{\hat{\sigma}^2}{(n-q-4)|\mathbf{R}_I|} & \{(n-q)(a_{11}+a_{22}+a_{33}+a_{44}-|\mathbf{R}_I|) \\ & +2(a_{12}t_{i_1}t_{i_2}+a_{13}t_{i_1}t_{i_3}+a_{14}t_{i_1}t_{i_4}+a_{23}t_{i_2}t_{i_3}+a_{24}t_{i_2}t_{i_4}+a_{34}t_{i_3}t_{i_4})\}\end{aligned} \tag{3.12}$$

と表現できる。(3.12)式の第 2 表現において，第 1 項は $m=4$ の場合を構成する第 $i_1$ 番目（第 $i$ 番目）の観測値，第 $i_2$ 番目（第 $j$ 番目）の観測値，第 $i_3$ 番目（第 $k$ 番目）の観測値それに第 $i_4$ 番目（第 $\ell$ 番目）の観測値を除去したときのそれぞれの誤差分散の推定量を基にして表すことができ，第 2 項は第 $i_1$ 番目（第 $i$ 番目）の観測値，第 $i_2$ 番目（第 $j$ 番目）の観測値，第 $i_3$ 番目（第 $k$ 番目）の観測値それに第 $i_4$ 番目（第 $\ell$ 番目）の観測値の組合せ効果とみなすことができる。

加えて，(3.12)式において，$r_{i_1i_2}=r_{i_1i_3}=r_{i_1i_4}=r_{i_2i_3}=r_{i_2i_4}=r_{i_3i_4}=0$ とすれば

$$\hat{\sigma}^{2*}_{(i_1i_2i_3i_4)} = \frac{n-q-1}{n-q-4}\{\hat{\sigma}^2_{(i_1)}+\hat{\sigma}^2_{(i_2)}+\hat{\sigma}^2_{(i_3)}+\hat{\sigma}^2_{(i_4)}\}-\frac{3(n-q)}{n-q-4}\hat{\sigma}^2 \tag{3.13}$$

と簡略化される。さらに加えて，(3.13)式において，$\hat{\sigma}^2_{(i_1)}=\hat{\sigma}^2_{(i_2)}=\hat{\sigma}^2_{(i_3)}=\hat{\sigma}^2_{(i_4)}=\hat{\sigma}^2$ とすれば，$\hat{\sigma}^{2**}_{(i_1i_2i_3i_4)}=\hat{\sigma}^2$ となり，$\sigma^2$ の不偏推定量と一致する。

## 4　実データへの適用

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[3]に掲げられている「配達時間データ（Delivery Time Data)」を取り上げる（データ分析の事例については竹内[7] などを参照)。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

表 4.1　配達時間データの分析結果：$m=1$ の場合

| No. | $t_i$ | $h_{ii}$ | $\hat{\sigma}^2_{(i)}$ |
|---|---|---|---|
| 1 | −1.628 | 0.102 | 9.790 |
| 2 | 0.365 | 0.071 | 11.063 |
| 3 | −0.016 | 0.099 | 11.130 |
| 4 | 1.580 | 0.085 | 9.868 |
| 5 | −0.142 | 0.075 | 11.120 |
| 6 | −0.091 | 0.043 | 11.126 |
| 7 | 0.270 | 0.082 | 11.093 |
| 8 | 0.367 | 0.064 | 11.062 |
| 9 | 3.214 | 0.498 | 5.905 |
| 10 | 0.813 | 0.196 | 10.795 |
| 11 | 0.718 | 0.086 | 10.869 |
| 12 | −0.193 | 0.114 | 11.111 |
| 13 | 0.325 | 0.061 | 11.077 |
| 14 | 0.341 | 0.078 | 11.071 |
| 15 | 0.210 | 0.041 | 11.108 |
| 16 | −0.223 | 0.166 | 11.105 |
| 17 | 0.138 | 0.059 | 11.120 |
| 18 | 1.113 | 0.096 | 10.503 |
| 19 | 0.579 | 0.096 | 10.961 |
| 20 | −1.874 | 0.102 | 9.354 |
| 21 | −0.878 | 0.165 | 10.740 |
| 22 | −1.450 | 0.392 | 10.066 |
| 23 | −1.444 | 0.041 | 10.076 |
| 24 | −1.496 | 0.121 | 9.998 |
| 25 | −0.068 | 0.067 | 11.128 |

$\hat{\sigma}^2=10.624$

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。つまり，説明変数の個数は $q=2+1$(定数項)$=3$ となる。

$n=25$ 個の全観測値について，$m=1$ の場合である個々の観測値に関する影響力評価をまとめた結果が表 4.1 である。基本的な影響力を評価する指標である標準化残差 $t_i$ およびてこ比 $h_{ii}$ を，全観測値 $(i=1, 2, \cdots, 25)$ について補足的に加えている。

つぎに，複数個の観測値集合について $m=2$ の場合は，観測値集合が全部で 300 組になるので，3 分割して表 4.2〜表 4.4 に誤差分散の推定量 $\hat{\sigma}^2_{(ij)}$（3 つの各表について，第 $(i, j)$ 成分，つまり第 $i$ 行と第 $j$ 列が交差した部分の数値）を示した。最小値は観測値集合

表 4.2　配達時間データの分析結果：$m=2$ の場合-(1)

| No. | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 10.243 | 10.262 | 8.979 | 10.268 | 10.265 | 10.259 | 10.213 | 5.286 | 10.115 |
| 2 | | 11.616 | 10.246 | 11.608 | 11.613 | 11.568 | 11.538 | 6.196 | 11.234 |
| 3 | | | 10.360 | 11.676 | 11.682 | 11.648 | 11.615 | 6.190 | 11.330 |
| 4 | | | | 10.361 | 10.361 | 10.276 | 10.234 | 5.473 | 10.017 |
| 5 | | | | | 11.671 | 11.640 | 11.608 | 6.118 | 11.322 |
| 6 | | | | | | 11.645 | 11.612 | 6.197 | 11.334 |
| 7 | | | | | | | 11.569 | 6.182 | 11.282 |
| 8 | | | | | | | | 6.172 | 11.267 |
| 9 | | | | | | | | | 5.508 |

表 4.3　配達時間データの分析結果：$m=2$ の場合-(2)

| No. | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10.042 | 10.268 | 10.252 | 10.261 | 10.271 | 10.167 | 10.273 | 9.592 | 10.111 | 8.272 |
| 2 | 11.341 | 11.598 | 11.550 | 11.545 | 11.589 | 11.594 | 11.603 | 10.938 | 11.421 | 9.751 |
| 3 | 11.413 | 11.667 | 11.630 | 11.625 | 11.663 | 11.659 | 11.676 | 11.029 | 11.509 | 9.822 |
| 4 | 10.073 | 10.359 | 10.268 | 10.284 | 10.323 | 10.325 | 10.326 | 9.511 | 10.070 | 8.491 |
| 5 | 11.405 | 11.653 | 11.622 | 11.615 | 11.654 | 11.650 | 11.667 | 11.028 | 11.504 | 9.807 |
| 6 | 11.410 | 11.661 | 11.627 | 11.622 | 11.659 | 11.655 | 11.673 | 11.028 | 11.507 | 9.811 |
| 7 | 11.374 | 11.630 | 11.584 | 11.580 | 11.622 | 11.623 | 11.635 | 10.966 | 11.453 | 9.776 |
| 8 | 11.332 | 11.601 | 11.553 | 11.551 | 11.589 | 11.587 | 11.601 | 10.916 | 11.418 | 9.766 |
| 9 | 4.841 | 6.177 | 6.190 | 6.070 | 6.090 | 6.045 | 6.199 | 5.793 | 6.198 | 5.733 |
| 10 | 11.051 | 11.300 | 11.254 | 11.225 | 11.300 | 11.334 | 11.324 | 10.722 | 11.159 | 9.524 |
| 11 | | 11.400 | 11.354 | 11.344 | 11.381 | 11.395 | 11.400 | 10.726 | 11.233 | 9.673 |
| 12 | | | 11.612 | 11.604 | 11.645 | 11.643 | 11.659 | 11.026 | 11.498 | 9.783 |
| 13 | | | | 11.561 | 11.604 | 11.608 | 11.618 | 10.955 | 11.439 | 9.771 |
| 14 | | | | | 11.597 | 11.607 | 11.613 | 10.965 | 11.441 | 9.779 |
| 15 | | | | | | 11.639 | 11.652 | 10.996 | 11.480 | 9.814 |
| 16 | | | | | | | 11.650 | 10.987 | 11.478 | 9.752 |
| 17 | | | | | | | | 11.002 | 11.491 | 9.816 |
| 18 | | | | | | | | | 10.776 | 9.205 |
| 19 | | | | | | | | | | 9.632 |

$I=\{9, 11\}$ のときに $\tilde{\sigma}^2_{(9,\,11)}=4.841$ であり，最大値は観測値集合 $I=\{3, 25\}$ のときに $\tilde{\sigma}^2_{(3,\,25)}=11.684$ となる。

なお，$m=3$ の場合は，観測値集合が全部で 2,300 組になるため，誤差分散の推定量が小さいもの（5 より小さい観測値集合の 18 組）だけを表 4.5 に示しておく（参考までに，最大値は観測値集合 $I=\{1, 6, 25\}$ のときに $\tilde{\sigma}^2_{(I)}=12.294$ となる）。

表 4.4　配達時間データの分析結果：$m=2$ の場合-(3)

| No. | 21 | 22 | 23 | 24 | 25 |
|---|---|---|---|---|---|
| 1 | 9.948 | 9.349 | 9.036 | 8.725 | 10.272 |
| 2 | 11.214 | 10.447 | 10.530 | 10.458 | 11.615 |
| 3 | 11.277 | 10.553 | 10.576 | 10.482 | 11.684 |
| 4 | 10.088 | 9.294 | 9.337 | 9.167 | 10.359 |
| 5 | 11.247 | 10.540 | 10.560 | 10.488 | 11.673 |
| 6 | 11.268 | 10.562 | 10.567 | 10.485 | 11.679 |
| 7 | 11.251 | 10.497 | 10.555 | 10.473 | 11.647 |
| 8 | 11.236 | 10.534 | 10.527 | 10.426 | 11.614 |
| 9 | 5.671 | 6.163 | 5.434 | 5.622 | 6.117 |
| 10 | 10.823 | 9.922 | 10.298 | 10.334 | 11.334 |
| 11 | 11.039 | 10.470 | 10.352 | 10.265 | 11.411 |
| 12 | 11.217 | 10.449 | 10.548 | 10.489 | 11.663 |
| 13 | 11.229 | 10.482 | 10.542 | 10.467 | 11.629 |
| 14 | 11.206 | 10.463 | 10.540 | 10.479 | 11.623 |
| 15 | 11.260 | 10.558 | 10.568 | 10.489 | 11.661 |
| 16 | 11.267 | 10.561 | 10.529 | 10.379 | 11.658 |
| 17 | 11.275 | 10.567 | 10.576 | 10.490 | 11.675 |
| 18 | 10.744 | 10.083 | 9.977 | 9.790 | 11.029 |
| 19 | 11.150 | 10.394 | 10.431 | 10.317 | 11.508 |
| 20 | 9.342 | 7.955 | 8.574 | 8.467 | 9.819 |
| 21 | | 9.745 | 10.126 | 10.179 | 11.268 |
| 22 | | | 9.361 | 9.546 | 10.568 |
| 23 | | | | 9.259 | 10.571 |
| 24 | | | | | 10.492 |

表 4.1 から $m=1$ の場合について，個々の観測値の影響力を評価すると，$\hat{\sigma}^2_{(i)}$ と $\hat{\sigma}^2$ の差から，明らかに観測値 No. 9 の影響力が大きいことがわかる。これ以外に差（の絶対値）が 1.000 を超えるのは観測値 No. 20 だけである。

$m=2$ の場合について，表 4.2～表 4.4 の誤差分散の推定量を算出した (3.5)式と簡略化した (3.6)式の差を計算すると概ね 0.100 程度（300 組中で 0.100 未満が 265 組［88.3%］）となり，300 組中で 0.500 以上となる組合せは 6 組（300 組中の 2.0%）だけである。つまり，残差の相関係数が小さい観測値の組合せが多いものといえる。ただし，$m=1$ の場合において影響力が大きい観測値と判定された観測値 No. 9 との組合せの中に差が大きいものが存在し，差が 1.000 を超えるケースがつぎの 3 組（300 組中の 1.0%）である。観測値集合 $I=\{9, 11\}$ の場合は差が 1.086，$I=\{9, 20\}$ の場合は差が 1.398（最大差），それに $I=\{9, 22\}$ の場合は差が 1.080 というものである。これらの観測値集合に関する残差の相関係数は，観測値集合 $I=\{9, 11\}$ の場合は $r_{ij}=-0.258$，$I=\{9, 20\}$ の場合は $-0.305$，それに $I=\{9, 22\}$

表 4.5　配達時間データの分析結果：$m=3$ の場合（一部抜粋）

| $I$ | $\bar{\sigma}^2_{(I)}$ | $I$ | $\bar{\sigma}^2_{(I)}$ | $I$ | $\bar{\sigma}^2_{(I)}$ |
|---|---|---|---|---|---|
| {1, 4, 9} | 4.792 | {4, 9, 11} | 4.378 | {9, 10, 11} | 4.271 |
| {1, 9, 11} | 4.263 | {4, 9, 23} | 4.994 | {9, 11, 14} | 4.923 |
| {1, 9, 20} | 4.958 | | | {9, 11, 15} | 4.919 |
| {1, 9, 23} | 4.645 | | | {9, 11, 16} | 4.790 |
| {1, 9, 24} | 4.705 | | | {9, 11, 18} | 4.644 |
| | | | | {9, 11, 20} | 4.869 |
| | | | | {9, 11, 21} | 4.620 |
| | | | | {9, 11, 22} | 4.533 |
| | | | | {9, 11, 23} | 4.422 |
| | | | | {9, 11, 24} | 4.622 |
| | | | | {9, 11, 25} | 4.997 |

の場合は −0.521 であり，残差の相関係数の絶対値が大きい上位の 3 組でもある（残差の相関係数の絶対値 $|r_{ij}|$ が 0.200 を超えているのは 300 組中で 8 組［2.7%］だけであり，観測値 No. 9 や No. 22 を含む観測値集合が多い）。

マスク効果が小さい，あるいは無視できる程度の場合であれば，簡略化した（3. 6）式を利用して観測値集合の影響力評価を行っても大きな問題はない。しかしながら，残差の相関係数の絶対値が大きいためにマスク効果も大きくなるものと想定される観測値集合 $I=\{20, 22\}$ の場合（$r_{ij}=-0.205$ の絶対値は 7 番目に大きい値）は差が 0.750 となり，観測値 No. 9 との組合せの上記 3 つのケースの次に大きな差となっている。

$m=3$ の場合についても，表 4. 5 の誤差分散の推定量を算出した（3. 8）式と簡略化した（3. 9）式の差を計算すると，2,300 組中で 0.500 以上となる組合せは 148 組（2,300 組の 6.4%），1.000 以上となる組合せは 64 組（同 2.8%），それに 1.500 以上となる組合せは 17 組（同 0.7%）だけである。差が 2.000 以上となる組合せは，観測値集合 $I=\{4, 9, 20\}$ の場合に差が 2.045 および $I=\{9, 20, 22\}$ の場合に差が 2.630（最大差）というものだけであるが，表 4. 5 の中の観測値集合の組合せにはない。けれども，$m=2$ の場合において，差が大きい観測値集合の中にある組合せに近いということもわかる。$m\geq 3$ の場合は，残差の相関係数についての組合せ効果も追加的に想定する必要があり，マスク効果自体も複雑な関係になることが予想される（が，十分に解明はされていない）。

したがって，今回のデータ分析からは，残差の相関係数の絶対値が比較的小さくマスク効果も小さいものとみなせる場合であれば，提案する誤差分散の推定量の簡略化された表現で十分に観測値集合の影響力評価が可能となる。しかしながら，残差の相関係数の絶対値が比較的大きく，これが主要因でマスク効果も大きくなるものと想定される場合は，正確な誤差分散の推定量に基づく評価式を利用した方がよいといえる。ただ，データ数 $n$ が大きくな

ると観測値集合の組合せも数多くなるが，残差の相関係数（の絶対値）は全体的に小さくなる傾向がある（てこ比の性質から全観測値のてこ比の合計は一定であるため，各観測値集合について残差の相関係数の分母を構成する $(1-h_{ii})(1-h_{jj})$ の値が大きくなるためである）ので，この場合でも，該当する観測値集合の組合せの数が多くなりにくいため，実際のデータ分析において大きな支障はないものと考えられる。

## 5　まとめと今後の課題

誤差分散の推定量が，これまで形式的な取り扱い方に留まることが多く，詳細な利用法が十分に検討されてこなかった。本論文では，この問題点を解決するために，回帰診断における診断統計量の基本的な構成要素の一つである誤差分散の推定量について，複数個の観測値集合の場合における具体的な計算表現を提案した。提案された新表現は，個々の観測値についての誤差分散の推定量を基にして，複数個の観測値集合の場合についての表現を可能とした。また，個々の観測値の成分と複数個の観測値間における組合せ効果（相乗効果）の２つの項に分離することが可能となった。この結果，複数個の観測値集合を構成する個々の観測値の効果と複数個の観測値固有の組合せ効果に分離することができ，この表現から，さらに簡略化した近似的な表現を導出することも可能となり，データ分析における実用面からも有用といえる。

しかしながら，複数個の観測値集合における問題点としてマスク効果が存在することを無視することはできない。このマスク効果もある程度考慮して評価できるような表現や新たな定式化が求められる。この点については，複数個の観測値の影響力評価への更なる拡張を含め今後の検討課題としたい。

## 付録：(3. 10)式の一般化表現の導出

(3. 10)式の一般化表現について，その導出過程を詳細に示す。

$$\begin{aligned}\hat{\sigma}^2_{(I)} &= \frac{\hat{\sigma}^2}{n-q-m}(n-q-\boldsymbol{t}_I'\mathbf{R}_I^{-1}\boldsymbol{t}_I)\\ &= \frac{\hat{\sigma}^2}{n-q-m}\Bigg[\frac{n-q-1}{\hat{\sigma}^2|\mathbf{R}_I|}(a_{11}\hat{\sigma}^2_{(i_1)}+a_{22}\hat{\sigma}^2_{(i_2)}+\cdots+a_{mm}\hat{\sigma}^2_{(i_m)})\\ &\qquad -\frac{2}{|\mathbf{R}_I|}(a_{12}t_{i_1}t_{i_2}+a_{13}t_{i_1}t_{i_3}+\cdots+a_{m-1,m}t_{i_{m-1}}t_{i_m})\\ &\qquad +\frac{n-q}{|\mathbf{R}_I|}(|\mathbf{R}_I|-a_{11}-a_{22}-\cdots-a_{mm})\Bigg]\end{aligned}$$

$$= \frac{\hat{\sigma}^2}{n-q-m}\left[\frac{n-q-1}{\hat{\sigma}^2|\mathbf{R}_I|}\sum_{j=1}^{m}a_{jj}\hat{\sigma}^2_{(i_j)}-\frac{2}{|\mathbf{R}_I|}\sum_{j=1}^{m-1}\sum_{j<k}^{m}a_{jk}t_{i_j}t_{i_k}+\frac{n-q}{|\mathbf{R}_I|}\left(|\mathbf{R}_I|-\sum_{j=1}^{m}a_{jj}\right)\right]$$

$$= \frac{\hat{\sigma}^2}{n-q-m}\left[\frac{n-q-1}{\hat{\sigma}^2|\mathbf{R}_I|}\sum_{j=1}^{m}a_{jj}\hat{\sigma}^2_{(i_j)}-\frac{2}{|\mathbf{R}_I|}\sum_{j=1}^{m-1}\sum_{j<k}^{m}a_{jk}t_{i_j}t_{i_k}-\frac{n-q}{|\mathbf{R}_I|}\left(\sum_{j=1}^{m}a_{jj}-|\mathbf{R}_I|\right)\right]$$

$$= \frac{\hat{\sigma}^2}{n-q-m}\left[\frac{n-q-1}{\hat{\sigma}^2|\mathbf{R}_I|}\sum_{j=1}^{m}a_{jj}\hat{\sigma}^2_{(i_j)}-\frac{1}{|\mathbf{R}_I|}\left((n-q)\left\{\sum_{j=1}^{m}a_{jj}-|\mathbf{R}_I|\right\}+2\sum_{j=1}^{m-1}\sum_{j<k}^{m}a_{jk}t_{i_j}t_{i_k}\right)\right]$$

$$= \frac{n-q-1}{n-q-m}\frac{1}{|\mathbf{R}_I|}\sum_{j=1}^{m}a_{jj}\hat{\sigma}^2_{(i_j)}-\frac{\hat{\sigma}^2}{n-q-m}\left[\frac{1}{|\mathbf{R}_I|}\left((n-q)\left\{\sum_{j=1}^{m}a_{jj}-|\mathbf{R}_I|\right\}+2\sum_{j=1}^{m-1}\sum_{j<k}^{m}a_{jk}t_{i_j}t_{i_k}\right)\right]$$

$$= \frac{n-q-1}{(n-q-m)|\mathbf{R}_I|}\sum_{j=1}^{m}a_{jj}\hat{\sigma}^2_{(i_j)}-\frac{\hat{\sigma}^2}{(n-q-m)|\mathbf{R}_I|}\left((n-q)\left\{\sum_{j=1}^{m}a_{jj}-|\mathbf{R}_I|\right\}+2\sum_{j=1}^{m-1}\sum_{j<k}^{m}a_{jk}t_{i_j}t_{i_k}\right)$$

**謝辞**：本研究は 2022 年度東京経済大学国内研究員制度の研究成果の一部である。

## 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 Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.

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

［4］ Nurunnabi, A. A. M. , Hadi, A. S. and Imon, A. H. M. R. (2014), Procedures for the identification of multiple influential observations in linear regression, *Journal of Applied Statistics*, **41**, 1315–1331.

［5］ Nurunnabi, A. A. M. , Nasser, M. and Imon, A. H. M. R. (2016), Identification and classification of multiple outliers, high leverage points and influential observations in linear regression, *Journal of Applied Statistics*, **43**, 509–525.

［6］ 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.

［7］ 竹内秀一（2020），説明変数空間における観測値の影響力評価，人文自然科学論集，**146** 号，3–13.

［8］ 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.

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