個票データにおける分類の併合モデル

岡山商科大学商学部 佐井 至 道

東京大学大学院経済学研究科 竹村 彰 通

Some Models for Merging Groups in Microdata

Shido SAI and Akimichi TAKEMURA

要 旨 個票データを公開するか否かを決定する際の最も重要な指標の一つが、レコードの項目の値の組み合わせが他のすべての個体と異なる一意な個体の数であり、標本調査では特に標本でも母集団でも一意である個体数が問題となる。一般に元の個票データではほとんどの個体が標本一意であり、そのまま公表することはできず、質的変数についてカテゴリーを合併したり、量的変数について丸めやトップコーディングを行ったりする秘匿措置を行う必要がある。これらは広い意味で項目の分類の併合と考えられるが、それによる一意な個体数の減少率は状況によってかなり異なる。本論文では、各分類ごとには従来用いられているポアソンガンマモデルを用いるとともに、分類間にも新たにモデルを導入することによって、分類の併合前に、併合後の一意な個体数を予測する方法について検討を行う。また提案した方法を労働力調査の個票データに対して適用し、その実用性を見る。

1. はじめに

標本調査によって得られた個票データを公開する場合には、その中に含まれる個体のプライバシーが保護される必要がある。その検討の際に最も重要な指標となるのが一意な個体数である。キー変数（地域や性別のように第三者が識別するのに用いることができる変数）の値の組み合わせが他のすべての個体と異なる個体を一意と呼ぶが、標本調査では標本でも母集団でも一意である個体数が特に問題となる。標本内で一意であっても母集団で一意でなければ、第三者は個票データの一意な個体が母集団のどれであるかを特定することはできず、危険性は大きく減少する。逆に、母集団内で一意であっても標本としてサンプリングされなければ問題はない。

一般に元の個票データではほとんどの個体が標本一意であり、そのまま公表することは危険と考えられる。そのため個票データの項目が質的変数であればカテゴリーを合併したり、量的変数であれば丸めやトップコーディングを行ったりして一意な個体数を十分安全な値まで減少させる対策をとらねばならない。これらは広い意味で項目の分類の併合と考えられる。このような手法については Willenborg and de Waal (1996) が最も詳しい。また一意に関するこれまでの研究

Key words: Poisson-gamma model, population unique, sample unique, statistical disclosure control

の流れについては佐井 (1998b) も参照されたい。分類を併合すると標本で一意な個体数は一般に減少するものの、減少率は状況によってかなり異なる。一意な個体数を大きく減少させる効果的な併合はその作業を行う者の勘に頼る部分が多分にあり、標本かつ母集団一意である個体数がどの程度減少するのかを事前に推測するのは困難であるのが現状である。

本論文では、各分類ごとには従来用いられているポアソングンマモデルを用いるとともに、分類間にも新たにモデルを導入することによって、分類の併合についての検討を行う。これは層化抽出における数個の層や、単に数個の集団を併合する場合にもそのまま適用できる。

まず第2節では一意などに関するいくつかの定義をするとともに、ポアソングンマモデルを紹介する。次に第3節では2つの両極端なモデルである完全連動型モデルと独立型モデルを分類間に導入し、その比較を行う。完全連動型モデルは異なる分類で対応する特性の個体数が極めて強く関係するモデルであり、独立型モデルは全く関係しないモデルである。この2つのモデルを比較することにより、併合による一意な個体数の変化の予測がある程度可能になると思われるが、実際のデータへの当てはめという点から見るとまだ十分とは言えない。そこで第4節では2つのモデルの間を埋めるために相関型モデルと部分連動型モデルを提案し、それぞれの性質について考える。相関型モデルは異なる分類で対応する特性の個体数がある程度の関係を持つモデルであり、部分連動型モデルでは分類がいくつかのグループに分かれており、グループ内の分類の対応する特性の個体数は強く関係し、異なるグループでは独立となるモデルである。前者の方が現実のデータへの適用範囲は広いと予想されるが、両者の性質を考え合わせれば、かなり多くの実データに対して分類の併合前に併合後の標本でも母集団でも一意である個体数の予測が可能になると思われる。さらに第5節では総務庁統計局統計基準部が中心となって開催された「標本データの秘密保護に関する研究会」において、佐井が竹村の研究協力者として目的外使用した労働力調査の個票の集計データに対して相関型モデルを中心に適用し、モデルの当てはまりを見るとともに、分類の併合前における予測の実用性について検討することにする。

2. 各分類におけるモデル

2.1. 一意に関する定義

個票データに含まれる個体がキー変数によって L 個 ($L \geq 2$) の集団に分類されているものとする。(以下単に分類と呼ぶ。) その第 l 番目の分類に含まれる母集団の個体数を N_l 、標本の個体数を n_l とし、抽出率 n_l/N_l は分類に共通で λ とする。例えば単純無作為抽出によって標本が得られ事後的に分類された場合、各分類における抽出率は一般に異なるが、母集団と標本が小さくなければ共通と考えても大きな問題とはならない。なお第3節以降の種々の性質の検討においては、各分類の母集団の大きさ N_l がすべての分類に共通で N とする。また Takemura (1999) で論じられているように、以下で扱うポアソングンマモデルでは母集団の個体数は確率変数となるため N_l は厳密には母集団の個体数の期待値であり、抽出率 λ はベルヌーイ抽出法における抽出確率であるが、議論の混乱は生じないであろう。

項目の中のキー変数が質的変数であればそのカテゴリー数、量的変数であればそのとり得る値の数を、すべてのキー変数についてかけあわせて組み合わせた数をセル数と呼ぶが、各分類においてセル数は共通で K として、各セルに1番から K 番までの通し番号をつけておく。例えば分類が都道府県単位で行われ個体が個人のとき、年齢や性別などの組み合わせのセル数は一般に都道

府県ごとに異なるが、その場合にはセル数が最も大きな分類に揃えることにする。また第 l 番目の分類の第 i 番目のセルに含まれる母集団と標本の個体数をそれぞれ $F_{l,i}, f_{l,i}$ とする。

分類の併合は、異なる分類で対応する番号のセルの合併と考えられるが、併合後の第 i 番目のセルに含まれる母集団と標本の個体数をそれぞれ F_i, f_i とする。すなわち $F_i = \sum_{l=1}^L F_{l,i}, f_i = \sum_{l=1}^L f_{l,i}$ である。

またキー変数の組み合わせが他のすべての個体と異なる個体を一意と呼び、標本でも母集団でも一意である個体数の合計を、分類の併合の前は S_a 、併合後は用いるモデル別に S_{b_1}, S_{b_2} などと表すことにする。

2.2. ポアソンガンマモデル

一意性の議論を理論的に行うためには、個体のセルへの入り方を記述する超母集団モデルを考えると便利である。モデルには多項分布やポアソン分布を単独で用いるものや、それぞれの分布のパラメータが他のモデルに従うことを仮定する多項ディリクレモデルやポアソンガンマモデルなどがある。ポアソンガンマモデルは Bethlehem et al. (1990) によって提案され、Keller and Bethlehem (1992), Skinner (1992), 佐井 (1998b) においても用いられている。なお多項ディリクレモデルについては Takemura (1999), Omori (1999) を、この2つのモデルを含む数種類のモデルの関係については Hoshino and Takemura (1998) を参照されたい。またモデルを用いないノンパラメトリックな方法を Greenberg and Zayatz (1992), 加納 (1997), 渋谷 (1999) がそれぞれ個別に提案している。

本論文では、各分類ごとにポアソンガンマモデルを仮定することにする。第 l 番目の分類における第 i 番目のセルに含まれる母集団の個体数 $F_{l,i}$ ($l=1, 2, \dots, L; i=1, 2, \dots, K$) が互いに独立にポアソン分布

$$p(F_{l,i}) = \frac{e^{-N_l \pi_{l,i}} (N_l \pi_{l,i})^{F_{l,i}}}{F_{l,i}!}$$

に従うが、各セルへの入りやすさを表すパラメータ $\pi_{l,i}$ が同じ分類の他のパラメータと独立にガンマ分布

$$\pi_{l,i} \sim \text{gamma}(\alpha, \beta), \quad f(\pi_{l,i}) = \frac{1}{\Gamma(\alpha) \beta^\alpha} \pi_{l,i}^{\alpha-1} e^{-\pi_{l,i}/\beta}$$

に従うとするのがポアソンガンマモデルである。ただし α, β は正のパラメータでポアソンガンマモデルでは $\alpha\beta=1/K$ という制約をおいている。

なお多項分布の方がポアソン分布よりも自然なモデルであるが、佐井 (1998a) に示されているように両者から導かれる結果は一般にほとんど変わらず、しかもポアソン分布の方が扱いやすいという利点がある。ただしポアソン分布そのものを用いたモデルではパラメータの数が極めて多く、個票データからの推定が困難であるため、ガンマ分布の仮定が必要となる。ポアソンガンマモデルは実質的に $\pi_{l,i}$ の散らばりを表す1個のパラメータ β しか持たないが、後の例などを見る限り実際のデータへの当てはまりはそれほど悪くないようである。ポアソンガンマモデルの性質については 佐井 (1998b) を参照されたい。

このモデルの下で、標本でも母集団でも一意な個体数の期待値のすべての分類についての和を考えると次のように計算される。

$$\begin{aligned}
 E(\underline{S}_a) &= \sum_{l=1}^L \sum_{i=1}^K p(F_{l,i}=f_{l,i}=1) \\
 &= \sum_{l=1}^L \sum_{i=1}^K p(F_{l,i}=1) \cdot \lambda \\
 &= \sum_{l=1}^L \sum_{i=1}^K E(e^{-N_l \pi_{l,i}} N_l \pi_{l,i}) \cdot \lambda \\
 &= \sum_{l=1}^L \sum_{i=1}^K \int_0^\infty e^{-N_l \pi_{l,i}} N_l \pi_{l,i} \frac{\pi_{l,i}^{\alpha-1} e^{-\frac{\pi_{l,i}}{\beta}}}{\Gamma(\alpha) \beta^\alpha} d\pi_{l,i} \cdot \lambda \\
 &= \sum_{l=1}^L \sum_{i=1}^K \int_0^\infty N_l \frac{\pi_{l,i}^\alpha e^{-\frac{\pi_{l,i}}{\beta(1+N_l \beta)}}}{\Gamma(\alpha) \beta^\alpha} d\pi_{l,i} \cdot \lambda.
 \end{aligned}$$

ここで $t = (1 + N_l \beta) \pi_{l,i} / \beta$ とおくと、和は次のように求められる。

$$\begin{aligned}
 E(\underline{S}_a) &= \sum_{l=1}^L \sum_{i=1}^K \frac{N_l \beta}{(1 + N_l \beta)^{1+\alpha}} \int_0^\infty \frac{t^\alpha e^{-t}}{\Gamma(\alpha)} dt \cdot \lambda \\
 &= \sum_{l=1}^L \sum_{i=1}^K \frac{N_l \alpha \beta}{(1 + N_l \beta)^{1+\alpha}} \cdot \lambda \\
 &= \sum_{l=1}^L \sum_{i=1}^K \frac{\frac{N_l}{K}}{(1 + N_l \beta)^{1+\frac{1}{K\beta}}} \cdot \lambda \\
 &= \sum_{l=1}^L \frac{N_l}{(1 + N_l \beta)^{1+\frac{1}{K\beta}}} \cdot \lambda.
 \end{aligned}$$

ここで $N_l \equiv N$ とすると

$$E(\underline{S}_a) = \frac{LN}{(1 + N \beta)^{1+\frac{1}{K\beta}}} \cdot \lambda \quad (1)$$

と変形される。

3. 完全連動型モデルと独立型モデルの性質

3.1. 完全連動型モデルと独立型モデル

この節では、分類の併合後の一意な個体数について議論するために、異なる分類で同じ順番のセルのパラメータ $\pi_{l,i}$ ($l=1, 2, \dots, L$) について、次のような両極端な2つのモデルを考えることにする。

完全連動型モデル

$$\pi_{1,i} = \pi_{2,i} = \dots = \pi_{L,i} \quad (i=1, 2, \dots, K)$$

独立型モデル

$$\pi_{l,i} \text{ と } \pi_{l',i'} \text{ (} l \neq l', i \neq i') \text{ は独立} \quad (i=1, 2, \dots, K)$$

なおどちらのモデルでも $\pi_{l,i}$ と $\pi_{l',i'}$ ($l \neq l', i \neq i'$) はすべて独立である。完全連動型のモデルは、一つの分類のあるセルに属する個体数が他の分類の同じ順番のセルに属する個体数と極めて強く関係するモデルであり、独立型モデルは全く関係を持たないモデルである。例えば分類が都道府県単位で行われ個体が個人の場合、ある県で40歳の男性が多ければ他の県でも多いというように、対応するセルのパラメータ同士が深く関係する状況は多数考えられるため、完全連動型モデルの方がより現実的と思われるが、関係が実際の状況よりも強く設定されすぎること容易に

予想される。

2つのモデルを用いた場合の、分類の併合後の標本でも母集団でも一意である個体数の期待値をそれぞれ $E(\underline{S}_{b1})$, $E(\underline{S}_{b2})$ とする。分類の併合後に $F_i = f_i = 1$ となるためには、併合前のある第 l 番目の分類に対して $F_{l,i} = f_{l,i} = 1$ かつ $F_{l',i} = f_{l',i} = 0$ ($l' \neq l$) であるという性質を用いると、完全連動型モデルの下での期待値は次のように求められる。

$$\begin{aligned}
 E(\underline{S}_{b1}) &= \sum_{i=1}^K p(F_i = f_i = 1) \\
 &= \sum_{i=1}^K \sum_{l=1}^L p(F_{l,i} = f_{l,i} = 1, F_{l',i} = f_{l',i} = 0; \text{ all } l' \neq l) \\
 &= \sum_{i=1}^K \sum_{l=1}^L E(N_l \pi_{l,i} e^{-N_l \pi_{l,i}} \cdot \lambda \cdot \prod_{l' \neq l} e^{-N_{l'} \pi_{l',i}}) \\
 &= \sum_{i=1}^K \sum_{l=1}^L E\{N_l \pi_{l,i} e^{-(\sum_{l'=1}^L N_{l'} \pi_{l',i})} \cdot \lambda \\
 &= \sum_{i=1}^K \sum_{l=1}^L \frac{\frac{N_l}{K}}{\{1 + (\sum_{l'=1}^L N_{l'} \beta)^{\frac{1}{1+\frac{1}{K\beta}}}\}} \cdot \lambda \\
 &= \frac{\sum_{l=1}^L N_l}{\{1 + (\sum_{l=1}^L N_l \beta)^{\frac{1}{1+\frac{1}{K\beta}}}\}} \cdot \lambda.
 \end{aligned} \tag{2}$$

また独立型モデルの下での期待値も同様に求めることができる。

$$\begin{aligned}
 E(\underline{S}_{b2}) &= \sum_{i=1}^K \sum_{l=1}^L E(N_l \pi_{l,i} e^{-N_l \pi_{l,i}} \cdot \lambda \cdot \prod_{l' \neq l} e^{-N_{l'} \pi_{l',i}}) \\
 &= \sum_{i=1}^K \sum_{l=1}^L \{E(N_l \pi_{l,i} e^{-N_l \pi_{l,i}}) \cdot \lambda \cdot \prod_{l' \neq l} E(e^{-N_{l'} \pi_{l',i}})\} \\
 &= \sum_{i=1}^K \sum_{l=1}^L \left\{ \frac{\frac{N_l}{K}}{(1 + N_l \beta)^{1+\frac{1}{K\beta}}} \cdot \prod_{l' \neq l} \frac{1}{(1 + N_{l'} \beta)^{\frac{1}{K\beta}}} \right\} \cdot \lambda \\
 &= \sum_{l=1}^L \frac{N_l}{1 + N_l \beta} \cdot \prod_{l'=1}^L \frac{1}{(1 + N_{l'} \beta)^{\frac{1}{K\beta}}} \cdot \lambda.
 \end{aligned} \tag{3}$$

ここで $N_l \equiv N$ とすると期待値はそれぞれ次のように書き直される。

$$E(\underline{S}_{b1}) = \frac{LN}{(1 + LN \beta)^{1+\frac{1}{K\beta}}} \cdot \lambda, \tag{4}$$

$$E(\underline{S}_{b2}) = \frac{LN}{(1 + N \beta)^{1+\frac{L}{K\beta}}} \cdot \lambda. \tag{5}$$

3.2. 2つのモデルの比較

この節では完全連動型モデルと独立型モデルの下での併合による一意な個体数の減少の程度を見るために $N_l \equiv N$ の場合の (4), (5) 式と $E(\underline{S}_a)$ ((1) 式) との比

$$R_1 = \frac{E(\underline{S}_{b1})}{E(\underline{S}_a)} = \left(\frac{1 + LN \beta}{1 + N \beta} \right)^{-1-\frac{1}{K\beta}},$$

$$R_2 = \frac{E(\underline{S}_{b2})}{E(\underline{S}_a)} = (1 + N \beta)^{-\frac{L-1}{K\beta}}$$

について、その性質を考えることにする。

まず β の値と R_1, R_2 それぞれとの関係について、それぞれ次のような性質が求められる。

定理 1

$\beta \rightarrow 0$ のとき $R_1 \rightarrow e^{-(L-1)N/K}$, $\beta \rightarrow \infty$ のとき $R_1 \rightarrow L^{-1}$ である。

(i) $K \leq (L \log L / (L-1))N$ のとき R_1 は β に関する単調増加関数である。

(ii) $(L \log L / (L-1))N < K < ((L+1)/2)N$ のとき, R_1 は $0 < \beta < \beta_1$ において単調増加で, $\beta_1 < \beta$ において単調減少である。

(iii) $((L+1)/2)N \leq K$ のとき R_1 は β に関する単調減少関数である。

ただし β_1 は次の方程式を満たす正の解 β である。

$$\log \frac{1+LN\beta}{1+N\beta} - (K\beta^2 + \beta) \left(\frac{LN}{1+LN\beta} - \frac{N}{1+N\beta} \right) = 0.$$

証明 まず明らかに $\lim_{\beta \rightarrow \infty} R_1 = L^{-1}$ である。また

$$R_1 = \left\{ 1 + \frac{(L-1)N\beta}{1+N\beta} \right\}^{\frac{1+N\beta}{(L-1)N\beta} \cdot \frac{-(L-1)N(1+K\beta)}{K(1+N\beta)}} \rightarrow e^{-\frac{(L-1)N}{K}} \quad (\beta \rightarrow 0)$$

となる。

次に $\log R_1$ を β に関して偏微分すると

$$\frac{\partial \log R_1}{\partial \beta} = \frac{1}{K\beta^2} \left\{ \log \frac{1+LN\beta}{1+N\beta} - (K\beta^2 + \beta) \left(\frac{LN}{1+LN\beta} - \frac{N}{1+N\beta} \right) \right\}$$

となるが、右辺の後半の括弧内を $f_1(\beta)$ とおき再び β で微分すると

$$\begin{aligned} f_1'(\beta) &= \beta \left(\frac{LN}{1+LN\beta} - \frac{N}{1+N\beta} \right) \left\{ (K\beta+1) \left(\frac{LN}{1+LN\beta} + \frac{N}{1+N\beta} \right) - 2K \right\} \\ &= \frac{(L-1)N\beta}{(1+LN\beta)(1+N\beta)} \cdot \frac{N(2LN - KL - K)\beta + (LN + N - 2K)}{(1+LN\beta)(1+N\beta)} \end{aligned}$$

と書くことができる。ここで右辺の積の前半は $L \geq 2$ より明らかに正であり、積の後半の分母も明らかに正である。そこで積の後半の分子を

$$f_2(\beta) = N(2LN - KL - K)\beta + (LN + N - 2K)$$

とおき、 β の 1 次式と見て、次のように場合分けして考える。

(I) β の係数が 0 以上で、定数項が正のとき

$K \leq (2L/(L+1))N$ かつ $K < ((L+1)/2)N$ となるが、 $L \geq 2$ で $2L/(L+1) < (L+1)/2$ より K の範囲は $K \leq (2L/(L+1))N$ となる。

このとき直線 $y = f_2(\beta)$ の形状から $f_2(\beta) > 0$ ($\beta > 0$) で、同じ範囲で $f_1(\beta) > 0$ であることがわかる。よって $f_1(\beta)$ は同じ範囲で単調増加で $f_1(0) = 0$ より $f_1(\beta) > 0$ ($\beta > 0$) となる。したがって R_1 も $\beta > 0$ において単調増加であることがいえる。

(II) β の係数が負で、定数項が正のとき

この場合 K の範囲は $(2L/(L+1))N < K < ((L+1)/2)N$ となる。

ここで $f_2(\beta) = 0$ を満たす β の値を β_0 と置くと、 $f_2(\beta)$ は $0 < \beta < \beta_0$ で正、 $\beta_0 < \beta$ で負となり、

$f_1(\beta)$ についても同じことがいえる。したがって $f_1(\beta)$ は $0 < \beta < \beta_0$ において増加し、 $\beta_0 < \beta$ において減少することがわかる。

ところで、 $f_1(0)=0$ と $\lim_{\beta \rightarrow \infty} f_1(\beta) = \log L - (L-1)K/(LN)$ とを考え合わせると、 $\log L - (L-1)K/(LN) \geq 0$ すなわち $K \leq (L \log L / (L-1))N$ であれば $f_1(\beta) > 0$ ($\beta > 0$) となり R_1 は $\beta > 0$ で単調増加となる。なお $2L/(L+1) < L \log L / (L-1) < (L+1)/2$ となることが 2 式ずつの差をそれぞれ 2 回微分することにより求まるため、結局この場合分けの K の範囲の中で $(2L/(L+1))N < K \leq (L \log L / (L-1))N$ については (I) と合わせて定理の (i) の場合となる。

また $(L \log L / (L-1))N < K < ((L+1)/2)N$ の場合には $f_1(\beta)=0$ を満たす正の数 β_1 が存在し、 $0 < \beta < \beta_1$ において $f_1(\beta)$ は正より R_1 は単調増加で、 $\beta_1 < \beta$ において $f_1(\beta)$ は負より R_1 は単調減少となることが証明できる。これが定理の (ii) の場合となる。

(III) β の係数が負で、定数項が 0 以下のとき

この場合 K の範囲は $K \geq ((L+1)/2)N$ となる。

このとき直線 $y=f_2(\beta)$ の形状から $f_2(\beta) < 0$ ($\beta > 0$) で、(I) と同様にして R_1 は $\beta > 0$ において単調減少であることがいえる。これが定理の (iii) の場合である。

なお β の係数が 0 以上で定数が 0 以下となることはない。

定理 2 R_2 は β に関する単調増加関数で、 $\beta \rightarrow 0$ のとき $R_2 \rightarrow e^{-(L-1)N/K}$ 、 $\beta \rightarrow \infty$ のとき $R_2 \rightarrow 1$ である。

証明 まず

$$R_2 = (1 + N\beta)^{\frac{1}{N\beta} \left(-\frac{(L-1)N}{K} \right)}$$

と変形されるので、 $\lim_{\beta \rightarrow 0} R_2 = e^{-(L-1)N/K}$ 、 $\lim_{\beta \rightarrow \infty} R_2 = 1$ となることは容易にわかる。

次に $\log R_2$ を β に関して偏微分すると

$$\frac{\partial \log R_2}{\partial \beta} = \frac{L-1}{K\beta^2} \left\{ \log(1 + N\beta) - \frac{N\beta}{1 + N\beta} \right\}$$

と書くことができる。右辺の積の前半は明らかに正である。積の後半を $f_3(\beta)$ とおき β で微分すると

$$f_3'(\beta) = \frac{N^2\beta}{(1 + N\beta)^2} > 0$$

となることより、 $f_3(\beta)$ は $\beta > 0$ で単調増加となる。また $f_3(0)=0$ より $f_3(\beta)$ は同じ範囲で正となり、 R_2 が $\beta > 0$ で単調増加となることが証明できる。

分類の併合前に、併合後の標本でも母集団でも一意である個体数の予測を行うような場合には、一般にパラメータ β を推定する必要があるが、独立型のモデルを用いる場合には R_2 が β に関して単調増加なので、 β を大きめに推定すれば減少率をやや控えめに見積もりすることができる。完全連動型のモデルでは、 K の値が十分小さければ β を大きめに、 K の値が十分大きければ小さめに推定すれば控えめな値が得られる。

また 2 つの期待値の比の大小関係について次のような性質が求められる。

定理 3

$$K_0 = \frac{L \log(1 + N\beta) - \log(1 + LN\beta)}{\beta \{\log(1 + LN\beta) - \log(1 + N\beta)\}}$$

とおくとき, R_1 と R_2 の大小関係は K_0 と K の大小関係に等しい.

証明 $R_1 > R_2$ のとき

$$\left(\frac{1 + LN\beta}{1 + N\beta} \right)^{-1 - \frac{1}{K\beta}} > (1 + N\beta)^{-\frac{L-1}{K\beta}}$$

の両辺の対数をとると

$$(K\beta + 1) \log(1 + LN\beta) < (K\beta + L) \log(1 + N\beta)$$

と変形することができ,

$$K < \frac{L \log(1 + N\beta) - \log(1 + LN\beta)}{\beta \{\log(1 + LN\beta) - \log(1 + N\beta)\}}$$

となることが容易に証明できる. $R_1 = R_2$, $R_1 < R_2$ の場合についても同様である.

定理 1, 2 より, β が小さい場合には完全連動型モデルと独立型モデルの結果はほぼ等しくなる. 逆に β が大きいときにはその差は大きくなり, 独立型モデルでは併合によって一意な個体数はほとんど減少しないことが分かる. また分類の併合前に併合後の一意な個体数の予測を行う場合には, K の値が十分小さければ完全連動型モデルを, K の値が十分大きければ独立型モデルを用いれば減少率の控えめな予測ができることが定理 3 から分かる.

4. 相関型モデルと部分連動型モデル

4.1. 相関型モデル

現実のデータでは, 異なる分類で同じ順番のセルのパラメータが等しくないものの, ある程度の関係を持っている場合が多く, 第 3 節で提案した 2 つの極端なモデルのみでは実データへのあてはめは不十分である. そこで完全連動型モデルと独立型モデルとの間を埋めるモデルとして, この節では相関型モデルを, 次節では部分連動型モデルを導入する.

前者は異なる分類で対応するセルのパラメータ $\pi_{l,i}$ ($l=1, 2, \dots, L$) 同士が相関係数 ρ ($0 \leq \rho \leq 1$) を持つモデルであるが, 次のようにパラメータに相関を入れることにする.

$$\pi_{l,i} = \pi_{\cdot,i} + \pi'_{l,i}$$

ただし $\pi_{\cdot,i}$ ($i=1, 2, \dots, K$) と $\pi'_{l,i}$ ($l=1, 2, \dots, L; i=1, 2, \dots, K$) はすべて互いに独立に次のようにガンマ分布に従う.

$$\begin{aligned} \pi_{\cdot,i} &\sim \text{gamma}(\rho\alpha, \beta), \\ \pi'_{l,i} &\sim \text{gamma}((1-\rho)\alpha, \beta). \end{aligned}$$

このときガンマ分布の再生性より $\pi_{l,i} \sim \text{gamma}(\alpha, \beta)$ となり, $\pi_{l,i}$ と $\pi_{l',i}$ ($l \neq l'$) は相関係数 ρ の関係があり, $\pi_{l,i}$ と $\pi_{l',i'}$ ($l \neq l', i \neq i'$) は独立となる. この種の相関の入れ方については Johnson and Kotz (1972) が詳しい. この相関型モデルは $\rho=1$ のときに完全連動型モデル, $\rho=0$ のときに独立型モデルに一致する.

このモデルの下で、分類の併合後に標本でも母集団でも一意である個体数の期待値 $E(\underline{S}_{b3})$ は (2), (3) 式と同様に、次のように求められる。

$$\begin{aligned}
 E(\underline{S}_{b3}) &= \sum_{i=1}^K \sum_{l=1}^L E(N_i \pi_{l,i} e^{-N_i \pi_{l,i}} \cdot \lambda \cdot \prod_{l' \neq l} e^{-N_{l'} \pi_{l',i}}) \\
 &= \sum_{i=1}^K \sum_{l=1}^L E\{N_i (\pi_{\cdot,i} + \pi'_{l,i}) e^{-N_i (\pi_{\cdot,i} + \pi'_{l,i})} \cdot \lambda \cdot \prod_{l' \neq l} e^{-N_{l'} (\pi_{\cdot,i} + \pi'_{l',i})}\} \\
 &= \sum_{i=1}^K \sum_{l=1}^L \{E\{N_i \pi_{\cdot,i} e^{-(\sum_{l=1}^L N_{l'} \pi_{l',i})}\} \cdot \prod_{l'=1}^L E(e^{-N_{l'} \pi_{l',i}}) \\
 &\quad + E(N_i \pi'_{l,i} e^{-N_i \pi'_{l,i}}) E\{e^{-(\sum_{l=1}^L N_{l'} \pi_{l',i})}\} \cdot \prod_{l' \neq l} E(e^{-N_{l'} \pi_{l',i}})\} \cdot \lambda \\
 &= \sum_{i=1}^K \sum_{l=1}^L \left\{ \frac{\frac{\rho N_i}{K}}{\{1 + (\sum_{l'=1}^L N_{l'} \beta)\}^{1+\frac{\rho}{K\beta}}} \cdot \prod_{l'=1}^L \frac{1}{(1 + N_{l'} \beta)^{\frac{1-\rho}{K\beta}}} \right. \\
 &\quad \left. + \frac{(1-\rho) N_i}{K} \cdot \frac{1}{\{1 + (\sum_{l'=1}^L N_{l'} \beta)\}^{\frac{\rho}{K\beta}}} \cdot \prod_{l' \neq l} \frac{1}{(1 + N_{l'} \beta)^{\frac{1-\rho}{K\beta}}} \right\} \cdot \lambda \\
 &= \sum_{i=1}^L \frac{\rho N_i (1 + N_i \beta) + (1-\rho) N_i \{1 + (\sum_{l'=1}^L N_{l'} \beta)\}}{\{1 + (\sum_{l'=1}^L N_{l'} \beta)\}^{1+\frac{\rho}{K\beta}} (1 + N_i \beta)} \cdot \prod_{l=1}^L \frac{1}{(1 + N_l \beta)^{\frac{1-\rho}{K\beta}}} \cdot \lambda. \quad (6)
 \end{aligned}$$

$N_i \equiv N$ のとき上式は次のように書き直される。

$$E(\underline{S}_{b3}) = \frac{LN \{\rho(1 + N\beta) + (1-\rho)(1 + LN\beta)\}}{(1 + N\beta)^{1+\frac{(1-\rho)L}{K\beta}} (1 + LN\beta)^{1+\frac{\rho}{K\beta}}} \cdot \lambda.$$

ここで前節と同様に、上の期待値と併合前の期待値 (1) 式) との比

$$R_3 = \frac{E(\underline{S}_{b3})}{E(\underline{S}_a)} = \frac{\rho(1 + N\beta) + (1-\rho)(1 + LN\beta)}{(1 + N\beta)^{\frac{(1-\rho)L-1}{K\beta}} (1 + LN\beta)^{1+\frac{\rho}{K\beta}}}$$

を考えると、次のような性質を得ることができる。

定理 4 $\beta \rightarrow 0$ のとき ρ の値に関わらず $R_3 \rightarrow e^{-(L-1)N/K}$, $\beta \rightarrow \infty$ のとき K, N の値に関わらず $R_3 \rightarrow (\rho + (1-\rho)L)/L$ である。

証明

$$R_3 = \frac{\rho(1 + N\beta) + (1-\rho)(1 + LN\beta)}{1 + LN\beta} \cdot \frac{1}{(1 + N\beta)^{\frac{1}{N\beta} \cdot \frac{N \cdot (1-\rho)L - 1}{K}} (1 + LN\beta)^{\frac{1}{LN\beta} \cdot \frac{\rho LN}{K}}}$$

より

$$\begin{aligned}
 \lim_{\beta \rightarrow 0} R_3 &= e^{-\frac{(L-1)N}{K}}, \\
 \lim_{\beta \rightarrow \infty} R_3 &= \frac{\rho + (1-\rho)L}{L}
 \end{aligned}$$

となることが示される。

分類の併合前に併合後の標本でも母集団でも一意である個体数を推定する際には相関係数 ρ を推定する必要があるが、 β が十分小さい場合にはその推定については高い精度が要求されないことがわかる。

次に相関係数 ρ に関する R_3 の増減について次のような関係が導かれる。

定理 5

- (i) $K \leq K_1$ のとき R_3 は $0 < \rho < 1$ において単調増加である。
 - (ii) $K_1 < K < K_2$ のとき, R_3 は $0 < \rho < \rho_1$ において単調増加で, $\rho_1 < \rho < 1$ において単調減少である。
 - (iii) $K_2 \leq K$ のとき R_3 は $0 < \rho < 1$ において単調減少である。
- ただし

$$K_1 = \frac{1+N\beta}{(L-1)N\beta^2} \log \frac{(1+N\beta)^L}{1+LN\beta},$$

$$K_2 = \frac{1+LN\beta}{(L-1)N\beta^2} \log \frac{(1+N\beta)^L}{1+LN\beta},$$

$$\rho_1 = \frac{1+LN\beta}{(L-1)N\beta} - \frac{K\beta}{\log \frac{(1+N\beta)^L}{1+LN\beta}}.$$

証明 $\log R_3$ を ρ に関して偏微分した式

$$\frac{\partial \log R_3}{\partial \rho} = \frac{(L-1)N\beta}{\rho(L-1)N\beta - (1+LN\beta)} + \frac{1}{K\beta} \log \frac{(1+N\beta)^L}{1+LN\beta}$$

を新たに $f_4(\rho)$ とおくと, その第1項は常に負でしかも ρ に関する単調減少関数であることを利用して, 次のように場合分けして考える。

- (i) $f_4(1) = (L-1)N\beta / ((L-1)N\beta - (1+LN\beta))$

$+ (1/(K\beta)) \log((1+N\beta)^L / (1+LN\beta)) \geq 0$ のとき

$K_1 = (1+N\beta) / ((L-1)N\beta^2 \log((1+N\beta)^L / (1+LN\beta)))$ とおいたとき $K \leq K_1$ となり, このとき $f_4(\rho) > 0$ ($0 < \rho < 1$) であるから R_3 は $0 < \rho < 1$ において単調増加である。

- (ii) $f_4(0) > 0$ かつ $f_4(1) < 0$ のとき

$K_2 = (1+LN\beta) / ((L-1)N\beta^2 \log((1+N\beta)^L / (1+LN\beta)))$ とおいたとき $K_1 < K < K_2$ となる。このとき $f_4(\rho_1) = 0$ となる ρ_1 ($0 < \rho_1 < 1$) が存在し, $\rho_1 = (1+LN\beta) / ((L-1)N\beta) - K\beta / \log((1+N\beta)^L / (1+LN\beta))$ となる。 $0 < \rho < \rho_1$ において $f_4(\rho) > 0$ より R_3 は単調増加で, $\rho_1 < \rho < 1$ において $f_4(\rho) < 0$ より R_3 は単調減少である。

- (iii) $f_4(0) = -(L-1)N\beta / (1+LN\beta) + (1/(K\beta)) \log((1+N\beta)^L / (1+LN\beta)) \leq 0$ のとき

K の範囲は $K \geq K_2$ となり, このとき $f_4(\rho) < 0$ ($0 < \rho < 1$) であるから R_3 は $0 < \rho < 1$ において単調減少である。

したがって分類の併合前に予測をする場合には, 例えば K の値が十分小さければ, 併合後の標本でも母集団でも一意である個体数の期待値の下限と上限を独立型モデルと完全連動型モデルによって得ることができる。また ρ の値をやや大きめに見積もることによって減少率のやや控えめな予測も可能となる。逆に K の値が十分大きければ下限が完全連動型モデルで上限が独立型モデルとなる。実際のデータでは十分な秘匿措置を行ってもまだ K の値は一般に極めて大きいため, 多くがこのケースに含まれるものと考えられる。したがって一般的には, 分類の併合の候補が複数ある場合, 相関係数が大きいと思われる分類同士を併合すれば一意な個体数を効果的に減少させることができる。なお (ii) のケースでは完全連動モデルも独立型モデルも楽観的すぎる見通しを導くことに注意しなければならない。

図 1. 相関型モデルでの一意な個体数の期待値の併合前後の比

図 1 に $L=2$, $N=10^6$, $\beta=10^{-4}$ の場合の相関係数 ρ と R_3 との関係を示す. なお K の値は $3 \cdot 10^4$, $6 \cdot 10^4$, $10 \cdot 10^4$ の 3 通りである. この場合 $K_1 \doteq 39662$, $K_2 \doteq 78931$ となるため, 上の 3 つの K の値はそれぞれ定理 5 の (i), (ii), (iii) に含まれる.

4.2. 部分連動型モデル

この節では L 個の分類が m 個ずつ L/m 個のグループに分かれており, グループ内の対応するパラメータは完全連動しており, 異なるグループ間のパラメータ同士は独立である部分連動型のモデルを考える. すなわち第 g 番目のグループの第 l 番目の分類における第 i 番目のセルのパラメータを新たに $\pi_{g,l,i}$ とすると, $\pi_{g,1,i} = \pi_{g,2,i} = \dots = \pi_{g,m,i}$ ($i=1, 2, \dots, K$) であり, $\pi_{g,l,i}$ と $\pi_{g',l',i}$ ($g \neq g'$; $l=1, 2, \dots, m$; $l'=1, 2, \dots, m$) は互いに独立である. ただし異なる i については, パラメータはすべて互いに独立である. このモデルは $m=1$, L のときにそれぞれ独立型モデルと完全連動型モデルに一致する.

第 g 番目のグループの第 l 番目の分類における母集団の大きさを $N_{g,l}$ とするとき, 分類の併合後に標本でも母集団でも一意である個体数の期待値 $E(\underline{S}_{b4})$ は次のように求められる.

$$\begin{aligned} E(\underline{S}_{b4}) &= \sum_{i=1}^K \sum_{g=1}^{L/m} \sum_{l=1}^m E(N_{g,l} \pi_{g,l,i} e^{-N_{g,l} \pi_{g,l,i}} \cdot \lambda \cdot \prod_{l' \neq l} e^{-N_{g,l'} \pi_{g,l',i}} \cdot \prod_{g' \neq g} \prod_{l'=1}^m e^{-N_{g',l'} \pi_{g',l',i}}) \\ &= \sum_{i=1}^K \sum_{g=1}^{L/m} \sum_{l=1}^m E(N_{g,l} \pi_{g,l,i} e^{-(\sum_{l'=1}^m N_{g,l'} \pi_{g,l',i})} \cdot \lambda \cdot \prod_{g' \neq g} E\{e^{-(\sum_{l'=1}^m N_{g',l'} \pi_{g',l',i})}\}) \\ &= \sum_{i=1}^K \sum_{g=1}^{L/m} \sum_{l=1}^m \frac{\frac{N_{g,l}}{K}}{\{1 + (\sum_{l'=1}^m N_{g,l'} \beta)\}^{1 + \frac{1}{K\beta}}} \cdot \prod_{g' \neq g} \frac{1}{\{1 + (\sum_{l'=1}^m N_{g',l'} \beta)\}^{\frac{1}{K\beta}}} \cdot \lambda \\ &= \sum_{g=1}^{L/m} \frac{\sum_{l=1}^m N_{g,l}}{1 + (\sum_{l=1}^m N_{g,l} \beta)} \cdot \prod_{g=1}^{L/m} \frac{1}{\{1 + (\sum_{l=1}^m N_{g,l} \beta)\}^{\frac{1}{K\beta}}} \cdot \lambda. \end{aligned}$$

$N_{g,i} \equiv N$ のとき上式は次のように書き直される。

$$E(\underline{S}_{b_4}) = \frac{LN}{(1+mN\beta)^{1+\frac{L}{mK\beta}}} \cdot \lambda.$$

また上の期待値と併合前の期待値 ((1) 式) との比は

$$R_4 = \frac{E(\underline{S}_{b_4})}{E(\underline{S}_a)} = \frac{(1+N\beta)^{1+\frac{1}{K\beta}}}{(1+mN\beta)^{1+\frac{L}{mK\beta}}} \quad (7)$$

と求められる。

前節と同様に (7) 式について次のような性質が得られる。

定理 6 $\beta \rightarrow 0$ のとき m の値に関わらず $R_4 \rightarrow e^{-(L-1)N/K}$ である。また $\beta \rightarrow \infty$ のとき $R_4 \rightarrow m^{-1}$ である。

証明

$$R_4 = \frac{(1+N\beta)(1+N\beta)^{\frac{1}{N\beta} \cdot \frac{N}{K}}}{(1+mN\beta)(1+mN\beta)^{\frac{1}{mN\beta} \cdot \frac{LN}{K}}}$$

より

$$\lim_{\beta \rightarrow 0} R_4 = e^{-\frac{(L-1)N}{K}},$$

$$\lim_{\beta \rightarrow \infty} R_4 = m^{-1}$$

となることが示される。

定理 7

(i) $K \leq K_3$ のとき R_4 は $1 < m < L$ において単調増加である。

(ii) $K_3 < K < K_4$ のとき, R_4 は $1 < m < m_1$ において単調増加で, $m_1 < m < L$ において単調減少である。

(iii) $K_4 \leq K$ のとき R_4 は $1 < m < L$ において単調減少である。

ただし

$$K_3 = \frac{1}{LN\beta^2} \{(1+LN\beta)\log(1+LN\beta) - LN\beta\},$$

$$K_4 = \frac{L}{N\beta^2} \{(1+N\beta)\log(1+N\beta) - N\beta\}$$

であり, m_1 は次の方程式を満たす正の解 m である。

$$L(1+mN\beta)\log(1+mN\beta) - mN\beta(mK\beta + L) = 0.$$

証明 $\log R_4$ を m に関して偏微分すると

$$\frac{\partial \log R_4}{\partial m} = \frac{1}{m^2 K \beta (1+mN\beta)} \{L(1+mN\beta)\log(1+mN\beta) - mN\beta(mK\beta + L)\}$$

となるが, 右辺の積の前半は明らかに正である。後半の括弧内を $f_5(m)$ とおき, 再び m で微分すると

$$f'_5(m) = N\beta\{L\log(1+mN\beta) - 2K\beta m\}$$

となる. ここで $f_5(m) = L\log(1+mN\beta)$ とおき, $y = f_5(m)$ と直線 $y = 2K\beta m$ の原点における傾きを比較する. $f'_5(m) = LN\beta/(1+mN\beta)$ より $f'_5(0) = LN\beta$ となるので, 次のように場合分けする.

(I) $LN\beta \leq 2K\beta$ すなわち $K \geq LN/2$ のとき

このとき $m > 0$ において $y = f_5(m)$ が直線 $y = 2K\beta m$ より下にあることより $f'_5(m) < 0$ ($m > 0$) である. よって $f_5(m)$ は同じ範囲で単調減少で $f_5(0) = 0$ より $f_5(m)$ も同じ範囲で負となる. よって R_4 は $1 < m < L$ で単調減少である.

(II) $LN\beta > 2K\beta$ すなわち $K < LN/2$ のとき

このとき $f'_5(m_0) = 0$ を満たす正の数 m_0 が存在し $0 < m < m_0$ において $f'_5(m) > 0$ より $f_5(m)$ は単調増加, $m_0 < m$ において $f'_5(m) < 0$ より $f_5(m)$ は単調減少である. ここで

$$\begin{aligned} f_5(1) &= L(1+N\beta)\log(1+N\beta) - (K\beta+L)N\beta, \\ f_5(L) &= L(1+LN\beta)\log(1+LN\beta) - (LK\beta+L)LN\beta \end{aligned}$$

の符号についてさらに次のように場合分けする.

(i) $f_5(L) \geq 0$ のとき

$K_3 = (1/(LN\beta^2))\{(1+LN\beta)\log(1+LN\beta) - LN\beta\}$ とおくと $K \leq K_3$ である. このとき $f_5(m) > 0$ ($1 < m < L$) であるから R_4 は同じ範囲で単調増加である.

(ii) $f_5(1) > 0, f_5(L) < 0$ のとき

$K_4 = (L/(N\beta^2))\{(1+N\beta)\log(1+N\beta) - N\beta\}$ とおくと $K_3 < K < K_4$ である. このとき $f'_5(m) = 0$ を満たす正の数 m を m_1 とすると $1 < m_1 < L$ である. $1 < m < m_1$ において $f'_5(m) > 0$ より R_4 は同じ範囲で単調増加である. また $m_1 < m < L$ において $f'_5(m) < 0$ より R_4 は同じ範囲で単調減少である.

(iii) $f_5(1) \leq 0$ のとき

$K \geq K_4$ である. このとき $f_5(m) < 0$ ($1 < m < L$) であるから R_4 は同じ範囲で単調減少である. なお $K_4 < LN/2$ となることが両者の比を微分することによって容易に求めることができるので, この場合分けを証明の最初の (I) と合わせて結局 (iii) の K の範囲は $K_4 \leq K$ となる.

定理 6, 7 から, 部分運動型モデルについても相関型モデルと同様の性質が得られるが, 定理 5 の境界値 K_1, K_2 と定理 7 の K_3, K_4 の間には次のような関係がある.

定理 8 $\beta > 0$ のとき $K_1 < K_3 < K_4 < K_2$ である.

証明 まず $K_1 < K_3$ について証明する.

$$\begin{aligned} & L(L-1)N\beta^2(K_3 - K_1) \\ &= (L-1)\{(1+LN\beta)\log(1+LN\beta) - LN\beta\} - L(1+N\beta)\log\frac{(1+N\beta)^L}{1+LN\beta} \end{aligned}$$

は $N\beta$ を x と置き直すことによって

$$f_7(x) = (L-1)\{(1+Lx)\log(1+Lx) - Lx\} - L(1+x)\log\frac{(1+x)^L}{1+Lx}$$

と書ける. ここで

表 1. 相関型モデルと部分連動型モデルにおける期待値の比の性質

K	R_3 (相関型モデル)	R_4 (部分連動型モデル)
0	$0 < \rho < 1$ において単調増加	$1 < m < L$ において単調増加
K_1		
K_3	$0 < \rho < \rho_1$ において単調増加 $\rho_1 < \rho < 1$ において単調減少	$1 < m < m_1$ において単調増加 $m_1 < m < L$ において単調減少
K_4		
K_2	$0 < \rho < 1$ において単調減少	$1 < m < L$ において単調減少
∞		

$$f'_7(x) = L^2 \left\{ \log \frac{1+Lx}{1+x} - \frac{(L-1)x}{1+Lx} \right\},$$

$$f''_7(x) = \frac{L^2 x (L-1)^2}{(1+Lx)^2 (1+x)}$$

と微分すると、 $x > 0$ において $f''_7(x) > 0$ で $f'_7(0) = 0$ より $f'_7(x)$ も同じ範囲で正である。したがって $x > 0$ において $f_7(x)$ は単調増加で $f_7(0) = 0$ より $f_7(x) > 0$ ($x > 0$) である。よって $K_1 < K_3$ ($\beta > 0$) であることが証明された。

$K_4 < K_2$ であることも同様に証明される。

したがって (ii) の期待値の比が極大値を持つ K の範囲は部分連動型モデルの $K_3 < K < K_4$ の方が相関型モデルの $K_1 < K < K_2$ よりも狭く、逆に (i) の単調増加の領域と、(iii) の単調減少の領域は部分連動型モデルの方が広い。 K の各範囲の相関型モデルと部分連動型モデルにおける期待値の比の性質を表 1 にまとめる。

また定理 7 の証明中にも述べたように $K_4 < LN/2$ という関係が常に成り立つため、 $K_4 < N$ となる。また K_2 と $LN/2$ とを比較すると、 $L \geq 3$ で $N\beta$ が十分小さければ $LN/2 < K_2$ となることはあるものの、実データから通常得られる値では $K_2 < N$ となることが多い。一般にセル数 K は各分類における母集団の個体数 N と比較して十分大きいと、実際のデータでは相関型モデルでも部分連動型モデルでもほとんどの場合が (iii) に含まれるものと考えられる。

なお相関型モデルや部分連動型モデルで、さらに細かい設定をしたり両者を併用することも可能である。しかし次節の応用例を見る限り、分類の併合の良し悪しを判断する目安として用いるのが目的であるならば、これまで述べた程度の結果を考え合わせることで十分と思われる。

5. 労働力調査の個票データに対する検討

この節では、総務庁統計局統計基準部が中心となって開催された「標本データの秘密保護に関する研究会」において、佐井が竹村の研究協力者として目的外使用した労働力調査の個票データに対して試算を行い、モデルの当てはまりとその実用性を見ることにする。なおここでは 2 つの

分類の併合のみを考えるため、部分連動型モデルは独立型モデルまたは完全連動型モデルに一致してしまう。したがって相関型モデルのみを用いることにするが、この節のような例では、仮に3つ以上の分類の併合でも相関型モデルの方が適していると思われる。

労働力調査は日本の就業及び不就業の状態を明らかにすることを目的とするもので、15歳以上の人を調査対象として、毎月約4万世帯の約10万人が調査されている。今回検討の対象とするのは1997年12月分の個票データで、その中には97,950人分の個人単位のレコードが含まれている。労働力調査の詳細については、総務庁統計局 統計調査部 労働力統計課 (1996)、総務庁統計局 (1998) を参照していただきたい。ただしキー変数と考えられる項目が20程度あり、元の個票ではすべての個体が標本で一意である。そこで分類の併合に関する検討を行う前に、例えば都道府県内の小地域に関する項目を削除して地域変数としては都道府県が最も細かな項目となるようにしたり、年齢を10歳刻みにするなどの秘匿措置を行って、標本でも母集団でも一意である個体数の期待値を100程度まで小さくしておいた。それでもセル数は $2.652 \cdot 10^{14}$ とかなり多い。その個票データに対して、都道府県、年齢階級に関してそれぞれ併合を行うことにする。なお結果を明確にするために、併合はすべて2つの分類に対して行う。

まず初めに表2のように主に隣接している2つの都道府県の併合を5通りについて行う。(山形県と石川県については母数の値がほぼ等しいため特に対象とした。) サンプルング方法などが原因で各都道府県における標本の抽出率が異なっているため、高い抽出率の都道府県の個体を無作為に削除して低い抽出率に揃えた。この操作の前後の個体数を、レコード数(元)とレコード数(修正後)の列に示している。また $\hat{\beta}$ の列はガンマ関数のパラメータ β のモーメント法による推定値を示しているが、その計算法については佐井(1998b)を参照されたい。また標本・母集団の一意数の列は標本でも母集団でも一意である個体数の期待値を表し、併合後の値については実際に分類を併合した個票データからポアソンガンマモデルを用いて求めたものである。なお各都道府

表2. 都道府県の併合の試算-1

都道府県	母集団の 大きさ	レコード数		$\hat{\beta}$	標本の一意数	標本・母集団の 一意数
		元	修正後			
秋田	1028000	1063	908	$2.9595 \cdot 10^{-4}$	771	2.975
山形	1056000	932	932	$2.3485 \cdot 10^{-4}$	798	3.746
秋田+山形	2084000	1995	1840	$2.2684 \cdot 10^{-4}$	1492	3.886
栃木	1674000	1504	1504	$2.3607 \cdot 10^{-4}$	1215	3.796
群馬	1701000	1707	1528	$2.9981 \cdot 10^{-4}$	1252	2.991
栃木+群馬	3375000	3211	3032	$2.3039 \cdot 10^{-4}$	2335	3.895
埼玉	5786000	3904	3904	$5.0600 \cdot 10^{-4}$	2778	1.333
千葉	4970000	3433	3353	$3.8265 \cdot 10^{-4}$	2446	1.762
埼玉+千葉	10756000	7337	7257	$4.3100 \cdot 10^{-4}$	4970	1.565
富山	962000	1237	1237	$3.0846 \cdot 10^{-4}$	1013	4.155
石川	1001000	1308	1287	$2.2338 \cdot 10^{-4}$	1077	5.731
富山+石川	1963000	2545	2524	$2.3232 \cdot 10^{-4}$	1988	5.523
山形	1056000	932	932	$2.3485 \cdot 10^{-4}$	798	3.743
石川	1001000	1308	883	$2.3856 \cdot 10^{-4}$	774	3.684
山形+石川	2057000	2240	1815	$2.0278 \cdot 10^{-4}$	1497	4.342

表 3. 2 都道府県のセルの同時分布 (縦: 秋田, 横: 山形)

	0	1	2	3	4	5	6	7	8	9	10
0	*	758	30	4	2						
1	734	28	5	4							
2	28	10	5	2	1						
3	1		1		1						
4	2	2	1		1						
5	1										
6											
7	1										
8											
9											
10											

県のセル数は $5.644 \cdot 10^{12}$ である。

次に個票データから相関係数 ρ の推定を行う。秋田県と山形県を例にとるとセルの同時分布は表 3 のようになっている。例えば両県の対応するセルのうちで、含まれる個体数が共に 1 のセルが 28 個あることになる。なお両県とも個体数が 0 のセルは総セル数 $5.644 \cdot 10^{12}$ から同時分布の他の度数の和 1622 を引いたものになる。秋田県を第 1 番目、山形県を第 2 番目の分類として、標本のセルに含まれる個体数 $f_{1,i}$ と $f_{2,i}$ の共分散の実現値を求めると $Cov(f_{1,i}, f_{2,i}) = (1/K) \sum_{i=1}^K (f_{1,i} - n_1/K)(f_{2,i} - n_2/K) = 2.8351 \cdot 10^{-11}$ となる。ここで標本共分散 $Cov(f_{1,i}, f_{2,i})$ と相関係数 ρ との間には次のような関係がある。ただし添え字の 1 はガンマ分布に関して、添え字の 2 は $\pi_{i,i}$ の条件付きのポアソン分布に関しての期待値, 分散, 共分散を表す。

$$\begin{aligned}
 Cov(f_{1,i}, f_{2,i}) &= E_1\{Cov_2(f_{1,i}, f_{2,i})\} + Cov_1\{E_2(f_{1,i}), E_2(f_{2,i})\} \\
 &= E_1(0) + Cov_1(n_1\pi_{1,i}, n_2\pi_{2,i}) \\
 &= n_1n_2\sqrt{V_1(\pi_{1,i})V_1(\pi_{2,i})}\rho \\
 &= \frac{n_1n_2\sqrt{\beta_1\beta_2}}{K} \cdot \rho
 \end{aligned}$$

この関係を用いて相関係数を推定するが、この計算においては β_1 と β_2 を異なる値として県別に用いることにする。相関係数は標本共分散の実現値と表 2 の値を用いて次のように推定される。

$$\begin{aligned}
 \hat{\rho} &= Cov(f_{1,i}, f_{2,i}) \cdot \frac{K}{n_1n_2\sqrt{\beta_1\beta_2}} \\
 &= 0.7171
 \end{aligned}$$

この値を表 4 の 相関係数 の列に示す。また $\hat{\rho}$ の列には、相関型モデルを用いた併合後の標本でも母集団でも一意である個体数の期待値((6) 式)にこの値を代入した結果を示す。なお β には 2 県の値の平均を用いた。どの併合についても、実際に都道府県を併合した個票データから求められる値に比較的近く、相関型モデルを用いることで併合前の予測が可能であると考えられる。

ところで相関係数の推定値 $\hat{\rho}$ を求めるためには表 3 のような分布を求める必要があり、分類の併合と比較してもそれほど作業が軽減されるとは言えない。しかし、例えばいくつかの典型的な都道府県の組に対して相関係数の推定値を求めておいて他の都道府県の併合の際には最も性質が

表 4. 都道府県の併合の試算-2

都道府県	相関係数	相関係数より併合前推定					併合後推定
		(独立) 0	$\hat{\rho} - \widehat{SD}$	$\hat{\rho}$	$\hat{\rho} + \widehat{SD}$	(完全連動) 1	
秋田+山形	0.7171	6.629	4.695	4.257	3.819	3.321	3.886
栃木+群馬	0.7229	6.691	4.626	4.275	3.924	3.349	3.895
埼玉+千葉	0.9274	3.036	1.861	1.628	1.518	1.518	1.565
富山+石川	0.7633	9.634	6.728	5.964	5.199	4.856	5.523
山形+石川	0.7136	7.425	5.250	4.781	4.311	3.720	4.342

似ていると思われる組の推定値で代用したり、同じ都道府県の組で多少異なるキー変数の組み合わせについて推定した相関係数をそのまま用いたり、または多少修正するなど、やや曖昧な推定を行うことは容易と思われる。そこでそのような状況を想定して、標本共分散 $Cov(f_{1,i}, f_{2,i})$ の実現値からその標準誤差の推定値

$$\frac{1}{K} \left\{ \frac{1}{K} \sum_{i=1}^K \left(f_{1,i} - \frac{n_1}{K} \right)^2 \left(f_{2,i} - \frac{n_2}{K} \right)^2 - \left\{ \sum_{i=1}^K \left(f_{1,i} - \frac{n_1}{K} \right) \left(f_{2,i} - \frac{n_2}{K} \right) \right\}^2 \right\}$$

を加減した値を共分散の代わりに用いた場合の一意的な個体数の期待値を表4の $\hat{\rho} + \widehat{SD}$ と $\hat{\rho} - \widehat{SD}$ の列にそれぞれ示す。埼玉県と千葉県については $\hat{\rho} + \widehat{SD}$ が1を超えるため相関係数が1の場合の結果を示す。相関係数の推定が曖昧になるに従って $\widehat{SD}$ の2倍や3倍を加減しなければならないだろう。そしてもはや相関係数に関する情報が全く得られない場合の対応策として、相関係数が0と1、すなわち独立型モデルと完全連動型モデルを用いた場合の値も同じ表に示す。セル数 K が十分大きく定理5の(iii)の場合分けに属するため、相関係数が大きいほど一意的な個体数の期待値は小さくなる。この結果から予測の目的では、低い精度で推定された相関係数を用いても十分実用に耐えうると考えることができる。

次に年齢階級の併合について同様の検討を行う。都道府県の併合では年齢を10歳刻みにするなどの秘匿措置を行った個票データを用いたが、ここでは年齢を5歳階級に分けたものを1つの分類として、2分類ずつ併合して10歳階級に戻した場合について考える。表5にパラメータ β の推定値や一意的な個体数などを、表6に相関型モデルを用いた推定結果をそれぞれ示す。なお各年齢階級のセル数は $2.947 \cdot 10^{13}$ である

20歳代と40歳代の相関係数が他の年齢階級と比較して低いのは、個票データに就業状態と15歳未満の子供の数に関する項目があることによっている。ただし20歳代については前半と後半のパラメータの推定値 $\hat{\beta}$ の値がかなり異なっているため、 β が等しいと仮定する相関型モデルの当てはめはやや問題があると思われる。

この結果からも都道府県の併合同様の結論が得られる。この場合も K の値が十分大きく定理5の(iii)に含まれ、標本でも母集団でも一意である個体数の期待値が相関係数に関して単調減少であるので、相関係数の小さい40歳代は他の年齢階級と比較して一意的な個体数の減少率の低いことを、事前に予測することも可能であろう。

なおどちらの例でも、相関型モデルから予測される一意的な個体数の期待値が、実際に併合された個票データから推定された値より全体的にやや大きいのは、ポアソンガンマモデルの当てはまりの良し悪しが原因かもしれないし、パラメータ β が各分類で共通という仮定の下で計算を行ったことによるものかもしれない。特にパラメータの値が大きく異なる場合には他のモデルを考え

表 5. 年齢階級の併合の試算-1

年齢階級	母集団の 大きさ	レコード数		$\hat{\beta}$	標本の一意数	標本・母集団の 一意数
		元	修正後			
20～24 歳	9620000	8298	7947	$1.3719 \cdot 10^{-4}$	6152	6.017
25～29 歳	9460000	7815	7815	$4.5125 \cdot 10^{-5}$	6796	18.264
20～29 歳	19080000	16113	15762	$6.5191 \cdot 10^{-5}$	12401	12.662
30～34 歳	8180000	6983	6983	$7.2761 \cdot 10^{-5}$	5995	11.713
35～39 歳	7790000	6833	6650	$5.4271 \cdot 10^{-5}$	5970	15.693
30～39 歳	15970000	13816	13633	$6.2069 \cdot 10^{-5}$	11544	13.740
40～44 歳	8290000	7507	7504	$1.7688 \cdot 10^{-5}$	7015	50.829
45～49 歳	10900000	9866	9866	$2.3485 \cdot 10^{-5}$	9075	38.393
40～49 歳	19190000	17373	17370	$1.7553 \cdot 10^{-5}$	15681	51.417
50～54 歳	8720000	8193	8193	$5.1516 \cdot 10^{-5}$	7295	18.198
55～59 歳	8290000	7894	7789	$8.2646 \cdot 10^{-5}$	6705	11.352
50～59 歳	17010000	16087	15982	$6.4262 \cdot 10^{-5}$	13601	14.608
60～64 歳	7690000	7427	7427	$1.5935 \cdot 10^{-4}$	5492	6.056
65～69 歳	6680000	6671	6452	$2.2288 \cdot 10^{-4}$	4352	4.330
60～69 歳	14370000	14098	13879	$1.8273 \cdot 10^{-4}$	9038	5.283

表 6. 年齢階級の併合の試算-2

年齢階級	相関係数	相関係数より併合前推定					併合後推定
		(独立) 0	$\hat{\rho} - \widehat{SD}$	$\hat{\rho}$	$\hat{\rho} + \widehat{SD}$	(完全連動) 1	
20～29 歳	0.4887	18.104	14.263	13.683	13.103	9.057	12.662
30～39 歳	0.9575	26.827	15.726	13.997	13.427	13.427	13.740
40～49 歳	0.6674	87.486	61.061	58.370	55.679	43.858	51.417
50～59 歳	0.9537	27.964	16.718	14.642	13.994	13.994	14.608
60～69 歳	0.9484	10.100	5.884	5.312	5.052	5.052	5.283

るか、または等しいとする仮定の結果に何らかの補正を行う必要があると思われ、今後の課題となろう。

6. おわりに

本論文では、個票データの個体が項目の値などによっていくつかの分類に分けられているとき、秘匿の目的で分類を併合した場合の標本でも母集団でも一意である個体数の変化の予測を目的として、いくつかの併合モデルを提案した。

その中心となるモデルは第4節で導入した相関型モデルであり、異なる分類で対応するセルのパラメータの相関係数を利用するものであり、相関係数が1と0の場合がそれぞれ第3節で述べた完全連動型モデルと独立型モデルである。この2つのモデルについては、パラメータ β に関する一意な個体数の期待値の増減と極限値を求めるとともに、その大小関係をセル数の値で場合分けして求めた。相関型モデルについては、 β に関する一意な個体数の期待値の極限と相関係数に関

する増減についての性質を考えた。例えばセル数が十分大きい場合には一意な個体数の期待値は相関係数に関して単調減少であり、独立型モデルが上限、完全連動型モデルが下限となる。労働力調査の個票データへの応用を見る限り、相関型モデルの実際のデータへの当てはまりは比較的良く、相関係数がやや曖昧な情報として得られた場合でも予測の目的では十分実用に耐えうると考えられる。特に、相関係数がほぼ等しく個体数が異なる多数の分類があり、その中でほぼ個体数の等しい分類の併合を多くの組で行おうとする場合に、予測は効率的に行える。さらに簡略化すれば、推定された相関係数をもとに独立型モデルと完全連動型モデルの値を補間することも考えられるが、セル数が定理5の(ii)に含まれる場合にはどちらのモデルの推定値も小さく、楽観的すぎる値が得られる点には注意しなければならない。

第4節では完全連動型モデルと独立型モデルが混在する部分連動型モデルについても考え、相関型モデルと同様の検討を行うとともに、両者の比較を行った。このモデルが適合するような実データの存在は考えにくい、相関型モデルを用いて多くの分類の併合を行う際には、このモデルの性質を考え合わせることも必要になると思われる。

この論文の結果から、いろいろな状況を想定した分類の併合後における一意な個体数の予測が、分類を実際に併合せずに、ある程度可能になったと思われる。

謝 辞 本論文の訂正にあたって査読者の方々からは適切なコメントを頂き、感謝致します。本論文は文部省科学研究費補助金、課題番号 11480055「調査データ公有化における理論的課題」、課題番号 11558026「統計データの個票開示における局所秘匿方法の確立」の研究成果に基づくものであり、統計数理研究所共同研究プログラム(12-共研-2022)の援助も受けている。

参 考 文 献

- Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990) : Disclosure Control of Microdata, *Journal of the American Statistical Association* 85, 38-45.
- Greenberg, B.V. and Zayatz, L.V. (1992) : Strategies for Measuring Risk in Public Use Microdata File, *Statistica Neerlandica* 46, 33-48.
- Johnson, N.L. and Kotz, S. (1972) : *Continuous Multivariate Distributions*, John Wiley.
- 加納 悟 (1997) : ミクロ個票データの公開におけるリスク評価の理論, 平成8年度科学研究費補助金報告書(加納班), 1-22.
- Keller, W.J. and Bethlehem, J.G. (1992) : Disclosure Protection of Microdata: Problems and Solutions, *Statistica Neerlandica* 46, 5-19.
- Hoshino, H. and Takemura, A. (1998) : Relationship between Logarithmic Series Model and Other Superpopulation Models Useful for Microdata Disclosure Risk Assessment, *Journal of the Japan Statistical Society* 28, 125-134.
- Omori, Y. (1999) : Measuring Identification Disclosure Risk for Categorical Microdata by Posterior Population Uniqueness, *Statistical Data Protection-Proceedings of the Conference, Lisbon, 25 to 27 March 1998-1999 edition*, Office for Official Publications of the European Communities, Luxembourg, 59-76.
- 佐井至道 (1998a) : 個票データにおける個体数とセル数との関係について, 平成9年度科学研究費補助金報告書(竹村班), 69-82.
- 佐井至道 (1998b) : 個票データにおける個体数とセル数との関係, 応用統計学 27, 127-145.
- 渋谷政昭 (1999) : size index の推測, 日本計量生物学会・応用統計学会合同年次大会予稿集, 11-14.
- Skinner, C.J. (1992) : On Identification Disclosure and Prediction Disclosure for Microdata, *Statistica Neerlandica* 46, 21-32.
- 総務庁統計局 統計調査部 労働力統計課 (1996) : 労働力調査・標本設計の解説.

総務庁統計局 (1998) : 平成 9 年労働力調査年報.

Takemura, A. (1999) : Some Superpopulation Models for Estimating the Number of Population Uniques, *Statistical Data Protection-Proceedings of the Conference, Lisbon, 25 to 27 March 1998-1999 edition*, Office for Official Publications of the European Communities, Luxembourg, 45-58.

Willenborg, L. and de Waal, T. (1996) : *Statistical Disclosure Control in Practice*, Springer.

(1999 年 7 月 21 日受付 2000 年 6 月 21 日最終修正)

著者連絡先 : 〒 700-8601 岡山市津島京町 2-10-1 岡山商科大学商学部 佐井至道