資料２

全国家計構造調査における ウエイトの推定方法について

令和元年11月27日

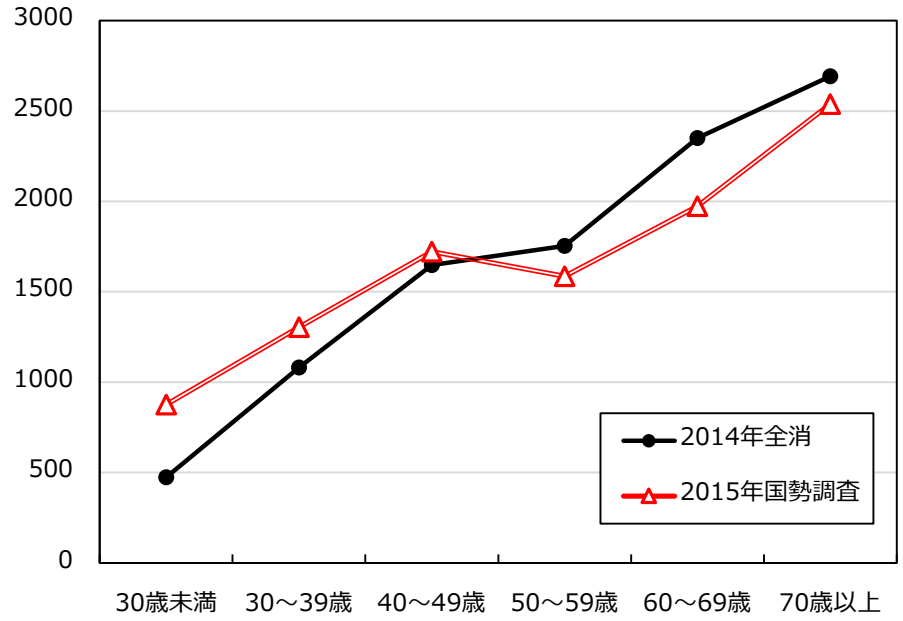
総務省統計局

前回の研究会における議論：ウエイト補正の必要性

世帯主の属性別世帯分布
（2014年全国消費実態調査結果，2015年国勢調査結果）

（10,000分比）

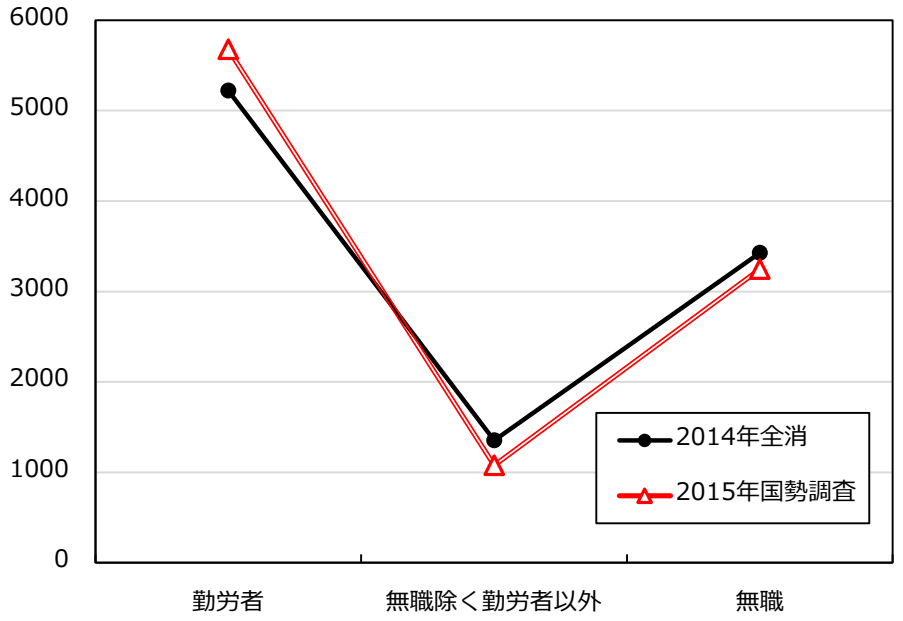
世帯主の年齢階級別



※国勢調査においては一般世帯のみ。年齢「不詳」を除く

（10,000分比）

世帯主の就業状況別



※国勢調査においては一般世帯のみ。全消の定義に合わせ特別集計を行った。

全国消費実態調査の結果は、国勢調査と比べ、以下のような特徴が見られる。

- 世帯主に若年者が少なく、高齢者が多い
- 世帯主に勤労者が少なく、勤労者以外が多い

前回の研究会における議論：IPF法の利用

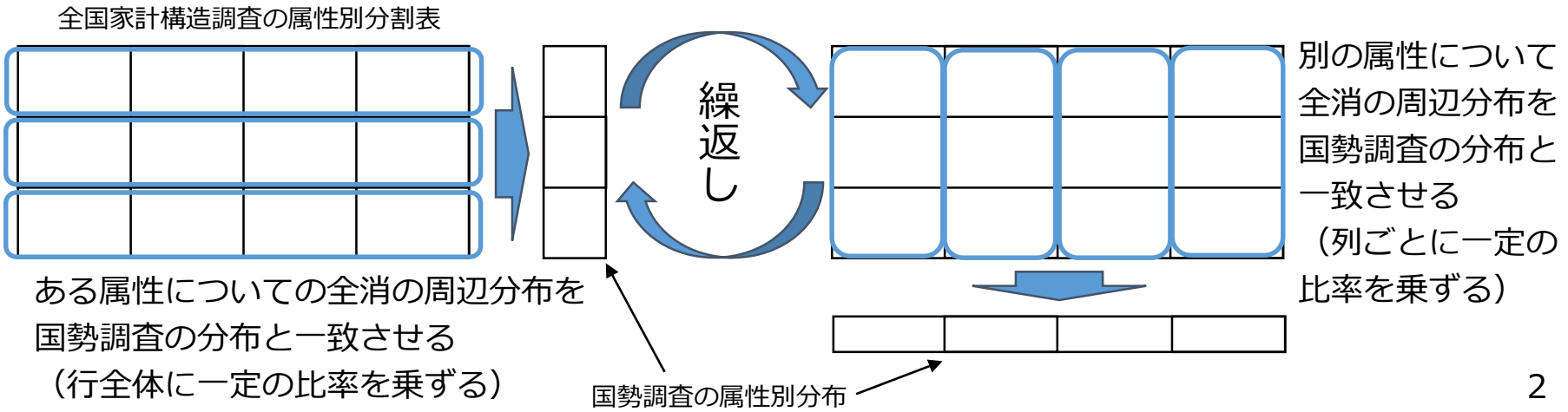
世帯の収支、資産等の分布に影響を与える世帯属性のうち、以下の基本的な属性の分布を全て都道府県別に補正したい。

- 世帯人員階級
- 世帯主の年齢階級
- 世帯主の就業状況（勤労、勤労以外、無職）

全国家計構造調査の世帯分布について、３つの属性による同時分布を、都道府県ごとに一致させるのは困難（調査世帯数が少ないため）

➡ 繰返し比例補正法（IPF法）により、３つの属性に対する周辺分布を満たす世帯分布を推定する

IPF法の計算イメージ（２次元の場合）



前回の研究会における議論：補正する変数の選択

階級の定義及び層の定義を以下のように変更

階級の定義（全ての都道府県で同じ・前回の計算から変更なし）

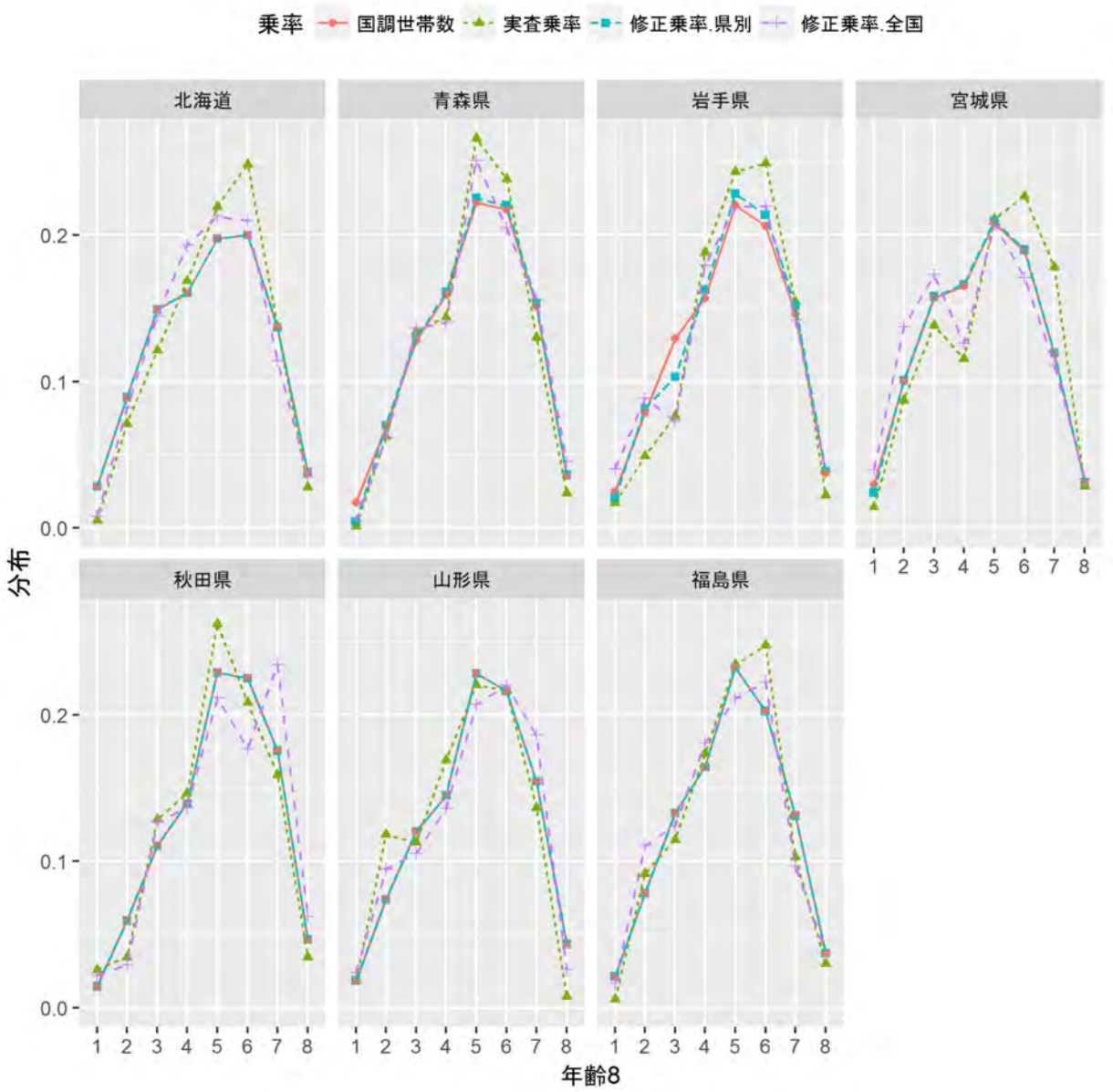
- 世帯主の年齢：25歳以下、25～34歳、…、75～84歳、85歳以上
- 世帯主の性別：男性、女性（単身世帯のみ）
- 世帯人員：１人、２人、３人、４人、５人以上
- 世帯主の就業状況：勤労世帯、無職世帯、その他

層の定義（全ての都道府県で同じ）

０層目　調整済調整係数
１層目　世帯人員（二人以上の世帯のみ）
２層目　世帯主の年齢（二人以上の世帯のみ）
３層目　世帯主の就業状況（二人以上の世帯のみ）
４層目　世帯主の性別×単身と二人以上の別（総世帯）
５層目　世帯主の年齢（総世帯）
６層目　世帯主の就業状況（総世帯）

二人以上の世帯と総世帯を別々に合わせることで、間接的に単身世帯の分布を合わせることを目指す。

前回の研究会における議論：補正結果①



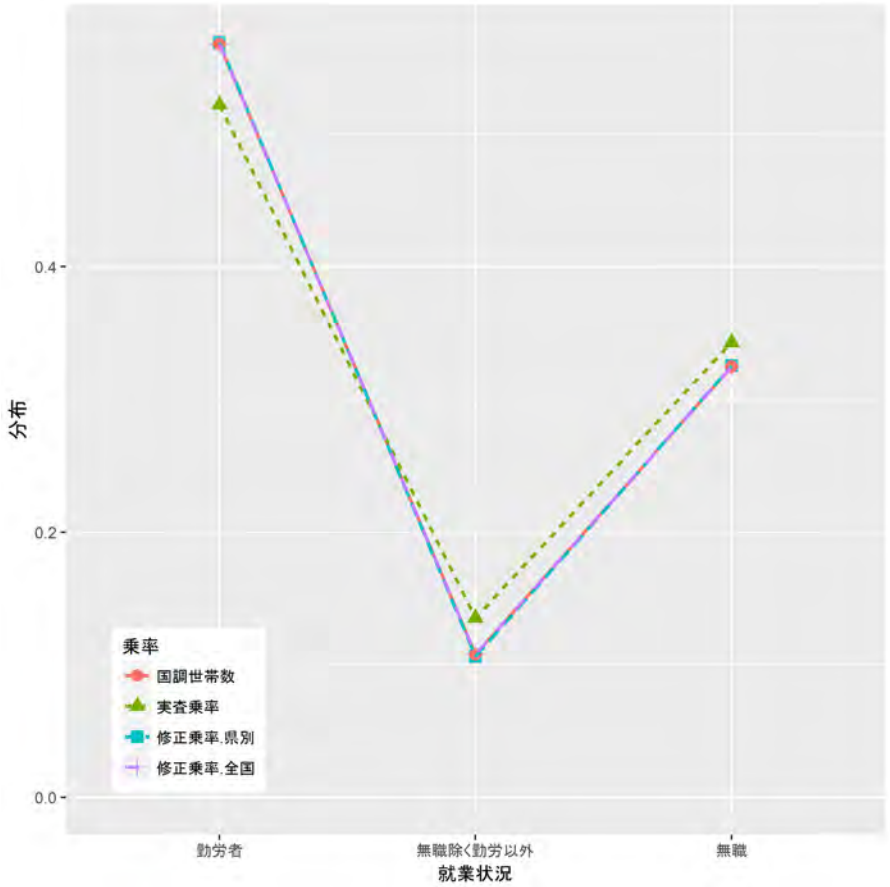
都道府県別年齢階級分布
（北海道・東北）

全国のデータで推定した結果もおおむね国勢調査の分布に合っているが、違いが大きくなる部分も存在する。都道府県別のデータで推定した結果の方は、国勢調査の分布とほぼ一致している。

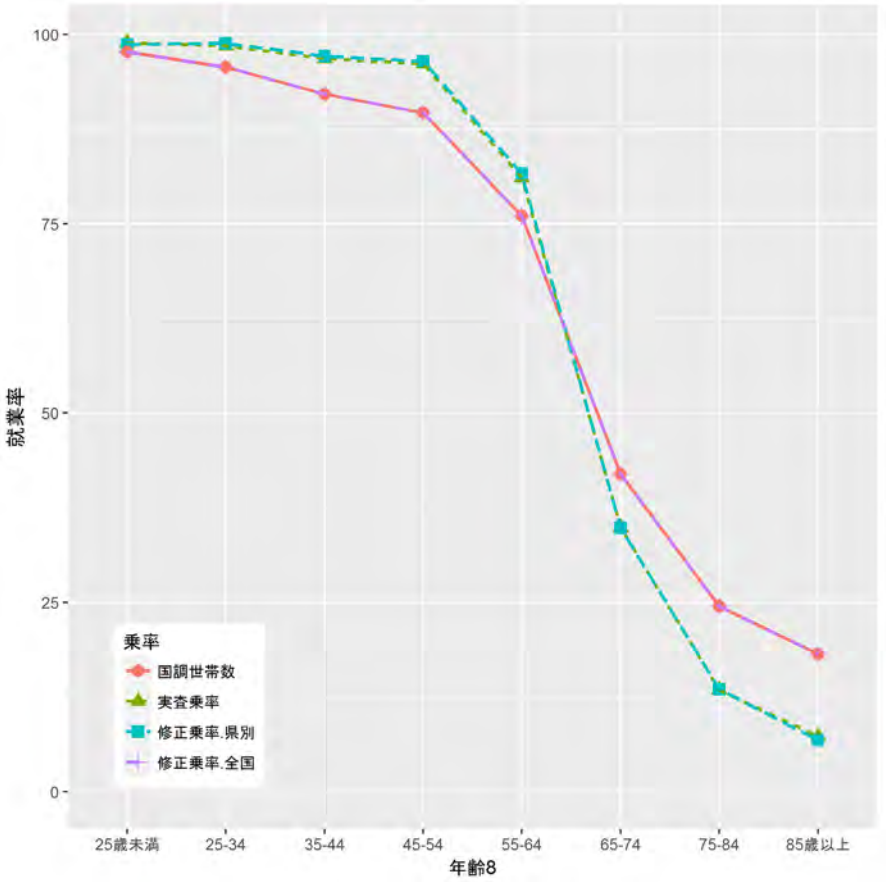
横軸の１～８はそれぞれ年齢階級の25歳未満～85歳以上に対応

前回の研究会における議論：補正結果②

就業状況別世帯分布



年齢階級別就業率



就業状況別の世帯分布（左図）についても、国勢調査をよく再現できている。
しかし、就業率（右図）については、県別での推定値は全く調整がされていない。
これは層の定義に年齢と就業状況の同時分布を用いることで解消される。

残る課題①：補正する変数の選択基準

同時分布を補正する必要がある変数の組合せ

Deville, et. al. (1993) における議論によると、あるサンプル調査における目的変数（消費支出など）と、サンプルの属性変数の間に以下の関係

（目的変数）＝（各属性変数の関数）＋（交差項）＋（確率項）

があるとき、

- ウエイト付与後に、交差項の存在する属性変数の同時分布が一致する
- 交差項がない

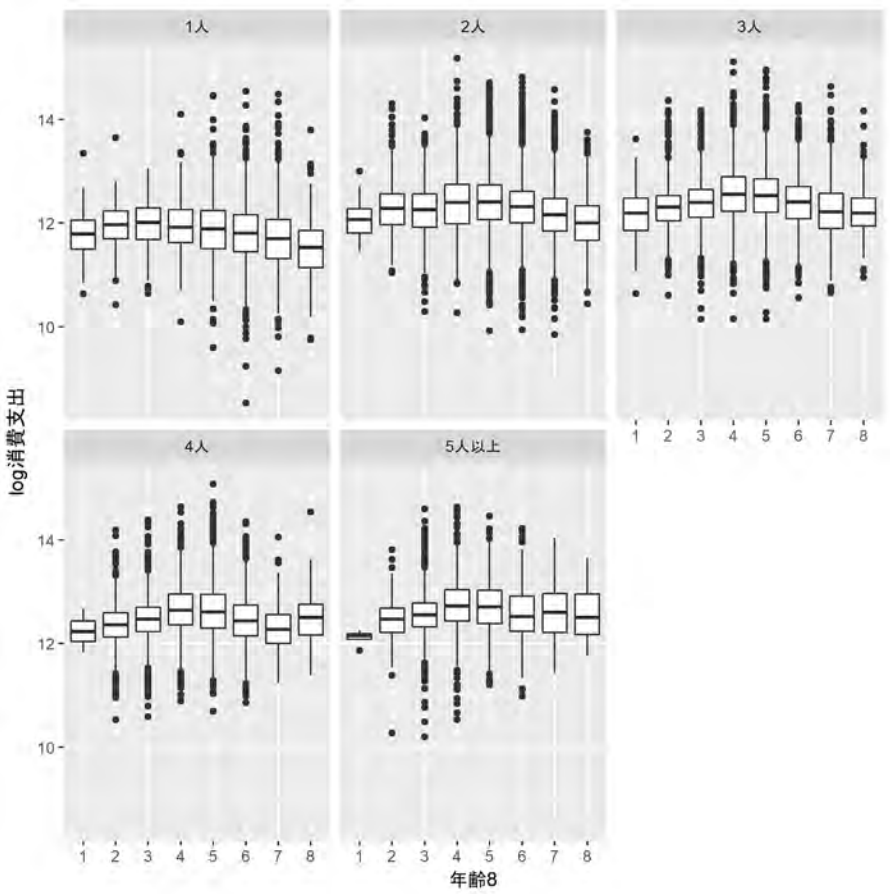
のどちらかが成立すれば、以下の式によりウエイトを利用した目的変数の推定結果にバイアスが生じないことが保証される（詳しくは付録参照）。

（バイアス）＝（同時分布のずれ）×（交差項）

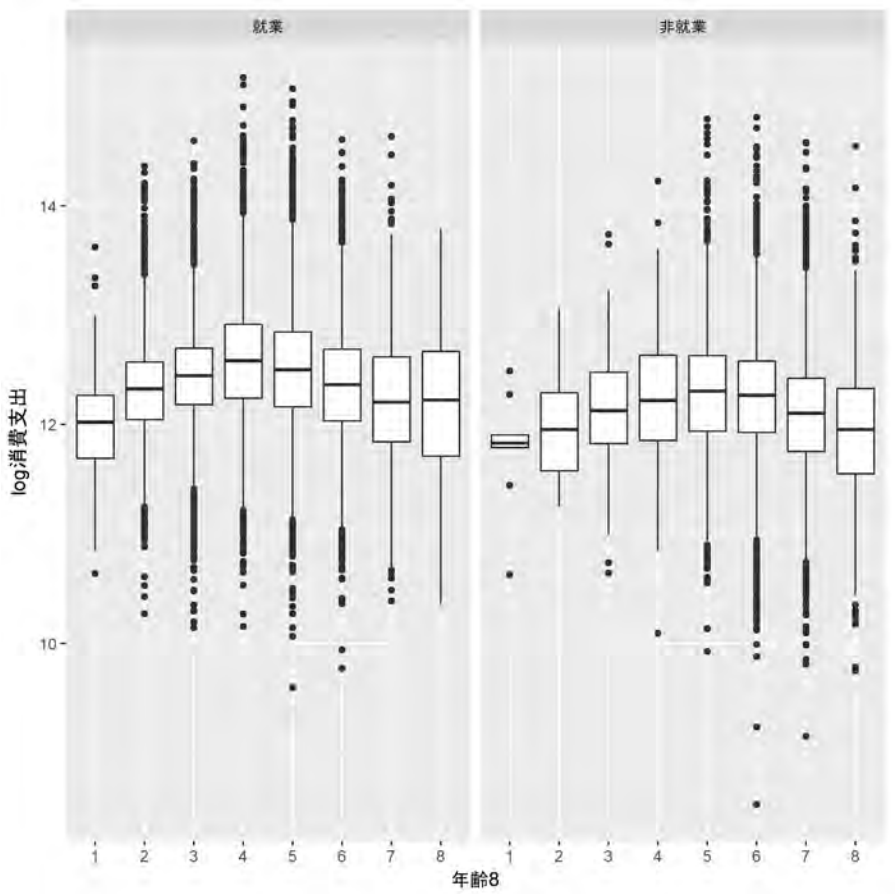
逆に言えば、目的変数の回帰式について交差項が存在するような属性変数がある場合、ウエイトによりその同時分布を再現できなければ、推定結果にバイアスが残ることとなる。

消費支出の分布と属性変数の関係

世帯人員別・世帯主の年齢階級別



就業状況別・世帯主の年齢階級別



世帯人員や世帯主の就業状況により、消費支出と世帯主の年齢階級の関係も変化

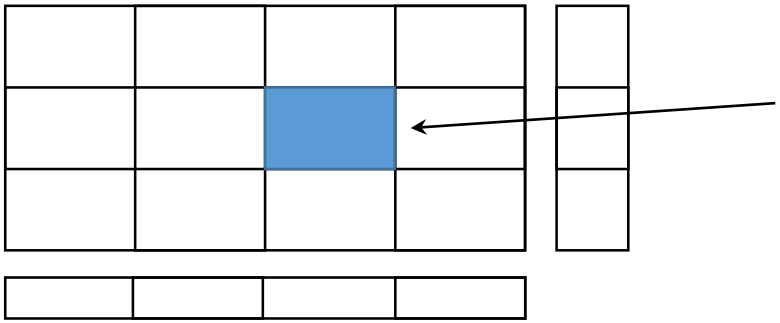
⇒　交差項が存在する可能性

（世帯人員と就業状況の間には明確な関係は見られない）

残る課題②：IPF法における分布の初期値

繰返し計算に用いる初期値の選択

IPF法における各層の定義について、含まれる世帯が0となるセルが存在すると、繰返し計算が収束しなくなる。そこで0となるセルにごく小さなウエイトを割り当てることで、計算が収束しない問題を回避することができる（Beckman, et. al., 1996)。



含まれる世帯数が0となる場合に
ごく小さなウエイト（0.01など）
を割り当てる

この方法は簡便である一方で推定結果にバイアスを生む可能性が、Guo, Bhat (2006) において指摘されている。
それに対しYe, et. al. (2009) において、割り当てるウエイトを任意の小さな値ではなく、調査地域における世帯分布を反映した値とする方法が提案されている。

前回の議論からの変更点①

(1)世帯数0のセルの初期ウエイトに国勢調査の分布を利用

Ye, et. al. (2009) における提案を参考に、IPF計算の初期分布において世帯数が0となるセルが発生した場合の対処法を変更

（旧方式）一律にごく小さなウエイト（他のセルの1万分の1以下）を入れる

↓

（新方式）国勢調査の世帯分布を基にしたウエイトを入れる

なお新方式でも、過大なウエイトが追加されることを防ぐため、追加したウエイトの合計は旧方式と一致するようにした。

(2)IPF法における層の定義の変更

Deville, et. al. (1993)における議論を踏まえ、消費支出の回帰関数を考えたときに、交差項が存在する可能性がある変数の組合せについて、IPF法における層の定義に追加

〈最終的な層の定義〉（全ての都道府県で同じ）

０層目　調整済調整係数
１層目　世帯人員×年齢（二人以上）
２層目　年齢×就業状況（二人以上）
３層目　世帯人員×年齢（総世帯）
４層目　年齢×就業状況（総世帯）

前回の議論からの変更点②

(3)全国単身世帯収支実態調査結果の統合

全国家計構造調査と同時に、全国単身世帯収支実態調査を実施中

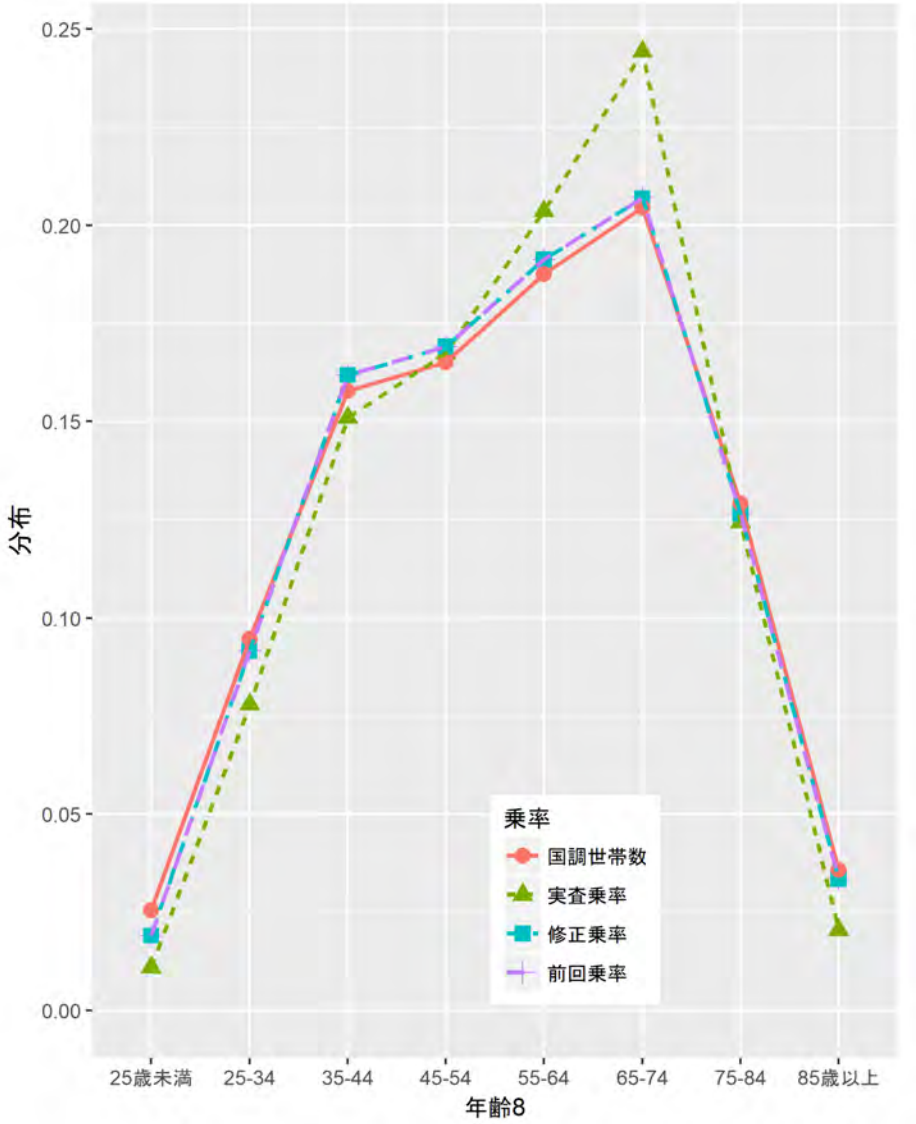
調査項目：全国家計構造調査と同じ
調査対象：全国の単身世帯　約2,000世帯
抽出方法：民間の調査会社に登録しているモニターから抽出

単身世帯のサンプルサイズを拡大するため、全国家計構造調査と全国単身世帯収支実態調査のデータを、以下の手順により統合する（詳細は付録参照）。

1. 全国単身世帯収支実態調査の調査世帯について調整係数を設定
2. 傾向スコアにより全国単身世帯収支実態調査の分布を補正（傾向スコアの推定には全国のデータを使用）
3. 全国家計構造調査のデータと全国単身世帯収支実態調査のデータを、適切なウエイトをつけて結合（今回はサンプルサイズの逆比を使用）
4. 結合後のデータについてIPF計算を行い、ウエイトを都道府県別に補正

ウエイト推定結果：初期分布の変更による影響

年齢階級別世帯分布（総世帯・全国）



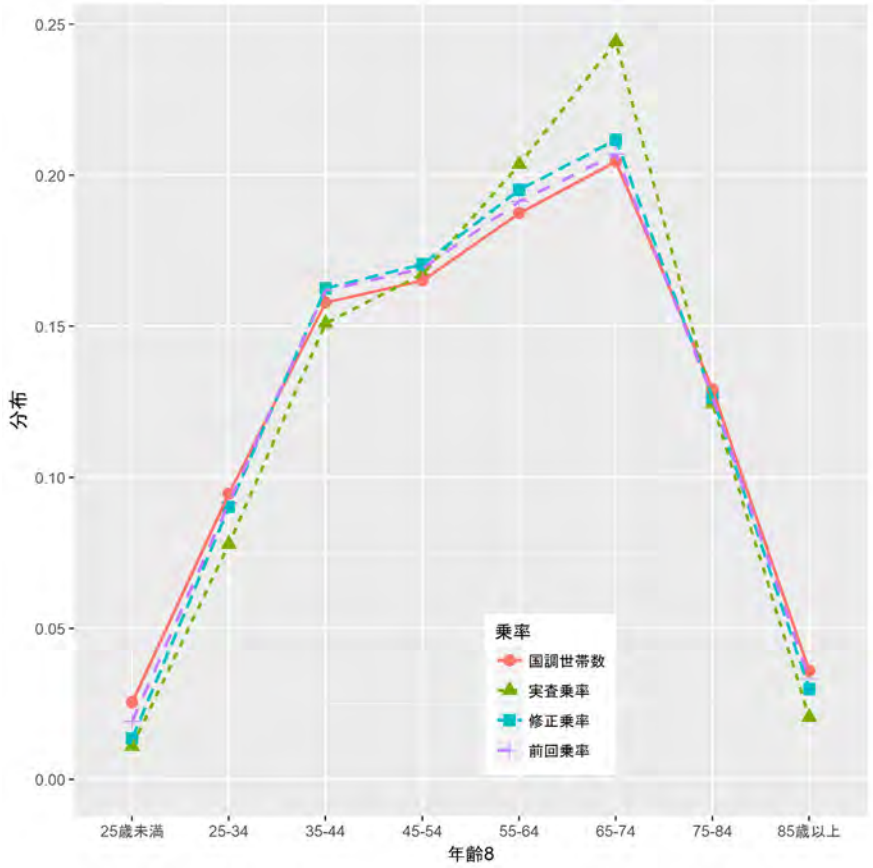
前回の研究会における結果と同じ層定義を用いてIPF法の計算を行った。

初期分布におけるウエイト補定の方法を変更したことによる分布への影響は見られなかった（県別でも同様）。

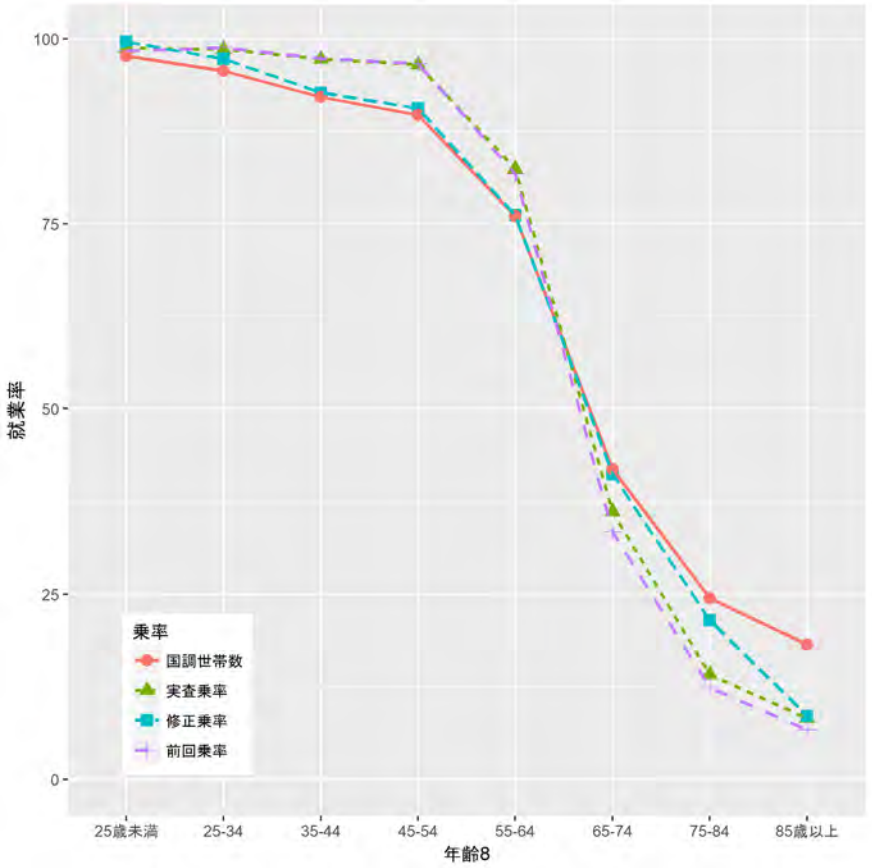
計算が収束する速さについては、過半数の都道府県では影響が見られなかったが、一部において収束が遅くなった。
ただし計算回数は最大で1.3倍程度の範囲にとどまっており、大きな影響を与える要因とはなっていない。

ウエイト推定結果：層の定義の変更による影響

年齢階級別分布（総世帯・全国）



年齢階級別就業率（総世帯・全国）

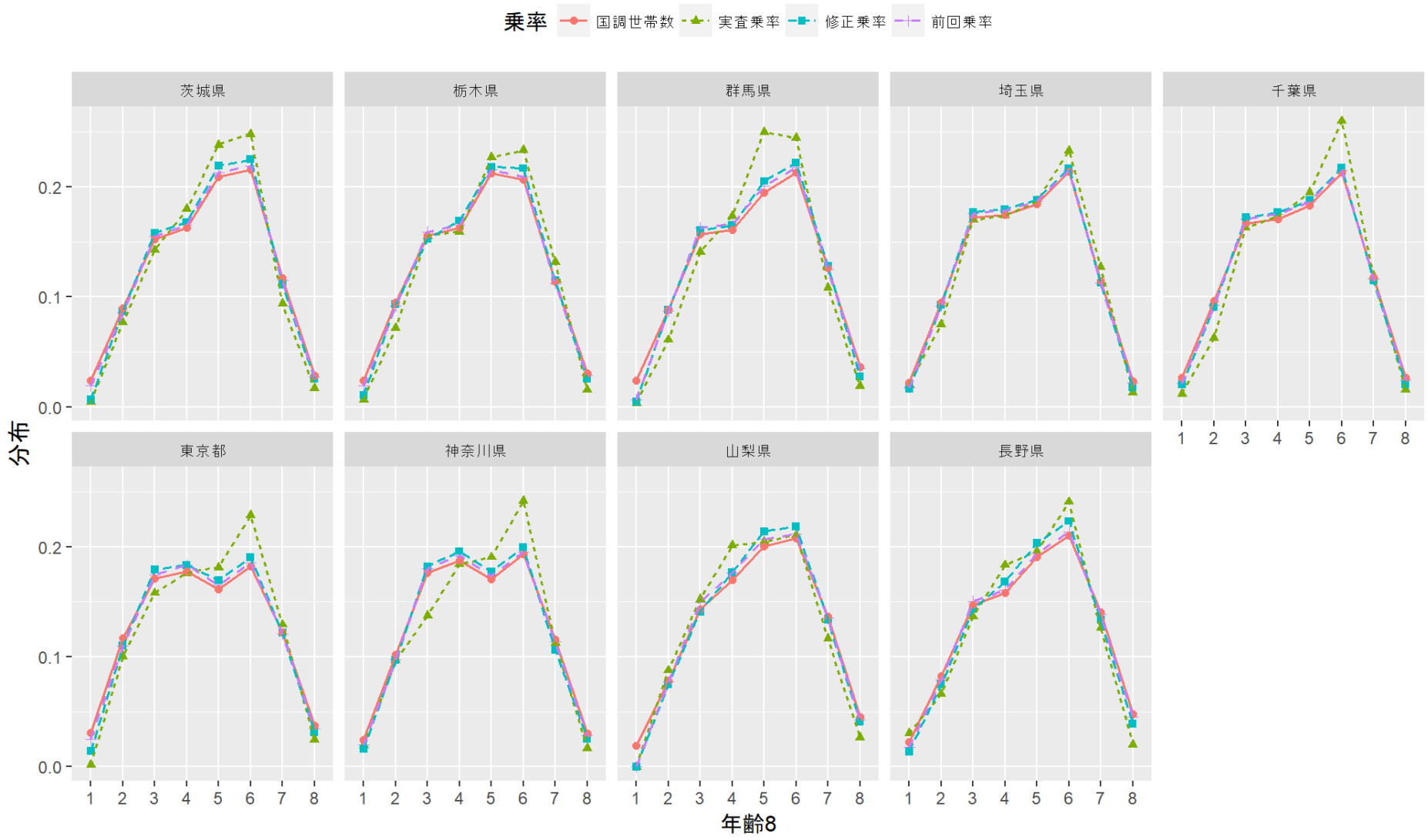


前回の研究会における結果と比較して

- 単一の変数に関する分布は当てはまりがやや悪化
- 同時分布に影響を受ける量の推定値は当てはまりが改善

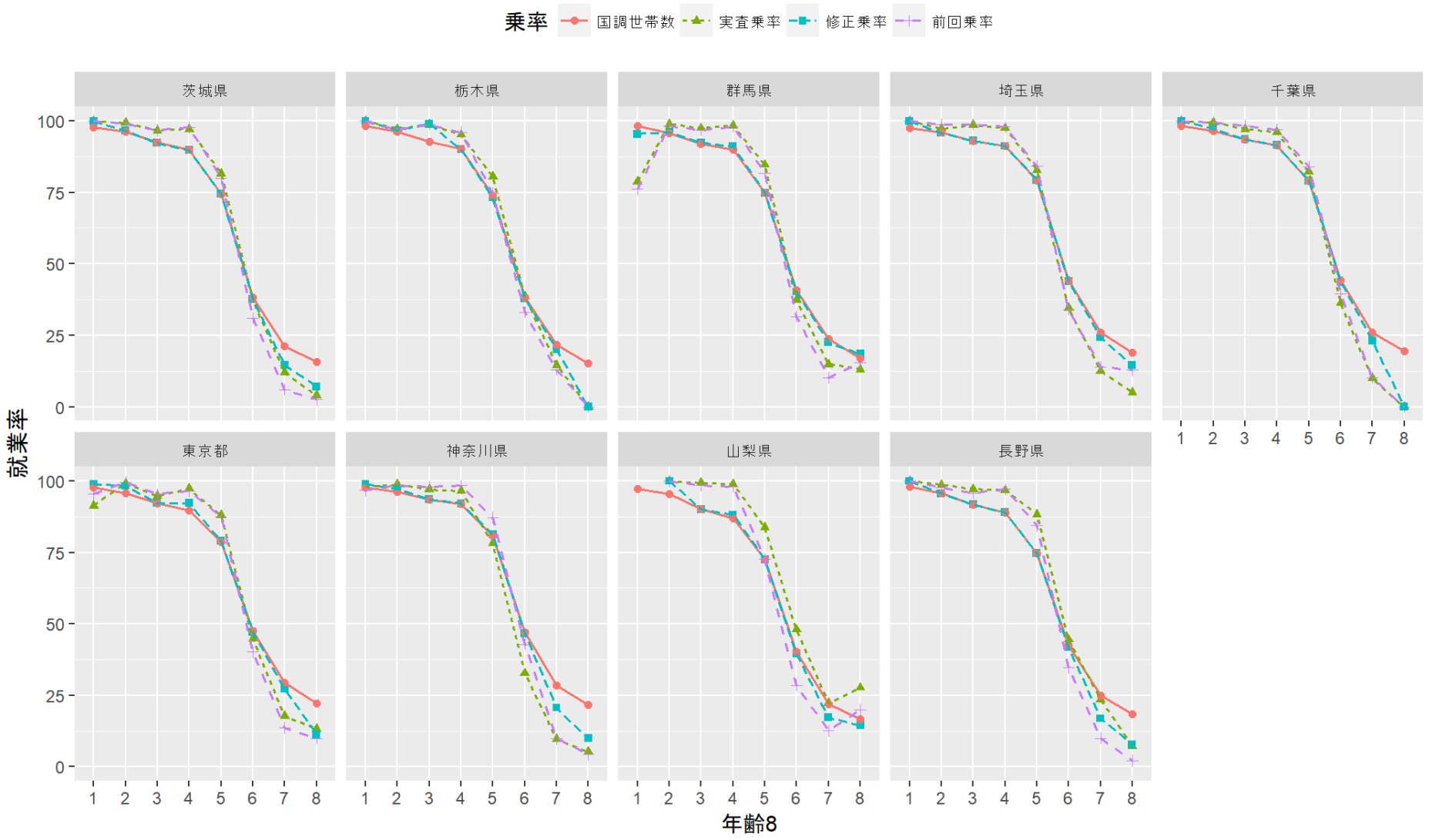
ウエイト推定結果：層の定義の変更による影響

年齢階級別分布（総世帯・関東）



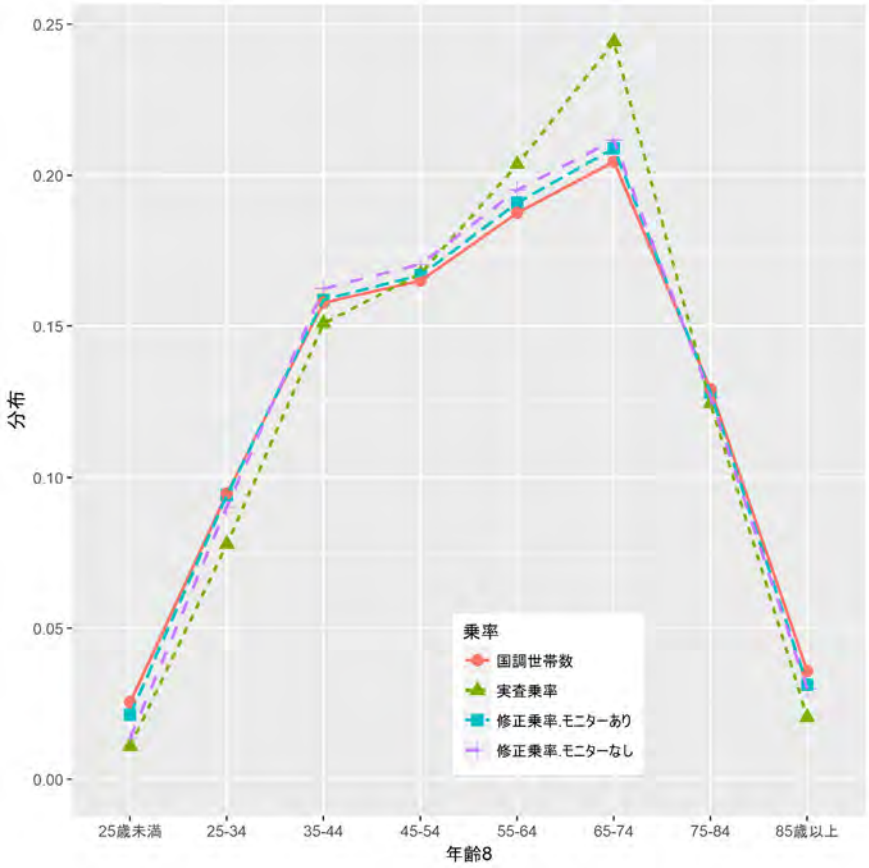
ウエイト推定結果：層の定義の変更による影響

年齢階級別就業率（総世帯・関東）

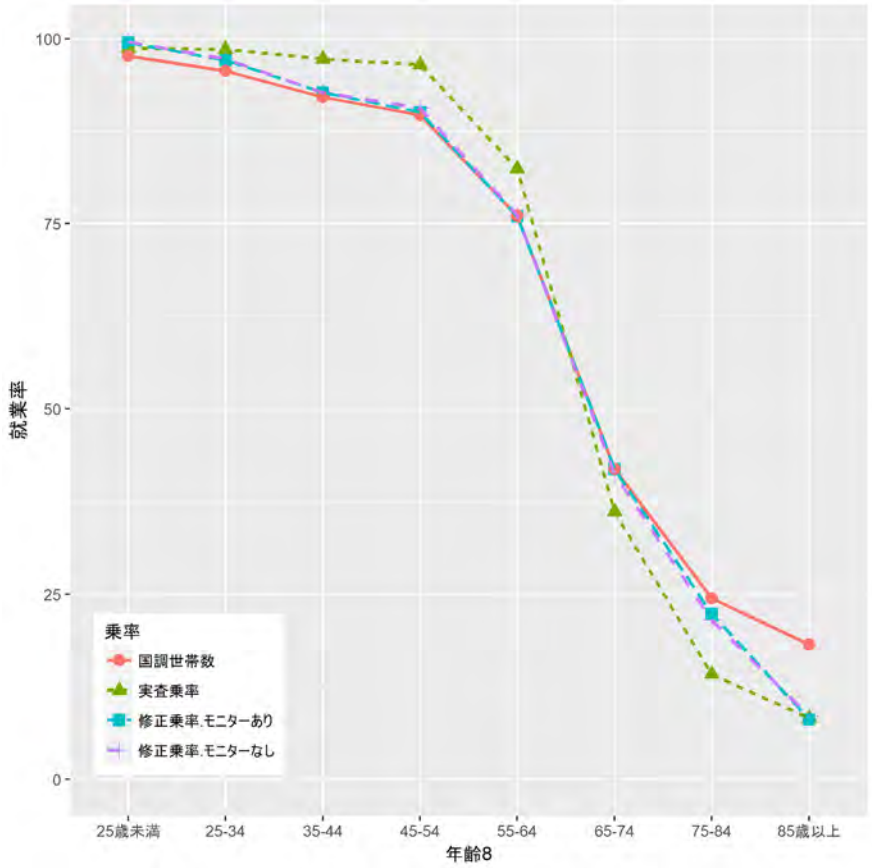


ウエイト推定結果：全単データの統合による影響

年齢階級別分布（総世帯・全国）



年齢階級別就業率（総世帯・全国）

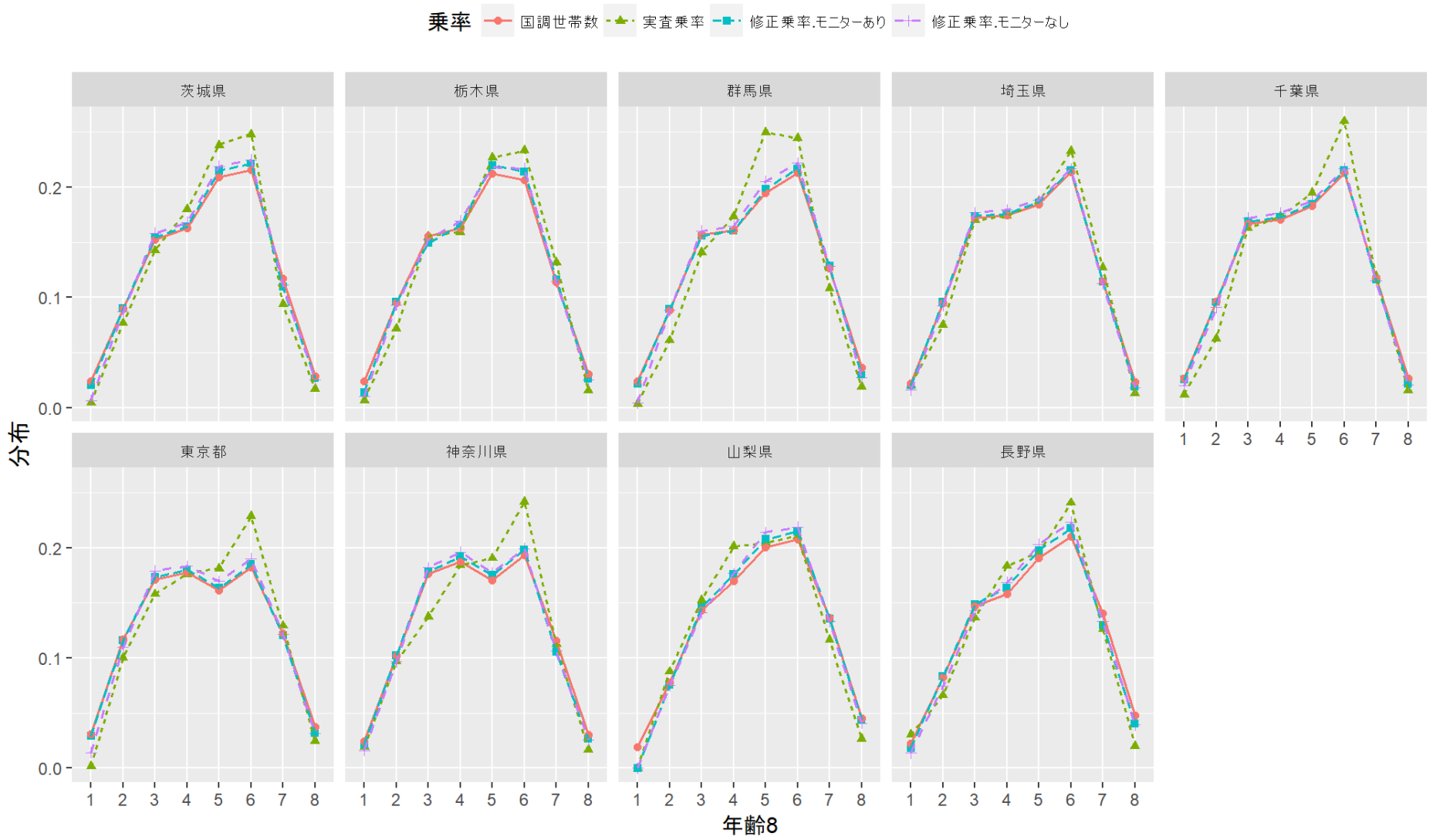


全国単身世帯収支実態調査のデータを用いない場合と比較して

- 主に若年層（全消で世帯が少ない部分）について世帯分布が改善
- 年齢階級別就業率に大きな変化はない　←傾向スコアによる分布補正の影響か

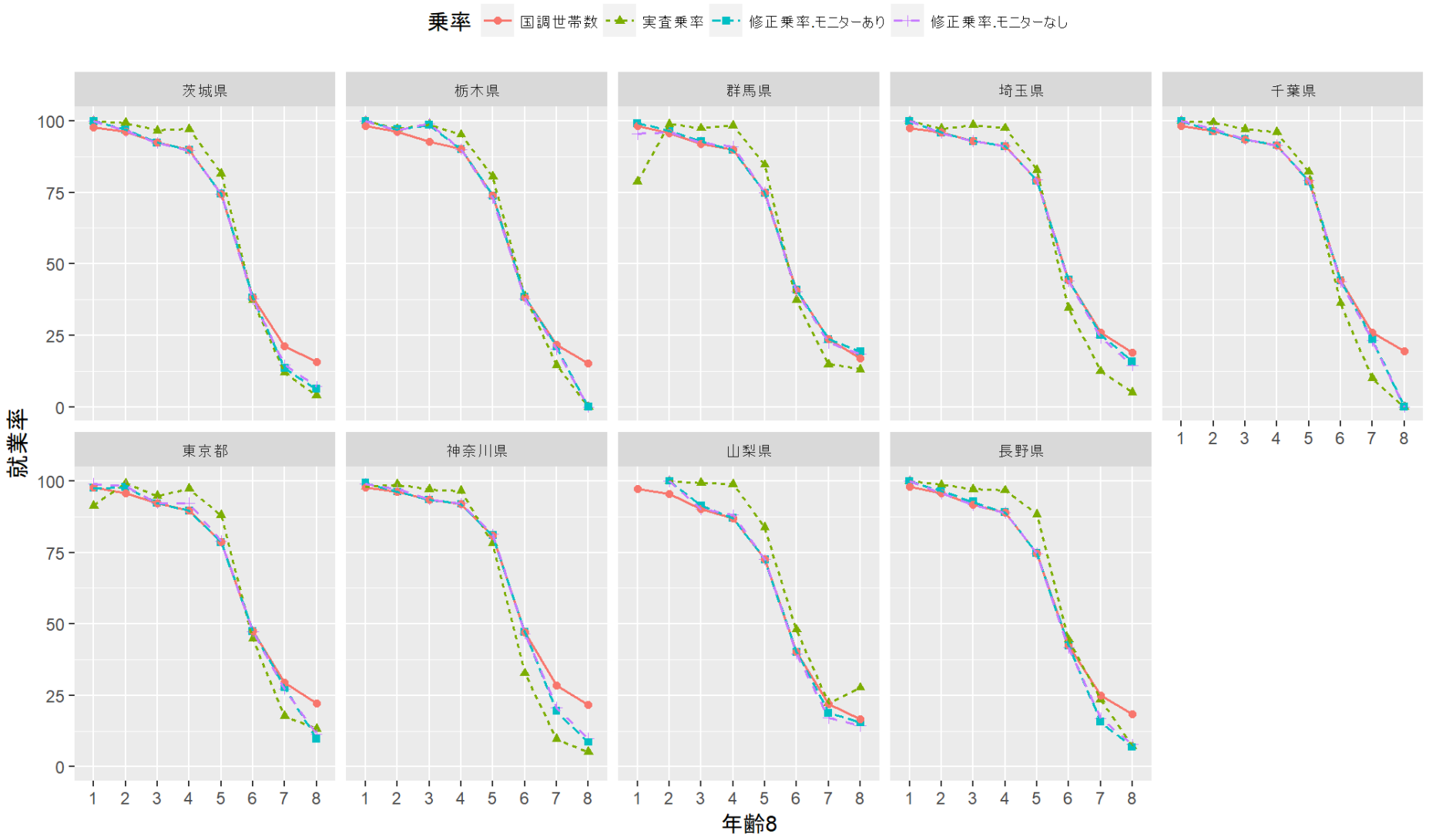
ウエイト推定結果：全単データの統合による影響

年齢階級別分布（総世帯・関東）



ウエイト推定結果：全単データの統合による影響

年齢階級別就業率（総世帯・関東）



結果のまとめ

前回の研究会での結果からの変更点とその影響は以下のとおり。

①IPF法の繰返し計算の初期値について、国勢調査の世帯分布を利用したウエイト補定を行った。
⇒計算の収束が少し遅くなる以外、大きな影響は見られなかった。

②消費支出の回帰式に交差項が存在すると考えられる世帯属性の組合せについて、補正後のウエイトによるバイアスの発生を防ぐため、IPF法の層の定義に同時分布を入れることで補正を行った。
⇒単一の世帯属性に関する世帯分布の当てはまりが少し悪くなるものの、２つの世帯属性の同時分布について当てはまりが改善した。

③サンプルサイズを増やすため、全国消費実態調査と全国単身世帯収支実態調査の結果を統合したデータを作成し、IPF法による補正を行った。
⇒全単データの導入により、全消において世帯の少ない属性部分のデータが補完され、より当てはまりのよい分布を得られた。

次回に向けた課題：参照データの時点調整

今回の分析では2015年国勢調査のデータのうち、年齢を１歳若くした分布を用いたが、全国家計構造調査での使用には、以下のような問題点がある。

○2019年全国家計構造調査の公表までに2020年国勢調査の結果は公表されないため、国勢調査については2015年以前のデータしか利用できない。

この点を解決するために、以下の方法を検討している。

- 2015年国勢調査結果における世帯属性別の世帯数を基に、労働力調査など他の世帯調査結果による「世帯数の増減率」を利用し、外挿する。
- 2015年国勢調査結果を初期値、2019年の労働力調査など他の世帯調査による結果を周辺分布とし、IPF法により2019年の国勢調査における分布を推定する。

これらの検討結果については、次回の消費統計研究会において議論する予定である。

参考文献

Beckman, R. J., Baggerly, K. A., McKay, M. D. (1996), “Creating Synthetic Baseline Populations”, Transportation Research Part A: Policy and Practice, 30(6), 415

Deville, J. C., Särundal, C. E., Sautory, O. (1993), “Generalized Raking Procedures in Survey Sampling”, Journal of the American Statistical Association, 88, 423

Guo, J.Y., Bhat, C.R. (2007), “Population Synthesis for Microsimulating Travel Behavior”, Transportation Research Record: Journal of the Transportation Research Board, 2014, 92-101

Ye, X., Konduri, K., Pendyala, R.M., Sana, B., Waddell, P. (2009) “A methodology to match distributions of both household and person attributes in the generation of synthetic populations”, 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

付録

Deville, et. al. における議論の詳細

①同時分布を補正する必要がある変数の組合せ

Deville, et. al. (1993)における議論

目的変数yと属性変数x, zに以下の関係があるとする。

$$y = m_x(x) + m_z(z) + b(x, z) + u$$

ここで$m_x(x), m_z(z), b(x, z)$はそれぞれx, zの関数、uは確率項

このとき、サンプル調査で得たデータ$\{y_i\}_{i=1,\cdots,n}$とウエイト$\{w_i\}_{i=1,\cdots,n}$を用いて得られる、母集団におけるyの総和t_yの推定値$\hat{t}_y = \sum_{i=1}^{n} w_i y_i$について、以下の式が成り立つ。

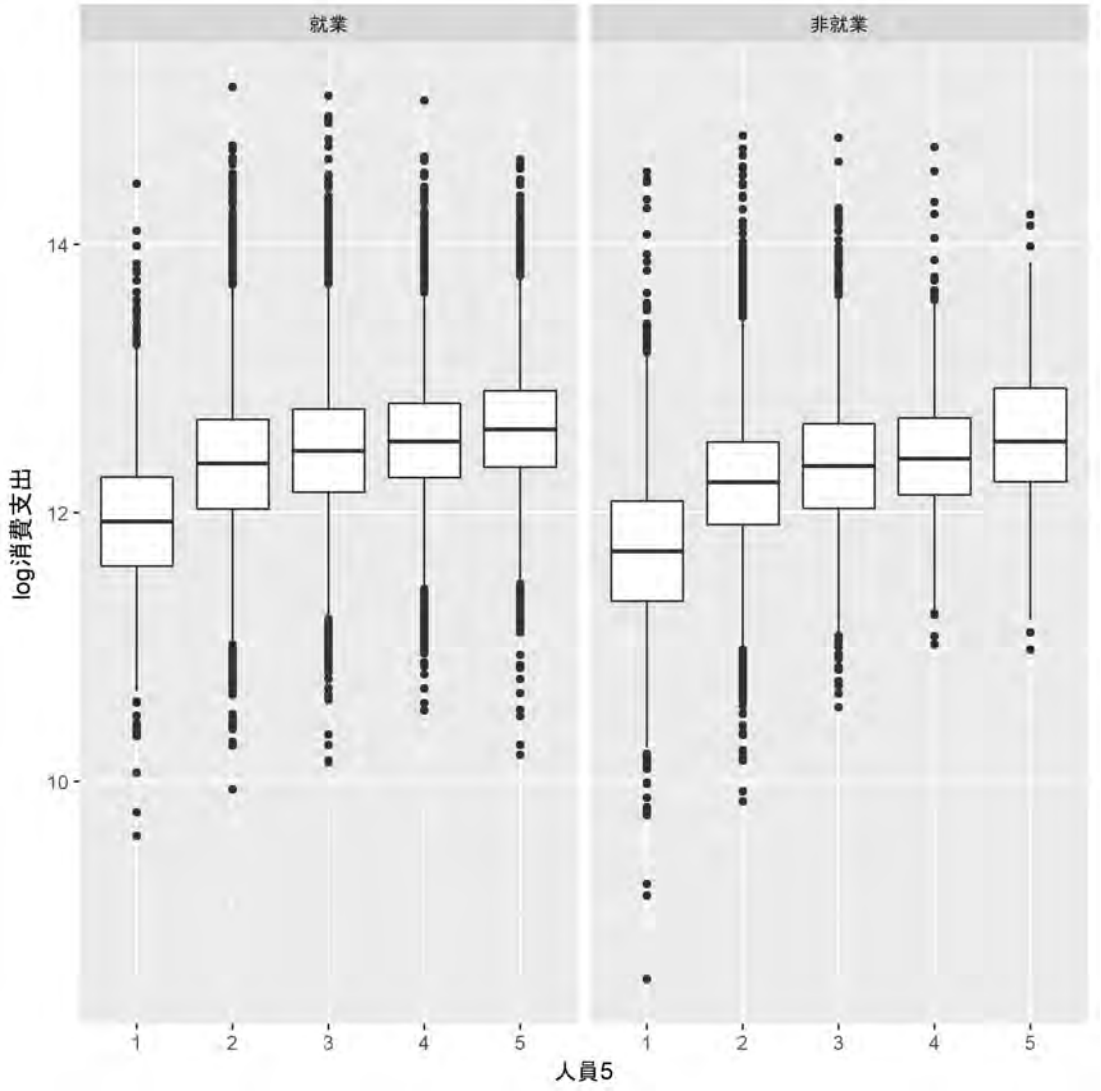
$$E\left[\hat{t}_y - t_y \middle| \widehat{N}\right] = \sum_{r,c} \left(\widehat{N}_{rc}^{w} - N_{rc}\right) \times b(r, c)$$

ここでr, cはそれぞれx, zの階級$(r = 1, \cdots, R, c = 1, \cdots, C)$、$\widehat{N}_{rc}^{w}$は対応する階級の世帯数の推定値、$\widehat{N} = (\widehat{N}_{11}, \widehat{N}_{12}, \cdots, \widehat{N}_{RC})$

またx, zを条件付けたとき、確率項の期待値は各世帯で0となる仮定をおく。

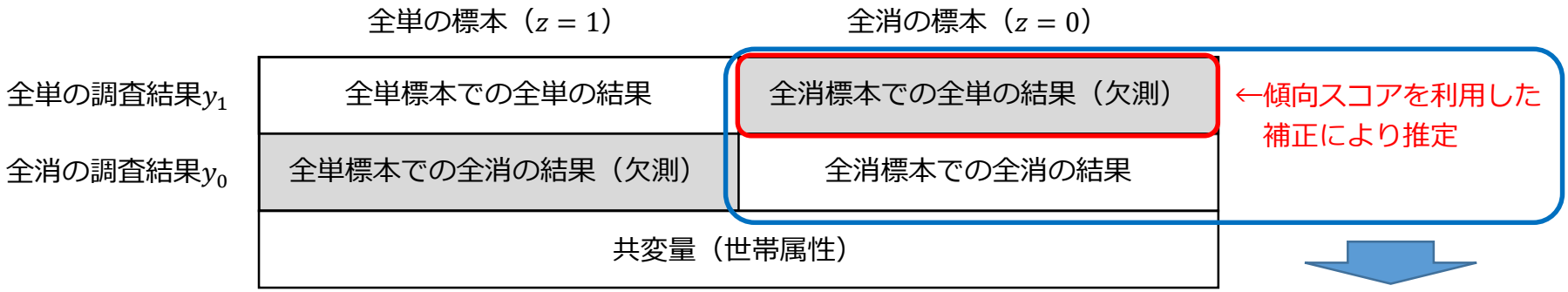
消費支出の分布と属性変数の関係

世帯主の就業状況別・世帯人員別



全消と全単を統合したウエイトの作成方法

全消の調整係数をα、全単の調整係数をβとする。



推定値を適切なウエイトで合成し、統合したサンプルによる推定値を求める

全消の世帯分布に対して全単の世帯分布がもつバイアスを補正するため、全消及び全単における調査世帯が全単標本に入る確率e（傾向スコア）を推定し、全単の各世帯に対し、傾向スコアによる補正ウエイト$\frac{1-e}{e}$を割り当てる。

さらに全消標本におけるy_0の推定量$\bar{y}_0$と、傾向スコアにより補正した全単標本によるy_1の推定量$\hat{E}[y_1|z=0]$を適切なウエイトw及び$1-w$で合成することにより、統合したサンプルによる推定値を求めることができる（さらに詳しくは前回研究会の資料２を参照）。

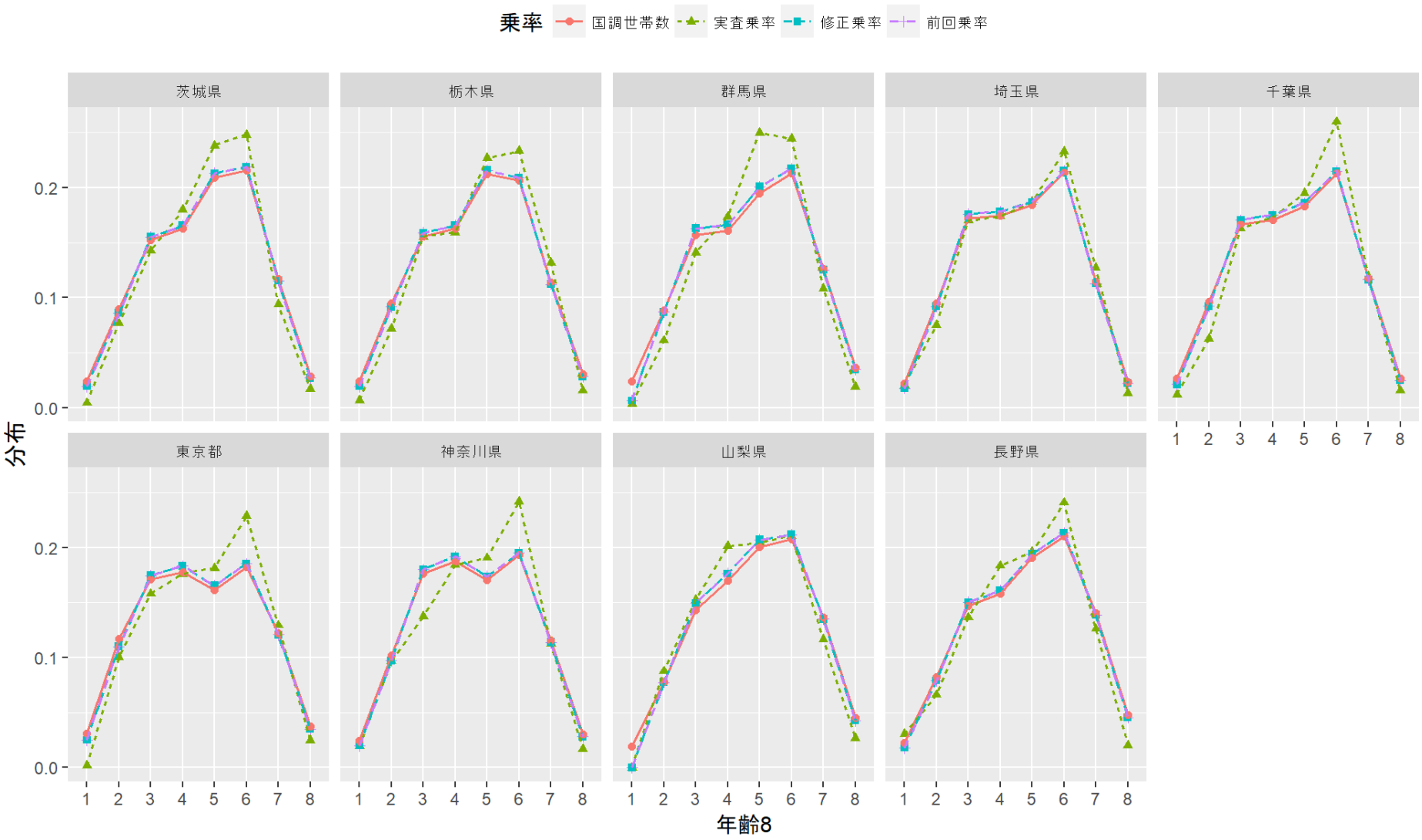
以上をまとめると、全消及び全単の統合データにおけるウエイトは以下のようになる。

全消の調査世帯：$w\alpha$

全単の調査世帯：$(1-w)\frac{1-e}{e}\beta$

} IPF法における初期値として使用

ウエイト推定結果：初期分布の変更による影響



参考：全国家計構造調査の設計概要