比推定と回帰推定について

多　賀　保　志

(1963 年 1 月 受 付)

On Unbiased Ratio and Regression Estimators

Yasushi TAGA

Various studies have been made on unbiased ratio and regression type estimators, since an unbiased ratio-type estimator was proposed by Hartley and Ross. (see [1], [2], [3], [4], [5], [6] and [7])

In this report, these methods will be surveyed and an unbiased regression type estimator will be proposed which is applicable even in case when the regression line in the population is nonlinear.

Institute of Statistical Mathematics

§1. はじめに

2 変量 X, Y についての有限母集団 (大きさ N) を考え，X の母平均 $\bar{X}$ は既知とするとき，それからとられた大きさ n のサンプルを観測して，Y の母平均 $\bar{Y}$ を推定する場合に，

(1)　$$\tilde{y}=\bar{y}\bar{X}/\bar{x}$$

(2)　$$\hat{y}=\bar{r}\bar{X}, \qquad \bar{r}=\frac{1}{n}\sum_{i=1}^{n} y_i/x_i$$

などがよく使われてきた．これらは偏りのある比推定量であり，その偏りの大きさも近似的に求められているが，Hartley と Ross は推定量 (2) の偏りを正確に求め，その補正を行なった不偏推定量として，

(3)　$$y'=\bar{r}\bar{X}+\frac{(N-1)n}{N(n-1)}(\bar{y}-\bar{r}\bar{x})$$

を得た．([1] 参照) この推定量の分散は，Goodman と Hartley によって漸近式が求められ ([2] 参照)，Robson によつて正確な式が与えられ，X をいくつかの成分にわけて推定の精度を上げる方法も考えられた ([3], [4] 参照)．また，(2) 式における補助変量 X を 2 次元以上の多変量に拡張する問題が Olkin によって論ぜられている ([5] 参照)．

さらに，(3) 式を重複抽出の場合に拡張することが，Sukhatme によって提案されている ([6] 参照)．そのほか，比推定と回帰推定の混合型の不偏推定量が Mickey によって考察されている ([7] 参照)．

これらの結果を概観してみると，比推定量を用いる場合には，偏りがあるものでもないものでも，母集団における Y の Xへの回帰線が原点を通る直線となっているときに限って有効であることがわかるし，回帰線が原点を通らない直線のときは回帰推定量を用いた方がよいという結論が下せる．さらに，Y の Xへの回帰が直線でないときは，上記の推定法は，単純なサンプル平均 $\bar{y}$ による方法とくらべて，精度の向上にはあまり役立たなくなってしまう．したがって，回帰線が直線である場合はもちろんのこと，直線でない場合にも適用できて，しかも推定の精度の向上が期待できる方法が見出せるならば，それが最も望ましい推定法といえるであろう．以下，そのような回帰推定法

について考察してみよう．

§2. 問題の設定と結果の記述

まずこれまでに考案されている推定量とその精度（分散）を概観しておく．サンプル数 n が十分大きいとき，推定量 (1) と (3) の漸近分散は，

$$\sigma_{\tilde{y}}^2 \sim \frac{1}{n}(\sigma_Y^2+Q^2\sigma_X^2-2Q\sigma_{XY})=\frac{1}{n}\{\sigma_Y^2(1-\rho_{XY}^2)+\sigma_X^2(Q-B)^2\} \tag{4}$$

$$\sigma_{y'}^2 \sim \frac{1}{n}(\sigma_Y^2+\bar{R}^2\sigma_X^2-2\bar{R}\sigma_{XY})=\frac{1}{n}\{\sigma_Y^2(1-\rho_{XY}^2)+\sigma_X^2(\bar{R}-B)^2\} \tag{5}$$

ここに $Q=\bar{Y}/\bar{X}$, $\bar{R}=\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i}$, $B=\frac{\sigma_{XY}}{\sigma_X^2}$

とかけるから，これらをふつうの回帰推定量

$$y''=\bar{y}-b(\bar{x}-\bar{X}), \qquad b=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{6}$$

の漸近分散

$$\sigma_{y''}^2 \sim \frac{1}{n}\sigma_Y^2(1-\rho_{XY}^2) \tag{7}$$

とくらべると，それぞれ

$$\frac{1}{n}\sigma_X^2(Q-B)^2,\ \frac{1}{n}\sigma_X^2(\bar{R}-B)^2$$

だけ大きくなっている．もし，母集団における Y の X への回帰線が原点を通る直線ならば，$Q=\bar{R}=B$ となるから，この場合には

$$\sigma_{\tilde{y}}^2 \fallingdotseq \sigma_{y'}^2 \fallingdotseq \sigma_{y''}^2 \tag{8}$$

とみなして差支えない．したがって，比推定量 $\tilde{y}$ と y' が有効に用いうるのは，“Y の X への回帰線が原点を通る直線とみなされる”場合に限られる（この場合は $\tilde{y}$ も不偏推定量になる）のであって，それ以外の場合には回帰推定量 y'' を用いた方がよいといえる．ついでに，$\tilde{y}$ と y'' の漸近分散 (4), (5) の差をみると，

$$\sigma_{\tilde{y}}^2-\sigma_{y'}^2 \sim \frac{1}{n}\sigma_X^2\{(Q-B)^2-(\bar{R}-B)^2\} \tag{9}$$

となるから，Q と $\bar{R}$ のどちらが回帰係数 B により近いかによって，漸近分散 $\sigma_{\tilde{y}}^2$ と $\sigma_{y'}^2$ の大小ができるのである．いずれにしても，比推定量 $\tilde{y}$ と y' は，母集団における Y の X への回帰線が原点またはその付近を通るときに限って用いるのが有効なのである（すなわち Q または $\bar{R}$ の値が B にごく近いときである）．くわしくいえば，

$$|Q-B| \text{ or } |\bar{R}-B| \leqq k\frac{\sigma_Y}{\sigma_X}\sqrt{1-\rho_{XY}^2}, \quad 0 \leqq k<1$$

となり，かつ k の値がごく小さいならば，比推定量 y または y' は，回帰推定量 y'' とほぼ同じ精度をもつと考えてよいが，k が 1 に近いときには比推定量の精度はぐんとわるくなってしまう．結論としていうならば，比推定量は回帰直線が原点のごく近くを通るときにのみ用いるべきであり，その場合でも回帰推定量を用いた方が安全だといえる．なお，2つの比推定量 $\tilde{y}$ と y' の優劣について考えてみると，後者は不偏であるという利点をもつが，$\bar{r}=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}$ の値を計算するのに非常に手数を必要とする一方，回帰直線が原点のごく近くを通るときは，$\tilde{y}$ の偏りも無視しうる程度であるから，むしろ $\tilde{y}$ の方が使いやすく精度も y' とあまり変らないとみてよい．

さて，以上によって回帰推定量 y'' の有越性がわかったのであるが，回帰線が直線でないときは，ρ_{XY}^2 の値が小さくなるから，y'' の分散 $\frac{1}{n}\sigma_y^2(1-\rho_{XY}^2)$ は単純なサンプル平均 $\bar{y}$ の分散 $\frac{1}{n}\sigma_y^2$ に近くなってしまい，手数をかけて y'' の値を計算する価値がうすくなってしまう．そのようなときには，どうしたらよいであろうか．問題を次のように定式化してみよう．

“母集団における Y の X への回帰線が，

(10)　$$\tilde{Y}_X=\varphi(X)$$

で表わされ．かつ X の分布（またはいくつかの絶対積率）がわかっているとしたとき，Y の母平均 $\bar{Y}$ の推定量（$\bar{y}$ より有効な）を求めよ”

ということになる．このとき一組の確率変数 (x, y) の間には，

(11)　$$y=\varphi(x)+\varepsilon(x)$$

なる関係があるから（$\varepsilon(x)$ は x を固定したとき，回帰線 $\varphi(x)$ のまわりの y の変動を表わす確率変数），両辺の期待値をとると，

(12)　$$\bar{Y}=E\{\varphi(x)\}\equiv\bar{\varphi}$$

となる．したがって，

(13)　$$y-(\varphi(x)-\bar{\varphi})=\bar{Y}+\varepsilon(x)$$

なる関係がえられる．この両辺の期待値をとると，

(14)　$$E\{y-(\varphi(x)-\bar{\varphi})\}=\bar{Y}$$

となるが，一般に $\varphi(x)$ は未知であるから，(13) 式よりすぐに $\bar{Y}$ の不変推定量はえられない．そこで，観測値 $\{x_i, y_i)$；$i=1, \cdots, n\}$ がえられたとき，それにもとづいて y の x への回帰線 $\varphi(x)$ の推定曲線を $\varphi_n(x)$ つくり（最小自乗法によって），それより統計量

(15)　$$t=\bar{y}-\frac{1}{n}\sum_{i=1}^{n}\{\varphi_n(x_i)-E(\varphi_n(x_i))\}$$

を求めると，これは $\bar{Y}$ の漸近的不変推定量となり，その漸近分散 σ_t^2 は，

(16)　$$\sigma_t^2=\frac{\sigma_Y^2}{n}(1-\eta_{XY}^2)$$

となることが見通せる（η_{XY} は母集団における相関比）．

回帰線 $\tilde{Y}_X=\varphi(X)$ が l 次の多項式

(17)　$$\varphi(X)=\sum_{k=0}^{l}A_kX^k$$

で表わされると仮定すると，サンプルの観測値 $\{(x_i, y_i)$；$i=1, 2, \cdots, n\}$ にもとずく $\varphi(X)$ の近似曲線は，

(18)　$$S=\frac{1}{n}\sum_{i=1}^{n}\left\{y_i-\sum_{k=0}^{l}A_kx_i^k\right\}^2$$

を最小にするように係数 A_k の値 a_i を求めることによって定まる：

(19)　$$\varphi_n(x)=\sum_{k=0}^{l}a_kx^k$$

実際に (18) を最小にする係数 a_k は，$(l+1)$ 元連立一次方程式

(20)　$$\sum_{k=0}^{l}m_{j+k}a_k=r_j,\quad j=0, 1, \cdots, l$$

ここに，$$r_j=\frac{1}{n}\sum_{i=1}^{n}x_i^jy_i,$$

$$m_{j+k}=\frac{1}{n}\sum_{i=1}^{n}x_i^{j+k},\quad j, k=0, 1, \cdots, l$$

の根としてえられ，この (20) 式は母集団における回帰方程式の係数 A_j がみたしている関係式

(21) $$\sum_{k=0}^{l}\mu_{j+k}A_j=\rho_j,\ j=0,\ 1,\ \cdots,\ l$$

ここに $\rho_j=E(x^jy)$, $\mu_{j+k}=E(x^{j+k})$

に対応する．(20) 式をみたす根 $\{a_k(n)\ ;\ k=0,\ 1,\ \cdots,\ l\}$ は，

(22) $$a_k(n)=\sum_{j=0}^{l}r_jD_{jk}/D,\ (k=0,\ 1,\ \cdots,\ l)$$

ただし，D_{jk} は行列式 $D=|m_{j+k}|$ の jk 余因子とする．

とかけ，この期待値は

(23) $$E\{a_k(n)\}=A_k$$

となる．くわしくいうと，$(x_1,\ \cdots,\ x_n)$ を固定して，$(y_1,\ \cdots,\ y_n)$ についての期待値をとる操作を $E_y(\cdot\,|x)$ とかくことにすると，

$$E_y\{a_k(n)|x\}=\sum_{j=0}^{l}E_y(r_j|x)D_{jk}/D,$$

とかけ，ここに

$$\begin{aligned}E_y(r_j|x)&=\frac{1}{n}\sum_{i=1}^{n}x_i^jE_y(y_i|x)\\&=\frac{1}{n}\sum_{i=1}^{n}x_i^j\left(\sum_{h=0}^{l}A_hx_i^h\right)\\&=\sum_{h=0}^{l}A_hm_{j+h}\end{aligned}$$

であり，

$$\begin{aligned}E_y\{a_k(n)|x\}&=\sum_{j=0}^{l}\left(\sum_{h=0}^{l}A_hm_{j+h}\right)D_{jk}/D\\&=\sum_{h=0}^{l}A_h\left(\sum_{j=0}^{l}m_{j+k}D_{jk}/D\right)\\&=\sum_{h=0}^{l}A_h\delta_{hk}=A_k\end{aligned}$$

となる．

さて，確率変数

(24) $$T=y-(\varphi(x)-\bar{\varphi})$$

の分散は，

(25) $$\sigma_T^2=E\{\varepsilon^2(x)\}=\sigma_Y^2(1-\eta_{YX}^2)$$

であるから，確率変数

(26) $$T_n=\bar{y}-\frac{1}{n}\sum_{i=1}^{n}\{\varphi(x_i)-\bar{\varphi}\}$$

の分散は，

(27) $$\sigma^2T_n=\frac{1}{n}\sigma_Y^2(1-\eta_{YX}^2)$$

とかける．したがって，回帰線 $\varphi(X)$ が (17) 式のような多項式であらわされるときは，

(26′) $$T_n=\bar{y}-\sum_{k=0}^{l}A_k(m_k-\mu_k)$$

となるから，係数 A_k をその不偏推定量 $a_k(n)$ でおきかえた統計量

(28) $$t_n=\bar{y}-\sum_{k=0}^{l}a_k(n)(m_k-\mu_k)=\sum_{k=0}^{l}a_k(n)\mu_k$$

ただし μ_k は既知とする $(k=0,\ 1,\ \cdots,\ l)$

は，$\bar{Y}$ の不偏推定量で，かつその漸近分散は (27) となることが期待される（実際 $a_k(n)$ は A_k に対応する一致統計量であるから，それがいえる）．この t_n の正確な分散は，

$$\sigma^2_{t_n}=\sum_{k=0}^{l}\sigma^2_{a_k(n)}\mu_k^2+\sum_{k\neq h}\sigma_{a_k(n),a_h(n)}\mu_k\mu_h \tag{29}$$

ここに $\sigma_{a_k(n),a_h(n)}$ は $a_k(n)$ と $a_h(n)$ との共分散をあらわす．

とかけるが，これをさらに精密に求めるのはむつかしい．

さて，A_k がみたしている関係式 (21) をながめると，別の推定量が考えられる．すなわち，(21) より

$$A_k=\sum_{j=0}^{l}\rho_j\Delta_{jk}/\Delta,\quad \Delta=|\mu_{j+k}|, \tag{30}$$

ここに，$\Delta=|\mu_{j+k}|$，Δ_{jk} は Δ の jk 余因子とする

とかけるから，ρ_j の不変推定量 $r_j(n)=\dfrac{1}{n}\sum_{i=1}^{n}x_i^jy_i$ を (30) に入れたものを

$$\tilde{\boldsymbol{a}}_k(n)=\sum_{j=0}^{l}r_j(n)\Delta_{jk}/\Delta, \tag{31}$$

とし，$\bar{Y}$ の推定量として，

$$\tilde{t}_n=\bar{y}-\sum_{k=0}^{n}\tilde{\boldsymbol{a}}_k(n)(m_k-\mu_k) \tag{32}$$

を考えると，その期待値は

$$E(\tilde{t}_n)=\bar{Y}+\frac{N-n}{n(N-1)}\sum_{k=0}^{l}\left(A_k\mu_k-\sum_{j=0}^{l}\rho_{j+k}\Delta_{jk}/\Delta\right) \tag{33}$$

となり，この第2項は推定の偏りを表わす．したがって，

$$t_n^*=\bar{y}-\sum_{k=0}^{l}\tilde{\boldsymbol{a}}_k(n)(m_k-\mu_k)-z(n), \tag{34}$$

$$\text{ここに } z(n)=\frac{N-n}{n(N-1)}\sum_{k=0}^{l}\left\{\tilde{\boldsymbol{a}}_k(n)\mu_k-\sum_{j=0}^{l}r_{j+k}(n)\Delta_{jk}/\Delta\right\}$$

は $\bar{Y}$ の不偏推定量となり，その漸近分散は (16) 式で与えられる．

§3. むすび

いままで用いられてきたさまざまな比推定量や回帰推定量を概観すると，"母集団における Y の X への回帰線が直線である" という前提のもとに，それらの推定量の有効性がみとめられるのであるが（比推定の場合にはさらに，"回帰直線が原点または原点の付近を通る" という条件がつけ加えられねばならない），一般に回帰線が直線であるという保証はないから，回帰線が直線である場合はもちろんのこと，曲線である場合にも適用しうる回帰推定量をつくるという点に重点をおいて考えてきた．一方，そのような場合に，層別抽出法や重複抽出法を加味して，従来の比推定量や回帰推定量の有効性を改良してみく方法も考えられている（[8]，[9]，[10] 参照）．しかし，母集団を層別するには，Y の X への回帰線がはっきりわかっていないと有効な層別ができないが，そのような知識があれば $\bar{Y}$ 自体がわかってしまい，$\bar{Y}$ の推定をいうことは意味がなくなってしうという矛盾をもつ．その欠点を改良するため，比較的多くの第1次サンプルをとって層別を行ない，その上で各層より第2次サンプルをとることにすれば，こんどは層別の基準（X の分点 x_i）が変動するから，各層にふくまれる第1次のサンプル数が変動することになり，推定量の精度の保証がむつかしくなるという欠点をもつようになる．それにくらべると，上に述べた一般化された回帰推定量の方があつかいやすく，しかも精度の保証が確実であり，層別という手数を省略できるという利点をもつといえる．

なお，問題を一般化すると，次のようになるであろう：

"母集団における2つの変量 X, Y が分布函数 $F(X, Y; \theta)$ にしたがっており, X についてのいくつかのモーメントがわかっているとき, その母集団からとられた大きさ n のサンプル $\{(x_i, y_i); i=1, \cdots, n\}$ の観測値にもとずいて, Y のモーメントを推定するために有効な推定量を求めたい".

この線にそった研究を今後もつづけてゆく予定である. (統計数理研究所)

参 考 文 献

[1] Hartley, H. O. and Ross, A., The unbiased ratio estimators, Nature, 174 (1954), pp. 270-271.

[2] Goodman, L. A., and Hartley, H. O., The precision of unbiased ratio-type estimators, J. Amer. Stat. Assoc., Vol. 53 (1958), p.p. 491-509.

[3] Robson, D. S., Application of multivariate polykays to the theory of unbiased ratio-type estimation, J. Amer. Stat. Assoc., Vol. 52 (1957), p.p. 511~522.

[4] Robson, D. S. and C. Vithayasai, Unbiased componentwise ratio estimator, J. Amer. Stat. Assoc. Vol. 56 (1961), p.p. 350~358.

[5] Olkin, Multivariate ratio estimation for finite populations, Biometrika, Vol. 45 (1958), p.p. 154~165.

[6] Sukhatme, B. V., Some ratio-type esimators in two-phase sampling, J. Amer. Stat. Assoc., Vol. 57 (1962), p.p. 628-632.

[7] Mickey, M. R., Some finite population unbiased ratio and regression estimators, J. Amer. Stat. Assoc., Vol. 54 (1959), p.p. 594-612.

[8] Raj, D., On the relative accuracy of some sampling techniques, J. Amer. Stat. Assoc., Vol. 53 (1958), p.p. 98-101.

[9] Mokashi, V. K., Efficiency of stratification in sub-sampling designs for the ratio method of estimation, J. Indian Soc. Agric. Stat., Vol. 6 (1954), p.p. 77-82.

[10] Rao, J. N. K., On the estimation of the relative efficiency of sampling procedures, A. Inst. Stat. Math., Vol. 14 (1962), p.p. 143-150.