2007

Person-level and household-level regression estimation in household surveys

David G. Steel
University of Wollongong, dsteel@uow.edu.au

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

Person-level and household-level regression estimation in household surveys

Abstract

A common class of survey designs involves selecting all people within selected households. Generalized regressionestimators can be calculated at either the person or household level. Implementing the estimator at the household level has the convenience of equal estimation weights for people within households. In this article the two approaches are compared theoretically and empirically for the case of simple random sampling of households and selection of all persons in each selected household. We find that the household level approach is theoretically more efficient in large samples and any empirical inefficiency in small samples is limited.

Keywords

Person, level, household, level, regression, estimation, household, surveys

Disciplines

Physical Sciences and Mathematics

Publication Details

Steel, D. G. & Clark, R. G. (2007). Person-level and household-level regression estimation in household surveys. Survey Methodology, 33 (1), 51-60.

Survey Methodology, June 2007
Vol. 33, No. 1, pp. 51-60
Statistics Canada, Catalogue No. 12-001

Person-level and household-level regression estimation in household surveys

David G. Steel and Robert G. Clark [1]

Abstract

A common class of survey designs involves selecting all people within selected households. Generalized regression estimators can be calculated at either the person or household level. Implementing the estimator at the household level has the convenience of equal estimation weights for people within households. In this article the two approaches are compared theoretically and empirically for the case of simple random sampling of households and selection of all persons in each selected household. We find that the household level approach is theoretically more efficient in large samples and any empirical inefficiency in small samples is limited.

Key Words: Contextual effects; Generalized regression estimator; Intra-class correlation; Sampling variance; Model-assisted; Household surveys.

1. Introduction

Many household surveys involve selecting a sample of households and then selecting all people in the scope of the survey in the selected households. Data on one or more variables of interest are collected for the people in the sample. There may be some auxiliary variables whose population totals and sample values are known; for example these may consist of population counts by geographic and demographic classifications. The generalized regression (GREG) estimator is often used to combine auxiliary information and sample data to efficiently estimate the population totals of the variables of interest.

The GREG estimator makes use of a regression model relating the variable of interest to the auxiliary variables. The standard approach is to fit this model using data for each person in the sample (*e.g.*, Lemaître and Dufour 1987, first paragraph). This person-level GREG estimator is equal to a weighted sum of the sample values of the variable of interest, where the weights are in general different for each person.

It is sometimes convenient to have equal weights for people within a household, for surveys which collect information on both household and person level variables of interest. The same weights can then be used for both types of variables. This ensures that relationships between household variables and person variables are reflected in estimates of total. If a household level variable is equal to the sum of person level variables (for example if household income is the sum of personal incomes), then the estimated total of the household variable will equal the estimated total of the person variable. This is not generally the case where separate weighting procedures are used for person and household variables. Similarly, if there is an inequality relationship between a household level variable and the sum of the person level variables, this will also be reflected in the estimates of the two variables. For example, the estimated number of households using child care centres should not exceed the estimated number of children using centres.

The household-level GREG estimator achieves equal weights within households by fitting the regression model using household totals of the variable of interest and the auxiliary variables (*e.g.*, Nieuwenbroek 1993). Weights with this property are called integrated weights.

An alternative approach would be to use different estimation methods for household-level and person-level variables, and then make an adjustment to force agreement of estimates which should be equal. This approach is sometimes called benchmarking and has mainly been used to achieve consistency between estimates from annual and sub-annual business surveys (*e.g.*, Cholette 1984). A benchmarking approach to household and person-level variables from household surveys would require explicit identification of which person and household-level variables should have equal population totals. In this article we concentrate on integrated weighting and do not consider benchmarking approaches.

Luery (1986); Alexander (1987); Heldal (1992) and Lemaître and Dufour (1987) discussed a number of methods which give integrated weights for person-level and household-level estimates. However, none of these authors evaluated the impact on the sampling variance of calculating the generalized regression estimator at the household level rather than the person level. This is an important issue in practice because the cosmetic benefit of integrated weighting must be balanced against any effect on sampling efficiency.

1. David G. Steel and Robert G. Clark, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australia. E-mail: David_Steel@uow.edu.au.

This article compares the design variance, which is the variance over repeated probability sampling from a fixed population, of the person-level and household-level generalized regression estimators. In Section 2, we prove that the large sample variance of the household-level estimator is less than or equal to that of the person-level estimator, by showing that the former is optimal in a large class of GREG estimators. We show that this is because the household-level estimator effectively models contextual effects whereas the person-level estimator does not. In Section 3 the two estimators are compared for a range of variables in a simulation study. Section 4 is a discussion. Three theorems are proved in an Appendix.

2. Theoretical comparison of person and household GREGs

2.1 The generalized regression estimator

In this subsection the generalized regression estimator is described for the general case of probability sampling from any population of units. Let U be a finite population of units and $s \subseteq U$ be the sample. The probabilities of selection are $\pi_i = \Pr[i \in s]$ for units $i \in U$. Let y_i be the variable of interest which is observed for units $i \in s$. Let $\boldsymbol{z_i}$ be the vector of auxiliary variables for unit i, which are observed for every unit in the population. The population totals of these variables are T_Y and $\boldsymbol{T_Z}$ respectively.

The generalized regression estimator of T_Y is based on a model relating the variable of interest to the auxiliary variables:

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\beta}^T \boldsymbol{z_i} \\ &\mathrm{var}_M[y_i] = v_i \sigma^2 \\ &y_i, y_j \text{ independent for } i \neq j \end{aligned}\right\} \quad (1)$$

where v_i are known variance parameters. Subscripts "M" refer to expectations under a model and subscripts "p" refer to design-based expectations, which are expectations over repeated probability sampling from a fixed population. For business surveys collecting continuous variables such as business income and expenses, v_i are often modelled as a function of business size. For household surveys, the variable of interest is often dichotomous, in which case v_i is usually set to 1 corresponding to a homoskedastic model.

Usually $\boldsymbol{z_i}$ have the property that there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T \boldsymbol{z_i} = 1$ for all $i \in U$. For example, this is true if the regression model (1) contains an intercept parameter.

Definition 1. generalized regression estimator

The generalized regression estimator for model (1) is defined as

$$\hat{T}_r = \hat{T}_\pi + \hat{\boldsymbol{\beta}}^T (\boldsymbol{T_Z} - \hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}}) \quad (2)$$

where

$$\hat{T}_\pi = \sum_{i \in s} \pi_i^{-1} y_i$$
$$\hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}} = \sum_{i \in s} \pi_i^{-1} \boldsymbol{z_i}.$$

and $\hat{\boldsymbol{\beta}}$ is a solution of

$$\sum_{i \in s} c_i \pi_i^{-1} \left(y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{z_i} \right) \boldsymbol{z_i} = \boldsymbol{0}$$

where c_i are regression weights. (Often c_i are set to $c_i = v_i^{-1}$.)

The coefficients $\hat{\boldsymbol{\beta}}$ are calculated from a weighted least squares regression of y_i on $\boldsymbol{z_i}$ for $i \in s$. The GREG estimator has low design variance if the model is approximately true but is design-consistent regardless of the truth of the model (*e.g.*, Särndal, Swensson and Wretman 1992, chapter 6).

For large samples the design variance of $\hat{T}_r$ is approximately equal to

$$\mathrm{var}_p[\hat{T}_r] \approx \mathrm{var}_p[\tilde{T}_r] \quad (3)$$

where

$$\tilde{T}_r = \hat{T}_\pi + \boldsymbol{B}^T (\boldsymbol{T_Z} - \hat{\boldsymbol{T}}_{\boldsymbol{Z\pi}})$$

and $\boldsymbol{B}$ is a solution of

$$\sum_{i \in U} c_i (y_i - \boldsymbol{B}^T \boldsymbol{z_i}) \, \boldsymbol{z_i} = \boldsymbol{0}$$

(Särndal *et al.* 1992, Result 6.6.1, page 235). The coefficients $\boldsymbol{B}$ are calculated from a weighted least squares regression of y_i on $\boldsymbol{z_i}$ for $i \in U$. The sample regression coefficients $\hat{\boldsymbol{\beta}}$ are design-consistent for $\boldsymbol{B}$.

2.2 Person and household level GREGs

We now consider the special case of household sampling, where the basic unit, i, is the person. Let $\boldsymbol{x_i}$ be the p-vector of auxiliary variables observed for all people $i \in U$. The elements of $\boldsymbol{x_i}$ may refer to characteristics of the person or of the household to which they belong. The population and sample of households will be denoted U_1 and s_1 respectively. The population of people in household $g \in U_1$ will be denoted U_g which is of size N_g. Let $y_{g1} = \sum_{i \in U_g} y_i$ and $\boldsymbol{x_{g1}} = \sum_{i \in U_g} \boldsymbol{x_i}$ be the household totals of y_i and $\boldsymbol{x_i}$. Let $\bar{\boldsymbol{x}}_{\boldsymbol{g}} = \boldsymbol{x_{g1}} / N_g$ be the household mean of $\boldsymbol{x_i}$.

We consider the common case where households are selected by probability sampling and all people are selected from selected households, so that $s = \bigcup_{g \in s_1} U_g$. Let

$\pi_{g1} = P[g \in s_1] > 0$ be the probability of selection for household g. It follows that $\pi_i = \pi_{g1}$ for $i \in U_g$.

The person-level GREG, $\hat{T}_P$, is the GREG under the following model:

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\beta}^T \boldsymbol{x_i} \\ &\text{var}_M[y_i] = v_i \sigma^2 \\ &y_i,\ y_j \text{ independent for } i \neq j. \end{aligned}\right\} \quad (4)$$

So the person-level GREG, $\hat{T}_P$, is given by substituting $\boldsymbol{x_i}$ for $\boldsymbol{z_i}$ in (2). Model (4) ignores any correlations between y_i and y_j for people i and j in the same household. These correlations were 0.3 or less in most of the variables considered by Clark and Steel (2002), although higher values occurred for variables related to ethnicity, such as Indigenous self-identification. Correlations of 1 could occur for environmental variables. Tam (1995) shows that the optimal model-assisted estimator for cluster sampling is robust to mis-specification of within-cluster correlations. One way of interpreting this result is that correlations within households are not relevant to estimating population totals, because all people are selected in selected households. So within-household correlations do not help to estimate for non-sample individuals, since the sampled and non-sampled people are in distinct households.

A number of methods have been suggested for GREG-type estimation with equal weights within households. Nieuwenbroek (1993) motivated an estimator by aggregating model (4) to household level:

$$\left.\begin{aligned} &E_M[y_{g1}] = \boldsymbol{\beta}^T \boldsymbol{x_{g1}} \\ &\text{var}_M[y_{g1}] = v_{g1} \sigma^2 \\ &y_{g1},\ y_{k1} \text{ independent for } g \neq k. \end{aligned}\right\} \quad (5)$$

where $v_{g1} = \sum_{i \in U_g} v_i$. The GREG estimator using sample data y_{g1} for $g \in s_1$ based on this model is $\hat{T}_H$:

$$\hat{T}_H = \hat{T}_\pi + \hat{\boldsymbol{\beta}}_{\boldsymbol{H}}^{\boldsymbol{T}} (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}}) \quad (6)$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{H}}$ is a solution of

$$\sum_{g \in s_1} \pi_{g1}^{-1} a_g \, (y_{g1} - \hat{\boldsymbol{\beta}}_{\boldsymbol{H}}^{\boldsymbol{T}} \boldsymbol{x_{g1}}) \, \boldsymbol{x_{g1}} = \boldsymbol{0}. \quad (7)$$

The regression coefficient $\hat{\boldsymbol{\beta}}_{\boldsymbol{H}}$ is a household level weighted least squares regression of the sample values of y_{g1} on $\boldsymbol{x_{g1}}$ with weights $\pi_{g1}^{-1} a_g$. The values of a_g could be set to v_{g1}^{-1}. If $v_i = 1$ then $v_{g1} = N_g$ so $a_g = N_g^{-1}$. Alternatively, $a_g = 1$ could also be used.

Several other equivalent integrated weighting methods have been used. Lemaître and Dufour (1987) constructed a generalized regression estimator at person level, using $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ instead of $\boldsymbol{x_i}$ as the auxiliary variables. Nieuwenbroek (1993) commented that this is equivalent to (6) if $c_i = a_g N_g$ for $i \in U_g$. Alexander (1987) developed closely related weighting methods using a minimum distance criterion.

Both the person and household level GREG can be written in weighted form $\sum_{i \in s} w_i Y_i$. The weights for both estimators can be written as $w_i = \pi_i^{-1} g_i$ where

$$g_i = 1 + (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})^T \left(\sum_{i \in s} c_i \pi_i^{-1} \boldsymbol{x_i} \boldsymbol{x_i^T} \right)^{-1} c_i \, \boldsymbol{x_i}$$

for $\hat{T}_P$ and

$$g_i = 1 + (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})^T \left(\sum_{g \in s_1} a_g \, \pi_{g1}^{-1} \boldsymbol{x_{g1}} \boldsymbol{x_{g1}^T} \right)^{-1} a_g \, \pi_g^{-1} \boldsymbol{x_{g1}}$$

for $\hat{T}_H$, where person i belongs to household g. (Superscript "-" stands for generalized inverse of a matrix).

2.3 Theoretical results

In this section, we show that $\hat{T}_H$ has the lowest possible large sample variance in a class of estimators which also includes $\hat{T}_P$, for the sample design where households are selected by simple random sampling without replacement. We will then explain this result by showing that $\hat{T}_H$ is equivalent to a regression estimator calculated using person level data, where the model includes contextual effects.

For large samples, $\hat{T}_P$ and $\hat{T}_H$ can be approximated by

$$\tilde{T}_P = \hat{T}_\pi + \boldsymbol{B_P^T} (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}});$$

and

$$\tilde{T}_H = \hat{T}_\pi + \boldsymbol{B_H^T} (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})$$

respectively, where $\boldsymbol{B_P}$ and $\boldsymbol{B_H}$ are solutions of

$$\left.\begin{aligned} &\sum_{i \in U} c_i (y_i - \boldsymbol{B_P^T x_i}) \, x_i = \boldsymbol{0} \\ &\sum_{g \in U_1} a_g (y_{g1} - \boldsymbol{B_H^T x_{g1}}) \, \boldsymbol{x_{g1}} = \boldsymbol{0} \end{aligned}\right\} \quad (8)$$

(Särndal *et al.* 1992, Result 6.6.1, page 235). Theorem 1 states the minimum variance estimator in a class including $\tilde{T}_P$ and $\tilde{T}_H$.

Theorem 1. Optimal estimator for simple cluster sampling

Suppose that m households are selected by simple random sampling without replacement from a population of M households, and all people are selected from selected households. Consider the estimator of T given by

$$\tilde{T} = \hat{T}_\pi + \boldsymbol{h}^T (\boldsymbol{T_X} - \hat{\boldsymbol{T}}_{\boldsymbol{X\pi}})$$

where $\boldsymbol{h}$ is a constant p-vector. It is assumed that there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda^T x_i} = 1$ for all $i \in U$. The

variance of this estimator is minimised by h^* which are solutions of

$$\sum_{g \in s_1} (y_{g1} - \boldsymbol{h}^T \boldsymbol{x}_{g1}) \boldsymbol{x}_{g1} = \mathbf{0}.$$

Hence $\tilde{T}_H$ with $a_g = 1$ for all g is the optimal choice of $\tilde{T}$.

Theorem 1 has the perhaps surprising implication that $\hat{T}_H$ (with $a_g = 1$ for all g) has lower variance than $\hat{T}_P$ for large samples. This is in spite of the fact that $\hat{T}_H$ discards some of the information in the sample, because it uses the household sums of $\boldsymbol{x}_i$ and y_i. The Theorem suggests that $\hat{T}_H$ is the appropriate GREG estimator for the cluster sampling design assumed here, and that the information discarded by summing to household level is not relevant when this design is used. To explain why $\hat{T}_H$ can perform better than $\hat{T}_P$, we will make use of a "linear contextual model" which is a more general model for $E_M[Y_i]$ than (4). The model is:

$$\left.\begin{aligned} &E_M[y_i] = \boldsymbol{\gamma}_1^T \bar{\boldsymbol{x}}_g + \boldsymbol{\gamma}_2^T \boldsymbol{x}_i \quad (i \in U_g) \\ &\mathrm{var}_M[y_i] = \sigma^2 \\ &y_i,\ y_j \text{ independent for } i \neq j. \end{aligned}\right\} \qquad (9)$$

Both $\bar{\boldsymbol{x}}_g$ and $\boldsymbol{x}_i$ are used as explanatory variables for y_i because the household mean of the person level auxiliary variables may capture some of the effect of household context (Lazarfeld and Menzel 1961). For example, if the elements of $\boldsymbol{x}_i$ are indicator variables summarising the age and sex of person i then $\bar{\boldsymbol{x}}_g$ are the proportions of people in the household falling into different age and sex categories. If the population of interest includes both adults and children, then $\bar{\boldsymbol{x}}_g$ includes the proportion of children in the household, which could be relevant to the labour force participation of adults in the household.

Theorem 2 shows that the improvement in the variance from using $\tilde{T}_H$ with $a_g = 1$ rather than using $\tilde{T}_P$ can be explained by the linear contextual model.

Theorem 2. Explaining the difference in the asymptotic variances

Suppose that households are selected by simple random sampling without replacement and all people are selected from selected households. Let $r_i = y_i - \boldsymbol{B}_P^T \boldsymbol{x}_i$, and let $\boldsymbol{B}_C$ be the result of regressing r_i on $\bar{\boldsymbol{x}}_g$ over $i \in U$ using weighted least squares regression weighted by N_g. Then

$$\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H] = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1} \boldsymbol{B}_C^T \left(\sum_{g \in U_1} \boldsymbol{x}_{g1} \boldsymbol{x}_{g1}^T\right) \boldsymbol{B}_C$$

where $\tilde{T}_H$ is calculated using $a_g = 1$ for all g.

The result shows that the reduction in variance from using $\tilde{T}_H$ (with $a_g = 1$) rather than $\tilde{T}_P$ is a quadratic form in $\boldsymbol{B}_C$. Hence the extent of the improvement depends on the extent to which $\bar{\boldsymbol{x}}_g$ helps to predict y_i after $\boldsymbol{x}_i$ has already been controlled for, *i.e.*, the extent to which a linear contextual effect helps to predict r_i over $i \in U$, using a weighted least squares regression weighted by N_g.

The proofs of Theorems 1 and 2 are very much dependent on the assumption of cluster sampling. The results would not be expected to apply if there was subsampling within households.

Theorems 1 and 2 only apply with $a_g = 1$ in the weighted least squares regression for $\hat{T}_H$. Other choices of a_g are often used, for example it would often be reasonable to assume that $v_{g1} = N_g$ in model (5), in which case it would be sensible to use $a_g = N_g^{-1}$. Theorem 3 shows that $\hat{T}_H$ is equivalent to a person-level GREG estimator fitted under the linear contextual model for other choices of a_g.

Theorem 3. The linear contextual GREG

For sample designs where all people are selected from selected households and $\pi_{g1} > 0$ for all $g \in U_1$, $\hat{T}_H$ with a given choice of a_g is the generalized regression estimator for model (9) where $c_i = a_g N_g$ for $i \in U_g$.

Theorem 3 means that $\hat{T}_H$ is the GREG under a more general model than $\hat{T}_P$. Nieuwenbroek (1993) showed that $\hat{T}_H$ is equal to a person-level GREG derived from regressing y_i on $\bar{\boldsymbol{x}}_g$. Theorem 3 states it is also equal to the person-level GREG from regressing y_i on both $\boldsymbol{x}_i$ *and* $\bar{\boldsymbol{x}}_g$, thereby automatically incorporating any household contextual effects. As a result, $\hat{T}_H$ would be expected to have lower variance than $\hat{T}_P$ for large samples. (In the case of $a_g = 1$, Theorem 1 stated that this is always the case). For small samples, however, a more general model may be counter-productive. Silva and Skinner (1997) showed for single-stage sampling that adding parameters to the model can increase the variance of the GREG estimator, although this effect is negligible for large samples. It is possible that the contextual effects have little or no predictive power for some variables. In this case, it would be expected that $\hat{T}_H$ would perform slightly worse than $\hat{T}_P$ for small samples, and about the same for large samples.

The contextual model, (9), includes all of the elements of $\boldsymbol{x}_i$ and all of the elements of $\bar{\boldsymbol{x}}_g$. An alternative would be to use only those elements of either $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_g$ which are significant, or which give improvements in the estimated variance of a GREG estimator. A GREG estimator based on

this type of model would probably have lower variance than the estimators considered in this paper, but would not give integrated weights unless the same elements of $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_g$ were used.

3. Empirical study

3.1 Methodology

A simulation study was undertaken to compare the person and household GREGs, $\hat{T}_P$ and $\hat{T}_H$, for a range of survey variables. We used two populations, consisting of 187,178 households randomly selected from the 2001 Australian Population Census and 210,132 households from the 1995 Australian National Health Survey. All adults and children in the households were included. The average household size was approximately 2.5.

We selected cluster samples from these populations, where households were selected by simple random sampling without replacement and all people from selected households were selected. We simulated samples of size m = 500, 1,000, 2,000, 5,000 and 10,000 households. In each case, 5,000 samples were selected. The auxiliary variables $\boldsymbol{x}_i$ consisted of indicator variables of sex by agegroup (12 categories). (This choice of $\boldsymbol{x}_i$ means that the GREG estimation is equivalent to post-stratification.) The person-level GREG with $c_i = 1(\hat{T}_P)$, the household-level GREG with $a_g = N_g^{-1}(\hat{T}_{H1})$, and the household-level GREG with $a_g = 1(\hat{T}_{H2})$ were all calculated. We also included the Hájek estimator

$$\hat{T}_1 = N\left(\sum_{i\in s}\pi_i^{-1}y_i\right)\Big/\left(\sum_{i\in s}\pi_i^{-1}\right)$$

which equals $N/n\sum_{g\in s_1}\sum_{i\in U_g}y_i$ for cluster sampling with simple random sampling of households, where n is the realized sample size of people.

The variables include labour force, health and other topics. All of the variables are dichotomous except for income (annual income in Australian dollars, based on range data reported from the Census). "Employment(F)" is the indicator variable which is 1 if a person is employed and female, and 0 otherwise. The first six variables are from the Census population and the remaining five variables are from the health population.

3.2 Results

Table 1 shows the relative root mean squared errors (RRMSEs) of $\hat{T}_1$, $\hat{T}_P$, $\hat{T}_{H1}$ and $\hat{T}_{H2}$, for a sample size of 1,000 households. The RRMSEs are expressed as a percentage of the true population total. The biases have not been tabulated because they were a negligible component of the MSE in all cases. The percentage improvements in MSE of $\hat{T}_{H1}$ and $\hat{T}_{H2}$ relative to $\hat{T}_P$ are also shown. The figures in brackets are the simulation standard errors of these percentage improvements.

For this sample size, $\hat{T}_{H1}$ and $\hat{T}_{H2}$ performed slightly worse than $\hat{T}_P$ for the health variables and slightly better for most other variables. The greatest gain was in estimating the number of sole parents; this variance was reduced by 10.8% and 16.3% by using the household-level GREGs. For all other variables, either the improvement was small or the household GREG was slightly worse than the person-level GREG. The inefficiency from using a household-level GREG rather than $\hat{T}_P$ was never more than 2.2%.

Table 2 shows the percentage improvement in MSE from using $\hat{T}_{H1}$ rather than $\hat{T}_P$ for different sample sizes. The simulation standard errors for each figure are shown in brackets. Table 3 shows the percentage improvements from using $\hat{T}_{H2}$ rather than $\hat{T}_P$. The asymptotic percentage improvements ($m = \infty$) are also shown, based on the large sample approximation to the variance of a GREG. For both household-level GREGs, the percentage improvements are generally increasing as the sample size increases. For m = 500, the household GREGs are generally worse than the person GREGs, although never more than 5% worse. For m = 10,000, an improvement is recorded for over half of the variables. The greatest improvements were for estimates of the number of sole parents (11.5%) and employed women (4.2%); all other improvements were small. $\hat{T}_{H1}$ and $\hat{T}_{H2}$ never had variances more than 0.2% higher than $\hat{T}_P$ for m = 10,000. Generally $\hat{T}_{H2}$ performs better than $\hat{T}_{H1}$ for larger sample sizes, as would be expected from Theorem 1, but the reverse is true for small sample sizes.

In practice estimates of subpopulation totals are often of as much interest as population totals. Table 4 shows the performance of the various estimators for age-sex domains (12 age categories) and region domains, for the sample size of 1,000 households. There were 49 regions in the census dataset. The health dataset did not contain a similar region variable, instead the socioeconomic quintile of the collection district (a geographical unit consisting of approximately 200 contiguous households) was used as the domain. The domain estimators were produced by calculating weights from each estimator and taking the weighted sum over the sample in the domain. This is equivalent to the domain ratio estimator described in Case 1, Section 2.1 of Hidiroglou and Patak (2004). We have used this method because it is the most commonly used in practice, as it enables all domains and population totals to be estimated with a single set of weights, although more efficient domain estimators exist (Hidiroglou and Patak 2004, cases 2-6).

In each case, the median RRMSE over the domains is shown. The table shows that there is not much difference

between the three GREG estimators. For age-sex domains, the household GREGs did slightly better than the person GREG for census variables and slightly worse for health variables. For region estimates, the household GREGs were slightly worse in all cases. Table 5 shows that the households GREGs performed very similarly to $\hat{T}_P$ for a sample size of 10,000 households. It is worth noting that Theorem 1 and 2 do not apply to the domain estimators we have used.

Table 1 Relative RMSEs for sample size of 1,000 households

Variable	RRMSE%				% improvement in MSE	
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
employed	2.62	2.09	2.09	2.10	0.20 (0.26)	-0.28 (0.27)
employed F	3.78	3.05	3.01	3.02	2.63 (0.33)	2.09 (0.33)
income	2.56	2.20	2.19	2.19	1.04 (0.25)	0.75 (0.24)
low income	5.04	4.87	4.89	4.90	-0.62 (0.20)	-1.12 (0.22)
hrs worked	3.08	2.54	2.53	2.53	0.94 (0.28)	0.70 (0.28)
sole parent	12.50	12.73	12.02	11.65	10.84 (0.62)	16.31 (0.49)
arthritis	5.52	4.50	4.53	4.53	-1.38 (0.17)	-1.57 (0.18)
smoker	4.73	4.57	4.60	4.61	-1.64 (0.18)	-1.81 (0.20)
high BPR	6.80	5.30	5.35	5.36	-1.70 (0.17)	-2.06 (0.18)
fair/poor hlth	9.79	9.42	9.47	9.47	-1.16 (0.16)	-1.07 (0.18)
alcohol	4.81	4.66	4.70	4.71	-1.77 (0.16)	-2.15 (0.18)

Table 2 Improvement in MSE of household GREG $\hat{T}_{H1}$ compared to $\hat{T}_P$

Variable	% improvement in MSE					
	m = 500	1,000	2,000	5,000	10,000	∞
employed	-0.65 (0.31)	0.20 (0.26)	1.02 (0.24)	0.90 (0.21)	2.17 (0.21)	1.85
employed F	1.22 (0.37)	2.63 (0.33)	2.59 (0.33)	3.53 (0.31)	4.24 (0.31)	4.13
income	-1.53 (0.31)	1.04 (0.25)	0.48 (0.24)	0.61 (0.19)	1.43 (0.19)	1.07
low income	-2.45 (0.27)	-0.62 (0.20)	0.02 (0.18)	0.18 (0.15)	0.00 (0.00)	0.65
hrs worked	-0.26 (0.34)	0.94 (0.28)	1.72 (0.27)	1.61 (0.24)	2.64 (0.24)	2.12
sole parent	7.81 (0.69)	10.84 (0.62)	10.74 (0.61)	10.23 (0.57)	11.50 (0.58)	11.21
arthritis	-3.01 (0.24)	-1.38 (0.17)	-0.34 (0.12)	-0.08 (0.09)	-0.13 (0.07)	0.08
smoker	-3.91 (0.25)	-1.64 (0.18)	-1.02 (0.12)	-0.26 (0.08)	-0.06 (0.07)	0.16
high BPR	-2.93 (0.24)	-1.70 (0.17)	-0.86 (0.12)	-0.31 (0.08)	-0.04 (0.06)	0.08
fair/poor hlth	-3.67 (0.25)	-1.16 (0.16)	-0.71 (0.12)	-0.05 (0.08)	0.03 (0.06)	0.10
alcohol	-4.22 (0.23)	-1.77 (0.16)	-0.77 (0.12)	-0.31 (0.08)	-0.21 (0.07)	0.14

Table 3 Improvement in MSE of household GREG $\hat{T}_{H2}$ compared to $\hat{T}_P$

Variable	% improvement in MSE					
	m = 500	1,000	2,000	5,000	10,000	∞
employed	-1.85 (0.35)	-0.28 (0.27)	1.25 (0.25)	1.05 (0.21)	2.22 (0.21)	1.98
employed F	0.28 (0.39)	2.09 (0.33)	2.71 (0.33)	3.55 (0.29)	4.50 (0.30)	4.31
income	-2.64 (0.31)	0.75 (0.24)	0.71 (0.22)	0.90 (0.17)	1.30 (0.16)	1.37
low income	-3.15 (0.30)	-1.12 (0.22)	-0.15 (0.18)	0.06 (0.15)	0.00 (0.00)	0.94
hrs worked	-1.51 (0.35)	0.70 (0.28)	1.98 (0.25)	1.79 (0.21)	2.57 (0.22)	2.26
sole parent	14.70 (0.53)	16.31 (0.49)	16.39 (0.47)	15.41 (0.44)	16.44 (0.44)	16.35
arthritis	-3.31 (0.26)	-1.57 (0.18)	-0.05 (0.13)	-0.12 (0.09)	-0.10 (0.07)	0.16
smoker	-3.82 (0.28)	-1.81 (0.20)	-0.69 (0.14)	0.21 (0.11)	0.28 (0.10)	0.57
high BPR	-3.20 (0.26)	-2.06 (0.18)	-1.12 (0.13)	-0.40 (0.09)	-0.05 (0.07)	0.12
fair/poor hlth	-4.02 (0.28)	-1.07 (0.18)	-0.57 (0.13)	-0.09 (0.09)	0.00 (0.07)	0.15
alcohol	-5.00 (0.26)	-2.15 (0.18)	-0.82 (0.13)	-0.49 (0.09)	-0.29 (0.08)	0.18

Table 4 Median relative RMSEs for domain estimators for sample size m = 1,000

Variable	Age-Sex Domains				Region Domains			
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
employed	12.74	7.92	7.93	7.90	29.89	29.92	30.20	30.34
employed F	13.12	8.32	8.36	8.34	34.64	34.65	35.03	35.16
income	13.25	8.43	8.49	8.47	28.04	28.12	28.43	28.51
low income	21.17	18.77	18.96	18.94	42.71	42.85	43.24	43.33
hrs worked	14.56	10.69	10.76	10.72	31.24	31.23	31.52	31.63
sole parent	96.20	96.33	97.64	96.69	92.99	93.30	94.37	93.50
arthritis	24.94	20.94	21.12	21.11	13.31	12.94	13.02	13.04
smoker	32.10	29.25	29.39	29.37	12.32	12.27	12.35	12.38
high BPR	27.01	23.80	23.97	23.95	15.83	15.31	15.44	15.45
fair/poor hlth	39.64	37.73	38.05	38.08	22.38	22.30	22.51	22.55
alcohol	25.58	21.42	21.53	21.58	12.73	12.70	12.80	12.82

Table 5 Median relative RMSEs for domain estimators for sample size m = 10,000

Variable	Age-Sex Domains				Region Domains			
	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$	$\hat{T}_1$	$\hat{T}_P$	$\hat{T}_{H1}$	$\hat{T}_{H2}$
employed	3.77	2.35	2.32	2.31	8.85	8.85	8.87	8.88
employed F	3.86	2.43	2.43	2.42	10.30	10.26	10.25	10.25
income	3.91	2.53	2.51	2.51	8.24	8.23	8.23	8.24
low income	6.31	5.63	5.62	5.61	12.67	12.68	12.69	12.69
hrs worked	4.29	3.15	3.15	3.12	9.26	9.25	9.27	9.27
sole parent	28.40	28.26	28.29	28.23	27.11	27.14	27.16	27.11
arthritis	7.40	6.26	6.27	6.27	3.98	3.85	3.85	3.85
smoker	9.53	8.58	8.58	8.57	3.69	3.67	3.68	3.67
high BPR	8.07	7.02	7.01	7.01	4.66	4.48	4.49	4.49
fair/poor hlth	11.69	11.02	11.02	11.01	6.75	6.69	6.69	6.69
alcohol	7.74	6.43	6.43	6.43	3.87	3.85	3.85	3.85

4. Discussion

The standard person-level GREG estimator produces unequal weights within households. Household-level GREG estimators can be used to give integrated household and person weights, which is beneficial for surveys collecting information on both household-level and person-level variables. This article demonstrated that there is little or no loss associated with the practical benefit of integrated weighting arising from using a household-level GREG estimator. For large samples, the household-level GREG has lower design variance than the person-level GREG. For smaller samples there is at most a small increase in variance for some variables from using the household GREG, because this estimator is equivalent to using a regression model containing more parameters. Therefore, if integrated weights would improve the coherence of a household survey's outputs, the household-level GREG can be adopted with little or no detriment to the variance and bias of estimators.

Acknowledgements

This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics. The views expressed here do not necessarily reflect the views of either organisation. The authors thank Julian England, Frank Yu and Ray Chambers for their thoughtful comments.

Appendix

Proof of theorems

Proof of theorem 1

Let $\bar{Y}_1 = T_Y / M$ and $\bar{\boldsymbol{X}}_1 = \boldsymbol{T}_X / M$ be the population means of y_{g1} and $\boldsymbol{x}_{g1}$ respectively. The variance of $\tilde{T}$ is

$$\begin{aligned}\mathrm{var}_p[\tilde{T}] &= \mathrm{var}[\hat{T}_\pi + \boldsymbol{h}^T(\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi})] \\ &= \mathrm{var}\left[\frac{M}{m}\sum_{g\in s_1}(y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1})\right] \\ &= \frac{M^2}{m}\left(1 - \frac{m}{M}\right)S_r^2\end{aligned}$$

where $S_r^2 = (M-1)^{-1}\sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}^2$. To minimise with respect to $\boldsymbol{h}$, we set the derivative of S_r^2 to zero:

$$
\begin{aligned}
0 &= (M-1)^{-1}\sum_{g\in U_1} \\
&\quad \{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\,(\boldsymbol{x}_{g1} - \bar{\boldsymbol{X}}_1) \\
0 &= \sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\,\boldsymbol{x}_{g1} \\
&\quad -\sum_{g\in U_1}\{(y_{g1} - \bar{Y}_1) - \boldsymbol{h}^T(\boldsymbol{x}_{g1} - \bar{\boldsymbol{X}}_1)\}\bar{\boldsymbol{X}}_1 \\
0 &= \sum_{g\in U_1}\{y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\}\boldsymbol{x}_{g1} \\
0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{h}^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1} - (\bar{Y}_1 - \boldsymbol{h}^T\bar{\boldsymbol{X}}_1)\,\boldsymbol{T}_X. \qquad (10)
\end{aligned}
$$

We now show that (10) is satisfied by $\boldsymbol{h}^*$. By assumption, $\boldsymbol{h}^*$ satisfies

$$\boldsymbol{0} = \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*)\,\boldsymbol{x}_{g1}. \qquad (11)$$

Hence the first sum in the right hand side of (10) is equal to zero for $\boldsymbol{h} = \boldsymbol{h}^*$. Premultiplying both sides of (11) by $\boldsymbol{\lambda}^T$ gives

$$
\begin{aligned}
0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*)\,\boldsymbol{\lambda}^T\boldsymbol{x}_{g1} \\
0 &= \sum_{g\in U_1}(y_{g1} - \boldsymbol{x}_{g1}^T\boldsymbol{h}^*) \\
0 &= T_Y - \boldsymbol{T}_X^T\boldsymbol{h}^*.
\end{aligned}
$$

Dividing by M gives $\bar{Y}_1 - \bar{\boldsymbol{X}}_1^T\boldsymbol{h}^* = 0$. Hence the rest of the right hand side of (10) is equal to zero. So $\boldsymbol{h}^*$ satisfies (10).

Proof of theorem 2

Let "-" denote a generalized inverse of a matrix. Then $\boldsymbol{B}_C$ is equal to

$$
\begin{aligned}
\boldsymbol{B}_C &= \left\{\sum_{g\in U_1}\sum_{i\in U_g} N_g\bar{\boldsymbol{x}}_g\bar{\boldsymbol{x}}_g^T\right\}^-\sum_{g\in U_1}\sum_{i\in U_g} N_g\bar{\boldsymbol{x}}_g r_i \\
&= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}r_{g1}. \qquad (12)
\end{aligned}
$$

Now, $r_i = y_i - \boldsymbol{B}_P^T\boldsymbol{x}_i$ so $r_{g1} = y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1}$. Hence (12) becomes

$$
\begin{aligned}
\boldsymbol{B}_C &= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}(y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1}) \\
&= \left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}y_{g1} \\
&\quad -\left\{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\right\}^-\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_P \\
&= \boldsymbol{B}_H - \boldsymbol{B}_P \qquad (13)
\end{aligned}
$$

since $\boldsymbol{B}_H = \{\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\}\sum_{g\in U_1}\sum_{i\in U_g}\boldsymbol{x}_{g1}y_{g1}$. The difference in the variances is given by

$$
\begin{aligned}
&\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H] = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1} \\
&\qquad \left\{\sum_{g\in U_1}(y_{g1} - \boldsymbol{B}_P^T\boldsymbol{x}_{g1})^2 - \sum_{g\in U_1}(y_{g1} - \boldsymbol{B}_H^T\boldsymbol{x}_{g1})^2\right\}
\end{aligned}
$$

which becomes

$$
\begin{aligned}
&\{\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H]\}\Big/\left\{\frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\right\} \\
&= \sum_{g\in U_1} r_{g1}^2 - \sum_{g\in U_1}(r_{g1} + \boldsymbol{B}_P^T\boldsymbol{x}_{g1} - \boldsymbol{B}_H^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1} r_{g1}^2 - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1} + \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 + \sum_{g\in U_1}(\boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&\quad + 2\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1}^T\boldsymbol{B}_C^T \\
&\quad - \sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})^2 \\
&= \sum_{g\in U_1}\boldsymbol{B}_C^T\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_C + 2\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1}^T\boldsymbol{B}_C. \qquad (14)
\end{aligned}
$$

Now, $\boldsymbol{B}_C$ is an ordinary least squares regression of r_{g1} on $\boldsymbol{x}_{g1}$ so

$$\sum_{g\in U_1}(r_{g1} - \boldsymbol{B}_C^T\boldsymbol{x}_{g1})\,\boldsymbol{x}_{g1} = \boldsymbol{0}.$$

Hence (14) becomes

$$
\begin{aligned}
&\mathrm{var}_p[\tilde{T}_P] - \mathrm{var}_p[\tilde{T}_H] = \\
&\qquad \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\boldsymbol{B}_C^T\sum_{g\in U_1}\boldsymbol{x}_{g1}\boldsymbol{x}_{g1}^T\boldsymbol{B}_C.
\end{aligned}
$$

Proof of Theorem 3

The GREG estimator is invariant under linear invertible transformations of the auxiliary variables. Hence model (9) can be re-parameterised to give

$$E_M[y_i] = \boldsymbol{\phi}_1^T \bar{\boldsymbol{x}}_g + \boldsymbol{\phi}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g) \tag{15}$$

or equivalently

$$E_M[y_i] = \boldsymbol{\phi}^T \boldsymbol{z}_i$$

where

$$\boldsymbol{z}_i = \begin{pmatrix} \bar{\boldsymbol{x}}_g \\ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_g \end{pmatrix}$$

and

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix}.$$

The parameters in model (15) are related to those in model (9) by $\boldsymbol{\phi}_1 = \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2$ and $\boldsymbol{\phi}_2 = \boldsymbol{\gamma}_2$.

From Definition 1, noting that

$$s = \bigcup_{g \in s_1} U_g$$

for the assumed design, the generalized regression estimator under model (15) is

$$\begin{aligned} \hat{T} &= \hat{T}_\pi + \sum_{i \in U} \hat{\boldsymbol{\phi}}^T \boldsymbol{z}_i - \sum_{i \in s} \pi_i^{-1} \hat{\boldsymbol{\phi}}^T \boldsymbol{z}_i \\ &= \hat{T}_\pi + \sum_{g \in U_1} \sum_{i \in U_g} \{\hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g + \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \\ &\quad - \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} \{\hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g + \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\}. \end{aligned} \tag{16}$$

However, $\sum_{i \in U_g} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g) = \boldsymbol{0}$ for each g. Hence (16) becomes

$$\begin{aligned} \hat{T} &= \hat{T}_\pi + \sum_{g \in U_1} \sum_{i \in U_g} \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_1^T \sum_{g \in U_1} \sum_{i \in U_g} \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_1^T \sum_{g \in s_1} \pi_{g1}^{-1} \sum_{i \in U_g} \bar{\boldsymbol{x}}_g \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_1^T \sum_{g \in U_1} \bar{\boldsymbol{x}}_{g1} - \hat{\boldsymbol{\phi}}_1^T \sum_{g \in s_1} \pi_{g1}^{-1} \bar{\boldsymbol{x}}_{g1} \\ &= \hat{T}_\pi + \hat{\boldsymbol{\phi}}_1^T (\boldsymbol{T}_X - \hat{\boldsymbol{T}}_{X\pi}). \end{aligned} \tag{17}$$

Notice that (17) does not include the estimator of $\boldsymbol{\phi}_2$. The least squares estimators

$$\hat{\boldsymbol{\phi}} = \begin{pmatrix} \hat{\boldsymbol{\phi}}_1 \\ \hat{\boldsymbol{\phi}}_2 \end{pmatrix}$$

are the solution of:

$$\sum_{i \in s} \pi_i^{-1} c_i (y_i - \hat{\boldsymbol{\phi}}^T \boldsymbol{z}_i)\, \boldsymbol{z}_i = \boldsymbol{0}$$

which is equivalent to:

$$\sum_{i \in s} \pi_i^{-1} c_i \{y_i - \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_1^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \begin{pmatrix} \bar{\boldsymbol{x}}_g \\ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_g \end{pmatrix} = \boldsymbol{0}.$$

By assumption, $c_i = a_g N_g$ so the first p elements of this equation are:

$$\begin{aligned} \boldsymbol{0} &= \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} a_g N_g \bar{\boldsymbol{x}}_g \{y_i - \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \\ \boldsymbol{0} &= \sum_{g \in s_1} \pi_{g1}^{-1} a_g N_g \bar{\boldsymbol{x}}_g \sum_{i \in U_g} \{y_i - \hat{\boldsymbol{\phi}}_1^T \bar{\boldsymbol{x}}_g - \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)\} \\ \boldsymbol{0} &= \sum_{g \in s_1} \pi_{g1}^{-1} a_g \boldsymbol{x}_{g1} \{y_{g1} - \hat{\boldsymbol{\phi}}_1^T \boldsymbol{x}_{g1} - \hat{\boldsymbol{\phi}}_2^T (\boldsymbol{x}_{g1} - \boldsymbol{x}_{g1})\} \\ \boldsymbol{0} &= \sum_{g \in s_1} \pi_{g1}^{-1} a_g \boldsymbol{x}_{g1} (y_{g1} - \hat{\boldsymbol{\phi}}_1^T \boldsymbol{x}_{g1}). \end{aligned}$$

Hence $\hat{\boldsymbol{\phi}}_1$ is a solution to (7). So the GREG estimator for model (9) is equal to $\hat{T}_H$ provided that $c_i = a_g N_g$.

References

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

Cholette, P. (1984). Adjusting sub-annual series to yearly benchmarks. *Survey Methodology*, 10, 35-49.

Clark, R.G., and Steel, D.G. (2002). The effect of using household as a sampling unit. *International Statistical Review*, 70 (2), 289-314.

Heldal, J. (1992). A method for calibration of weights in sample surveys. In *Workshop on uses of auxiliary information in surveys*. University of Orebro, Sweden.

Hidiroglou, M., and Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 67-78.

Lazarfeld, P.F., and Menzel, H. (1961). On the relation between individual and collective properties. In *Complex Organizations*: *A Sociological Reader*. Holt, Reinhart and Winston. 422-440.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

Luery, D.M. (1986). Weighting sample survey data under linear constraints on the weights. In *Proceedings of the Social Statistics Section*, American Statistical Association, (Alexandria, VA), 325-330.

Nieuwenbroek, N. (1993). *An integrated method for weighting characteristics of persons and households using the linear regression estimator*. Netherlands Central Bureau of Statistics.

Särndal, C., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Silva, P.L.N., and Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.

Tam, S.M. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90, 379-382.