線型回帰分析における観測誤差の影響と誤差の評価について

三 枝 義 清

一 問 題

線型回帰モデルを実際に推定しようとする場合、考慮すべき説明変数のなかには観測不能のものや、利用しうるデータがないものが出てくる。それらの変数を説明変数として回帰モデルの中に組み入れようとすれば、観測可能な別の変数で代理させねばなるまい。たとえば期待価格や期待収量を説明変数にした農産物の供給分析では、これらの変数を過去数ヵ年の実現値の加重平均値で代用させるだろう。農産物の価格指数や気象指数を説明変数にしようとする場合──これらの変数は観測不能ではないが──適当な統計データがないために、入手可能な別の指数で代用することがしばしば行なわれる。厳密に考えれば、トレンド項を除けば大半の説明変数は真の変数に対する代

理変数（proxy variable）とみなければならない。多くの場合、作付面積や収量に関する統計データは標本誤差や回答誤差などの観測誤差を含むものであるから、われわれの利用するこれらの観測値も true variable の proxy variable とみることができる。proxy variable と true variable の関係はさまざまで、

proxy variable＝true variable＋observation error ……………………（1・1）

という簡単なものもあるし、期待価格の場合のように複雑な関係をもつものもある。

回帰分析で説明変数に proxy variable を使った場合、推定されたパラメーターは真のパラメーター（true variable を説明変数にした時の）とどのような関係をもつだろうか。二の㈠では proxy variable を確定変数とみなした場合を取り扱い、二の㈡では proxy variable が一個で、特に（1・1）のような関係がある場合についてパラメーターの推定値の分布を論じた。三は観測誤差を評価する際の一つの試みを述べたものである。

二　線型回帰分析における観測誤差の問題

従属変数 y は次のような線型回帰式にしたがうものとする。

$$y_j=\sum_{i=1}^{k}\beta_i z_{ji}+u_j \quad \cdots\cdots\cdots\cdots (2\cdot 1)$$

ただし，$j=1,2,\cdots,n$

説明変数はすべて確定変数で、誤差項 u_i は

$$E(u_j)=0 \quad E(u_j^2)=\sigma_2^2 \quad E(u_j u_{j'})=0 \quad (j\neq j')$$

とする。2・1式を行列形式にして次のように表わすことにする。

$$y=\sum_{i=1}^{k}\beta_i z_i+u \qquad \cdots\cdots (2\cdot 2)$$

あるいは

$$y=Z\beta+u \qquad \cdots\cdots (2\cdot 3)$$

ただし
$$y=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad z_i=\begin{pmatrix} z_{1i} \\ \vdots \\ z_{ni} \end{pmatrix} \qquad i=1,2,\cdots,k$$

$$u=\begin{pmatrix} u_1 \\ \vdots \\ n_n \end{pmatrix} \qquad \beta=\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad Z=[z_1,\cdots,z_k]$$

true variable z_i の代わりに別の変数 x_i を proxy variable に使って $(i=1,2,\cdots,k)$ β の最小二乗推定を行なったとしよう。最小二乗推定値を $\hat{\beta}=[\hat{\beta}_1,\cdots,\hat{\beta}_k]'$ とすれば

$$\hat{\beta}=[X'X]^{-1}X'y=[X'X]^{-1}X'(Z\beta+u)$$

ただし $x_i=[x_{1i},\cdots,x_{ni}]' \quad (i=1,2,\cdots,k)$

$$X=[x_1,\cdots,x_k]$$

(一) Proxy variables を確定変数とみなした場合

x_i $(i=1,2,\cdots,k)$ を確定変数とみることにすれば

$$[X'X]^{-1}X'Z \equiv \begin{pmatrix} a_{11} & \cdots\cdots & a_{k1} \\ \vdots & & \vdots \\ a_{1k} & \cdots\cdots & a_{kk} \end{pmatrix} = \Lambda \qquad \cdots\cdots (2\cdot4)$$

とおくと

$$E(\hat{\beta}) = \Lambda\beta \qquad \cdots\cdots (2\cdot5)$$

となる。上記の $[a_{1i},\cdots,a_{ki}]'$ は

$$[a_{1i},\cdots,a_{ki}]' = [X'X]^{-1}X'z_i$$

であるから z_i を $x_1\cdots x_k$ に回帰させた時の回帰係数ベクトルである。

もし $x_i = z_i$ であれば

$$a_{ii}=1,\quad a_{ji}=0 \quad (j \neq i)$$

となる。たとえば $x_1 = z_1$ とすれば

$$X'X = \begin{bmatrix} z_1'z_1 & z_1'\tilde{X} \\ \tilde{X}z_1 & \tilde{X}'\tilde{X} \end{bmatrix} \quad \text{ただし} \quad \tilde{X} = [x_2,\cdots,x_k]$$

で

$$\begin{bmatrix} a_{11} \\ \vdots \\ a_{k1} \end{bmatrix} = \begin{bmatrix} z_1'z_1 & z_1'\tilde{X} \\ \tilde{X}'z_1 & \tilde{X}'\tilde{X} \end{bmatrix}^{-1} \begin{bmatrix} z_1'z_1 \\ \tilde{X}'z_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0_{k-1} \end{bmatrix}$$

ただし 0_{k-1} は $k-1$ 次のゼロベクトル

となる。

従って特に $x_i = z_i$ $(i=1,2,\cdots,k-1)$, $x_k \neq z_k$ とすれば $E(\hat{\beta})$ と β の関係は次のように表わされる。

$$E(\hat{\beta}) = \begin{bmatrix} I_{k-1} & \tilde{a}_k \\ 0 & a_{kk} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \beta_k \end{bmatrix}$$

あるいは

$$E(\hat{\tilde{\beta}}) = \tilde{\beta} + \beta_k \tilde{a}_k \quad \cdots\cdots (2\cdot 6)$$

$$E(\hat{\beta}_k) = \beta_k a_{kk} \quad \cdots\cdots (2\cdot 7)$$

ただし　$\tilde{\beta} = [\beta_1, \cdots, \beta_{k-1}]'$　$\hat{\tilde{\beta}} = [\hat{\beta}_1, \cdots, \hat{\beta}_{k-1}]'$

$\tilde{a}_k = [a_{1k}, \cdots, a_{k-1k}]'$

従って

$$z_k = \sum_{i=1}^{k-1} z_i a_{ik} + x_k a_{kk} + \text{residual}$$

ともかける。

2・5式にみるように proxy variables により推定された係数の期待値 $E(\hat{\beta}_i)$ は、一般にオリジナルな係数 $\beta_1, \cdots, \beta_k$ の一次結合である。従って2・4式の行列 A に関する知識がない限り β に関する先験的情報（符号条件とか係数間の制約条件など）が与えられたとしても、それを係数の推定に利用することはできない。

$\hat{\beta}$ と β の関係を更に追求するには proxy variable と true variable の関係を特定化せねばならないが、ここでは1・1式のような関係がある場合について考察を進めてみよう。問題を単純化するために説明変数の中の一個に

ついて proxy variable が使われているものとする。そして proxy variable x は次のようなモデル（errors in variables model）に従う確率変数であると想定する。

$$x_j=\xi_j+v_j \quad \cdots\cdots (2\cdot 8)$$

$j=1,2,\cdots,n$

ただし　$E(v_j)=0$　$E(v_j^2)=\sigma_1^2$　$E(v_j v_{j'})=0$　$(j \neq j')$

ξ_j は確定変数だが，その値は未知な変数.

従属変数 y_j の従う線型回帰式を書き改めて次のように表わしておく。説明変数 $z_1,\cdots,z_k$ はその値が既知な確定変数であるが，ξ はその値が未知な確定変数である。誤差項 u_j についての仮定は２・１式と同じ。

$$y_j=\alpha_0+\sum_{i=1}^{k}\alpha_i z_{ji}+\beta\xi_j+u_j \quad \cdots\cdots (2\cdot 9)$$

$j=1,2,\cdots,n$

この場合には観測系列 y_i, x_j は２・９式，２・８式という equation systems で規定される内生変数であるが，ξ_j を x_j で代用して最小二乗法で２・９式の係数を推定したとすれば最小二乗推定値 $\hat{\beta}$ はどのような分布に従うであろうか。

(II) errors in variables model

1　単純な場合について

まず次のような単純な場合を考察してみる。

$$y=\beta\xi+u \quad\cdots\cdots\cdots\cdots (2\cdot 10)$$

$$x=\xi+v \quad\cdots\cdots\cdots\cdots (2\cdot 8)$$

ただし，u, v は正規変数ベクトルで

$$E(u)=E(v)=0_n \quad E(u_j v_{j'})=0 \quad (j \neq j')$$

u, v の共分散行列 $\sum_{uu}$, $\sum_{vv}$ は

$$\textstyle\sum_{uu}=\sigma_2^2 I_n \qquad \sum_{vv}=\sigma_1^2 I_n$$

とする.

ξ の代わりに proxy variable x を使った時の β の最小二乗推定値 β は

$$\hat{\beta}=\sum_{j=1}^{n} y_j x_j \Big/ \sum_{j=1}^{n} x_j^2$$

であるが、$\hat{\beta}$ の確率密度関数やモーメントについては既に Richardson and De-Min Wu〔3〕Sawa〔4〕並びに Takeuchi〔5〕によって明らかにされている。$\hat{\beta}$ のモーメントは〔5〕の結果を利用すれば直ちに導けるが、それによると、$\hat{\beta}$ の期待値 $E(\hat{\beta})$ は

$$E(\hat{\beta})=\beta 2z g_n(z) \quad ただし，n>1$$

で、2・7式に対応する関係式が得られる。従って $\hat{\beta}$ の偏りは

$$E(\hat{\beta}-\beta)=-\beta[1-2z g_n(z)] \quad\cdots\cdots\cdots\cdots (2\cdot 11)$$

上式の z および $g_n(z)$ の定義は次の通りである。

$$z \equiv \sum_{j=1}^{n} \xi_j^2 / 2\sigma_1^2$$

$$g_m(z) \equiv e^{-z} \sum_{k=0}^{\infty} \frac{1}{(n+2k)} \frac{z^k}{k!}$$

$\hat{\beta}$の平均二乗誤差（M. S. E）は

$$E(\hat{\beta}-\beta)^2 = \beta^2[1+(2z^2-5z)g_n(z)+zg_{n-2}(z)-2z^2g_{n+2}(z)] + \frac{\sigma_2^2}{\sigma_1^2}g_{n-2}(z) \qquad (2\cdot12)$$

ただし、$n>2$

あるいは

$$E(\hat{\beta}-\beta)^2 = \frac{(n-3)\beta^2}{2} + \left\{\frac{\sigma_2^2}{\sigma_1^2} - \beta^2\left[(n-3)z + \frac{(n-2)(n-5)}{2}\right]\right\} g_{n-2}(z) \qquad (2\cdot13)$$

上記の $g_n(z)$ は合流型超幾何関数（confluent hypergeometric functiou）${}_1F_1(a, b, z)$[一] と次の関係があるから

$$ne^z g_n(z) = {}_1F_1\left(\frac{n}{2}, \frac{n}{2}+1, z\right)$$

2・11式および2・13式は次のようにも書ける。

$$E(\hat{\beta}-\beta) = -\beta\left[1-\frac{2z}{n}e^{-z}{}_1F_1\left(\frac{n}{2}, \frac{n}{2}+1, z\right)\right] \qquad (2\cdot14)$$

あるいは[2]

$$E(\hat{\beta}-\beta)=-\beta e^{-z}{}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2},\ z\right)\cdots\cdots(2\cdot15)$$

そして

$$E(\hat{\beta}-\beta)^2=\frac{(n-3)}{2}\beta^2+\frac{1}{n-2}\left\{\frac{\sigma_2{}^2}{\sigma_1{}^2}-\beta^2\left[(n-3)z+\frac{(n-5)(n-2)}{2}\right]\right\}$$
$$\times e^{-z}{}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2},\ z\right)\cdots\cdots(2\cdot16)$$

あるいは[3]

$$E(\hat{\beta}-\beta)^2=\frac{1}{n-2}\left(\frac{\sigma_2{}^2}{\sigma_1{}^2}+\beta^2\right)e^{-z}{}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2},\ z\right)$$
$$+\beta^2\left(\frac{n-3}{n-2}\right)e^{-z}{}_1F_1\left(\frac{n}{2}-2,\ \frac{n}{2},\ z\right)\cdots\cdots(2\cdot17)$$

$\hat{\beta}$の確率密度関数については〔3〕および〔4〕を参照されたい[4]。2・15式、2・17式と合流型超幾何関数の性質を利用することによって$\hat{\beta}$の相対的な偏りや M.S.E についていろいろな事柄を導くことができる。

(1) $\hat{\beta}$の相対的偏りについて

(イ) $z>0$, $n>1$であれば

$${}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2},\ z\right)>0$$

だから一般に

$$\frac{E(\hat{\beta}-\beta)}{\beta}<0$$

である。

(ロ)　nを固定して$\hat{\beta}$の相対的偏りをzの関数とみると、その絶対値はzが増加するにしたがって減少する。何故ならば

$$\tau\equiv\sum_{j=1}^{n}\xi_j^2/n\sigma_1^2 \qquad \text{すなわち} \quad z=2n\tau$$

とおくと

$$\frac{\partial}{\partial\tau}\left[e^{-z}{}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2},\ z\right)\right]=-4e^{-z}{}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2}+1,z\right)<0$$

ただし，$n>1$

であるから

$$\frac{\partial}{\partial\tau}\left|\frac{E(\hat{\beta}-\beta)}{\beta}\right|<0$$

となる。

(ハ)

$$\lim_{\tau\to\infty}\frac{E(\hat{\beta}-\beta)}{\beta}=0,\ \lim_{\tau\to0}\frac{E(\hat{\beta}-\beta)}{\beta}=-1$$

であるので、如何なるnに対しても$\hat{\beta}$の相対的偏りは常に0と-1の間にあることがわかる。従って$E(\hat{\beta})$の符

号はβの符号と一致することになる。

これは合流型超幾何関数の積分表示式2・18式を使って2・14式を書き改めれば直ぐに導ける。

$$ {}_1F_1\left(\frac{n}{2},\ \frac{n}{2}+1,\ z\right)=\frac{n}{2}z^{-\frac{n}{2}}\int_0^z e^t t^{\frac{n}{2}-1}dt \quad\cdots\cdots(2\cdot 18)$$

$$\frac{E(\hat{\beta}-\beta)}{\beta}=-1+z^{1-\frac{n}{2}}e^{-z}\int_0^z e^t t^{\frac{n}{2}-1}dt \quad\cdots\cdots(2\cdot 19)$$

$$\lim_{\tau\to\infty} z^{1-\frac{n}{2}}e^{-z}\int_0^z e^t t^{\frac{n}{2}-1}dt=1$$

だから

$$\lim_{\tau\to\infty}\frac{E(\hat{\beta}-\beta)}{\beta}=0$$

(二)　2・19式を使って、nが充分大きい場合の、$\hat{\beta}$の相対的偏りの漸近的近似式を求めると

$$e^{-z}z^{1-\frac{n}{2}}\int_0^z e^t t^{\frac{n}{2}-1}dt=\int_0^z e^t\left(1-\frac{t}{z}\right)^{\frac{n}{2}-1}dt=\int_0^{2n\tau} e^{-t}\left(1-\frac{t}{2n\tau}\right)^{\frac{n}{2}-1}dt$$

ただし，$z=2n\tau$

$\lim_{n\to\infty}\sum_{j=1}^{n}\xi_j^2/n$ が収束するとせば、nが充分大きい場合には

$$\int_0^{2n\tau} e^{-t}\left(1-\frac{t}{2n\tau}\right)^{\frac{n}{2}-1}dt \sim \int_0^{\infty} e^{-\left(1+\frac{1}{\tau}\right)t}\,dt=\frac{\tau}{1+\tau}$$

従って

$$\frac{E(\hat{\beta}-\beta)}{\beta}\sim\frac{-1}{1+\tau}$$

ただし，$\tau=\sum_{j=1}^{n}\xi_j^2/n$.

なお Richardson and De-Min Wu〔3〕は次のような，より詳しい漸近式を導いている。

$$\frac{E(\hat{\beta}-\beta)}{\beta}=\frac{-1}{1+\tau}\left[1-\frac{2}{n}\frac{\tau^2}{(1+\tau)^2}+\cdots\cdots\right]$$

(2) $\hat{\beta}$の平均二乗誤差（M.S.E）について

(イ) 2・17式より知れるように M.S.E は n, β^2, σ_2^2/σ_1^2 および τ の四つの要因に依存している。

$$\frac{\partial E(\hat{\beta}-\beta)^2}{\partial\tau}=\frac{2n}{n-2}\left[-\frac{2}{n}e^{-z}{}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2}+1,z\right)-\frac{4(n-3)}{n}{}_1F_1\left(\frac{n}{2}-2,\ \frac{n}{2}+1,z\right)\right]<0$$

ただし，$n>2$

であるから，M.S.E は τ の減少関数である。他方 β^2, σ_2^2/σ_1^2 に関して増加関数であることは2・19式より明らかである。

(ロ)〔3〕には $\hat{\beta}$ の分散 $V(\hat{\beta})$ について次のような漸近式が導出されている。

$$V(\hat{\beta})=\frac{1}{n-2}\left[\frac{\sigma_2{}^2/\sigma_1{}^2}{1+\tau}+\frac{\beta^2\tau(1+\tau^2)}{(1+\tau)^3}\right]+\cdots\cdots$$

$\hat{\beta}$の相対的偏りや M.S.E の数値的な評価にあたっては〔3〕に掲げられている数値表が便利である。

2　外成変数が追加された場合

観測値 (x_j, y_j) $j=1,2,\cdots,n$ が2・8式および2・9式にしたがう場合を考察する。行列形式に書き改めると

$$y=Z\alpha+\beta\xi+u \cdots\cdots\cdots\cdots (2\cdot20)$$

$$x=\xi+v \cdots\cdots\cdots\cdots (2\cdot21)$$ [5]

ただし，$Z=\begin{pmatrix} 1 & z_{11}\cdots\cdots z_{1k} \\ \vdots & \vdots \quad\quad \vdots \\ 1 & z_{n1}\cdots\cdots z_{nk} \end{pmatrix}$　$\alpha=\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$

Zの階数は $(k+1)$ とする.

2・20式でξの代わりにxを proxy variable に使って，βの最小二乗推定値$\hat{\beta}$を求めたとすれば，$\hat{\beta}$はどんな分布に従うだろうか。以下に述べるようにこの問題は二の(一)の1で扱った問題に還元できるので，そこで導出した結果を少し修正するだけで処理できる。

$\hat{\beta}$を書き改めると

$$\hat{\beta}=\frac{x'Qy}{x'Qx} \cdots\cdots\cdots\cdots (2\cdot22)$$

ただし，$Q=I_n-Z(Z'Z)^{-1}Z'$

他方ξを $z_1, \cdots, z_k$ に回帰させた時の最小二乗法的な回帰式を

$$\xi = Zr + e \qquad (2\cdot 23)$$

とおくと，2・22式の$\hat{\beta}$は次のように表わせる。

$$\hat{\beta} = \frac{(e'+v')Q(\beta e+u)}{(e'+v')Q(e+u)} \qquad (2\cdot 24)$$

Zの階数が $k+1$ であれば，$Z(Z'Z)^{-1}Z'$ は階数が $k+1$ のベキ等行列 (idempotent matrix) であるから

$$Z(Z'Z)^{-1}Z' = P'\begin{bmatrix} I_{k-1} & 0 \\ 0 & 0 \end{bmatrix}P$$

ただし，Pはn次の直交行列．I_{k+1} は $(k+1)$ 次の恒等行列．

従って $a'Qb$ という双1次形式は常に2・25式のように変形できる。[6]

$$a'Qb = c'_{(2)}d_{(2)} \qquad (2\cdot 25)$$

ただし，$c_{(2)}, d_{(2)}$ は $c \equiv Pa$, $d \equiv Pb$ をそれぞれ次のような二つの部分に分割してえられる $(n-k-1)$ 次の部分ベクトルである．

$$c = \begin{bmatrix} c_{(1)} \\ c_{(2)} \end{bmatrix} \quad c_{(1)} = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix} \quad c_{(2)} = \begin{bmatrix} c_{k+1} \\ \vdots \\ c_n \end{bmatrix}$$

$$d = \begin{bmatrix} d_{(1)} \\ d_{(2)} \end{bmatrix} \quad d_{(1)} = \begin{bmatrix} d_1 \\ \vdots \\ d_k \end{bmatrix} \quad d_{(2)} = \begin{bmatrix} d_{k+1} \\ \vdots \\ d_n \end{bmatrix}$$

2・25式を使って2・24式を更に変形すると

$$\hat{\beta}=\frac{h'_{(2)}k_{(2)}}{h'_{(2)}h_{(2)}} \qquad \cdots\cdots (2 \cdot 26)$$

ただし，$h \equiv P(e+v) \equiv [h_{(1)}, h_{(2)}]'$

$k \equiv P(\beta e+u) \equiv [k_{(1)}, k_{(2)}]'$

ベクトル h, k は，お互いに独立に正規分布にしたがう確率ベクトルであって，期待値が

$$E(h)=Pe \equiv [\varphi_{(1)}, \varphi_{(2)}]'$$

$$E(k)=\beta Pe \equiv [\beta\varphi_{(1)}, \beta\varphi_{(2)}]'$$

共分散行列 $\sum_{hh}, \sum_{kk}$ が

$$\sum_{hh}=\sigma_1^2 I_n \qquad \sum_{kk}=\sigma_2^2 I_n$$

であるから

$$k_{(2)}=\beta\varphi_{(2)}+u \qquad \cdots\cdots (2 \cdot 27)$$

$$h_{(2)}=\varphi_{(2)}+v$$

ただし，u，v は (2・8)，(2・10) の時と同じ確率ベクトル．

となる。従って $\hat{\beta}$ は2・27式の β を最小二乗法で推定していることにほかならない。

$$\varphi'_{(2)}\varphi_{(2)}=e'Qe=e'e$$ [7]

であるから二の(二)の1で述べた結果を使えば，$\hat{\beta}$ は次のような偏りをもつことがわかる。

$$\frac{E(\hat{\beta}-\beta)}{\beta}=-exp(-z)F_1\left(\frac{n-k-1}{2}-1,\ \frac{n-k-1}{2},\ z\right)\cdots\cdots(2\cdot28)$$

ただし，$n>k+2$

$$z\equiv e'e/2\sigma_1^2 \cdots\cdots(2\cdot29)$$

他方 $\hat{\beta}$ のM.S.Eは2・16式あるいは2・17式の n を $n-k-1$，で，z を2・29式で置換するだけで評価できる。$e'e$ は ξ の回帰式2・22式の残差平方和だから，k 個の説明変数による決定係数を R^2 とおけば

$$z=\frac{\sum_{j=1}^{n}(\xi_j-\bar{\xi})^2(1-R^2)}{2\sigma_1^2}$$

従って，その他の条件を一定とすれば R^2 が大きい程 $\hat{\beta}$ の相対的偏りの絶対値は増大することになる。たとえば

$$n=11,\ k=2,\ \sum_{j=1}^{n}(\xi_j-\bar{\xi})^2/(n-1)\sigma_1^2=4$$

とする。〔3〕の数値表によれば $R^2=0,\ 0.5,\ 0.8$ の時の $\hat{\beta}$ の相対的偏りは，それぞれ $-0.14, -0.28, -0.47$ となる。

注(1)　${}_1F_1(a,b,z)\equiv\sum_{k=0}^{\infty}\frac{(a)_k}{(b)_k}\frac{z^k}{k!}=\sum_{k=0}^{\infty}\frac{a(a+1)\cdots(a+k-1)}{b(b+1)\cdots(b+k-1)}\frac{z^k}{k!}$

(2)(3)　合流型超幾何関数についての次の関係式を利用する。

$$z\,{}_1F_1(a+1,b+1,z)=b[{}_1F_1(a+1,b,z)-{}_1F_1(a,b,z)]$$

$$\left(\frac{n-2}{n}\right)e^{z}=\left(z+\frac{n-4}{2}\right){}_1F_1\left(\frac{n}{2}-1,\ \frac{n}{2},\ z\right)+{}_1F_1\left(\frac{n}{2}-2,\ \frac{n}{2},\ z\right)$$

（4） いずれも

$$y=\alpha+\beta\xi+u,\quad x=\xi+v$$

というモデルのもとでの $\hat{\beta}$ の分布である。

（5） $x=r\xi+v$ であれば $r\xi=\eta$ とおくことにより

$$y=\alpha_0+\sum_{i=1}^{k}\alpha_i z_i+\left(\frac{\beta}{r}\right)\eta+u$$

$$x=\eta+v$$

となって2・20式、2・21式と同じ形式のモデルになる。

（6） $a=P'c,\ b=P'd$ だから

$$a'Qb=c'P[I_n-Z(Z'Z)^{-1}Z']P'd=c'd-c'\begin{bmatrix}I_{k+1} & O\\ 0 & 0\end{bmatrix}d=c'_{(2)}d_{(2)}$$

（7） 2・25式より

$$\varphi'_{(2)}\varphi_{(2)}=e'Qe=\xi'Q\xi=e'e$$

三　観測誤差の評価——農業生産統計を対象にして

proxy variable が2・8式のようなモデルに従うものとすれば、観測誤差に関する情報をもつことが望ましい。たとえば大雑把にでも σ_1 の大きさがわかれば、それを利用して β のより偏りの小さい最小二乗推定値を得ることができる。

最近、東南アジア諸国の農産物供給を対象にした計量経済学的な研究が発表されているが、Behrman〔1〕の計

測にみるように農業統計の不備と不正確さが一つの障害になっているように思える。生産統計の多くは表式調査や素朴な面接調査を通じて作成されているので、調査誤差の影響を無視できない。第三節では既成の統計データをもとにして調査誤差を暫定的に評価しようとする試みを述べたものである。事例としてタイ、フィリピンの米の生産統計を扱ってある。

(一) 農業センサスとの比較

Behrman〔1〕はタイの主要農産物の生産量統計の信頼性を評価するに当たって、次の三通りの方法を採用している。(イ)一つは農林省の生産量統計から誘導される消費量を別の消費量統計と比較する方法。(ロ)その統計が表式調査や面接調査によっている場合には、標本実測調査による推計値と比較する。たとえば単位面積当たり収量を坪刈りによる推計値と比較してチェックする法。第一表はP.C. Mahalanobis がインド・ベンガル州のジュートの生産量について三種類の統計数字を比較した結果を紹介したものであるが、上記の方法の典型的な例である。

第1表　ベンガル州のジュート生産量に関する諸統計の比較（1944/45 および 1945/46）

（単位：1,000 bales）

	1944/45	1945/46
1. 消費量		
工場で	6,000	6,308
移　出	1,050	2,213
村　で	600	600
2. 計	7,650	9,121
3. 前年よりの繰越し	324	697
4. 移　入	598	862
5. （バランス）消費統計による生産量	6,728	7,562
6. 農林省推計（悉皆）	4,895	6,304
7. 標本調査	6,480	7,540
8. 6—5; 5に対する%	−27.2	−16.6
9. 7—5;　〃	−3.6	−0.3

資料：Mahalanobis, P. C. and Lahiri, D. B., "Analysis of errors in censuses and Survey with special reference to experience in India," *Bulletin of the International Statistical Institute*, Vol. 38, Part 2, 1961.

農林省推計はジュート畑を悉皆的に見回って検見した結果にもとづいている。第一表の七の標本調査の推計量は坪刈りと面積実測をもとにしたものである。(ハ)第三の方法は特定年次に限られるが、農業センサスによる結果との比較によるものである。

以下述べるのはこの方法によって調査誤差を検出しようとする試みである。事例としてタイ、フィリピンの米生産統計を扱うので、対象とする統計系列の概略を述べておきたい。フィリピンの一九五三年までの公式統計は *Phillipin Agriculture Statistics*, Vol. 1, II に集録されているデータが唯一のものであるが、これはそれまで各省各部局が業務統計として発表していたものを農業天然資源省(DANR)が統一集大成したものである。*phillipin Agricultune Statistics* の一九四八年度(作物年度)の米の作付面積、生産量、単位面積当たり収量を一九四八年に行なわれた農業センサスの結果と比較するわけであるが、DANR が行なった生産統計の編集手続きは明らかでない。農業センサスはセンサス局によるもので、一九四八年度(一九四七・七・一〜一九四八・六・三〇)の生産活動を対象にして農場面積が一〇ヘクタール以上の面積をもつ農家を調査している。一九五四年からは DANR によって標本面接調査による作物家畜調査が開始され、この調査から各作物に関する生産統計や家畜飼育頭数などのカレントな統計が作成されるようになっている。

タイの農産物に関するカレントな生産統計は muban-tambal-amphur-changwad という行政機構を通り、農林省の Rice Department や Division of Agricultural Economics で集計されるわけであるが、米の生産統計は muban (village) の長である phuyaibau が彼の village について述べた報告数字が基礎になっている。タイについては一九六〇年度の米の生産量統計を農業センサスの結果と比較する。

(三) 比較の方法

二つの統計系列を比較するには、農家あるいは村といった最終の調査単位で行なうのが望ましいが、公表されているデータはたかだか province といった大きな行政区轄に関するものであるから、実際には province レベルの比較しかできない。フィリピンでは一九四八年度について province レベルの比較を、タイでは一九六〇年度について changwad レベルの比較をすることができる。

ここでは province (changwad) 間で二つの統計系列がどのような相違をしめすかに注目する。なお、フィリピンにおける province の数は五〇で、これらの province が九個の region に分類されている。タイでは七一の changwad が四個の region に分類されている。

ところで比較した結果の判定基準であるが、たとえば米の作付面積を province レベルで比較した場合、二つの統計系列の間の一致度が高ければカレントな生産統計は一致度の低いものに比べて better であると考える。センサスの結果が正確であるとする保証はないから、一致度の高いことはカレントな生産統計の正確さを意味するものでないが、一致度が高ければそれだけ信頼しうるものと認めようという考えである。

1 不一致度の尺度

二つの統計系列の一致度の測り方にはいろいろあるが、ここでは次のように表わす。i 番目の province (changwad) のカレントな生産統計によるデータを X_i で、農業センサスの結果を Y_i で表わす。

$$g=\frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i)^2 \quad \cdots\cdots (3\cdot1)$$

ただし，n は province (changwad) の数.

3・1式のgは面接調査における回答誤差の変動を分析するのによく使われる統計量であるが、その場合にはX_i、Y_iは同じ調査条件のもとで面接を二回反覆してえられる被調査者の回答を意味する。米国のセンサス局が最近統計調査の信頼性に関する組織的研究を実施しているが、3・1式のgを gross difference rate と呼んでいる。またこの統計量は予測の正確度を測る場合にも使われている。この時にはX_iが予測値を、Y_iが実際値を表わす。

統計系列の比較を年次ごと作物ごとあるいは国ごとに行なってその結果を比べようとするには、3・1式のgを指数形式にしておくのが便利である。Theil〔6〕の手法に従って次のように指数化する。Theil は "inequality coefficient"（不等度）と呼びUという記号を使っている。

$$U=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i)^2}\Bigg/\left[\sqrt{\frac{\sum_{i=1}^{n}X_i^2}{n}}+\sqrt{\frac{\sum_{i=1}^{n}Y_i^2}{n}}\right]\quad\cdots\cdots\cdots\cdots(3\cdot2)$$

Uが大きい程二つの統計系列の間の不一致度は大きいと判断するわけだが、この根拠はU-指数の幾何学的性質にもとづくものである。(一)

ところで3・1式のgには種々の要因が関与している。それぞれの統計系列を生成する調査手続きや調査環境の相違、標本抽出法が使われておれば抽出誤差、面接調査であれば被調査者の回答誤差、その他種々の要因が影響している。米国センサス局の回答誤差の研究ではgの要因分析をするために〃センサス局モデル〃が設定されているが、ここでもそれに相当する errors model を前提することになる。その議論は後回しにして、まず3・1式のgを Theil〔6〕に従って次のように分解してみる。

$$g=(\bar{X}-\bar{Y})^2+(S_X-S_Y)^2+2(1-r)S_XS_Y \cdots\cdots \quad (3・3)$$

ただし，r は相関係数．S_X, S_Y はそれぞれ標準偏差．

3・3式の右辺の各項を$\left\{\frac{\sum X_i^2}{n}+\frac{\sum Y_i^2}{n}\right\}$で除したものをそれぞれ U_M^2，U_S^2 および U_C^2 とすれば，3・2式は次のように表わされる。

$$U^2=U_M^2+U_S^2+\bar{u}_C^2$$

Theil〔6〕は U_M^2，U_S^2 および U_C^2 より次の諸量を誘導している。

$U_M=\sqrt{U_M^2}$：中心値の不等による偏不等度

$U_S=\sqrt{U_S^2}$：変動の不等による偏不等度

$U_C=\sqrt{U_C^2}$：共変関係の不等による偏不等度

$U^M=U_M^2/U^2$：bias proportion

$U^S=U_S^2/U^2$：Variance 〃

$U^C=U_C^2/U^2$：Covariance 〃

U_M が調査手続きや調査環境などの相違による系統的な不整合を表わす成分であることは明らかであろう。もし二つの統計系列が，同じ調査手続きに従って類似の調査環境のもとで作成されたものとすれば，bias proportion の U^M は $U^M \fallingdotseq 0$ となろう。他方，調査手続きや調査環境が均しくても U^S や U^C が消失することはないだろう。二つの統計系列は同じ項目を観測しているから，調査がお互いに無関係に行なわれたとしても統計系列の間にはあ

第2表　タイ，フィリピンの米生産統計の U-指数

	タイ（1960年度）			フィリピン（1948年度）			日本（1965年度）
	作付面積	生産量	収量	作付面積	生産量	収量	作付面積
U	0.0572	0.0685	0.0985	0.1652	0.2001	0.1160	0.0539
U_M	0.0160	0.0011	0.0228	0.0316	0.0549	0.0506	0.0471
U_S	0.0010	0.0007	0.0030	0.0317	0.0130	0.0020	0.0035
U_C	0.0549	0.0047	0.0958	0.1590	0.1920	0.1044	0.0259
U^M	0.0785	0.0003	0.0536	0.0365	0.0752	0.1904	0.7646
U^C	0.0003	0.0001	0.0009	0.0367	0.0042	0.0002	0.0041
U^S	0.9212	0.9996	0.9455	0.9268	0.9206	0.8094	0.2313
n	50			71			45

注 1．タイの収量は収穫面積当たり，フィリピンの収量は作付面積当たりの収穫量である．

2　日本では農業センサスの収穫面積（X）を作物調査の作付面積（Y）と比較．

る直線的な相関関係が成立する。実際の観測過程には偶然変動を含めて種々の観測誤差が混入するから、二つの観測系列間の相関関係は誤差の程度に応じて不完全なものになる。この不完全さはU_CあるいはU^Cを通じて測ることができる。U_SあるいはU^Sは観測誤差の大きさの差を表わすものであるが、次に述べる計測例ではこの成分の寄与は無視しうる程度である。

2　計測結果

米の作付面積、生産量および収量について二つの統計系列の間のU-指数を計算してみる。各成分も併せて第二表に掲げてあるが、参考のためにわが国の水稲についてのU-指数を掲げてある。日本のU-指数は一九六五年度について農業センサスの収穫面積を作物調査による作付面積と比較したもので、北海道を除く四五都府県を対象にしている。ただし、農業センサスの結果をX_tで、作物調査の推計値をY_tで表わしている。タイやフィリピンのU-指数を検討するための対照として役立つであろう。

作付面積のU-指数は日本とタイの間ではその差は僅少であるが、指数の成分をみると著しい相違のあることがわかる。タイで

は bias proportion の U^M が〇・〇七八五であるのに対して、日本の U^M は〇・七六四六となっている。日本の場合、作物調査による推計は実測によっているので、二つの統計系列間では調査方法の相違による影響が大きく、この相違が規則だった系統的誤差を生成する主たる源泉になっている。しかも、統計系列間の共変関係の程度が高いので、その結果として bias proportion はタイに比べて著しく増大するわけである。タイの場合、調査手続きでは、一方は面接調査、他は表式調査といった差があるわけだが、その差は g を構成する主要因にはなっていない。系統的な不整合を生成するにしてもその程度は小さいか、あるいはその現われ方が changwad を通じて一様ではない。

タイの場合、作付面積だけでなく生産量、収量についても共変関係の不完全さが、U-指数の主要因になっている。この点はフィリピンでも同じだが、タイに比べて U^C がより大きい点が特長である。

(三) errors model の設定

(二)の2でみたようにタイ、フィリピンの指数は主として二つの統計系列間の共変関係の乱れによっている。この傾向は米以外の作物についても同じであろう。作物間でどのような差が出てくるかを検討するのも興味ある問題であるが、ここでは共変関係の不完全さを別の観点から分析してみたい。

各 province (changwad) の真の米作付面積を Z_i, $i=1,2,\cdots,n$ とおくと、X_i-Z_i および Y_i-Z_i は一般に Z_i の大きさに依存するであろう。この関係を次のようにモデル化する。

$$X_i=(\alpha+\Delta\alpha_i)Z_i$$

$$Y_i=(\beta+\Delta\beta_i)Z_i$$

$$i=1,2,\cdots,n$$

$\Delta\alpha_i$, $\Delta\beta_i$ はいずれも期待値がゼロ、ある分散をもつ確率変数。つまり X_i, Y_i の Z_i に対する偏倚率はそれぞれ期待値 α、β をもつ分布に従って変動するものと想定する。これらの期待値は地域ごとに異なるのであるが、まず各地域を通じて一定とみなして議論を進めてゆく。

$\{X_i\}$ 系列についてみると

$$\log X_i=\log_e X_i=\log_e Z_i+\log_e \alpha+\log_e\left(1+\frac{\Delta\alpha_i}{\alpha}\right)$$

であるが、

$$\log_e\left(1+\frac{\Delta\alpha_i}{\alpha}\right)\sim\frac{\Delta\alpha_i}{\alpha}-\frac{1}{2}\left(\frac{\Delta\alpha_i}{\alpha}\right)^2$$

と近似し、さらに $\Delta\alpha_i/\alpha$ の分散は province (changwad) を通じて一定と仮定して $C_\alpha{}^2$ とおけば

$$E\left\{\log_e\left(1+\frac{\Delta\alpha_i}{\alpha}\right)\right\}\sim 1-\frac{1}{2}C_\alpha{}^2$$

$$V\left\{\log_e\left(1+\frac{\Delta\alpha_i}{\alpha}\right)\right\}\sim C_\alpha{}^2$$

ただし、$C_\alpha{}^2=V\left(\frac{\Delta\alpha_i}{\alpha}\right)$ $i=1,2,\cdots,n$

$C_\alpha{}^2$ は $\Delta\alpha_i$ の relative variance にほかならない.

したがって $\{X_i\}$ 系列について次のような errors model が導かれる。

$$\log_e X_i = \log_e Z_i + \phi + u_i \cdots\cdots (3 \cdot 4)$$

ただし，$\phi \fallingdotseq \log_e \alpha - \frac{C_\alpha^2}{2}$

$$E(u_i)=0 \quad E(u_i^2) \fallingdotseq C_\alpha^2$$

同じ要領で $\{Y_i\}$ 系列についても次のような errors model を設ける。

$$\log_e Y_i = \log_e Z + \varphi + v_i \cdots\cdots (3 \cdot 5)$$

ただし，$\varphi \fallingdotseq \log_e \beta - \frac{C_\beta^2}{2}$

$$E(v_i)=0 \quad E(v_i^2) \fallingdotseq C_\beta^2$$

$$C_\beta^2 = V\left(\frac{\Delta\beta_i}{\beta}\right) \quad i=1,2,\cdots,n$$

統計系列の $\{X_i\}$, $\{Y_i\}$ が表式調査や面接調査のような観測過程を通じて生成される場合には、α、β は一般に１から乖離した値をとるだろう。

３・４式、３・５式の ϕ、φ はそれぞれの観測過程の系統的な偏りを表わす項である。C_α^2, C_β^2 は真値 Z_i に対する偏倚率の province (changwad) 間変動を表わす relative Variance であるが、この変動は調査手続きや被調査者側の知識状態のあいまいさによって生ずるものであるから、観測過程の不規則性または不安定性をしめす尺度とみることができる。C_α, C_β をそれぞれの統計系列の不安定度と呼ぶことにする。

わが国の米作付面積のように $\{Y_i\}$ 系列が標本実測調査で推定されている場合を考えてみよう。

推定値が不偏性をもてば $\beta=1$ とおける。$C_\beta{}^2$ は標本抽出誤差による Y_i の relative Variance であるから、(2) わが国の県段階の水稲作付面積の C_β は大体〇・五%とみてよい。あとでしめすように農業センサスによる水稲収穫面積の不安定度 C_α は約四%とみられる。この場合には悉皆調査であるから不安定度の主源泉は回答誤差における不規則性である。

C_α および C_β を求める手続きは次の通りである。

$$\log_{10} X_i=x_i,\quad \log_{10} Y_i=y_i$$

として

$$y_i-x_i=d_i$$

を求めると、3・4式、3・5式から $d_1, d_2, \cdots, d_n$ は3・6式にしめすところの期待値 $\tilde{d}$、分散 σ^2 をもつ母集団分布からの任意標本とみなすことができる。

$$\tilde{d}=\varphi-\phi=(\log_{10}\beta-\log_{10}\alpha)-\frac{M}{2}(C_\beta{}^2-C_\alpha{}^2)\cdots\cdots\cdots\cdots(3\cdot6)$$

$$\sigma^2=M^2(C_\beta{}^2+C_\alpha{}^2-2C_{\alpha\beta})$$

ただし，$C_{\alpha\beta}=E(\varDelta\beta_i\varDelta\alpha_i)/\alpha\beta$

$$M=0.43429$$

$\tilde{d}$, σ^2 は $d_1, d_2, \cdots, d_n$ より計算する標本平均 $\bar{d}$ および標本分散 $S_d{}^2$ で推定できる。第三表の(1)、(3)欄にそれ

第3表　不安定度の計算（米の作付面積）

	$\bar{d}$	β/α	S_d	不安定度	備考
日本	0 04960	1.120	$180,673\times10^{-7}$	$41\ 3\times10^{-3}$	Y_j：実測調査 X_j：センサス
タイ	0.02013	1.0480	$7,428\times10^{-5}$	$110\ 3\times10^{-3}$	Y_j：センサス X_j：表式調査
フィリピン	0.01275	1.030	$16,966\times10^{-5}$	276.2×10^{-3}	Y_j：センサス X_j：DANR による.

らの計測結果が掲げてある。$\bar{d}$, S_d^2 を3・6式の $\tilde{d}$, σ^2 に代入することにより β/α や不安定度を近似的に評価できる。

1　計測結果

日本の場合には既述の通り　$C_\beta=5\times10^{-3}$　$C_{\alpha\beta}=0$　とおけるから

$$C_\alpha^2=\frac{\sigma^2}{M}-C_\beta^2$$

σ は $180,673\times10^{-7}$ と推定されるから、農業センサスの不安定度（水稲面積に関する）C_α は 41.3×10^{-3} と推定できる。

$$\log_{10}\left(\frac{\beta}{\alpha}\right)=\tilde{d}+\frac{M}{2}(C_\beta^2-C_\alpha^2)$$

より $\log_{10}\left(\frac{\beta}{\alpha}\right)$ は 4924×10^{-5} と推定される。

タイ、フィリピンの場合には $C_{\alpha\beta}=0$ とおくことは正しくないであろう。それぞれの統計系列における観測過程がどのように関連しているかによるわけだが、$C_{\alpha\beta}>0$ とみなすのが普通であろう。また $C_\beta>C_\alpha$ と考えられるが、ここでは

$$C_\beta=C_\alpha=C,\quad C_{\alpha\beta}=0$$

という仮定を置いて、$S_d/\sqrt{2M}$ で共通の不安定度を求めることにする。$C_{\alpha\beta}>0$ であれば不安定度を過小に評価している結果になる。第三表の(4)欄は以上のような手続きで

求めたものである。タイ、フィリピンの β/α は $\bar{d}=\log_{10}\left(\dfrac{\beta}{\alpha}\right)$ にしたがって求めた。

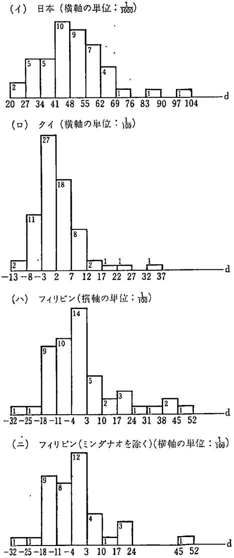
第二表の U_c で見ても明らかであるが、タイ、フィリピンに比べてわが国の農業センサスにおける水稲収穫面積の不安定度は小さい。過小率の全国平均は〇・八九であるが、県間変動は relative Standard deviation でみて約四%である。タイ、フィリピンのカレントな生産統計における米作付面積の偏倚率を知る客観的なデータはないが、Behrman〔1〕の計算によると一九六二年度のタイの米生産量の過小率は八〇〜八五%とみられている。フィリピンの過小率は村岡〔2〕によると一九五一年から一九六五年までの年平均で八六%と推計されている。タイ、フィリピンのカレントな生産統計で問題になるのは全国平均の偏倚率よりも、偏倚率がしめす province (changwad) 間の変動の大きさではないかと思う。先に導いた不安定度でみるとタイが日本の三倍、フィリピンが七倍とかなりの大きさになっている。

第一図は参考として各国の d_c に関するデータをヒストグラムにまとめたものである。

2　地域差について

これまで述べた不安定度の計測では各 province (changwad) の X_c, Y_c の偏倚率の期待値は一定とみたが、これらの期待値は地域間で異なるだろう。その結果3・6式の $\varphi-\phi$ は地域間で変動する。たとえばフィリピンにおいて province をミンダナオとその他の二つの地帯に分けて、地帯別に $\bar{d}$ を求めると、ミンダナオが $\bar{d}=0.14965$, その他が $\bar{d}=-0.02232$ となる（第一図の(ハ)、(ニ)を参照）。これは極端なケースであるが、地域差の存在は無視できない。この影響は S_d の中に混入するから不安定度を推定するためには地域差の影響をできるだけ除去しておくの

第1図　米作付面積に関する d_i のヒストグラム



が望ましい。フィリピンでは九地域に、タイでは四地域に分類[3]されているのでこれらの地域を対象にしてその影響を検討してみる。

地域 h に属する province (changwad) j に関する測定値を X_{hj}, Y_{hj} と区別し、

$$d_{hj}=\log_{10}Y_{hj}-\log_{10}X_i$$

とおく。そして地域差の影響を次のようにモデル化する。

$$d_{hj}=\delta+\tau_h+\epsilon_{hj} \quad \cdots\cdots\cdots\cdots (3\cdot7)$$

ただし，$h=1,2,\cdots,m$

$j=1,2,\cdots,n_h$

3・7式のδは二つの統計系列の log-difference の全国平均，τ_hは地域差をしめす項であって$\sum_{h=1}^{m}\tau_h=0$とおく。

ϵ_{hj}は残差項で　$E(\epsilon_{hj})=0$，$E(\epsilon_{hj}^2)=\sigma^2$と仮定する。

地域間変動およびσ^2に関する推定は次のような分散分析表に従って行なうことができる。

分　散　分　析　表

要　因	平　方　和	自由度	平均平方	期 待 値
地域間	$\sum_h n_h(\bar{d}_{h\cdot}-\bar{d}_{\cdot\cdot})^2$	$m-1$	S_b^2	$\sigma^2+\sum_h n_h\tau_h^2/(m-1)$
地域内	$\sum_h\sum_j(\bar{d}_{hj}-d_{h\cdot})^2$	$n-m$	S_W^2	σ^2
総　計	$\sum_h\sum_j(d_{hj}-\bar{d}_{\cdot\cdot})^2$	$n-1$	S_d^2	

ただし，$\bar{d}_{h\cdot}=\frac{1}{n_h}\sum_{j\sim1}^{n_k}d_{hj}$

$$\bar{d}_{\cdot\cdot}=\frac{1}{n}\sum_h\sum_j d_{hj}$$

3・7式と3・4式，3・5式および3・6式との関連は次の通りである。X_{hj}の真値Z_{hj}に対する偏倚率が従う分布の平均をα_h，relative Variance をC_α^2，一方Y_{hj}の偏倚率の分布の平均をβ_h，relative Variance

をC_{β}^{2}とすると

$$\delta \fallingdotseq \log_{10}\left(\frac{\beta}{\alpha}\right)-\frac{M}{2}(C_{\beta}^{2}-C_{\alpha}^{2})$$

ただし，β は β_h の，α は α_h の幾何平均……………………………(3・8)

$$\tau_h=\log_{10}\left(\frac{\beta_h}{\alpha_h}\Big/\frac{\beta}{\alpha}\right)$$

$$V(\epsilon_{hj})=\sigma^{2}\fallingdotseq M^{2}(C_{\beta}^{2}+C_{\alpha}^{2}-2C_{\alpha\beta})$$

3・8式にみるように3・7式の地域差を表わす項は β_h と α_h の間の log-difference として定義されるものである。

第四表および第五表はタイ、フィリピンに関する分散分析表を掲げたものである。有意水準一%で地域差の有意性検定をすれば、タイでは F-value が五・七一で有意差が認められる。フィリピンについてはミンダナオとその他に分けて地域差をテストした結果が第五表であって明らかに差が認められる。然しミンダナオを除いた八地域間では有意な差は検出できなかった。

まえと同じように

$$C_{\alpha}=C_{\beta}=C,\ C_{\alpha\beta}=0$$

の仮定を置けば分散分析表で求めた S_W^{2} を用い $S_W/\sqrt{2}M$ で不安定度Cを推定できる。地域差が消去されているので、三の㈢の1で求めた不安定度よりも小さく出てくる。地域差を除去したあとの不安定度は、タイが 110×10^{-3} フィリピンが 255×10^{-3} である。それにしてもわが国の農業センサスの四%と対照するとかなり大きい。日本の水

第4表　分散分析表（タイの米作付面積，1960年度）

要　　因	平方和	自由度	平均平方	
地域間	0.0786,754	3	0.0262,251	$F=5.71$
地域内	0.3075,265	67	0.0045,899	$F_{0.05}=2.75$　$F_{0.01}=4.10$
総　　計	0.3862.019	70	0.0055,172	

第5表　分散分析表（フィリピンの米作付面積，1948年度）

要　　因	平方和	自由度	平均平方	
地域間	0.2365,876	1	0.2365,876	$F=9.67$
地域内	1.1739,053	48	0.0244,564	$F=_{0.05}=4.04$
総　　計	1,4104,928	49	0.0287,856	$F_{0.01}=7.19$

稲にみるような例は特異かも知れない。農家自身水田面積について明確なデータをもっているので、農業センサスにおける農家の回答は実測値より偏倚しても、その偏倚の仕方は規則的である。タイの〔X_i〕系列は村長による村レベルの報告にもとづいているから、不安定度が日本に比べて大きいという事実は、偏倚の仕方がより不規則であるということにほかならないが、これは報告調査のもつ一般的な欠陥ともみられる。観測過程として恣意的な要素が多く、各種の調査誤差に曝され易いので、偏倚の仕方が単一のパターンをもたぬのであろう。三の(一)で指摘したが、フィリピンの一九四八年度の〔X_i〕系列はデータ・ソースが唯一でなくDANRによって編集されたものである。フィリピンの不安定度が二五%と大きいのはそれが関係しているものと思われる。

3　不安定度の役割

終わりに三の(一)のところで述べた統計データの評価の方法(イ)との関連を明らかにしておきたい。

方法(イ)によって主要作物に関するカレントな生産量統計の偏倚率がえられる。三の(二)および(三)で述べた手続きによって生産量のU-指数と近似的な不安定度がえられるが、不安定度は方法(イ)で求めた偏倚率に対す

る標準偏差の役割をもつものとみられる。偏倚率はいくつかの仮定のもとで導かれるものであるから、その正確度は客観的に評価し難い。客観的評価は望めないにしても主観的な判断にもとづく評価は行なわねばならない。偏倚率の幅についての知識がないとすれば、三の㈢で求めた不安定度が手掛りになる。たとえば方法(イ)によってタイのある年度の米生産量の偏倚率が八五%、すなわち $\alpha=0.85$ と定められたとしよう。もし

$$C_\alpha=10\%,\quad n=71 \text{ (changwad の数)}$$

とすれば、α の従う分布の relative な標準偏差は(4)

$$\frac{C_\alpha}{\sqrt{n}}=\frac{0.10}{\sqrt{71}}=0.0119$$

従って α の標準偏差は $0.85\times0.0119=0.01.$

α に対する信頼係数九五%の信頼区間は

$$0.85-0.02<\alpha<0.85+0.02$$

で、$0.83<\alpha<0.87$ となる。この信頼区間は n についての事前的な確信度を表わしたものであるから、方法(ロ)によって生産量に関する客観的データが入手されたならばそれを使って修正すればよい。上記の信頼区間はそれまでの暫定値として意味をもつであろう。

注(1)　観測値の組 $(X_1,\cdot\cdot,X_n)$, $(Y_1,\cdots,Y_n)$ を表わす n 次元空間の点を P, Q とする（次頁の図を参照）。線分 OP, OQ および PQ の長さをそれぞれ $|OP|, |OQ|, |PQ|$ とすると

$$|OP|=\sqrt{\sum X_i^2},\quad |OQ|=\sqrt{\sum Y_i^2},\quad |PQ|=\sqrt{\sum(X_i-Y_i)^2}$$

だから

$$U=\frac{|PQ|}{|OP|+|OQ|}$$

点Pと点Qの距離が二辺の和に比べて小さければ一致度は高いとみなすわけである。

なお $0\leqq U\leqq 1$ である。

（2） $V(Y_i)=Z_iV(\Delta\beta_i)$.

$\beta=1$ だから

$$\frac{V(Y_i)}{Z_i^2}=V\left(\frac{\Delta\beta_i}{\beta}\right)=C_\beta^2$$

（3） 地域分類（括孤内の数字は province あるいは changwad の個数）

フィリピン…Ilcos (5) cagayan (4) Central luzon (7) Southern Tagalog (8) Bicol (6) Eastern Visayas (4) Western Vissyas (6) N. & E. Mindanao (6) S. & W. Mındanao (4)

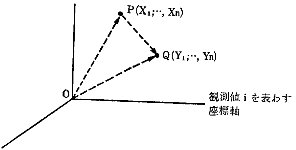
タイ………Central Region (35) North-eastern Region (15) Northern Region (7) Southern Region (14)

（4） $X_i=(\alpha+\Delta\alpha_i)Z_i,\ i=1,2,\cdots,n$

$\Delta\alpha_i$ と Z_i とが無相関とすれば

$$\sum_{i=1}^{n}X_i=(\alpha+\overline{\Delta\alpha})\sum_{i=1}^{n}Z_i$$

ただし、$\overline{\Delta\alpha}=\frac{1}{n}\sum_{i=1}^{n}\Delta\alpha_i$



引用文献

〔一〕 Jere R. Behrman, *Supply Response in Underdeveloped Agriculture*—A Case study of four major annual crops in Thailand, 1937-1963, North-Holland, 1968, Chapter 7.

〔2〕 村岡徳人「東南アジアの統計評価試論(IV)――フィリピンの米――」（アジア経済研究所編『アジア研究』第一一巻第三号、昭和四五年四月）、九〇～九九頁。

〔3〕 David H. Richardson and De-Min Wu, "least Squares and Grouping Method Estimators in the Errors in Variables Model," *Journal of the American Statistical Association*, Vol. 65, No. 330, 1970, pp 724～748.

〔4〕 Takamitsu Sawa, "The Exact Sampling Distribution of Ordinary least Squares and Two-Stage least Squares Estimators," *Journal of the American Statistical Association*, Vol. 64, No. 327, 1969, pp. 923～937.

〔5〕 Ke Takeuchi, "Exact Sampling Moments of The Ordinary least Squares, Instrumental Variable, and Two-Stage least Squares Estimators," *International Economic Review*, Vol. 11, No 1, 1970, pp. 1～12.

〔6〕 Henri Theil, *Economic Forecasts and Policy*, North-Holland, 1958, pp. 31～42.

（研究員）