論　文

線形回帰分析における部分影響力評価

竹　内　秀　一

Assessment of Partial Influence in Linear Regression

Hidekazu TAKEUCHI

There are two procedures to assess the influence of observations for explanatory variables in linear regression. One is based on the case deletion procedure in variable selection problems, and the other on selecting some variables before the case deletion. The latter is called the partial influence procedure. This paper gives a new expression of the partial influence measure proposed by Cook and Weisberg [5] to assess the influence of observations for the selected variables. The new expression of the partial influence measure consists of Cook's distance and a similar influence measure. Furthermore a cut-off point for the new expression is derived by using that for Cook's distance. For the single case with one observation and one variable deleted, the cut-off point for the new expression is also compared with the size-adjusted cut-off point proposed by Belsley, Kuh and Welsch [2] and Chatterjee and Hadi [4].

1　はじめに

線形回帰分析においては，説明変数に対する観測値の影響力評価（assessment of influence）を検討する場合に大きく分けて二つの方法がある。一つは説明変数の選択問題（以下では「変数選択問題（variable selection problems)」とする）として観測値の影響力評価を考える場合であり，影響力を調べたい観測値集合を除去してから変数選択された説明変数集合に関する影響力を評価する方法である。もう一つは，手順としてはこの逆になるが，説明変数集合を選択（選定）してから影響力を調べたい観測値集合を除去して影響力の評価をする方法（以下では「部分影響力（partial influence）評価」とする）である。この二つの方法は，除去される対象の順序が異なるだけであるが，影響力評価の立場がまったく異質で

あるので，一般に影響力評価の結果も異なる。

前者の影響力評価方法については，Léger and Altman［7］や Takeuchi［10］あるいは竹内［11］などの研究例があるが，後者についてはあまり研究されていない。もちろん，Belsley, Kuh and Welsch［2］や Cook and Weisberg［6］などの影響力評価（または回帰診断）に関する代表的な著書においては紹介されているが，研究論文としては Cook and Weisberg［5］や Chatterjee and Hadi［4］くらいしか見受けられない。最近では，Castillo, Hadi, Conejo and Fernández-Canteli［3］の論文中で提案されている新規の方法の比較対象として取り上げられているが，影響力評価方法として踏み込んだ議論はされていない。

そこで本論文では，従来から提案されている部分影響力を評価するための診断統計量（influence measure）を再検討し，新たな表現を導入する。また，この新表現に対してデータ数に基づいて調整された（size-adjusted）打切り点（たとえば，竹内［12］を参照）を提案する。さらに，通常，一つの説明変数を除去し，そのつぎに一つの観測値を除去することにより部分影響力を評価するための診断統計量を算出しているが，これを一般化して複数の説明変数集合および複数個の観測値集合を除去する場合へと拡張することについても検討をする。

本論文の構成は以下のとおりである。2 節では，線形回帰モデルおよび観測値の影響力評価で用いられる各種の定義を与える。3 節において，従来から提案されている部分影響力を評価するための診断統計量を与える。4 節では，その診断統計量の新表現を導入する。5 節では，提案する新表現に基づく打切り点を導出し，従来の打切り点との比較をする。最後の 6 節は全体のまとめである。

2　定義

ここでは，線形回帰モデルとして，

$$y = \mathrm{X}\beta + \varepsilon$$

を考える。このとき，y は $n\times 1$ の目的変数ベクトル，X は $n\times q$ のフルランクの説明変数行列，β は $q\times 1$ 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 $N(0, \sigma^2\mathrm{I}_n)$ に従うものとする。ただし，I_n は n 次の単位行列を表す。また，β の最小 2 乗推定量は $\hat{\beta}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y$ として得られ，σ^2 の不偏推定量は $\hat{\sigma}^2=e'e/(n-q)$ となる。ただし，「′」は行列あるいはベクトルの転置を表し，e は残差ベクトルであり，$e=y-\mathrm{X}\hat{\beta}=(\mathrm{I}_n-\mathrm{H})y$ である。このとき，H は説明変数行列から構成されるハット行列（hat matrix）$\mathrm{H}=\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ であり，その第 i 対角成分 h_{ii} がてこ比である。ただし，$1/n\leq h_{ii}<1$ 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 $(\cdot)$ は n 個の観測値の中から除去される観測値番号または観測値集合を表し，$\hat{\sigma}^2$ および $\hat{\sigma}^2_{(i)}$ の関係式は，

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

である。よって，t_i および t_i^* の関係式は，

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

となる。

つぎに，観測値の影響力評価を行うときに必要な各種の定義を与える。Cook and Weisberg［6］から，観測値の影響力評価を行うための典型的な診断統計量である Cook の距離は，除去される m 個の観測値集合 $I=\{i_1, i_2, \cdots, i_m\}$ に対して，$y'=(y'_{(I)}\, y'_I)$，$\mathrm{X}'=(\mathrm{X}'_{(I)}\, \mathrm{X}'_I)$，それに $e'=(e'_{(I)}\, e'_I)$ と分割することにより

$$\mathrm{CD}_I = \frac{(\hat{\beta}-\hat{\beta}_{(I)})'\mathrm{X}'\mathrm{X}(\hat{\beta}-\hat{\beta}_{(I)})}{q\hat{\sigma}^2} = \frac{(\hat{y}-\hat{y}_{(I)})'(\hat{y}-\hat{y}_{(I)})}{q\hat{\sigma}^2} = \frac{e'_I(\mathrm{I}_m-\mathrm{H}_I)^{-1}\mathrm{H}_I(\mathrm{I}_m-\mathrm{H}_I)^{-1}e_I}{q\hat{\sigma}^2} \tag{2.2}$$

と定義される。ただし，

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

であり，$\mathrm{H}_I=\mathrm{X}_I(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'_I$ とする。

さらに，(2.2)式の代替表現として，Takeuchi［8］は以下のような表現を提案した。

$$\mathrm{CD}_I = c'_I\Pi_I^{-\frac{1}{2}}\mathrm{R}_I^{-1}(\Pi_I+\mathrm{I}_m-\mathrm{R}_I)\mathrm{R}_I^{-1}\Pi_I^{-\frac{1}{2}}c_I \tag{2.3}$$

ただし，

$$\Pi_I = [diag(\mathrm{H}_I)][diag(\mathrm{I}_m-\mathrm{H}_I)]^{-1}$$

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

$$c_I = \frac{1}{\sqrt{q}}\Pi_I^{-\frac{1}{2}}t_I$$

であり，

$$t_I = \frac{1}{\hat{\sigma}}[diag(\mathrm{I}_m-\mathrm{H}_I)]^{-\frac{1}{2}}e_I$$

である。ここで，$diag(\cdot)$ は $(\cdot)$ 内の正方行列の対角成分を取り出し，非対角成分をすべて0とする対角行列を表す。

特に，$I=\{i\}$ の場合，標準化残差およびてこ比を利用して(2.2)式は

$$\mathrm{CD}_i = \frac{t_i^2}{q}\cdot\frac{h_{ii}}{1-h_{ii}} \tag{2.4}$$

と簡略化して表現される。また，(2.3) 式の代替表現については，

$$\mathrm{CD}_i = c_i^2$$

と表現される。さらに，m 個の観測値に関する影響力の単純な和については，

$$\sum_{i\in I}\mathrm{CD}_i = c_I' c_I \tag{2.5}$$

と表記することもある。

3　部分影響力を評価するための診断統計量

部分影響力を評価するための診断統計量として，代表的な二つの指標の定義式を与える。どちらの定義も，説明変数集合 J（または第 j 番目の説明変数）を除去した後に，観測値集合 I（または第 i 番目の観測値）を除去することにより部分影響力を評価するという手法を採用している。これは結果的に，説明変数集合 J に対する観測値集合 I の部分影響力の評価指標に相当する。

3.1　Belsley, Kuh and Welsch の定義

Belsley, Kuh and Welsch［2］は第 j 番目の説明変数に対する第 i 番目の観測値の影響力を評価するために，以下のような診断統計量を提案した。それは，第 i 番目の観測値を除去したときに，第 j 番目の説明変数に対する回帰係数に生じる変化量を，その標準偏差（分散の平方根）で規準化した指標として

$$\mathrm{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{(i)j}}{\sqrt{\mathrm{var}(\hat{\beta}_j)}} = \begin{cases} \dfrac{t_i}{\sqrt{1-h_{ii}}}\cdot\dfrac{w_{ij}}{\sqrt{\mathrm{W}_j'\mathrm{W}_j}} \\ \dfrac{t_i^*}{\sqrt{1-h_{ii}}}\cdot\dfrac{w_{ij}}{\sqrt{\mathrm{W}_j'\mathrm{W}_j}} \end{cases} \tag{3.1}$$

と与えられる。ただし，$\mathrm{var}(\cdot)$ は $(\cdot)$ 内の分散を表し，$\mathrm{W}_j = (\mathrm{I} - \mathrm{H}_{[j]})\mathrm{X}_j$ である。ここで，

$$\mathrm{H}_{[j]} = \mathrm{X}_{[j]}(\mathrm{X}_{[j]}'\mathrm{X}_{[j]})^{-1}\mathrm{X}_{[j]}' \quad \text{および} \quad w_{ij} = x_{ij} - x_{i[j]}(\mathrm{X}_{[j]}'\mathrm{X}_{[j]})^{-1}\mathrm{X}_{[j]}'\mathrm{X}_j$$

である。このとき，添字の $[\cdot]$ は q 個の説明変数の中から除去される説明変数番号または説明変数集合を表す。

(3.1)式の定義において，第一表現が Belsley, Kuh and Welsch［2］の定義であり，簡略化された第二表現は Chatterjee and Hadi［4］の定義である。正確には，第二表現の第二式（下）は，平方した場合が定義されているが，定義式を統一的に扱うためにこのように再定義する。また，Belsley, Kuh and Welsch［2］では(3.1)式の定義において，回帰係数（ベクトルの第 j 成分）の添字は $\hat{\beta}_{(i)j}$ ではなく $\hat{\beta}_{j(i)}$ と表記している。けれども，本論文では，説明変数行列 X やハット行列 H などの添字と同様に，観測値番号（または観測値集合）のつぎに説明変数番号（または説明変数集合）の順に統一した。

(3.1)式の第二表現において，第一式（上）および第二式（下）の相違は，σ^2 の推定量とし

て，$\hat{\sigma}^2$ を適用するのか $\hat{\sigma}^2_{(i)}$ を適用するのかという点である。この相違が t_i と t_i^* の差異として現れている。t_i および t_i^* の関係式は(2.1)式のとおりであり，$|t_i| \leq 1$ 程度であれば両者に大きな違いはない（逆に，$|t_i|>1$ の場合は変動が大きくなりやすくなる）。一般に，t_i を利用した場合は保守的な結果になり，t_i^* を利用した場合は劇的に大きな変化をもたらす結果になることがある。このため，通常の部分影響力評価においては，解析結果が安定している第一式を利用することが多い。しかしながら，Belsley, Kuh and Welsch［2］は第二式を部分影響力の評価式として定義している。

3.2　Cook and Weisberg の定義

Cook and Weisberg［5］［6］は，Belsley, Kuh and Welsch［2］とは異なる観点から，一般化された部分影響力評価のための診断統計量を提案した。説明変数集合を特定する行列 L に基づき，観測値集合 I の影響力を調べるために，以下のような診断統計量（距離規準）を導入したのである。二つのベクトル $\hat{\Psi}=\mathrm{L}\hat{\beta}$ および $\hat{\Psi}_{(I)}=\mathrm{L}\hat{\beta}_{(I)}$，ただし，行列 L は大きさが $\ell \times q$ でランクは ℓ であるものと定義し，これらの間の距離を $\mathrm{D}_I(\hat{\Psi})$ とする。つまり，

$$\mathrm{D}_I(\hat{\Psi}) = \frac{(\hat{\Psi}-\hat{\Psi}_{(I)})'[\mathrm{L}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}']^{-1}(\hat{\Psi}-\hat{\Psi}_{(I)})}{\ell\hat{\sigma}^2} \tag{3.2}$$

とする。特に，$I=\{i\}$ および $\mathrm{L}=[0\cdots01]$，つまり $\ell=1$ の場合に，たとえば，第 j 列（ここでは，最後の q 列目とみなす）についての部分影響力の評価式は

$$\mathrm{D}_i(\hat{\Psi}) = \frac{(\hat{\beta}-\hat{\beta}_{(i)})'\mathrm{L}'[\mathrm{L}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}']^{-1}\mathrm{L}(\hat{\beta}-\hat{\beta}_{(i)})}{\hat{\sigma}^2} = \frac{t_i^2}{1-h_{ii}}(h_{ii}-h_{ii[j]}) \tag{3.3}$$

となる。(3.3)式の簡単な式変形により第 j 番目の説明変数に対する $\mathrm{D}_i(\hat{\Psi})$ が $\mathrm{DFBETAS}_{ij}$（の 2 乗，正確には第二表現の第一式の 2 乗）に一致することがわかる（付録 A を参照）。

4　診断統計量の新表現

前節の部分影響力を評価するための診断統計量との比較をするために，(3.2)式の $\mathrm{D}_I(\hat{\Psi})$ の表記を変えて D_U とし，全説明変数の数 q と関連付けて以下のように定義し直す。すなわち，ℓ 個の説明変数集合 $J=\{j_1, j_2\cdots, j_\ell\}$ に対する m 個の観測値集合 I の部分影響力を測定するための診断統計量は，$\mathrm{X}=(\mathrm{X}_{[J]}\,\mathrm{X}_J)$ とし，

$$\hat{\beta}_{(I)} = \begin{pmatrix} \hat{\beta}_{(I)[J]} \\ \hat{\beta}_{(I)J} \end{pmatrix}$$

と分割するとき

$$D_{IJ} \equiv \frac{(\hat{\beta}_J - \hat{\beta}_{(I)J})'[\mathrm{Var}(\hat{\beta}_J)]^{-1}(\hat{\beta}_J - \hat{\beta}_{(I)J})}{\ell}$$
$$= \frac{q}{\ell}\cdot\frac{(\hat{\beta}-\hat{\beta}_{(I)})'X'(H-H_{[J]})X(\hat{\beta}-\hat{\beta}_{(I)})}{q\hat{\sigma}^2} = \frac{q}{\ell}\cdot\frac{(\hat{y}-\hat{y}_{(I)})'(H-H_{[J]})(\hat{y}-\hat{y}_{(I)})}{q\hat{\sigma}^2} \quad (4.1)$$

と定義される（式変形は付録 B を参照）。ただし，Var(・)は (・) 内のベクトルの分散共分散行列を表し，$H_{[J]}=X_{[J]}(X'_{[J]}X_{[J]})^{-1}X'_{[J]}$ である。

(4.1)式は(2.2)式における Cook の距離を部分影響力評価のために拡張しているとも考えられる。ベクトルに挟まれた重み行列の形式は異なるが，よく似た表現になっている。この点を明確にするために，さらに，(4.1)式に対して別の式変形を試みる。ハット行列は $H=H_{[J]}+W_J(W'_JW_J)^{-1}W'_J$ と分解できる。ただし，$W_J=(I_n-H_{[J]})X_J$ である。これを利用すると(4.1)式の第二表現から，

$$D_{IJ} = \frac{q}{\ell}\cdot\frac{(\hat{\beta}-\hat{\beta}_{(I)})'X'W_J(W'_JW_J)^{-1}W'_JX(\hat{\beta}-\hat{\beta}_{(I)})}{q\hat{\sigma}^2}$$
$$= \frac{q}{\ell}\cdot\frac{(\hat{\beta}-\hat{\beta}_{(I)})'\begin{bmatrix}0 & 0\\ 0 & X'_J(I_n-H_{[J]})X_J\end{bmatrix}(\hat{\beta}-\hat{\beta}_{(I)})}{q\hat{\sigma}^2} \quad (4.2)$$
$$= \frac{q}{\ell}(CD_I - CD^*_{IJ})$$

と変形することができる。ただし，

$$CD^*_{IJ} \equiv \frac{e'_I(I_m-H_I)^{-1}H_{I[J]}(I_m-H_I)^{-1}e_I}{q\hat{\sigma}^2} \quad (4.3)$$

であり，

$$H_{I[J]} = X_{I[J]}(X'_{[J]}X_{[J]})^{-1}X'_{I[J]}$$

とする。このとき，

$$X_{[J]} = \begin{pmatrix} X_{(I)[J]} \\ X_{I[J]} \end{pmatrix}$$

と分割している。

この結果，D_{IJ} は典型的な 2 次形式であるから，$D_{IJ}\geq 0$ となる。つまり，(4.2) 式の第三表現から

$$CD_I \geq CD^*_{IJ} (\geq 0)$$

となることがわかり，(2.2)式の Cook の距離 CD_I の最小値は(4.3)式の CD^*_{IJ} のとる値によって決まることになる。従来の研究においては，Barret and Gray [1] が CD_I の上限について議論をしているが，この点を含めた研究については別の機会に検討する。

5　打切り点の比較

部分影響力を評価するための診断統計量である(3.2)式についての一般的な打切り点は提案されていない。けれども，特定の観測値一つを除去した場合の(3.1)式については，Belsley, Kuh and Welsch［2］や Chatterjee and Hadi［4］により打切り点が導出されている。(3.1)式の打切り点として，第 j 番目の説明変数に対する第 i 番目の観測値が

$$\mathrm{DFBETAS}_{ij}^2 = \mathrm{D}_i(\hat{\Psi}) > \frac{4}{n} \tag{5.1}$$

となる場合を影響力の大きい観測値と判定するものとして導かれている。(5.1)式の導出は，(3.1)式と本質的に同じ(3.3)式において，標準化残差が $|t_i|>2$（スチューデント化残差の場合も $|t_i^*|>2$），てこ比については $h_{ii}=1/n$ および $h_{ii}-h_{ii[j]}=1/n$，その上で，$n-1\cong n$ の近似が適用されている。てこ比への代入については，最初の「$h_{ii}=1/n$」はてこ比 h_{ii} の平均 q/n について $q=1$，つまり一つの説明変数を想定した場合であり，つぎの「$h_{ii}-h_{ii[j]}=1/n$」は $h_{ii}-h_{ii[j]}$ の n 個の観測値についての合計が 1 になることから，その平均を適用した場合であるので，導出経過はまったく異なるのである。このため，てこ比に関するこれら二つの打切り点の導出自体に問題点（不自然さ）があることは否定できない。

そこで，こうした問題点を改善するために，提案する新表現に基づく診断統計量について，まずこの新表現の特徴を利用した打切り点を一般的な場合について導出する。つぎに観測値と説明変数がそれぞれ一つずつ除去される場合の部分影響力評価において，従来の打切り点および新表現に基づく打切り点の具体的な比較検討をする。

(4.2)式の第三表現に着目し，Cook の距離 CD_I および CD_{ij}^* の打切り点をそれぞれ導出する。Cook の距離 CD_I については，いくつかの打切り点が提案されているが，ここでは，Takeuchi［9］における Welsch-Kuh の距離の打切り点導出と同様の方法を与える。その導出方法に従えば，(2.3)式において，$\mathrm{CD}_I\cong c_I'c_I$ と近似することになる。つまり，m 個の観測値集合 I を構成する観測値それぞれの単独の Cook の距離を単純に合計したものとして近似をするのである。すると，(2.4)式および(2.5)式から

$$\mathrm{CD}_I \cong c_I'c_I = \sum_{i\in I}\frac{t_i^2}{q}\cdot\frac{h_{ii}}{1-h_{ii}}$$

となる。ここで，標準化残差（t 分布に従う統計量）についてはその 2 乗（つまり F 分布に従う統計量）の期待値 $t_i^2=(n-q)/(n-q-2)$ を，てこ比についてはその平均 $h_{ii}=q/n$ を代入して打切り点を導く。よって，

$$\sum_{i\in I}\frac{t_i^2}{q}\cdot\frac{h_{ii}}{1-h_{ii}} = \frac{m}{q}\cdot\frac{n-q}{n-q-2}\cdot\frac{\frac{q}{n}}{1-\frac{q}{n}} = \frac{m}{n-q-2} \tag{5.2}$$

となる。

(5.2)式の導出方法と同様に，(4.3)式の CD^*_{IJ} についても打切り点を類推する。Cook の距離の定義式である(2.2)式の第三表現における H_I が(4.3)式における $H_{I[J]}$ に置き換わっただけであり，$H_{[J]}$ の対角成分の和が $q-\ell$ であるので，この平均 $(q-\ell)/n$ を利用すればよいのである。よって，CD^*_{IJ} の近似式から

$$CD^*_{IJ} \cong \sum_{i\in I} CD^*_{ij} = \sum_{i\in I} \frac{t_i^2}{q}\cdot\frac{h_{ii[J]}}{1-h_{ii}} = \frac{m}{q}\cdot\frac{n-q}{n-q-2}\cdot\frac{\dfrac{q-\ell}{n}}{1-\dfrac{q}{n}} = \frac{m}{q}\cdot\frac{q-\ell}{n-q-2} \tag{5.3}$$

となる。

したがって，上記のことから，(4.2)式の D_{IJ} の打切り点は，係数 q/ℓ を考慮して(5.2)式および(5.3)式から

$$D_{IJ} > \frac{m}{n-q-2} \tag{5.4}$$

となる。(5.4)式は一般的な場合であるので，第 j 番目の説明変数に対する第 i 番目の観測値に限定して比較を行うことにする。つまり，$I=\{i\}$ および $J=\{j\}$ の場合に対して比較をするので，(5.4)式は，

$$D_{ij} > \frac{1}{n-q-2} \tag{5.5}$$

となる。この(5.5)式と従来の打切り点である(5.1)式のそれぞれの右辺についての差を計算すると，

$$\text{(5.1)式} - \text{(5.5)式} = \frac{3n-4(q-2)}{n(n-q-2)} \tag{5.6}$$

となる。(5.6)式の分母は明らかに正の値であるので，分子の大小関係だけが問題になる。分子についても，$q\geq 2$($q=2$ は「定数項＋説明変数」の単回帰）であるので，

$$n > \frac{4}{3}(q-2)$$

の場合に，提案する新しい打切り点(5.5)式が従来の(5.1)式よりも小さくなり，より厳密な打切り点になるといえる。

一般的な場合については比較対象となる打切り点が存在しないので，明確なことは言えないが，特定の一つの説明変数に対する複数個の観測値の影響力評価については，一つの観測値の場合と同様の結果になることが予想される。だが，複数の説明変数に対する複数個の観測値の影響力評価については，CD^*_{IJ} の打切り点をより精密に検討する必要があると思われるので，簡単に判断をすることはできないであろう。

6　まとめ

本論文では，部分影響力を評価するための診断統計量について，従来の指標に関する新表現を提案し，その新表現に基づく打切り点を導出した。新表現は観測値の影響力評価（回帰診断）においてよく利用される Cook の距離とよく似た表現形式になっている。このため，通常の回帰診断を実施すればそれと連動させて部分影響力を測定し評価することも可能となる。また，この新表現に対して新たに導出された打切り点が，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が示された。従来の打切り点は，大き目に設定されており，実際のデータ解析での影響力評価において，あまり有効ではないという欠点が指摘されている。けれども，この新しい打切り点により部分影響力評価のための診断統計量が，実用上，より利用しやすいものになったと考えられる。

今後の課題としては，変数選択問題と部分影響力の評価方法における説明変数集合と観測値集合の除去手順の違いによる立場の違いをより厳密に検討し，それぞれの相違点あるいは類似点を明確にする必要がある。また，二つの評価方法の相補性についても，一般的な診断統計量との関連性から，実用上の観点も視野に入れて検討することが必要であると考えられる。加えて，この二つの課題とは異なるが，Cook の距離の上限と下限の議論についても，部分影響力評価と関連付けて研究を進めていきたいと考えている。

付録 A：(3. 1) 式および (3. 3) 式の同一性

(3. 1)式の第二表現の第一式を 2 乗すると

$$\mathrm{DFBETAS}_{ij}^2 = \frac{t_i^2}{1-h_{ii}} \cdot \frac{w_{ij}^2}{\mathrm{W}_j'\mathrm{W}_j}$$

となる。(3. 3)式への式変形を考える上で，$w_{ij}^2/\mathrm{W}_j'\mathrm{W}_j$ をハット行列 H の成分により表現することが問題になる。そこで，$\mathrm{X}=(\mathrm{X}_{[j]}\,\mathrm{X}_j)$ とし，

$$\mathrm{X} = \begin{bmatrix} \mathrm{X}_{(i)[j]} & \mathrm{X}_{(i)j} \\ x_{i[j]} & x_{ij} \end{bmatrix}$$

と分割すると

$$\mathrm{H} = \mathrm{H}_{[j]} + \frac{(\mathrm{I}_n - \mathrm{H}_{[j]})\mathrm{X}_j\mathrm{X}_j'(\mathrm{I}_n - \mathrm{H}_{[j]})}{\mathrm{X}_j'(\mathrm{I}_n - \mathrm{H}_{[j]})\mathrm{X}_j} = \mathrm{H}_{[j]} + \frac{\mathrm{W}_j\mathrm{W}_j'}{\mathrm{W}_j'\mathrm{W}_j}$$

であるので，

$$\mathrm{H} - \mathrm{H}_{[j]} = \frac{\mathrm{W}_j\mathrm{W}_j'}{\mathrm{W}_j'\mathrm{W}_j}$$

となる。この両辺の行列成分は当然等しいので，第 i 対角成分についても

$$h_{ii}-h_{ii[j]}=\frac{w_{ij}^2}{\mathrm{W}_j'\mathrm{W}_j}$$

となる。ただし，

$$\mathrm{W}_j=\begin{bmatrix}\mathrm{X}_{(i)j}-\mathrm{X}_{(i)[j]}(\mathrm{X}_{[j]}'\mathrm{X}_{[j]})^{-1}\mathrm{X}_{[j]}'\mathrm{X}_j\\ x_{ij}-x_{i[j]}(\mathrm{X}_{[j]}'\mathrm{X}_{[j]})^{-1}\mathrm{X}_{[j]}'\mathrm{X}_j\end{bmatrix}=\begin{pmatrix}\mathrm{W}_{(i)j}\\ w_{ij}\end{pmatrix}$$

である。したがって，

$$\mathrm{DFBETAS}_{ij}^2=\frac{t_i^2}{1-h_{ii}}(h_{ii}-h_{ii[j]})=\mathrm{D}_i(\hat{\Psi})$$

となり，(3.1)式の第二表現の第一式を2乗したものと(3.3)式は一致することがわかる。

付録 B：(4.1) 式の導出

(3.2)式を $\hat{\beta}$ および $\hat{\beta}_{(I)}$ を使って表すと

$$\mathrm{D}_I(\hat{\Psi})=\frac{(\hat{\beta}-\hat{\beta}_{(I)})'\mathrm{L}'[\mathrm{L}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}']^{-1}\mathrm{L}(\hat{\beta}-\hat{\beta}_{(I)})}{\ell\tilde{\sigma}^2} \tag{B.1}$$

となる。ここで，$\mathrm{L}=(\mathrm{O}\quad \mathrm{I}_\ell)$，$\mathrm{A}=(\mathrm{X}_{[J]}'\mathrm{X}_{[J]})^{-1}$ それに $\mathrm{B}=[\mathrm{X}_J'(\mathrm{I}_n-\mathrm{H}_{[J]})\mathrm{X}_J]^{-1}$ とすると，

$$\mathrm{L}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}'=(\mathrm{O}\quad \mathrm{I}_\ell)\begin{bmatrix}\mathrm{A}+\mathrm{A}\mathrm{X}_{[J]}'\mathrm{X}_J\mathrm{B}\mathrm{X}_J'\mathrm{X}_{[J]}\mathrm{A} & -\mathrm{A}\mathrm{X}_{[J]}'\mathrm{X}_J\mathrm{B}\\ -\mathrm{B}\mathrm{X}_J'\mathrm{X}_{[J]}\mathrm{A} & \mathrm{B}\end{bmatrix}\begin{pmatrix}\mathrm{O}\\ \mathrm{I}_\ell\end{pmatrix}$$

$$=\mathrm{B}=\frac{1}{\sigma^2}\cdot\sigma^2(\mathrm{X}'\mathrm{X})_{JJ}^{-1}=\frac{1}{\sigma^2}\mathrm{Var}(\hat{\beta}_J)$$

であり，

$$\mathrm{L}(\hat{\beta}-\hat{\beta}_{(I)})=(\mathrm{O}\quad \mathrm{I}_\ell)\begin{pmatrix}\hat{\beta}_{[J]}-\hat{\beta}_{(I)[J]}\\ \hat{\beta}_J-\hat{\beta}_{(I)J}\end{pmatrix}=\hat{\beta}_J-\hat{\beta}_{(I)J}$$

である。よって，

$$\mathrm{D}_I(\hat{\Psi})=\frac{(\hat{\beta}_J-\hat{\beta}_{(I)J})'[\mathrm{Var}(\hat{\beta}_J)]^{-1}(\hat{\beta}_J-\hat{\beta}_{(I)J})}{\ell}=\mathrm{D}_{IJ}$$

である。ただし，σ^2 をその不偏推定量 $\tilde{\sigma}^2$ で置き換えている。これで(3.2)式と(4.1)式の第一表現は一致することが証明された。

つぎに，(3.2)式から(4.1)式の第二表現および第三表現への式変形を示す。(3.2)式の分子に着目すると(B.1)式の分子から

$$\begin{aligned}&(\hat{\beta}-\hat{\beta}_{(I)})'\mathrm{L}'[\mathrm{L}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}']^{-1}\mathrm{L}(\hat{\beta}-\hat{\beta}_{(I)})\\ &\quad=[\mathrm{X}(\hat{\beta}-\hat{\beta}_{(I)})]'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}'[\mathrm{L}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}']^{-1}\mathrm{L}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'[\mathrm{X}(\hat{\beta}-\hat{\beta}_{(I)})]\\ &\quad=(\hat{y}-\hat{y}_{(I)})'\mathrm{C}(\mathrm{C}'\mathrm{C})^{-1}\mathrm{C}'(\hat{y}-\hat{y}_{(I)})\end{aligned} \tag{B.2}$$

となる。ただし，$\hat{y}=\mathrm{X}\hat{\beta}$, $\hat{y}_{(I)}=\mathrm{X}\hat{\beta}_{(I)}$ それに $\mathrm{C}=\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{L}'$ である。ここで，上記の式変形過程の結果から

$$(C'C)^{-1} = X_J'(I_n - H_{[J]})X_J = B^{-1}$$

であり，また

$$C = X(X'X)^{-1}L' = (I_n - H_{[J]})X_J[X_J'(I_n - H_{[J]})X_J]^{-1} = (I_n - H_{[J]})X_JB$$

であるので，

$$\begin{aligned} C(C'C)^{-1}C' &= (I_n - H_{[J]})X_J[X_J'(I_n - H_{[J]})X_J]^{-1}X_J'(I_n - H_{[J]}) \\ &= W_J(W_J'W_J)^{-1}W_J' = H - H_{[J]} \end{aligned}$$

となる。よって，(B. 2)式は

$$\begin{aligned} (\hat{\beta} - \hat{\beta}_{(J)})'L'[L(X'X)^{-1}L']^{-1}L(\hat{\beta} - \hat{\beta}_{(J)}) &= (\hat{y} - \hat{y}_{(J)})'(H - H_{[J]})(\hat{y} - \hat{y}_{(J)}) \\ &= (\hat{\beta} - \hat{\beta}_{(J)})'X(H - H_{[J]})X(\hat{\beta} - \hat{\beta}_{(J)}) \end{aligned}$$

と式変形でき，(3. 2)式の分子は(4. 1)式の定数項 q/ℓ 部分を除き，それぞれ第三表現および第二表現の分子と一致する。

以上のことより，(3. 2)式から(4. 1)式を導出できることが証明された。

参 考 文 献

[1] Barrett, B. E. and Gray, J. B. (1992), Efficient computation of subset influence in regression, *Journal of Computational and Graphical Statistics*, 1, 271-286.

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

[3] Castillo, E., Hadi, A. S., Conejo, A. and Fernández-Canteli, A. (2004), A general method for local sensitivity analysis with application to regression models and other optimization problems, *Technometrics*, 46, 430-444.

[4] Chatterjee, S. and Hadi, A. S. (1986), Influential observations, high leverage points and outliers in linear regression, *Statistical Science*, 1, 379-416.

[5] Cook, R. D. and Weisberg, S. (1980), Characterizations of an empirical influence function for detecting influential cases in regression, *Technometrics*, 22, 495-508.

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

[7] Léger, C. and Altman, N. (1993), Assessing influence in variable selection problems, *Journal of the American Statistical Association*, 88, 547-556.

[8] Takeuchi, H. (1991), Detecting influential observations by using a new expression of Cook's distance, *Communications in Statistics—Theory and Methods*, 20, 261-274.

[9] Takeuchi, H. (1992), Regression diagnostics using a new expression of Welsch-Kuh distance, 静岡県立大学経営情報学部報「経営と情報」, 4, 17-26.

[10] Takeuchi, H. (2002), Assessment of influence of individual observations on prediction mean square errors in variable selection problems, *Journal of the Japan Statistical Society*,

32, 43-55.
［11］ 竹内秀一 (2003)，変数選択問題における観測値除去法に基づく診断統計量, 人文自然科学論集，116 号，23-36.
［12］ 竹内秀一 (2005)，線形回帰における尤度距離による影響力評価, 人文自然科学論集，119 号，19-30.