資料２

（2）全国家計構造調査と全国単身世帯収支実態調査の統合集計について
ー傾向スコアを用いた全国消費実態調査（単身世帯）と全国単身世帯収支実態調査の合成ー

慶應義塾大学 経済学部　星野崇宏

慶應義塾大学大学院経済学研究科　博士課程/Graduate School of Economics, University of Wisconsin-Madison　清水祐弥

研究の概要

- 目的

サンプルサイズは大きいが単身若年層の少ない全国消費実態調査（以下, 全消）と, 若年層が多い全国単身世帯収支実態調査（以下, 全単）のデータを融合する[1].

⇒単身若年層の地域別など粒度の細かい分析を行うためには意義がある

- 手法

集団の違いがある可能性があることから強く無視できる割り当てを仮定した, 傾向スコアを用いた重みづけ.

- 注意点：全国単身世帯収支実態調査の対象者

総務省統計局が指示する地域別調査世帯配分数に基づき，業務を受託した民間調査機関が保有・管理する登録モニター等の調査協力世帯の中から選定した全国の単身世帯約2,000世帯⇒集団の違い

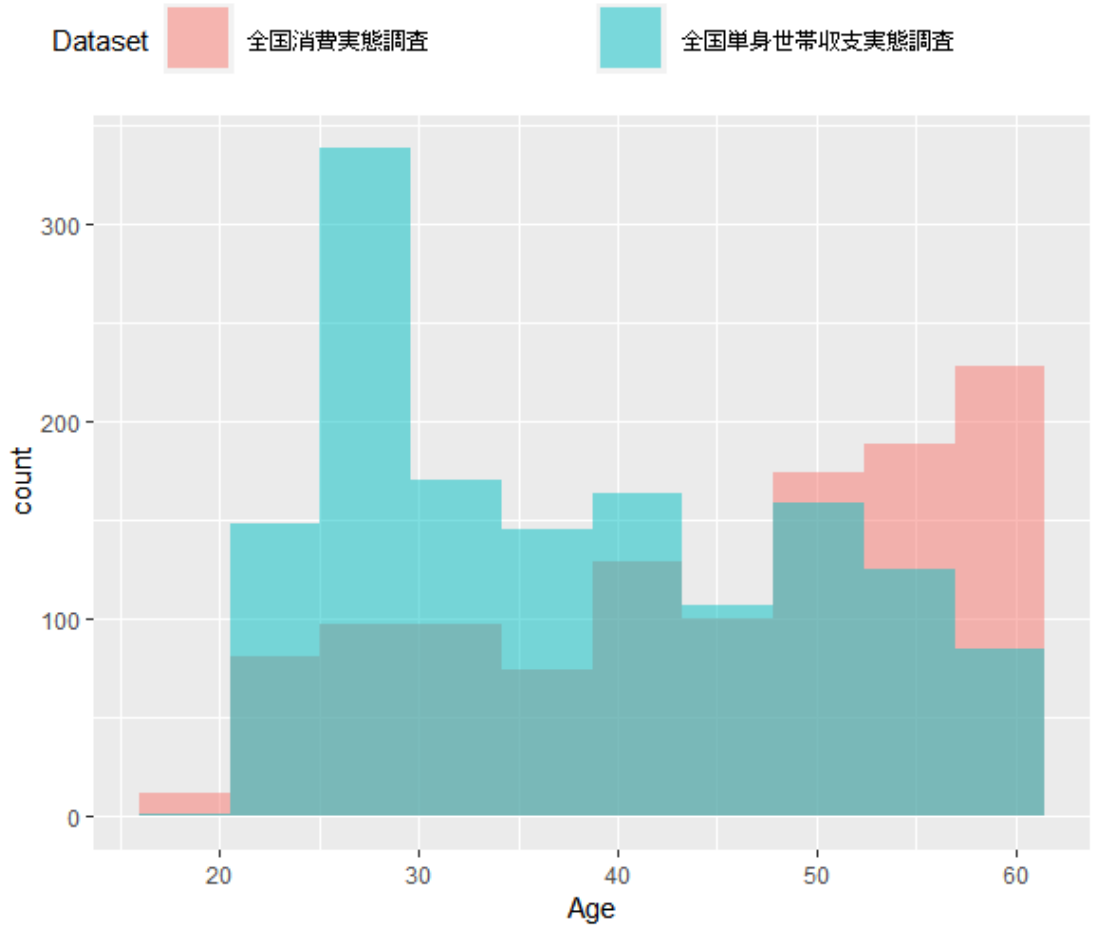
1. ただし, 年齢は60歳以下に限定.

総務省統計局が指示する地域別調査世帯配分数

地方，男女，年齢階級別調査世帯数

		計	30歳未満	30～39	40～49	50～59	60歳以上
男女計	全国	2,000	598	390	290	322	400
	北海道地方	103	28	20	16	18	21
	東北地方	109	32	17	13	23	24
	関東地方	845	257	189	138	120	141
	北陸地方	60	23	8	5	10	14
	東海地方	195	60	40	30	29	36
	近畿地方	324	92	58	48	51	75
	中国地方	101	34	17	12	14	24
	四国地方	54	14	8	4	12	16
	九州地方	191	54	30	21	40	46
	沖縄地方	18	4	3	3	5	3
男	全国	1,144	351	254	198	209	132
	北海道地方	54	16	12	10	10	6
	東北地方	58	17	10	9	15	7
	関東地方	509	156	124	95	81	53
	北陸地方	34	14	5	4	6	5
	東海地方	121	37	30	22	20	12
	近畿地方	178	52	36	33	32	25
	中国地方	55	19	11	8	10	7
	四国地方	28	8	5	3	8	4
	九州地方	97	31	19	12	23	12
	沖縄地方	10	1	2	2	4	1
女	全国	856	247	136	92	113	268
	北海道地方	49	12	8	6	8	15
	東北地方	51	15	7	4	8	17
	関東地方	336	101	65	43	39	88
	北陸地方	26	9	3	1	4	9
	東海地方	74	23	10	8	9	24
	近畿地方	146	40	22	15	19	50
	中国地方	46	15	6	4	4	17
	四国地方	26	6	3	1	4	12
	九州地方	94	23	11	9	17	34
	沖縄地方	8	3	1	1	1	2

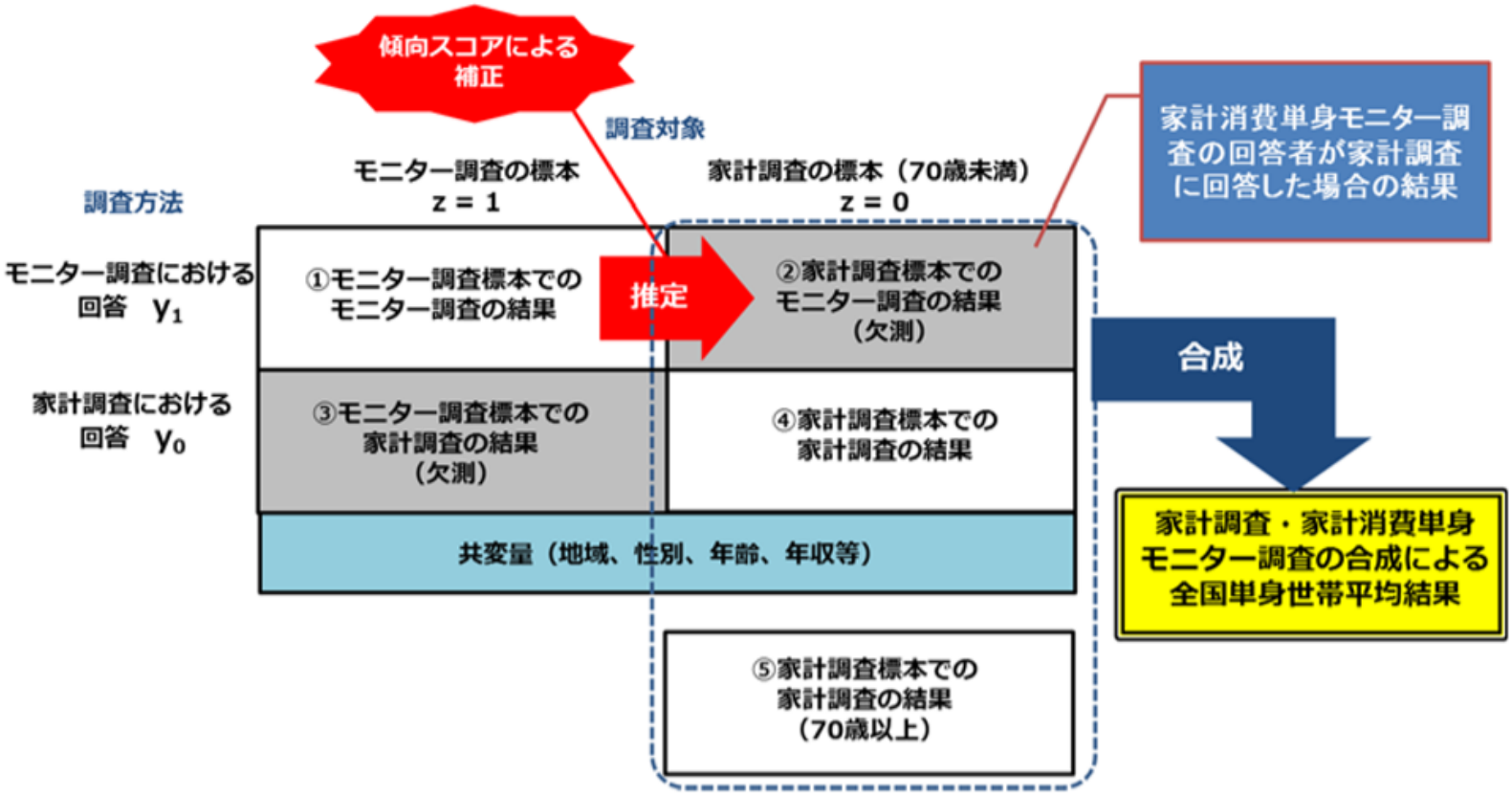
年齢の分布


Dataset
全国消費実態調査
全国単身世帯収支実態調査
300
200
100
0
count
20
30
40
50
60
Age

世帯消費動向指数（CTIミクロ）のモニター融合

https://www.stat.go.jp/data/cti/pdf/micro20180309.pdf

2400人程度の単身モニター対象者の偏りを考慮



手法のイメージ

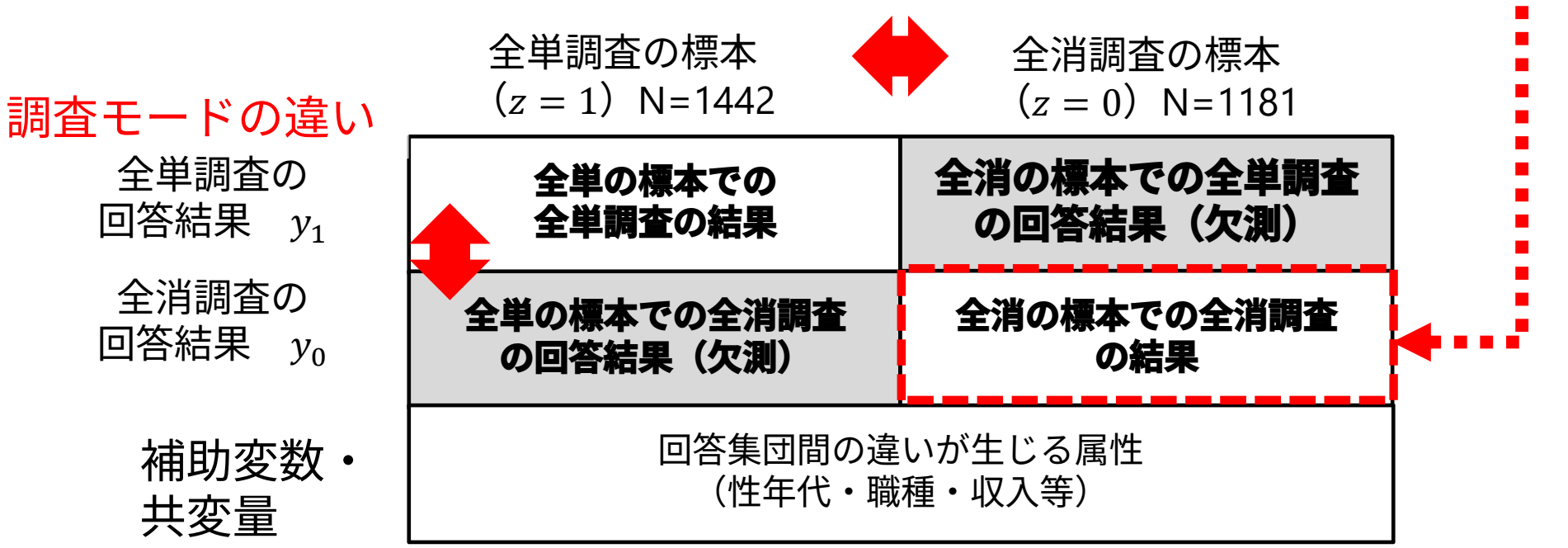
“選択バイアス”＝回答集団の違い

“調査・データ取得モードの違い”＝取り方の違い

⇒両者の違いが交絡しているので

分離して議論したい

今回の目的は
全単調査を用いて
こちらのサンプル
を増やすこと



手法のイメージ

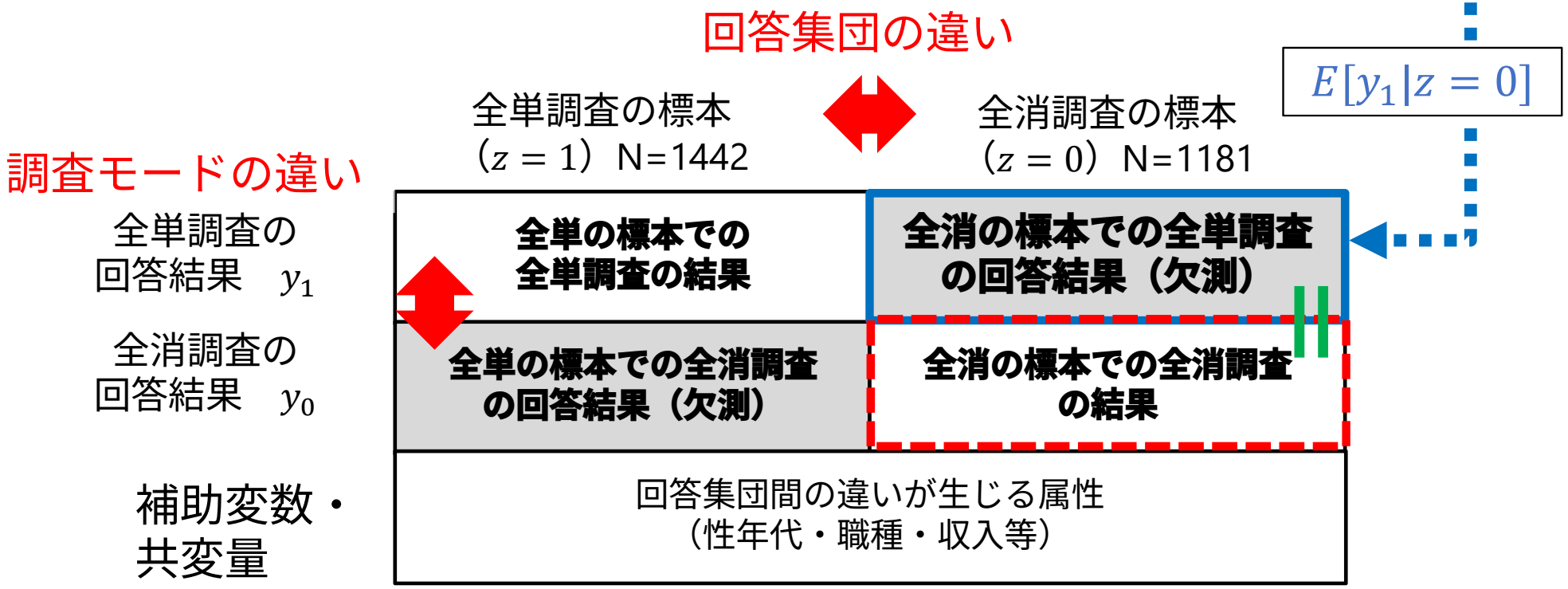
“調査・データ取得モードの違い”

＝取り方の違い（今回は項目の違い）

がないと仮定すると

もしも年齢の偏りが少ない全単の標本が, 全消の標本だった場合の結果
→傾向スコア$p(z=1|x)$
による重みづけにより推定

$E[y_1|z=0]$

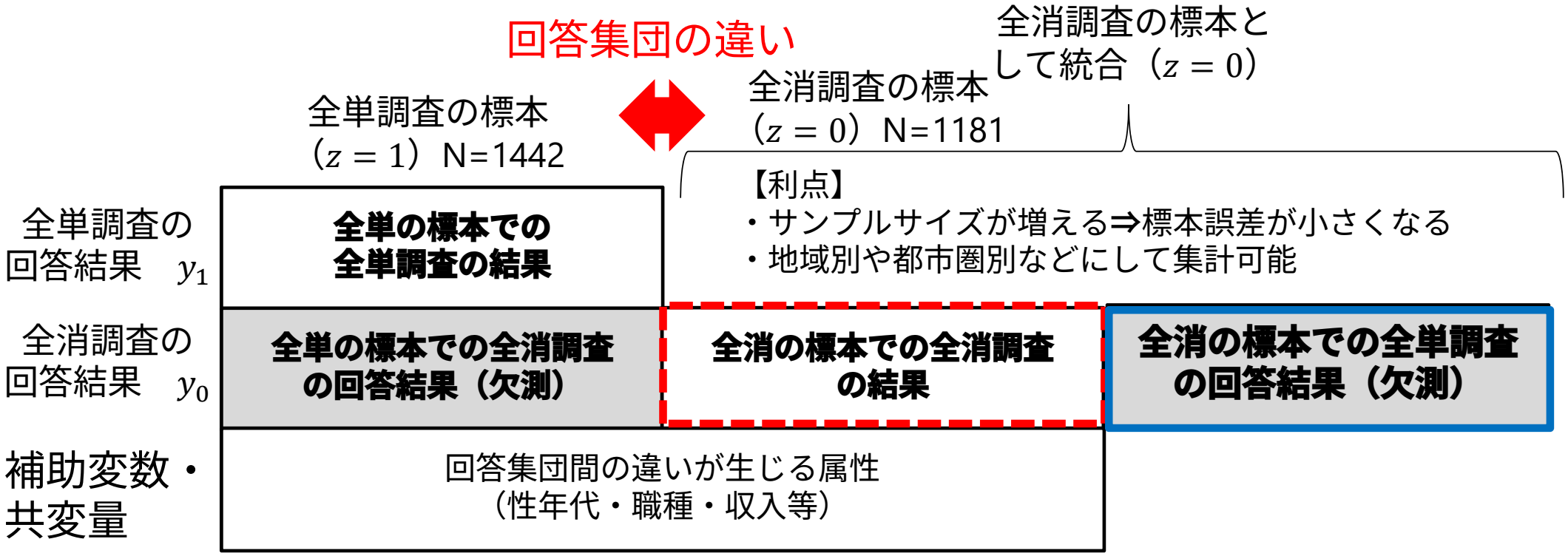


手法のイメージ

“調査・データ取得モードの違い”

＝取り方の違い（今回は項目の違い）

がないと仮定すると



手法の説明

- Goal: $E[y_1|z=0]$ の推定
  （もしも年齢の偏りが少ない全単の標本が, 全消の標本だった場合の結果）

- 直感的には, 全単に所属しそうな人ほど割り引いて, 全消に所属しそうな人ほど大きく重み付ければよい.

- $= \dfrac{\sum_{i=1}^{N} \frac{z_i(1-e_i)}{e_i} y_{1i}}{\sum_{i=1}^{N} \frac{z_i(1-e_i)}{e_i}}$ によって推定.

ただし, e_iは傾向スコア $p(z=1|x)$ の推定値.

今回の傾向スコアの変数選択の候補

消費統計研究会(平成29年度第1回)

資料4 別紙「傾向スコアを用いた家計調査（単身世帯）と家計消費単身モニター調査の合成に関する試算結果」に存在する変数を利用

性別ダミー　　年齢（連続）

職業分類区分（13区分）　住居の所有関係（6区分）

居住面積

都市階級（5区分）

地方区分（10区分）

世帯年収の対数値 （連続）⇒世帯年間収入では使わない

純資産総額 （連続）

消費支出の対数値（連続）⇒消費支出では使わない

variable	変数	符号表の名前	値	説明	分析用の区分	変数
gender	性別	Seibetsu	1	男		1
			2	女		0
age	満年齢	Nenrei	△△0～116	0～116歳		
job	職業符号	Shokugyo	△1	常用労務作業者	労務作業者	JOB1
			△2	臨時及び日々雇労務作業者	労務作業者	JOB1
			△3	民間職員	民間職員	基準
			△4	官公職員1	官公職員	JOB2
			△5	官公職員2	官公職員	JOB2
			△6	商人及び職人	民間職員	基準
			△7	個人経営者	経営者	JOB3
			△8	農林漁業従事者	民間職員	基準
			△9	法人経営者	経営者	JOB3
			10	自由業者	経営者	JOB3
			11	その他	経営者	JOB3
			12	無職	無職	JOB4
			13	家族従業者	民間職員	基準
housing	住居の所有関係	JukyoShoyu	1	持ち家	持ち家	HOU1
			2	民営の賃貸住宅	民営の賃貸住宅	基準
			3	都道府県･市区町村営賃貸住宅	公営の賃貸住宅･社宅等	HOU2
			4	都市再生機構･公社等の賃貸住宅	公営の賃貸住宅･社宅等	HOU2
			5	社宅･公務員宿舎（借上げの社宅を含む）	公営の賃貸住宅･社宅等	HOU2
			6	借間	民営の賃貸住宅	基準
space	住居の敷地面積	ShikichiMenseki	△△△0.1～9999.9	住居の敷地面積		
city	都市階級	C_ToshiKaikyu	1	大都市	大都市	基準
			2	中都市	中都市	CIT1
			3	小都市A	小都市A	CIT2
			4	小都市B	小都市B･町村	CIT3
			5	町村	小都市B･町村	CIT3
region	地方区分	ChihoKubun	01	北海道	北海道･東北	REG1
			02	東北	北海道･東北	REG1
			03	関東	関東	基準
			04	北陸	北陸･東海	REG2
			05	東海	北陸･東海	REG2
			06	近畿	近畿	REG3
			07	中国	中国･四国	REG4
			08	四国	中国･四国	REG4
			09	九州	九州･沖縄	REG5
			10	沖縄	九州･沖縄	REG5

	consum	消費支出	Bdy083	△△△△△△△0.000000～	金額（円）
	income	世帯の年間収入	M_Nenshu	△△△△△△△△1～	金額（万円）
	asset	純資産総額	BdyShisan011	-9999999.999999～	金額（千円）
	debt	負債現在高	BdyShisan020	△△△△△△△0.000000～	金額（千円）
	saving	貯蓄計	ChochikuTotal	△△△△△1～999999	金額（万円）
	engel	エンゲル係数			consum/income
	ln_con	消費支出の対数値			log(consum)
	ln_inc	世帯年収の対数値			log(income)
	ln_sav	貯蓄計の対数値			log(saving)

傾向スコアの変数選択

- 傾向スコアのモデル

$$logit[p(r=1|x)] = \boldsymbol{x}^t\boldsymbol{\alpha}$$

- 理論や先行研究で関連すると考えられる変数を, 説明変数$\boldsymbol{x}$に含める.
- Brookhart et al. (2006) のシミュレーション結果

（たとえ,割り当て z には相関しなくても）回答結果 y を説明する変数は常に傾向スコアのモデルに含まれるべきである.

（今回は）関心のある変数を5%有意に説明する共変量（ダミー変数のグループの場合は一つでも有意なら）を含める

＊消費支出はこの基準だと共変量がバランスしなかったので追加の変数を加えて分析

参考：Imbens and Rubin(2014)による共変量選択

共変量候補K個のうち

(1) 先行研究の知見等から必ず入れるK_B個を指定

(2) K_L-K_B個についてロジスティック回帰分析で割り当てに関連するかという観点からK_L-K_B個選択

変数増加法

(3) K_L個の変数の2次項とその交互作用項を含めた$\frac{K_L \times (K_L+1)}{2}$個の項から$K_Q$個選択し、1次項を含めて$K_Q$＋$K_L$個選択

その際には変数増加法

(4) 推定した傾向スコアで層別して共変量の分布の違いがなくなっているかをチェックする

合成値の考え方

調査モードがないとするので$\theta_0 = E[y_0|z=0] = E[y_1|z=0]$には
二つの推定量$\bar{y}_0$と$\hat{E}[y_1|z=0]$が存在。これらに重みwをつけて

$$\hat{\theta}_0 = w\bar{y}_0 + (1-w)\hat{E}[y_1|z=0]$$とするときのwは？

【1:最適な重みにする】

$V(\hat{\theta}_0)$を最小にする重みは（両推定量がほぼ無相関のため）

$$w = \frac{V(\hat{E}[y_1|z=0])}{V(\bar{y}_0) + V(\hat{E}[y_1|z=0])}$$

【2:人数比で考える】

全消調査の標本のサンプルサイズ（z=0）N_0= 1181と
全単調査の標本のサンプルサイズ（z=1）N_1= 1442より

$$w = \frac{N_0}{N_0 + N_1}$$

＊その後の層別の分析の際の重みの分かりやすさという点では2か？

推定: 消費支出

使用した変数

- 性別ダミー
- 年齢
- 居住面積
- 職業分類ダミー（4変数）
- 住居の所有関係ダミー（2変数）
- 都市階級ダミー（3変数）
- 地方ダミー（5変数）
- 世帯年収の対数値
- 純資産総額

合成値１：全消単体の場合に比べて推定値の分散が1/1.876に減少
⇒（全消）1181人のデータを2216人分に増やしたのと同じ

合成値２：1/1.308に減少
⇒（全消）1545人分に増やしたのと同じ

	消費支出
全消平均	176014.2
(s.e.)	(3301.6)
全単平均	177504.2
(s.e.)	(3315.4)
推定値	169888.2
(s.e.)	(5948.0)
合成値１	1745471.3
(s.e.)	(2410.2)
合成値２	172646.4
(s.e.)	(2886.7)

推定: 消費支出

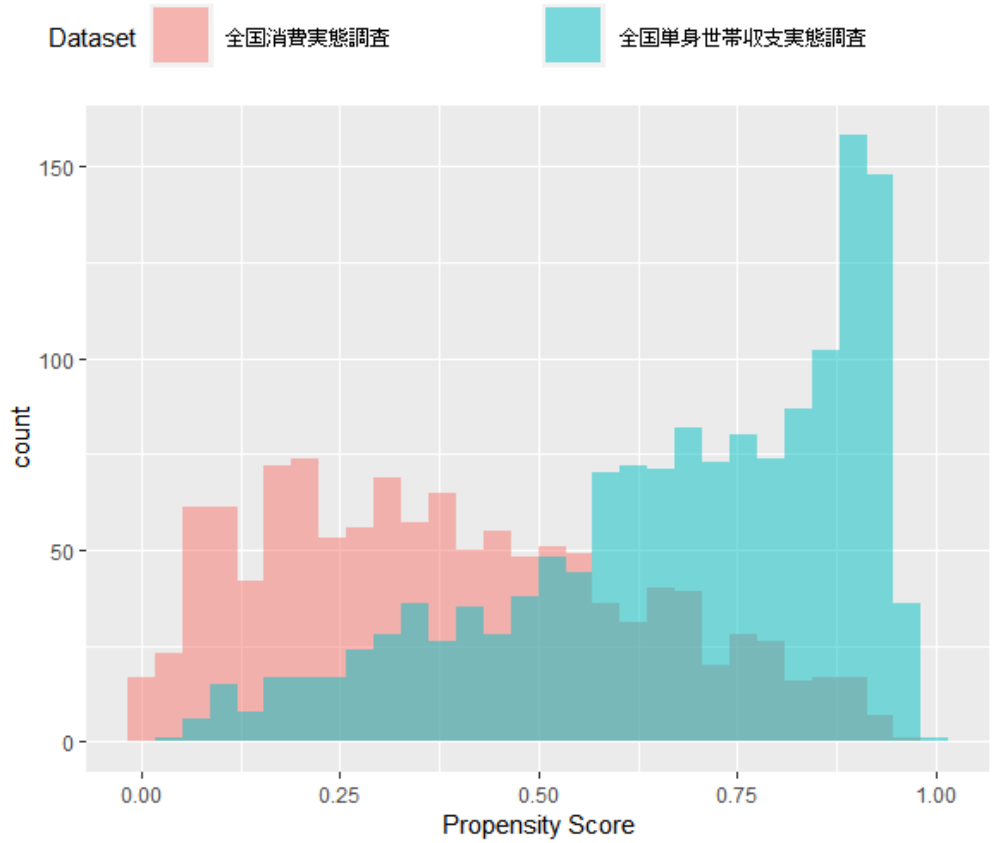


図: 傾向スコア $p(z = 1|x)$ の推定値の分布

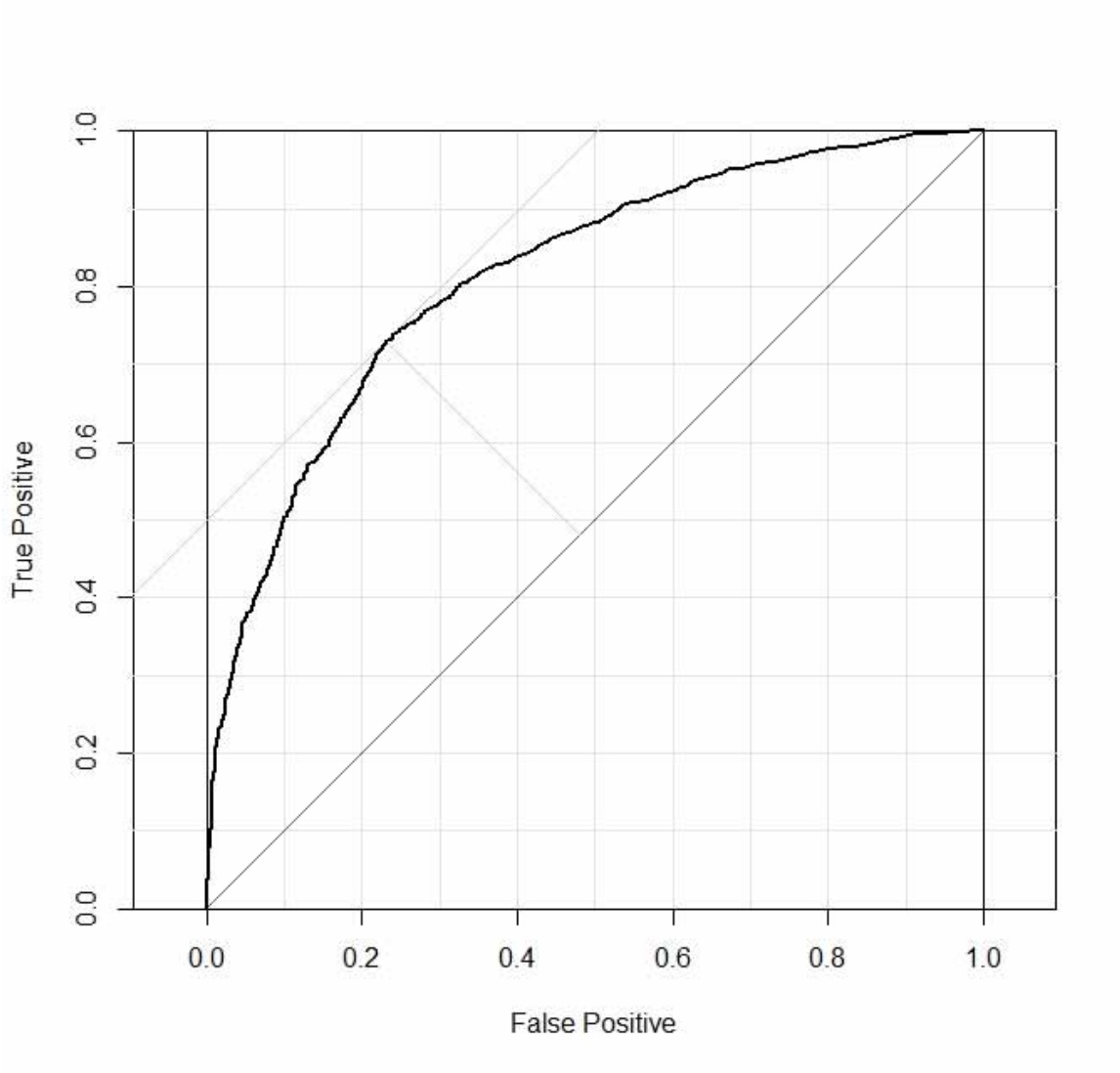
推定: 消費支出

- z=0とz=1での共変量の平均値の絶対差

変数名	絶対差（傾向スコアによる調整前）	絶対差（傾向スコアによる調整後）
年齢	0.806	1.827
性別ダミー	0.105	0.021
居住面積	11.046	4.549
世帯年収の対数値	0.634	0.287
純資産総額	349.560	825.901

ROC曲線：AUC

AUC（C統計量）
0.8129
一定程度以上の予測力を
有すると考える



推定: 年間収入

使用した変数

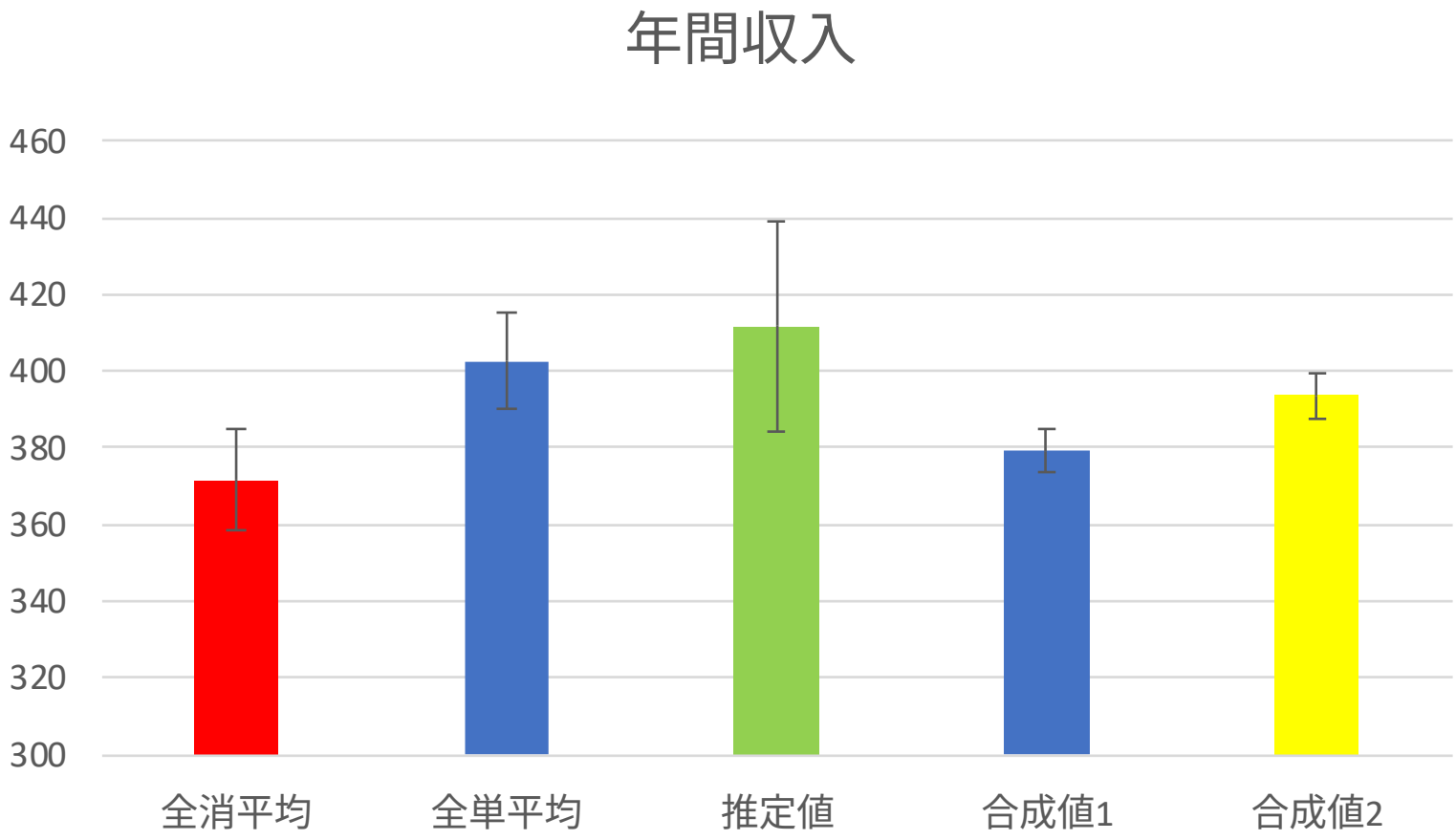
- 性別ダミー
- 年齢
- 居住面積
- 職業分類ダミー（4変数）
- 住居の所有関係ダミー（2変数）
- 都市階級ダミー（3変数）
- 地方ダミー（5変数）
- 消費支出の対数値

合成値１：全消単体の場合に比べて分散が1/1.475に減少
⇒（全消）1181人のデータを1741人分に増やしたのと同じ

合成値２：同様に分散が1/1.243に減少
⇒（全消）1468人に増やしたのと同じ

	年間収入
全消平均	371.5
(s.e.)	(6.8)
全単平均	402.7
(s.e.)	(6.4)
推定値	411.7
(s.e.)	(14.1)
合成値１	379.0
(s.e.)	(5.6)
合成値２	393.6
(s.e.)	(6.1)

結果の比較


年間収入
460
440
420
400
380
360
340
320
300
全消平均
全単平均
推定値
合成値1
合成値2

推定: 年間収入

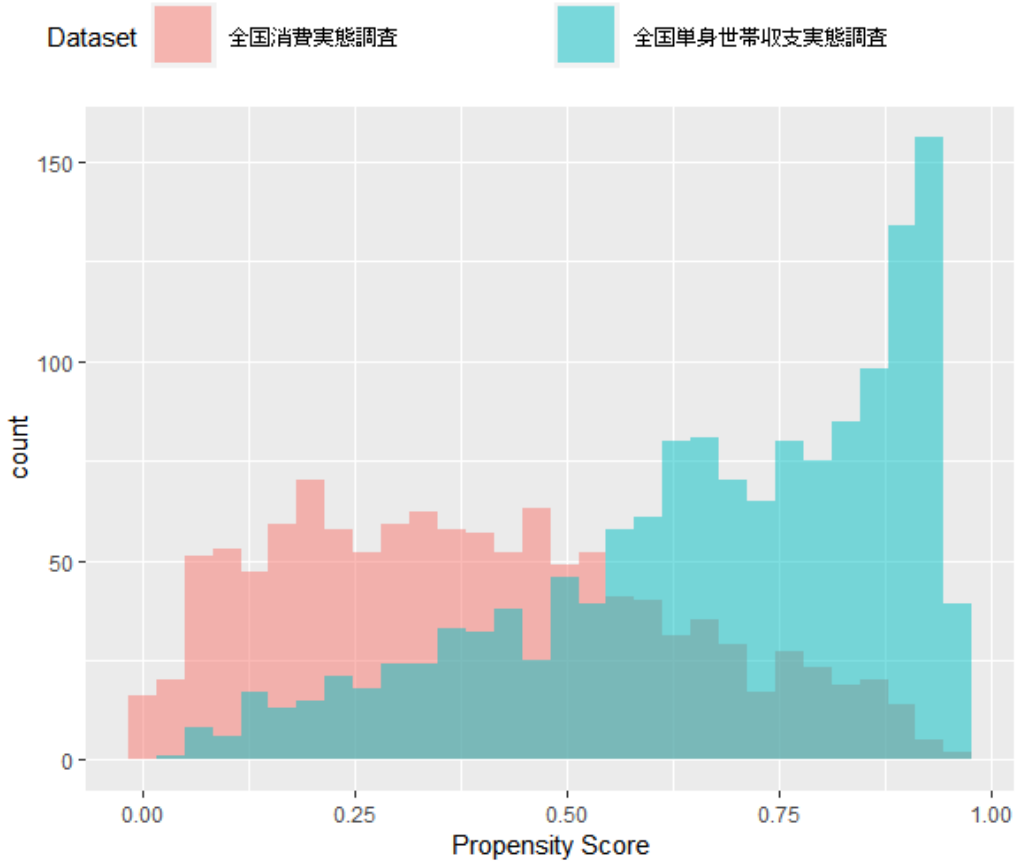


図: 傾向スコア $p(z = 1|x)$ の推定値の分布

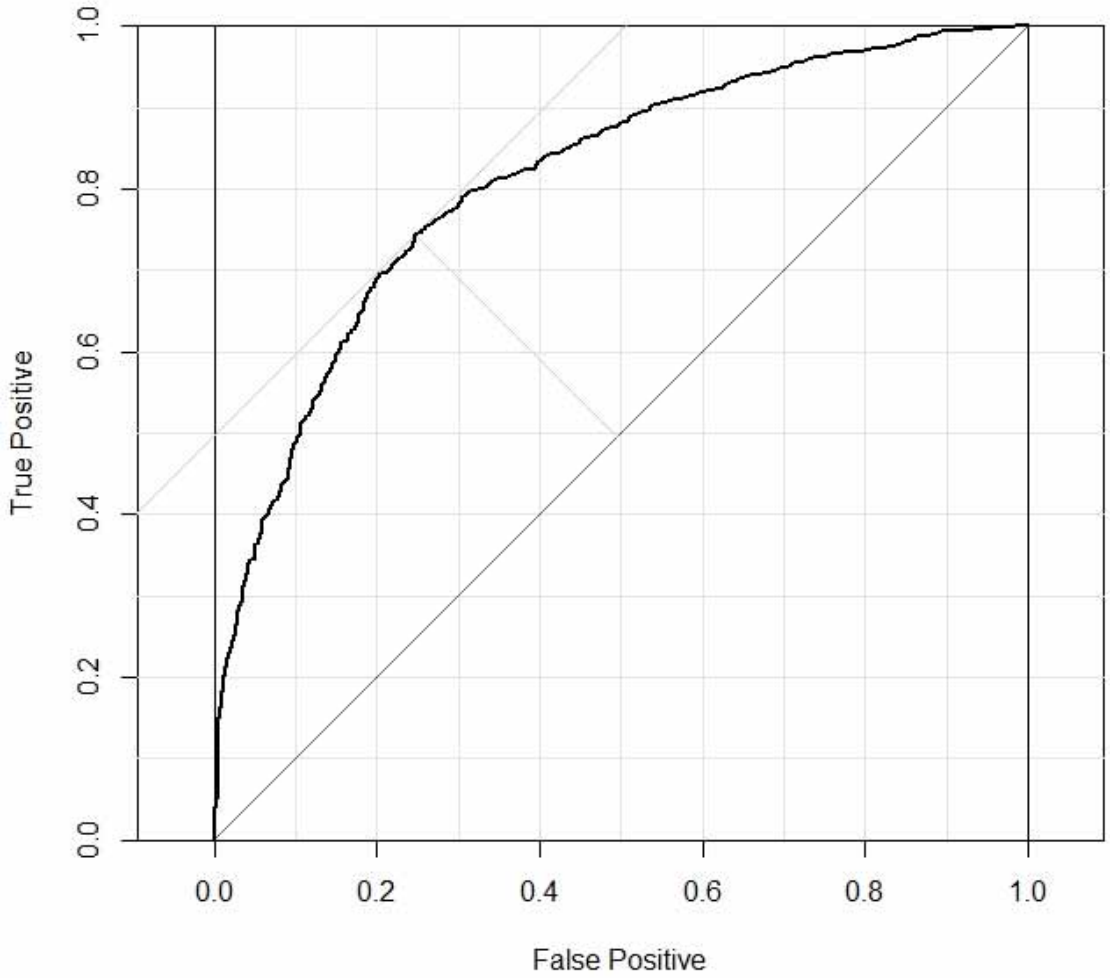
推定: 年間収入

- z=0とz=1での共変量の平均値の絶対差

	絶対差（傾向スコアによる調整前）	絶対差（傾向スコアによる調整後）
年齢	0.806	2.191
性別ダミー	0.105	0.022
居住面積	11.047	5.759
消費支出の対数値	1.189	0.629

ROC曲線：AUC

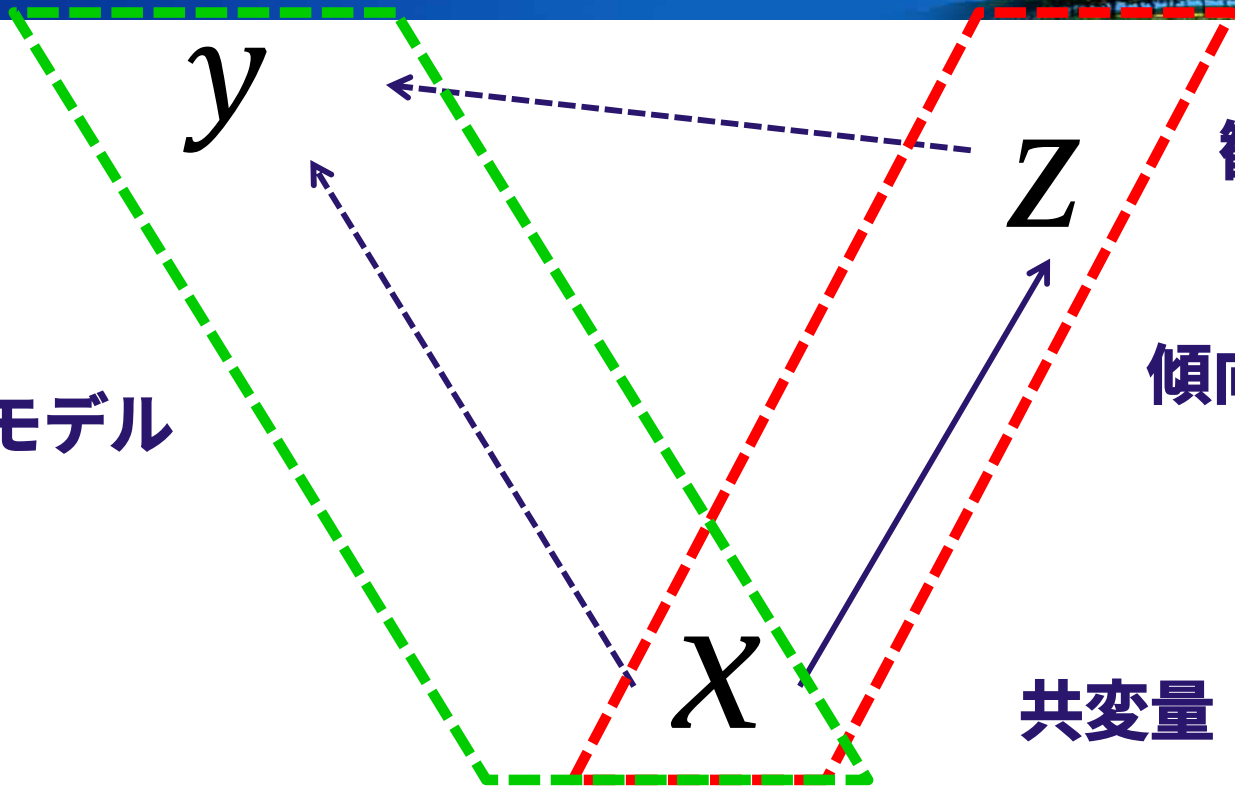
AUC（C統計量）
0.8100
一定程度以上の予測力を
有すると考える



議論

- サンプルサイズは大きいが単身若年層の少ない全国消費実態調査（以下, 全消）と, 若年層が多い全国単身世帯収支実態調査（以下, 全単）のデータを融合する方法論を提示し、消費と年間収入について実施した。
- 両者の集団集団の違いがある可能性があることから強く無視できる割り当てを仮定した, 傾向スコアを用いた重みづけを実施した。
- 共変量としては消費統計研究会(平成29年度第1回)資料4 別紙「傾向スコアを用いた家計調査（単身世帯）と家計消費単身モニター調査の合成に関する試算結果」に存在する変数を利用した。
- 融合時の重み付けは「サンプルサイズ比率で足し上げる」方法以外に「分散を最小化させる」ものも存在する方法も利用した。
- どちらの方法が良いかは目的に依存する（分かりやすさか精度か）
- 乗率（国勢調査へ）の利用や地域別集計なども考慮する必要
- さらに精度を高めるには全単のバイアス補正に二重にロバストな推定法を利用することが考えられる

Doubly Robust Estimation



因果効果推定のためには「outcome Yとcovariate Xの回帰モデル」か「indicator Zとcovariate Xの(離散)回帰モデル」かを推定すればよい
⇒両方のモデルも用意し“どちらかが正しければ”推定可能

二重にロバストな推定

従属変数　y　のDoubly robust (DR) estimator

Covariate Xによる回帰関数 $g(x_i, \boldsymbol{\beta})$ もPSも推定して

$$\hat{E}^{DR}(y) = \frac{1}{N}\sum^{N}\left[\frac{z_i y_i}{e(\mathbf{x}_i)} + \left(1 - \frac{z_i}{e(\mathbf{x}_i)}\right) g(x_i, \hat{\boldsymbol{\beta}})\right]$$

$e(x)$ が正しいと $g(x, \boldsymbol{\beta})$ 誤設定でもゼロ

$$= \frac{1}{N}\sum^{N}\left[y_i + \frac{z_i - e(\mathbf{x}_i)}{e(\mathbf{x}_i)}\left(y_{i1} - g(x_i, \hat{\boldsymbol{\beta}})\right)\right]$$

$g(x, \boldsymbol{\beta})$ が正しいと $e(x)$ 誤設定でもゼロ

・一般に推定量の分散がsemiparametricモデルの中で最小になる

・一般にどちらのモデルもmis-specifyした場合ではバイアスが小さい

⇒ “モデル比較や検定”を行わずに複数のモデルの組み合わせを行う方法論として近年発展(例:Han,2014,JASA)

二重にロバストな推定も入れた結果

単純には全消と全単で合計2623人
実際には全単のバイアス補正で
推定が不安定に
⇒いかに2623人に近づけるか？

合成値１とDR合成値１の性能が高い

	消費支出	何人分?
全消平均	176014.2	1181
(s.e.)	3301.6	
全単平均	177504.2	1442
(s.e.)	3315.4	
推定値	169888.2	
(s.e.)	5948	
合成値１	1745471.3	2216
(s.e.)	2410.2	
合成値２	172646.4	1545
(s.e.)	2886.7	
DR推定値	175808.3	
(s.e.)	4308.2	
DR合成値１	175903.2	2375
(s.e.)	2328.4	
DR合成値２	174788.5	1948
(s.e.)	2570.6	

Appendix: $E[y_1|z=0]$の分散の推定

$\hat{E}[y_1|z=0] = \frac{1}{N}\sum_{i=1}^{N} \frac{z_i(1-e_i)}{e_i \cdot p(z=0)} y_{1i}$ の漸近分散を考える.
また, 簡便のため傾向スコアの推定による分散は考慮しない.

本研究で用いた推定量は真値に対して一致性, 漸近正規性を持ち, その漸近分散は,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{z_i\,(1-e_i)}{e_i \cdot p(z=0)}\left(y_{1i} - \hat{E}\,[y_1|z=0]\right)\right)^2$$

によって推定でき, 推定値の標準誤差は, N で割った値の平方根,

$$\sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{z_i\,(1-e_i)}{e_i \cdot p(z=0)}\left(y_{1i} - \hat{E}\,[y_1|z=0]\right)\right)^2}$$

により推定する.

詳しい証明は星野 (2005) 等を参照されたい. 本研究の漸近分散を導出するには, 星野 (2005) で $\frac{\partial}{\partial\theta}\frac{z}{e}\frac{(1-e)}{p(z=0)} m\,(y_1|\theta, z=0) = \frac{z}{e}\frac{(1-e)}{p(z=0)}\,(y_1 - E\,[y_1|z=0])$ とすればよい.