名古屋女子大学　紀要　38（家・自）119～127　1992

回帰診断の１つの見方

田中浩光

A New Look at the Regression Diagnostics

Hiromitsu TANAKA

序

回帰分析に対する広範・多岐に亘る強い支持は，モデルのもつ見通しの良さにもとづく解釈の容易さと統計的性質の明瞭さによる．実地における適用では，回帰プログラムの登録化に伴う利便さによる乱用・誤用も見うけられるが，回帰係数あるいは応答値に非負の現象的制約があるときに負の推定値を得るなど推定結果と現象が必ずしも整合しない適用上の問題点が指摘されている．近年，所謂“回帰診断（Regression Diagnostics）”の研究が，これらの解釈にまつわる問題点の解消に向けて精力的に展開されている．しかし，その多くの関心は所与のデータに想定モデルをあてはめ，残差（応答値とその推定値の差）の吟味に基づいて，誤差に関する前提条件（分散均一性，系列相関の無相関性，正規性）の充足の点検に集中している　残差の吟味・点検による診断方式は，データとモデルの対比に力点がおかれ，診断後の考慮がなく治療法（Remdies）を診断と切り離して点検後の作業として捉える．Belsley, Kuh and Welsch (1980)，Velleman, Cook and Weisberg (1982)，Atkinson (1982)，Weisberg (1985)，Williams (1987)，あるいは Simonoff (1988) に代表されるように，診断に関する論文・成書が公表されているが，必ずしも診断・治療の一連の過程で捉えられていないのが現状である　Weisberg (1985) は望ましいとする回帰診断の５原則を提示したが，そこでは診断と治療の結びつきを示唆するにとどまり両者が data investigation の過程を踏まえる診断方式としては位置づけられていない

ここでは，一貫して回帰モデルを受容する立場を採り，モテルを評価するのでなくデータをモデルに基づいて診断する観点に立ち論旨を展開する．回帰データの診断方式は，データの背後に潜むサンプリング構造と data investigation の過程の相補関係に着目することではじめて診断・治療過程の中で確立することになる．すなわち，提案する診断方式では data investigation における特定の“処方（do something）”がサンプリングの不備と結びつき，その補いが診断・治療過程の中で適応的な治療を可能にする　ここでは，提案方式の１つの接近法として，Suich & Derringer (1977) の予測指標規準に基づく点検方式（以降，SD 方式と呼ぶ）を導入して，data investigation の過程を考慮する診断方式を展開する．しかし，この診断方式は data investigation の過程とサンプリングを対応づけることなく，処方のみが診断・治療過程で独り歩きしていると考えられる　この SD 方式の例示として，田中・勘場・後藤 (1981) に従い，はずれ値の削除を含む１つの回帰診断方式を与える

回帰モデルの構成

回帰モテルは，説明変数，応答変数の観測機構に応して5分類される（Press（1963））　ここでは，説明変数を数理変数とする標準回帰モテルに限定し，通常，次のように与えられる

$$Y = X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 In)$$

Xは，p個の説明変数からなる$n \times p$行列　βは，未知パラメータからなる$p \times 1$ヘクトル，Yは，応答からなる$n \times 1$ベクトルである　εは，その成分が平均ゼロで一定分散σ^2の正規分布に，独立に従う誤差の$n \times 1$ベクトルである　Inは$n \times n$恒等行列である　以降の議論を容易にするため定数β_0を含まない

回帰モテルは，X，β，εの3要素で構成されるので，診断対象には，これら3要素が絡むことになる

・ε：誤差に関する前提条件（等分散性，無相関性，正規性）

・β，σ^2：回帰性（$\gamma^2 = (X\beta)^T(X\beta)/p\sigma^2$）

・X：観測点の中心点からの乖離（マハラノヒス距離で規準化），直交性

診 断 統 計 量

診断統計量は，基本的には回帰モテルの構成要素てあるX，β，εに注目し，各要素別に用意される　すなわち，Xで多重共線性を，βで回帰性を，εで誤差の前提条件を点検する各種診断統計量が用意される　また，包括的な診断法としてテータ変換をあげることができる

(1)　観測点行列Xの影響

Xがβの最小二乗推定量$\hat{\beta}$に与える影響を$\hat{\beta}$の平均平方誤差で測ると

$$\begin{aligned} \mathrm{MSE}(\hat{\beta}) &= E(\hat{\beta}-\beta)^t(\hat{\beta}-\beta) \\ &= \sigma^2(\mathrm{X}^T\mathrm{X})^{-1} \\ &= \sigma^2 \sum_{i=1}^{p} (1/\lambda_i) \end{aligned}$$

となる　ここに，$\lambda_1 \geqq \lambda_2 \geqq \quad \geqq \lambda_p$は$X^TX$の固有値である　また，$X\hat{\beta}$の平均平方誤差は

$$\mathrm{MSE}(X\hat{\beta}) = \sigma^2 H$$

となる　ここに，Hとその要素はそれぞれハノト行列，梃率である　X_iはXの第i行からなる$p \times 1$ヘクトルである

$$H = \{h_{ij}\} = X(X^tX)^{-1}X^T$$
$$h_{ij} = X_i^T(X^TX)^{-1}X_j$$

多重共線性を測る尺度として，第i説明変数の分散膨張因数（Variance Inflation Factor）

$$VIF_i = (1 - R^2_i)^{-1}$$

が用いられる　ここに，R_iは第i説明変数の重相関係数である

(2)　残差による診断

通常，残差e_iは次のように定義される

$$e_i = y_i - \hat{y}_i$$
$$= (1 - h_{ii})\varepsilon_i - \sum_{k \neq i} h_{ik}\varepsilon_k$$

ここに，$y_i = x_i{}^T\beta$，残差 e_i は重みに梃率の関係項をもつ誤差$\{\varepsilon_i\}$の線形結合で表現されることに注目する　とくに e_i の分散 $V(e_i)$ は

$$V(e_i) = \sigma^2(1 - h_{ii})$$

と表現され，h_{ii} の影響を受けていることがわかる．

通常残差 e_i を標準化したスチューデント化残差は

$$t_i = e_i / \{\hat{\sigma}(1 - h_{ii})^{1/2}\}$$

と与えられる．ここに $\hat{\sigma}^2$ は σ^2 の不偏推定量である．

(3)　影響関数による診断

影響力のある観測値を検出するために影響関数が用意され，観測値の推定結果に及ぼす影響を測る　この影響関数は残差の関係項と観測点の中心点からのずれを示す項 h_{ii} の2要素の積として表現できるため，その影響度は残差項のみに拠るものでないことから，1つの診断基準として有用されている　ここに，残差関係項は

$$t_i = e_i / \{\hat{\sigma}(1 - h_{ii})^{1/2}\}$$

と表される　たとえば，第 i 観測値を削除したときの影響を Cook (1979)の提案する交叉確認形式（cross-validatory form）で測る Cook 統計量は

$$D_{(i)} = (X\hat{\beta}_{(i)} - X\hat{\beta})^T(X\hat{\beta}_{(i)} - X\hat{\beta})/ps^2$$
$$= t^2{}_i h_{ii}/\{p(1 - h_{ii})\}$$

と表現できる　$\hat{\beta}_{(i)}$は第 i 番目のテータを削除したときのβの最小二乗推定量である

(4)　変換による診断（Atkinson (1982)）

変換パラメータλを伴うモデルを

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

とする　ある変換パラメータ値 λ_0 の周りで1次項まで展開する

$$Y^{(\lambda_0)} \cong Y^{(\lambda)} + (\lambda_0 - \lambda)G_Y^{(\lambda_0)}$$
$$\cong X\beta + (\lambda_0 - \lambda)G_Y^{(\lambda_0)} + \varepsilon$$

ここに，$G_Y^{(\lambda_0)} = \partial Y/\partial\lambda \mid_{\lambda=\lambda_0}$ である

このとき，λの簡便推定量は

$$\tilde{\lambda} = \lambda_0 - \hat{\phi}$$

と推定される　ここに，$\hat{\phi}$ は ϕ $(=\lambda_0 - \lambda)$ の最小二乗推定量である

仮説$\lambda = \lambda_0$の検定については構成変数 $G^{(\lambda_0)}{}_Y$ の追加に関する t 統計量が自由度 $n-p$ の t 分布になることを利用して検定を行う　この変数には Box & Cox の提案したベキ変換も含む

診断方式

診断（作業）は診断・治療過程て捉えられるべきものであり，診断指標，診断基準の測定もその過程の中で考察されなければならない　診断の目的を明らかにして，診断対象を限定しその後の治療法を考慮にいれた診断・治療過程の中で検討されるべきものである　すなわち，推定回帰式の適切性を崩す原因とその発生形態の関係を摑むことで適切な治療（法）につなく診断方式が期待できる　とくに，治療が data investigation の過程における特定の“処方（do something)”として捉えられること，更にその処方がサンプリンクの不備を補う手立てに結びつくことで診断・治療過程における診断方式が確立するといえる

Weisberg（1983）は望ましい回帰診断の5原則を提示している

- 診断方式の挙動が想定モテルの下で，また，ある特定の前提条件を変更したモテルの下で少なくとも近似的に把握できること
- 前提条件をパラメータ化し，それにより臨界問題を少なくとも近似的にパラメータの推測の問題に帰着することで有用な診断が得られること
- 診断方法の計算は簡単であること
- 診断はグラフィカルに行えること
- 診断方式から対処の仕方が得られること

診断と治療

診断・治療の構造をより明確に調べるために，診断の目的と対象，推定回帰式の適切性を崩す原因とその形態，診断方式，そして，診断後の治療について各項目別に吟味して整理する（表1）　なお，図1に回帰テータの治療を含む診断過程を示す流れ図を与える

(1)　目的と対象

診断の目的が回帰の利用目的にあるのか，あるいは回帰の一般的な前提条件を満足している

表1　診断と治療

崩壊原因（欠患）	崩壊形態（症状）							原因対策（根治）療法
	モテルの前提条件崩壊	準特異(X)	はすれ値	回帰奇与小	モテル不適合	過度のあてはめ	現象と不整合	
偏サンプリンク	*	*	−	*	*	*	*	計画された追加観測
拡大サンプリンク	*	−	*	*	*	−	*	背景因子による観測値集合の層別
分布の端値	*	*	*	*	*	−	*	観測値の削除
標本サイス小	*	*	*	*	*	*	*	追加観測
説明変数過多	−	*	−	−	−	*	*	説明変数の選択
説明変数不足	*	−	*	*	*	−	*	説明変数の追加
異質標本の併合	*	−	*	*	*	−	*	観測値集合の分割
診断方法	分散均一性検定 無相関性検定 正規性検定	多重共線性検定	はすれ値検定	回帰性検定 回帰奇与率	モテル偏検定	偏係数有意性	解釈可能性	
個別対策（対処）療法	テータ変換 観測値の削除	変数選択 リッチ回帰 下成分回帰 観測値の追加	テータ変換 観測値の削除	テータ変換 変数の追加 交互作用 観測値の追加	変数の追加 交互作用	変数の追加 リッチ回帰 下成分回帰	変数の追加 下成分回帰	

(1)　表中の＊、－はそれぞれ疾患か症状に影響を及ぼす、あるいは及ぼさないことを表す。

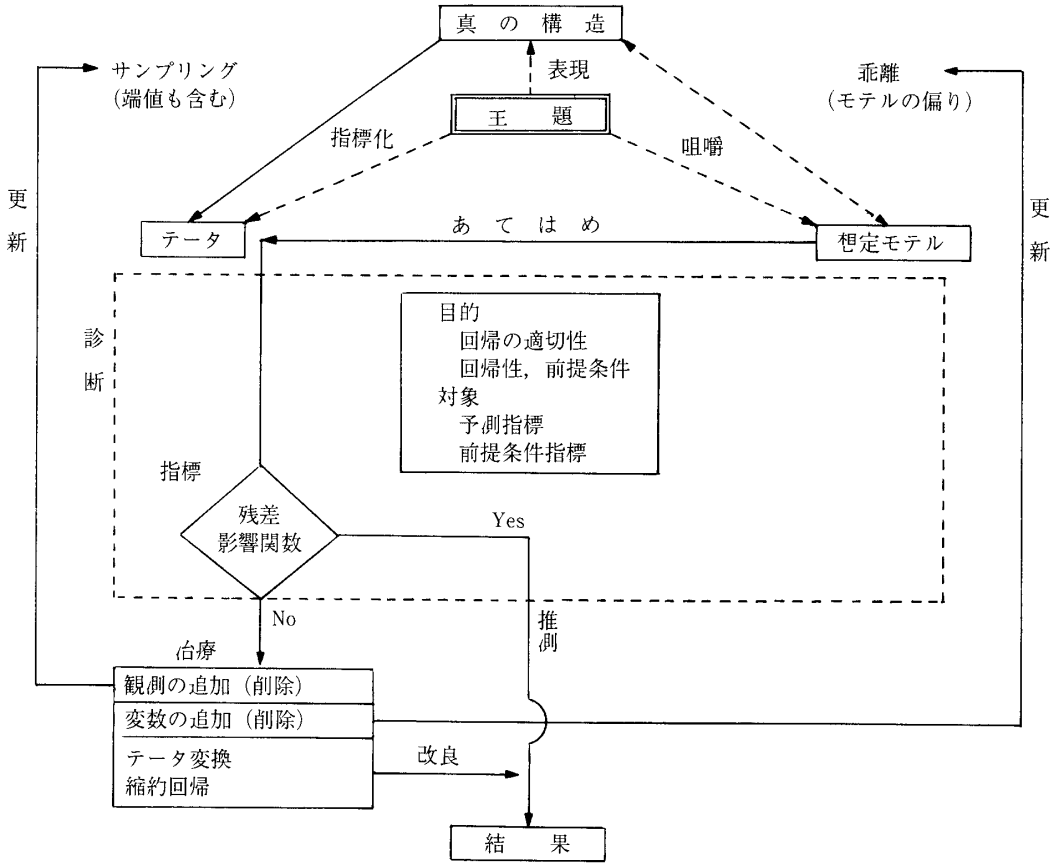

図1　診断過程（「治療」を含む）

か否かの点検にあるのかを明確にすべきである．診断の所在に応じて診断方法とその後の治療が異なる　また，診断対象を限定し，目的が予測にあるときは，推定回帰式の予測指標を対象としてとりあげる　モデルの前提条件の充足の点検のときは，個々の誤差条件を明確に対象として設定する．

(2)　崩れの想定原因

・母集団から偏りをもって標本抽出がなされる，あるいは母集団より拡大した集団から標本抽出がなされる
・母集団分布の端値に相応する標本の抽出がなされる．
・標本サイズが小さい．
・説明変数が過不足である
・異質な母集団を併合した集団から標本抽出がなされる．

(3)　崩れの形態

・データ全体が想定モデルの前提条件を満足しない
・特定の観測値集合（1観測値も含む）が想定モデルからのはずれ値となる．
・回帰式が回帰性を有しない（標本変動が回帰変動に比して大きい）．
・観測点行列 X が多重共線性（非直交性）を呈示する．
・回帰モデルの不適合が生じる　回帰寄与の高い変数が隠れている場合が多い
・応答値や回帰係数に非負の現象的制約がある場合に負の推定値を得るなど推定結果と現象が整合しない

(4)　診断方法と図示

ここでは，診断方法に対する2つの接近を与える

・所与のテータに想定モテルをあてはめて，その残差を吟味する　診断の狙い（例えば，誤差分布の正規性など）を限定した上で適切な診断統計量を評価する　影響関数の適用も考えられる

・診断とする前提条件をパラメータ表現することで，想定モテルを包含するモテル（族）を構成する　誤差の分布が規定できれば，尤度比検定が適用できる

以下に代表的なクラフィカル手法を挙けて，診断用途を併せて付記する

・散布図　・テータ全体の把握

・標準化残差・スチューテント化残差プロット　・残差の点検（はずれ値の検出など）

・部分残差プロット　・説明変数の影響

・偏回帰プロット　・　説明変数の影響

・確率プロット（P-P, Q-Q）　・・誤差分布における正規性，等分散性の点検

・影響関数の図示　・・・観測値の影響

・リッチ・トレース　　多重共線性の点検

(5)　診断後の治療

診断の狙いと崩れの想定原因に応じて診断後の治療が異なる

・テータ全体が想定モテルの前提条件を満たさないと診断された後の治療（対処）には，テータに適応的させて前提条件を満たすようにテータ変換（例えは，Box-Cox のヘキ変換）を行うことが考えられる

・テータの本体と異なる特定の観測値集合が幾つかの前提条件もしくは利用目的てある予測に影響を及ぼすと診断された後の治療には，その観測値集合の削除が考えられる

・回帰性を有しないと診断された後の治療には，観測値あるいは説明変数（交互作用も含めた）の追加が考えられる

・とくに，X が準特異であると診断された後の治療には，X の直交変換による縮約回帰（直交回帰，主成分回帰），相関の高い変数を落とす所謂“説明変数の変数選択”あるいは最小二乗推定係数値の縮小をはかるリッシ回帰の適用が考えられる　また，サンプリングに偏りがあるときは観測の追加も考えられる

・回帰性を認めるがモデル不適合と判断された後の治療には，交互作用の追加，新たな変数の追加が考えられる

・過度のあてはめと診断された後の治療には，変数選択もしくはリッチ回帰，主成分回帰の適用が考えられる

・現象の制約に不整合であり解釈が困難と判断された後の治療には，変数選択もしくは主成分回帰の適用が考えられる

1つの予測指標

回帰分析の利用目的が予測にあるとする　いま，“diagonostics”を主柱部分の診断として捉えると，その目的はモテルの前提条件を充足しているか否かを問うのでなく，推定回帰式が予測に耐えうるものかを診るものとなる　ここでは SD 方式を導入することにより診断規準を考察する　SD 方式は，利用目的を予測に限定した上で，Wetz（1964）の提案した一つの予測指標 γ^2（回帰変動対誤差変動の比：SN 比）に基づく検定方式を再検討し，回帰式の適切性の点

検方式を提案している．SD 方式では，１つの予測指標を応答の平均周りの変化測度$\sigma^2\{E(Y)\}$対応答の推定値 $Y=X\hat{\beta}$ の分散 $\sigma^2(\hat{Y})$ の比

$$\gamma^2=\sigma^2\{E(Y)\}/\sigma^2(\hat{Y})$$
$$=(X\beta)^T(X\beta)/P\sigma^2$$

で定義し，帰無仮説 H_0：$\gamma=\gamma_0$を対立仮説 H_1：$\gamma>\gamma_0$に対して検討する　このとき，統計量は，$F=(X\hat{\beta})^T(X\hat{\beta})/ps^2$で与えられ，$H_0$の F で非心パラメータ$\lambda=p\lambda^2{}_0/2$ をもつ自由度 $(p, n-1)$ の非心 F 分布に従う　s^2はσ^2の不偏推定量である　従来の回帰の有意性検定は$\gamma=0$ に対応する　SD 方式では，検定結果が有意であれば，応答の推定値に $\hat{Y}=X\hat{\beta}$ が採用され，有意でなければ，経験則に委ねる data investigation の処方（do something）が推奨されることにる

はずれ値の検出と削除：適切性の診断方式

一般に，はずれ値の意味するところは，多様であり，少なくとも３通りの発生形態が考えられる

・分布の端値

・モデルの偏り（規定誤差）に基づく特異な値

・実験，あるいは観測の失敗もしくは転記ミスなどの人為的な不確定要素による特異な値.

ここでは，はずれ値として分布の端値を考える　適切性の点検過程におけるはずれ値として，予測規準に照らして最も影響のある観測値を同定するために Cook 統計量 $D_{(i)}$を利用する　このとき，$D(i^*)=\max D_{(i)}$となる i^*に対応する観測値がはずれ値となる

ここでは，田中・勘場・後藤（1981）に従い，はずれ値の点検を SD 方式に基づいて診断作業を実施する　つまり，SD 方式で回帰性が認められないとき，分布の端値が存在すると考え，予測規準の下で，最も影響のある観測値を削除し，改めて SD 方式の点検から回帰式の適切性を診断する　“適切性の診断方式”からβの予備検定推定量$\tilde{\beta}$を構成し，その危険関数を予測平均平方誤差の意味で評価する．ここに，

$$\tilde{\beta}=I_{(0\ C_1)}(F)\,(\text{do something})+I_{(C_1,\infty)}(F)\,\hat{\beta}$$

であり，$I_{(a\ b)}(F)$ は，F が (a, b) 内にあれば１，そうでなければ０をとる指標関数である　C_1は，非心パラメータλ $(=p\gamma_0{}^2/2)$ をもつ自由度 $(p, n-p)$ の非心 F 分布の上側５％点である　この際の“do something”は，はずれ値の同定と，その削除から$\hat{\beta}_{(i)}$を得る作業を含む．ここに，

$$\tilde{\beta}=I_{(0\ C_1)}(F)\{I_{(C_2,\infty)}(F)\,\hat{\beta}_{(i)}+I_{(0,C_2)}(F_{(i)})\,(\text{do something})\}+I_{(C_1,\infty)}(F)\,\hat{\beta}$$

である　$F_{(i)}$は第 i 観測値を削除したときの F 値を示す　はずれ値の削除後の SD 方式で，有意な回帰が得られないときにはβの推定値に０を与え，data investigation はくり返さない．ここに，C_2は自由度 $(p, n-p-1)$ の中心 F 分布の上側５％点とする

$$R(\gamma, \gamma_0)=\mathrm{MSEP}(\tilde{Y}/\gamma, \gamma_0)$$
$$=n\sigma^2+E(\tilde{\beta}-\beta)^TX^TX(\tilde{\beta}-\beta)$$

となる　ここに，$\tilde{Y}=X\tilde{\beta}$である　ここでは，経験的にも事前にもγについて何の知識も持たない場合を考える　Hahn-Bancroft（1968）に従い仮説値γ_0（>0）に基づく$\tilde{Y}$の，中心 F 検定に

基づく$\tilde{Y}$（仮説値0に対応する）に対する相対効率を予測平均平方誤差の比

$$e(\gamma, \gamma_0) = \mathrm{MSEP}(\tilde{Y}/\gamma, 0)/\mathrm{MSEP}(\tilde{Y}/\gamma, \gamma_0)$$

で定義する　あるγ_0に対して未知のγを動かすとき，その最大相対効率 e^*と最小相対効率 e_0を

$$e^* = \mathrm{Max}_{\gamma>0}\ e(\gamma, \gamma_0)$$
$$e_0 = \mathrm{Min}_{\gamma>0}\ e(\gamma, \gamma_0)$$

と表す　相対効率 $e(\gamma, \gamma_0)$ の挙動をみるとすべてのγに対して一様に良好となるγ_0をもつ予備検定推定量は存在しない　そのため，少なくともある一定の相対効率以上の保証を得る限界相対効率に対応するγ_0の選定が望まれることになる　以下に，$p=1$て，$n=10$，20のときのγ_0の数値結果を表2に与える　Wetz の推奨する$\gamma_0=2$は，$n=20$のとき最大相対効率1 90で，最小相対効率0.81を与えることがわかる

表2　最大相対効率 e *と最小相対効率 e_0

	n＝10		n＝20	
γ_0	e *	e_0	e *	e_0
0 0	1 00	1 00	1 00	1 00
0 5	1 09	0 97	1 21	0 97
1 0	1 60	0 92	1 33	0 91
1 5	2 06	0 86	1 66	0 86
1 8	2 06	0 82	1 90	0 81
2 0	2 33	0 79	1 90	0 80
2 3	2 79	0 77	2 41	0 79
2 5	2 79	0 75	2 41	0 76
3 0	3 07	0 68	2 41	0 67
4 0	3 29	0 58	5 84	0 50
5 0	3 29	0 53	5 84	0 45

ま　と　め

回帰診断の概観の紹介と併せて回帰診断の構造の解明に考察を加えた　その際に，回帰の利用目的を予測におき，診断の目的が推定回帰式の適切性にあるとして議論をすすめた　サンプリング構造と data investigation の過程に注視した上で診断後の治療を考慮にいれた診断方式を示唆した　この適切性の診断方式が統計的妥当性の評価を得るべく定式化が可能になるときより実践に即した診断方式の確立が期待できる　回帰診断の1つの接近法として，“処方（do something）”の活用を図る，はずれ値の点検（削除）を考慮した SD 方式は，不完全ながらも data investigation 過程を踏まえた実践方式といえる　処方には削除以外に観測値の追加，説明変数の追加・削除，交互作用の追加そして観測値集合の層別が含まれる　より実践での診断方式の確立を意図するときは，上記の処方を考慮する適切性の診断方式の検討も必要である　また，たとえ診断目的が，回帰モデルの前提条件の点検に力点をおくときでも，予測を目的とする限りにおいては，上述の SD 基準のような予測指標の下での検討が望まれる　最後に，ここでの診断方式は推定回帰式の適否に統計的側面からの1つの目安を与えるが，診断結果での“青信号”は，必ずしも現象との整合性の確認あるいは注意深いテータの見直しといった利用者としての当然の心得を無視することを意味しないことに留意したい

文　　献

1）Atkinson, A C　*J Roy, Statist Soc*, B44, 1-36 (1982)
2）Belsley, D A, kuh, E, and Welsch, R E　*Regression Diagnostics* Wiley (1980)
3）Cook, R D　*J Amer statist Assoc*, **74** 169-174 (1979)
4）Cook, R D and Weisberg S　*Residuals and Influence in Regression*, Chapman and Hall (1982)

5) Hahn, C P, and Bancroft, T A *J Amer Statist Assoc*, **63**, 133-134 (1968)
6) Suich, R, and Derringer, G C *Technometrics*, **19**, ・213-216 (1977)
7) Weisberg, S *Applied Linear Regression*, Wiley (1985)
8) Williams, D A *Applied Statistics*, **36**, 181-191 (1987)
9) Simonoff, J F *Technometrics*, **30**, 205-214 (1988)
10) 田中告光，勘場　貢，後藤昌司：第20回日本品質管理学会予稿集，5-8 (1981)