安全基準としての閾値と ホッケー・スティック回帰法*

岡山理科大学　山　本　英　二
慶応義塾大学　竹　島　克　朗**
統計数理研究所　柳　本　武　美

(1977年7月　受付)

Threshold Values as Safe Levels and the Hockey Stick Regression Method

Eiji Yamamoto
(Okayama College of Science)
Katsuro Takeshima
(Keio University)
Takemi Yanagimoto
(The Institute of Statistical Mathematics)

The hockey stick regression method, a kind of two phase regression ones, is used to obtain estimators of safety levels. It is based on the hypothesis of the existence of physiological threshold values to chemical compounds. The purpose of this paper is to give remarks on the validity of using it, as well as to present statistical properties. The lower limit of a confidence interval is recommended to a value to estimate a safety level. Problems about selections among different regression models are studied.

Data related to the air quality standard of sulfer oxides in Japan and to Los Angeles student nurse study are exemplified.

1. 序

科学技術のめざましい進歩の中で，多種，多量の化学物質の人体への害作用が重要な問題になってきた．環境汚染，食品添加物等の安全性に典型的に現われている．害作用も発癌性，催奇型性，呼吸器症状の増加等と多岐に亘っている．

こうした害作用を防ぐために既存の物質あるいは新しく人体に摂取される物質に対し，行政当局によって規制値が設定されている．この規制値は行政上のいわゆる線引きとして，また一方人体に悪影響を及ぼさない，あるいは許容すべき安全基準に基づいて設定される．

安全基準の設定を行政の問題としてではなく，科学の問題として把えると，一つの過程として実験あるいは調査データから安全基準としての数値を導出する．この過程では諸々の問題が起るが，データ解析はその中で重要な一つの環である．本稿ではその一つの手法であるホッケー・スティック回帰法について考察する．

ホッケー・スティック回帰法は特に大気汚染物質の安全基準（環境基準）の設定の基礎資料

* この研究は昭和51年度特別事業「安全性の評価と安全基準に関する統計学的研究」の一環として行なわれた．
** 現在　清水建設．

を提供する方法として着目されている（たとえば [1]）．また現行のSOx（硫黄酸化物）の環境基準の設定にも利用された．そこで実際のデータに則して，ホッケー・スティック回帰法の性質，モデル設定の妥当性を調べた．

2節ではホッケー・スティック回帰法の概略について，3節ではその基礎となっている閾値について述べる．4節では2つの例に則してその性質を調べる．5節ではモデルの設定について考察し，関連した例が6節にあげられる．

2. ホッケー・スティック回帰

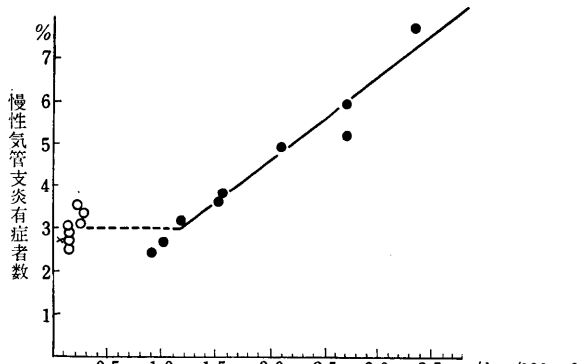
安全基準の設定の議論ではしばしば閾値の存在が仮定される．閾値については3節で改めて述べられる．生体に及ぼす刺激量がある一定の量以下であれば生体にはなんら害作用が現われないとみなし，その量を閾値，x_0 と呼ぶ．ホッケー・スティック回帰法では回帰線は

$$\begin{aligned} y &= \beta_0 & x &\leq x_0 \\ &= \beta_1 + \beta_2 x & x &> x_0 \end{aligned} \tag{1}$$

とする．データから x_0 を推定し，それを安全基準の目安にする．β_0 は当面対象にしている刺激以外で起る，いわゆる自然有病率を表わす．

回帰線を2つあるいは多くのフェイズに分ける方法は古くから使われていたと思われるが，安全基準との結びつきで調べられた研究として，[2]，[3] がある．それらの数学的な基礎は [4]，[5] である．

一方刺激と生体の反応を表わす，用量・反応関係としてはプロビット・モデルを代表として種々工夫されている．安全基準を求めるためのモデルをたてる場合には閾値の評価と扱いが重要な問題となるが [1]，ホッケー・スティック回帰法は明確にその存在を設定したモデルを採用するラフなモデルであるために余り良く研究されていないが，モデルの意味が明確であることから実際に使われている．



×：四日市非汚染地区調査結果（三重県立医大）
●：大阪市内調査結果（大阪府成人病センター）
○：赤穂市内調査結果

図1 慢性気管支炎有症者率と亜硝酸ガス濃度（PbO_2 法）

本稿で主として扱うデータは図1に示される例 [6] である．その例は SOx の環境基準の設定の基礎となったもので，日本の大気汚染の環境基準をめぐる解析では典型的な例である．横軸は SOx の平均濃度がとられ，縦軸は持続性せき・たんの有症者率である．持続性せき・たんは Fletcher の定義に基づいて疫学調査によって得られている．図1のデータの数値は表1である．ここで疫学調査の標本サイズが正確には不明であるが必要な場合には 2000 として扱った．

表1

亜硝酸ガス濃度 (mg/day/100 cm²)	0.21	0.28	0.27	0.15	0.15	0.14	0.13	0.14	3.4	2.75	2.75	2.1	1.6	1.55	1.15	1.0	0.9
有症者率 (%)	3.5	3.3	3.1	3.0	2.9	2.7	2.7	2.5	7.8	5.9	5.2	4.8	3.8	3.7	3.2	2.7	2.4

同時に上記の例の原型である [2] の例についても扱う．Los Angeles の看護学校の生徒に身体症状を毎日記入してもらい，その日の Ox（オキシダント）の最高濃度と比較している．データは表 2，図 2 に示される．

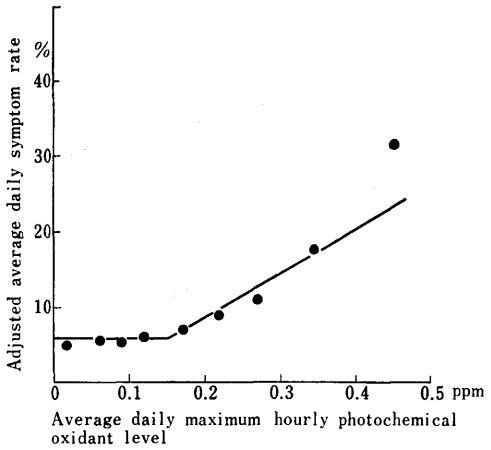


図 2　Ox 濃度と目の痛み訴症率

3. 閾値，その存在の仮定の妥当性

ここでいう閾値は生理学での定義である．生物の 1 器官，例えば筋肉に物理的あるいは化学的刺激，例えば電気刺激，を加えたとき，0-1 反応を起す．電気刺激を徐々に増加させたとき，初めは全く反応せず

表 2

		平均訴症率		
Ox 最大濃度（ppm）	日　数	平均報告件数	目のいたみ	胸部不快感
≤0.04	229	64	5.0	1.8
0.05-0.08	184	59	5.4	1.8
0.09	35	62	5.6	1.9
0.10-0.14	176	58	5.9	1.8
0.15-0.19	144	60	6.9	1.7
0.20-0.24	63	60	9.2	1.6
0.25-0.29	25	60	11.2	2.0
0.30-0.39	9	67	17.8	2.3
0.40-0.50	3	53	31.8	5.8

弛緩したまゝの筋肉が，あるレベル以上，このレベルを閾値と呼ぶ，加えられて初めて反応を起し，筋肉が収縮する（図 3）．それ以上の刺激を与えても同じ反応を示すにとどまる．

上の意味での閾値の存在は明らかであるが，我々が扱う例では人間の全身に対する害作用で

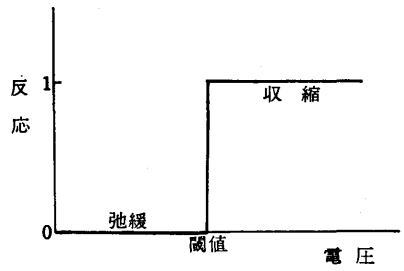


図 3　筋繊維の収縮

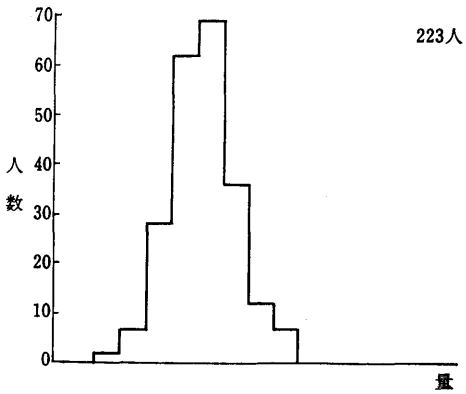


図 4　キニーネの閾値の分布 [7]

あり，また人間集団に対する閾値の存在が問われる．生体の器官は孤立してはいないから，その用量・反応関係は複雑になる．また個々人に対しては明確に閾値が仮定できる場合にも，生体には当然のことながら個体差が存在する．図4はキニーネの味覚の閾値の分布の例 [7] である．安全基準として閾値は集団の閾値の最下限，いわば閾値の閾値として取り扱われる必要がある．

ホッケー・スティック回帰法に限らず，統計的モデルをたてる場合には，そのモデルが現実的に妥当であるか否かに敏感でなければならない．安全基準をめぐる議論では実際的閾値をどのように理解するかがキー・ポイントになっている．厳密な意味での閾値の存在を仮定せず，閾値をより広く把え，操作的に用いられるモデルも考えられている [8] [9]．この場合閾値を回帰線の急増点，立ち起り点として漠然と理解する．

我々が主として扱う例では，目的変数となる有症者率の疾患が，赤痢のように原因と結果が1対1に対応するような特異的な疾患でなく，非特異的であるために複雑になる．非特異的な疾患では当面対象としている刺激以外によっても症状を呈する．この場合 (1) の x_0 はより把え難い．

閾値の存在の仮定が妥当か否かは本質的であるが，我々は一応この議論を保留して議論を進める．

4. ホッケー・スティック回帰法の適用と注意

2節で述べた2つの例について検証する．この節ではモデルの妥当性については余り問わない．モデルの妥当性の数理的な検証は次節以降で行なう．

4-1 モ デ ル

2つの例では次のモデルを与える．ここで x 軸は刺激としての大気汚染物質の濃度を，y 軸は有症者率を表わす．

$$y = \begin{cases} \beta_0 + \varepsilon & x \leq x_0 \\ \beta_1 + \beta_2 x + \varepsilon & x > x_0 \end{cases} \tag{2}$$

ただし

$$\beta_0 = \beta_1 + \beta_2 x_2$$
$$\varepsilon \sim N(0,\ \sigma^2)$$

誤差項は通常の回帰の場合と同様に分散が共通な正規分布に従うと仮定する．β_0 は自然有症率を表わし，x_0 が閾値を表わす．極めて簡単なモデルである．ホッケー・スティック回帰法の名は回帰線がホッケー競技の杖に似ていることからきている．

(2) のモデルの下で，データから最小2乗法，同時に最尤推定法によって母数を推定し，x_0 を推定する．しかしデータのとられた状況から，更に母数の推定方法が異なる．我々が扱っている問題の場合，図1に示したデータでは非汚染（低汚染）であると見なされた8地域と，汚染されていると見なされた9地域の違いがある．しかし図2の場合にはそのような区別はない．

4-2 予めデータを2群に分ける場合

(2) で回帰線が2つの部分から成っている．図1ではデータを両部分に対応して，非汚染地域と汚染地域に分ける．母数 β_0 と β_1, β_2 を別々に推定する．図1の例での致命的な欠陥はデータのばらつきが評価されず，区間推定がなされていないことである．

4-2-1 閾値 x_0 の推定量 $\hat{x}_0$

データ (x_1, y_1), $\cdots$, (x_m, y_m), (x_{m+1}, y_{m+1}), $\cdots$, (x_{m+n}, y_{m+n}) が得られ，前の m 個が (2) で x 軸に並行な回帰線に従い，後の n 個が直線回帰に従うとする．図1の場合，m

$=8$, $n=9$ である．このとき

$$\boldsymbol{y} = X_H \cdot (\beta_0,\ \beta_1,\ \beta_2)' + \boldsymbol{\varepsilon} \tag{3}$$

ただし

$$X_H = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & x_{m+1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{m+n} \end{pmatrix} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \\ y_{m+1} \\ \vdots \\ y_{m+n} \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \\ \varepsilon_{m+1} \\ \vdots \\ \varepsilon_{m+n} \end{pmatrix}$$

と表わされる．最小2乗法によって β_0, β_1, β_2 を推定し，$\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ とおく．x_0 の推定量 $\hat{x}_0$ は

$$\hat{\beta}_0 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \hat{x}_0 \tag{4}$$

で与えられる．$\hat{x}_0$ は

$$\hat{x}_0 = (\bar{y}_1 - \bar{y}_2) \cdot \frac{\sum_{j=1}^{n} (x_{m+j} - \bar{x}_2)^2}{\sum_{j=1}^{n} (x_{m+j} - \bar{x}_2)\ (y_{m+j} - \bar{y}_2)} + \bar{x}_2$$

ただし (5)

$$\bar{y}_1 = \frac{\sum_{i=1}^{m} y_i}{m},\quad \bar{y}_2 = \frac{\sum_{j=1}^{n} y_{m+j}}{n},\quad \bar{x}_2 = \frac{\sum_{j=1}^{n} x_{m+j}}{n}$$

予めデータを2群に分けることの妥当性が問われるが，図1のデータの場合は妥当であると思われる．データに伴なう事前の情報を利用している点で良いモデルと思われる．

4-2-2　推定量 $\hat{x}_0$ の分布

モデルの母数 β_0, β_1, β_2, σ_2 が既知であるとして，$\hat{x}_0$ の分布を求める．線型回帰の理論から

$$\begin{pmatrix} \hat{\beta}_0 - \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 - \beta_1 \\ \beta_2 \end{pmatrix},\ \sigma^2 \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} (X'_H X_H)^{-1} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}' \right) \tag{6}$$

が得られる．$\hat{x}_0 = (\hat{\beta}_0 - \hat{\beta}_1)/\hat{\beta}_2$ は正規分布と異なり，いわば，非心非心 t-分布（自由度1）とでも呼ばれる分布になって平均，分散は存在しない．

実際

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \begin{pmatrix} \sigma_1^2, & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2, & \sigma_2^2 \end{pmatrix} \right) \tag{7}$$

であるとき

$$Z = \frac{X_1}{X_2} \sim \frac{\sigma_1 \rho}{\sigma_2} + \frac{(\mu_1 - \frac{\sigma_1}{\sigma_2}\rho\mu_2) + \sigma_1\sqrt{1-\rho^2}\,N_1}{\mu_2 + \sigma_2 N_2} \tag{8}$$

ここで確率変数 N_1, N_2 は共に標準正規分布 $N(0,\ 1)$ に従い，互いに独立である．

Fieller [10] は，その密度関数の表現を与えている．また，$\hat{\beta}_2$ の変動係数が小さい時の近似的な密度関数と分布関数の表現を与えている．

1例として図1のデータから推定される $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}^2$ を真の値とした場合の $\hat{x}_0$ の分布の密度関数と分布関数は図5，図6で与えられる．勿論，母数は未知であるから，$\hat{x}_0$ の真の分布

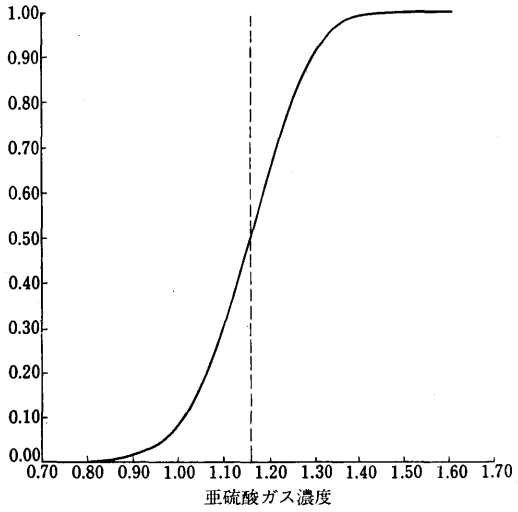


図5 分布関数

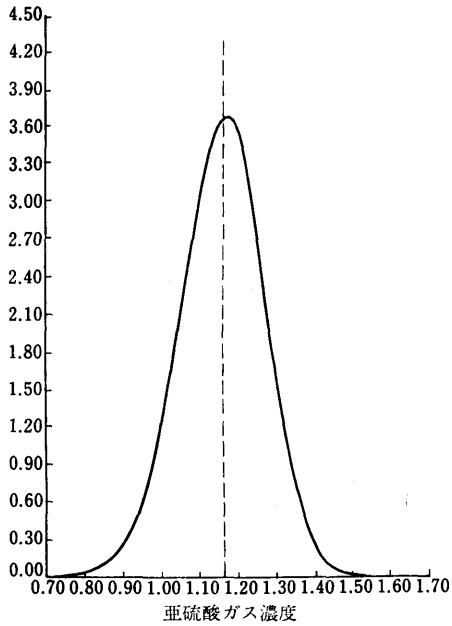


図6 密度関数

を与えているのではなく，単に視覚的に概略を示すだけである．数値計算では密度関数はFiellerの式から，分布関数は統計数値表 [11] の L 関数を用いた [12]．大略，正規分布に近く，単峰な分布をしているが，正規分布と比較すると非対称で，分布の裾が重い．

4-2-3 閾値 x_0 の区間推定

4-2-2でも示したように $\hat{x}_0$ は変動し，その分布の裾は重い．安全基準の目安として閾値 x_0 を推定する場合，点推定よりも区間推定の方が良い．区間推定を行い，安全サイドである下限値を目安とするのが妥当である．

回帰分析の逆推定にあたっているが，常法（例えば竹内 [13]）に従い信頼区間 $(\hat{x}_L, \hat{x}_U)$ を求めると，

$$\hat{x}_U,\ \hat{x}_L = \frac{(|\Sigma^{-1}|(\hat{\beta}_0-\hat{\beta}_1)\hat{\beta}_2 - \frac{mn}{m+n}\bar{x}_2\delta^2) \pm \sqrt{\delta^2|\Sigma^{-1}|(\hat{\beta}'\Sigma^{-1}\hat{\beta}-\delta^2)}}{|\Sigma^{-1}|\hat{\beta}_2^2 - \frac{mn}{m+n}\delta^2}$$

ただし

$$\Sigma = \begin{pmatrix}1 & -1 & 0\\ 0 & 0 & 1\end{pmatrix}(X'_H X_H)^{-1}\begin{pmatrix}1 & -1 & 0\\ 0 & 0 & 1\end{pmatrix}' \tag{9}$$

$$\delta^2 = F_{1,n+m-3;2\alpha}\cdot S_1/(m+n-3)$$

$$S_1 = (\boldsymbol{y}-X_H\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-X_H\hat{\boldsymbol{\beta}})$$

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_3)'$$

となり，図1の場合は表3になる．

水準の選択が残るが，いずれにせよ，$\hat{x}_0 = 1,160$ より小さな値を目安とするのが妥当である．

4-3 一般の場合

データを2群に分けられる場合は比較的特殊な場合である．一般の場合には説明変数 x_i を

順次2群に分けて解析する．図2のデータの場合には予め2群を設定する根拠はない．このデータでは大きさ868としても利用できる．これを適当にまとめた場合には標本の大きさの重みを加える．

表3　信頼区間の上限と下限

α（水準）	$\hat{x}_L$	$\hat{x}_U$
0.1	1.004	1.299
0.05	0.953	1.339
0.01	0.834	1.422

4-3-1　閾値 x_0 の推定量 $\hat{x}_0$

この場合閾値 x_0 の推定量 $\hat{x}_0$ は，ホッケー・スティック回帰線の中で残差平方和を最小にするものを探すことによって求められる．

x に順序を入れたデータのペアを $(x_1, y_1), \cdots, (x_N, y_N)$ とする．今，整数 k $(1 \leqq k \leqq N-1)$ に対して

$$\boldsymbol{y} = X^k{}_H (\beta_0, \beta_1, \beta_2)' + \boldsymbol{\varepsilon}$$

$$X^k{}_H = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & x_{k+1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_N \end{pmatrix} \tag{10}$$

とする．k を与えたとき，(10) から最小2乗法で $(\beta_0, \beta_1, \beta_2)$ を推定する．このとき制約条件

$$x_k \leqq \frac{\beta_0 - \beta_1}{\beta_2} < x_{k+1} \tag{11}$$

を付ける．各々の k に対し (11) を満足する最小2乗誤差を計算し，これら $N-1$ 通りの最小2乗誤差の中で最小のものを探す．得られた k, $\beta' = (\beta_0, \beta_1, \beta_2)$ の推定値 $\hat{k}$, $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ とする．即ち推定量は

$$(\boldsymbol{y} - \hat{X^k{}_H}\hat{\boldsymbol{\beta}})' (\boldsymbol{y} - \hat{X^k{}_H}\hat{\boldsymbol{\beta}}) = \underset{1 \leqq k \leqq N-1}{\text{Min}} \quad \underset{x_k \leqq \frac{\beta_0 - \beta_1}{\beta_2} < x_{k+1}}{\text{Min}} (\boldsymbol{y} - X^k{}_H \boldsymbol{\beta})' (\boldsymbol{y} - X^k{}_H \boldsymbol{\beta}) \tag{12}$$

を満たすよう選ぶ．x_0 の推定値 $\hat{x}_0$ は次式で与えられる．

$$\hat{x}_0 = (\hat{\beta}_0 - \hat{\beta}_1)/\hat{\beta}_2 \tag{13}$$

(12) は比較的簡単な計算によって解かれる (Hudson [4], Hasselblad ら [3])．

4-3-2　推定量 $\hat{x}_0$ の分布

$\hat{x}_0$ の分布は4-2の場合と異なり，制約条件 (11) が入り，順次最小2乗法を適用しているので，正確な分布は分からない．区間推定も正確には求められず，漸近的に正規分布に従う (Hinkley [5]) ことを用いて近似的に求める [3]．しかし大気汚染と疫学調査のような場合，漸近的な性質は，データの同質性を維持するために標本の大きさが限られるので，使い難い．図2の場合も同様である．より精密な区間推定の研究が望まれる．

5. モデルの選択

この節では，ホッケー・スティック回帰法の妥当性について数理的な面から検討する．モデル (2) は2つの仮定から成り立っている．1つは回帰線に対する仮定であり，他は誤差項に対する仮定である．データがモデルと適合するか否かの観点から，モデルの選択を行なう．

5-1　誤差項の評価

(2) では誤差として正規分布を仮定したが，直接的には2項分布が仮定される．汚染濃度 x のときの母有症者率を $p(x)$ として，出現確率 $p(x)$ の2項分布と見なされる．濃度 x_i での調査対象者の数を n_i とすると，有症者率 y_i の分散は $p_i=p(x_i)$ とすると，$p_i(1-p_i)/n_i$ である．このことは (2) で仮定された分散一定の仮定と矛盾する．しかし図1のデータでは n_i も p_i も余り差がないので近似的に分散一定とみなされ得る．

正規誤差は2項分布からの近似として扱われると共に，$p(x)$ が変動する場合にもロバストである．母有症者率 $p(x)$ が地域変動，調査誤差を伴なって，

$$p(x, a) = p(x) + \varepsilon \tag{14}$$

と表わされる場合である．ε は各々の調査地点で被調査者共通に作用する個別要因を表わす．実験に比べて，調査では管理が不十分だから母有症者率が x のみでは表現されないことが多い．(14) で仮定したとき，y_i の分散は

$$V(y_i) = p(x_i)(1-p(x_i))/n_i + V(\varepsilon)/(1-1/n_i) \tag{15}$$

で表わされる．(14) の仮定をおくと変動が大きくなる．

この2つのモデルをデータから比較するとき，一般に誤差は回帰線のずれをも含むので厄介である．6節で議論されるように，2項誤差を用いてデータが説明されるので，特に (14) で付加される誤差を入れる必要はなさそうである．4-2で扱ったモデルでの平均2乗誤差は 0.130×10^{-4} で，一方 $p=0.03$, $n=2,000$ とおくと $p(1-p)/n=0.145\times10^{-4}$ に比べて大きくはない．このことは図1のデータの質の良さを示している．

結局誤差としては正規分布よりも2項分布を仮定する方が自然になる．図1のデータのように n_i, p_i が余り変らず，(14) のような誤差がまぎれ込むときには，分散共通の正規分布が近似的に仮定される．

5-2 回帰線の選択

次に回帰線の選択を扱う．2節で扱った実質科学からのモデルの設定と共に，データからモデルを選択することは多くの示唆を与える．誤差としては (2) の分散共通の正規分布を仮定する．従って上の考察から実例としては図1のデータについてのみ行なう．先ず検定論の立場から比較する．

比較するモデルとして (3) の X_H の代りに

$$X_S = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{m+n} \end{pmatrix} \tag{16}$$

及び

$$X_B = \begin{pmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 0 & 0 \\ 0 & 0 & 1 & x_{m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{m+n} \end{pmatrix} \tag{17}$$

とおいた2つの場合を考える．前者は回帰線として直線を仮定した単回帰モデルである．後者は回帰線として折れ線を仮定している．以後折れ線回帰モデルと呼ぶ．

単回帰，ホッケー・スティック回帰，折れ線回帰で最小2乗法による残差平方和を各々 S_1, S_2, S_3 と書く．S_1/σ^2, S_2/σ^2, S_3/σ^2 は各々，自由度 $m+n-2$, $m+n-3$, $m+n-4$ の χ^2 分布

に従う．図1のデータでは $S_1=6.44\times10^{-4}$, $S_2=1.81\times10^{-4}$, $S_3=1.39\times10^{-4}$ である．未知母数の数が少ない簡単なモデルを帰無仮説とする2つの検定問題を考える．水準は $\alpha=0.05$ として議論する．

$$\text{検定問題 1}\quad \begin{matrix} H_0: \text{単回帰} \\ H_1: \text{ホッケー・スティック回帰} \end{matrix} \tag{18}$$

$$\text{検定問題 2}\quad \begin{matrix} H_0: \text{ホッケー・スティック回帰} \\ H_1: \text{折れ線回帰} \end{matrix} \tag{19}$$

検定問題 1，検定問題2に対して棄却域は検定量 $(S_1-S_2)/(S_2/(m+n-3))$ と $(S_2-S_3)/(S_3/(m+n-4))$ を用いて構成される．後者の検定量の分布は自由度 $(1,\ m+n-4)$ の F 分布である．前者の検定量の分布は F 分布ではない．特殊な分布で説明変数の値にも依存する [14] のでパーセント点を求めるとすれば，その時々計算する必要がある．

データについて計算すると検定問題1では棄却域を正確に計算するまでもなく，帰無仮説は棄却されると予想される．一方検定問題2では $(S_2-S_3)/(S_3/(m+n-4))=3.906$ で $F_{1,13,0.05}=4.667$ に近いが，帰無仮説は棄却されない．

5-3 AIC に基づくモデルの選択

文字通りデータからのモデルの選択の規準として赤池によって AIC に基づく方法が提案されている [15]．この方法は大標本に基づく近似的なモデルであるが，適用範囲は極めて広い．

AIC に基づく方法では，自由な母数の数を k とすると，異なるモデルの間で

$$\text{AIC}\ (k) = -2\log(\text{最大尤度}) + 2k \tag{20}$$

を最小にするモデルを選択する．この規準を (2) のような回帰モデルにあてはめると，$N=m+n$ とし，残差平方和を S とすると，

$$\text{AIC}\ (k) = N\log\frac{2\pi S}{N} + N + 2k \tag{21}$$

と表わされる．簡単のため，共通な定数を除いて

$$\text{AIC}'\ (k) = N\log S + 2k \tag{22}$$

とする．5-2 で扱った検定論に比べて，AIC (k) の分布を求める必要がないので簡便である．従って 4-3 の場合のような制限条件のある場合にも使える．また検定論の規準に比べて複雑なモデルを選好する．

5-2 での検定問題1では単回帰と，ホッケー・スティック回帰を比べた．AIC′ は $k=3$, $k=4$ だから，$\text{AIC}'(3)=N\log S_1+2\times3$ と $\text{AIC}'(4)=N\log S_2+2\times4$ を比較することになる．$N\log S_1/S_2-2$ の符号によって正ならばホッケー・スティック回帰，負ならば単回帰が選ばれる．

検定問題2では同様に，折れ線回帰では $\text{AIC}'(5)=N\log S_3+2\times5$ となる．やはり $N\log S_2/S_3-2$ の符号によって，正ならば折れ線回帰が，負ならばホッケー・スティック回帰が選択される．選択の規準を検定の基準に合わせて表現すると，

$$(S_2-S_3)/(S_3/(N-4)) \geq (e^{-2/N}-1)\ (N-4) \tag{23}$$

となる．検定の場合右辺は $F_{1,N-4;\alpha}$ であった．図1のデータにならって $N=17$ として左辺を計算すると約 1.623 でほぼ，$F_{1,13,0.15}$ にあたる．普通に用いられる検定の水準例えば $\alpha=0.05$ と比べると折れ線回帰が選択されやすくなる．実際 $\text{AIC}'(3)\fallingdotseq16.11$, $\text{AIC}'(4)=13.64$ となり折れ線回帰が選択される．

形式的にデータについて AIC′ を計算した結果が表4である．

ただし累積対数正規曲線は，$y=\beta_0+(1-\beta_0)\,\Phi\,(\beta_1+\beta_2\log x)$，$\Phi$ は標準正規分布の分布関

表 4 モデルと AIC

モデル	残差平方和	R	AIC′
定数	29.33 ($\times 10^{-4}$)	2	61.43($-4N\log 10$)
単回帰	6.44	3	37.66
ホッケー・スティック	1.81	4	18.11
2次曲線	1.94	4	19.28
累積対数正規	2.12	4	20.77
折れ線	1.39	5	15.64
3次曲線	1.94	5	21.25

数である．

実際データでは，これらのモデルの中では折れ線回帰がベストと判断される．しかし折れ線回帰線は，$y=2.23+3.98x$ と $y=0.76+1.90x$ となって $\hat{x}_0=-0.70$ となってしまう．この結果を防ぐには (10) の制約条件をつける必要がある．

5-4 2項誤差の下での解析

5節でのこれまでの考察によれば，数値的にみた場合誤差として2項分布を仮定した方が自然である．回帰線もなめらかで，concave な単調増加関数を仮定出来そうである．この仮定の下に簡単にホッケー・スティック回帰法を調べる．我々が扱っている問題では，汚染濃度の評価にも疑問がある [16] ので，細かい議論には立ち入らない．

あてはめる曲線としては5-3で用いた累積対数正規曲線を仮定する．誤差として2項分布を仮定するので，薬理学で用いられるプロビット解析になる．同時にホッケー・スティック回帰線をもあてはめた．

2つのデータについてあてはめた結果が図7，図8である．累積対数正規曲線のあてはまりが良い．適合度をみる χ^2-値は，プロビット解析でいう homogeneity の検定量，図7の場合ホッケー・スティック回帰線で 9.138(自由度 14)，累積対数正規曲線で 4.761(自由度 14) となって共に大変良くフィットしている．しかも後者の方が良くフィットしている．図8の場合

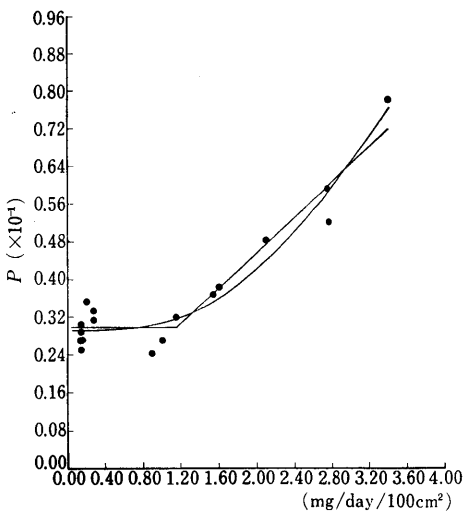


図 7 亜硫酸ガス濃度と慢性気管支炎有症者率

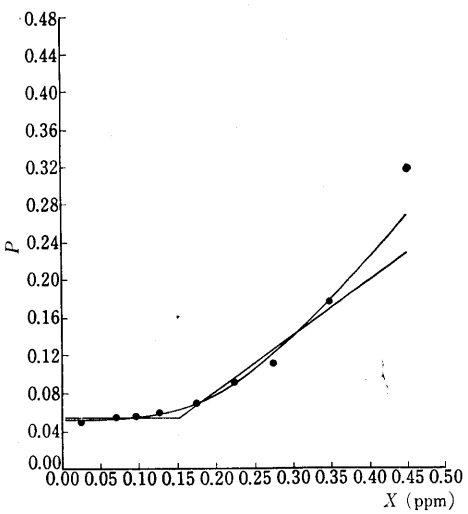


図 8 Ox 濃度と目の痛み訴症率

ホッケー・スティック回帰線で 21.534 (自由度 6)，累積対数正規曲線で 6.436 (自由度 6) である．この場合ホッケー・スティック回帰線は水準1%で棄却される．

図8の Los Angeles でのデータでは水準1%で棄却された．その理由は回帰線が妥当でなかったためであるとみられる．5-1 の誤差項の考察で扱った (14) による誤差も大きいとは考え難い．図をみても concave な曲線を想定されそうである．

累積対数正規曲線のようになめらかな曲線を仮定した場合，閾値の存在は否定され，無視されたモデルになる．安全基準の推定として解析する場合には閾値とは違ったアプローチが必要となる．それは曲線の立ち上り点，あるいは許容すべきリスクとなる．この問題は多くの人によって研究されている重要なテーマである [8]，[9]，[17]，[18].

6. 関連した例

ホッケー・スティック回帰法は閾値の存在の仮定の上に立っているだけに，特殊な手法である．しかし直観的なイメージがはっきりしているだけに，公表される例は多くないが，使用される例は多い．以下では関連した2つの例をあげる．最初の例は折れ線回帰線をあてはめ，その交点を漠然と相の違いとみなしている．

例 1.　牧野ら [19] は東京都目黒区の小学校，中学校，高等学校の児童を対象にして身体症状の調査を行なった．また同時に目黒区役所で測定された汚染物質の濃度を利用し，その関係を調べた．汚染濃度は PI と呼ばれる指数にまとめ，児童の訴症率 (症状を訴えた児童の比率) とが比較された．図9がその1例である．

それ以前の文献によって PI＝4 が閾値あるいは急増点とみなされていたので，それを検証するために，折れ線回帰線を引いている．ここでは予め2群には分けず，順次最小2乗法により2本の回帰線を引いて，残差最小を与える結果が図中の2本の線である．

2本の線が共に直線で，x 軸に平行でないので3節の閾値を仮定していない．閾値は広い概念として把えられている．この解析はデータのばらつきを無視した極めて直観的な例である．

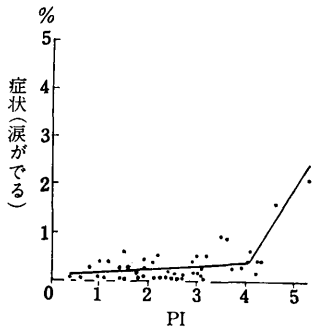


図 9

例 2.　法律の改正が行なわれて，飼料の安全性が要求される中で，吉田 [20] は飼料添加物への製品への移行，具体的には例えば鶏の飼料に添加した抗生物質の鶏卵への移行を論じた．飼料中の抗生物質の濃度が低いときは鶏卵中には検出されないことから，高濃度の飼料を与えて，鶏卵中の濃度を測定し，そのデータから回帰直線を推定し x 軸との交点 x_0 を，移行0の飼料含量と呼んでいる (図10)．そして信頼区間を求める必要があると述べている．

この例では $\beta_0=0$ が既知である場合のホッケー・スティック回帰とみなされる．しかしこの場

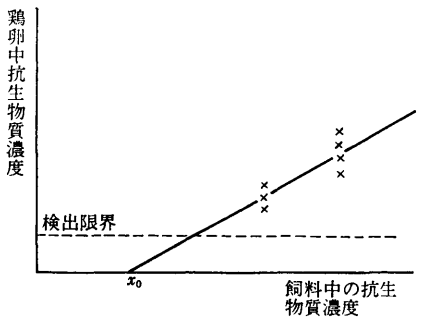


図 10

合も連続的に増加する上に concave な回帰線が想定出来そうである．実験した濃度にも依存するとみられる．x_0 を移行0の飼料含量と名づけるのも妥当ではない．操作的な目安とみなされる．

また x_0 より低い用量での測定値が検出限界以下となって，0か否かが分からない．このため解析が外挿を用いるため，より信頼性が悪くなる．

大塚 [21] は最近稲の登熟歩合を解析するために，折れ線回帰の適用について調べている．

謝　辞

研究の段階で清水忠彦氏（近畿大学，医）を初め多くの方から有益な助言を頂いたことに感謝します．またレフェリーの方々のコメントによって改善することができました．多くの計算は坂本淑子さん（統計数理研究所）によって手際良く行なわれました．併せてお礼します．

参　考　文　献

[1] 鈴木武夫 (1976) 一般公害に関する量一効果，反応，環境放射能研究会シンポジウム（線量一効果関係と閾値の問題）報告書 (KURRI-TR-147) 27-37.
[2] Hammer, D.I. et al (1974) Los Angeles student nurse study 'Daily symptom reporting and photochemical oxidents', *Arch. Environ. Health* **28**, 255–260.
[3] Hasselblad, V. et al (1974) Regression using "hockey stick" functions, Read before the Statistical Section, 101st Annual Meeting of the American Public Health Association, Nov. 8, 1973, San Francisco.
[4] Hudson, D.J. (1966) Fitting segmented curves whose join points have to estimated, *JASA* **61**, 1097–1129.
[5] Hinkley, D.V. (1969) Inference about the intersection in two-phase regression, *Biometrika* **56**, 495–504.
[6] 高田亘啓ら (1970) 赤穂市における大気汚染と慢性気管支炎について，兵庫県公害研究報告 第1号 25-34.
[7] Smith, S.E. et al (1973) *Variability in Human Drug Response*, Butlerworth.
[8] Hartley, H.O. et al (1977) Estimation of "safe doses" in carcinogenic experiment, *Biometrics* **33**, 1–30.
[9] Mantel, N. et al (1975) An improved Mantel-Bryan procedure for "safety" testing of carcinogens, *Canser Research* **35**, 865–872.
[10] Fieller, E.C. (1932) The distribution of the index in a normal bivariate population, *Biometrika* **24**, 428–441.
[11] 日本規格協会 (1972) 統計数値表.
[12] 竹島克朗 (1977) 安全性評価のための統計解析，慶応大学修士論文.
[13] 竹内 啓 (1966) 数理統計学，東洋経済新報社.
[14] 柳本武美 (1976) 折れ線回帰と単回帰，未発表.
[15] 赤池弘次 (1976) 情報量 AIC とは何か，数理科学 153号，5-11.
[16] 塚谷恒雄ら (1977) Dosage 解析と環境基準への応用，昭和52年度大気汚染研究全国大会予稿集.
[17] 竹内 啓 (1973) 許容基準の定め方，応用統計学 **3**, 1-8.
[18] 山本英二ら (1977) 安全基準のとしての閾値とホッケー・スティック回帰法の検討，昭和51年度統計数理研究所特別研究報告書，1-13.
[19] 牧野国義ら (1975) 光化学スモッグによる自覚症調査について，日本公衛誌 **22**, 421-430.
[20] 吉田 実 (1975) 抗生物質の畜産食品への残留，畜産の研究 **29**, 1号，3-9.
[21] 大塚雍雄ら (1976) 1ないし2の折曲点をもつ折れ線モデルのあてはめ，応用統計学 **5**, 1号，29-39.