Thesis for the degree of Master of Science

Iterative Calibration Approach to Integrated Weighting for Household Surveys

by

Sujin Kim

Department of Statistics

The Graduate School

Pukyong National University

February 22, 2019

Iterative Calibration Approach to Integrated Weighting for Household Surveys

(가구조사의 통합가중치 산출을 위한 반복적 칼리브레이션 방법)

Advisor: Prof. Inho Park

by

Sujin Kim

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in Department of Statistics, The Graduate School,
Pukyong National University

February 2019

Iterative Calibration Approach to Integrated Weighting for Household Surveys

A dissertation
by
Sujin Kim

Approved by:

_____
(Daeheung Jang)

_____
(Seongbaek Yi)

_____
(Inho Park)

February 22, 2019

Contents

List of Tables

List of Figures

가구조사의 통합가중치 산출을 위한 반복적 칼리브레이션 방법

김 수 진

부 경 대 학 교 대 학 원 통 계 학 과

요약

가구조사에서 가구 및 개인가중치는 각 수준별로 추정할 수 있도록 만들어진다. 이 때, 이미 알려져 있는 모집단에 대한 보조정보가 주어진다면 모집단의 총계와 추정치가 일치하도록 가중치를 보정하는 방법이 적용된다. 개인들이 모여 가구를 구성하는 구조를 고려하면, 두 수준의 보조정보를 함께 활용하는 경우 가구와 개인의 구조적 관계를 유지하면서 더욱 대표성 있는 가중치를 산출할 수 있다. 이를 통합가중치라고 칭하며, 가구와 개인가중치간의 개념구조는 보통 몇몇 통합의 형식으로 정해진다. 본 논문에서는 각 수준별 보조정보를 이용하여 가중치보정을 조정된 가중치가 수렴할 때까지 반복하는 통합가중치 산출 방법을 제안한다. 먼저, 개인설계가중치를 개인수준 보조변수를 이용해 보정된 개인가중치를 산출한 후 각 가구 내 보정된 개인가중치의 평균값을 가구수준 보조변수와 함께 보정된 가구가중치를 산출한다. 다음으로, 보정된 개인가중치에 조정계수를 곱하여 가구 가중치의 변화를 반영한다. 그리고 보정 전 가중치와 보정 후 가중치가 수렴할 때 까지 위의 보정과정을 반복한다. 제안하는 방법과 문헌에서 제시된 일부 방법들의 비교결과를 제시하기위해 ACS(American Community Survey)를 통한 사례연구 및 모의실험을 수행하였다.

주요용어 : 통합가중치, 반복절차, 캘리브레이션, 일반화 회귀 추정량, 가구조사.

1. Introduction

In household surveys, both household and individual weights are commonly developed to allow estimation at each level. This double objective urges the need for an integrated weighting for calibration when the auxiliary information is available at different levels together. Since households and individuals are not independent and have a hierarchical structure, we expect to produce improved weights by using a combination of household and individual auxiliary information to be used for calibration. The conceptual structure between household and individual weights is commonly imposed by adopting some form of integration.

In this paper, we propose an iterative method of integrated weighting that calibrates on the auxiliary information at each level at a time until convergence. For example, the individual weights can first be obtained by calibrating its base weights on the corresponding auxiliary variables and then their averages within households are to be calibrated to produce the household weights on the corresponding auxiliary variables. Next, the adjustment factor reflecting changes in household weights is used to recalibrate individual weights to obtain weights until weights are similar after adjusting the previously presented calibration process.

The paper is organized as follows. Section 2 reviews the definition of a calibration estimator and the distance minimizing approach to calculate the

calibrated weights. Chapter 3 introduces methods of calibration for the integrated weights in the literature (Estevao and Särndal, 2006) and presents a proposed calibration method. An empirical study and a simulation study using ACS are to be carried out to present the distribution of weights produced by each method and the bias as well as variation of these calibrated estimators for a number of survey variables of interest in Section 4. Section 5 summarizes the results of the study.

2. Weight Calibration

In this chapter, we review the calibration to adjust the weights using auxiliary information related to the survey variables of interest for the purpose of enhancing the efficiency of estimation.

2.1 Calibration Estimator

Consider a probability sample s drawn from a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$ according to a sampling design, denoted by $p(s)$. Let $\pi_i = P(i \in s)$ define the inclusion probabilities of element i. The design weights are the inverses of the inclusion probabilities $d_i = 1/\pi_i$. For estimating a population total of a survey variable that we are interested $Y = \sum_{i \in U} y_i$, the Horvitz-Thomson (HT) estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{n} d_i y_i \tag{2.1}$$

can be used (Horvitz and Thompson, 1952).

When auxiliary information x related to the survey variable y is available, the estimator can be more precise by using the relationship between x and y than HT estimator. Among the potential benefits of the calibration estimator are decrease in variances, bias correction for frame coverage, nonresponse adjustment (Valliant et al., 2013). Let $\mathbf{x}$ be a $p-$dimensional

auxiliary vector and $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ be a vector for element i. The calibration estimator is

$$\hat{Y}_{cal} = \sum_{i=1}^{n} w_i y_i, \tag{2.2}$$

where the weights w_i have been calibrated to the population total $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$ to satisfy the calibration equation

$$\sum_{i=1}^{n} w_i \mathbf{x}_i = \mathbf{X} \tag{2.3}$$

and the calibration weights are as close as possible to the initial weights. The calibrated weights close to the initial weights can be obtained by defining a distance measure.

2.2 Distance Minimization Approach

The distance minimization approach was proposed by Deville and Särndal (1992). A unique set of calibrated weights is obtained by minimizing the distance measure between the initial weights and the calibrated weights, subject to the calibration constraint (2.3).

For element i, we consider the distance function $G_i(w, d)$ is nonnegative, strictly convex and twice continuously differentiable with respect to w for every fixed $d > 0$ such that $G_i(d, d) = 0$ and $G_i'(d, d) = 0$. Let $E_p(\cdot)$ denote expectation with respect to the sample design $p(s)$. To minimize $E_p(\sum_s G_i(w_i, d_i))$ subject to (2.3) for all s equals to minimize $\sum_s G_i(w_i, d_i)$

subject to (2.3) for any particular s. Minimizing $\sum_s G_i(w_i, d_i)$ subject to (2.3) gives

$$g_i(w_i, d_i) - \mathbf{x}_i'\lambda = 0$$

with $g_i(w, d) = \partial G_i(w, d)/\partial w$ and a vector of Lagrange multipliers λ. Let $F(\cdot)$ denote the inverse function of $g(\cdot)$ and assume that a unique solution exists, then calibrated weights

$$w_i = d_i F_i(\mathbf{x}_i'\hat{\lambda}), \tag{2.4}$$

where $\hat{\lambda}$ obtained as a solution from (2.3).

Various distance functions are available for finding new weights (Deville and Särndal, 1992). For example, if we define a distance function $G_i(w, d)$ as

$$G_i(w, d) = (w_i - d_i)^2/2d_i q_i, \tag{2.5}$$

it gives $q_i g_i(w_i, d_i) = w_i/d_i - 1$ and $F(q_i\mathbf{x}_i'\lambda) = 1 + q_i\mathbf{x}_i'\lambda$. The resulting weights from the defined distance function is the form of

$$w_i = d_i(1 + q_i\mathbf{x}_i'\lambda) \tag{2.6}$$

with $\lambda' = (\mathbf{X} - \hat{\mathbf{X}})'(\sum_s d_i q_i \mathbf{x}_i\mathbf{x}_i')^{-1}$. In this case, the calibrated weights can be positive or negative. Another distance function is

$$G_i(w, d) = \{w_i log(w_i/d_i) - w_i + d_i\}/q_i \tag{2.7}$$

with positive constants q_i. It gives $q_i g_i(w_i, d_i) = log(w_i/d_i)$ and $F(q_i\mathbf{x}_i'\lambda) = exp(q_i\mathbf{x}_i'\lambda)$ and the resulting weights known as the exponential weights are

$$w_i = d_i exp(q_i\mathbf{x}_i'\hat{\lambda}), \tag{2.8}$$

where the solution of (2.3) is $\hat{\lambda}$ obtained by iterative methods. The resulting weights in (2.8) can have extremely large value.

2.3 The Generalized Regression Estimator

The generalized regression (GREG) estimator concept was presented by Cassel et al. (1976). The estimator of $Y = \sum_{i \in U} y_i$ estimated by weights in (2.6) can be expressed in the form of the GREG estimator

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{B}} \tag{2.9}$$

with $\hat{\mathbf{X}} = \sum_s d_i \mathbf{x_i}$ and $\hat{\mathbf{B}} = (\sum_s d_i q_i \mathbf{x}_i \mathbf{x}_i')^{-1}(\sum_s d_i q_i \mathbf{x}_i y_i)$. A property of $\hat{Y}_{GREG}$ is design unbiased (Deville and Särndal, 1992). We derive The GREG estimator with the distance measurement approach, but it can also be derived with the model-assisted approach (Särndal, Swensson and Wretman, 1992).

In Deville and Särndal (1992), under the mild condisions on $F_i(\mathbf{x}_i'\lambda)$, all estimators estimated by weights getting from distance measure approach are asymptotically equivalent to the GREG estimator in (2.9). They also have the same asymptotic variance of $\hat{Y}_{GREG}$

$$AV(\hat{Y}_{GREG}) = \sum \sum_U (\pi_{ij} - \pi_i \pi_j) d_i E_i d_j E_j, \tag{2.10}$$

where π_{ij} is the joint inclusion probability of i and j and $E_i = y_i - \mathbf{x}_i'\mathbf{B}$ with $\mathbf{B}$ satisfying the normal equation $(\sum_U q_i \mathbf{x}\mathbf{x}')\mathbf{B} = \sum_U q_i \mathbf{x}_i y_i$. The variance estimator is given by

$$\hat{V}(\hat{Y}_{GREG}) = \sum \sum_s (\pi_{ij} - \pi_i \pi_j) w_i e_i w_j e_j / \pi_{ij}, \tag{2.11}$$

where sample-based residuals is $e_i = y_i - \mathbf{x}_i'\hat{\mathbf{B}}$ and $\hat{\mathbf{B}}$ are the values satisfying the sample-based normal equation $(\sum_s w_i q_i \mathbf{x_i x_i}')\hat{\mathbf{B}} = \sum_s w_i q_i \mathbf{x}_i y_i$. (2.11) is a design-consistent variance estimator and nearly model-unbiased for the model mean squared error.

3. Integrated Weighting

Many household surveys set the objective to produce both household estimates and individual estimates. To allow consistent estimation at both levels, integrated weighting is often used in many national statistical offices such as Eurostat, Statistics Canata, Statistics New Zealand, etc. Integrated weighting is computed using auxiliary information at both individual and household levels in some composite fashions. A common approach to integrated weighting is to give all members of a selected household equal weight, which is a weight also used for producing household statistics. In this chapter, we first briefly review integrated weighting methods for two-stage sampling in the literature (*e.g.*, Estevao and Särndal, 2006) and propose a new method based upon an iterative calibration approach.

3.1 Calibrated Weights in Two-Stage Sampling

Consider the population of clusters $U_I = \{1, 2, \ldots, j, \ldots, N_I\}$ of size N_I and the population of units $U = \{1, 2, \ldots, k, \ldots, N\}$ of size N. In two-stage sampling, a sample of clusters s_I is first drawn from U_I with inclusion probabilities π_j for $j \in U_I$. Units within each of selected clusters are then sampled with inclusion probabilities $\pi_{k|j}$ for $k \in U_j$. Then the design weights are defined by $d_j = 1/\pi_j$ for cluster j and $d_k = d_j d_{k|j}$ for unit k.

Let $y_{(c)}$ denote a survey variable at cluster level and let $y_{(c)j}$ denote its value for cluster j. Also, let $y_{(u)}$ denote a survey variable at unit level and let $y_{(u)k}$ denote its value for unit k. Similarly, let $\mathbf{x}_{(c)j}$ and $\mathbf{x}_{(u)k}$ denote the auxiliary vector values for cluster j and unit k, respectively.

Assume that the survey objective is to estimate both cluster total $Y_{(c)} = \sum_{U_I} y_{(c)j}$ and unit total $Y_{(u)} = \sum_U y_{(u)k}$. Then, calibration estimators are given as $\hat{Y}_{I,cal} = \sum_{s_I} w_{Ij} y_j$ for household statistics and $\hat{Y}_{cal} = \sum_s w_k y_k$ for individual statistics, respectively, with household weights w_{Ij} satisfying the calibration equation

$$\sum_{s_I} w_{Ij} \mathbf{x}_{(c)j} = \sum_{U_I} \mathbf{x}_{(c)j} \tag{3.1}$$

and individual weights w_k satisfying the calibration equation

$$\sum_{s} w_k \mathbf{x}_{(u)k} = \sum_{U} \mathbf{x}_{(u)k}. \tag{3.2}$$

For integrated weighting, two options may be considered to maintain a structural relationship between a household and persons.

(1) $\sum_{s_j} w_k = N_j w_j$ for every selected household j of size N_j.

(2) $w_k = d_{k|j} w_j$ for every selected person k in household $j \in s_I$.

Under the option (1), household weights and individual weights produce the same estimated total for each household. For a one-stage cluster sampling where all individuals are selected, the average of the individual weights within

the household is set as the household weight. Under the option (2), individual weights are calculated in a manner similar to how individual weights are computed as the product of household weights and person's conditional weights.

We consider the following four methods of computing weights of the calibration estimators (Estevao and Särndal, 2006).

(i) Non-integrated calibration

Calculate the calibrated weights for each level using the corresponding auxiliary information. That is, from d_j, compute household weight w_j calibrated to satisfy the constraint (3.1). Similarly, from $d_k = d_j d_{k|j}$, compute individual weights w_k calibrated to satisfy the constraint (3.2).

(ii) Single step calibration with integration option (1)

Conduct a person-level calibration by combining the auxiliary vector into the person level by personalizing the household value. Assign the divided the auxiliary value $\mathbf{x}_{(c)k} = \mathbf{x}_{(c)j}/N_j$ on selected household by number of people in the same household and define the stacked auxiliary vector by $\mathbf{x}_{(cu)k} = \begin{pmatrix} \mathbf{x}_{(c)k} \\ \mathbf{x}_{(u)k} \end{pmatrix}$. From the individual input weights $d_k = d_j d_{k|j}$, calculate individual weights w_k satisfying the constraint $\sum_s w_k \mathbf{x}_{(cu)k} = \sum_U \mathbf{x}_{(cu)k}$. Then, compute the household weights as $w_j = \sum_{s_i} w_k / N_j$.

(iii) Single step calibration with integration option (2)

Conduct a household-level calibration by combining the household and

person's auxiliary vector into the household level. Define the stacked auxiliary vector by $\mathbf{x}_{(cu)j} = \begin{pmatrix} \mathbf{x}_{(c)j} \\ \hat{\mathbf{x}}_{(u)j} \end{pmatrix}$, where $\hat{\mathbf{x}}_{(u)j} = \sum_{s_i} d_{k|j}\mathbf{x}_{(u)k}$ is the unbiased estimator of the household total $\mathbf{x}_{(u)j} = \sum_{U_i} \mathbf{x}_{(u)k}$. From d_j, calculate household weights w_j to satisfy the constraint $\sum_{s_i} w_j \mathbf{x}_{(cu)j} = \sum_{U_i} \mathbf{x}_{(cu)j}$. Then, compute the individual weights as $w_k = d_{k|j} w_j$.

(iv) Two step calibration with integration option (1)

In step one, compute household weights w_j from d_j calibrated to the household information to satisfy the constraint (3.1). In step two, calculate the individual weights w_k from $d_k = d_j d_{k|j}$, calibrated to satisfy (3.2) such that $\sum_{s_i} w_k = N_j w_j$ for every $j \in s_I$.

Weights by methods (i) and (iv) for each level satisfy only the calibration constraint of the corresponding level, but those by method (iv) are different in that they are created based upon the options (1). On the other hand, those by methods (ii) and (iii) satisfy both calibration constraints (3.1) and (3.2). Method (ii) and (iv) both differ in that they satisfy the integration option (1), but (iv) uses personalized household variables $x_{(c)k}$ rather than true household variables $x_{(c)j}$. Method (iii) uses the strict integration option (2). In one stage sampling, every household member get the same conditional weights $d_{k|j} = 1$, which implies $w_k = w_j$. In method (iv), to keep the constraint (3.1) is so stringent that the variation in the individual weights can be significantly increased, and for a one-person household survey there is a

problem where individual weights that meet the calibration equation (3.2) and the option (1) cannot be obtained.

3.2 Proposed Method

In this section, we propose a new integrated weighting method, which adjusts weights for each level in a way to retain the multivariate relationship among the auxiliary information.

Method (ii) and (iii) presented in Section 3.1 carry out a single step calibration using household and individual auxiliary information at the same time. However, when applying a raking ratio adjustment instead of a GREG as a calibration method, the sum of the marginal distributions should be the same. Therefore, we need to reconstruct the control total and know the joint distribution of households and individuals in this process. In general, joint distribution of household and individual auxiliary information may not be availiable to the public. We devised a method that can be applied even when the joint distribution of axuiliary information at both levels is not available. Also, it utilizes true values $x_{(c)j}$ and $x_{(u)k}$, not redefined values. In method (iv), the individual weights increase in variability because they attempt to satisfy option (1) at once. The proposed method satisfies option (1) through the iterative process and calculates the household weights that reflect the adjustments at the individual level using the adjustment factors.

Firstly, starting from initial individual weights $w_k^{(r-1)} = d_k$, compute intermediate individual calibrated weights $a_k^{(r)}$ to satisfy (3.2), where $r = 1$. Next, take the average of intermediate individual weights as the intermediate

household weights denoted by $a_j^{(r)} = N_j^{-1} \sum_{s_j} a_k^{(r)}$ and compute household calibrated weights w_j^r to satisfy (3.1). Finally, obtain the initial individual weights $w_k^{(r)} = a_k^{(r)} c_j^{(r)} = w_j^{(r)} / a_j^r$, where $c_j^{(r)}$ are adjustment factors for the r-th iteration, Let $r = r + 1$ and repeat the aforementioned steps until $w_j^{(r)}$ and $r_k^{(r)}$ converge.

For simplicity's sake, let $d \overset{x}{\Longrightarrow} w$ to denote the calibration process of finding a new set of calibrated weights $w = \{w_i, i \in s\}$ that are near the initial weights $d = \{d_i \in s\}$ subject to the calibration equation with an auxiliary vector $\mathbf{x}$. The proposed process can be expressed as follows:

(1) Let $r = 1$ and let $w_k^{(r-1)} = d_k$.

(2) Do $w_k^{(r-1)} \overset{x_{(u)k}}{\Longrightarrow} a_k^{(r)}$.

(3) Let $a_j^{(r)} = N_j^{-1} \sum_{s_j} a_k^{(r)}$.

(4) Do $a_j^{(r)} \overset{x_{(c)j}}{\Longrightarrow} w_j^{(r)}$.

(5) Set $w_k^{(r)} = a_k^{(r)} c_j^{(r)}$ with adjustment factors $c_j^{(r)} = w_j^{(r)} / a_j^{(r)}$.

(6) Let $r = r + 1$ and repeat steps (2)-(5)
until $w_j^{(r)}$ and $w_k^{(r)}$ converge.

4. Application to ACS

4.1 Empirical Study

An empirical study was conducted using 2012 American Community Survey (ACS) data from the IPUMS.org website to evaluate the performance of the proposed integrated weighting method in comparison with other methods in the literature. Basic data settings refer to the paper by Kolenikov and Hammer (2015). The ACS survey data, conducted by the United States Census Bureau, produces the detailed population and houshold information. The data is comprised of 2,294,898 adults over the age of 18 in 1,207,415 households. We took this data into account as a population.

A description of variables used in the analysis is as follows Table 1.

A sample was drawn from the data under a sampling design to include 5,000 households randomly selected and all adults theirin. To produce nonresponse, sequential logistic response models with coefficients as listed in Table 2 are assumed so that if a household did not respond, individuals in the household did not respond, and if all individuals did not respond, the household did not either. As a result, we got 3,368 respondents in 2,474 households.

Table 3 describes the distribution of population and sample by some variables. There is an imbalance between the population and the sample that cannot be ignored. Male and the age group 2 had fewer responses and

Table 1. Description of variables for ACS data

Level	Variable	Description
Household	HHSZ	Household size (1 : one-person household, 2: two-person household, 3: three-person household, 4: household with more than four people)
	HHINCOME	Household income (Continous)
	HHIC	Household income (1: under 20,000, 2: 20,000 to under 40,000, 3: 40,000 to under 65,000, 4: 65,000 under 100,000, 5: 100,000 and above)
	HISPRE	Hispanic present (1: present, 2: not present)
	LINGISOL	Linguistically isolated (1: not linguistically isolated, 2: linguistically isolated)
Individual	SEX	Sex (1: male, 2: female)
	AGE	Age (1: 18-29, 2: 30-44, 3: 45-54, 4:55-64, 5: 65 and above)
	RACE	Race (1: white only, 2: black/african americal only, 3: other)
	EDUC	Educational attainment (1: below high school, 2: high school/ general education deploma, 3: some college/associate degree, 4: bachelor's degree, 5: graduate/professional degree)
	MARST	Marital status (1: married, 2: not married)
	EMPSTAT	Employment (1: employed, 2: unemployed, 3: not in labor force)
	INCTOT	Individual income (Continous)

Table 2. The response model: P(response)=$(1 + e^{x'\beta})^{-1}$

a) The household response model

Variable	Category & transformation	Coefficient
HHSZ	1	-0.5
HHSZ	3 and 4	-0.7
HHINCOME	ln(HHINCOME+20,000)	0.1

b) The individual response model

Variable	Category	Coefficient
SEX	1	-0.2
AGE	2	-0.5
RACE	1	0.25
EDUC	1	-0.4
EDUC	4	0.1
EDUC	5	0.3

the race group 1 responded well. We assumed a model in which non-response occurs frequently when the household size group is 3 or 4. However, in the case of single-person household, the non-response of the household occurred if the individual did not respond, resulting in a large number of non-response of the one-person household, and in the end, the households with two member answered the most. We use the categorical variables (Household size, Sex, Age, Race and Education) and the continuous variable (Household income) as auxiliary variables.

Generalized regression(GREG) weighting were adopted for calibration for each of the methods on the consideration given,

W1 : Non-integrated calibration

W2 : Single step calibration with integration option (1)

Table 3. Distribution of population and sample by variables

Variable	Category	Population total	Population %	Sample count	Sample %
Households		1207415	100.00%	2474	100.00%
Household size	1	388470	32.17%	531	21.46%
	2	629353	52.12%	1516	61.28%
	3	131801	10.92%	293	11.84%
	4	57791	4.79%	134	5.42%
Total household income			$86,277,024,521		$191,813,775
Hispanic present		145173	12.02%	264	10.67%
Linguistically isolated		47061	3.90%	106	4.28%
Individuals		2294898	100.00%	3368	100.00%
SEX	1	1085531	47.30%	1511	44.86%
	2	1209367	52.70%	1857	55.14%
AGE	1	395250	17.22%	547	16.24%
	2	528792	23.04%	678	20.13%
	3	437672	19.07%	648	19.24%
	4	428807	18.69%	673	19.98%
	5	504377	21.98%	822	24.41%
RACE	1	1814707	79.08%	2786	82.72%
	2	227826	9.93%	288	8.55%
	3	252365	11.00%	294	8.73%
EDUC	1	299730	13.06%	370	10.99%
	2	656608	28.61%	1006	29.87%
	3	697947	30.41%	985	29.25%
	4	399943	17.43%	600	17.81%
	5	240670	10.49%	407	12.08%
Married		1297358	56.53%	2015	59.83%
Employment	1	1342689	58.51%	1960	58.19%
	2	122905	5.36%	163	4.84%
	3	829304	36.14%	1245	36.97%
Total individual income			$86,161,238,287		$129,039,467

W3 : Single step calibration with integration option (2)

W4 : Two step calibration with integration option (1)

W5 : Proposed calibration

In Figure 1, The distribution of the adjustment factors by iteration for W5 is illustrated because the weights produced by the proposed method are calculated through iteration. The figure shows that the values of the adjustment factor converge to 1. This means that the degree of variation in household weights by calibration is reduced by iteration.

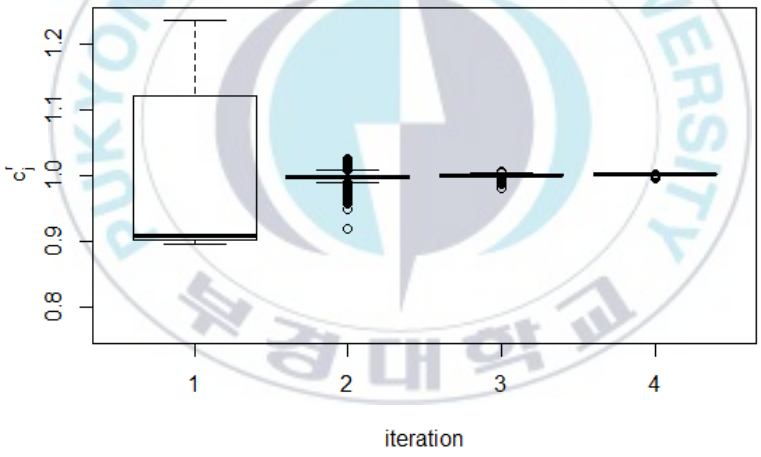
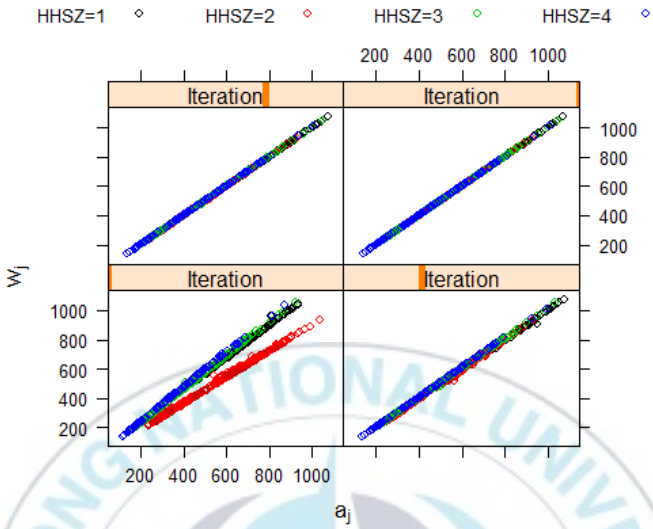
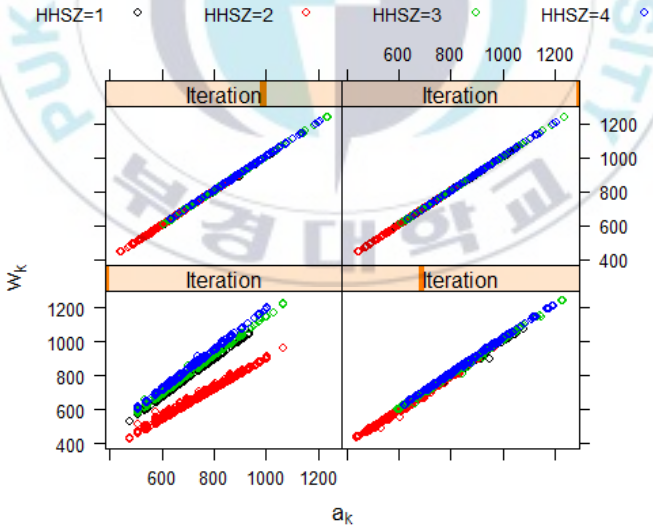


Figure 1. Box plot of adjustment factor by iteration

In Figure 2, each panel is divided by iteration and each point is colored by the household size(HHSZ) with a categorical variable. The figure shows that the intermediate weights and the calibrated weights become similar by iteration and household non-response by household size is also adjusted by repeated weighting.



a) Changes in household weights by iteration



b) Changes in individual weights by iteration

Figure 2. Scatter plot of $a^{(r)}$ vs $w^{(r)}$ by iteration

Table 4 compares the distribution of weights calculated for each calibration method. The spread of weights can be assessed with simple descriptive statistics min, man and max. The variation in weights can be also evaluated through the amount of variance increase due to unequal weighting $L_w = 1 + cv_w^2$, where cv_w is the coefficient of variation in weights and the sum of distance measurement $G(w, d)$ in (2.5). At the household level, the sum of weights is consistent with the number of households in the population and the average weight of all households is equal to 488.0. Table 4 shows that the variation of W1 and W4 which are identical is the smallest because they do not consider individual characteristics in the calibration process. Among W2, W3, W5 taking into account both household and individual characteristics as auxiliary variables for calibration, W3 has the largest variation and W2 and W5 distribute similarly. In the W3 method, 25% of the weights is negative. At the individual level, the sum of W4 and W5 is slightly inconsistent with the number of individuals. This inconsistency can be corrected with rescaling. W3 and W4 have some negative weights. In terms of the variation in weights, W1 using the least variable in weighting has the smallest variation, and W3 has the largest variation among integrated weights. On the other hand, the W2 and W5 can be found to have relatively small variations.

Figure 3 is a density plot of the distribution of household weights according to each calibration method. W2, W3, W5 are more widely distributed than W1 and W4, especially W3 is heavy-tailed. W2 and W5 in particular are similarly distributed. Figure 4 is a density plot of the distribution of individual weights according to each calibration method. W3 is distributed in a similar form to household weights. W1, W2 and W5 are similarly adjusted,

Table 4. Summary statistics of weights

Level	Method	Mean	Sum	Min	Q1	Q2	Q3	Max	Std	Deff	G(d,w)
House	W1	488.0	1207415	341.0	415.7	418.9	452.9	734.2	127.9	1.07	7.90
-hold	W2	488.0	1207415	127.3	298.8	490.9	649.3	1031.6	197.1	1.16	10.21
	W3	488.0	1207415	-1142.0	-22.3	614.8	938.1	5134.2	603.4	2.53	43.52
	W4	488.0	1207415	341.0	415.7	418.9	452.9	734.2	127.9	1.07	7.90
	W5	488.0	1207415	139.5	300.1	488.1	644.1	1075.2	197.6	1.16	10.23
Indivi	W1	681.4	2294898	474.6	605.5	671.8	743.2	1068.7	105.8	1.02	28.55
-dual	W2	681.4	2294898	431.2	571.6	668.6	768.0	1180.0	135.7	1.04	29.56
	W3	681.4	2294898	-1142.0	60.3	740.2	1082.9	5134.2	688.7	2.02	93.12
	W4	682.4	2298233	-2231.0	501.8	732.9	834.2	3319.4	389.1	1.33	48.22
	W5	682.1	2297367	443.9	576.9	661.8	761.9	1238.4	136.3	1.04	29.67

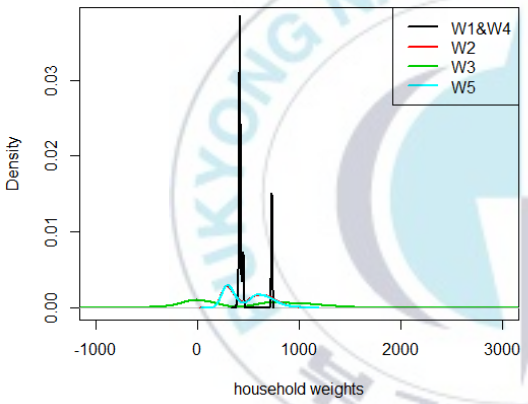


Figure 3. Density plot of household weights by methods

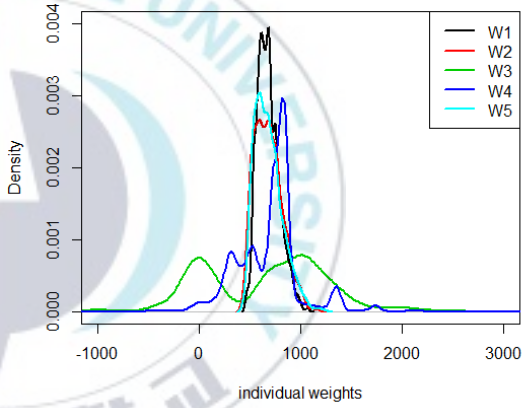


Figure 4. Density plot of individual weights by methods

especially 2 and 5 are more similar. W4 becomes more diverse for individual characteristics under option (1).

Table 5 shows estimated totals and percentages of household variables considered as auxiliary information. W1 to W5 estimate the variable equal to the target population value.

As shown in Table 6, estimated totals and percentages based on W1, W2

Table 5. Estimated total and percentage of household variables

Variable	Category	Population total	Population %	[W1 W2 W3 W4 W5] Weighted count	Weighted %
HHSZ	1	388470	32.17%	388470	32.17%
	2	629353	52.12%	629353	52.12%
	3	131801	10.92%	131801	10.92%
	4	57791	4.79%	57791	4.79%
Total household income			$86,277,024,521		$86,277,024,521

and W3 are equal to the target population. W4 is slightly different in total and percentage. W5 is different in total but the percentage is the same. As mentioned before, the slight inconsistency in total estimates based on W5 can be resolved through rescaling.

Table 6. Estimated total and percentage of individual variables

				[W1 W2 W3]		[W4]		[W5]	
Variable	Category	Population total	Population %	Weighted count	Weighted %	Weighted count	Weighted %	Weighted count	Weighted %
SEX	1	1085531	47.30%	1085531	47.30%	1088866	47.38%	1086699	47.30%
	2	1209367	52.70%	1209367	52.70%	1209367	52.62%	1210668	52.70%
AGE	1	395250	17.22%	395250	17.22%	398585	17.34%	395675	17.22%
	2	528792	23.04%	528792	23.04%	528792	23.01%	529361	23.04%
	3	437672	19.07%	437672	19.07%	437672	19.04%	438143	19.07%
	4	428807	18.69%	428807	18.69%	428807	18.66%	429268	18.69%
	5	504377	21.98%	504377	21.98%	504377	21.95%	504920	21.98%
RACE	1	1814707	79.08%	1814707	79.08%	1818042	79.11%	1816659	79.08%
	2	227826	9.93%	227826	9.93%	227826	9.91%	228071	9.93%
	3	252365	11.00%	252365	11.00%	252365	10.98%	252637	11.00%
EDUC	1	299730	13.06%	299730	13.06%	303065	13.19%	300052	13.06%
	2	656608	28.61%	656608	28.61%	656608	28.57%	657314	28.61%
	3	697947	30.41%	697947	30.41%	697947	30.37%	698698	30.41%
	4	399943	17.43%	399943	17.43%	399943	17.40%	400373	17.43%
	5	240670	10.49%	240670	10.49%	240670	10.47%	240929	10.49%

4.2 Simulation Study

We conducted the simulation study to assess the performance of the estimators based on each weight. Under the same sampling design as the empirical study, we selected 200 household samples of size 5000 each. we assumed the same response model in Table 2.

The performance of the estimators was examined by the Chi-squared statistic X^2

$$\mathrm{X}^2 = \sum_{g=1}^{G} \frac{(\hat{Y}_g - Y_g)^2}{Y_g},$$

where $\hat{Y}_g$ is the estimator for categorical variable with the number of groups g.

Table 7 shows the distribution of the Chi-squared statistics of the estimators based on w_k. The estimators are for combinations of variables using an auxiliary variable. This is the result of the individual level estimation for the combination of the household variable HHSZ and the individual variables SEX, AGE, RACE and EDUC. Overall, W5 is good in terms of the adequacy of the total estimators and the W2 as well. W4 also estimates better than W1 which produced by the non-integration method. The estimators based on W3 are the most unstable except for SEX.

Table 8 reports X^2 of the individual estimators for combination of the household variable HHIC and the individual variables. W2 and W5 are also good when assessing the adequacy of the total estimators. In case of W3, it can be seen that the variation in the Chi-squared statistics is greatest.

Figure 5 shows a boxplot of X^2 presented in Tables 7 and 8.

Table 7. X^2 of the individual estimators by combinations of auxiliary variables (1)

Variable		Statistic	Method				
Household	Individual	X^2	W1	W2	W3	W4	W5
HHSZ	SEX	Mean	23650	2044	12000	4312	2017
		Min	3456	24	706	80	79
		Q1	17020	882	6347	2066	909
		Q2	22780	1571	9960	3511	1578
		Q3	30200	2622	16120	5931	2630
		Max	50990	8263	48910	16650	8331
	AGE	Mean	31310	10530	33890	27890	10030
		Min	10260	2537	5951	5325	2835
		Q1	25030	7002	23080	18380	6896
		Q2	30300	9976	31000	26230	9389
		Q3	37610	13620	41050	35440	12870
		Max	65040	29920	91420	86750	25850
	RACE	Mean	29930	8891	35260	10790	8864
		Min	10510	557	6377	876	476
		Q1	22890	4514	22480	5885	4487
		Q2	28610	7820	31500	9583	7385
		Q3	36820	12450	44500	13870	12110
		Max	57680	26340	153600	35150	24630
	EDUC	Mean	31450	11080	35460	19040	10790
		Min	8189	1789	5316	4363	1395
		Q1	24640	7534	24220	13890	7669
		Q2	31190	10780	32700	17810	10230
		Q3	37640	13820	43660	23190	13180
		Max	57700	29340	96040	49640	27540

Table 8. X^2 of the individual estimators by combinations of auxiliary variables (2)

Variable		Statistic	Method				
Household	Individual	X^2	W1	W2	W3	W4	W5
HHIC	SEX	Mean	6931	6373	16240	8134	6376
		Min	1126	809	1775	1576	858
		Q1	4441	4190	9232	5702	4097
		Q2	6257	5836	14020	7741	5913
		Q3	9070	8143	19950	9884	8148
		Max	20990	17660	53910	27010	17380
	AGE	Mean	18710	17680	43900	21440	17620
		Min	6641	5465	14840	7646	5356
		Q1	15010	13960	31050	16530	13890
		Q2	18350	17240	41550	20690	17280
		Q3	21750	20880	55230	24820	20540
		Max	34970	36620	90220	42460	35980
	RACE	Mean	15320	14860	43990	15890	14980
		Min	3655	3561	15110	2777	3730
		Q1	10760	10170	31220	11020	10700
		Q2	14620	14360	41180	14790	14550
		Q3	18950	18350	53310	19110	18660
		Max	37560	36610	115000	40490	36030
	EDUC	Mean	18020	17400	44430	23060	17400
		Min	4236	4148	13720	7250	3281
		Q1	13390	12810	32450	17080	12910
		Q2	17100	16190	42660	22150	16050
		Q3	22070	21990	54180	27020	21760
		Max	37240	39950	100300	62140	39940

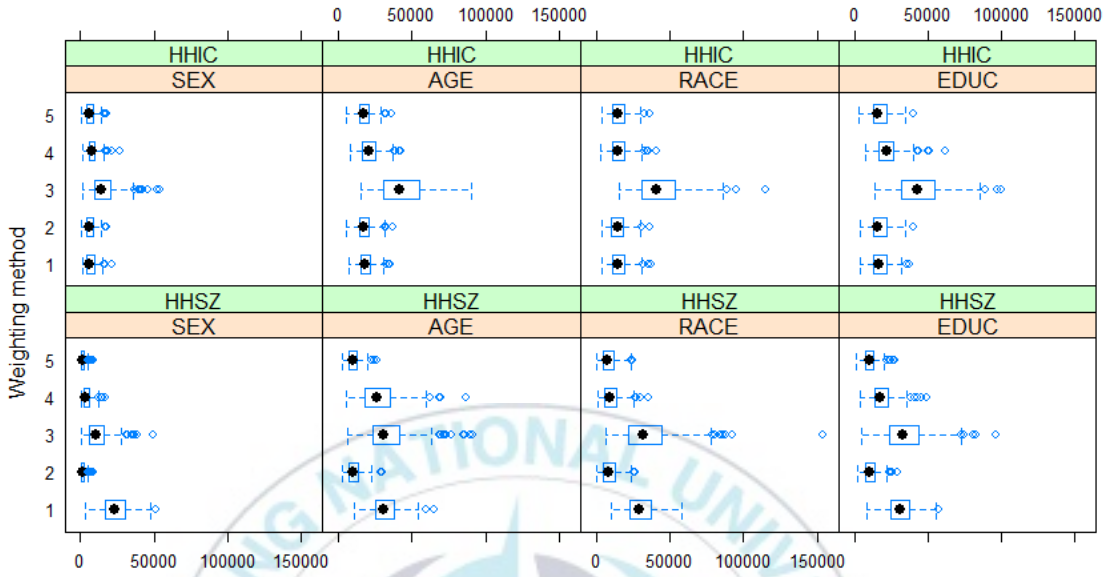


Figure 5. Box plot of X^2 of the individual estimators

We examined the relative bias(RB)(%) and the relative root mean squared errors(RRMSE)(%) for comparisons of each method for variables not used as auxiliary information.

- $\mathrm{RB}(\%) = (E(\hat{Y}) - Y)/Y \times 100$; relative bias
- $MSE = E[(\hat{Y} - Y)^2] = V(\hat{Y}) + Bias^2$; mean squared error
- $RMSE = \sqrt{MSE}$; root mean squared error
- $RRMSE(\%) = RMSE/Y \times 100$; relative root mean squared error

In addition, the distribution of estimators for each method is expressed by the box plot.

Table 9. Relative bias and root mean squared error of estimators for household level variables

Variable	Statistic	W1	W2	W3	W4	W5
Hispanic present	RB(%)	-14.5	-7.6	1.7	-14.5	-7.7
	RRMSE(%)	15.30	9.25	8.72	15.30	9.27
Linguistically isolated	RB(%)	-9.1	-0.6	3.0	-9.1	-0.7
	RRMSE(%)	13.83	11.28	15.90	13.83	11.35

Table 9 reports the relative bias and the relative root mean squared error of the household level estimators. For Hispanic present, W3 has the least RB and RRMSE, but the variation of the estimators is somewhat large. In case W1 and W4, there is the overall underestimation. W2 and W5 are slightly underestimated and have a moderate variation than W3. For Lilnguistically isolated, RB is the largest for W1 and W4. W3 has the largest RRMSE, W2 has the smallest RRMSE, and W5 is the next smallest.

In the same manner, Figure 6 and 7 show the distribution of each estimator by methods. The horizontal line of the box plot means the known population value.

Table 10 reports the relative bias and the relative root mean squared error of the individual level estimators. W5 is the smallest RB when Married and Total individual income are estimated, and RRMSE is the smallest when Married, Employed, and Total individual income are estimated. W3 is the smallest RB for Employed, Unemployed, and Not in labor force, but the RRMSE is large overall because the variance of the estimators is large. W2 has the smallest RRMSE in Unemployed, Not in labor force.

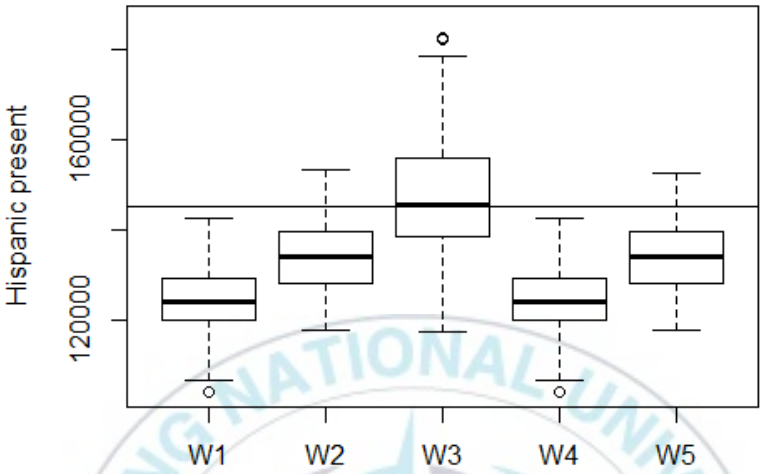


Figure 6. Box plot of the estimators (Hispanic present) by methods

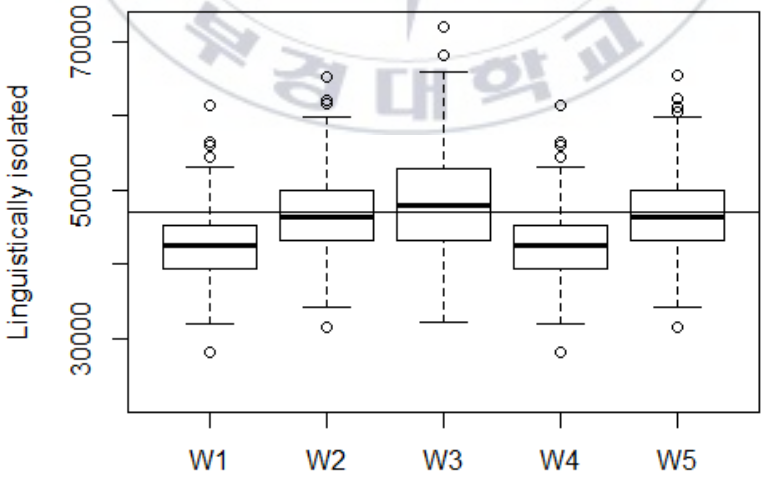


Figure 7. Box plot of the estimators (Linguistically isolated) by methods

Table 10. Relative bias and root mean squared error of estimators for individual level variables

Variable	Statistic	W1	W2	W3	W4	W5
Married	RB(%)	2.95	0.25	-0.46	-0.25	0.24
	RRMSE(%)	3.35	1.39	2.26	1.45	1.39
Employment	RB(%)	0.41	0.20	0.06	-0.27	0.17
- Employed	RRMSE(%)	1.39	1.32	1.90	1.62	1.32
Employment	RB(%)	-2.50	-1.07	-0.33	0.34	-1.02
- Unemployed	RRMSE(%)	8.48	8.34	12.29	9.30	8.40
Employment	RB(%)	-0.30	-0.16	-0.05	0.49	-0.14
- Not in labor force	RRMSE(%)	2.03	1.98	2.83	2.40	2.00
Total individual income	RB(%)	1.50	0.04	0.11	-1.43	-0.01
	RRMSE(%)	2.61	1.06	1.35	2.26	1.04

In Figure 8, W2, W4 and W5 deliver quite stable estimators in the sense that RB and RRMSE are small in estimating the Married variable. The estimators using W3 have the greatest variation and W1 tends to overestimate.

In Figure 9, the estimators based on W2 and W5 are similarly calculated. They have the small variance and the bias. For the estimators based on W3, the bias is the smallest, but the variance is the greatest.

In Figure 10, W3 and W4 have small RB, but the variation of estimators is greater than the others. W2 and W5 also produce similar estimators.

In Figure 11, The estimators based on W3 produce the best results considering RB, although the range of estimators is wide. The next best results are produced by W5, W2, W1 and W4.

The Figure 12 are presented that W2 and W5 well estimate the variable (Total individual income). The estimators based on W1 tend to be overestimated and have the widest range.

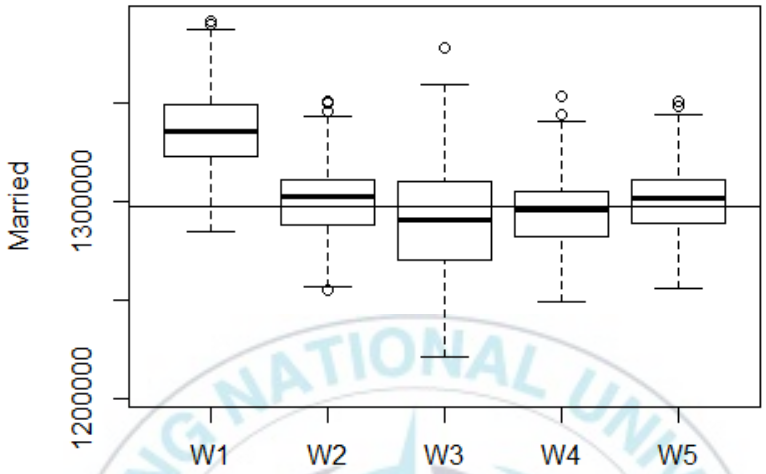


Figure 8. Box plot of the estimators (Married) by methods

To summarize, in estimating the joint distribution of households and individual variables used in the calibration process, W5 estimates the population value most accurately. In estimating variables that were not used as auxiliary information, W3 has a small RB and a large variation of estimators in many cases. W2 and W5 produce efficient estimates in terms of MSE considering bias and variance.

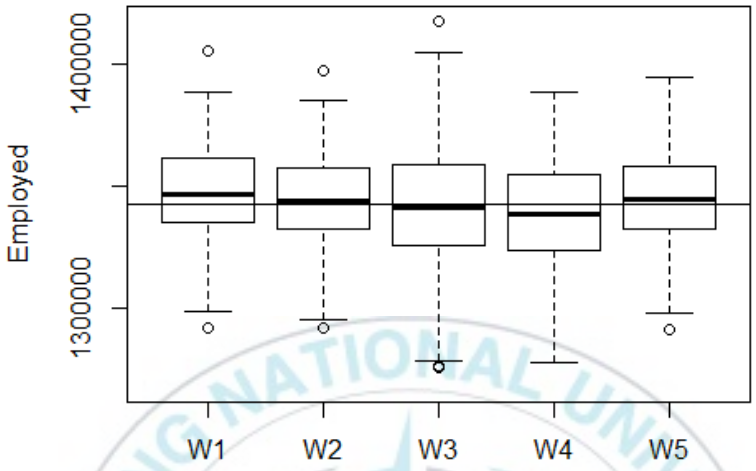


Figure 9. Box plot of the estimators (Employment - Employed) by methods

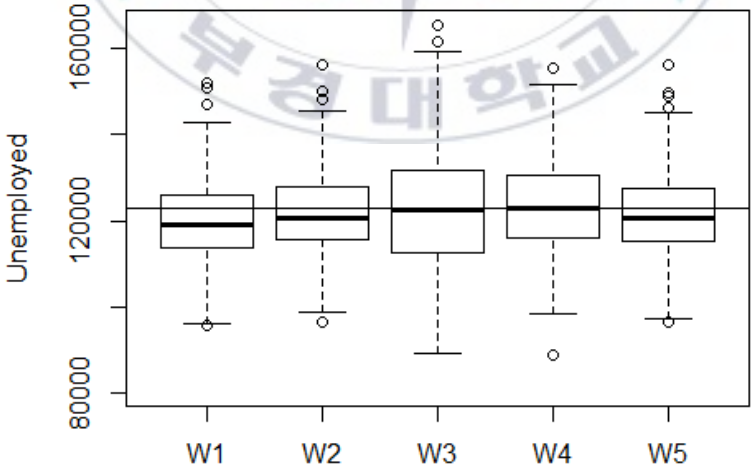


Figure 10. Box plot of the estimators (Employment - Unemployed) by methods

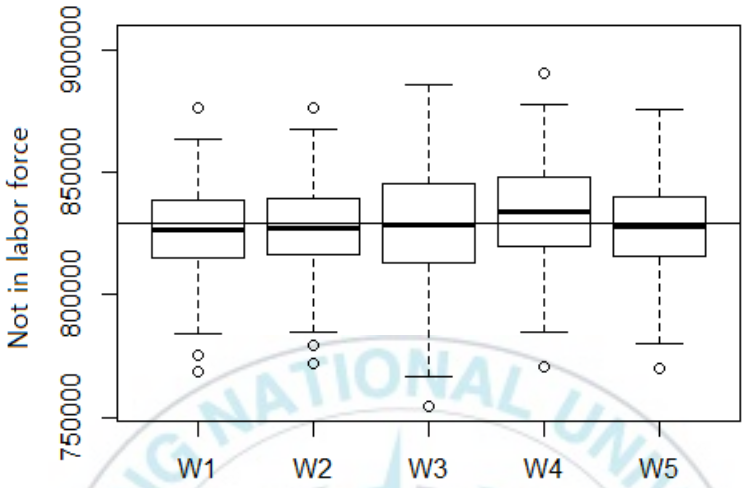


Figure 11. Box plot of the estimators (Employment - Not in labor force) by methods

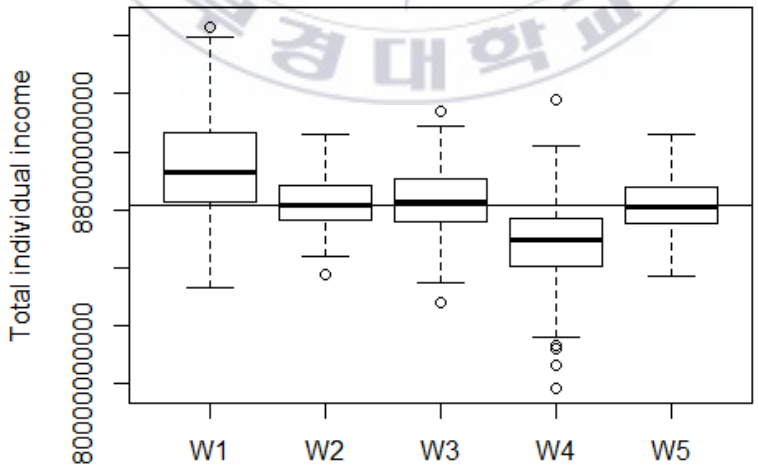


Figure 12. Box plot of the estimators (Total individual income) by methods

5. Conclusions

In household surveys, estimates at different levels are needed according to the circumstances. Estimates for each level can be further improved when household and individual weights are well calculated to reflect each other's multivariate structures. Therefore, the objective of our study was to produce improved weights by using auxiliary information of both levels.

In this paper, we proposed the iterative composite fitting approach that repeats the calibration at the household and individual level in an effort to maintain a structural relationship between them. The existing calibration methods for integrated weights in the literature and proposed method were compared and analyzed using both empirical study and simulation study.

Under our sampling design and response mechanism, we have identified that the proposed method and the single step calibration with integration option (1) were similarly distributed. The difference between the two methods is that the single step calibration can only be calculated when the joint distribution of households and individuals is available when applying raking ratio adjustment as a calibration method, but the proposed method can be produced even when the joint distribution can not be accessed. Plus, the proposed method differs in that it does not redefine auxiliary information but uses auxiliary information for each level as it is. Our method is calibrated at each level, so it is possible to adjust non-response at each level. Though Non-integrated calibration produced weights with the smallest variation, variance

of estimators based on the weights could be increased. Single step calibration with integration option (2) yielded weights with wide range and negative. This in turn increased inefficiency in the estimation. Two step calibration with integration option (1) was unable to produce the household weights reflected the auxiliary information of the individual level. Individual weights computed by two step calibration were negative and wide because the option (1) must be keep in the calibration process.

For the future study, our simulation will be extended to more practical situations where stratified multistage sample designs is used. Weight trimming can also be considered to solve negative weights for our study. Further research on the estimation of variance for the proposed weights is needed.

References

[1] Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376-382.

[2] Estevao, V. M. and Särndal, C. E. (2006). Survey estimates by calibration on complex auxiliary information. *Internatinal Statistical Review* **74**, 127-147.

[3] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

[4] Kolenikov, S. and Hammer, H. (2015). Simultaneous raking of survey weights at multiple levels. *Survey Methods: Insights from the Field, Special issue: 'Weighting: Practical Issues and 'How to' Approach. Survey Methods: Insights from the Field. from http://surveyinsights.org/?p=5099.*

[5] Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer-Verlag.

[6] Valliant, R., Dever, J. A. and Kreuter, F. (2013). *Practical tools for designing and weighting survey samples.* New York: Springer.

Appendix.
R Code

```
library(sas7bdat)
library(survey)
library(sampling)
library(questionr)

# ACS2012 ( hh level )
hh=read.sas7bdat("F:\\논문\\data\\hh1205.sas7bdat")
hh$HHSZ<- replace(hh$num_adults,hh$num_adults>=4,4)
hh$HHIC<-
  ifelse(-20000<=hh$HHINCOME&hh$HHINCOME<=19999,1,
         ifelse(20000<=hh$HHINCOME&hh$HHINCOME<=39999,2,
                ifelse(40000<=hh$HHINCOME&hh$HHINCOME<=64999,3,
                       ifelse(65000<=hh$HHINCOME&hh$HHINCOME<=99999,4,5))))

# ACS2012 ( ps level )
ps=read.sas7bdat("F:\\논문\\data\\ps1205.sas7bdat")
ps$RACE <- ifelse(ps$RACE>=3,3,ps$RACE)
ps$EDUC <- ifelse(1<=ps$EDUCD&ps$EDUCD<=61,1,
                  ifelse(63<=ps$EDUCD&ps$EDUCD<=64,2,
                         ifelse(65<=ps$EDUCD&ps$EDUCD<=99,3,
                                ifelse(100<=ps$EDUCD&ps$EDUCD<=110,4,5))))
ps$AGE <- ifelse (18<=ps$AGE&ps$AGE<=29,1,
```

```
                    ifelse (30<=ps$AGE&ps$AGE<=44,2,
                            ifelse (45<=ps$AGE&ps$AGE<=54,3,
                                    ifelse (55<=ps$AGE&ps$AGE<=64,4,5))))

ps$HHSZ<- replace(ps$num_adults,ps$num_adults>=4,4)
ps$INCTOT5<-
  ifelse(ps$INCTOT<=9999,1,
         ifelse(10000<=ps$INCTOT&ps$INCTOT<=19999,2,
                ifelse(20000<=ps$INCTOT&ps$INCTOT<=32749,3,
                       ifelse(32750<=ps$INCTOT&ps$INCTOT<=49999,4,5))))

# pop totals of hh var
N_hh <- nrow(hh)
x.HHSZ <- table(hh$HHSZ)[-1]
x.HHINCOME <- sum(hh$HHINCOME)
pop.hh_w1 <- c('(Intercept)' = N_hh,
               HHSZ = x.HHSZ,
               HHINCOME = x.HHINCOME)

# pop totals of ps var
N_ps <- nrow(ps)
x.SEX <- table(ps$SEX)[-1]
x.RACE <- table(ps$RACE)[-1]
x.EDUC <- table(ps$EDUC)[-1]
x.AGE <- table(ps$AGE)[-1]
pop.ps_w1 <- c('(Intercept)' = N_ps,
               SEX = x.SEX,
               RACE = x.RACE,
               EDUC = x.EDUC,
```

```
                AGE = x.AGE
)

pop.ps_w2<- c(
  '(Intercept)' = N_ps,
  SEX = x.SEX,
  RACE = x.RACE,
  EDUC = x.EDUC,
  AGE = x.AGE,
  HHSZ=table(hh$HHSZ),
  HHINCOME <- sum(hh$HHINCOME)
)


pop.hh_w3<- c(
  '(Intercept)' = N_hh,
  SEX = table(ps$SEX),
  RACE = x.RACE,
  EDUC = x.EDUC,
  AGE = x.AGE,
  HHSZ=x.HHSZ,
  HHINCOME=sum(hh$HHINCOME)
)
pop.hh_w4 <- c('(Intercept)' = N_hh,
                HHSZ = x.HHSZ,
                HHINCOME=sum(hh$HHINCOME)
)
pop.hh_w5 <- c('(Intercept)' = N_hh,
                HHSZ = x.HHSZ,
                HHINCOME=sum(hh$HHINCOME)
```

```
)
pop.ps_w5 <- c('(Intercept)' = N_ps,
               SEX = x.SEX,
               RACE = x.RACE,
               EDUC = x.EDUC,
               AGE = x.AGE
)

set.seed(2018200)
n=1
I<-list()
w_hh_1<-list();w_ps_1<-list();w_hh_2<-list();w_ps_2<-list();
w_hh_3<-list();w_ps_3<-list();w_hh_4<-list();w_ps_4<-list();
w_hh_5<-list();w_ps_5<-list();sam_hh<-list();sam_ps<-list()

repeat{
  ########################################################################
  # randomly select 5000 households
  n_hh <- 5000
  p_hh <- rep(n_hh/N_hh,N_hh)
  samh <- sample(1:N_hh, n_hh)
  samdat_hh <- hh[samh, ]
  samdat_hh$d_hh <- 1/p_hh[samh]

  # non-response model
  h_respro=1/(1+exp(-(-0.5*(samdat_hh$HHSZ==1)
                      -0.7*(samdat_hh$HHSZ>=3)
                      +0.1*log(samdat_hh$HHINCOME+20000))))
  h_responded<-runif(5000)<h_respro
  samdat_hh<-samdat_hh[h_responded,]
```

```
samp<-merge(samdat_hh[,c(1,11)],ps,by='SERIAL',all.x=T)
ps_respro=1/(1+exp(-(-0.2*(samp$SEX==1)-0.5*(samp$AGE==2)
                     +0.25*(samp$RACE==1)-0.4*(samp$EDUC==1)
                     +0.1*(samp$EDUC==4)+0.3*(samp$EDUC==5)
)))
ps_responded<-runif(length(ps_respro))<ps_respro
# respondents
samdat_ps<-samp[ps_responded,]
samdat_hh <- samdat_hh[with(samdat_hh,order(SERIAL)),]
samdat_ps <- samdat_ps[with(samdat_ps,order(SERIAL)),]
samhh_<-data.frame(unique(samdat_ps$SERIAL)
 ,rep(1,length(unique(samdat_ps$SERIAL))))
colnames(samhh_)<-c("SERIAL","I")
#responded household
samdat_hh<-merge(samhh_,samdat_hh,by='SERIAL',all.x=T)[,-2]
samdat_ps$d_ps<-samdat_ps$d_hh

#########################################################################

# w1
acs.dsgn_hh <- svydesign(ids = ~0, # no clusters
                         strata = NULL, # no strata
                         data = data.frame(samdat_hh),
                         weights = ~ d_hh)
hh.lin_w1 <- calibrate(design = acs.dsgn_hh,
                       formula = ~as.factor(HHSZ) + HHINCOME,
                       population = pop.hh_w1,
                       calfun="linear")
```

```
wt_hh_1<-weights(hh.lin_w1)

acs.dsgn_ps <- svydesign(ids = ~0, # no clusters
                         strata = NULL, # no strata
                         data = data.frame(samdat_ps),
                         weights = ~ d_ps)

ps.lin_w1 <- calibrate(design = acs.dsgn_ps,
                       formula = ~as.factor(SEX) + as.factor(RACE)
                       + as.factor(EDUC)
                       + as.factor(AGE),
                       population = pop.ps_w1,
                       calfun="linear")

wt_ps_1 <- weights(ps.lin_w1)

#########################################################################


#########################################################################
# w2

samdat_ps_w2<-samdat_ps
samdat_ps_w2$HHSZ<- replace(samdat_ps$num_adults
                                    ,samdat_ps$num_adults>=4,4)
samdat_ps_w2$HHIC<-
  ifelse(-20000<=samdat_ps$HHINCOME&samdat_ps$HHINCOME<=19999,1,
  ifelse(20000<=samdat_ps$HHINCOME&samdat_ps$HHINCOME<=39999,2,
  ifelse(40000<=samdat_ps$HHINCOME&samdat_ps$HHINCOME<=64999,3,
  ifelse(65000<=samdat_ps$HHINCOME&samdat_ps$HHINCOME<=99999,4,5))))
samdat_ps_w2<-transform(samdat_ps_w2,
```

```
                        HHSZ_1=ifelse(HHSZ==1,1,0),
                        HHSZ_2=ifelse(HHSZ==2,1,0),
                        HHSZ_3=ifelse(HHSZ==3,1,0),
                        HHSZ_4=ifelse(HHSZ==4,1,0)
)
samdat_ps_w2<-transform(samdat_ps_w2,
                        HHIC_1=ifelse(HHIC==1,1,0),
                        HHIC_2=ifelse(HHIC==2,1,0),
                        HHIC_3=ifelse(HHIC==3,1,0),
                        HHIC_4=ifelse(HHIC==4,1,0),
                        HHIC_5=ifelse(HHIC==5,1,0))
samdat_ps_w2$HHINCOME<-samdat_ps$HHINCOME/samdat_ps$num_adults
for ( i in 25:28 ){
  samdat_ps_w2[,i]<-samdat_ps_w2[,i]/samdat_ps_w2$num_adults
}
for ( i in 29:33 ){
  samdat_ps_w2[,i]<-samdat_ps_w2[,i]/samdat_ps_w2$num_adults
}
acs.dsgn_ps <- svydesign(ids = ~0, # no clusters
                         strata = NULL, # no strata
                         data = data.frame(samdat_ps_w2),
                         weights = ~ d_ps)
# Compute ps GREG weights
ps.lin_w2 <- calibrate(design = acs.dsgn_ps,
                       formula = ~as.factor(SEX)+ as.factor(RACE)
                       + as.factor(EDUC)
                       + as.factor(AGE)
                       +HHSZ_1+HHSZ_2+HHSZ_3
                       +HHSZ_4+HHINCOME,
```

```
                        population = pop.ps_w2,
                        calfun="linear")

samdat_ps_w2$wt_ps_2<-weights(ps.lin_w2)
wt_ps_2<-weights(ps.lin_w2)
wt_hh_2 <- aggregate(samdat_ps_w2$wt_ps_2
               ,by=list(samdat_ps_w2$SERIAL),sum)[,2]/samdat_hh$num_adults

#############################################################################


#############################################################################
# w3
samdat_ps_w3<-samdat_ps
samdat_ps_w3<-transform(samdat_ps_w3,
                        SEX_1=ifelse(SEX==1,1,0),
                        SEX_2=ifelse(SEX==2,1,0))
samdat_ps_w3<-transform(samdat_ps_w3,
                        AGE_1=ifelse(AGE==1,1,0),
                        AGE_2=ifelse(AGE==2,1,0),
                        AGE_3=ifelse(AGE==3,1,0),
                        AGE_4=ifelse(AGE==4,1,0),
                        AGE_5=ifelse(AGE==5,1,0)
)
samdat_ps_w3<-transform(samdat_ps_w3,
                        RACE_1=ifelse(RACE==1,1,0),
                        RACE_2=ifelse(RACE==2,1,0),
                        RACE_3=ifelse(RACE==3,1,0)
)
samdat_ps_w3<-transform(samdat_ps_w3,
```

```
                        EDUC_1=ifelse(EDUC==1,1,0),
                        EDUC_2=ifelse(EDUC==2,1,0),
                        EDUC_3=ifelse(EDUC==3,1,0),
                        EDUC_4=ifelse(EDUC==4,1,0),
                        EDUC_5=ifelse(EDUC==5,1,0)
)
head(samdat_ps_w3,7)
samdat_hh_w3<-samdat_hh
SEX_1<-aggregate(as.numeric(samdat_ps_w3$SEX_1)
                 , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
SEX_2<-aggregate(as.numeric(samdat_ps_w3$SEX_2)
                 , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
AGE_1<-aggregate(as.numeric(samdat_ps_w3$AGE_1)
                 , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
AGE_2<-aggregate(as.numeric(samdat_ps_w3$AGE_2)
                 , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
AGE_3<-aggregate(as.numeric(samdat_ps_w3$AGE_3)
                 , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
AGE_4<-aggregate(as.numeric(samdat_ps_w3$AGE_4)
                 , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
AGE_5<-aggregate(as.numeric(samdat_ps_w3$AGE_5)
                 , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
RACE_1<-aggregate(as.numeric(samdat_ps_w3$RACE_1)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
RACE_2<-aggregate(as.numeric(samdat_ps_w3$RACE_2)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
RACE_3<-aggregate(as.numeric(samdat_ps_w3$RACE_3)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
```

```
EDUC_1<-aggregate(as.numeric(samdat_ps_w3$EDUC_1)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
EDUC_2<-aggregate(as.numeric(samdat_ps_w3$EDUC_2)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
EDUC_3<-aggregate(as.numeric(samdat_ps_w3$EDUC_3)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
EDUC_4<-aggregate(as.numeric(samdat_ps_w3$EDUC_4)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
EDUC_5<-aggregate(as.numeric(samdat_ps_w3$EDUC_5)
                  , by=list(samdat_ps_w3$SERIAL), FUN=sum)[,2]
dummy<-data.frame(SEX_1,SEX_2,AGE_1,AGE_2,AGE_3,AGE_4,AGE_5
                  ,RACE_1,RACE_2,RACE_3,EDUC_1,EDUC_2,EDUC_3,EDUC_4,EDUC_5)
samdat_hh_w3[,12:26]<-dummy

acs.dsgn_hh <- svydesign(ids = ~0, # no clusters
                         strata = NULL, # no strata
                         data = data.frame(samdat_hh_w3),
                         weights = ~ d_hh)
hh.lin_w3 <- calibrate(design = acs.dsgn_hh,
                       formula = ~SEX_1+SEX_2+RACE_2+RACE_3
                                 +EDUC_2+EDUC_3+EDUC_4+EDUC_5
                                 +AGE_2+AGE_3+AGE_4+AGE_5
                       +as.factor(HHSZ)+HHINCOME,
                       population = pop.hh_w3,
                       calfun="linear")

samdat_hh_w3$wt_hh_3<-weights(hh.lin_w3)
wt_hh_3<-weights(hh.lin_w3)
wt_ps_3<-(merge(samdat_hh_w3[,c(1,27)],samdat_ps_w3,key="CLUSTER",all.y=T))[,2]
```

```
samdat_ps_w3$wt_ps_3<-wt_ps_3

#########################################################################

#########################################################################
# w4
acs.dsgn_hh <- svydesign(ids = ~0, # no clusters
                         strata = NULL, # no strata
                         data = data.frame(samdat_hh),
                         weights = ~ d_hh)
hh.lin_w4 <- calibrate(design = acs.dsgn_hh,
                       formula = ~ as.factor(HHSZ)+HHINCOME,
                      population = pop.hh_w4,
                       calfun="linear")

wt_hh_4<-weights(hh.lin_w4)

constr<-wt_hh_4*samdat_hh$num_adults
acs.dsgn_ps <- svydesign(ids = ~0, # no clusters
                         strata = NULL, # no strata
                         data = data.frame(samdat_ps),
                         weights = ~ d_ps)
x.constr <- constr
pop.ps_w4 <- c(constr = x.constr,
               SEX = x.SEX,
               RACE = x.RACE,
               EDUC = x.EDUC,
               AGE = x.AGE
)
# Compute ps GREG weights
ps.lin_w4 <- calibrate(design = acs.dsgn_ps,
```

```
                        formula = ~as.factor(SERIAL) + as.factor(SEX)
                        + as.factor(RACE) + as.factor(EDUC)
                        + as.factor(AGE)+0,
                        population = pop.ps_w4,
                        calfun="linear")
wt_ps_4 <- weights(ps.lin_w4)

#############################################################################

#############################################################################
# w5

iter=1
c_j <- data.frame(rep(0,dim(samdat_hh)[1])) # adjustment factor
colnames(c_j)<- c("V1")
w_hh<-data.frame(V1=rep(0,nrow(samdat_hh)))
w_ps<-data.frame(V1=rep(0,nrow(samdat_ps)))
tw_hh<-data.frame(V1=rep(0,nrow(samdat_hh)))
tw_ps<-data.frame(V1=rep(0,nrow(samdat_ps)))

repeat{
  acs.dsgn_ps <- svydesign(ids = ~0, # no clusters
                           strata = NULL, # no strata
                           data = data.frame(samdat_ps),
                           weights = ~ d_ps)
  ps.lin_w5 <- calibrate(design = acs.dsgn_ps,
                         formula = ~as.factor(SEX) + as.factor(RACE)
                         + as.factor(EDUC)
                         + as.factor(AGE),
                         population = pop.ps_w5,
```

```
                        calfun="linear")

samdat_ps$twt_ps <- weights(ps.lin_w5)
samdat_hh$twt_hh <- aggregate(samdat_ps$twt_ps
            ,by=list(samdat_ps$SERIAL),sum)[,2]/samdat_hh$num_adults
tw_hh[,iter] <- samdat_hh$twt_hh
tw_ps[,iter] <- samdat_ps$twt_ps
acs.dsgn_hh <- svydesign(ids = ~0, # no clusters
                         strata = NULL, # no strata
                         data = data.frame(samdat_hh),
                         weights = ~ twt_hh)
hh.lin_w5 <- calibrate(design = acs.dsgn_hh,
                       formula = ~as.factor(HHSZ) + HHINCOME,
                       population = pop.hh_w5,
                       calfun="linear")
samdat_hh$wt_hh<-weights(hh.lin_w5)

if (iter==1) {
samdat_ps <- merge(samdat_hh[,c(1,12,13)]
                   ,samdat_ps,key="CLUSTER",all.y=T)
} else {
samdat_ps <- merge(samdat_hh[,c(1,12,13)]
                   ,samdat_ps[,-c(2,3)],key="CLUSTER",all.y=T)
}
samdat_ps$wt_ps <- samdat_ps$wt_hh*samdat_ps$twt_ps/samdat_ps$twt_hh

c_j[,iter] <- samdat_hh$wt_hh/samdat_hh$twt_hh
w_hh[,iter] <- samdat_hh$wt_hh
w_ps[,iter] <- samdat_ps$wt_ps
```

```
    if((max(abs(samdat_ps$d_ps-samdat_ps$wt_ps))<0.00001)|iter>=100){break}
    iter<-iter+1
    samdat_ps$d_ps <- samdat_ps$wt_ps
  }

  wt_hh_5<-w_hh[,dim(w_hh)[2]]
  wt_ps_5<-w_ps[,dim(w_hh)[2]]

  I[[n]]<-iter
  w_hh_1[[n]]<-wt_hh_1
  w_ps_1[[n]]<-wt_ps_1
  w_hh_2[[n]]<-wt_hh_2
  w_ps_2[[n]]<-wt_ps_2
  w_hh_3[[n]]<-wt_hh_3
  w_ps_3[[n]]<-wt_ps_3
  w_hh_4[[n]]<-wt_hh_4
  w_ps_4[[n]]<-wt_ps_4
  w_hh_5[[n]]<-wt_hh_5
  w_ps_5[[n]]<-wt_ps_5
  sam_hh[[n]]<-samdat_hh
  sam_ps[[n]]<-samdat_ps
  if(n==200){break}
  n<-n+1
}
```