回帰分析におけるグラフィカル接近法

松 原 義 弘*, 後 藤 昌 司*

GRAPHICAL APPROACHES IN REGRESSION ANALYSIS

Yoshihiro MATSUBARA, Masashi GOTO

In this paper, we describe and discuss some useful graphical approaches in regression analysis from two view points; one is to diagnosis certain structures underlying in fitted regression model and another is to check or evaluate sampling behaviors of the model, including check of outliers and influence observations and so on. Further, we introduce some recent graphical techniques based on computer graphic display. Most of the graphical methods are illustrated on two well-known examples, Hald's cement data and Brownlee's stack loss data.

1. 序

多変量解析や他の多くの統計解析の基盤を与える回帰分析は、そのモデルの簡明さと結果の表現・解釈の容易さから多くの応用分野で古くから汎用されており、最近の広範なコンピュータ・ネットワークの発達と容易に入手できるコンピュータプログラムがその適用範囲を益々広範多岐なものとしている。けれども、それらの利便性はどちらかといえば数値解析的な方法に偏しており、ともすればデータや仮定の点検、モデルの診断といった本来のデータ解析の基本的側面が軽くなり扱われている感がある。

一般に、回帰分析では想定モデルのあてはめと吟味から、残差の点検そしてモデルの修正あるいは調整に進む一連の過程が考えられる。Daniel & Wood (1980)を始めとする多くの研究者が回帰式のあてはめ過程で生じる多様な点検と診断の役割や方法を提示している (Cook (1977, 1979), Hoaglin & Welsch (1978), Belsley, Kuh & Welsch (1980), Cook & Weisberg (1980), Goto (1981), Atkinson (1981, 1982))。そこでは散布図や確率プロットを主体とするグラフィカル表現の諸法が有効に働いている。

本稿では主として線形回帰モデルのあてはめ過程において診断を行うのに有効な数個のグラフィカル表現

法を平均的 (集団的) 側面を指向する立場と個々の対象を (最終的に) 問題にする立場に分けて提示する。平均的な集団の代表指標を問題にする前者の立場は所謂、モデルの安定性 (ロバスト概念などを含む) を指向し、個々の対象を問題とする後者の方向は感度分析や適応的な調整の立場につながる。最近の回帰分析の展開の主流は後者の方向に変化してきている。予測や制御が結局は個々の対象に対して行われることに注目すると、この流れは当然ともいえる。なお、ここで与えるグラフィカル表現には周知の二つの文献例を一貫して用いる。

2. モデルと記法

回帰分析では通常、線形モデル

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

が仮定される。ここに $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T = \{x_{ij}\}$ は p_0 個の説明変数上の観測値からなる $n \times p_0$ 行列、 $\mathbf{y} = (y_1, \dots, y_n)^T$ はこれに対応する応答からなる $n \times 1$ ベクトル、 $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ は誤差を表す $n \times 1$ ベクトルであり、その成分は平均ゼロ、分散 σ^2 の正規分布に独立に従うと仮定する。 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{p_0})^T$ は未知のパラメータからなる $p_0 \times 1$ ベクトルである。

このとき、 $\boldsymbol{\beta}$ の最小二乗推定量 $\hat{\boldsymbol{\beta}}$ は

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (2)$$

で与えられ、予測応答は

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \equiv \mathbf{H}\mathbf{y} \quad (3)$$

* 塩野義解析センター (Shionogi Kaiseiki Center)

と得られる。ここに、 $H = X(X^T X)^{-1} X^T$ はそれが y を $\hat{y}$ へ射影する役割を強調してハット行列と呼ばれる (Hoaglin & Welsch (1978))。この行列表示を用いると $\hat{y}$ と残差 $r = y - \hat{y}$ の分散共分散行列はそれぞれ

$$\text{Var}(\hat{y}) = \sigma^2 H \quad (4)$$

$$\text{Var}(r) = \sigma^2 (I_n - H) \quad (5)$$

で与えられる。ここに I_n は $n \times n$ 恒等行列である。実際には、 σ^2 は未知であるため、その不偏推定量 $s^2 = r^T r / (n - p_0 - 1)$ が用いられることが多い。

本稿では二つの方向から回帰分析の診断過程を考える。一つはモデルの選択や回帰係数の安定性の点検あるいは調整といった X の列方向に対応する構成的評価である。これには平均応答の予測性能に対する安定性や説明変数間の病的状態の点検・調整なども含まれる。これに対し、個々の対象の予測・制御が最終的な目的であることに注目し、はずれ値や影響ある観測値などの点検を含む標本変動の吟味・評価が考えられる。すなわち X の行方向からの診断である。

3. X の列方向に対応する診断

回帰分析の適用に際して古くから利用されているものに散布図があるが、これは2変数間の関連表示に限定される。

脇本・田栗 (1974) は説明変数と応答変数の回帰関係の2次元平面上への表示を試み、単回帰と重回帰に対応させてそれぞれ連結ベクトルグラフおよび重連結ベクトルグラフを与えている。これらの方法の特徴は連結されたベクトルパターンから説明変数と応答変数間の関連の程度および変移点が視覚的に判断できることと、その関連が非線形である場合にも利用できることである。他方、その難点は説明変数の加重ベクトルを作成する際に付される規準化重みが乱数で与えられるために、満足な結果が得られるまでに試行錯誤的なグラフを作成する必要がある、しかも最適関連を示しているかどうかの判定基準がないことである。これらの難点はすでに回帰パラメータが推定され、その利用が回帰関係の図示だけに限定される場合には問題とならないであろう。

未知パラメータ β の推定の観点からみると、説明変数の行列 X や誤差の性質に影響をうけ、種々の問題が提起される。とくに、 X の「病的状態」の診断には説明変数間の相関とか、固有値のプロフィールが注目され、相関係数の確率プロット (例えば、Hills (1969)) やパイプロット表示 (例えば、Gabriel (1971),

Corsten & Gabriel (1976)) が提案されている。

いま、 $n \times p_0$ 行列 X の特異値分解はそのランクを r におさえるとき

$$X = \sum_{i=1}^r \lambda_i^\dagger p_i q_i^T = P \Lambda^\dagger Q, \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0 \quad (6)$$

と書き表される。ここに λ_i は $X^T X$ の第 i 固有値であり、 p_i は λ_i に対応する $X X^T$ の第 i 固有ベクトル、 q_i は λ_i に対応する $X^T X$ の第 i 固有ベクトルである。パイプロット表示では X がランク2で近似できるとき、 $H = (q_1, q_2)$, $G = (\lambda_1^\dagger p_1, \lambda_2^\dagger p_2)$ あるいは $H = n^\dagger (\lambda_1^\dagger q_1, \lambda_2^\dagger q_2)$, $G = n^\dagger (p_1, p_2)$ ととられ、前者の場合は G の n 個の要素の対 $(\lambda_1^\dagger p_{1i}, \lambda_2^\dagger p_{2i})$ ($i=1, 2, \dots, n$)、後者の場合は H の p 個の要素の対 $((\lambda_1/n)^\dagger q_{j1}, (\lambda_2/n)^\dagger q_{j2})$ ($j=1, \dots, p_0$) が対応して2次元座標上にプロットされる。

説明変数と応答変数の関連構造の探索には $n \times (p_0 + 1)$ 行列 $Z = (y : X)$ のパイプロット表示が考えられる。Gabriel (1978) はランク2行列

$$Z = [\hat{y} : (y^T \hat{y})^{-1} (\hat{y} y^T) X + X^*] = G H^T \quad (7)$$

のパイプロット表示が y と回帰変数の近似的な表示を与えることを示している。ここに、 X^* は

$$X^* = X - (y^T \hat{y})^{-1} (\hat{y} y^T) X \quad (8)$$

のランク1近似行列である。

また病的状態の診断とその対策が同時に行われる場合もある。 X が病的状態の場合、未知パラメータの最小二乗推定量は膨張して不安定となり、実地での解釈が不自然となる。 $X^T X$ のランクを p_0 とし、 β の最小二乗推定量は (6) 式の記法を用いて

$$\hat{\beta} = Q \hat{\gamma}, \quad Q = (q_1, q_2, \dots, q_{p_0}) \quad (9)$$

と得られる。 $\hat{\gamma}$ は $\hat{\beta}$ の直交成分よりなるベクトル $\gamma = Q^T \beta$ の最小二乗推定量である。 $\hat{\beta}$ の総平均平方誤差 (MSE; Mean Square Error) は $\sigma^2 (\lambda_1^{-1} + \lambda_2^{-1} + \dots + \lambda_{p_0}^{-1})$ であるから、固有値が極端に小さいことは MSE の膨張の因となる。 β のリッジ推定量は縮小因数、 $\delta_1, \delta_2, \dots, \delta_{p_0}$ ($0 \leq \delta_i \leq 1$) を用いて

$$\hat{\beta}^* = Q \Delta \hat{\gamma} = \sum_{i=1}^{p_0} \delta_i \gamma_i q_i, \quad \Delta = \text{diag}(\delta_1, \delta_2, \dots, \delta_{p_0}) \quad (10)$$

と定義される。このとき、 $\delta_i \gamma_i$ の平均平方誤差を最小にする δ_i は $\delta_i^* = \lambda_i (\lambda_i + \sigma^2 / \gamma_i^2)^{-1}$ で与えられる。Hoerl & Kennard (1970) はこの形式の類推より、 $\delta_i = \lambda_i (\lambda_i + k)^{-1}$ の形式のリッジ推定量

$$\hat{\beta}^* = (X^T X + kI)^{-1} X^T y \quad (11)$$

を考え、 k の選定について $k=0$ から $(\lambda_1+\lambda_2+\dots+\lambda_{p_0})/p_0$ までの範囲で $\hat{\beta}^*$ の p_0 個の成分をプロットし、それらが安定化する値を選ぶ「リッジ・トレース」を提案している。しかしながら、この k 上でのリッジ・トレースは完全直交な X が得られている場合でも、 k が大きくなれば $\hat{\beta}^*$ は益々安定になる傾向を示し、望ましくない。そのため、Obenchain (1977) は

$$m = p_0 - \delta_1 - \delta_2 - \dots - \delta_{p_0} \quad (12)$$

の上で $\hat{\beta}^*$ の p_0 個の成分をプロットすることを提案している。なお、 m は多重共線性許容度と呼ばれている。この m 上のトレースは、一般的な適用性、比較可能性、線形安定性およびランク落ちの解釈可能性などの良好な性質をもっている。

上記の点検あるいは診断の後に回帰モデルを修正したり、説明変数を選択してモデルを縮約する方法も数多く提案されている。変数選択のグラフィカルな基準としては、Mallows の C_p プロット (Mallows (1973), Gorman & Toman (1966)) がとくに有名である。 p ($\leq p_0$) 個の説明変数で形成される部分回帰モデル上の未知パラメータを β_p 、その最小二乗推定量を $\hat{\beta}_p$ で表すと、残差平方和は

$$RSS_p = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\beta}_p)^2 \quad (13)$$

で与えられ、このとき、Mallows の C_p 基準は

$$C_p = RSS_p / s^2 - n + 2p \quad (14)$$

で定義される。 C_p プロットは固定された p のあらゆる部分集合で得られる C_p の値を各 p ($1 \leq p \leq p_0$) に対してプロットする。このとき、直線 $C_p = p$ に近い布置をとる最小な p の点とそれに対応する部分集合、あるいは最小 C_p の点とそれに対応する部分集合が最良な説明変数の集合として選択される。ここに

$$C_p = p + \{RSS_p / s^2 - (n - p)\} = p + (\text{偏りの平方和部分}) \quad (15)$$

と表されることに注意すると、直線 $C_p = p$ より上側にある点ほど、それに対応する部分集合回帰は偏りをもつ。また、偏りの平方和がゼロに近いとき、

$$C_p - p = (n - p) \{s_p^2 / s^2 - 1\}, \quad s_p^2 = RSS_p / (n - p - 1) \quad (16)$$

となる。つまり、ランダムなバラツキが大きいと、 $s_p^2 / s^2 < 1$ 、つまり $C_p < p$ となる。いいかえれば、直線の下側はランダムな変動の大きさを示す。したがって、 $C_p = p$ に近い布置をとる点に対応する説明変数の部分集合が総体的に適合のよい回帰を与える。最小 C_p 基準に対応するものとして Hocking (1972) は基準

$$J_p = (n + p)s^2 / (n - p) \quad (17)$$

を提案している。この J_p 基準は Akaike (1974) の AIC 基準

$$AIC_p = n \ln(s^2) + 2p \quad (18)$$

と同等である。これらの基準に対するグラフィカル表示は C_p の場合と同様、1 から p_0 までの p の値に対してもプロットされる。

C_p プロットに関しては種々の変法が提案されている。Spjøtvoll (1977) は回帰係数の有意性を検定する通常の F 統計量が C_p 統計量を用いて

$$F_p = 1 + (C_p - p) / (n - p + 1) \quad (19)$$

で表されることに注目し、その有意確率 P_p の値が適切でない部分回帰 (集合) に対して小さくなることを示唆している。そして F_p 対 p および P_p 対 p のプロット (それぞれ F_p プロットと P_p プロット) を奨めている。また、Weisberg (1981) は C_p 統計量を

$$C_p = \sum_{i=1}^n \frac{(\hat{y}_i - \mathbf{x}_i^T \hat{\beta}_p)}{s^2} + v_{ii} - (u_{ii} - v_{ii}) = \sum_{i=1}^n C_{pi} \quad (20)$$

により、 n 例の対象に分解し、 C_{pi} 対 v_{ii} のプロット (C_{pi} プロット) を提案している。ここに、 v_{ii} と u_{ii} はそれぞれ行列

$$\begin{aligned} V &= X_1 (X_1^T X_1)^{-1} X_1^T \\ U &= (X_1 X_2) \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix} \end{aligned} \quad (21)$$

の対角成分であり、 X_1 は部分回帰 (集合) で組込まれる変数に対応する $n \times p$ 説明変数行列、 X_2 は残りの $n \times (p_0 - p)$ 説明変数行列である。この C_{pi} プロットから以降で述べるはずれ値や影響観測値の点検が行える。

変数選択基準の多くは推定 (または予測) 平均平方誤差の関数として与えられ、偏りの平方和と分散のトレードオフとなっている。Young (1982) は偏りの平方和と分散の加重和に基づく測度

$$H_p(w) = \sum_{i=1}^n [\mathbf{x}_i^T \beta - \mathbf{x}_i^T E(\hat{\beta}_p)]^2 / \sigma^2 + w \sum_{i=1}^n [\mathbf{x}_i^T E(\hat{\beta}_p) - \mathbf{x}_i^T \beta]^2 / \sigma^2 \quad (22)$$

を BIVAR (BIAs-VARiance) 測度と呼び、この測度に基づく変数選択基準を提案している。ここに w は非負の定数である。この基準は

$$C_p(w) = w RSS_p / s^2 + (1 + w)p - wn \quad (23)$$

で定義され、 w を固定したときに最小な $C_p(w)$ の値をとる部分集合が選択される。彼はまた他の多くの基

準がこの基準の特殊な場合となることを示している。とくに、

$$C_p(w) = wC_p + (1-w)p \quad (24)$$

である事実に注目すると、Mallows の C_p 統計量は $w=1$ のときに該当する。(24) 式は

$$C_p(w) = p + wB_p, \quad B_p = C_p - p \quad (25)$$

と書き直すことができる。したがって、 p 個の説明変数からなる部分集合に対する $C_p(w)$ の w に関するプロットは傾き $B_p = C_p - p$ の直線となる。このことから、 C_p プロット上で $C_p = p$ より下側に布置する点に対する直線は負の傾きをもち、上側に布置する点に対応する直線は正の傾きをもち、これらの直線の中で $C_p = p$ に最も近い点に対応する直線が最小の傾きをもつ。

この $C_p(w)$ プロットからは利用目的による最良部分集合の選定が可能である。すなわち、 $C_p(w)$ 基準が BIVAR 測度の一つの推定量であることから、0 に近い w の値に対して選択される部分集合は予測に最良であり、1 を起える大きな w の値に対して選択される部分集合は平均応答を推定するのに最良である。 $w=1$ を含む中間の区間で選択される部分集合は良好な予測と安定な平均応答を得る折衷的な目的に対して最良である。

4. 残差の点検； X の行方向からの診断

回帰の適合結果を診断する際に、最も普通に用いられるのが残差の点検である。表 1 には各種の残差の定義とそれらの関連を与える。これらの残差を視察するために各種のグラフィカル表現法が提案されている。その目的の大半が残差に潜む規則性、非正規性、不等分散性、はずれ値、などの点検にある。その典型は散布図と確率プロット（本紙の田崎・後藤を参照）の適用であり、そのプロフィールから規則性あるいは基本的な仮定からのズレが検出される。

通常、散布図では残差対予測応答、または残差対説明変数のプロットが行われる。一般に通常残差の場合、非規準化残差をプロットするよりも規準化残差をプロットする方が良いといわれている。これは非規準化残差が等分散性を満足しないのに対し、規準化残差は等分散性を満足するためである。偏残差対説明変数のプロットは Larsen & McCleary (1972) によって提案された。彼らは傾きゼロの直線のまわりにプロットする他の残差プロットよりも、傾き $\beta_i (i=1, \dots, p_0)$ の直線まわりにプロットされる偏残差プロットが非線形

性、はずれ値および不等分散性などとの関連で線形性の広がりや方向が評価しやすいことを示唆している。

あてはめた式が強い観測値の近くを通ることから、説明変数の粗悪な値は、通常残差プロットからは容易に検出されない。また、影響観測値を検出する統計量も多数提案されている (Cook (1977), Horglin & Welsch (1978), Andrews & Pregibon (1978), Belsley, Kuh & Welsch (1980), Atkinson (1981), 勘場・田中・後藤 (1982))。これらの統計量を表 2 に与える。Horglin & Welsch (1978) はハット行列の対角成分を幹葉表示 (Tukey (1977)) で提示し、Atkinson (1981) は彼が提案した統計量とジャックナイフ残差の半正規プロットを行い、そのプロット上にシミュレートしたデータから得られる範囲を付している。

粗悪な値の存在が応答変数と説明変数の両方で疑われるならば、表 1 と表 2 で与えた統計量の 2 個のプ

表 1. 各種残差とその関連

1) 非規準化残差 (通常残差)

$$r_i = y_i - \hat{y}_i, \quad \hat{y}_i = \mathbf{x}_i^T \hat{\beta}$$

2) 規準化残差

$$r'_i = \frac{r_i}{s\sqrt{1-h_i}}$$

$s^2: \sigma^2$ の不偏推定量

h_i : ハット行列の第 i 対角成分

$$(\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i)$$

3) 規準化ジャックナイフ残差 (交叉確認残差)

$$r_i^* = \frac{y_i - \mathbf{x}_i^T \hat{\beta}_{(i)}}{s_{(i)} \{1 + \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i\}^{\frac{1}{2}}} = \frac{r_i}{s_{(i)} \sqrt{1-h_i}}$$

$\hat{\beta}_{(i)}$: 第 i 観測値を除いたときの β の最小二乗推定量

$s_{(i)}^2: \hat{\beta}_{(i)}$ に対応する残差分散

$$(n-p-1)s_{(i)}^2 = (n-p)s^2 + r_i^2 / (1-h_i)$$

4) バランスドジャックナイフ残差

$$r_{bi}^* = r_i (1-nh_i)$$

5) アンバランスドジャックナイフ残差

$$r_{ui}^* = r_{bi}^* (1-h_i)^{-1}$$

6) 偏 残 差

$$r_{ij} = r_i + x_{ij} \hat{\beta}_j$$

表2. 影響観測値を検出するための統計量

1) Cook (1977) の統計量

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{ps^2} = \frac{r_i'^2 h_i}{p(1-h_i)}$$

2) Atkinson (1981) の統計量

$$T_i = \left\{ \left(\frac{n-p}{p} \right) \left(\frac{h_i}{1-h_i} \right) \right\}^{\frac{1}{2}} |r_i^*|$$

3) Andrews & Pregibon (1978) の統計量

$$AP_i = \left(1 - \frac{r_i'^2}{n-p} \right) (1-h_i)$$

4) Belsley, Kuh & Welsch (1980) の統計量

$$BKWi = \frac{r_i'^2 h_i}{1-h_i}$$

5) 勘場一田中一後藤 (1982) の統計量

$$KTGi = \frac{(\hat{\beta}_{(i)}^* - \hat{\beta})^T X^T X (\hat{\beta}_{(i)}^* - \hat{\beta})}{ps^2}$$

$\hat{\beta}_{(i)}^* : \beta$ の偏りのあるジャックナイフ推定量
(リッチ回帰推定量, 主成分回帰推定量
など)

ロットの視察が必要であろう。

5. コンピュータ・グラフィックス

最近における安価で高性能なデスクトップコンピュータ (あるいはミニコンピュータ) の出現は従来の方法論の展開と応用面での新しい方向と可能性を提供しつつある。とくに、デスクトップコンピュータの際立った特徴である優れたグラフィカル表示機能は2次元グラフから3次元グラフへ、さらには静的なグラフから動的なグラフへの展開を可能にしている。

Andrews (1981) はグラフィカル端末の特徴をうまく利用した新奇なグラフィカル表現法を与えている。その着想は2次元座標の回転に基づいている。一つの有用な例は数個のモデル間の特性表示である。

いま、残差 r と予測応答 $\hat{y}$ の通常の残差プロットを45度だけ時計と逆方向に回転させると新しい軸はそれぞれ $(2\hat{y}-y)$ と y になる。したがって、標準プロットでの $(\hat{y}, r)$ は新しい座標上で $(2\hat{y}-y, y)$ となる。このとき、残差が小さいことはプロットされる点が45度線上の直線に近いことを意味する。誤差の正規性の仮定の下で残差分散と対応する予測応答の分散の和は

一定である。したがって、残差分散を小さくする試みは予測応答の分散を大きくするだけであり、残差だけに焦点をあてることはモデルの過剰なあてはめに通じる。このような観点から、予測応答 y の標準誤差に比例した水平線をプロット点に加える。例えば、 $\hat{y}_U$ と $\hat{y}_L$ がそれぞれ予測応答 (期待値) の信頼区間を表すとするれば、これらの直線の端点は $(2\hat{y}_U - y, y)$ と $(2\hat{y}_L - y, y)$ となる。したがって、適切なモデルは45度線まわりの短い水平線の集りとして図示される。

2個のモデルの比較では3次元空間上の $(2\hat{y}_{U1} - y, y, 2\hat{y}_{U2} - y)$ と $(2\hat{y}_{L1} - y, y, 2\hat{y}_{L2} - y)$ を結ぶ線分を考える。ここに、 $\hat{y}_{Ui}$ と $\hat{y}_{Li}$ はモデル i における予測応答 (期待値) 信頼区間を示す。この3次元構造の表示はモデル1からモデル2までの連続回転表示によって得られる。したがってモデル間の相異は回転表示における動きの変化として捉えられる。Andrews (1981) はまた y の x への従属性および要因配置モデル上での平方和についての回転表示も与えている。

Andrews & Manchester (1982) は共変量が存在する場合に説明変数 x と応答変数 y の間の従属性を表示する分散共分散プロットを与えている。いま、 y と x の関連が $y = \alpha + \beta x$ で近似できるとする。このとき、 C を x と y の共分散、 V を x の標本分散とすると、 C が βV で近似できることを利用し、データを共変量に関して均質にグループ化した上で、グループ毎に C 対 V のプロットが考えられる。これが Andrews の CV プロットである。

6. 例示と解釈

ここでは回帰分析における二つの典型的な文献例に上述のグラフィカル諸法を適用し、回帰分析の直截的な解釈を与える。

6.1 Hald (1952) のセメントデータ

説明変数の選択問題あるいは回帰係数の調整問題に頻繁に用いられる例として13個の観測値からなる Hald (1952) のデータがある。説明変数はセメント合成に必要な4種の化学成分の構成比率(パーセント)であり、それらの和は理論的に100と想定される。応答はセメント合成に必要な熱量である。

図1から図5はこのデータに3節で記述した種々のグラフィカル表現法での結果を示している。連結ベクトルグラフとパイプロット表示から x_1 と x_2 が y に強く関連していることがうかがえる。また x_3 は x_1 で、 x_4 は x_2 でそれぞれ代用できることも読み取ることがで

きる。これらの解釈は C_p プロットと $C_p(w)$ プロットからも明らかである。 C_p プロットで $C_p=p$ に最も近いのは部分集合 (x_1, x_2) であり、この部分集合は $C_p(w)$ プロット上の広い範囲の w に対して最良であることを示している。最小 C_p 基準で部分集合 (x_1, x_2) の好敵手と考えられる部分集合 (x_1, x_2, x_3) と (x_1, x_2, x_4) は $C_p(w)$ プロットから $w > 1.5$ で最良である。したがって、これらの部分集合は安定な平均応答を得ることに興味がある場合には最良な部分集合といえる。

図1. 連結ベクトルグラフ：セメントデータ

図3. C_p プロット：セメントデータ

係数の安定性からみると、 m スケール上のリッジトレースで比較的安定なプロファイルを示すのは $\hat{\beta}_1^*$ と $\hat{\beta}_2^*$ である。また、すべての回帰係数が安定化するのは $m \approx 2$ のときである。したがって、このグラフからも説明変数 x_1 と x_2 が選択され、しかもこのデータはランク 2 で近似できることがわかる。

6.2 硝酸製造プラントの貯槽損失データ

表 3 に与えたデータは硝酸製造プラントにおけるアンモニアの貯槽損失率に関するデータであり、Brownlee (1965) によって最初に公表された。Daniel & Wood (1980) はこのデータの詳細な解析からはずれ値

図2. バイプロット表示：セメントデータ

図4. $C_p(w)$ プロット: セメントデータ 図5. リッジ回帰係数の m トレース: セメントデータ

の存在に注目し, Andrews (1974) はロバストな回帰分析の例示にこのデータを引用している。

図6はこのデータの通常残差を $\hat{y}$ に対してプロットしたものである。図から4個の観測値が他の観測値に比して異常に大きな残差をもつことが見られる。図7と図8にこれらの残差に関する全正規 $Q-Q$ プロットと半正規 $Q-Q$ プロットを示す。全正規プロットからは異常な残差が観測値21でしか見られないのに対し, 半正規プロットでは1, 3, 4, 21の観測値が異常とみなされる。このことはこのデータの残差が非対称に

表3. アンモニアの貯槽損失率データ

実験No.	空気流 x_1	温度 x_2	酸濃度 x_3	貯槽損失率 y
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

図6. 通常残差対 $\hat{y}$ のプロット: アンモニアの貯槽損失率データ

図7. 通常残差の全正規 Q-Q プロット：アンモニアの貯槽損失率データ

図9. AP 統計量の半正規 Q-Q プロット：アンモニアの貯槽損失率データ

図8. 通常残差の半正規 Q-Q プロット：アンモニアの貯槽損失率データ

ここで得られた通常残差や AP 統計量に見られる傾向は他の種々の残差や影響観測値の検出統計量でも同様に得られる。

7. 結 び

本稿では線形回帰モデルのあてはめ過程において、診断に有用な数個のグラフィカル表現法を平均的側面を指向する立場と個々の対象を問題にする立場に分けて提示した。回帰分析が適用される実地の場面では、適用目的を考慮した上でこれらの接近法が組合わされて利用されることが多いと考えられる。

グラフィカル表現法はそれ自体、データ解析における一つの有力な手段であり、定型的手法との併用でその威力はさらに増加する。ここに提示したグラフィカル表現法も定型的な回帰分析を一層強力なものにする。

参 考 文 献

- Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions On Automatic Control*, AC-19, 716-723.
- Andrews, D.F. (1979) *The robustness of residual displays*. *Robustness in Statistics*, edited by Launer, R.L. & Wilkinson, G.N., Academic Press. 19-32.
- Andrews, D.F. (1981) Dynamic displays for interactive statistical analysis. Presentation at Round Table Meeting of Japan Statistical

分布していることを示している。一般に残差に対する正規確率プロットは全正規プロットよりも半正規プロットの方が感度が良いといわれており、この例はその典型であるといえる。図9はCook (1977)のAP統計量を半正規プロットで示したものである。先の通常残差の確率プロットで見られなかった観測値1が影響観測値として浮き彫りされ、この結果はDaniel & Wood (1980)あるいはAndrews (1974)で得られた結果とも一致する。

- Society.
- Andrews, D.F. & Manchester, L. (1982) Variance covariance plots: Displays of dependence. Presentation at Pacific Area Statistical Conference.
- Andrews, D.F. & Pregibon, D. (1978) Finding outliers that matter. *J. Roy. Statist. Soc.*, **B40**, 85-93.
- Atkinson, A.C. (1981) Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.
- Atkinson, A.C. (1982) Regression diagnostics, transformations and constructed variables (with discussion). *J. Roy. Statist. Soc.*, **B44**, 1-36.
- Belsley, D.A., Kuh, E., & Welsch, R.E. (1980) *Regression Diagnostics*. New York Wiley.
- Brownlee, K.A. (1965) *Statistical Theory and Methodology in Science and Engineering*, 2nd Edn. John Wiley & Sons, New York.
- Cook, R.D. (1977) Detection of influential observation in linear regression. *Technometrics*, **19**, 15-18.
- Cook, R.D. (1979) Influential observations in linear regression. *J. Amer. Statist. Assoc.*, **74**, 169-174.
- Cook, R.D. & Weisberg, S. (1980) Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, **22**, 495-508.
- Corsten, L.C.A. & Gabriel, K.R. (1976) Graphical exploration in comparing variance matrices. *Biometrics*, **32**, 851-863.
- Daniel, C. & Wood, F. (1980) *Fitting Equations to Data*, 2nd Edn. John Wiley & Sons.
- Gabriel, K.R. (1971) The biplot display of matrices with application to principal component analysis. *Biometrika*, **58**, 458-467.
- Gabriel, K.R. (1978) Least squares approximation of matrices by additive and multiplicative models. *J. Roy. Statist. Soc.* **B40**, 186-196.
- Gorman, J.W. & Torman, R.J. (1966) Selection of variables for fitting equations to data. *Technometrics*, **8**, 27-51.
- Goto, M. (1981) Some devices in graphical data analyses. Invited Paper at the 43rd Session of the International Statistical Institute. Buenos Aires, Argentina.
- Hald, A. (1952) *Statistical Theory with Engineering Applications*. John Wiley & Sons.
- Hills, M. (1969) On looking at large correlation matrices. *Biometrika*, **56**, 249-253.
- Hoaglin, D.C. & Welsch, R.E. (1978) The hat matrix in regression and ANOVA. *Amer. Statistician*, **32**, 17-22.
- Hoerl, A.E. & Kennard, R.W. (1970) Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55-67.
- Hocking, R.R. (1972) Criteria for selection of a subset regression: which one should be used. *Technometrics*, **14**, 967-970.
- Larsen, W.A. & McCleary, S.J. (1972) The use of partial residual plots in regression analysis. *Technometrics*, **14**, 781-790.
- Mallows, C.L. (1973) Some comments on C_p . *Technometrics*, **15**, 661-675.
- Obenchain, R.L. (1977) Classical F-tests and confidence regions for ridge regression. *Technometrics*, **19**, 429-439.
- Spijøtvoll, E. (1977) Alternatives to plotting C_p in multiple regression. *Biometrika*, **64**, 1-8.
- Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison Wesley.
- Weisberg, S. (1981) A statistic for allocating C_p to individual cases. *Technometrics*, **23**, 27-31.
- Young, A.S. (1982) The bivar criterion for selecting regressors. *Technometrics*, **24**, 181-189.
- 勘場 貢, 田中浩光, 後藤昌司(1981) 回帰における影響力のある観測値の検出について. 日本品質管理学会, 第11回年次大会研究発表要旨.
- 田崎武信, 後藤昌司(1978) 回帰におけるJackknifing. 第46回日本統計学会発表要旨.
- 脇本和昌, 田栗正章(1974) 連結ベクトルパターンによる重相関度の表現. 日本統計学会誌, **5**, 1, 9-24.