名古屋女子大学　紀要　**39**（家・自）63～70　1993

回帰における影響関数

田中浩光

Influence Function in Linear Regression

Hiromitsu TANAKA

序

回帰分析の実践においては，現象と合致しない不都合な解釈を強いられるなど適用上の問題点が指摘されている．推定結果が現象に整合しない１つの大きな原因には回帰残差（観測値と想定回帰モデルによるそのあてはめた値の差）に影響を与える，所謂「はずれ値」の存在があげられる。はずれ値を検出するための方策には回帰残差の吟味，あるいは個々の観測値の推定回帰式への影響を測るCook型距離統計量の利用が，推定回帰式の適切性を診断する観点に基づき提示される．回帰残差の直接吟味による点検，あるいは検定統計量に基づくはずれ値の検出はモデルに付随する前提条件を問い，点検・検出過程が分析目的に沿っていないこともあり，推定結果が意図と合致しない．

一方，分析目的に適う形式を採るCook (1977)の距離関数を一般化したCook型距離統計量が，摂動(perturbation)の影響を測ることを主眼とする，Hampelの影響関数に源を置くことに注目する．Cook型距離統計量を影響関数と対比することで有用な解釈を得るべく再表現する．とくに，Draper and John(1981)は，Cook距離統計量が残差項と観測点集合の中心点からのずれを示す項の２要素の積として表現できることから，その影響度を必ずしも残差項のみに拠るものでないと指摘している．この指摘は，はずれ値の検出過定を構成的に捉える意味で重要である．

ここでは，勘場・田中・後藤(1981)に従い，一連のCook型距離統計量が回帰係数の最小二乗推定量を基軸にして構成されることに着目し，縮小推定量であるSchlove (1974)の推定量の導入が影響観測値の検出において非直交性の影響を緩和することを若干の数値実験で与えている．とくに偏りの推定量の導入の視点を再確認し，その意義を影響関数の検出過程を通して探る．最後に，実践のデータ解析に対する適用上の問題を言及し，併せてその解決策に若干の示唆を与える．

回帰モデル

線形な回帰モデルは

$$Y = X\beta + \underline{\varepsilon},\ \underline{\varepsilon} \sim N(\underline{0},\ \sigma^2 In) \qquad (1)$$

と表される．ここに，Xはp個の説明変数からなる$n \times p$行列，βは未知パラメータ（定数項も含む）からなる$p \times 1$ベクトル，Yは応答からなる$n \times 1$ベクトルである．Inは$n \times n$の恒等行列である．$\underline{\varepsilon}$は誤差を示す$n \times 1$ベクトルであり，その分布は正規分布とする．

影　響　関　数

Hampel (1968)は標的関数の摂動(perturbation)をみるために，観測値$(y_i, \underline{x}_i)$を得たときの分布Fのずれ$\varepsilon \cdot \delta x_i, y_i + (1-\varepsilon)F$が与える影響として捉える影響関数(Influence Function)を，

$$IF(i)=\lim_{\varepsilon\to 0}\frac{\mathrm{T}[(1-\varepsilon)\mathrm{F}+\varepsilon\delta x_i, y_i]-\mathrm{T}[\mathrm{F}]}{\varepsilon} \qquad (2)$$

で定義する．ここに，Fは累積分布関数，$\delta x_i, y_i$は(y_i, x_i)のみで1，それ以外で0をとる指標関数，T(·)はベクトル値統計量とする．この影響関数は感度分析(sensitivity analysis)に有用である．

$IF(i)$がベクトル値統計量であることから位置の尺度に対する比に不変性条件を要請することでノルム化する。次の2次形式の距離統計量

$$D(i/M, C^*)=\frac{1}{C^*}IF(i)\cdot M\cdot IF(i) \qquad (3)$$

を定義する．ここに，M, C^*は適当に選択するものとする．

第i観測値の推定回帰式に与える影響を測るため，Cook (1977)は式(3)の類推から第i観測値を削除するときの交叉確認方式(cross-validatory form)で測る距離統計量(Cook統計量)を提案した．

$$C(i)=\frac{(\hat{\beta}-\hat{\beta}_{(i)})X^TX(\hat{\beta}-\hat{\beta}_{(i)})}{ps^2} \qquad (4)$$

ここに，$\hat{Y}_{(i)}=X\hat{\beta}_{(i)}$であり，$s^2$は$\sigma^2$の不偏定量である．因に，Cook距離統計量は(3)式の表理を借りれば$D(i|X^TXi^{(n-1)}ps^2)$で表すことができる．$C(i)$が観測点の推定回帰係数に与える影響を差の形式で捉えているのに対し，Andrews and Pregibon (1977)はその影響を比の形式で表現したAP統計量を提示している．第i観測値の影響を測るAP統計量は

$$\mathrm{AP}(i)=|\mathrm{x}^{*T}_{(i)}\mathrm{x}^*_{(i)}|/|\mathrm{x}^{*T}\mathrm{x}^*| \qquad (5)$$

で与えられる．ここに，$X^*=(X:Y)$であり，$X^*_{(i)}$は第i観測値を削除したXである．同様に，Belsley, Kun, and Welsch (1980)は$\hat{\beta}_{(i)}$の分散行列Var $(\hat{\beta}_{(i)})$の行列式のVar $(\hat{\beta})$の行列式に対する比を影響関数とする検出統計量CVR (i)を提示している．

Cook型距離統計量の検出機構

Cook距離統計量(4)は$\underline{\beta}$に対する$\hat{\underline{\beta}}_{(i)}$の100$(1-\alpha)$%信頼楕円体に基づく影響関数として解釈できる．すなわち，(4)式を一般化することで，$C_{(i)}$をあらたに

$$C(i|M,C^*)=\frac{(\hat{\beta}-\hat{\beta}_{(i)})^{\mathrm{T}}\mathrm{M}(\hat{\beta}-\hat{\beta}_{(i)})}{C} \qquad (6)$$

として，この距離統計量をCook型距離統計量として呼ぶ．このM, Cはそれぞれ重みつき距離を表す因数，基準化因数と呼ぶ．Cook統計量の修正を意図して，Welsch and Kuh (1977), Belsley, Kun and Welsch (1980), Atkison (1981), Welsch (1982)がそれぞれM, Cを適当に選択することでCook距離統計量の枠の中で一連の修正Cook距離統計量を提案している．ここに，$X^{\mathrm{T}}_{(i)}X_{(i)}$, $\mathrm{S}^2_{(i)}$はそれぞれX^TX, S^2に対して第i観測値を削除して得られたものである．

Xの非直交性の影響

$\hat{\beta}$の分散行列，分散は，それぞれ

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{7}$$

$$tr\,(\mathrm{Var}\,(\hat{\beta})) = \sigma^2 \sum_{j=1}^{p} (1/\lambda j) \tag{8}$$

される．ここに，λ_1，……，λ_pは，$X^T X$の固有値である．

推定回帰式の適切性の評価尺度には，上記の分散の他に，回帰目的が回帰構造の解釈にあるときは真値$\underline{\beta}$に対する距離を測る平均平方誤差基準，あるいは予測にあるときは将来の観測値yに対する距離を測る予測平均平方誤差基準が有用となる．$\hat{\underline{\beta}}$の平均平方誤差は，

$$\begin{aligned} \mathrm{MSE}\,(\hat{\underline{\beta}}) &= \sigma^2 tr\,(X^T X)^{-1} \\ &= \sigma^2 \sum_{j=1}^{p} (1/\lambda_j) \end{aligned} \tag{9}$$

$\hat{\beta}$の予測平均平方誤差は，

$$\mathrm{MSEP}\,(\hat{Y}/X_0) = \sigma^2 (1 + X_0 (X^T X)^{-1} X_0) \tag{10}$$

となる．ここに，X_0は将来の観測点である．λ_1，……，λ_pは，$X^T X$の固有値である．上記の式は，いずれも$X^T X$が含まれていることから，Xに特異行列のときは，非常に不安定な推定量となる．前述のCook距離統計量とAP統計量は，

$$C(i) = t^2_i h_{ii} / \{p\,(1 - h_{ii})\} \tag{11}$$

$$\mathrm{AP}\,(i) = \left(1 - \frac{ti^2}{n-p}\right)(1 - h_{ii}) \tag{12}$$

と再表現される．ここに，

$$h_{ii} = X_i^T (X^T X^T) X_j$$

式(11)と式(12)のいずれの影響関数も残差の関係項と観測点集合の全体配置でのずれを示す項h_{ii}の2要素の積として表現できる．

これらの表現では，Xの非直交性を表す情報は含まれていないことに注意を要する．表1には，Cook型距離関数，AT統計量，CVR統計量の各種影響観測値の検出統計量が上述の視点に基づく2要素の積での再表現される．

表1．影響関数と再表現

	影響関数 M	影響関数 c	再表現	
C(i)	$X^T X$	ps^2	$\frac{1}{p} t_i^2 \frac{h_{ii}}{1-h_{ii}}$	Cook(1977)
WK(i)*	$X^T X$	$s^2(i)$	$t_i^{*2} \frac{h_{ii}}{1-h_{ii}}$	Welsh and Kuh(1977)
BKW(i)*	$X^T X$	$ps^2(i)$	$\frac{1}{p} t_i^{*2} \frac{h_{ii}}{1-h_{ii}}$	Belsley, Kuh, and Welsh(1980)
AT(i)	$X^T X$	$\frac{p}{n-p} s^2(i)$	$\frac{n-p}{p} t_i^{*2} \frac{h_{ii}}{1-h_{ii}}$	Atkinson(1981)
W(i)*	$X^T(i) X(i)$	$s^2(i)$	$t_i^{*2} \frac{h_{ii}}{(1-h_{ii})^2}$	Welsh(1982)
AP(i)	$\frac{\lvert X(i)^{*T} X(i)^* \rvert}{\lvert X^{*T} X^{*T} \rvert}$		$(1 - \frac{t_i^2}{n-p})(1-h_{ii})$	Andrews and Pregibon(1978)
CVR(i)	$\frac{\lvert \mathrm{Var}(\hat{\beta}(i)) \rvert}{\lvert \mathrm{Var}(\hat{\beta}) \rvert}$		$(\frac{n-p-t_i^2}{n-p-1})^p \frac{1}{1-h_{ii}}$	Belsley, Kuh, and Welsh(1980)

※ $X^* = (X : Y)$

偏りのある推定量の導入

回帰問題では，Xの非直交性が高いとき，所謂「悪条件の問題（条件付回帰モデルでは多重共線性の問題）」が生じるが，影響関数の場合も例外でない．ここではその対応策として，影響観測値の検出統計量の構成において最小二乗推定量$\hat{\underline{\beta}}$の代替推定量である縮小推定量の導入を考察する．文献的には，リッヂ回帰推定量，Schlove推定量を代替する縮小推定量として用いたCook距離統計量を調整した検出統計量が提示されている．多重共線性問題の解消に有用とされるリッヂ回帰推定量は

$$\hat{\underline{\beta}}_R = (X^TX + \mathrm{kI})^{-1}X^TY \qquad (13)$$

である(Hoerl and Kennard(1970))．ここに，Kはリッヂ因数である．Walker and Birch (1988)は，この$\hat{\beta}_R$を導入することで1つの調整Cook距離統計量

$$\mathrm{BR}(i) = (\hat{\underline{\beta}}_R - \hat{\beta}_{R(i)})^TX^TX(\hat{\beta}_R - \hat{\underline{\beta}}_{R(i)}) \,/\, ps^2$$

を提示している．しかし，この導入の視点は，偏に$\hat{\underline{\beta}}$の改良を意図し，リッヂ回帰推定量の性能の評価に基づき提示されている．$\hat{\underline{\beta}}_{R(i)}$は第$i$観測値を削除するときの$\hat{\beta}_R$である．一方，Cook型距離統計量が$\hat{\underline{\beta}}$を基軸にしていることに着目し，偏りのある縮小推定量としてSchlove推定量の導入は影響観測値の検出状況の変化をもたらすことを期待している．（勘場・田中・後藤（1981））．この視点に沿えば，影響関数は「悪い」観測値の検出，削除から「良好な」観測値の検出残存に対する機能を有するとして理解できることになる．

ここでは勘場・田中・後藤（1981）に従い，偏りのある推定量として(Schlove)の推定量

$$\hat{\underline{\beta}}^* = \left(1 - \frac{p-2}{n-p+2} \cdot \frac{(n-p)s^2}{\hat{\underline{\beta}}'\hat{\underline{\beta}}}\right)\hat{\underline{\beta}}$$

を導入して，Cook距離推定量を調整する．この$\hat{\underline{\beta}}^*$は偏りを有するが，説明変数の数が3以上のとき$\underline{\beta}$に対する平均平方誤差基準の意味で$\hat{\underline{\beta}}$よりも良好となることが知られている．このとき，調整Cook距離統計量は

$$BC(i) = (\hat{\underline{\beta}}^* - \hat{\underline{\beta}}(i))^TX^TX(\hat{\underline{\beta}}^* - \hat{\underline{\beta}}^*(i)) \,/\, ps^2 \qquad (14)$$

と定義される．$\hat{\underline{\beta}}^*(i)$は第$i$観測値を削除するときの$\hat{\underline{\beta}}^*$である．

特定の観測値を観測点集合からの削除対象とすべきか否か，次の3状況で検討する(図1)．

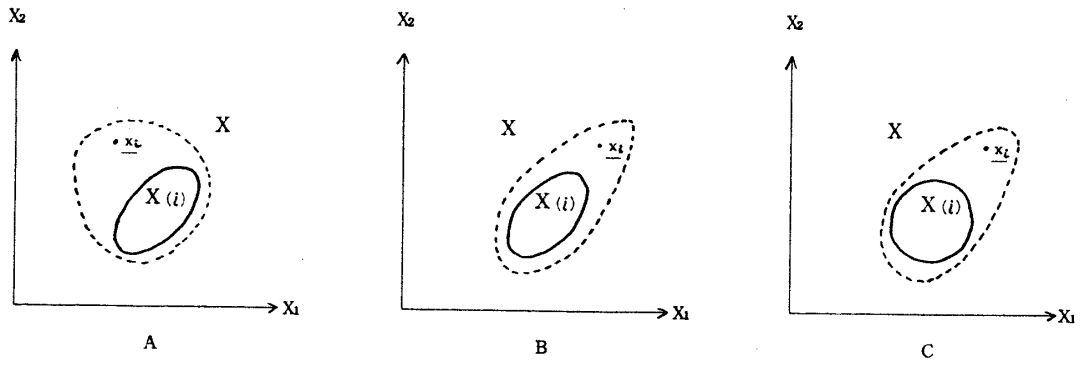

図1．$\underline{X}(i)$の削除と観測点集合

(1) 特定の観測点が非直交性の高い残余観測点集合の主軸方向と法線上にある状況では，特定の観測点を削除無しとすれば非直交性の程度は緩和される（図1のAに相当）.

(2) 特定の観測点が非直交性の高い残余観測点集合の主軸方向の延長線上にある状況では，特定の観測点を削除有りとすれば非直交性の程度は緩和されるがその程度は小さい（図1のBに相当）.

(3) 特定の観測点が直交性の高い残余の観測点集合から離れている場合は，特定の観測点を削除無しとすれば直交性が崩れる（図1のCに相当）.

以下に数値実験により調整Cook距離統計量の性能を評価する．特定の観測点が残余観測点集合の中心点から乗離する状況を設定し，影響関数による検出，削除が非直交性の緩和に有効か否かを検証することを狙いとする.

ここでは，上記の意図の下に数値実験のデザインを次の2通り（A，B）に分けて設計する．実験Aでは特定の観測点（1，－3，3）の削除はXの非直交性を残すと期待できる（図2）．実験Bでは特定の観測点（1，3，3）の削除はXの非直交性を緩和するがその程度は小さいことを予想する（図3）．実験Aは状況(1)に，実験Bは状況(3)に対応する．状況(2)については特定の観測点の削除が非直交性を緩和する意味で状況(1)と同じであり，その程度が小さいことから数値実験を省略する.

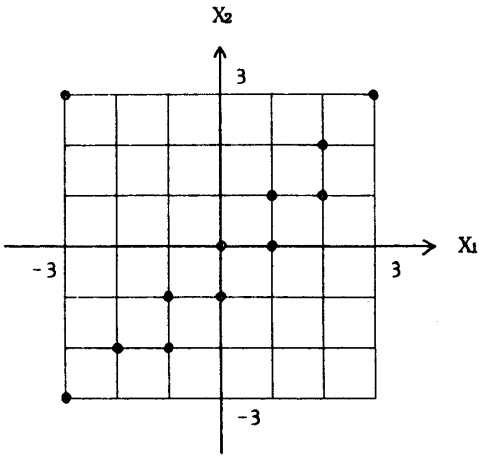

図2．観測点の配置（実験A）

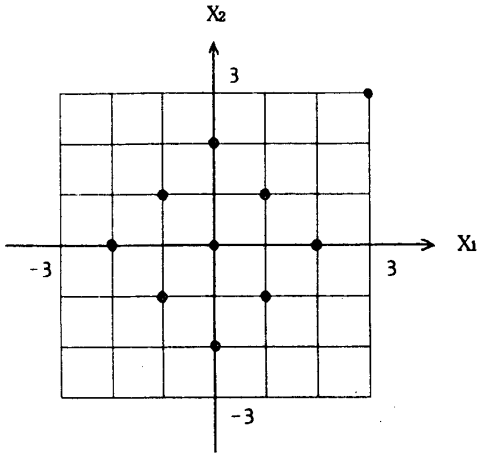

図3．観測点の配置（実験B）

(1) 実験A

・X：説明変数からなる$n \times p$計画行列（観測数$n=12$，説明変数の数$p=3$）

・β：回帰係数ベクトル$\beta^T=(\beta_1,\beta_2,\beta_3)$，$\beta_1=0.0$，$\beta_2=1.0$，$\beta_3=1.0$

・ε：正規誤差$(N(0,\sigma^2 In))$であり分散σ^2は25.0である

・数値検証に要した反復回数は 100回とした.

(2) 実験B

・X：説明変数からなる$n \times p$計画行列（観測数$n=10$，説明変数の数$p=3$）

・β：回帰係数ベクトル$\beta^T=(\beta_1,\beta_2,\beta_3)$，$\beta_1=0.0$，$\beta_2=1.0$，$\beta_3=1.0$

・ε：正規誤差$(N(0,\sigma^2 In))$であり分散σ^2は1.0である

・数値検証に要した反復回数は100回とした.

ここでは，各種統計量の性能を，観測値(点)の検出頻度(最大統計値，AP統計量AP(i)では最小統計値に対応する観測点の検出）で評価する.

実験Aでは，観測点(1, − 3, 3)に注目すると，この観測点を検出する割合は調整Cook統計量BC (i)(77%)，C (i)(80%) およびAP (i)(100%) の順に小さい．このことはBC (i)がXの直交性を考慮に入れた改良になっていることを示唆する．AP (i)の適用はすべての反復で観測点（1, − 3, 3）を検出している．このことはAP統計量自体がXのずれの大きさに影響される事実を反映している（表2）．

実験Bでは，観測点（1, 3, 3）に注目すると，この観測点を検出する割合はAP (i)(56%)，BC (i)(50%)，C (i)(44%)の順に小さい．BC (i)がAP (i)に次いで検出し，C (i)より検出が高いことは観測点（1, 3, 3）が削除の対象にあるが，非直交性には影響を与えないことがわかる（表3），BC (i)は平均平方誤差で比較すると他の2個の統計量に比し，良好な性能を示している．因に，BC (i)，AP (i)，C (i)の平均平方誤差はそれぞれ22.29, 24.83, 24.80である．

表2．影響観測値としての検出頻度（実験A）

観測点	AP (i)	C (i)	BC (i)
(1, 3, 3)	0	10	9
(1, 2, 2)	0	3	3
(1, 2, 1)	0	1	3
(1, 1, 0)	0	0	0
(1, 1, 1)	0	1	0
(1, 0, 0)	0	0	0
(1, 0, − 1)	0	1	2
(1, − 1, − 2)	0	2	4
(1, − 1, − 1)	0	0	0
(1, − 2, − 2)	0	0	0
(1, − 3, − 3)	0	2	2
(1, − 3, 3)	100	80	77

表3．影響観測値としての検出頻度（実験B）

観測点	AP (i)	C (i)	BC (i)
(1, 2, 0)	2	4	3
(1, 1, 1)	2	0	0
(1, 1, − 1)	4	5	5
(1, 0, − 2)	8	17	15
(1, 0, 0)	4	1	1
(1, 0, 2)	10	9	8
(1, − 1, − 1)	1	2	1
(1, − 1, 1)	5	7	6
(1, − 2, 0)	8	11	11
(1, 3, 3)	56	44	50

若干の問題

Cook (1977), Belsch, Kuh and Welsch (1980), Welsch (1982)等の一連のCook型距離統計量は，回帰における影響力のある観測値の検出に対して有効な1つの診断道具としての役割を確保している．しかし，実践のデータ解析では，これらの統計量が有用な道具として効果的に作動するには幾つかの問題点を含む．ここでは，適用の限界を探ることで残された問題点を洗い出す．

第1点は，影響力の強い観測値が各種検出統計量で特定できるが，その削除の有無に対する決定的意味づけはない．これは，これらCook型距離統計量が影響関数の枠組の中で構成されていることに拠る．ある合理的リスク関数の構成が可能になるとき最適な臨海水準の決定問題に帰着できることになる．第2点は，重回帰ではXの非直交性の影響が大きい．Cook型距離統計量を含む各種検出統計量は，残差の関係項t_iと観測点集合からの乖離を示す梃率h_{ii}の2要素の積で分解されている．しかし，非直交性の影響を組み込むことで3要素の積の分解は影響力をモデルに即して構成的に把握することができる．推定回帰式の適切性を探る意味で有用であると期待できる．第3点は，影響関数の構成では観測値だけでなく，説明変数の検出にも注意が要る．観測値の影響を測ることは，結果的に推定回帰式の適切性を損うはずれ値，分布の端値を検出することが狙いとなる．また，回帰寄与の小さい説明変数，あるいは多重共線性を惹起する説明変数は冗長な説明変数集合，偏サンプリングに関係する説明変数，意味のない説明変数等の検出を意味し，推定回帰式の改良を導く．このように，観測値，説明変数の両側面により観測状況に即して解析を進めていくことで，データへの愛着が生まれ，さらにグラフィカル表現法を加ることで，実効のある回帰の実行が可能となり，より生産的な知見が期待できる．最後に，実践に絡む問題でよく散見するが適用する回帰モデルが観測機構と合致しないことの多いことを指摘する．実践の回帰分析では同時観測データの適用に対して，定型的に標準回帰分析の実行をみることが多い．条件付回帰モデルを基礎にする影響観測値の検出についての検討が望まれる．

要　　約

本論ではCook距離統計量が摂動の影響を測ることを主眼とするHampelの影響関数に源を置くことに注目し，Cook距離統計量を影響関数から得る再表現形式を整理する．とくにCook型距離統計量が最小二乗推定量を基軸に構成されていることに注目し，縮小推定量であるSchlove推定量の導入が影響観測値の検出問題でXの非直交性の影響を緩和することを指摘し若干の数値実験を与える．

本論の考察から，実践のデータ解析における適用上の問題を指摘し，その効果的な解決策として若干の示唆を与える．

- 特定の影響観測値の有無に関する検定問題はある合理的リスク関数の構成のもとに最適な臨界水準の決定問題に帰着できる。
- 影響関数を非直交性を組み入れる再表現形式はグラフィカル表現など加えてより実効のあるデータ解析が期待できる。
- 実践では同時観測データに対し定型的に実行する標準回帰分析が多いことから，観測機構を考慮する影響観測値の検出統計量の構成が望まれる。

文　　献

1）Chalton, D.O. and Troskie, C.G. : Commum. Statist. Simul., 21 (3), 607-626 (1992)
2）Chatterjee, S. and Hadi, A.S. : Statistical Science., **3**, 379-416 (1986)
3）Farebrother, R.W. : Commun. Statist. Simul., 21 (3), 709-710 (1992)
4）Cook, R.D. : J.R. Statist. Soc., 48 (2), 133-169 (1986)
5）Little, J.K. : Technometrics, 27 (1), 13-15 (1985)
6）Farebrother, R.W. : Technometrics, 27 (1), 85-86 (1985)
7）Sclove, S.L., Morris, C. and Radhakrishnam, R. : Ann. Math. Statist., 43, 1481-1490 (1972)
8）Hampel, F.R. : J.Amer. Statist. Assoc.; 69, 383-393, (1974)
9）Cook, R.D. : Technometrics, 19, 15-18 (1977)
10) Cook, R.D. : J. Amer. Statist. Assoc., 74, 169-174 (1979)
11) Cook, R.D. and Weisberg, S. : Residuals and Influence in Regression, Chapman and Hall (1982)
12) Weisberg, S. : Applied Linear Regression, Wiley (1985)
13) 勘場貢，田中浩光，後藤昌司：品質管理学会第11回年次大会研究発表要旨集，5-8 (1981)