個票レベルのグループデータを用いた市区町村別所得分布の推定

川久保友超 [1]　　小林弦矢 [2]

[1] 千葉大学

[2] 明治大学

2024 年 11 月 18 日

公的統計ミクロデータ利活用に関する研究集会

グループデータ

- 家計調査
  - 月次
  - 五分位数と階級内平均値
- 住宅・土地統計調査
  - 5 年に 1 度実施（本研究の分析対象は 2018 年）
  - 市区町村ごとに世帯収入の推計度数分布が公表

住宅・土地統計調査におけるグループデータ

調査票情報

地域	世帯	0-300万円	300-500万円	500-700万円	700-1000万円	1000万円超
1	1		✓			
1	2	✓				
⋮	⋮					
1	n_1				✓	
2	1			✓		
2	2			✓		
⋮	⋮					
2	n_2					✓
⋮	⋮					

推計

公表データ（市区町村ごとの度数分布）

地域	0-300万円	300-500万円	500-700万円	700-1000万円	1000万円超
1	11250	33130	38660	24290	9540
2	22450	69570	80010	57340	23840
⋮					

住宅・土地統計調査におけるグループデータ

調査票情報

地域	世帯	0-300万円	300-500万円	500-700万円	700-1000万円	1000万円超
1	1		✓			
1	2	✓				
⋮	⋮					
1	n_1				✓	
2	1			✓		
2	2			✓		
⋮	⋮					
2	n_2					✓
⋮	⋮					

個票レベルグループデータ

推計

公表データ（市区町村ごとの度数分布）

地域	0-300万円	300-500万円	500-700万円	700-1000万円	1000万円超
1	11250	33130	38660	24290	9540
2	22450	69570	80010	57340	23840
⋮					

地域レベルグループデータ

地域レベルグループデータ

- Sugasawa, Kobayashi and Kawakubo (2020, CSDA)
- Kawakubo and Kobayashi (2023, CSDA)
- Kobayashi, Sugasawa and Kawakubo (2022, arXiv)

個票レベルグループデータ

- Walter, Groß, Schmid and Tzavidis (2021, JRSSA)

有限母集団パラメータの推定

- $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{i,N_i})^\top$: i 番目の地域の N_i 個の世帯の収入
- $h(\boldsymbol{z}_i)$: i 番目の地域の何らかの有限母集団パラメータ（平均収入，ジニ係数，貧困指標など）
- $\boldsymbol{z}_i$ を超母集団 f_i からの実現値とみる．すなわち $\boldsymbol{z}_i \sim f_i(\cdot)$
- $\boldsymbol{z}_i$ の長さ n_i（サンプルサイズ）の部分ベクトルが，標本調査によって観測される．
- 未観測の $(N_i - n_i)$ 個の z_{ij} たちを，超母集団分布 f_i にもとづいて予測する．

Nested error regression model (NERM)

$$\log(z_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad b_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2), \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

- 小地域推定における超母集団分布 $\boldsymbol{z}_i \sim f_i(\cdot)$ の最も基本的なモデル
- 補助変数 $\boldsymbol{x}_{ij}$ は，全世帯（$j = 1, \ldots, N_i$）において観測されていると仮定.
  - $\boldsymbol{x}_{ij}$ の確率分布は仮定しない（する必要がない）
  - $z_{ij} \mid \boldsymbol{x}_{ij}$ の条件付分布をモデリングする（比較的シンプルなモデルで）
  - z_{ij} の周辺分布（おそらく複雑）もモデリングする必要がない
- $\{z_{ij}\}_{j=1}^{n_i}$ が標本調査によって観測されていると仮定する．有限母集団パラメータ $h(\boldsymbol{z}_i)$ を推定するには，未観測の $\{z_{ij}\}_{j=n_i+1}^{N_i}$ を予測しなければならない.
- Molina and Rao (2010, CJS) は，NERM を仮定し，地域ごとの貧困指標を推定した.

提案モデル

$$\log z_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + b_i + \varepsilon_{ij} \quad (i = 1, \ldots, m; \quad j = 1, \ldots, N_i),$$
$$b_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2), \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

- z_{ij}: 地域 i の世帯 j の収入
- 補助変数ベクトル $\boldsymbol{x}_{ij}$ は，すべての世帯について観測されている.
- z_{ij} を直接観測できない代わりに，予め定められた境界値 $0 = c_0 < c_1 < \cdots < c_{G-1} < c_G = \infty$ たちによって区切られた G 個の区間のどこに，標本となった世帯の収入が属すかを観測する．つまり観測 y_{ij} は,

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} < c_1 \\ 2 & \text{if } c_1 \le z_{ij} < c_2 \\ \vdots & \\ G & \text{if } c_{G-1} \le z_{ij} \end{cases}$$

ベイズ推定

- 事前分布:

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, c_\beta \boldsymbol{I}_p), \quad \tau^2 \sim \mathrm{IG}(a_\tau, b_\tau), \quad \sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)$$

- 階層表現:

$$\begin{aligned}
y_{ij} \mid \tilde{z}_{ij} &\overset{\text{indep}}{\sim} \sum_{g=1}^{G} I(\log c_{g-1} \le \tilde{z}_{ij} < \log c_g) I(y_{ij} = g), \\
\tilde{z}_{ij} \mid \boldsymbol{\beta}, b_i, \sigma^2 &\overset{\text{indep}}{\sim} \mathrm{N}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i, \sigma^2), \\
b_i \mid \tau^2 &\overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2), \\
\boldsymbol{\beta} &\sim \mathrm{N}(\mathbf{0}, c_\beta \boldsymbol{I}_p), \\
\tau^2 &\sim \mathrm{IG}(a_\tau, b_\tau), \\
\sigma^2 &\overset{\text{iid}}{\sim} \mathrm{IG}(a_\sigma, b_\sigma),
\end{aligned}$$

ただし, $\tilde{z}_{ij} = \log(z_{ij})$.

Posterior computation

- full conditional distributions:

$$
\begin{aligned}
\tilde{z}_{ij} \mid - &\overset{\text{indep}}{\sim} \mathrm{TN}_{[\log c_{g-1}, \log c_g)}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i, \sigma_i^2) \quad \text{if } y_{ij} = g, \\
b_i \mid - &\overset{\text{indep}}{\sim} \mathrm{N}\left(\frac{\tau^2(\overline{\tilde{z}}_i - \overline{\boldsymbol{x}}_i^\top \boldsymbol{\beta})}{\tau^2 + \sigma_i^2/n_i}, \frac{\tau^2 \sigma_i^2/n_i}{\tau^2 + \sigma_i^2/n_i} \right), \\
\sigma^2 \mid - &\sim \mathrm{IG}\left(\frac{n + 2a_\sigma}{2}, \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\tilde{z}_{ij} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} - b_i)^2 + b_\sigma \right), \\
\tau^2 \mid - &\sim \mathrm{IG}\left(\frac{m + 2a_\tau}{2}, \frac{1}{2} \sum_{i=1}^{m} b_i^2 + b_\tau \right), \\
\boldsymbol{\beta} \mid - &\sim \mathrm{N}\left(\widetilde{\boldsymbol{\beta}}, (c_\beta^{-1} \boldsymbol{I}_p + \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \right),
\end{aligned}
$$

ただし，$\overline{\tilde{z}}_i = n_i^{-1} \sum_{j=1}^{n_i} \tilde{z}_{ij}$, $\widetilde{\boldsymbol{\beta}} = (c_\beta^{-1} \boldsymbol{I}_p + \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} (\tilde{\boldsymbol{z}} - \boldsymbol{J}\boldsymbol{b})$, $\boldsymbol{\Sigma} = \mathrm{diag}\,(\sigma_i^2)$, $\boldsymbol{J} = \mathrm{diag}\,(\boldsymbol{1}_{n_i})$ and $n = \sum_{i=1}^{m} n_i$.

- 上記の full conditional distributions にもとづいて，Gibbs サンプリングによって事後分布をシミュレーションできる.

有限母集団パラメータの推定

- $\boldsymbol{y}$ を所与とした $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{i,N_i})^\top$ の任意の関数（有限母集団パラメータ）の条件付分布 $f(h(\boldsymbol{z}_i) \mid \boldsymbol{y})$ は，以下のようにシミュレーションする.

  1. $\boldsymbol{y}$ を所与とした標本抽出された世帯の z_{ij}（$j = 1, \ldots, n_i$）の条件付分布は，MCMC のアウトプットからシミュレーションできる.
  2. $\boldsymbol{y}$ を所与とした out-of-sample の世帯の z_{ij}（$j = n_i + 1, \ldots, N_i$）の条件付分布:

$$f(z_{ij} \mid \boldsymbol{y}) = \iiint f(z_{ij} \mid \boldsymbol{\beta}, b_i, \sigma_i^2) \pi(\boldsymbol{\beta}, b_i, \sigma_i^2 \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}b_i \mathrm{d}\sigma_i^2,$$

  これは，$\{\boldsymbol{\beta}, b_i, \sigma_i^2\}$ の MCMC アウトプットからシミュレーションできる.

sampling design の考慮

- 住宅・土地統計調査は単純無作為抽出ではなく，バイアスのかかった標本調査のデザイン
- 住宅・土地統計調査の質問票情報は，sampling design の情報として survey weight を含んでいる.
- survey weight w_{ij} は，包含確率 π_{ij} に反比例する量，すなわち $w_{ij} \propto 1/\pi_{ij}$.
- survey weight w_{ij} は，（対数）尤度を重み付けて pseudo-likelihood を構成するのに用いられる（Parker, Janicki and Holan, 2019, arXiv）.
- survey weight w_{ij} は「標本 z_{ij} が母集団のいくつの世帯を代表しているか」に比例する量と解釈される.
- しばしば $\sum_{j=1}^{n_i} w_{ij} = n_i$ と基準化されて用いる.

Pseudo-likelihood and pseudo-posterior

- $\tilde{z}_{ij}$ の尤度への contribution を以下のように修正する.

$$\tilde{z}_{ij} \mid \boldsymbol{\beta}, b_i, \sigma^2 \sim f(\tilde{z}_{ij} \mid \boldsymbol{\beta}, b_i, \sigma^2)^{w_{ij}},$$

ただし $f(\tilde{z}_{ij} \mid \boldsymbol{\beta}, b_i, \sigma^2) = \mathrm{N}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i, \sigma^2)$.

- このとき, pseudo-posterior は以下の full conditional distributions にもとづいてシミュレーションできる.

$$\tilde{z}_{ij} \mid - \overset{\text{indep}}{\sim} \mathrm{TN}_{[\log c_{g-1}, \log c_g)}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + b_i, \sigma^2 / w_{ij}) \quad \text{if } y_{ij} = g,$$

$$b_i \mid - \overset{\text{indep}}{\sim} \mathrm{N}\left(\frac{\tau^2(\overline{\tilde{z}}_{iw} - \overline{\boldsymbol{x}}_{iw}{}^\top \boldsymbol{\beta})}{\tau^2 + \sigma^2/n_i}, \frac{\tau^2 \sigma^2 / n_i}{\tau^2 + \sigma^2/n_i} \right),$$

$$\sigma^2 \mid - \sim \mathrm{IG}\left(\frac{n + 2a_\sigma}{2}, \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} (\tilde{z}_{ij} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} - b_i)^2 + b_\sigma \right),$$

$$\tau^2 \mid - \sim \mathrm{IG}\left(\frac{m + 2a_\tau}{2}, \frac{1}{2} \sum_{i=1}^{m} b_i^2 + b_\tau \right),$$

$$\boldsymbol{\beta} \mid - \sim \mathrm{N}\left(\widetilde{\boldsymbol{\beta}}, (c_\beta^{-1} \boldsymbol{I}_p + \boldsymbol{X}^\top \boldsymbol{\Sigma}_w{}^{-1} \boldsymbol{X})^{-1} \right),$$

ただし $\overline{\tilde{z}}_{iw} = n_i^{-1} \sum_{j=1}^{n_i} w_{ij} \tilde{z}_{ij}$, $\overline{\boldsymbol{x}}_{iw} = n_i^{-1} \sum_{j=1}^{n_i} w_{ij} \boldsymbol{x}_{ij}$ and $\boldsymbol{\Sigma}_w = \mathrm{diag}\,(\sigma_i^2 / w_{ij})$.

pseudo-posterior 推定の問題点

- 住宅・土地統計調査では，w_{ij} のばらつきが大きい
- w_{ij} が小さい個票における $\tilde{z}_{ij}$ のサンプリングで，稀に極端に大きな値が生成されてしまう

$$\tilde{z}_{ij} \mid - \overset{\text{indep}}{\sim} \mathrm{TN}_{[\log c_{g-1}, \log c_g)}(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + b_i, \sigma^2 / w_{ij}) \quad \text{if } y_{ij} = g$$

- 分析する際，極端に大きな $\tilde{z}_{ij}$ のシミュレーション値は捨てるというアドホックな対応を行った

実データ解析

- 2018 年住宅・土地統計調査の調査票情報を用いて，市区町村ごとの平均所得とジニ係数を推定
- 補助変数
  - 「市町村税課税状況等の調」にもとづいた「1 人あたり課税対象所得額」
  - 「住宅・土地統計調査」と「国勢調査」にもとづいた，持ち家か否かのダミー変数

考察

- 本研究では，補助変数を所与とした所得（の対数変換）の条件付き分布を，線形混合モデルでモデリングした
- 補助変数の有限母集団分布が既知（全数で値が観測されている）という仮定により，所得分布を予測するという戦略
- しかしながら本研究では，補助変数として「1 人あたり課税対象所得額」と「持ち家ダミー」の 2 変数しか利用できず，モデルが脆弱である可能性が否定できない
- 補助変数の全数調査もしくはクロス集計表（同時分布）が利用可能である場合は，有効な手法

謝辞

この研究は，JSPS 科研費と，日本経済研究センター研究奨励金の支援を受けています.

- Kawakubo, Y. and Kobayashi, G. (2023). Small area estimation of general finite-population parameters based on grouped data. *Computational Statistics and Data Analysis*, **184**, 107741.

- Kobayashi, G., Sugasawa, S. and Kawakubo, Y. Spatio-temporal smoothing, interpolation and prediction of income distributions based on grouped data. *arXiv preprint*, arXiv: 2207.08384.

- Molina, I. and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, **38**, 369–385.

- Parker, P.A., Janicki, R. and Holan, S.H. (2019). Unit level modeling of survey data for small area estimation under informative sampling: A comprehensive overview with extensions. *arXiv preprint*, arXiv:1908.10488.

- Sugasawa, S., Kobayashi, G. and Kawakubo, Y. (2020). Estimation and inference for area-wise spatial income distributions from grouped data. *Computational Statistics and Data Analysis*, **145**, 106904.

- Walter, P., Groß, M., Schmid, T. and Tzavidis, N. (2021). Domain prediction with grouped income data. *Journal of the Royal Statistical Society. Series A*, **184**, 1501–1523.