集計データによる偏よりについて

中 山 慶 一 郎

1．問題の所在

回帰分析において，個別データの使用が可能でない時，縮約された集計データを用いることが多い。例えば，個々の世帯の調査データを用いてエンゲル関数を推定する場合のモデルは

$$Y_i=\alpha+\beta X_i+\varepsilon_i \quad i=1,2,\cdots,n \qquad (1.1)$$

で表わされる。X_i は独立変数で各世帯の所得を示し，Y_i は X_i に対応する世帯の支出である，ε_i は誤差項で，α，β は推定すべきパラメータである。集計データでは，通常平均データを用いることが多い。そのモデルは

$$\overline{Y}_g=\alpha+\beta\overline{X}_g+\bar{\varepsilon}_g \quad g=1,2,\cdots,n \qquad (1.2)$$

であり，独立変数 $\bar{X}_g$ は，あるグループ g の世帯の平均所得であり，従属変数 $\bar{Y}_g$ は同じグループの平均支出である。

集計データでは情報が縮約されているためにパラメータの推定に偏より（bias）が生ずる。これを集積による偏より（aggregation bias）という。集団レベルのモデル（macro model）から，個人レベルのモデル（micro model）を推論するときには誤まった推論を行うことが多い。この小論はこの点を明確にしようと試みるものである。

2．モデルの設定

ここで取りあげるモデルは独立変数が1つのときと，2つの場合に限定して議論する。

最も単純な場合として1変数モデルを考える。ミクロモデルとして

$$Y_i=\alpha+\beta X_i+\varepsilon_i \quad i=1,2,\cdots,n \qquad (2.1)$$

ただし，i はデータ数を示す。

マクロモデルとして

$$\overline{Y}_g=\alpha+\beta\overline{X}_g+\bar{\varepsilon}_g \quad g=1,2,\cdots,G \qquad (2.2)$$

とする。ただし，g は分類のグループとする。ここで誤差項 ε_i，$\bar{\varepsilon}_g$ について，以下の仮定を定める。

(1) $E(\varepsilon_i)=E(\bar{\varepsilon}_g)=0$

(2) $E(\varepsilon_i\varepsilon_j)=\begin{cases}0 & i\neq j\\ \sigma^2 & i=j\end{cases}$

$$E(\bar{\varepsilon}_g\bar{\varepsilon}_k)=\begin{cases}0 & g\neq k\\ \dfrac{\sigma^2}{n_g} & g=k\end{cases}$$

(3) $E(X_i\varepsilon_i)=0$

$E(\bar{X}_g\bar{\varepsilon}_g)=0$

これらの仮定は通常誤差項に設定されるものと同じである。(2)は誤差項が互いに独立で，分散は

---

homoscedasticity の場合パラメータの推定量は有効（efficient）であるが，heteroscedasticity のときは，有効ではない。

パラメータ β の推定量 b_{YX} が不偏推定であることは，次のようにして証明される。$Y_i=\alpha+\beta X_i+\varepsilon_i$，$\bar{Y}=\alpha+\beta\bar{X}+\bar{\varepsilon}$ とすれば，$Y_i-\bar{Y}=y$，$X_i-\bar{X}=x$，$\varepsilon_i-\bar{\varepsilon}=\varepsilon$ とすれば，$y=\beta x+\varepsilon$ となる。従って

$$b_{YX}=\frac{\Sigma xy}{\Sigma x^2}=\frac{\Sigma x(\beta x+\varepsilon)}{\Sigma x^2}=\beta+\frac{\Sigma x\varepsilon}{\Sigma x^2}=\beta+\frac{s(x,\varepsilon)}{s(x^2)}$$

となる。ここで $s(x,\varepsilon)$ は独立変数Xと誤差項の共分散であり，$s(x^2)$ はXの分散である。

$$E(b_{YX})=\beta+E\left[\frac{s(x,\varepsilon)}{s(x^2)}\right]$$

従って仮定(3)が成立すると，上式の右辺の第2項は零となり，$E(b_{YX})=\beta$ となって不偏推定であることがわかる。

マクロモデルも同様に $E(\bar{b}_{YX})=\beta$ となる。

個々のミクロモデルで同一である(homoscedasticity)。マクロモデルでは分散は$\frac{\sigma^2}{n_g}$でグループ毎に異なる（heteroscedasticity)。各グループのデータ数が等しいとき，$N_g=n$ $(g=1,2,\cdots,G)$ならば等分散性（homoscedasticity）が成立する。仮定(3)は独立変数X が固定した数である非確率変数であることを意味している。この仮定が成立しないときは，パラメータβの推定量は偏よりが生じ，誤差項εと独立変数Xに相関のある変数がモデルから落ちていることを意味する。

モデルの設定には定式化誤差（specification error）が発生する。モデルが定式化誤差を持たないということは，方程式の誤差項と独立変数とが相関を持たないことであり，仮定(3)が成立する。従って定式化誤差を持つモデルは必要な独立変数がモデルにとり入れられていない場合であり，仮定(3)が成立しない。この定式化誤差の大きさは，モデルが正しいときのパラメータの不偏推定量と，モデルが正しくないときのパラメータの偏よった推定量の期待値の差に等しくなる。

次にデータのグループ化による集計の偏よりについて考えよう。モデル(2.2)は集計データによるものであるが，データのグループ化には四つのパターンが考えられる［3］。

(1)独立変数Xの値によるグループ化。

(2)誤差項εの値によるグループ化。

(3)従属変数Yの値によるグループ化。

(4)XとYと相関のある変数Zの値によるグループ化。

(1)と(2)によるグループ化は定式化誤差をもたらさないが，(3)と(4)によるものは定式化誤差をもたらす。したがって集計による偏よりは定式化誤差によって表わされる。

以下ではミクロモデルが正しい場合と正しくない場合について，四つのグループ化に関する偏より（bias）について考察することにする。

3．正しく設定されたミクロモデルにおける集計の偏より

ミクロレベルにおける正しいモデルを

$$Y=\beta X+\varepsilon \qquad (3.1)$$

とし，マクロレベルにおけるモデルを

$$\bar{Y}=\beta\bar{X}+\bar{\varepsilon} \qquad (3.2)$$

とする。誤差項εについての仮定は

$$E(\varepsilon)=0,\ V(\varepsilon)=\sigma^2,\ E(X,\varepsilon)=0$$

である。(3.1)のβの推定量b_{YX}を最小2乗法で求めると，

$$b_{YX}=\beta+\frac{\Sigma x\varepsilon}{\Sigma x^2}$$

となり，期待値をとると

$$E(b_{YX})=\beta+E\left[\frac{s(x,\varepsilon)}{s(x^2)}\right] \qquad (3.3)$$

従って正しいミクロモデルは

$$E(b_{YX})=\beta \qquad (3.4)$$

となり不偏推定量が得られる。集計データを用いるマクロモデル(3.2)のβの推定量も

$$E(\bar{b}_{YX})=\beta+E\left[\frac{s(\bar{x},\varepsilon)}{s(\bar{x}^2)}\right] \qquad (3.5)$$

となり，$E[s(\bar{x},\varepsilon)]$が零であるかないかによって偏よりがないか，偏よりが生じるかということになる。

(a). 独立変数xの値によるグループ化。

独立変数Xによる分類は，ミクロレベルではXとεは相関をもたないと仮定されているので独立変数と誤差項は独立である。マクロレベルでもXによって分類された$\bar{X}$と$\bar{\varepsilon}$は相関をもたず独立である。例えば所得水準Yが両親の社会経済的地位Xによって決まるモデルを考えたとき，住居地域で分類したデータを用いると，住居地域による分類が社会経済階層による分類に近似しているものと考えられる。同じ地域のデータは比較的同質で誤差項と独立変数Xとは相関を持たないと仮定出来る。即ち，$E[s(\bar{x},\bar{\varepsilon})]=0$ で $E[b_{YX}]=E[\bar{b}_{YX}]=\beta$となる。

(b). 誤差項εの値による分類。

データが誤差項の値によって分類することは，データをランダムに分類することを意味する。実験データの分類のような特殊な場合しか，このような分類は行われない。社会調査データにこのような分類がなされることはほとんどない。ランダム分類であるので誤差項の仮定(3)は満されている。$E[s(\bar{x},\bar{\varepsilon})]=0$ でβの推定量は不偏である。

(c). 従属変数Yの値による分類。

前述の例で従属変数である所得水準Yでデータを分類すれば，両親の社会経済的地位Xと相関が

生じ，その結果 $\bar{X}$ と $\bar{\varepsilon}$ との間に相関が生じる。従って，$E[s(\bar{x}, \bar{\varepsilon})] \neq 0$ で β の推定量 $\bar{b}_{YX}$ の期待値は，

$$E(\bar{b}_{YX}) = \beta + E\left[\frac{s(\bar{x}, \bar{\varepsilon})}{s(\bar{x}^2)}\right]$$

であるから，$\bar{X}$ と $\bar{\varepsilon}$ の相関が正か負になるにつれて，$E(\bar{b}_{YX})$はパラメータ β より大きくなったり，小さくなったりする偏よりを生ずることになる。その偏よりの大きさは

$$|\beta - E(\bar{b}_{YX})| = \left| E\left[\frac{s(\bar{x}, \bar{\varepsilon})}{s(\bar{x}^2)}\right] \right| \qquad (3.6)$$

となる。Yによる分類はYの値の変動に応じてXと ε の値が変動する。

(d). 従属変数Yと独立変数Xによらない変数Zの値による分類。

X，Y以外の第3の変数Zに従って分類された場合，マクロモデルの正しい定式化は独立変数 $\bar{Z}$ を導入して

$$\bar{Y} = \alpha + \beta_{YX \cdot Z}\bar{X} + \beta_{YZ \cdot X}\bar{Z} + \bar{\varepsilon} \qquad (3.7)$$

となる。このとき

$$E(\bar{b}_{YX \cdot Z}) = E(b_{YX}) = \beta$$

となるので，$\beta_{YX \cdot Z}$ の推定量 $\bar{b}_{YX \cdot Z}$ は不偏でありミクロモデルの推定量に等しい。ここでマクロモデルで独立変数 $\bar{Z}$ を除いて

$$\bar{Y} = \alpha + \beta_{YX}\bar{X} + \bar{\varepsilon} \qquad (3.8)$$

とすれば，マクロモデルは誤まって定式化されたものとなる。すなわち，適切な説明変数を省略した為に生じる定式化誤差（specification error）が発生する。

$Y = \alpha + \beta_{YX \cdot Z}X + \beta_{YZ \cdot X}Z + \varepsilon$ とすると

$$E(b_{YZ \cdot X}) = E\left[\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}\right] \qquad (3.8)$$

$$Y - \bar{Y} = \beta_{YX \cdot Z}(X-\bar{X}) + \beta_{YZ \cdot X}(Z-\bar{Z}) + (\varepsilon_i - \bar{\varepsilon})$$

を(3.8)に代入すると

$$E(b_{YZ \cdot X}) = \beta_{YX \cdot Z} + \beta_{YZ \cdot X}E\left[\frac{\Sigma(X-\bar{X})(Z-\bar{Z})^2}{\Sigma(X-\bar{X})^2}\right] \qquad (3.9)$$

従って誤差の大きさは

$$E(b_{YX}) - \beta_{YX \cdot Z} = \beta_{YZ \cdot X}b_{ZX} \qquad (3.10)$$

となる。ミクロモデルでは，$Y = \alpha + \beta_{YX}X + \varepsilon$ が正しいモデルであるので $\beta_{YZ \cdot X} = 0$ となるが，マクロモデルでは $\bar{X}$ と $\bar{Z}$ との間に相関が生ずるので，$\bar{b}_{ZX} \neq 0$ でなく，定式化誤差が生じ，その大きさは

$$E(\bar{b}_{YX}) - \beta_{YX \cdot Z} = \beta_{YZ \cdot X}\bar{b}_{ZX}$$

となる。マクロモデルではZの大きさに従ってXが分類されるために $\bar{b}_{ZX}$ は零にならない。マクロモデルより除いた説明変数Zと，モデルに含まれている説明変数が独立であれば，$\bar{b}_{YX}$ は不偏であるが，Zによって分類されるため一般にXとZは相関をもつ。$\bar{b}_{YX}$ の偏よりの方向は，$\beta_{YZ \cdot X}$ の符号によって決まる。

4．誤まって設定されたミクロモデルにおける集計の偏より。

ミクロレベルにおける正しいモデルは

$$Y = \alpha + \beta_{YX \cdot Z}X + \beta_{YZ \cdot X}Z + u \qquad (4.1)$$

で，誤まったモデルは

$$Y = \alpha + \beta_{YX}X + v \qquad (4.2)$$

とする。

マクロレベルにおける正しいモデルは

$$\bar{Y} = \alpha + \beta_{YX \cdot Z}\bar{X} + \beta_{YZ \cdot X}\bar{Z} + \bar{u} \qquad (4.3)$$

で，誤まったモデルは

$$\bar{Y} = \alpha + \beta_{YX}\bar{X} + \bar{v} \qquad (4.4)$$

とする。このときミクロモデルの定式化誤差は

$$E(b_{YX}) - \beta_{YX \cdot Z} = \beta_{YZ \cdot X}b_{ZX} \qquad (4.5)$$

で，マクロモデルの定式化誤差は

$$E(\bar{b}_{YX}) - \beta_{YX \cdot Z} = \beta_{YZ \cdot X}\bar{b}_{ZX} \qquad (4.6)$$

である。集計によるマクロモデルの偏よりは，二つのモデルの定式化誤差の差で表わされ，

$$\beta_{YZ \cdot X}(\bar{b}_{ZX} - b_{ZX}) \qquad (4.7)$$

である。この場合定式化誤差(specification bias)は常に存在し，分類による集計の偏より(aggregation bias)はマクロレベルで発生し，$\beta_{YZ \cdot X}$ と $\bar{b}_{ZX}$ と b_{ZX} の差に依存する。

(a)独立変数Xによる分類

モデルに含まれている独立変数Xによって分類されたデータを考える。この場合 $\bar{b}_{zx}$ と b_{zx} は等しいので集計による偏よりは生じない。この点を(4.1)，(4.2)で考えてみる。まずZのXに対する回帰式を

$$Z_i = \beta X_i + w_i \qquad (4.8)$$

とする。(4.8)に最小2乗法を適用して，β の推定量 b_{zx} は

$$b_{zx} = \frac{\Sigma X_i Z_i}{\Sigma X_i^2} \qquad (4.9)$$

となる。ここで

$$E(b_{zx})=E\left[\frac{\Sigma X_i Z_i}{\Sigma X_i^2}\right]=E\left[\frac{\Sigma X_i(\beta X_i+w_i)}{\Sigma X_i^2}\right]$$
$$=\beta+E\left[\frac{\Sigma X_i w_i}{\Sigma X_i^2}\right]$$

XはWと独立であるので

$$E(b_{zx})=\beta$$

となる。マクロモデルも同様に

$$E(\bar{b}_{z\bar{x}})=\beta$$

となる。従って，$\beta_{YZ\cdot X}(\bar{b}_{zx}-b_{zx})=0$ となって，集計による偏よりは存在しない。

(b)省略された独立変数Zによる分類

この場合，モデル(4.2)，(4.4)は定式化誤差が生じその大きさは，$\beta_{YX\cdot Z}b_{zx}$，$\beta_{YZ\cdot X}\bar{b}_{zx}$である。独立変数XはZに従って分類されるので，(a)のときとは異なりZに関連した誤差項wとXは独立にならないので，$b_{zx}\neq\bar{b}_{zx}$ で分類による誤差(aggregation bias)が生ずる。分類による誤差の大きさは(4.7)によって定まる。

(c). 独立変数と相関をもつ変数による分類。

一般に考えられる分類は他に第三の変数によるものであるが，XとZに相関関係のある変数Aに基づく分類であることが多い。この場合，ミクロの定式化誤差は

$$E(b_{YX})-\beta_{YX\cdot Z}=\beta_{YZ\cdot X}b_{ZX}$$

で，マクロの定式化誤差は

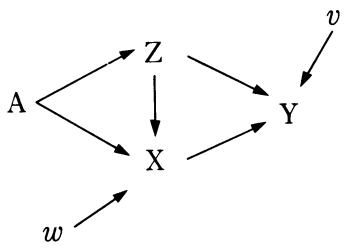
$$E(\bar{b}_{YX})-\beta_{YX\cdot Z}=\beta_{YZ\cdot X}\bar{b}_{ZX}$$

であって，分類による集計の偏よりは，

$$\beta_{YZ\cdot X}(\bar{b}_{ZX}-b_{ZX})$$

となる。

このときのモデルとして



と考えると，独立変数間に次のような回帰モデルが考えられる。

$$Z=\beta_{ZX\cdot A}X+\beta_{ZA\cdot X}A+w \qquad (4.10)$$
$$Z=\beta_{ZX}X+u \qquad (4.11)$$

(4.11)から，β_{ZX}の推定量 b_{ZX} は

$$b_{ZX}=\frac{\Sigma XZ}{\Sigma X^2}=\frac{\Sigma X(\beta_{ZX\cdot A}X+\beta_{ZA\cdot X}A+w)}{\Sigma X^2}$$
$$=\beta_{ZX\cdot A}+\beta_{ZA\cdot X}\cdot\frac{\Sigma XA}{\Sigma X^2}+\frac{\Sigma Xw}{\Sigma X^2}$$

となり，XとAは相関があり，Xとwは独立であるので，

$$E(b_{ZX})=\beta_{ZX\cdot A}+\beta_{ZA\cdot X}b_{AX} \qquad (4.12)$$

となる。マクロレベルも同様に考えると，

$$E(\bar{b}_{ZX})=\beta_{ZX\cdot A}+\beta_{ZA\cdot X}\bar{b}_{AX} \qquad (4.13)$$

となる。

(4.13)を利用すると集計の偏よりは

$$\beta_{YZ\cdot X}(\bar{b}_{ZX}-b_{ZX})=\beta_{YZ\cdot X}\beta_{ZA\cdot X}(\bar{b}_{AX}-b_{AX}) \qquad (4.14)$$

となる。

推定量の変動について

推定量の偏よりの他に考察する必要があるものは推定値の変動の大きさである。この点についてはCramer［1］によって論じられている。

モデルを次のように設定しよう。元のミクロモデルを

$$Y_{ig}=\alpha+\beta X_{ig}+\varepsilon_{ig}$$
$$i=1,2,\cdots,n_g;\ g=1,2,\cdots,G \qquad (5.1)$$

とする。添字igはg番目のグループのi番目のデータを意味する。従って，$N=\Sigma_{ng}$である。

マクロモデルは

$$\bar{Y}_g=\alpha+\beta\bar{X}_g+\bar{\varepsilon}_g \qquad (5.2)$$

であり，$\bar{Y}_g$，$\bar{X}_g$はg番目のグループの平均値である。データをグループ化することにより全体のデータの変動は，分散分析の考えに従うと，グループ内分解とグループ間分散との和に分解出来る。

$$\sum_i\sum_g(X_{ig}-\bar{X})^2=\sum_i\sum_g(X_{ij}-\bar{X}_g)^2+\sum_g n_g(\bar{X}-\bar{X})^2 \qquad (5.3)$$

(5.3)を利用すると，(5.1)からβの推定量$\hat{\beta}$の分散は，

$$Var(\hat{\beta})=\frac{\sigma^2}{\sum_i\sum_g(X_{ig}-\bar{X})^2} \qquad (5.4)$$

となり，(5.2)からβの推定量$\tilde{\beta}$の分散は，

$$Var(\tilde{\beta})=\frac{\sigma^2}{\sum_g n_g(\bar{X}_g-\bar{X})^2} \qquad (5.5)$$

となる。

この二つの分散の比は

$$\frac{\mathrm{Var}(\tilde{\beta})}{\mathrm{Var}(\hat{\beta})}=1+\frac{\sum_i\sum_g(X_{ig}-\bar{X}_g)^2}{\sum_g n_g(\bar{X}_g-\bar{X})^2} \qquad (5.6)$$

右辺の第2項が分類による推定値の変動の大きさ，即ちデータからの情報量の損失を表わすもので，この比率はグループ間変動とグループ内変動の比である。

次に定式化誤差をもつモデルの推定値の変動について，竹内［2］に従って考察することにする。

正しいモデルを

$$Y_i=\beta_1X_{1i}+\beta_2X_{2i}+\varepsilon_i \quad i=1,\cdots,n \qquad (5.7)$$

とし，誤まったモデルを

$$Y_i=\alpha X_{1i}+u_i \quad i=1,\cdots,n \qquad (5.8)$$

とする。u_iは単なる誤差項でないので，X_2に対するX_1の回帰を考える。

$$X_{2i}=\gamma X_{1i}+v_i \qquad (5.9)$$

(5.9)に最小2乗法を適用し

$$\gamma=\frac{\Sigma X_{1i}X_{2i}}{\Sigma X_{1i}^2} \qquad (5.10)$$

となる。

$$\Sigma v_iX_{1i}=\Sigma(X_{2i}-\gamma X_{1i})X_{1i}=0$$

であるから，v_iとX_{1i}は独立である。

(5.9)を(5.7)に代入して

$$Y_i=\beta_1X_{1i}+\beta_2(\gamma X_{1i}+v_i)+\varepsilon_i$$
$$=(\beta_1+\beta_2\gamma)X_{1i}+\beta_2v_i+\varepsilon_i \qquad (5.11)$$

$\beta_1+\beta_2\gamma$の最小2乗推定量は

$$(\hat{\beta}_1+\hat{\beta}_2\gamma)=\frac{\Sigma Y_iX_{1i}}{\Sigma X_{1i}^2}=\hat{\alpha} \qquad (5.12)$$

で，(5.8)のαの推定量に等しい。

$$\mathrm{Var}(\hat{\alpha})=\frac{\sigma^2}{\Sigma X_{1i}^2}=\mathrm{Var}(\beta_1+\beta_2\gamma)$$

従って，X_1とX_2が独立であれば，$\gamma=0$となって$\mathrm{Var}(\hat{\alpha})=\mathrm{Var}(\hat{\beta}_1)$であるが，独立でないときは，$\mathrm{Var}(\hat{\alpha})>\mathrm{Var}(\hat{\beta}_1)$となる。

6．おわりに

本稿は主として「生態学的推論」［3］に従って，集計の偏よりについて統計学的な意味づけを述べたものである。Langbein と Lichtman はこの問題の解決に生態学的回帰を提案しているが，この点はここではふれない。

参考文献

［1］ Cramer, J.S.(1964), "Efficient grouping, regression and corelation in Engel curve analysis", J.A.S.A., 59

［2］ 竹内　啓（1973）数理統計学の方法的基礎　東洋経済

［3］ L.I. Langbein and A. J. Lichtman(1978), Ecological Inference. Sage Pub.

（生態学的推論　長谷川政美訳　朝倉書店）