資料３

家計調査を利用した
全国消費実態調査の時点間補完
に向けて
―枠組みの提示と解析例の紹介―

慶應義塾大学
経済学部・大学院経済学研究科
（兼）理研AIP　経済経営情報融合分析チーム
（兼）総務省統計研究研修所
星野崇宏

問題意識

全国消費実態調査のメリット／デメリット

　サンプルサイズが大きい→詳細な区分での分析

　2か月(前回3か月)　5年に一度

家計調査のメリット／デメリット

　通年で結果が得られるローテーションパネル

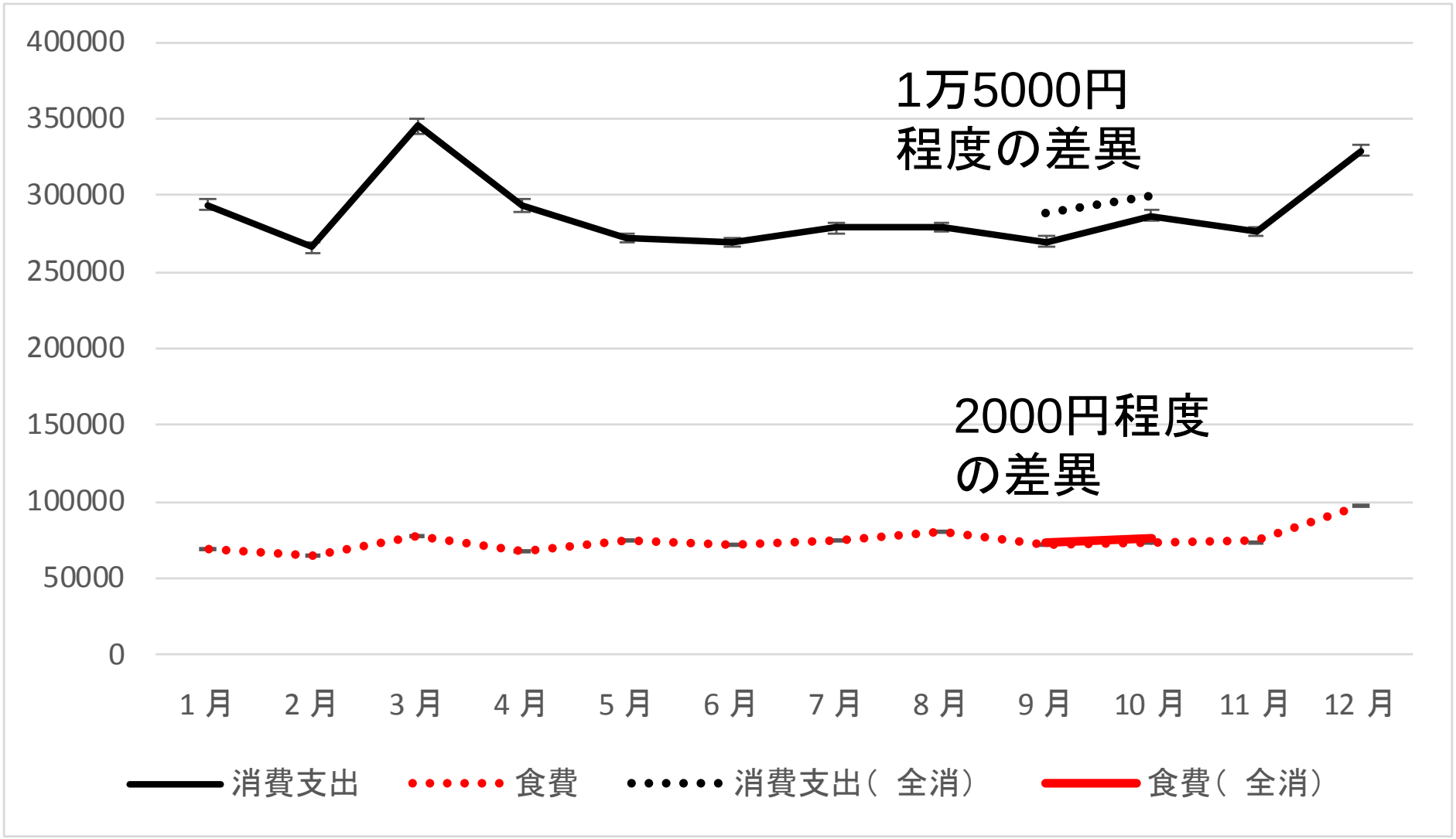
　比較的サンプルサイズが小さい

食費など月次の変動が大きいものについて両者を融合させて通年での推測をしたい

〇個票データでの補完　×時系列データとしての補完

⇒両調査の調査モードや標本の違いを反映し精度の高い推定

例：月次での支出の変動



＊消費や食費で家計調査が低いのは過小記入バイアスや倹約化か？
(過小記入バイアスDeaton&Irish,1984;牧,2007; 倹約化・調査疲れStephens&Unayama,2011)

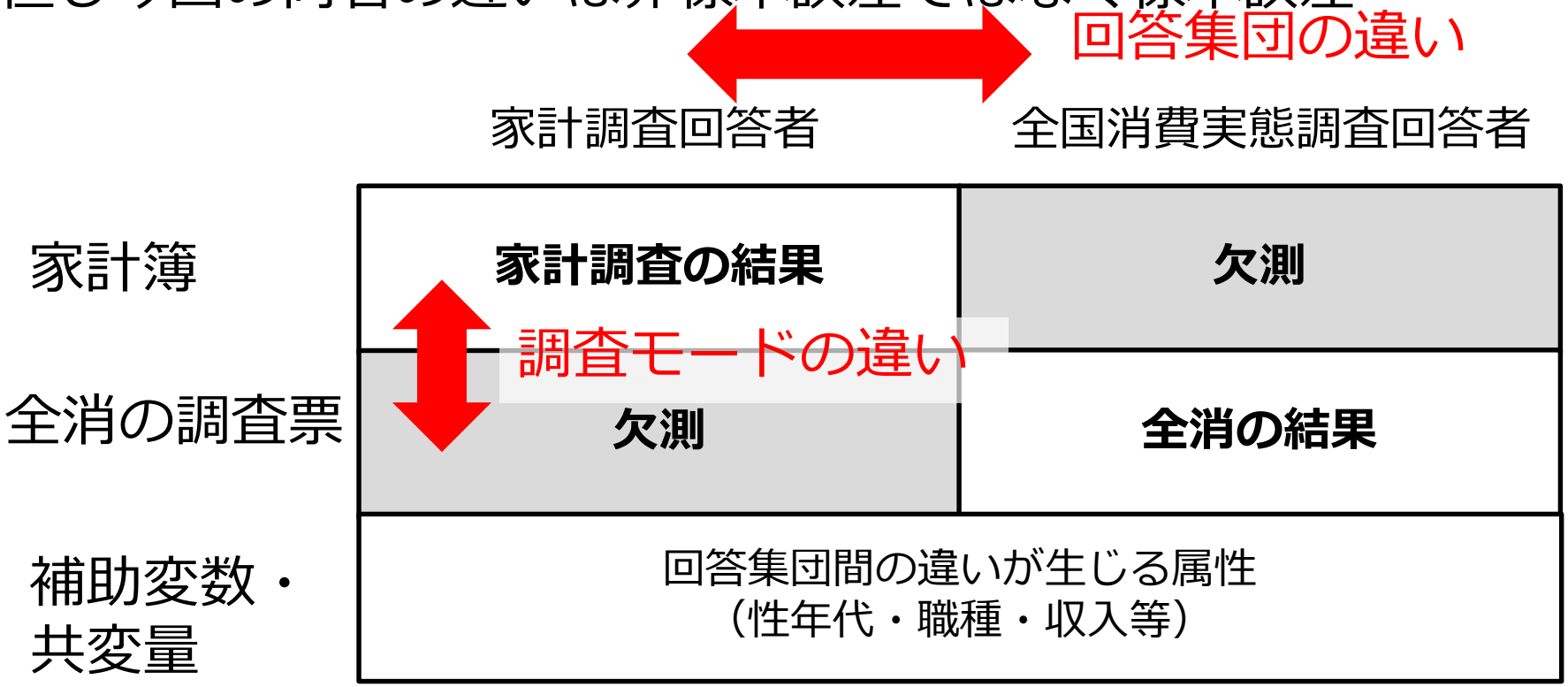
家計調査と全国消費実態調査の調査の違い

“選択バイアス”＝回答集団の違い

“調査・データ取得モードの違い”＝取り方の違い(今回は質問)

⇒両者の違いが混ざっているので分離して議論したい

＊但し今回の両者の違いは非標本誤差ではなく標本誤差



目的と方法

【目的】月次の変動が大きい消費について両調査を融合させて通年での推測／年平均の計算

【方法論】個票データで欠測が存在していることを仮定した

欠測データ解析

具体的にはEMアルゴリズムを用いた平均と相関構造の推定

【注意点】

１）実用的な方法としては6月中に完成とのことのため、今回は枠組みの提示と「可能性」の検討のレベル

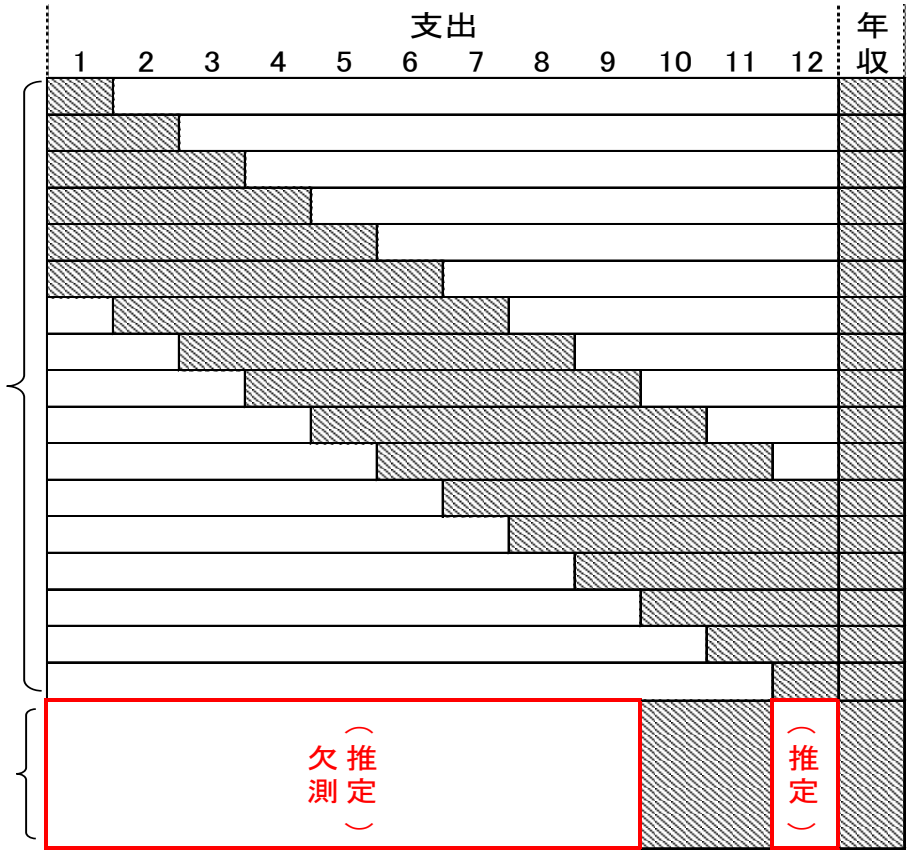
２）集計時系列解析での季節性を考慮する方法とは異なる

３）家計調査と全国消費実態調査の回答者の標本誤差や過小記載の構造・項目（以後”調査モード”と呼称）などが同じとする仮定と異なるとする仮定で実施

提案手法

ローテーショナルパネルである
家計調査と全国消費実態調査の
関係を右図のように整理し、
欠測データの構造であること
を考慮し両者を併合させた分析
仮定A)欠測が「ランダムな欠測」
仮定B）両者は共に代表性が
あるとして月次の平均は共通

＊仮定Bは無くても識別可能



上図）2人以上世帯の場合
「全国消費実態調査・年平均値推定ロードマップ案」より引用

【利点】

・家計調査単体より精度が向上　・全消の諸変数と相関算出可能

消費支出と食費についての情報

今回は以下の情報を利用

【家計調査】※家計簿の種類は２種類有、同時記入

①家計簿A（二人以上の世帯用）

Ⅰ 口座自動振替による支払

Ⅱ 口座への入金(給与・年金等)[世帯主,世帯主の配偶者,その他（間柄を記入）] ※世帯収入・消費は構成員の収入の合計

III 現金収入又は現金支出

IV クレジット・電子マネ ーなど現金以外による購入

②家計簿B（二人以上の世帯用）

・口座自動振替による支払、現金収入又は現金支出

・クレジットカード,掛買い,月賦による購入又は現物

消費支出と食費についての情報

今回は以下の情報を利用

【全国消費実態調査】※家計簿の種類は２種類有

①家計簿A（９月、１０月分）
②家計簿B（１１月分）

Ⅰ口座振替による自動支払

Ⅱ現物(現物支給,貰い物・もてなし,自家産,自分の店の商品)

III 現金収入又は現金支出

IV クレジットカード,掛買い,月賦,電子マネーによる購入

消費支出：各項目の合計

食費：各項目のうち食料に分類されているものの合計

いくつかの分析

（１）2調査の異質性を考慮しない場合

（２）2調査の異質性を考慮した場合

（３）年収を共変量として利用する場合（2調査の異質性なし）

（４）年収を共変量として利用する場合（2調査の異質性考慮）

＊2調査の異質性

⇒「調査モード」と「2調査の標本の標本誤差」両者が存在

（１）2調査の異質性を考慮しない場合

- 家計調査のデータと全国消費実態調査のデータ(9・10月)を融合したことの効果としての平均値と標準誤差の変化

（家計調査単独）

（全消と家計調査との融合）

12か月の各平均、その標準誤差

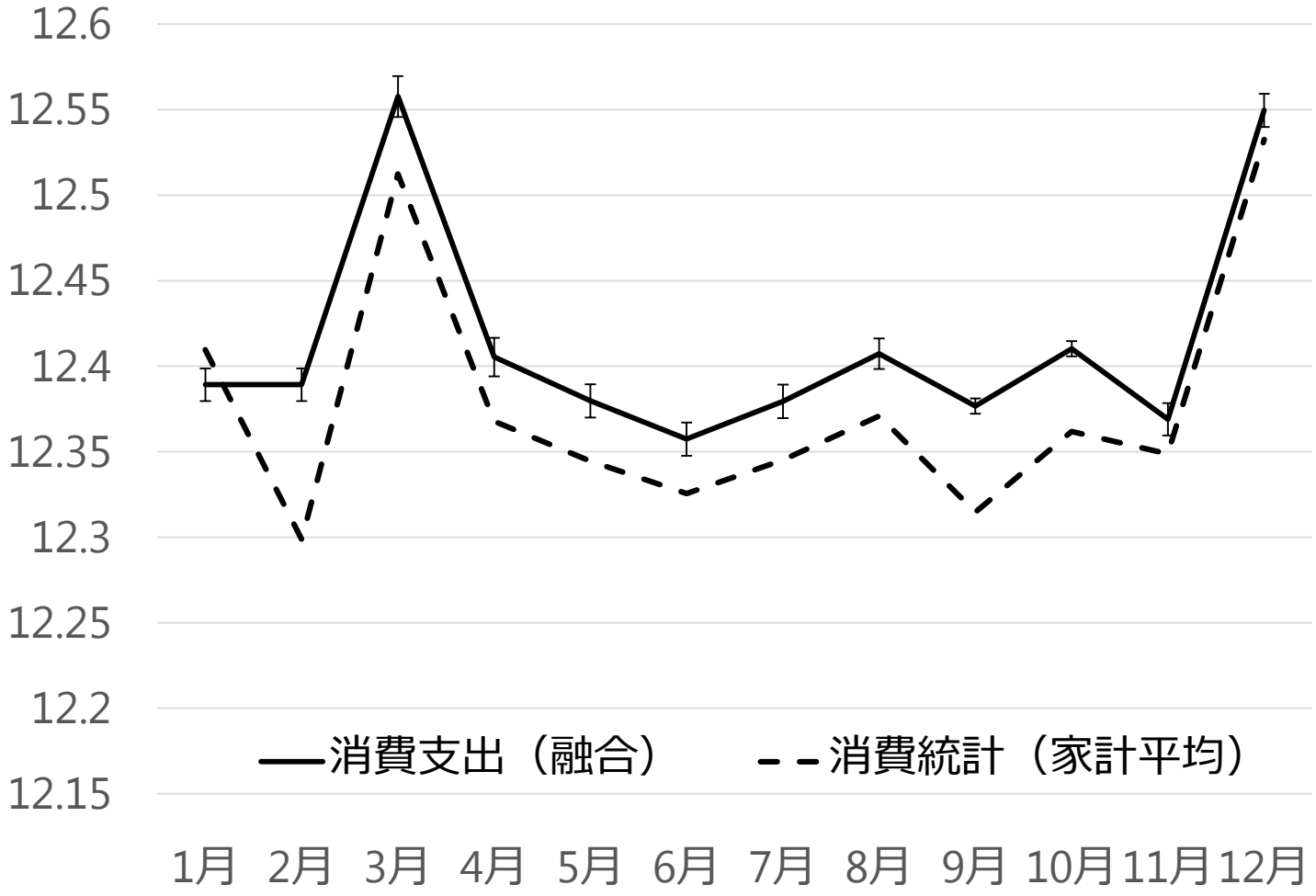
- ここでのモデルは家計調査も全国消費実態調査も同じ分布に従うと仮定

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{12} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{12} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1,12} \\ \vdots & \ddots & \vdots \\ \sigma_{1,12} & \cdots & \sigma_{12}^2 \end{bmatrix} \right)$$

解析結果（消費支出）

横軸月、縦軸対数値の平均、

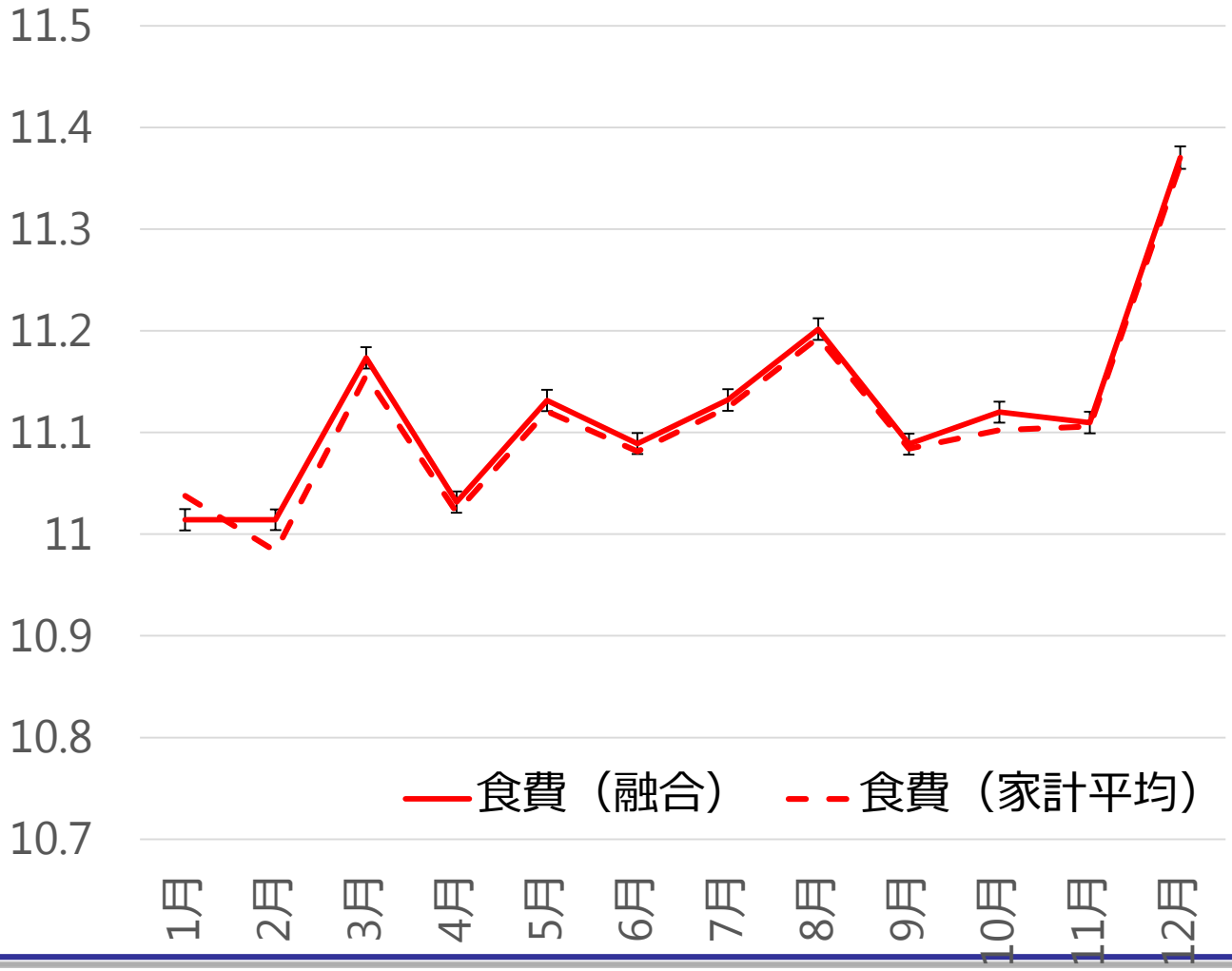
融合のみエラーバーで95%信頼区間



解析結果（食費）

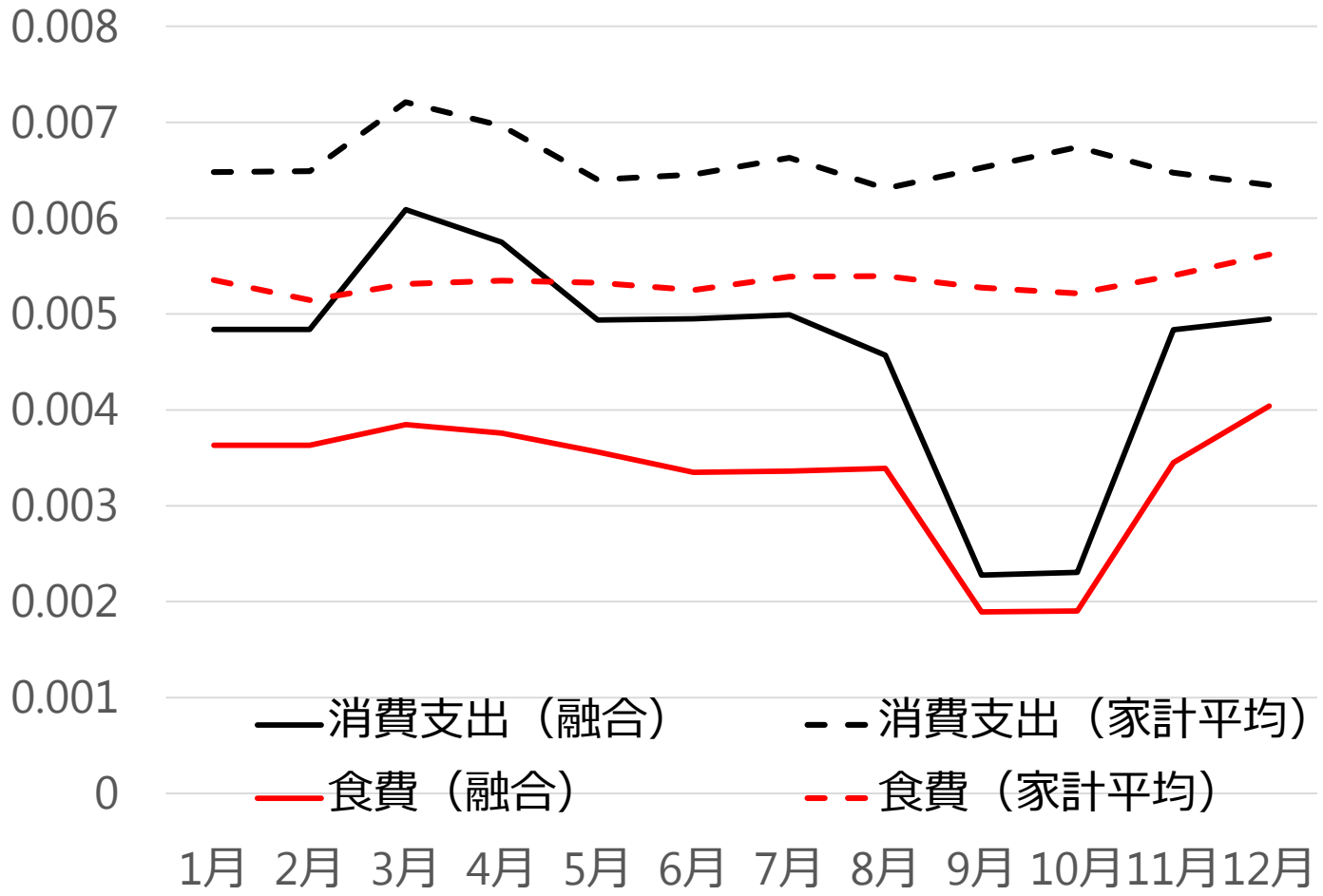
横軸月、縦軸対数値の平均、

融合のみエラーバーで95%信頼区間



解析結果（標準誤差：両者共通平均共分散）

標準誤差について



全国消費実態調査が加わる9・10月はもちろん、他の月も改善。

解析結果（比にしたもの）

標準誤差について

前ページグラフを比にしたもの

融合後の推定値の標準誤差 ÷ 融合前（家計調査の標本平均）の標準誤差

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
消費支出	0.747	0.746	0.845	0.826	0.772	0.767	0.753	0.724	0.349	0.342	0.747	0.78
食費	0.678	0.706	0.724	0.703	0.668	0.637	0.624	0.628	0.359	0.365	0.638	0.719

全国消費実態調査が加わる9・10月はもちろん、他の月も改善。

（２）2調査の異質性を考慮する場合

全国消費実態調査単体での公表と、家計調査と融合した結果を別々に月単位で公表する場合

「二調査で平均が異なるが共分散行列は共通」の仮定

＊どちらも共通としないと融合の効果はない

（家計調査）

$$\begin{bmatrix} y_{A1} \\ \vdots \\ y_{A12} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{A1} \\ \vdots \\ \mu_{A12} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1,12} \\ \vdots & \ddots & \vdots \\ \sigma_{1,12} & \cdots & \sigma_{12}^2 \end{bmatrix} \right)$$

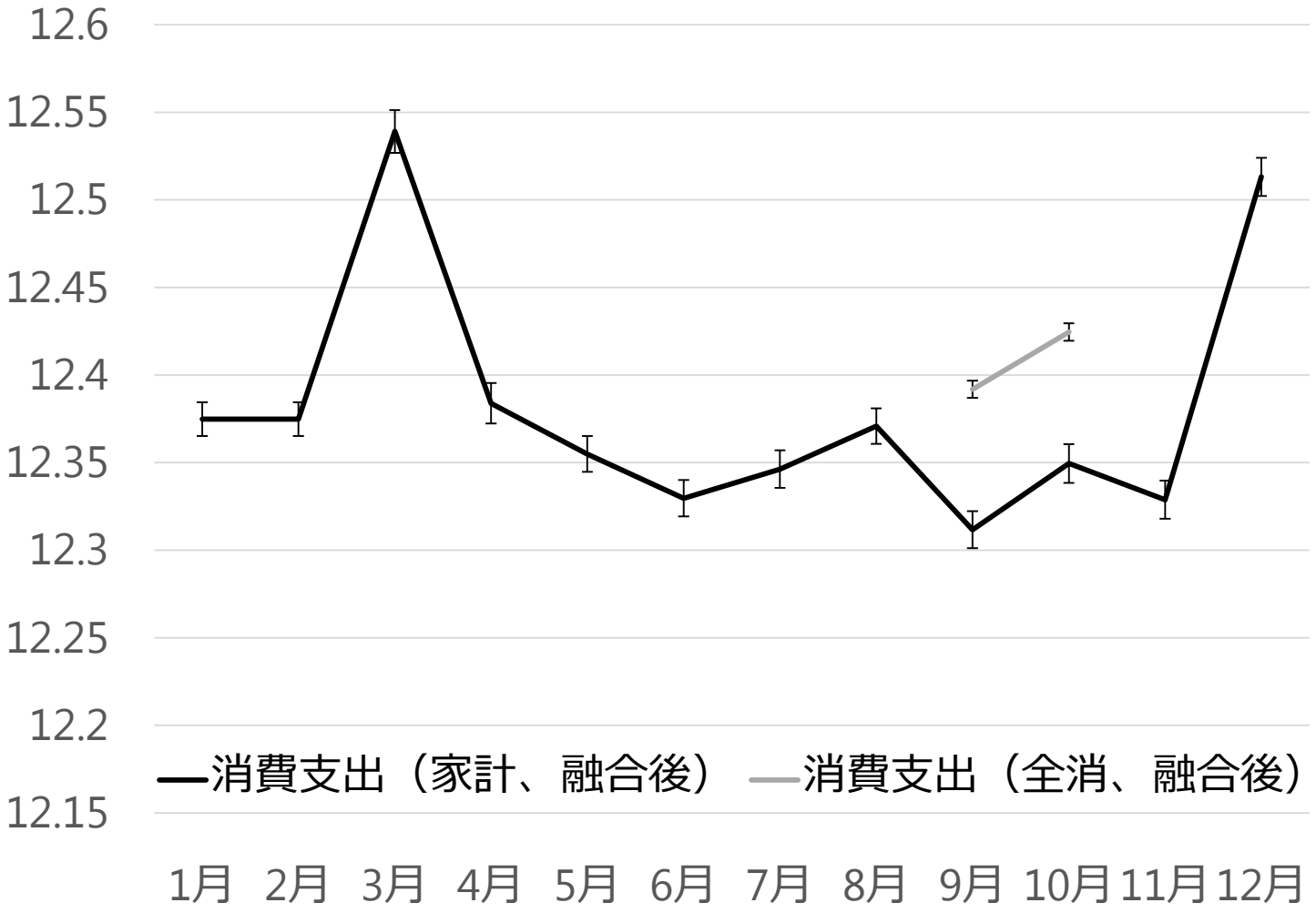
（全国消費実態調査）

$$\begin{bmatrix} y_{B1} \\ \vdots \\ y_{B12} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{B1} \\ \vdots \\ \mu_{B12} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1,12} \\ \vdots & \ddots & \vdots \\ \sigma_{1,12} & \cdots & \sigma_{12}^2 \end{bmatrix} \right)$$

解析結果（平均が2調査で違う場合：消費支出）

平均共通の仮定を置かなかった場合の推定値。両データ使用。

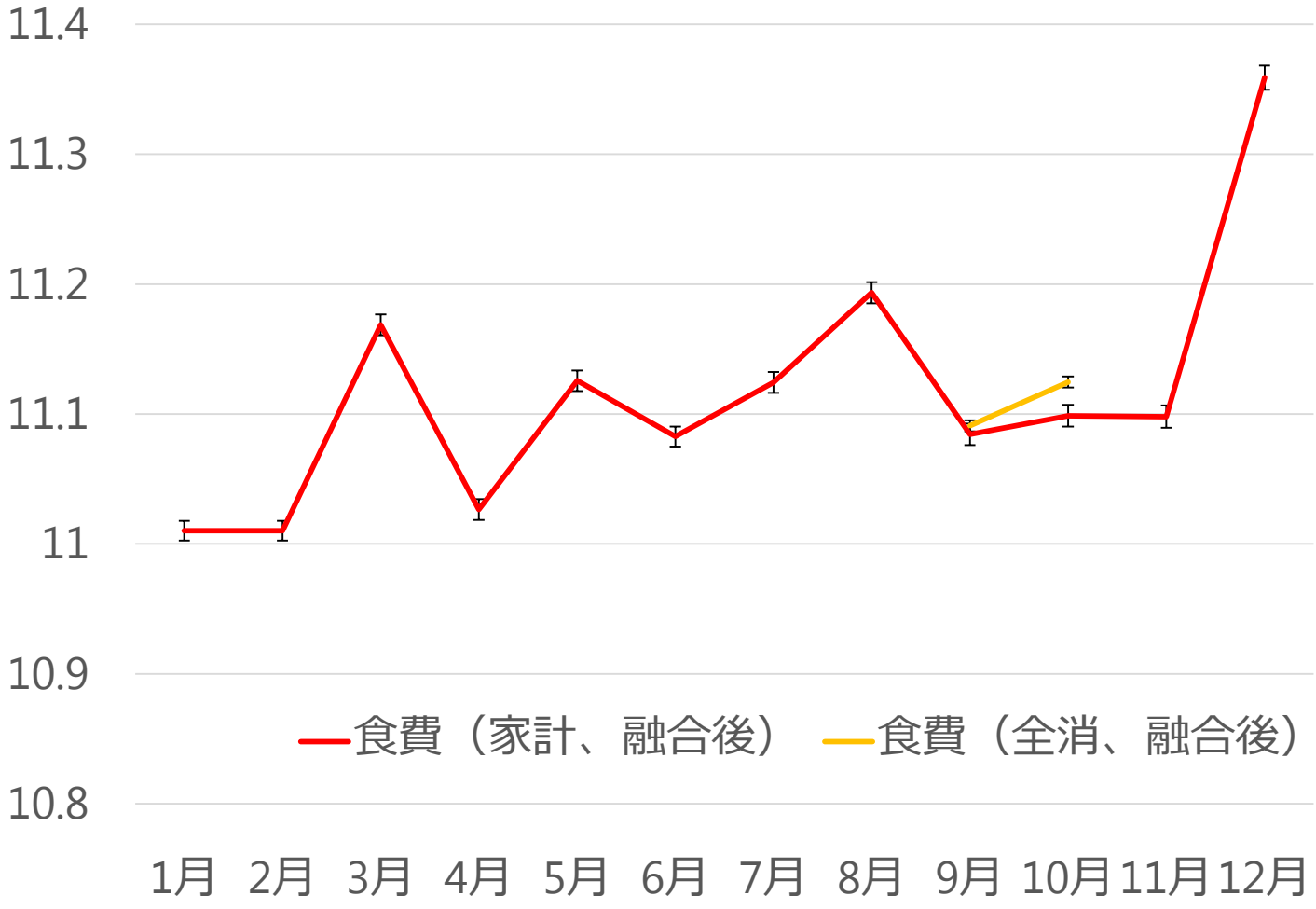
横軸月、縦軸対数値の平均、エラーバーで95%信頼区間



解析結果（平均が2調査で違う場合：食費）

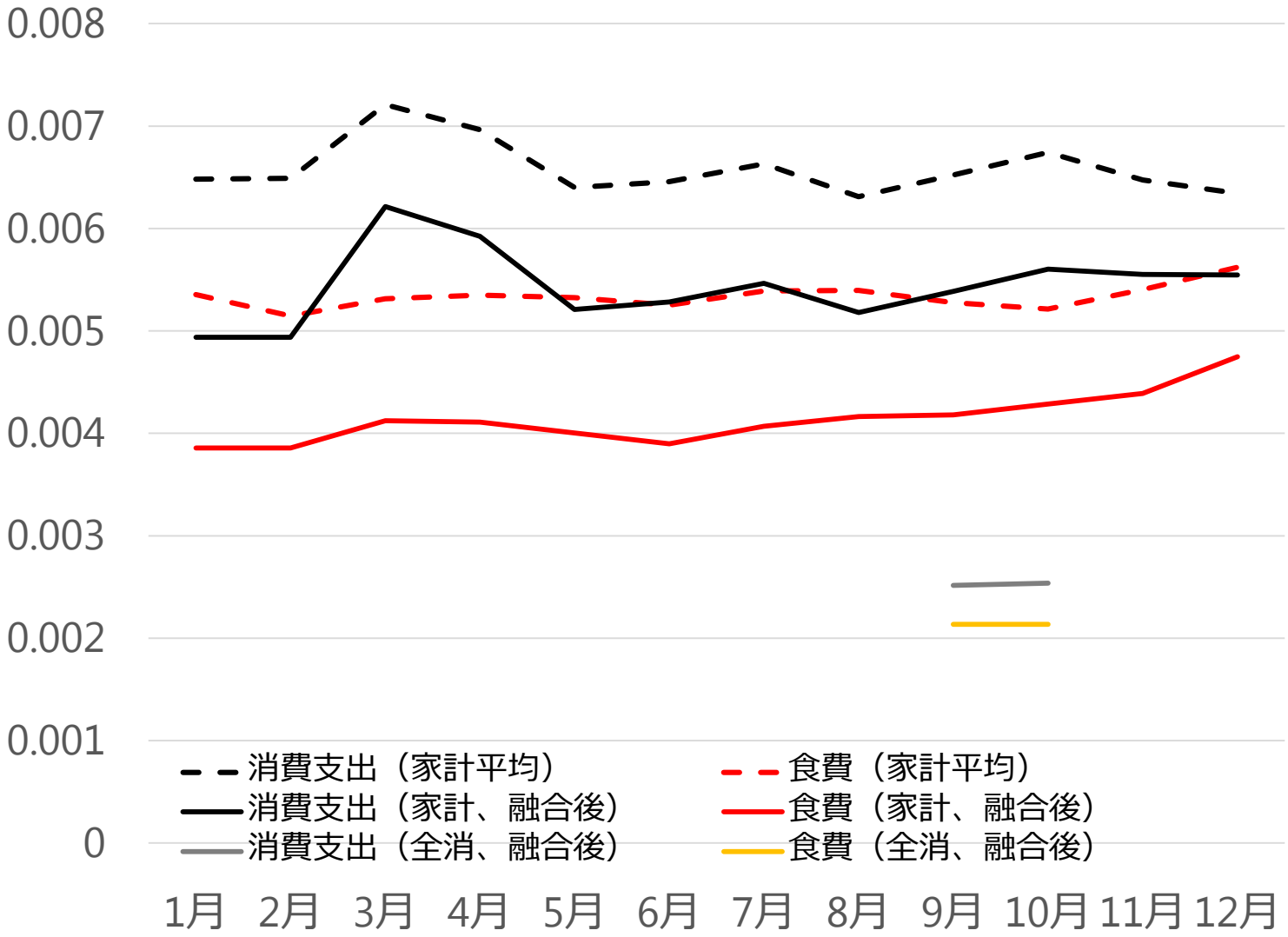
平均共通の仮定を置かなかった場合の推定値。両データ使用。

横軸月、縦軸対数値の平均、エラーバーで95%信頼区間



解析結果（平均が2調査で違う場合：標準誤差）

標準誤差について、平均共通の仮定を置かなかった場合



融合後の全国消費実態調査だけでなく、家計調査も改善。

解析結果（平均が2調査で違う場合：標準誤差）

標準誤差について、平均共通の仮定を置かなかった場合

前ページグラフを比にしたもの

融合後の推定値の標準誤差 ÷ 融合前（家計調査の標本平均）の標準誤差

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
消費支出（家計）	0.762	0.761	0.862	0.851	0.814	0.818	0.824	0.821	0.826	0.831	0.858	0.874
消費支出（全消）									0.386	0.376		
食費（家計）	0.72	0.749	0.776	0.768	0.752	0.742	0.755	0.771	0.792	0.822	0.812	0.844
食費（全消）									0.405	0.409		

融合後の全国消費実態調査だけでなく、家計調査も改善。

（３）年収等の共変量を利用する場合

年収の対数値で消費支出や食費の対数値を説明する回帰分析。

欠測を考慮。その際、誤差の共分散共通、回帰の傾きは共通と仮定。

切片に関して共通の場合と共通でない場合の両方を実施。

↑今までの平均共通の仮定を置いた場合と置かなかった場合に相当。

【両調査とも共通の場合】

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{12} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 + \beta_1 x \\ \vdots \\ \mu_{12} + \beta_{12} x \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1,12} \\ \vdots & \ddots & \vdots \\ \sigma_{1,12} & \cdots & \sigma_{12}^2 \end{bmatrix} \right)$$

年収情報

【利用した情報】今回年収については以下の情報を利用

（家計調査）

(1)勤め先年間収入

(2)営業年間利益

(3)内職年間収入

(4)公的年金・恩給

(5)農林漁業収入

(6)その他の年間収入

(7)現物消費の見積額

※各項目の合計を収入として利用

年収情報

【利用した情報】今回年収については以下の情報を利用

（全国消費実態調査）

年収・貯蓄等調査票「Ⅰ 年間収入について」

(1)勤め先年間収入

(2)農林漁業収入

(3)農林漁業以外の事業収入

(4)内職などの年間収入

(5)家賃・地代の年間収入

(6)公的年金・恩給

(7)企業年金・個人年金

(8)利子・配当金

(9)親族などからの仕送り金

(10)その他の年間収入

(11)現物消費の年間見積額

※各項目の合計を収入として利用

年収情報

【利用した情報】今回年収については以下の情報を利用

（全国消費実態調査）

世帯収入は、世帯主、世帯主の配偶者、その他世帯構成員の収入を合計したものを使用。

【両者の群間差】2調査間で年収については以下のような違い

	家計調査	全国消費実態調査
平均	597.4113	568.5028
標準偏差	376.52	361.46

単位：万円

＊今回はデータ整形の都合上家計調査については中間欠測は除外

標本変動の範囲か？項目の違いか？毎回答えて頂いているご家庭なので？

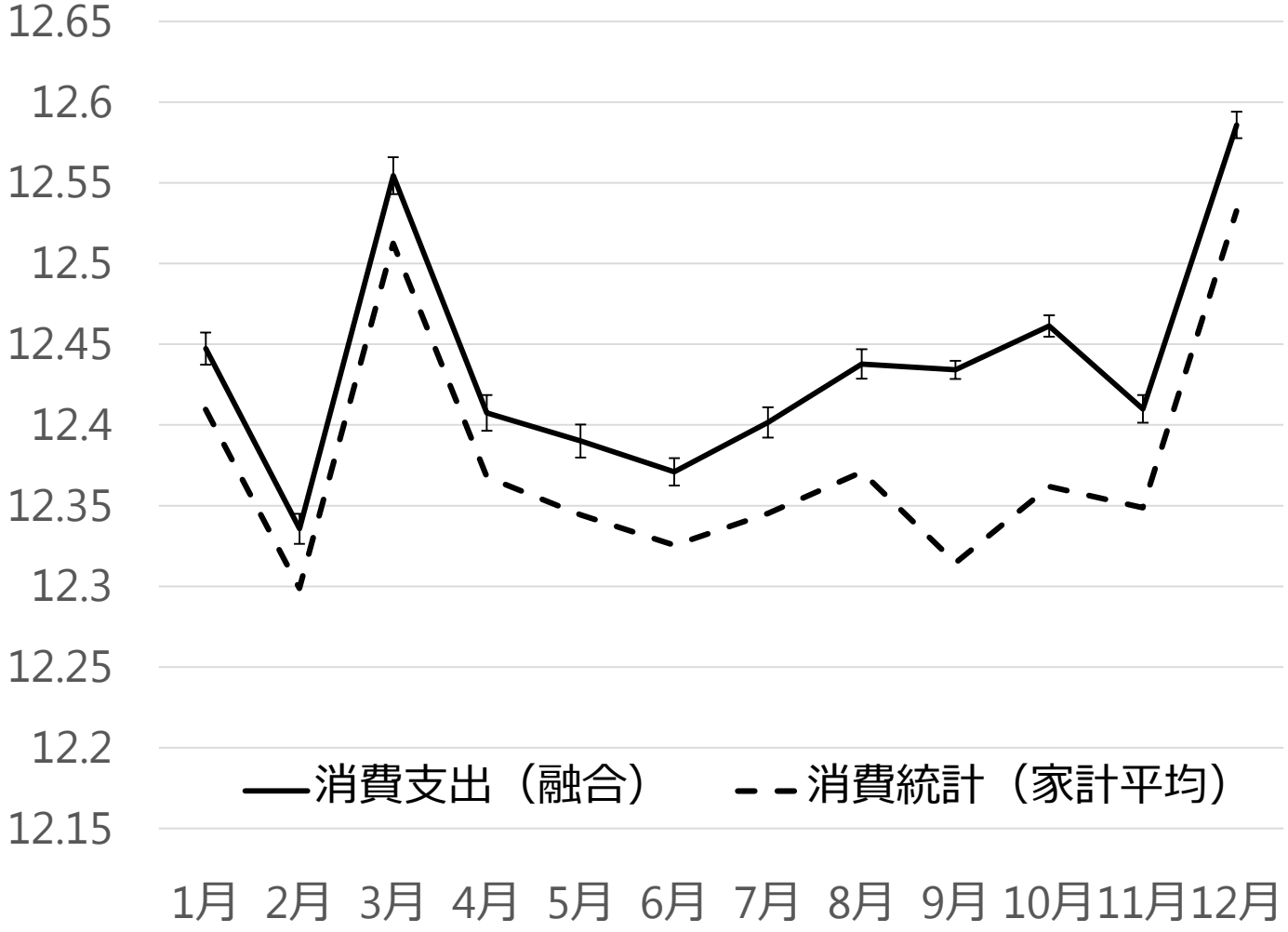
＊消費や食費で家計調査が低いのは過小記入バイアスや倹約化か？

(過小記入バイアスDeaton&Irish,1984;牧,2007; 倹約化・調査疲れStephens&Unayama,2011)

解析結果（収入で回帰：消費支出、切片共通）

横軸月、縦軸対数値の平均、

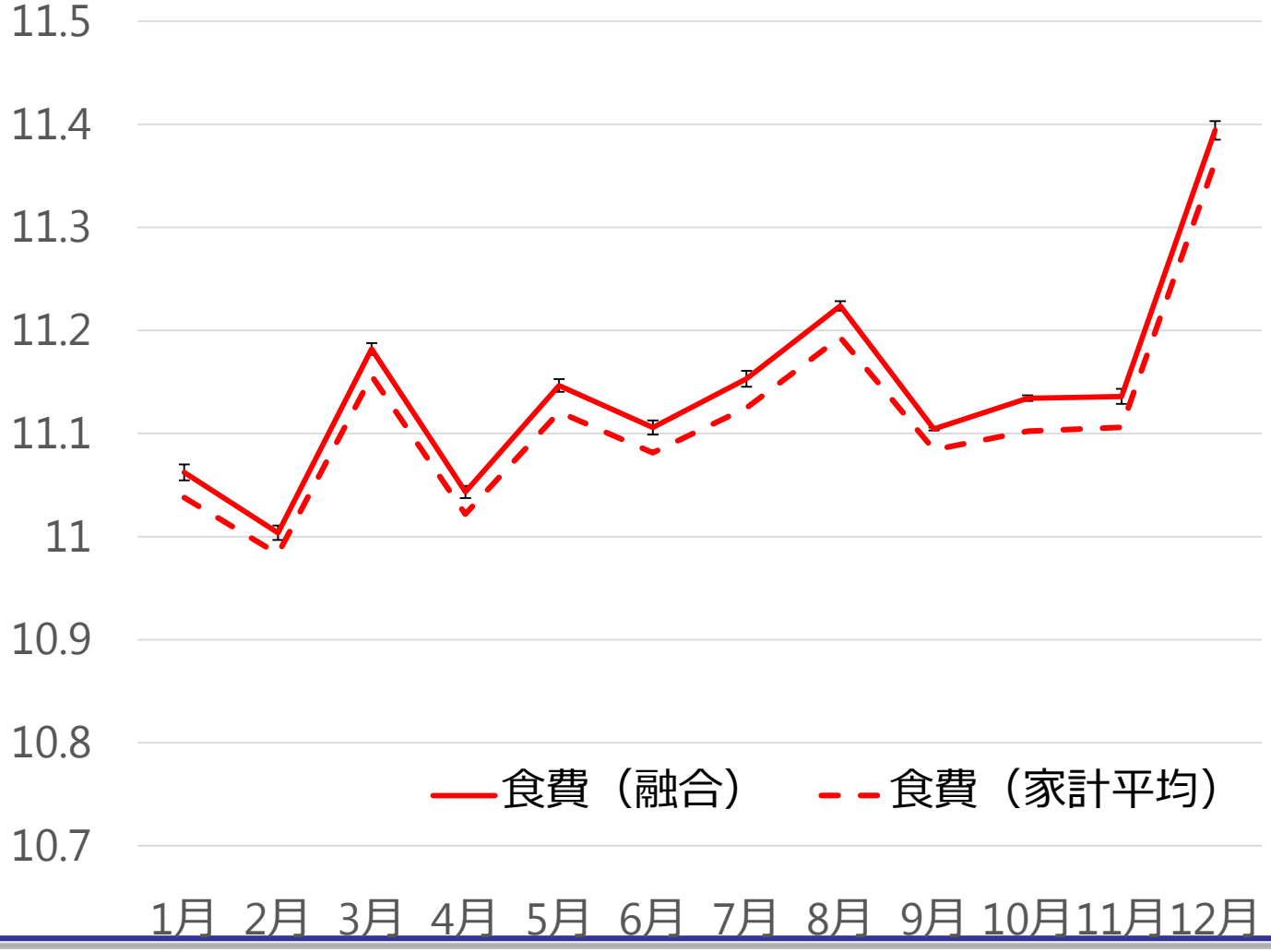
融合のみエラーバーで95%信頼区間



解析結果（収入で回帰：食費、切片共通）

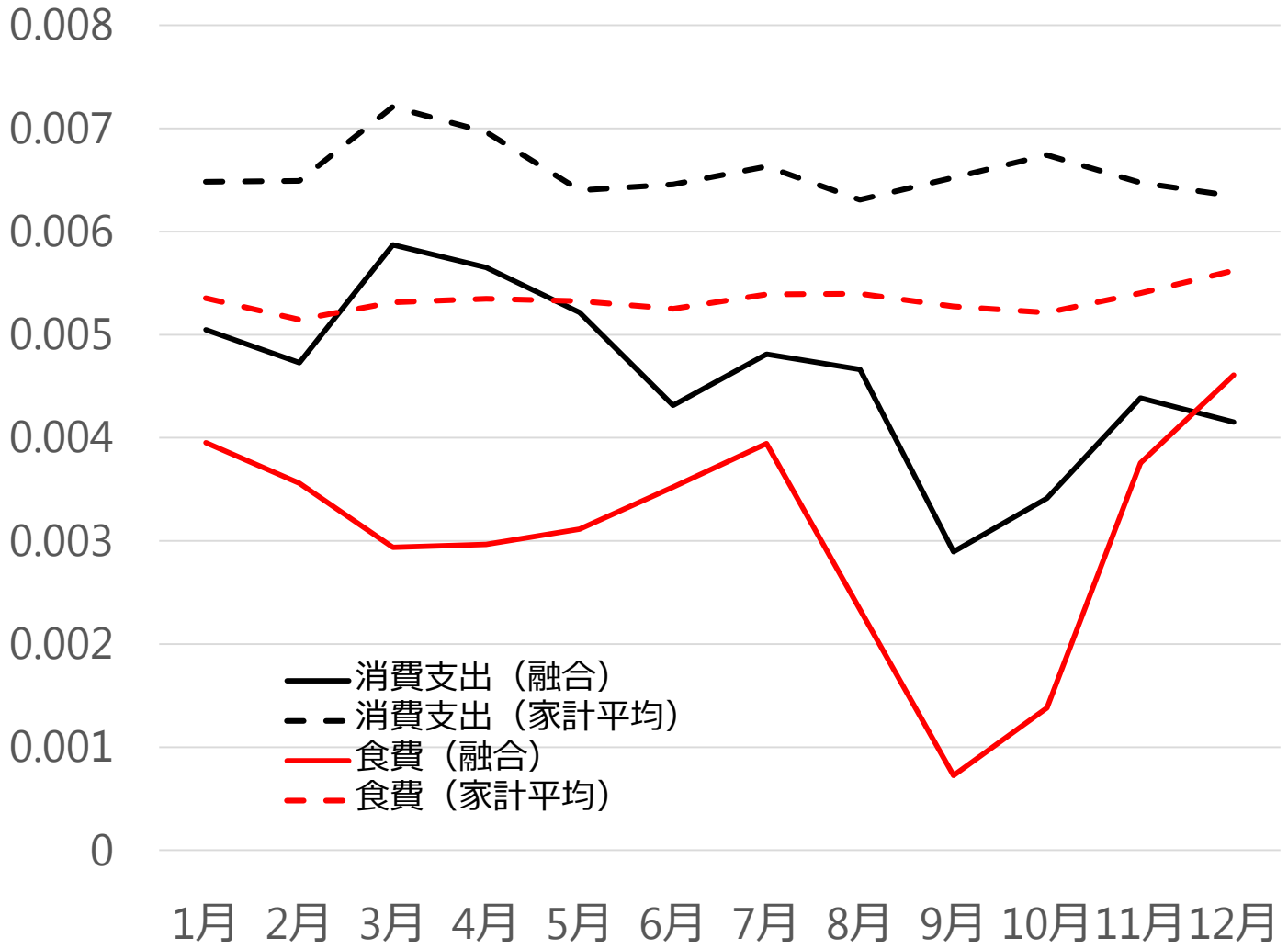
横軸月、縦軸対数値の平均、

融合のみエラーバーで95%信頼区間



解析結果（収入で回帰：標準誤差、切片共通）

標準誤差について



全国消費実態調査が加わる9・10月はもちろん、他の月も改善。

解析結果（収入で回帰：比にしたもの、切片共通）

標準誤差について

前ページグラフを比にしたもの

融合後の推定値の標準誤差 ÷ 融合前（家計調査の標本平均）の標準誤差

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
消費支出	0.779	0.729	0.814	0.811	0.815	0.668	0.725	0.739	0.444	0.506	0.678	0.655
食費	0.738	0.692	0.553	0.555	0.585	0.671	0.732	0.432	0.137	0.265	0.695	0.82

全国消費実態調査が加わる9・10月はもちろん、他の月も改善。

(4)調査モードを考慮しかつ2調査が異なる仮定

- 家計調査は以下の分布に従うと仮定

$$\begin{bmatrix} y_{A1} \\ \vdots \\ y_{A12} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_A + \beta_1 x \\ \vdots \\ \mu_A + \beta_{12} x \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1,12} \\ \vdots & \ddots & \vdots \\ \sigma_{1,12} & \cdots & \sigma_{12}^2 \end{bmatrix} \right)$$

但しxは収入

- 全国消費実態調査は以下の分布に従うと仮定

$$\begin{bmatrix} y_{B1} \\ \vdots \\ y_{B12} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_B + \beta_1 x \\ \vdots \\ \mu_B + \beta_{12} x \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1,12} \\ \vdots & \ddots & \vdots \\ \sigma_{1,12} & \cdots & \sigma_{12}^2 \end{bmatrix} \right)$$

実際は y_{B9} y_{B10} しか観測されないが、観測されている要素から

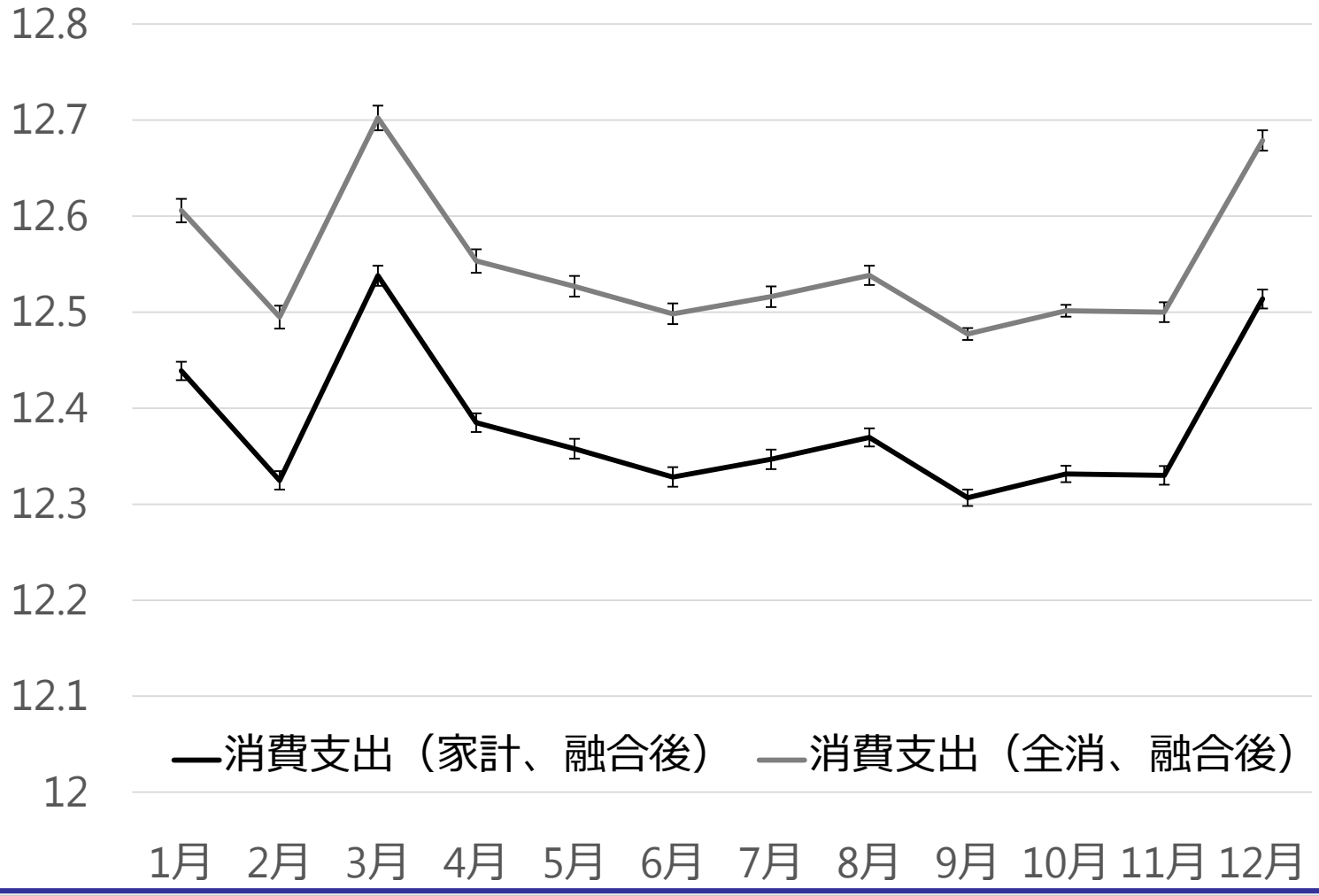
$$E[y_{Bt}] = \mu_B + \beta_t E(x)$$ の推定が可能

＊切片が調査モードの差　xの分布差が両調査の標本誤差に対応

解析結果（収入で回帰：消費支出、切片が2標本で違う場合）

切片共通の仮定を置かなかった場合の推定値。両データ使用。

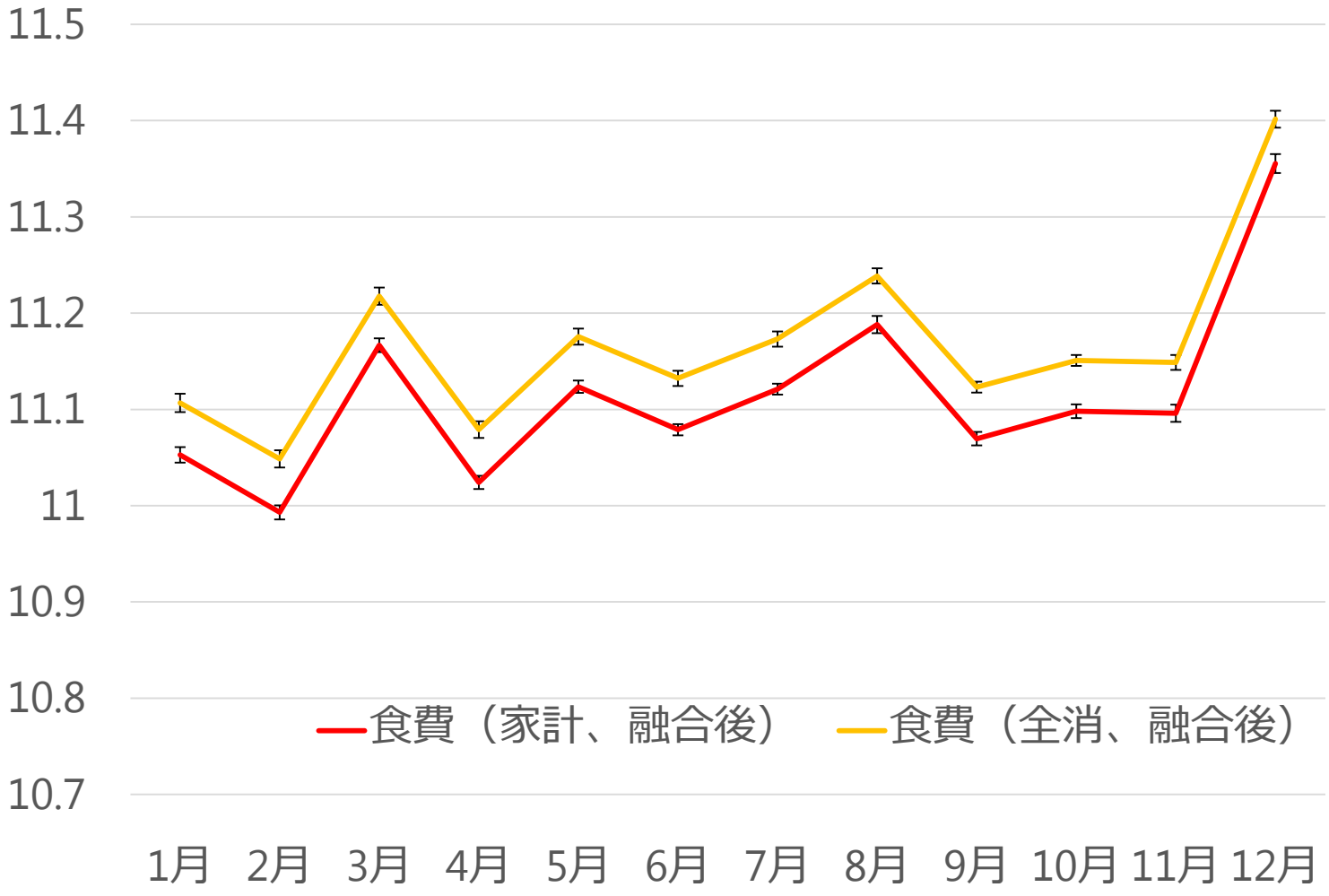
横軸月、縦軸対数値の平均、エラーバーで95%信頼区間



解析結果（収入で回帰：食費、切片が2標本で違う場合）

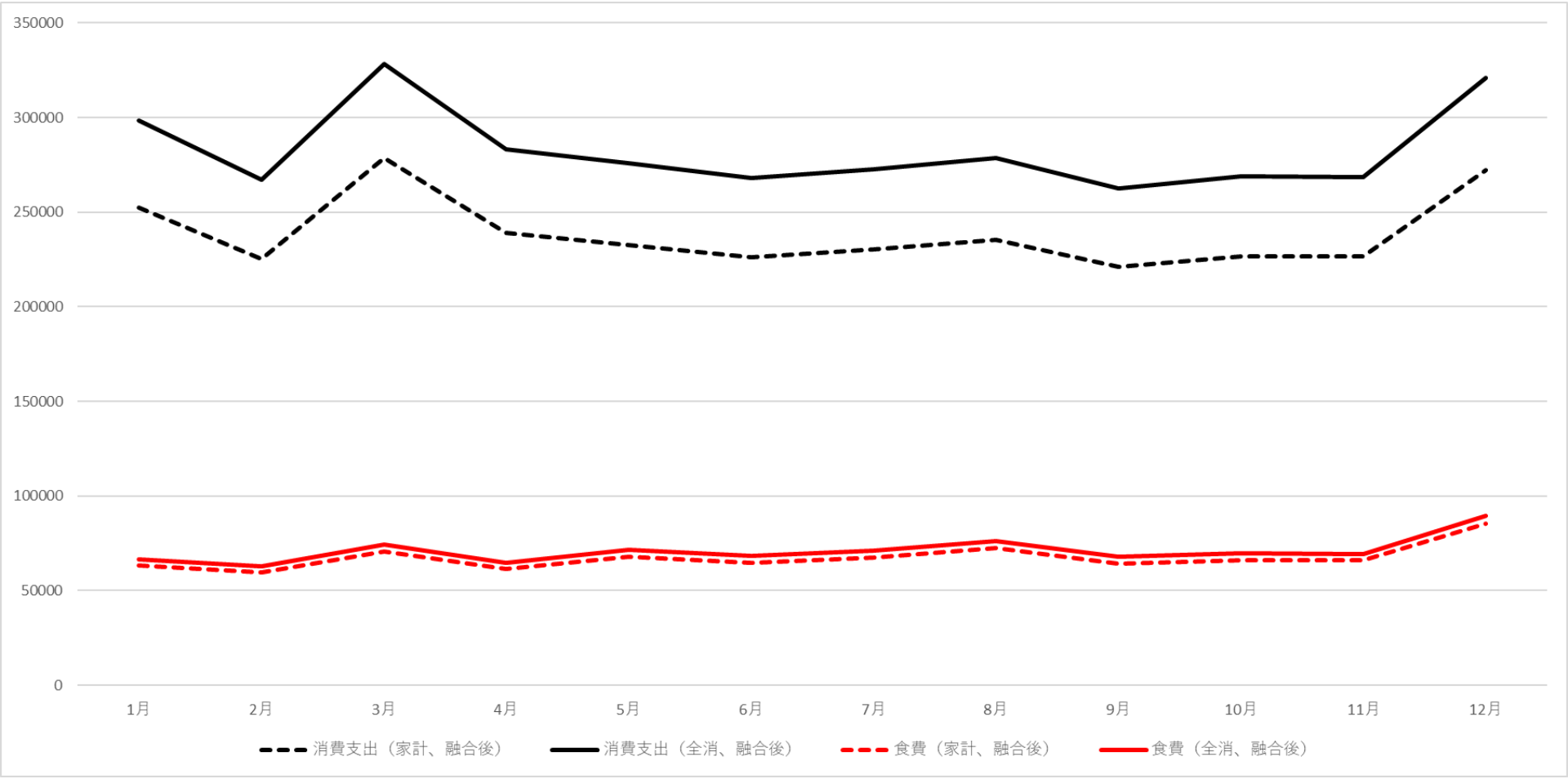
切片共通の仮定を置かなかった場合の推定値。両データ使用。

横軸月、縦軸対数値の平均、エラーバーで95%信頼区間



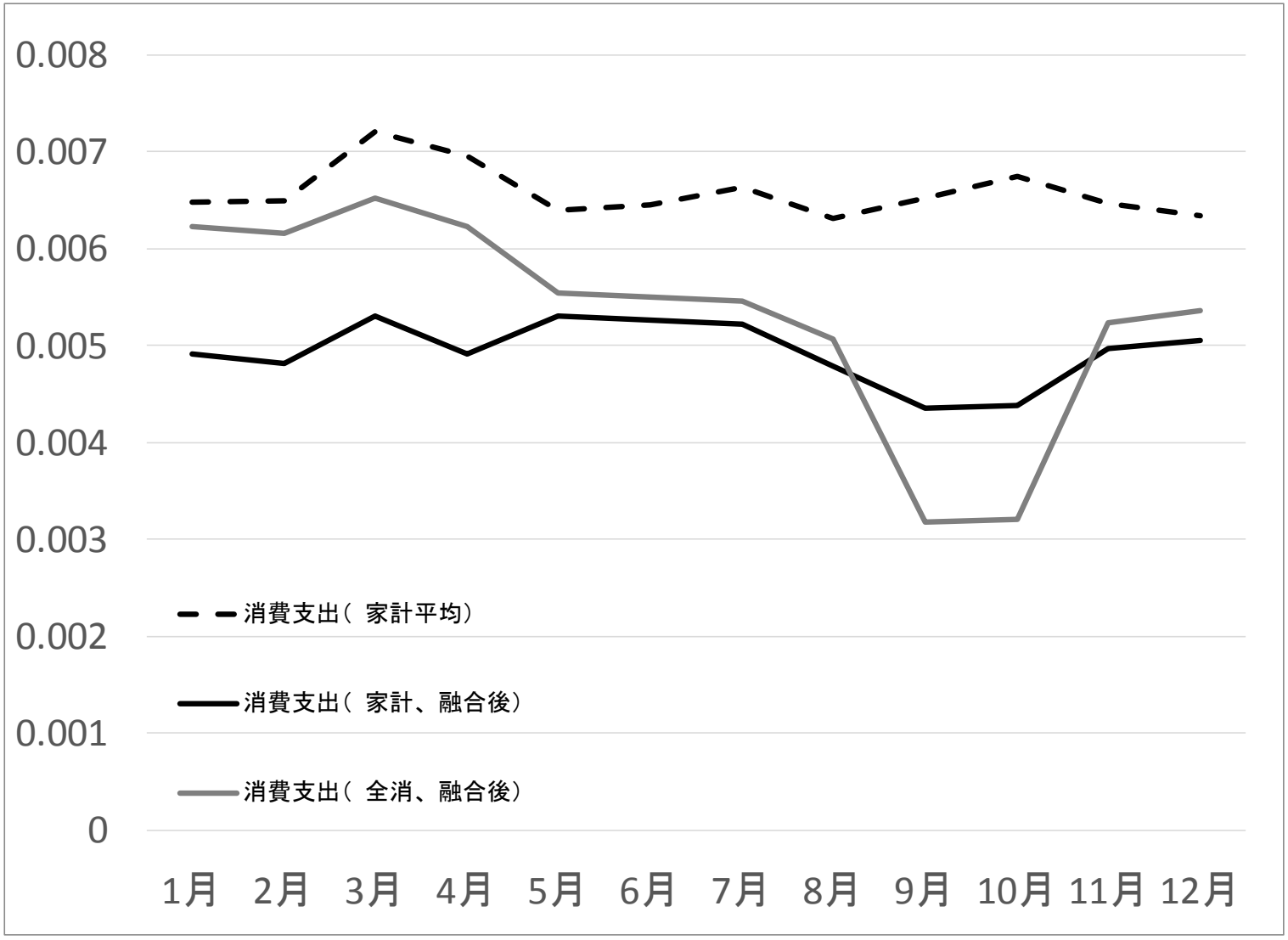
解析結果（収入で回帰：切片が2標本で違う場合）

対数を実数値に戻した場合



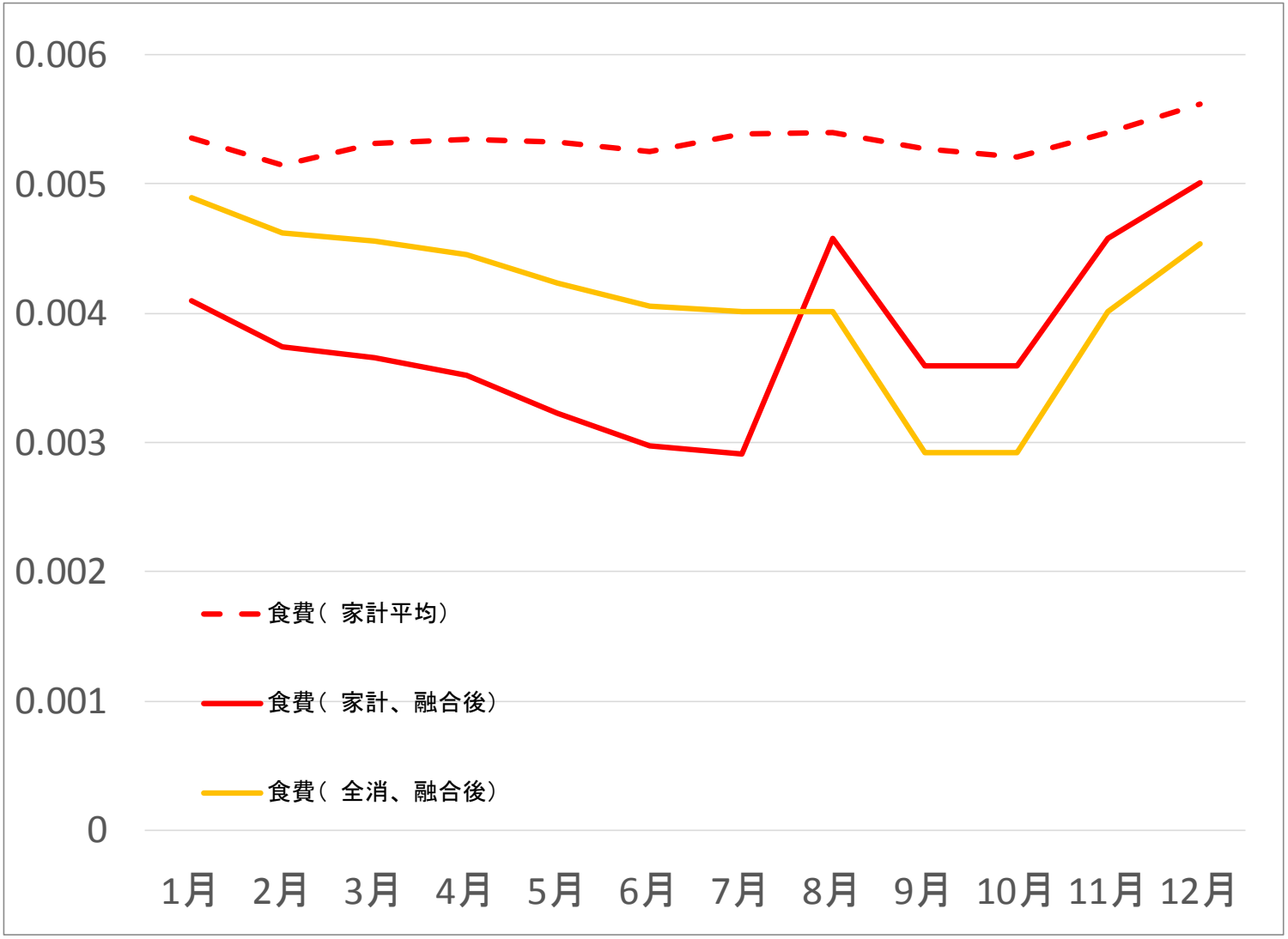
解析結果（収入で回帰：標準誤差、切片が2標本で違う場合）

消費支出について



解析結果（収入で回帰：標準誤差、切片が2標本で違う場合）

食費について



解析結果（収入で回帰：標準誤差、切片が2標本で違う場合）

標準誤差について、切片共通の仮定を置かなかった場合

前ページグラフを比にしたもの

融合後の推定値の標準誤差 ÷ 融合前（家計調査の標本平均）の標準誤差

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
消費支出（ 家計）	0.758	0.743	0.735	0.705	0.829	0.815	0.788	0.759	0.667	0.649	0.767	0.796
消費支出（ 全消）	0.96	0.949	0.904	0.894	0.865	0.852	0.824	0.804	0.487	0.476	0.808	0.846
食費（ 家計）	0.765	0.727	0.689	0.658	0.606	0.566	0.539	0.848	0.682	0.69	0.847	0.891
食費（ 全消）	0.915	0.898	0.858	0.833	0.796	0.773	0.744	0.743	0.553	0.559	0.742	0.807

融合後の全国消費実態調査だけでなく、家計調査も改善。

年平均の値について

(A)全消での9・10月の平均（月）　(B)家計調査での9・10月の平均

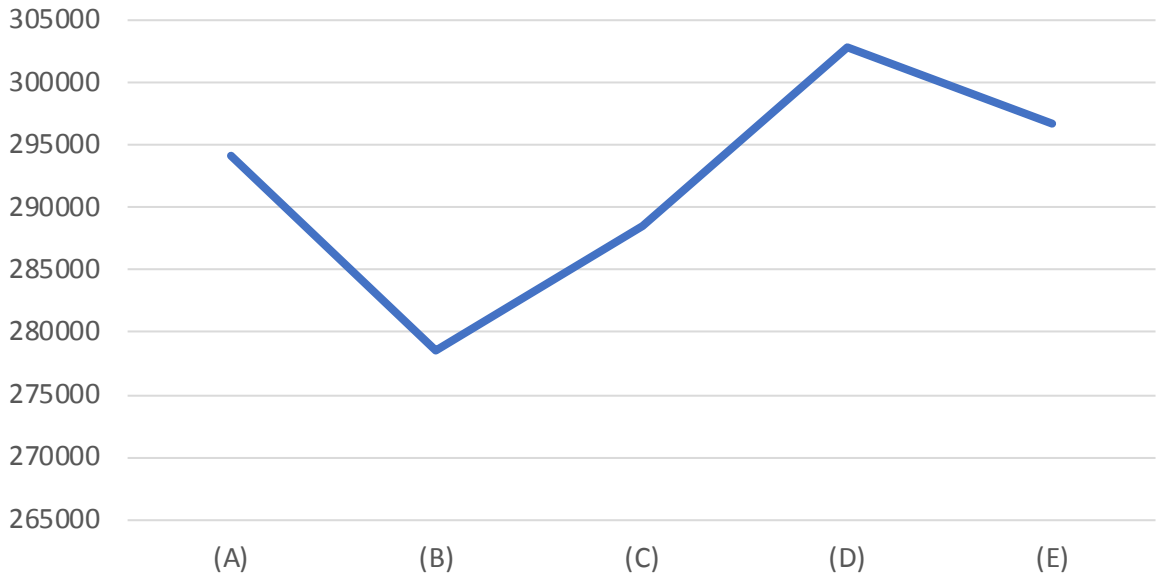
(C)家計調査での年平均

(D)上記を融合した結果から“全国消費実態調査”の調査対象者が年間で答えたと仮定した場合の推計値　⇒方法(1)と方法(2)の融合

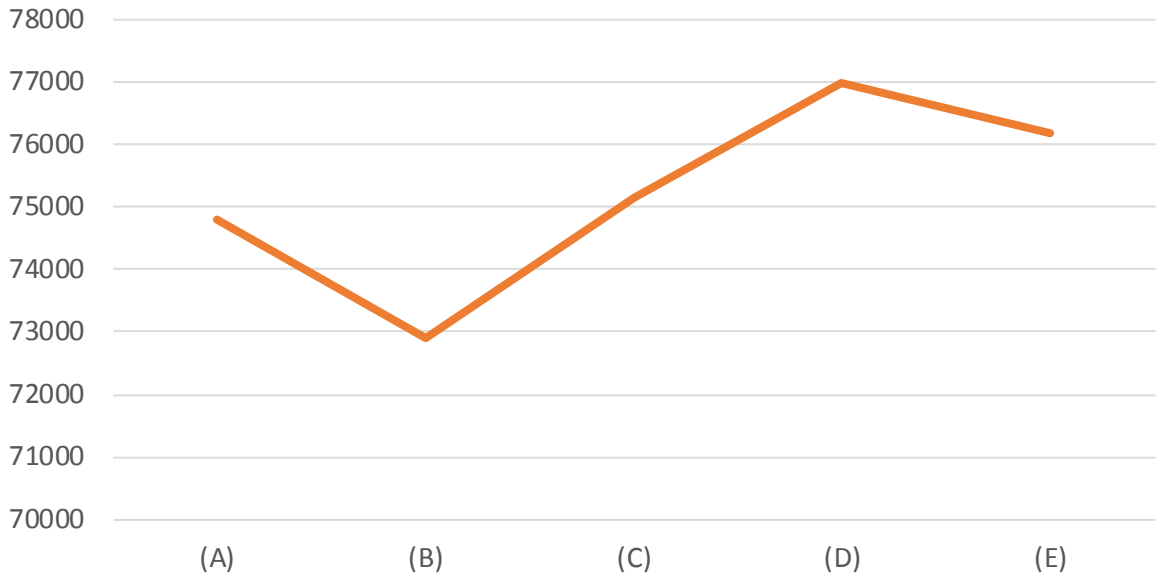
(E)上記を融合：収入を共変量とし調査モードが両者が異なることを仮定した場合　⇒方法(4)の場合

方法	（A）	（B）	（C）	（D）	（E）
消費支出	294090	278534	288536	302886	296716
食費	74803	72900	75168	76979	76196

消費支出


305000
300000
295000
290000
285000
280000
275000
270000
265000
(A)
(B)
(C)
(D)
(E)

食費


78000
77000
76000
75000
74000
73000
72000
71000
70000
(A)
(B)
(C)
(D)
(E)

まとめと今後の予定

【まとめ】

・全国消費実態調査は9・10月の実施だが消費の季節変動は大

⇒単にこれを6倍して年間の消費額や食費とすると生活保護世帯への支給額の算定など様々な政策的決定を誤る危険性

・今回は集計時系列上での季節調整ではなくミクロレベルでの調整

⇒当初の目的を実現するだけではなく融合した家計調査にもメリット

【今後実施すべき項目】

・多変量t分布の実装

・単身世帯でのプログラム作成と2人以上との統合

・代入の可能性の検討(多重代入でないと分散過小評価)

⇒4月中に程度めどをつけたい

最後に

実際のプログラム作成や分析は以下の学生が実施しました。

清水祐弥君（慶應義塾大学大学院経済学研究科修士1年）

慶野有輝君（慶應義塾大学経済学部4年）

（参考文献）

Kim. JK. and Shao, J. (2014) Statistical Methods for Handling Incomplete Data. Chapman & Hall.

Little R.A.J and Rubin, D.B. (2002) Statistical Analysis with Missing Data, 2nd ed. Wiley.

Ridder, G. & Moffitt, R. (2007) The Econometrics of Data Combination, in J. J. Heckman & E. E. Leamer, eds, `Handbook of Econometrics', Vol. 6, Elsevier, chapter 75, p.5469-5547.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data, Chapman & Hall.

星野(2009)「調査観察データの統計科学」岩波書店

高井・星野・野間(2016)「欠測データの統計科学」岩波書店

資料編

欠測データ分析での記号の定義

y　：関心のある変数（ベクトル）

y_{obs} ： y　のうち観測されている部分

y_{mis} ：観測されていない部分

つまり　$y = (y_{obs}^t, y_{mis}^t)^t$

ここで　m　を 欠測インディケーターとする

例）N人のデータ、Unit nonresponseがある場合

$$m = (m_1, \cdots, m_N)^t$$

最尤法と欠測

関心のある変数と欠測インディケータの同時分布の
モデリングの代表的なものとして

選択モデル(Selection model)

$$p(y,m\mid\theta,\phi)=p(y\mid\theta)\,p(m\mid y,\phi)$$

「関心のある変数の周辺分布」に関心がある場合

パターン混合モデル(Pattern mixture model)

$$p(y,m\mid\xi,\omega)=p(y\mid m,\xi)\,p(m\mid\omega)$$

例）

$$y=\begin{bmatrix}y_1\\ y_2\end{bmatrix}$$

$m=1$の時
y_2 が欠測

θ はyの平均と分散共分散行列

ξ は「m=1の群」での平均と共分散行列と「m=0の群」での平均と共分散行列

選択モデルでの尤度

3つの尤度の違いの理解

完全データの尤度(Complete data likelihood)

$$p(y,m\mid\theta,\phi)=p(y\mid\theta)\,p(m\mid y,\phi)$$

＊全データが観測できる場合の尤度

完全尤度(Full likelihood)

$$p(y_{obs},m\mid\theta,\phi)=\int p(y\mid\theta)\,p(m\mid y,\phi)dy_{mis}$$

＊観測された関心のある変数と欠測インディケータの尤度

観測データの尤度(Observed likelihood)直接尤度(Direct～)

$$p(y_{obs}\mid\theta)$$

＊観測された関心のある変数だけの周辺尤度

ランダムな欠測

「選択モデル」において“ランダムな欠測”
Missing at Random(Rubin,1976)とは

$$p(m \mid y, \phi) = p(m \mid y_{obs}, \phi)$$

欠測するかどうかが欠測値に依存しない

この場合「完全尤度」は

$$p(y_{obs}, m \mid \theta, \phi) = \int p(y \mid \theta)\, p(m \mid y_{obs}, \phi) dy_{mis}$$
$$= \mathrm{p}(\mathrm{y}_{\mathrm{obs}} \mid \theta)\ p(m \mid y_{obs}, \phi)$$

したがって「観測データの尤度」$p(y_{obs} \mid \theta)$の最大化
によってθ の最尤推定量が得られる

⇒「欠測インディケータ」mのモデルは考えなくてよい

今回はランダムな欠測を仮定する（計画的欠測のため）

＊家計調査と全消の対象者と過小記載バイアスは別に考慮?

今回のモデルでの分析：ランダムな欠測での解析

確率変数ベクトルyがp次元　$y=(y_1,\cdots,y_p)$ とする

欠測パターンを最大Q通りとする(mの取る値がQ通り)

例）p=3、あらゆる組み合わせ可⇒Q=8(通り)

第k欠測パターンについて

$$y_{obs}=y^{(k)},\quad y_{mis}=y^{(-k)}$$

とすると、yをソーティングすれば　$y=(y^{(k)},y^{(-k)})$

さらにインディケータ変数として

$$I^{(k)}=\begin{cases}1 & (\text{第}k\text{パターンの時})\\ 0 & (\text{それ以外})\end{cases}$$

このとき独立に同一な分布に従う場合には

欠測インディケータベクトルmをy同様に分割する

$$m = (m^{(k)}, m^{(-k)})$$

【完全尤度】

$$\prod_{i=1} p(y_{obsi}, m_i \mid \theta, \phi) = \prod_{i=1} \prod_{k=1} p(y_i^{(k)}, m^{(k)} \mid \theta, \phi)^{I_i^{(k)}}$$

対数尤度は

$$\sum_{i=1}^{N} \sum_{k=1}^{Q} I_i^{(k)} \log p(y_i^{(k)}, m^{(k)} \mid \theta, \phi)$$

【観測データの尤度】

$$\prod_{i=1} p(y_{obsi} \mid \theta) = \prod_{i=1} \prod_{k=1} p(y_i^{(k)} \mid \theta)^{I_i^{(k)}}$$

対数尤度は

$$\sum_{i=1}^{N} \sum_{k=1}^{Q} I_i^{(k)} \log p(y_i^{(k)} \mid \theta)$$

観測データの尤度を用いたEMアルゴリズム

Expectationステップ

$$Q(\theta|\theta^{(l)}) = \sum_{i=1}^{N}\sum_{k=1}^{Q} I_i^{(k)} \int \log p\left(y_i \ |\theta\right) p\left(y_i^{(-k)}\middle|y_i^{(k)}, \theta^{(l)}\right) dy_i^{(-k)}$$

Maximaizationステップ

$$\theta^{(l+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(l)})$$

これを繰り返すことで収束した値を最尤推定値とみなすことができる