ダミー変数を用いた線型回帰モデルとそのＡＩＣによる表層地盤におけるＳ波速度の推定

河合伸一*

Estimation of the Velocity of S Wave on the Surface Ground with Linear Regression Models Using Dummy Variables and Their Comparison by AIC

By

Shinichi KAWAI*

**National Research Institute for Earth Science and Disaster Prevention*

Abstract

The velocity of S wave is one of main indices characterizing the response of the surface ground at the time of earthquake. In this report, estimation of the velocity of S wave from other related indices in described.

As predictor variables, soil texture, depth and N-value are considered. These variables are selected appropreately in order to make linear regression models. In these models, there are two quantification indices, depth and N-value, and one qualification index, soil texture. When taking qualification variables into the model, it is desirable to use dummy variables. Fitting data to some models, and they are compared by AIC (Akaike Information Criterion).

Transformation of predictor variables and a response variable is possible when using a linear regression model. We have to consider the transformation of a response variable when using AIC.

As a result, the combination of depth and N-value as predictor variables has a good fit for the data. The models including soil texture are less fit than the others and t-tests also indicate that soil texture should not be included in the model.

Key words：the velocity of S wave, N-value, linear regression model, dummy variable, AIC, t-test

キーワード：S波速度，N値，線型回帰モデル，ダミー変数，AIC，t-検定

*防災科学技術研究所　先端解析技術研究部　数理解析研究室

1　はじめに

地震が起きたときの表層地盤の応答特性を知るためには，各地点の地層での地震波速度，特にS波伝播速度を知る必要がある．しかしながら，実際にS波速度を測るのは費用の点からもむずかしく，得られているデータも多くはない．一方，S波速度を他の比較的測定容易な地盤の指標から推定する試みが行われている（太田・後藤（1976））．これらの指標として，各地点におけるある深さでの土質，および地盤の硬さを表すN値と呼ばれる指標を使う．一般にN値が小さければ地盤は柔らかく，N値が大きければ地盤は硬い．この（深さ，土質，N値）の3つの指標から適当なものを選んで線型回帰モデルを作る．このモデルには説明変数として，質的指標（土質）と量的指標（深さ，N値）があるため，ダミー変数を用いた線型回帰モデルを作る必要がある．

本報告では，ダミー変数を用いた線型回帰モデルを幾つか作り，S波速度が測定してある地点のデータを使って各モデルに適用する．更にモデルどうしを比較するためにＡＩＣ（Akaike Information Criterion, 赤池情報量基準）を用いる．

2　ダミー変数を用いた線型回帰モデル

質的指標と量的指標の両方を使って線型回帰モデルを作る場合，各項は，目的変数の項と誤差項を除いて，定数項，量的指標のみを説明変数として用いる項，質的指標のみを説明変数として用いる項，量的指標と質的指標の両方を説明変数として用いる項の4つに分けることができる．また，線型モデルの「線型」とは，パラメータに関する線型性であるから，説明変数として，質的指標と量的指標をどのように組み合わせても良い．また，量的指標や目的変数も適当に変数変換して良い．

いま，量的指標の項目数を q，質的指標の項目数を r とする．各量的指標を $Z_1, \ldots, Z_q$ とする．i 番目の質的指標が m_i 個の属性に分かれているとする（$i=1, \ldots, r$）．ある観測値がこの i 番目の質的指標の j 番目の属性を持つとき，$n = q + \sum_{k=1}^{i-1} m_k + j$ として，$Z_n = 1$，属性を持たないとき，$Z_n = 0$ であるとする（$j = 1, \ldots, m_i$）．ここで，$Z_{q+1}, \ldots, Z_{q+m_1+\cdots m_r}$ をダミー変数という．つまり，1番目の質的指標は m_1 個の属性を持ち，これに対応するダミー変数は $Z_{q+1}, Z_{q-2}, \cdots, Z_{q+m_1}$ であり，2番目の質的指標は m_2 個の属性を持ち，これに対応するダミー変数は $Z_{q+m_1+1}, Z_{q+m_1+2}, \cdots, Z_{q+m_1+m_2}$ であり，…，r 番目の質的指標は m_r 個の属性を持ち，これに対応するダミー変数は $Z_{q+m_1+\cdots+m_{r-1}+1}, Z_{q+m_1+\cdots+m_{r-1}+2}, \cdots, Z_{q+m_1+\cdots+m_r}$ である．

目的変数 Y を変数変換したものを $\phi(Y)$ とし，これを新しい目的変数とする．変数変換を行わないときは $\phi(Y) = Y$ である．また，説明変数については，全ての指標を使う場合，一部の指標を使う場合，量的指標を変数変換する場合，交互作用を考える場合などいろいろな場合が考えられるので，ここでは各説明変数を $Z_1, \ldots, Z_{q+m_1+\cdots+m_r}$ の関数と考えることにす

る．つまり，説明変数の数を p とし，i 番目の説明変数を

$$X_i = X_i(Z_1, \ldots, Z_{q+m_1+\cdots+m_r}) \ (i = 1, \ldots, p)$$

とする．このとき線型回帰モデル

$$\phi(Y) = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

を考える．ここで，ϵ は平均 0，分散 σ^2 の正規分布に従うと仮定する．

(1) が定数項を含む場合は $X_1=1$ とすればよい．

(1) において量的指標のみを説明変数として用いる項がある場合は，通常の回帰モデルと同じように，例えば第 i 項がそのような項であるときは，$X_i = X_i(Z_1, \ldots, Z_q)$ とすれば良い．

(1) において質的指標のみを説明変数として用いる項がある場合は，ダミー変数を用いて説明される．

1つの質的指標を説明変数として用いるときは次のようにする．ある質的指標が m 個の属性に分かれているとする．このときダミー変数を $A_1, \ldots, A_m$，対応するパラメータを $\beta'_1, \ldots, \beta'_m$ とするとき，

$$\beta'_1 A_1 + \cdots + \beta'_m A_m$$

で説明される．実際は制約条件 $A_1+\cdots+A_m=1$ より，

$$\beta_1 A_1 + \cdots + \beta_{m-1} A_{m-1} + \beta'_m$$

で説明され，β'_m の項は定数項に含まれることになる．ここで，$\beta_i=\beta'_i-\beta'_m \ (i=1, \ldots, m-1)$ である．

2つの質的指標を説明変数として用いるときは次のようにする．まず，クロネッカー積を定義する．

いま，p 行 q 列の行列 A と r 行 s 列の行列 B があるとする．このとき A と B のクロネッカー積 $A \otimes B$ は

$$A \otimes B \equiv \begin{pmatrix} a_{11}B & \cdots & a_{1q}B \\ \vdots & & \vdots \\ a_{p1}B & \cdots & a_{pq}B \end{pmatrix}$$

で定義される $p \times r$ 行 $q \times s$ 列行列である．

ある2つの質的指標がそれぞれ m, n 個の属性に分かれているとする．このときダミー変数をそれぞれ $\boldsymbol{A}=(A_1, \ldots, A_m)^T$, $\boldsymbol{B}=(B_1, \ldots, B_n)^T$ とする．ここで ^{T} は行列の転置を表す．$l=mn$ とし，対応するパラメータを $\beta'=(\beta'_1, \ldots, \beta'_l)^T$ とするとき，

$$\beta'^T(\boldsymbol{A}\otimes\boldsymbol{B})$$

で説明される．実際は制約条件 $\sum_{i=1}^{m} A_i=\sum_{i=1}^{n} B_i=1$ より，$k=(m-1)(n-1)$, $\boldsymbol{A}^{(m)}=(A_1, \ldots, A_{m-1})^T$, $\boldsymbol{B}^{(n)}=(B_1, \ldots, B_{n-1})^T$, $\beta=(\beta_1, \ldots, \beta_k)^T$ として，

$$\beta^T(\boldsymbol{A}^{(m)}\otimes\boldsymbol{B}^{(n)})$$

で説明され，$\beta'_{k+1}, \ldots, \beta'_l$ の項は定数項に含まれることになる．3つ以上の質的指標を説明変数として用いるときも同様に説明される．

(1) において質的指標と量的指標の両方を説明変数として用いる項がある場合は，次のように説明される．

例えば，2つの質的指標と量的指標 $X=X(Z_1, \ldots, Z_q)$ を説明変数として用いるときは次のようにする．質的指標について，上と同じ記号を用いると，制約条件のもとで，

$$\beta^T(\boldsymbol{A}^{(m)}\otimes\boldsymbol{B}^{(n)})X$$

で説明され，$\beta'_{k+1}X, \ldots, \beta'_lX$ の項は X の1次の項に含まれることになる．

以上のことを考慮して，モデル (1) を作り，正規方程式を解いて $\beta=(\beta_1, \ldots, \beta_p)^T$ の最小二乗推定量 $\hat{\beta}=(\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ を求める．

3　目的変数の変換を考慮したＡＩＣの導出

(1) に属するモデルを幾つか考え，ＡＩＣ (Akaike Information Criterion, 赤池情報量基準) による比較を行う．ＡＩＣについては，赤池 (1976)，坂元・他 (1983) に詳しい．

標本の数を n とし，i 番目の標本に対する目的変数を Y_i, 説明変数を $x_{i1}, x_{i2}, \ldots x_{ip}$ とする ($i=1, \ldots, n$). 各標本に (1) を適用すると，

$$\phi(Y_i)=\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\epsilon_i \qquad (i=1, \ldots, n)$$

となる．ここで，$\epsilon_1, \ldots, \epsilon_n$ は独立に平均0, 分散 σ^2 の正規分布に従うとする．このとき，$Y_1, \ldots, Y_n$ の同時密度関数（尤度関数）は，

$$\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\sum_{i=1}^{n}\{\phi(y_i)-\sum_{j=1}^{p}\beta_ix_{ij}\}^2/(2\sigma^2)}\prod_{i=1}^{n}|\phi'(y_i)|$$

となる．$\beta=(\beta_1, \ldots, \beta_p)^T$, $\boldsymbol{y}=(y_1, \ldots, y_n)^T$ とすると，対数尤度 $\ln L(\beta, \sigma^2|\boldsymbol{y})$ は，

$$\ln L(\beta, \sigma^2 | \boldsymbol{y}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\phi(y_i) - \sum_{j=1}^{p}\beta_i x_{ij}\right\}^2 + \sum_{i=1}^{n}\ln|\phi'(y_i)|$$

となる．ここで，lnは自然対数である．$X=(x_{ij})_{1\leq i\leq n,\ 1\leq j\leq p}$，$\boldsymbol{W}=(\phi(Y_1), \ldots, \phi(Y_n))^T$ とし，β，σ^2 の最尤推定量を $\hat{\beta}$，$\hat{\sigma}^2$ とすると，

$$\hat{\beta} = (X^T X)^{-1} X^T \boldsymbol{W}$$
$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{W} - X\hat{\beta})^T(\boldsymbol{W} - X\hat{\beta})$$

である．$\hat{\beta}$ は β の最小二乗推定量に等しい．残差 $\boldsymbol{e}=(e_1, \ldots, e_n)^T$ とすると，$\boldsymbol{e} = \boldsymbol{W} - X\hat{\beta}$ より，

$$\hat{\sigma}^2 = \frac{1}{n}\boldsymbol{e}^T\boldsymbol{e} = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

となる．従ってＡＩＣは

$$\begin{aligned} AIC &= -2\ln L(\hat{\beta}, \hat{\sigma}^2 | \boldsymbol{y}) + 2(p+1) \\ &= n\ln 2\pi + n\ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + n - 2\sum_{i=1}^{n}\ln|\phi'(y_i)| + 2(p+1) \end{aligned} \qquad (2)$$

となる．なお，(2) の導出は，竹内 (1976) に詳しい．

4　S波速度推定への適用

実際にS波速度が測定してある地点のデータを使って各モデルに適用する．ここでは，千葉県浦安地区の7地点における441個の（深さ，土質，N値，S波速度）のデータの組を使用する．各地点における，深さに対する土質，N値，S波速度の散布図を図1に示す．この図において，ある地点におけるある深さでの土質，N値，S波速度の値の組が1個のデータとなる．

量的指標である深さ，N値をそれぞれ X_1，X_2 とする．ここで，深さについては，その大きさをとることにする．つまり，$X_1 > 0$ である．N値については $X_2 \geq 0$ である．また質的指標の土質については，土，砂，シルト，礫，粘土の5つの属性に分ける．従ってダミー変数は X_3，X_4，X_5，X_6 の4つである．また，定数項があると仮定する．定数項のパラメータを β_0 とし，$X_1, \ldots, X_6$ に対するパラメータを $\beta_1, \ldots, \beta_6$ とする．

いま，表1のような20個のモデルを考え，それぞれのモデルのパラメータ数 p（σ^2 は除く），決定係数 R^2（R は重相関係数（$R > 0$)），ＡＩＣを表す．ここで，$\bar{w} = \sum_{i=1}^{n} w_i/n$ として，

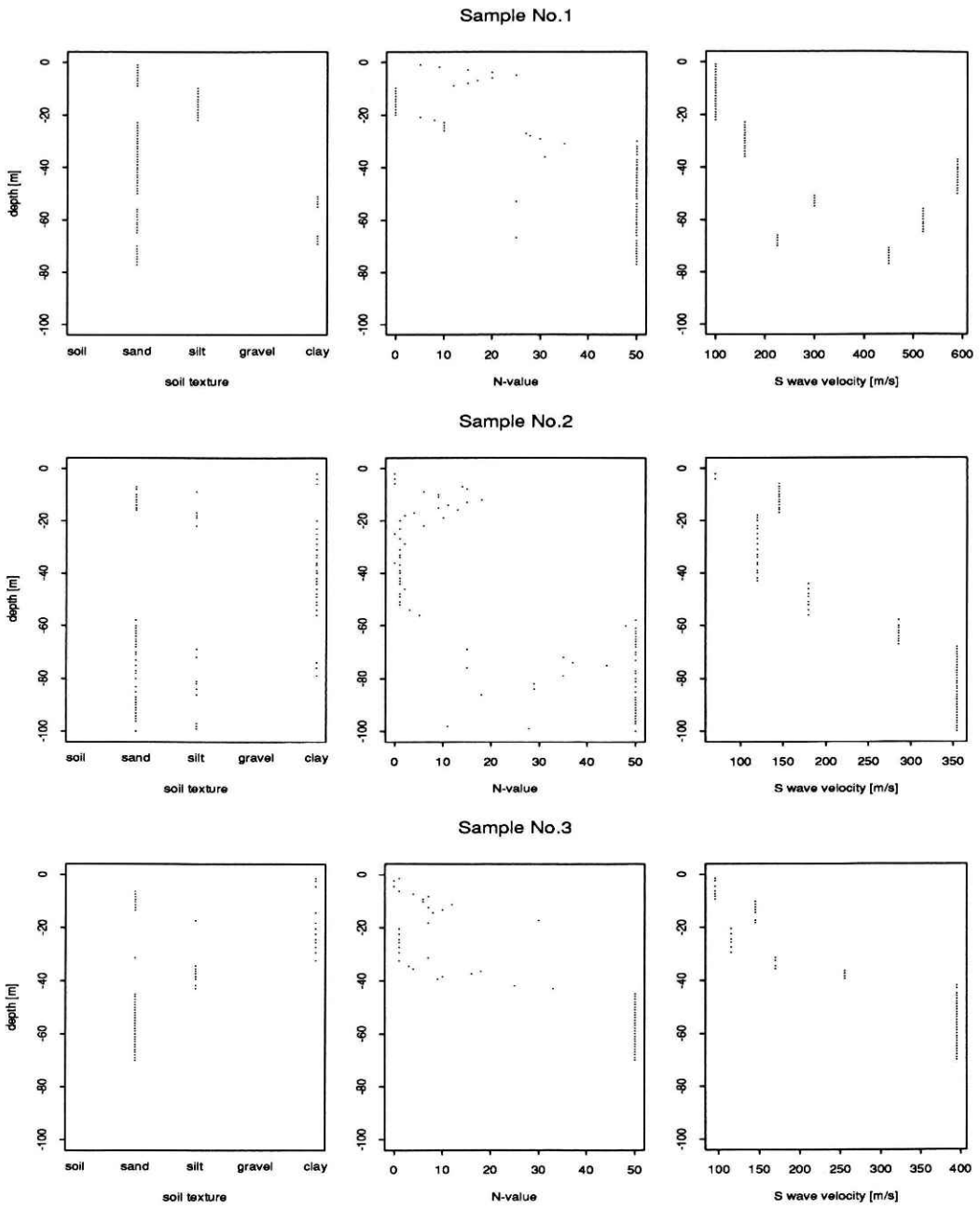

図1　　各地点における深さに対する土質，N値，S波速度の散布図.

Fig. 1　　Scatter plots of soil texture, N-value, S wave velocity for depth at each point.

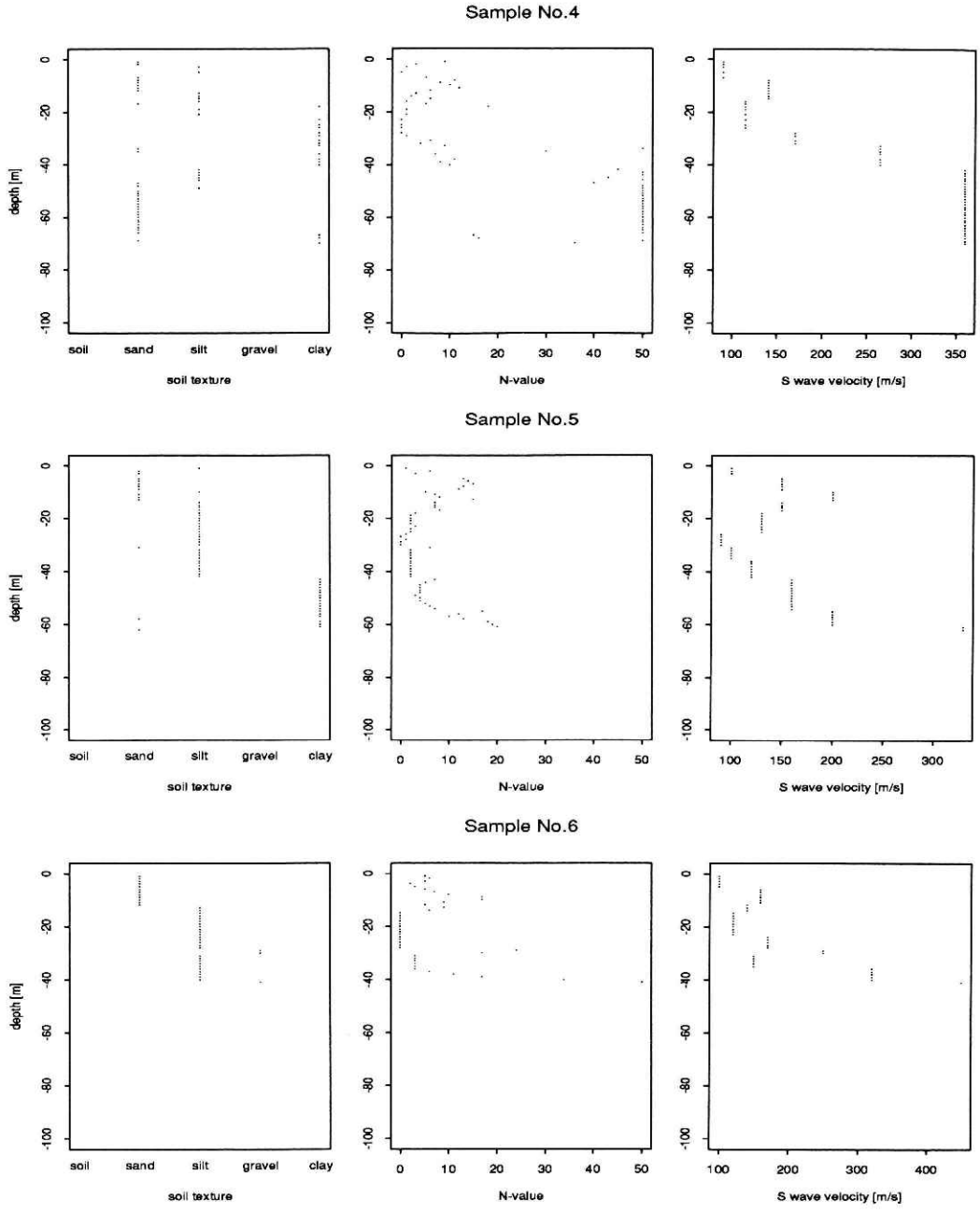

図1　各地点における深さに対する土質，N値，S波速度の散布図(続き)．

Fig. 1　Scatter plots of soil texture, N-value, S wave velocity for depth at each point (continued).

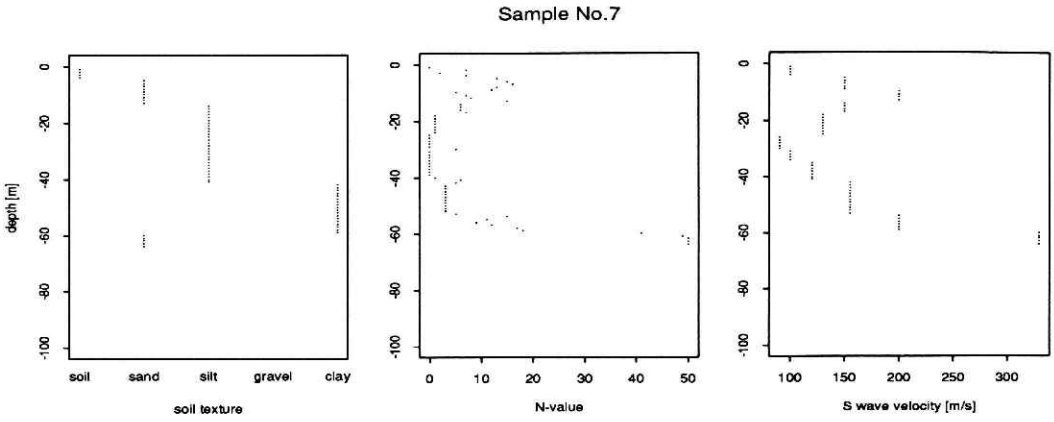

図1　各地点における深さに対する土質，N値，S波速度の散布図(続き)．

Fig. 1　Scatter plots of soil texture, N-value, S wave velocity for depth at each point (continued).

R^2 の定義は次の通りである．

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{w}_i - \bar{w})^2}{\sum_{i=1}^{n}(w_i - \bar{w})^2}$$

X_1, X_2 の自然対数をとるとき，$X_1 > 0$ であるので，そのまま $\ln X_1$ をとるが，X_2 は0の値をとることもあるので，$\ln(1+X_2)$ をとることにした．

各モデルの誤差項 ϵ は，平均が0の正規分布を仮定しているが，各モデルの残差が正規分布にあてはまっているかどうかを見る方法の1つに正規 $Q-Q$ プロットがある．もし残差が正規分布にあてはまっていれば，正規 $Q-Q$ プロットによって打たれた点は直線に近くなる．図2に各モデルの正規 $Q-Q$ プロットを示す．

いくつかのモデルを比較するとき，一般にAICの値が小さい方が良いモデルと言われている．表1でAICが最も小さいのはmodel 17である．また，深さとN値の組み合わせのモデルが全般にあてはまりが良い（model 4，11，17）．また，説明変数，目的変数に対数をとったほうがあてはまりが良い．

一方，質的指標である土質についてはどのモデルもあてはまりが悪い．model 7，14，20のAICの値は，それぞれのモデルから土質の項を除いたmodel 4，11，17より大きくなっている．model 20 の t -検定表を表2に示す．model 20はAICの値が2番目に小さいが，t -検定からは土質の項に対するパラメータを0とするという仮説を棄却できない．土質の項を含む他のモデルについても同じ結果がでている．従って，このままの属性の分け方では，土質をモデルに取り入れることはできない．

表 1　各モデルの比較.
Table 1　Comparison of each model.

No.	Model	p	R^2	AIC
1	$Y=\beta_0+\beta_1X_1+\epsilon$	2	46.97	5284.8
2	$Y=\beta_0+\beta_2X_2+\epsilon$	2	72.22	4999.8
3	$Y=\beta_0+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	5	22.44	5458.5
4	$Y=\beta_0+\beta_1X_1+\beta_2X_2+\epsilon$	3	74.45	4964.9
5	$Y=\beta_0+\beta_1X_1+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	6	61.87	5147.4
6	$Y=\beta_0+\beta_2X_2+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	6	73.18	4992.2
7	$Y=\beta_0+\beta_1X_1+\beta_2X_2+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	7	74.64	4969.5
8	$\ln Y=\beta_0+\beta_1X_1+\epsilon$	2	55.73	5014.3
9	$\ln Y=\beta_0+\beta_2X_2+\epsilon$	2	73.97	4780.1
10	$\ln Y=\beta_0+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	5	21.81	5271.2
11	$\ln Y=\beta_0+\beta_1X_1+\beta_2X_2+\epsilon$	3	78.94	4688.5
12	$\ln Y=\beta_0+\beta_1X_1+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	6	68.29	4875.2
13	$\ln Y=\beta_0+\beta_2X_2+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	6	76.16	4749.3
14	$\ln Y=\beta_0+\beta_1X_1+\beta_2X_2+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	7	79.24	4690.3
15	$\ln Y=\beta_0+\beta_1\ln X_1++\epsilon$	2	44.90	5110.8
16	$\ln Y=\beta_0+\beta_2\ln(1+X_2)+\epsilon$	2	67.53	4877.5
17	$\ln Y=\beta_0+\beta_1\ln X_1+\beta_2\ln(1+X_2)+\epsilon$	3	78.98	4687.9
18	$\ln Y=\beta_0+\beta_1\ln X_1+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	6	66.80	4895.4
19	$\ln Y=\beta_0+\beta_2\ln(1+X_2)+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	6	69.74	4854.5
20	$\ln Y=\beta_0+\beta_1\ln X_1+\beta_2\ln(1+X_2)+\sum_{i=3}^{6}\beta_iX_i+\epsilon$	7	79.33	4688.4

表 2　model 20のt-検定.
Table 2　t-test for model 20.

	推定値	標準誤差	t 値	$*Pr(>\|t\|)$
$\hat{\beta}_0$	3.8428	0.0566	67.8367	0.0000
$\hat{\beta}_1$	0.2458	0.0173	14.1932	0.0000
$\hat{\beta}_2$	0.2385	0.0147	16.2233	0.0000
$\hat{\beta}_3$	0.2174	0.1492	1.4571	0.1458
$\hat{\beta}_4$	0.0833	0.0410	2.0336	0.0426
$\hat{\beta}_5$	0.0509	0.0335	1.5196	0.1293
$\hat{\beta}_6$	0.2535	0.1357	1.8686	0.0624

* t 値の絶対値より大きくなる確率

従って 20 個のモデルの中で一番良いのは model 17 である．残差に対する正規 $Q-Q$ プロットは図 2 よりだいたい直線になっている．model 17 に対する t -検定表を表 3 に示す．model 17 の各データの予測値に対する観測値の散布図を図 3 (a)に示す．また，予測値に対する残差の絶対値の散布図を図 3 (b)に示す．図 3 (a)では観測値が 5.7(S 波速度に換算すると約 300 m/s) くらいのところまでは予測値が観測値のまわりをだいたい均等に分布している．

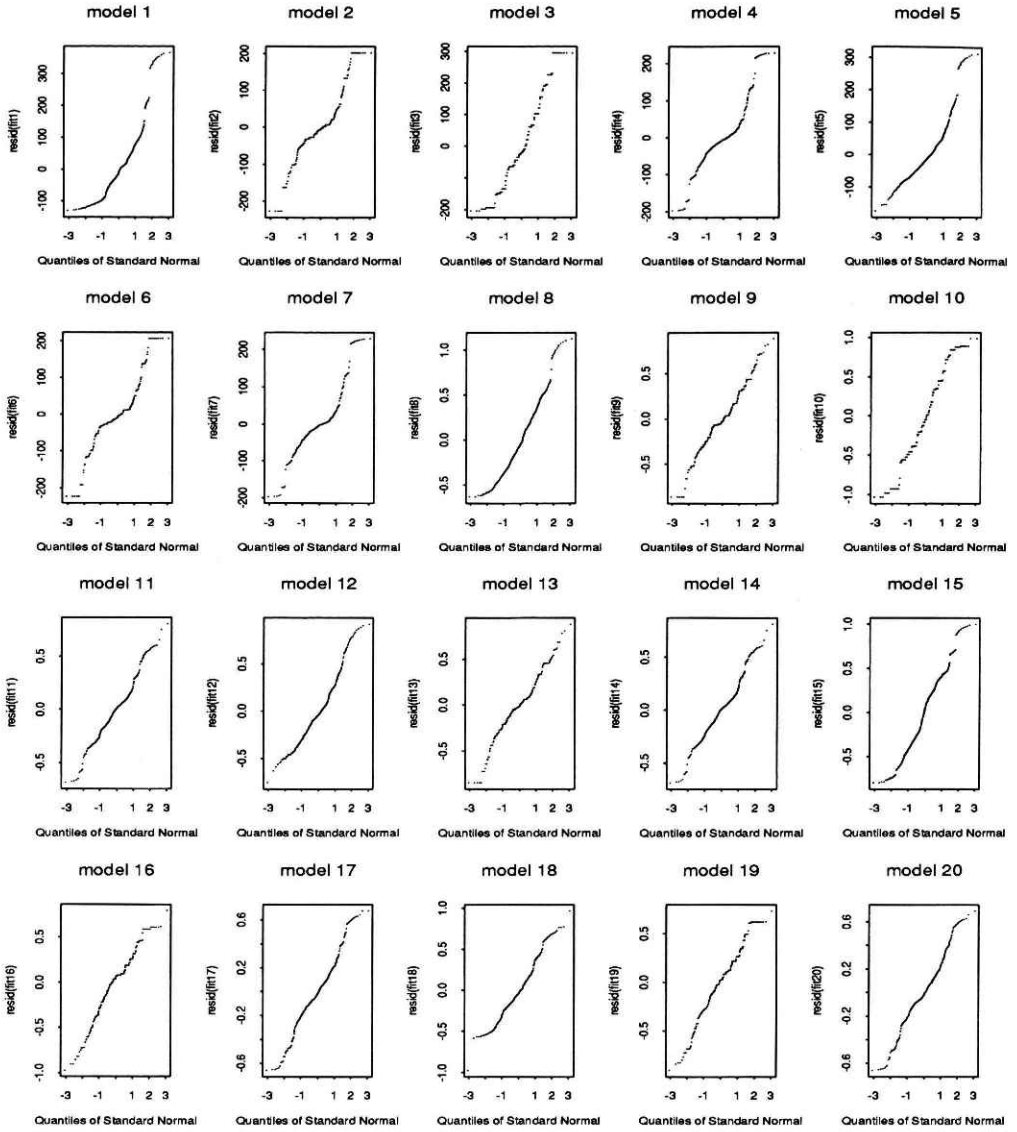

図2　各モデルの残差に対する正規Q-Qプロット．

Fig. 2　Normal probability plots of residuals for the fit of each model by Q-Q plot.

図3 (b)についても，予測値が5.7くらいのところまでは残差は均等に分布している．予測値が5.7より大きいところではやや偏りがある．よって，モデルを改良する余地が残っているようである．

ここで扱ったデータとmodel 17によるS波速度の推定式は

表3　model 17のt-検定.

Table 3　t-test for model 17.

	推定値	標準誤差	t 値	$*Pr(>\|t\|)$
$\hat{\beta}_0$	3.9290	0.0452	86.9840	0.0000
$\hat{\beta}_1$	0.2252	0.0146	15.4397	0.0000
$\hat{\beta}_2$	0.2556	0.0096	26.6449	0.0000

$*t$ 値の絶対値より大きくなる確率.

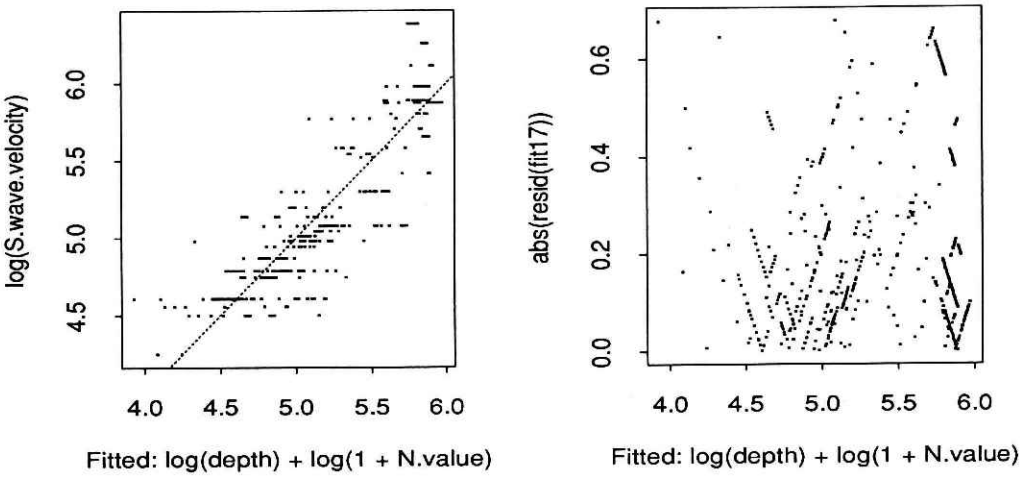

図3　model 17における各データの(a)予測値に対する観測値の散布図，(b)予測値に対する残差の絶対値の散布図.

Fig. 3　Scatter plots of (a) observed values and (b) absolute values of residuals versus predicted values for the data with model 17.

$$\ln \hat{Y} = 3.9290 + 0.2252X_1 + 0.2556\ln(1 + X_2)$$

である.

5　おわりに

S波速度を他の地盤の指標から推定する線型回帰モデルの開発を試みた．まず，ダミー変数を含む線型回帰モデルをいくつか作成し，AICによるモデルの比較を行った．また，各モデルの残差について正規 $Q-Q$ プロットを行い，誤差の正規性を調べた．さらに，各モデルについて t-検定を行い，どの指標を説明変数として選ぶとよいかを調べた．その結果，AICによって，目的変数，量的変数に自然対数をとったものがあてはまりがよいことが示された．また，t-検定によって，土質をモデルに取り入れることに否定的な結果が出された．

最終的に選ばれたモデルはAICが最小のものに一致したが，モデルの改良の余地はまだあるようである．

実際には，適用したデータの地点数が少なく，これだけからモデルを決定するのは難しいことである．しかしながら，ある程度の傾向は出ているようである．

ここで適用した方法は，量的指標と質的指標を含むどんなデータにも適用できる．今後も，様々なデータについて，線型回帰モデルの適用の可能性を探っていきたい．

謝　辞

本研究を行うに際し，筑波大学数学系の赤平昌文教授には，データ解析手法について有用な助言を頂きました．ここに深く感謝いたします．

参考文献

1) 赤池弘次（1976）：情報量基準ＡＩＣとは何か．数理科学，**153**, 5-11.
2) 太田　裕・後藤典俊（1976）：Ｓ波速度を他の土質的指標から推定する試み．物理探鉱，第24巻．
3) 坂元慶行・石黒真木夫・北川源四郎（1983）：情報量統計学，共立出版．
4) 竹内　啓（1976）：情報統計量の分布とモデルの適切さの基準．数理科学，**153**, 12-18.