ブートストラップ法の回帰モデル安全管理への応用について

熊本大学工学部　学生員　○白岩　正憲
熊本大学工学部　正　員　小林　一郎
熊本大学工学部　正　員　三池　亮次

1．はじめに　アーチダムのたわみδがクラウンに沿う2、3の標高における堤体温度t_i、温度勾配α_i(i=1,2,..)と貯水池水位$h-h_0$の要因に支配されて、次式のような線形回帰モデルが設定されるものとする。

$$\delta = k_0 + \Sigma a_i t_i + \Sigma b_i \alpha_i + \Sigma e_i (h-h_0)^i + e \qquad (1)$$

ただし、eは偏差である。このような重回帰モデルにおいて、その回帰の構造の経年変化を推定する方法として、さきに区間推定による方法と差の検定による方法を提案した[1)]。前者の方法では、安全性の判断が即刻可能であるが、偏差eが正規分布に従わないとき、管理限界の統計学的意義が不明確となる。後者においては、偏差が正規分布に従わなくても回帰関数の推定値$\{\hat{\mu}\} = [X]\{\hat{\beta}\}$は中心極限定理に従って、データサイズが十分に大きいとき、正規分布に近づくので、より正確な管理限界を得ることを指摘した。後者の差の検定による手法において偏差eが正規分布に従わないとき、どの程度のデータサイズに対して回帰係数$\{\beta\}$や回帰関数$\{\mu\}$の正規性、また不偏分散Se/σ^2のχ^2分布への収束性が保証されるかを検証することは興味あることである。

2．単回帰モデルへのブートストラップ法の適用　記号の説明をかねて次のように単回帰モデルを設定する。

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (2)$$

ただし、eは次式に従うものとする。

$$E[e_i] = 0 \ , \quad V[e_i] = \sigma^2 \qquad (3)$$

回帰係数β_0、β_1の推定値$\hat{\beta}_0$、$\hat{\beta}_1$　の不偏分散の平方根(RMS)をε_0、ε_1とすると

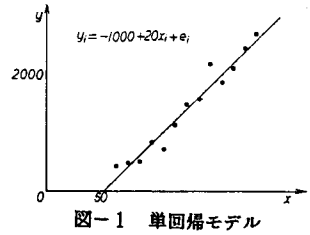
$$\varepsilon_0 = \sqrt{((1/n)+(x_m/S_{11})) \times (S_E/(n-2))} \qquad (4)$$

$$\varepsilon_1 = \sqrt{(1/S_{11}) \times (S_E/(n-2))} \qquad (5)$$

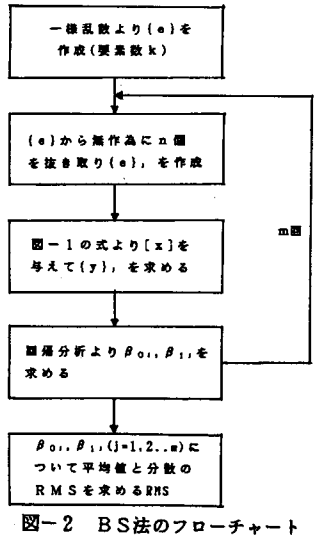
ここで、x_m, y_m は x_i および y_i の算術平均値で、n-2は自由度である。また、S_{11}とS_Eは次の通りである。

$$S_{11} = \Sigma (x_i - x_m)^2 \qquad (6)$$

$$S_E = \Sigma (y_i - y_m)^2 - \beta_1 \Sigma (x_i - x_m)(y_i - y_m) \qquad (7)$$



図－1　単回帰モデル



図－2　BS法のフローチャート

ブートストラップ(BS)法は、1つのサンプルのデータから母集団の分散や相関係数といった値の確からしさをコンピュータによる大量の数値計算によって評価しようとするものである。図－1の単回帰モデルを用いてBS法の適用について述べるが、図－2はその概要を示したものである。説明変数x_iと従属変量y_i, i=1,2,...nの各測定の組(x_i, y_i)について多くのコピーを作る。その中からn組のデータをat random に取り出し、回帰係数$\{\beta\}$を求める。Efron は、これをm回だけ繰り返し$\{\beta\}$の度数分布を調べると、中央の68%の幅の1/2の値σ_Bが母集団の$\{\beta\}$のRMSの値と極めて良く一致することを示している[2)]。ただし、データのサンプル数およびBS法の繰り返し回数の取り方については明確な規準が示されていないので以下に数値実験を行った結果を示す。

3．数値計算例　回帰分析によって求まる式(4)、(5)のRMSの値と

ＢＳ法で得られた値を比較し、データサンプルの数ｎとＢＳ法の繰り返し回数ｍについて検討する。

式(2)の偏差 e_i について、次の２通りのデータを用いる。

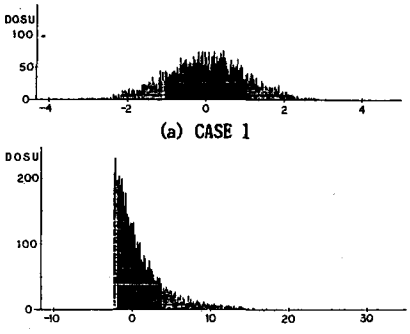
CASE 1 --- e_i は N[0,1] の正規分布に従う。

CASE 2 --- e_i は $e_i = \alpha \log_e(1-u_i) - \alpha \log_e 2$ の指数分布に従う。ただし、$\alpha = 10/3$ とし、u_i は一様乱数である。

表－1

単回帰モデルのデータ(CASE 2,n=5)

i	x_i	y_i	e_i
1	53.522	40.482	0.0340
2	56.991	143.902	4.0696
3	52.907	56.557	-1.5750
4	55.291	105.215	-0.6016
5	50.592	9.872	-1.9625

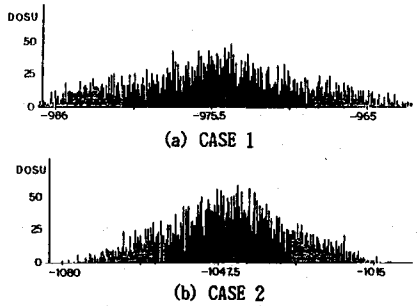


(a) CASE 1

(b) CASE 2

図－３　誤差の分布形

データのサンプル数は n = 5,10,15,20,50 の５通りとした。表－１はn=5 のデータである。図－３(a)、(b)はCASE-1およびCASE-2の e_i の度数分布を示したものである。また、図－４、５はCASE-1、CASE-2についての回帰係数 β_0 の度数分布である。n=5の場合は両ケースともまだ正規分布と言えないが、 n=20の場合は正規分布になっており、中心極限定理が数値的に確かめられたといえる。表－２、３は各回帰係数のＲＭＳの値を比較したものであり、n=15以上では回帰分析による値とＢＳ法による値はほぼ等しい。



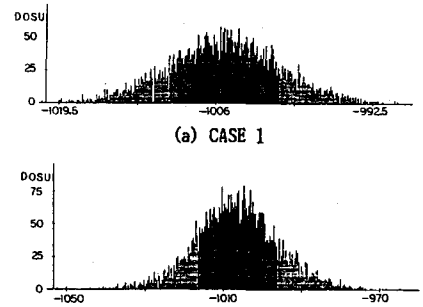
(a) CASE 1

(b) CASE 2

図－４　回帰係数 β_0 の度数分布(n=5)

ＢＳ法の繰り返し回数ｍについてのデータは割愛したがｍ=1000以上であれば結果に大差ないようである。

以上の結果より、データのサンプル数が10以下の場合は平均値、分散といった統計量について母集団の値を正く推定しているとはいい難い。しかし、逆にサンプル数が20以上あれば、偏差ｅがCASE-2のように正規分布に従わなくても、中心極限定理が成立し回帰係数の推定値{β}は正規性があるといえる。このため、筆者らが文献1)で提案した差の検定法は、データのサンプル数が20以上あれば信頼できる結果が得られるといえる。今回は単回帰モデルについての結果をまとめたが、さらに重回帰モデルについてもブートストラップ法による同様の検討を行う予定である。



(a) CASE 1

(b) CASE 2

図－５　回帰係数 β_0 の度数分布(n=20)

表－２　回帰係数のRMS(CASE 1)

	回帰分析		ＢＳ法			
n	ε_0	ε_1	ε_0^*	ε_1^*	$\varepsilon_0/\varepsilon_0^*$	$\varepsilon_1/\varepsilon_1^*$
5	7.3658	0.1366	5.5715	0.1034	1.3220	1.3213
10	6.9360	0.1291	6.0514	0.1126	1.1462	1.1465
15	5.0074	0.0932	4.6004	0.0857	1.0885	0.1089
20	5.3590	0.0993	5.1103	0.0947	1.0487	1.0488
50	2.9022	0.0531	2.8629	0.0524	1.0137	1.0134

表－３　回帰係数のRMS(CASE 2)

	回帰分析		ＢＳ法			
n	ε_0	ε_1	ε_0^*	ε_1^*	$\varepsilon_0/\varepsilon_0^*$	$\varepsilon_0/\varepsilon_1^*$
5	16.5171	0.3064	12.4475	0.2311	1.3269	1.3256
10	16.6352	0.3096	14.8436	0.2763	1.1207	1.1207
15	15.8648	0.2953	14.5665	0.2711	1.0891	1.0892
20	11.2191	0.2079	10.5207	0.1949	1.0664	1.0669
50	7.2903	0.1334	7.2856	0.1333	1.0006	1.0011

参考文献 1)Miike,Kobayashi:Safety Control of Dams by Multivariate Regression Model,Proc. of ICOSSAR 1986 2)Efron:Censored Data and the Bootstrap,Jour. of American Statistical Association,1981