資料２－３

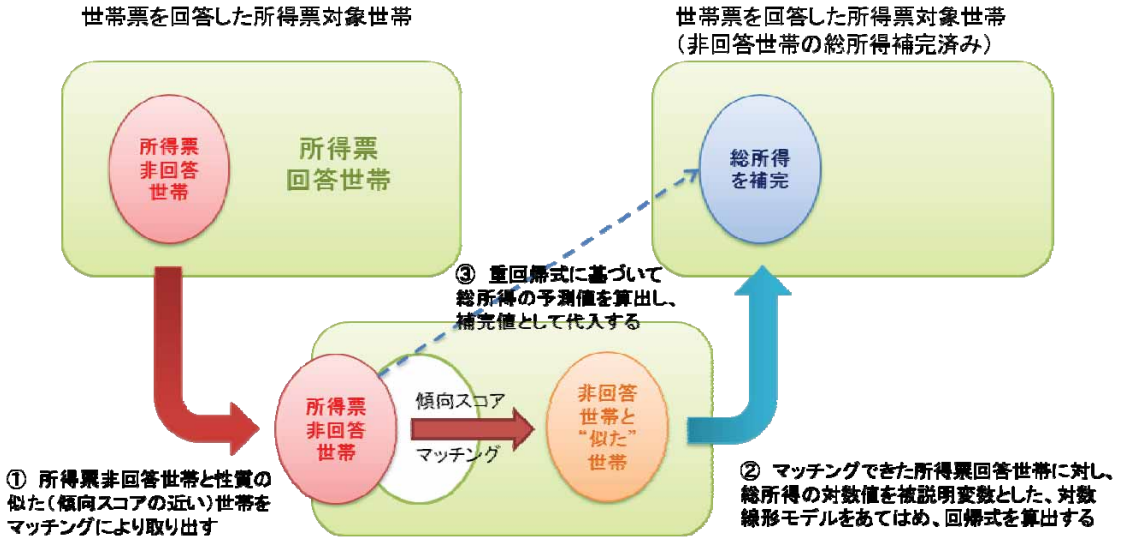
世帯票情報を用いた傾向スコアによる総所得の推定について

1　本試算の概要

本試算では，まず平成 19 年国民生活基礎調査の所得票対象世帯の傾向スコア（回答確率）の推定を行い，その推定値を用いて，IPW（Inverse Probability Weighting）推定及び傾向スコアマッチングによる総所得の代入を行った。後者においては，まず傾向スコア推定値による非回答世帯と回答世帯のマッチングを行い，マッチングした回答世帯の総所得を用いて，非回答世帯の総所得にノンパラメトリックな代入及びパラメトリックな代入を行った。

なお、「所得票対象世帯」には、住み込み又はまかない付きの寮・寄宿舎など所得票の調査対象外世帯（112 世帯）を含むため、概況・報告書記載の調査客体数（36,285 世帯）とは世帯数が異なる。

（傾向スコアマッチングとパラメトリックな代入のイメージ）



2　傾向スコアの推定

所得票対象世帯のうち，回答世帯で **1**，非回答世帯で **0** をとる欠測インディケータを z_i として，n 個の世帯票項目の変数（共変量）を $x_{1i}, x_{2i}, \cdots, x_{ni}$ としたとき，第 i 世帯の回答確率 e_i を第 i 世帯の**傾向スコア**（Propensity score）とよぶ。

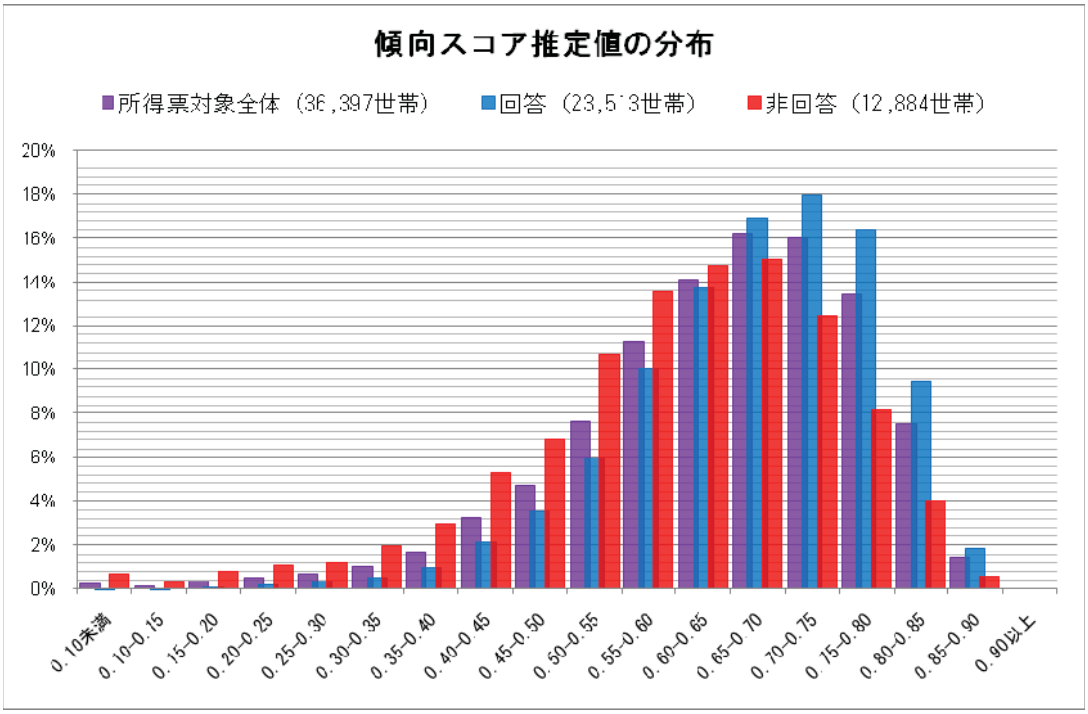
$$e_i = p\left(z_i = 1 \middle| x_{1i}, x_{2i}, \cdots, x_{ni}\right)$$

ただし、e_i の真値はわからないため，その推定値をロジスティック回帰モデルにより求めた。

$$\hat{e}_i = \frac{1}{1+\exp\left(-\left(\hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \cdots + \hat{\alpha}_n x_{ni}\right)\right)}$$

共変量には，“市郡別，地域ブロック，世帯人員数，有業人員数，世帯構造，世帯類型，世帯業態，家計支出額，住居の種類，世帯主年齢階級”を用いた。傾向スコア推定値の分布は次のとおり。

○　傾向スコア推定値の分布

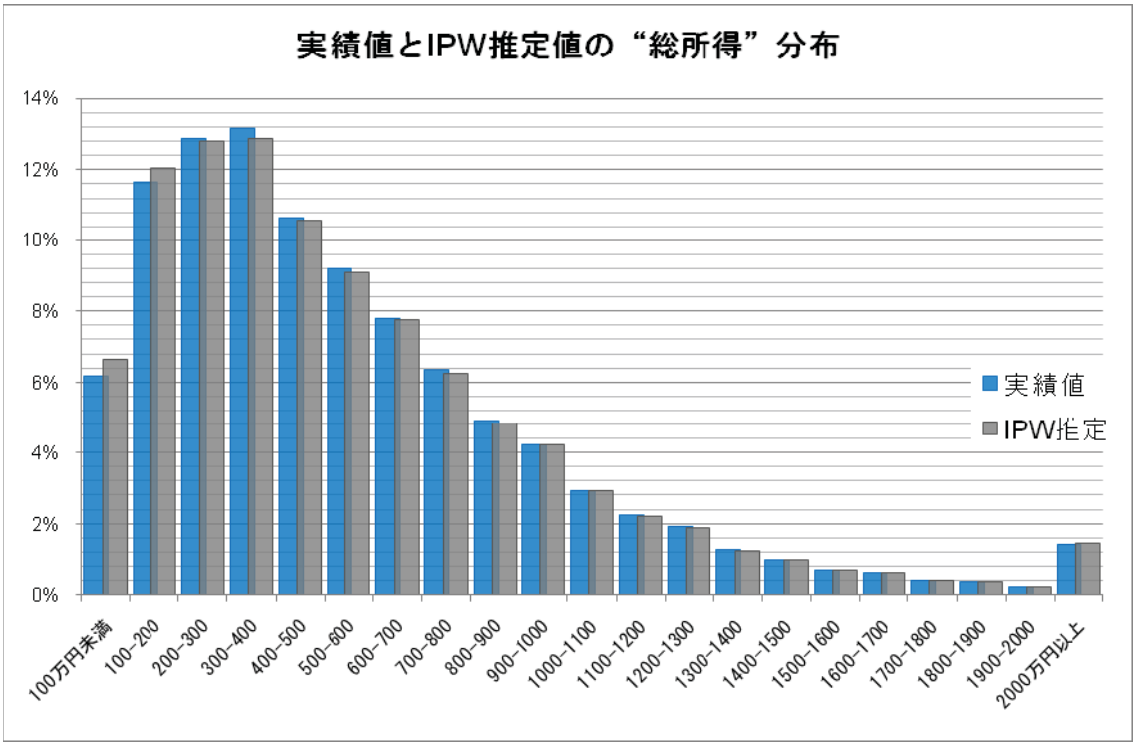


3　IPW（Inverse Probability Weighting）推定

回答世帯の回答結果は，その世帯の回答確率の高低によってウェイトが異なると考えて，世帯毎に傾向スコア推定値の逆数でウェイト付けする。すなわち，現行の拡大乗数 w_i を $\frac{w_i}{\hat{e}_i}$ に置き換えて総所得の推定を行った。

表2：　実績値及びIPW推定値における平均所得額と所得分位値

一世帯あたり平均所得（万円）			所得分位値（万円）		
	実績値	IPW推定値		実績値	IPW推定値
全世帯	566.8	564.0	中央値	451	450
高齢者世帯	306.3	303.5	第1五分位	214	207
母子世帯	236.7	236.4	第2五分位	365	360
父子世帯	515.9	515.9	第3五分位	554	550
その他の世帯	646.8	634.6	第4五分位	838	834



4　マッチングによる代入

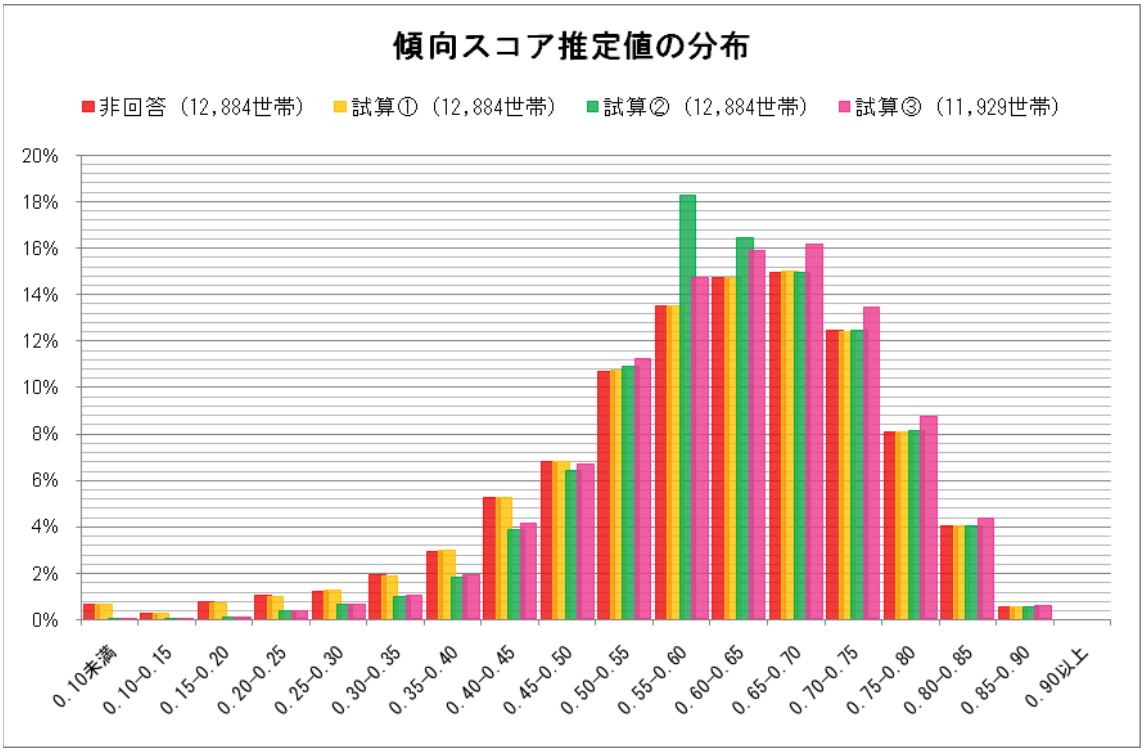
4.1　マッチング

傾向スコア推定値のマッチング方法（１対１）は次の３パターンについて試算した。

- 試算①：最近傍マッチング（復元）
- 試算②：最近傍マッチング（非復元）
- 試算③：キャリパーマッチング（非復元）

（傾向スコア推定値の絶対値の差が 0.001 未満の世帯がなければマッチングせず）

各試算において，非回答世帯の相手としてマッチングされた回答世帯について，集団としての各共変量の構成割合を調べたところ，マッチングされた回答世帯の集団は非回答世帯によく似た集団となった。



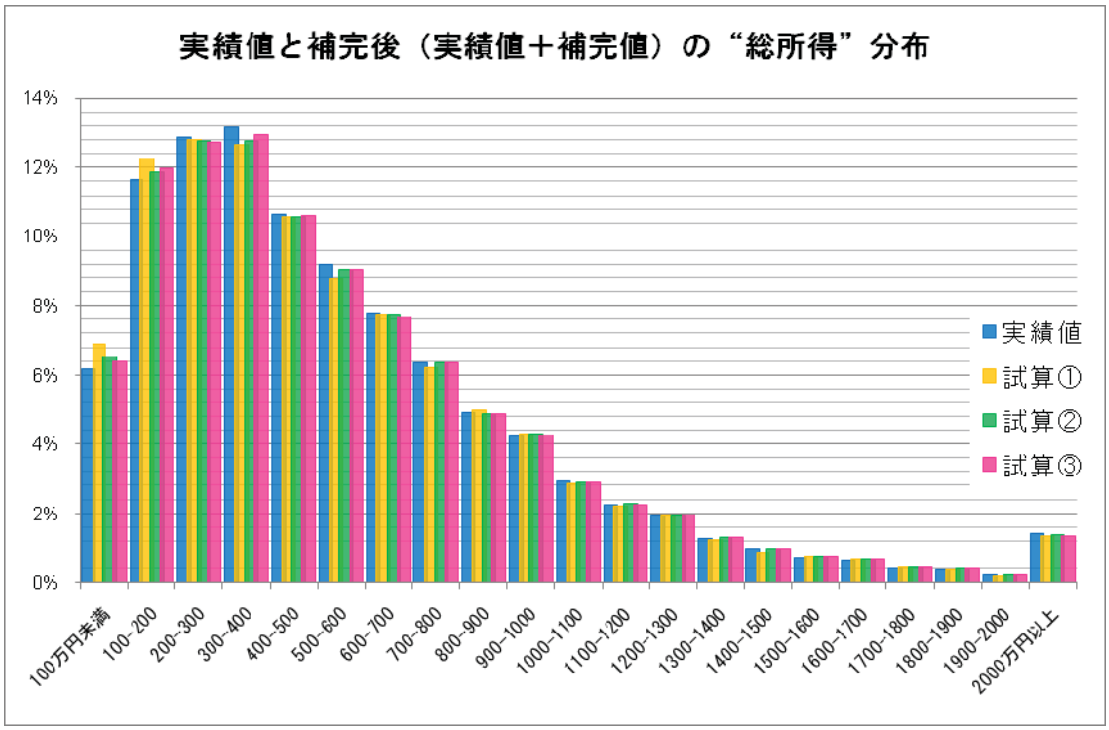
4.2　ノンパラメトリックな代入

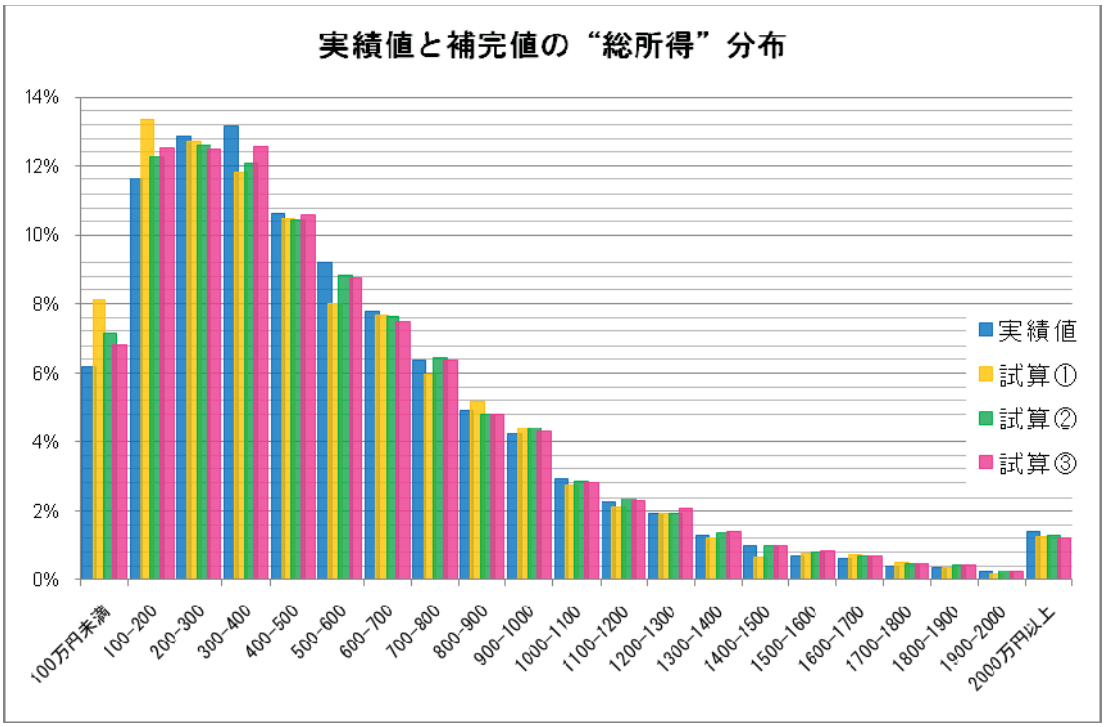
最近傍マッチングにおいて，非回答世帯にマッチングした回答世帯の総所得をそのまま非回答世帯の総所得として代入した。（キャリパーマッチングの結果は参考値として掲載）

表3：　実績値及びノンパラメトリック代入における平均所得額と所得分位値

	実績値	試算①		試算②		試算③（参考値）	
		補完値	補完後	補完値	補完後	補完値	補完後
一世帯あたり平均所得（万円）							
全世帯	566.8	545.1	558.7	564.5	566.0	561.2	564.8
高齢者世帯	306.3	375.1	327.2	452.5	350.9	420.5	340.5
母子世帯	236.7	552.3	361.1	597.4	379.0	581.3	366.3
父子世帯	515.9	384.7	451.0	540.6	528.1	528.3	521.7
その他の世帯	646.8	578.8	620.4	585.9	623.2	590.5	626.2
所得分位値（万円）							
中央値	451	430	446	450	450	450	450
第1五分位	214	189	202	200	210	200	210
第2五分位	365	342	358	360	364	360	363
第3五分位	554	539	550	557	555	550	553
第4五分位	838	830	834	850	842	847	840

一世帯あたり平均所得は，全世帯でみると実績値に近い値で補完されている一方，世帯構造別にみると実績値から大きく乖離した値で補完されているものもある。これは，傾向スコア推定値の“近さ”のみでマッチングするため，出現頻度の低い世帯構造では，同じ世帯構造の回答世帯がマッチングされるとは限らないためである。





4.3　パラメトリックな代入

マッチングした回答世帯の集団における総所得と世帯票情報との関係から，総所得に関する対数線形モデルを構成し，非回答世帯の総所得を推計した。

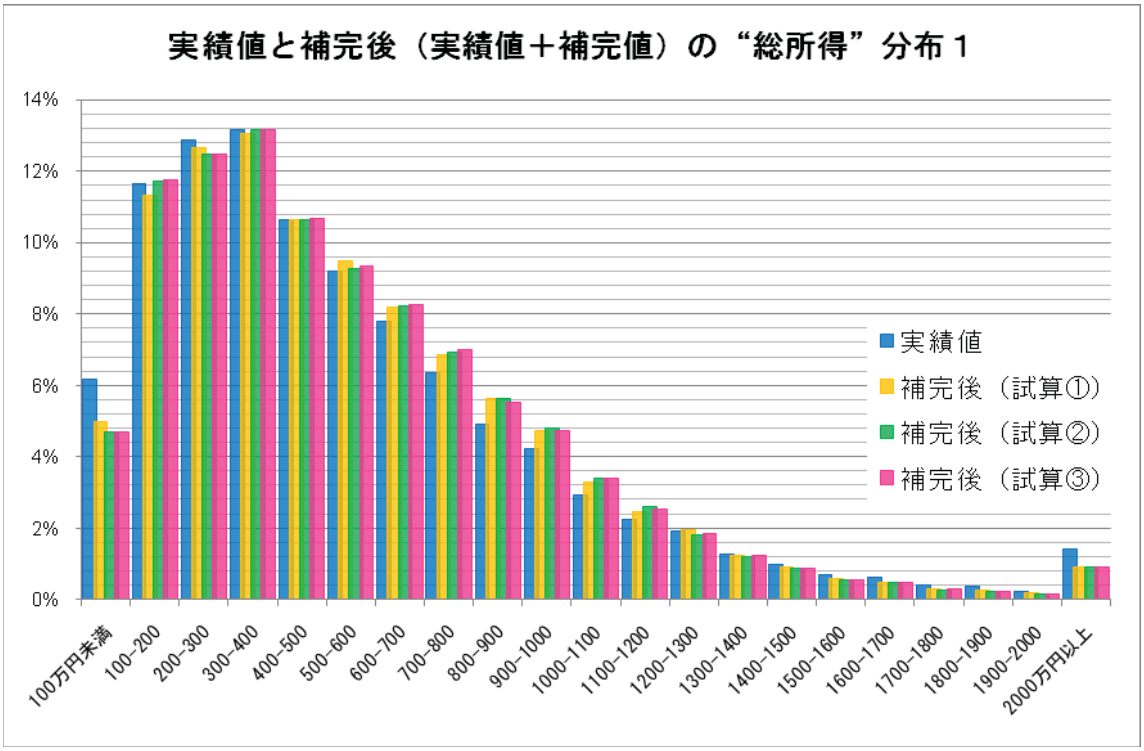
総所得の対数値 $\log(Y_k)$ を被説明変数，世帯票情報 $X_{1k}, X_{2k}, \cdots, X_{mk}$ を説明変数として重回帰分析を行い，重回帰モデル $\log(Y_k) = \beta_0 + \beta_1 X_{1k} + \cdots + \beta_m X_{mk} + \varepsilon_k$ のパラメータの推定値 $\hat{\beta}_l$ を最小二乗法により決定して，欠測した総所得の補完値として代入した。総所得の推計値に用いる推定量によって，以下の３パターンを試算した。

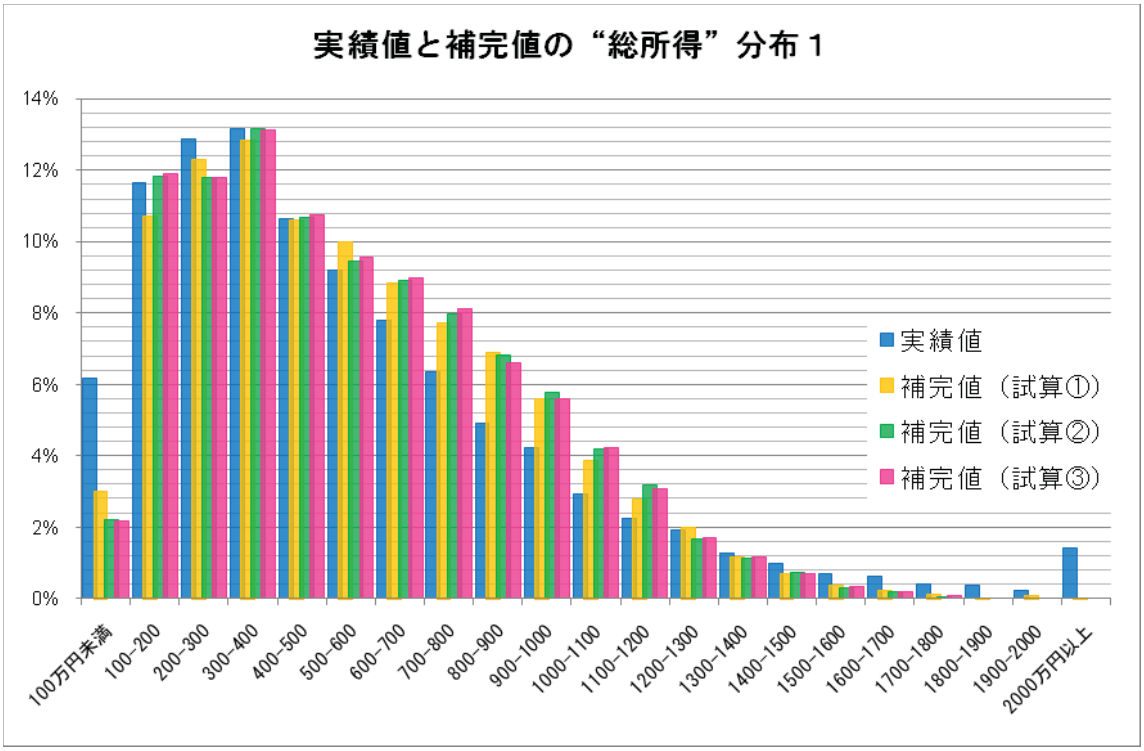
- 不偏推定量：$\hat{Y}_{1t} = \left\{1 + \frac{1}{2}s^2\right\} \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \cdots + \hat{\beta}_m X_{mt}\right)$
- 最尤推定量：$\hat{Y}_{2t} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \cdots + \hat{\beta}_m X_{mt}\right)$
- 確率的回帰代入：$\hat{Y}_{3t} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \cdots + \hat{\beta}_m X_{mt} + \hat{\varepsilon}_t\right)$

表5：所得票の実績値及び補完後の平均所得額と所得分位値

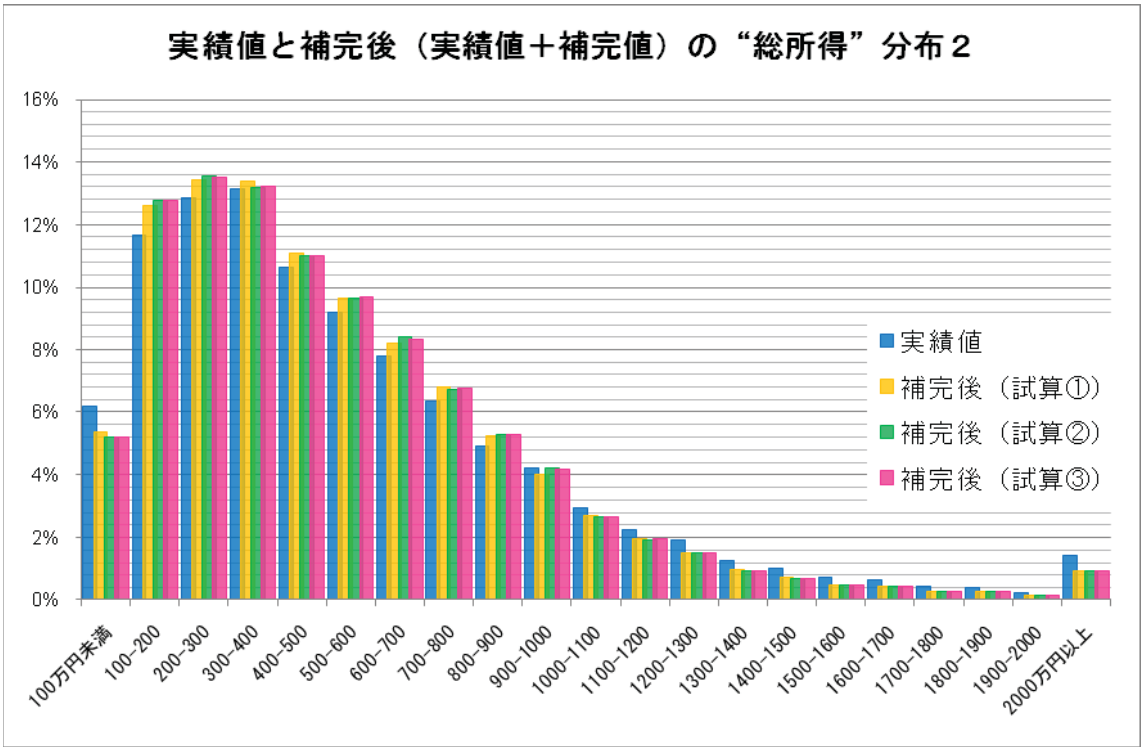
	実績値	回帰代入（不偏推定）			回帰代入（最尤推定）			確率的回帰代入		
		試算①	試算②	試算③	試算①	試算②	試算③	試算①	試算②	試算③
一世帯あたり平均所得（万円）										
全世帯	566.8	565.0	564.2	564.0	534.1	533.6	533.9	568.9	568.2	567.9
高齢者世帯	306.3	304.7	304.1	304.3	291.2	290.7	291.1	305.7	305.0	305.2
母子世帯	236.7	239.8	238.8	240.0	225.5	224.8	226.0	239.7	238.8	240.1
父子世帯	515.9	512.9	538.3	519.6	475.7	497.6	482.2	601.3	632.0	606.3
その他の世帯	646.8	636.1	635.2	634.9	600.6	600.1	600.3	640.6	639.9	639.5
所得分位値（万円）										
中央値	451	472	473	472	442	445	444	438	438	438
第1五分位	214	227	226	226	212	213	213	200	200	200
第2五分位	365	382	382	382	360	360	360	350	350	350
第3五分位	554	575	575	574	540	540	540	542	542	541
第4五分位	838	848	847	846	789	790	790	849	849	848

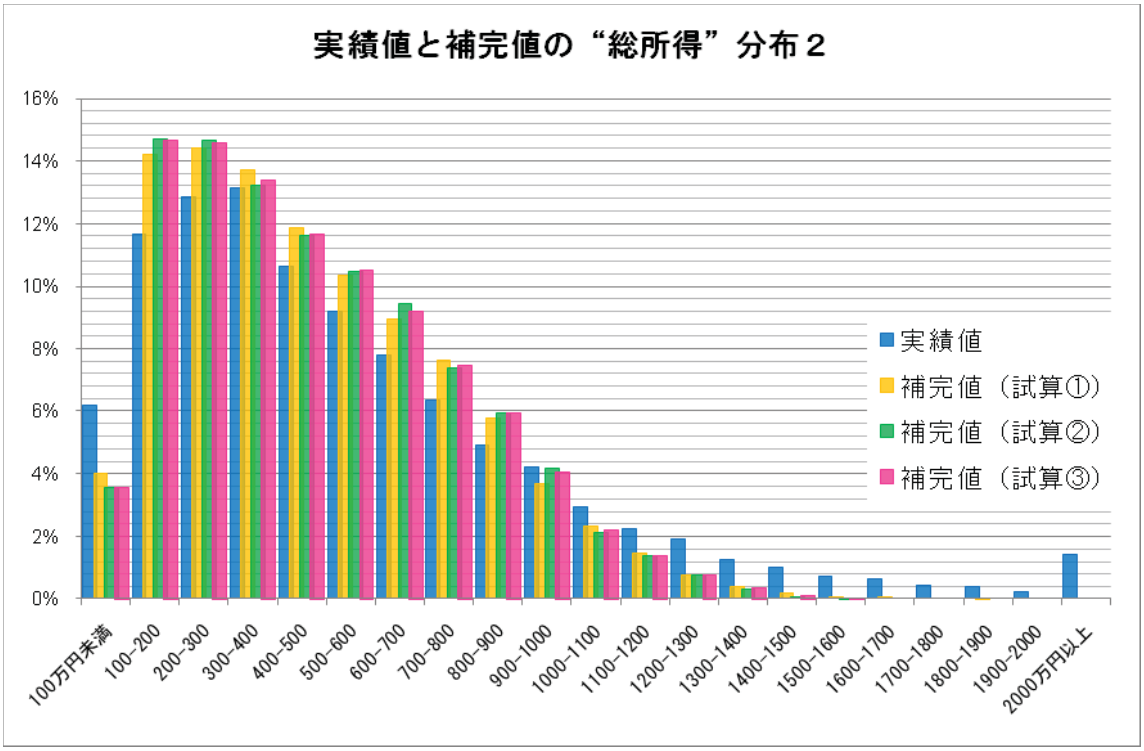
（ア）不偏推定


実績値と補完後（実績値＋補完値）の“総所得”分布１
14%
12%
10%
8%
6%
4%
2%
0%
実績値
補完後（試算①）
補完後（試算②）
補完後（試算③）
100万円未満
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
1000-1100
1100-1200
1200-1300
1300-1400
1400-1500
1500-1600
1600-1700
1700-1800
1800-1900
1900-2000
2000万円以上

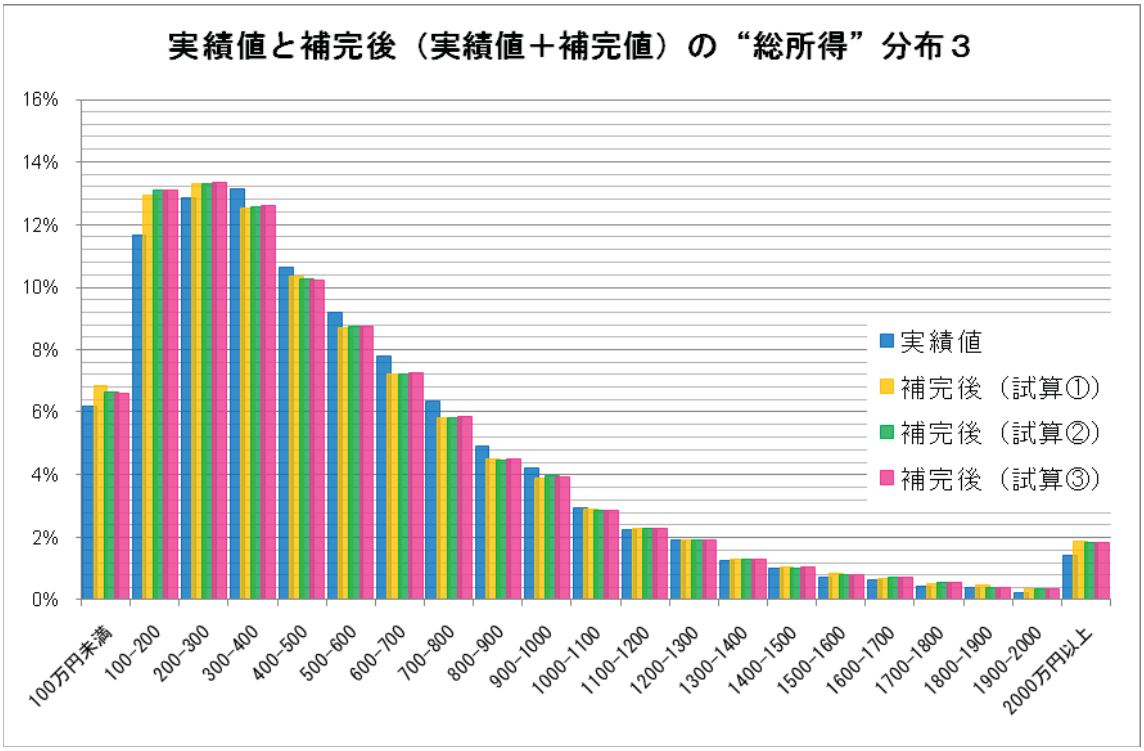

実績値と補完値の“総所得”分布１
14%
12%
10%
8%
6%
4%
2%
0%
実績値
補完値（試算①）
補完値（試算②）
補完値（試算③）
100万円未満
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
1000-1100
1100-1200
1200-1300
1300-1400
1400-1500
1500-1600
1600-1700
1700-1800
1800-1900
1900-2000
2000万円以上

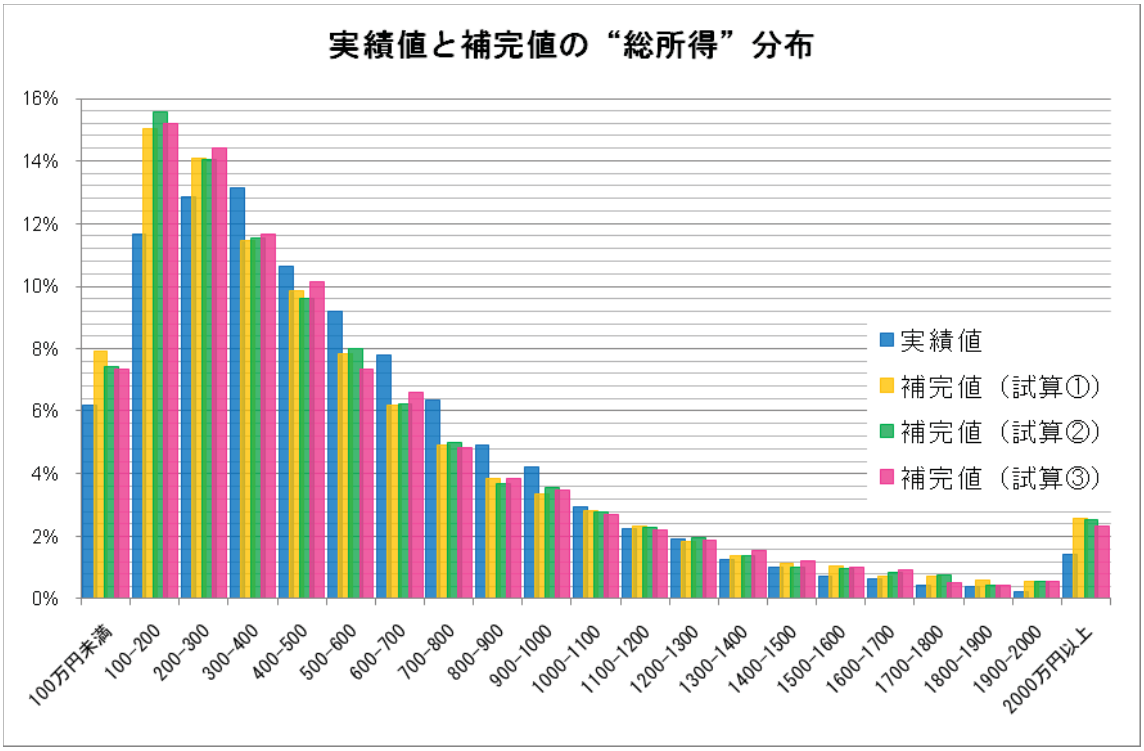
（イ）最尤推定


実績値と補完後（実績値＋補完値）の“総所得”分布２
16%
14%
12%
10%
8%
6%
4%
2%
0%
実績値
補完後（試算①）
補完後（試算②）
補完後（試算③）
100万円未満
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
1000-1100
1100-1200
1200-1300
1300-1400
1400-1500
1500-1600
1600-1700
1700-1800
1800-1900
1900-2000
2000万円以上


実績値と補完値の“総所得”分布２
16%
14%
12%
10%
8%
6%
4%
2%
0%
実績値
補完値（試算①）
補完値（試算②）
補完値（試算③）
100万円未満
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
1000-1100
1100-1200
1200-1300
1300-1400
1400-1500
1500-1600
1600-1700
1700-1800
1800-1900
1900-2000
2000万円以上

（ウ）確率的回帰代入


実績値と補完後（実績値＋補完値）の“総所得”分布３
16%
14%
12%
10%
8%
6%
4%
2%
0%
実績値
補完後（試算①）
補完後（試算②）
補完後（試算③）
100万円未満
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
1000-1100
1100-1200
1200-1300
1300-1400
1400-1500
1500-1600
1600-1700
1700-1800
1800-1900
1900-2000
2000万円以上


実績値と補完値の“総所得”分布
16%
14%
12%
10%
8%
6%
4%
2%
0%
実績値
補完値（試算①）
補完値（試算②）
補完値（試算③）
100万円未満
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
1000-1100
1100-1200
1200-1300
1300-1400
1400-1500
1500-1600
1600-1700
1700-1800
1800-1900
1900-2000
2000万円以上

5　まとめ

以上の１２通りの試算結果はいずれも一長一短があり，真の所得分布を知り得ないこともあって，各方法の優劣を判断することは困難である。また，各方法は次のような仮定に基づいているため，妥当性の評価等を含めさらなる検討が必要であり，現時点で集計に適用可能であると結論づけることはできず，政府統計への採用の判断は時期尚早である。

【各モデルの前提となる仮定】

- 所得票の非回答が総所得に依存しないランダムな欠測（Missing at random）である。
- 必要十分な共変量が適切に選択できる。
- 傾向スコアは真の回答確率に一致する。
- マッチングされた両集団の所得分布は同一である。（ノンパラメトリックな代入）
- 総所得が世帯票情報に基づく重回帰モデルで計算できる。（パラメトリックな代入）