資料１－２　別紙１

家計調査の改善に伴う断層の推定方法及び試算結果

１．分析の概要

(1) 推定方法の案

・以下の３種類の方法について、断層を生じさせたデータを基に検証を行う。

①重回帰モデルに基づく方法

②差分の差（difference in differences）推定に基づく方法

③傾向スコア・IPW（Inverse Probability Weighting）推定に基づく方法

(2) 検証方法

・調査モードの違いで断層が生じている状況を再現したデータで検証を実施。

←新調査票の対象地域の世帯に似ている家計消費状況調査の世帯について統計的にマッチングを行い、分析用データを作成する。

２．データ

(1) 分析に使用したデータの概要：

・家計調査：９月分結果（8,424世帯：新調査票対象は4,175世帯）

及び平成29年10月分結果（8,413世帯：新調査票対象は4,183世帯）

・家計消費状況調査：９月分結果（22,020世帯）及び平成29年10月分結果（22,205世帯）

・使用した調査事項：家計消費状況調査の家計調査補完品目及び対応する家計調査の支出項目

(2) 家計調査と家計消費状況調査の結合

・家計消費状況調査の特定50品目への消費支出額と、対応する家計調査の消費支出額との差を、家計調査の世帯の消費支出に加算。

・世帯間の「類似」度については、共通の調査事項に基づく「Gower距離」で計測。

←計算には R のパッケージ StatMatch の NND.hotdeck 関数を使用。

Gower 距離：$D_{ij} = \left(\sum_{k=1}^{K} D_{ijk}\right)/K$

- X_{ik}, X_{jk} が連続変数の場合：

  $D_{ijk} = \left|X_{ik} - X_{jk}\right|/R_k$（$R_k$：$k$ 番目の変数のレンジ（最大値－最小値））
- X_{ik}, X_{jk} がカテゴリ変数の場合：

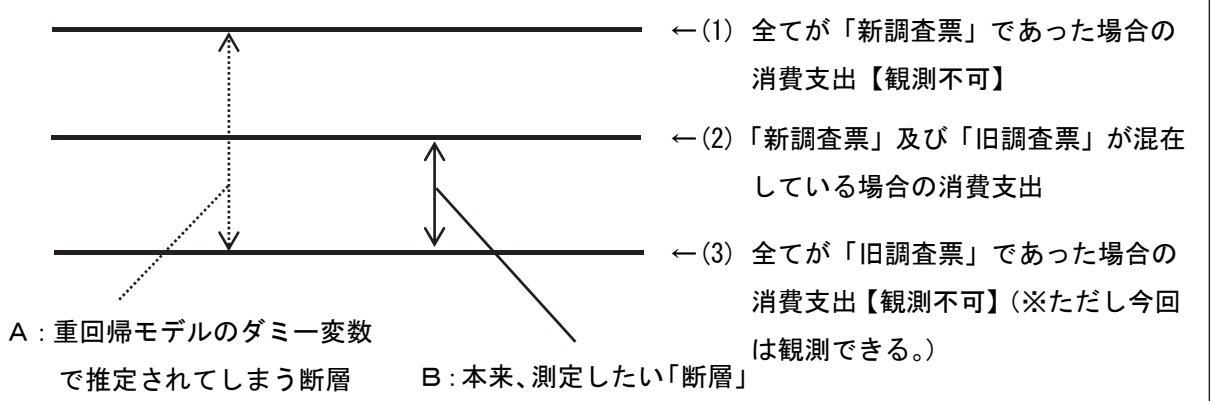
  $D_{ijk} = 0\ (X_{ik} = X_{jk}),\ = 1\ (X_{ik} \neq X_{jk})$

(3) 断層の考え方について

・重回帰モデルのダミー変数など、「差」を直接推定しにいく方法では、下の図１の「Ａ」を推定してしまい、断層の過大評価になる。

・本来推定したいのは下図の「Ｂ」の断層。これには、両方とも「旧」（図１の(3)）の水準を、別途推定する必要がある。

図１　断層のイメージ



・今回のデータでは、(2) が 257,090円 (3) が 238,030円（本来は観測できない）
←「真」の断層（(2)/(3)）は、 1.08007　これらの結果をベースに評価を行う。
・「旧」の対象のみで平均を計算すると　237,144円
←「(3)両方とも旧」の 238,030円 よりも低い値となっている。
・断層の推定結果は　1.08411（=238,030円 / 237,144円）
← 962円（(3)×0.00404）程度、差がある。

３．断層の推定方法

(1) 重回帰分析による方法

・世帯主の属性変数を用いて消費支出を予測する線形重回帰モデルを構築。
・旧調査票（新調査票）に割り当てられた場合に 0（1）となるダミー変数 D_i を導入。
・ダミー変数の係数 $\hat{\gamma}$ が、調査票の違いによる消費支出の差の推定量となる（星野（2009））。
（※全てが「新調査票」の場合と全てが「旧調査票」の場合との差になることに注意）
【モデル式】

$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} \cdots + \beta_p x_{ip} + \gamma D_i + \varepsilon_i$$

y_i：消費支出
$x_{i1}, x_{i2}, \cdots, x_{ip}$：属性変数（log(年収)、性別、年齢、職業、地域　等）
D_i：＝0（旧調査票）＝1（新調査票）

・デザインに基づく重回帰モデルの考え方（土屋（2009））に基づき、乗率をウエイトとした加重最小自乗法を用いて推定を行う（推定には R の関数 lm を使用。）。
・年収や消費支出は、対数変換を行う。目的変数を対数変換しているため、$\exp(\hat{\gamma})$ が新・旧の断層（倍率）の推定値となる。
・断層の推定結果は　1.10223　← 5,275円（(3)×0.02216）程度、差がある。

(2) 差分の差（difference in differences）推定

・２時点間のデータを使用し、調査票違いによる影響を取り除く。

図２　差分の差推定のイメージ

	切替え：あり		切替え：無し	
2017年9月 (T = 0)		旧【B】 (D = 0)		旧【D】 (D = 0)
2017年10月 (T = 1)	新【A】 (D = 1)			旧【C】 (D = 0)

Ａ－Ｂ＝「新調査票の効果」＋「9月データから10月データへの変化」

Ｃ－Ｄ＝「9月データから10月データへの変化」

（Ａ－Ｂ）－（Ｃ－Ｄ）＝新調査票の効果・・・以下のモデルで推定する。

【モデル式】

$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} \cdots + \beta_p x_{ip} + \gamma D_i + \lambda T_i + \delta D_i T_i + \varepsilon_i$$

$\exp(\delta)$ が新調査票の効果の推定量となる。

・断層の推定結果は　1.07491　← 1,228円（(3)×(−)0.00516）程度、差がある。

(3) 傾向スコアに基づく方法

・新調査票（旧調査票）に割り当てられた場合に1(0)となる割当て変数 Z_i を導入。

・世帯主の属性変数を用いて新調査票への割当て確率を推定するロジスティック回帰モデルを構築。

【モデル式】

$$log(e_i/(1-e_i)) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$e_i = p(z_i = 1|x_{i1}, x_{i2}, \cdots, x_{ip})$：新調査票への割当て確率（傾向スコア）

$x_{i1}, x_{i2}, \cdots, x_{ip}$：属性変数（年収、性別、年齢、職業、地域　等）

Z_i：＝0（旧調査票）＝1（新調査票）

・傾向スコアの推定値（$\hat{e}_i$）を基にしたIPW 推定量（Inverse Probability Weighting Estimator）により、全てが旧調査票であった場合の平均値 $\hat{E}^{IPW}(y_0)$ を、以下のとおり推定。

【IPW推定量】

$$\hat{E}^{IPW}(y_0) = \sum_{i=1}^{N} \frac{(1-z_i)y_i}{(1-e_i)} \Big/ \sum_{i=1}^{N} \frac{1-z_i}{1-e_i}$$

・断層の推定結果は　1.08247　← 571円（(3)×0.0024）程度、差がある。

（$\hat{E}^{IPW}(y_0)$ ＝ 237,504円　は、真の断層 238,030円 と差がある。）