

HOUSEHOLD-LEVEL VERSUS PERSON-LEVEL REGRESSION WEIGHT CALIBRATION FOR HOUSEHOLD SURVEYS

S. Wu¹, B. Kennedy² and A.C. Singh¹

ABSTRACT

In household surveys, the design weights are generally common for all persons in the same household. These weights are usually adjusted for nonresponse, yielding a set of new weights, called subweights. When the auxiliary information in the form of population totals is available, weight calibration is introduced to satisfy the auxiliary controls. The calibration adjusts the subweights to improve the efficiency of the estimates as well as to reduce coverage error. The calibrated weights need not be common for persons in the same household because auxiliary controls may include person-level characteristics such as age and sex. If estimates based on households are needed, each household must have a single weight. To achieve this, two methods are commonly in use: (i) each person in the household is assigned a household average value of the auxiliary variable, and then calibration is performed at the person level; (ii) each household is assigned the aggregate value for the auxiliary variable and the calibration is performed at the household level. In this paper, the above two methods are compared in the context of regression estimation by noting that they are special cases of the generalized regression estimators. Also, using the theory of optimal regression, properties of these estimators are considered for special designs, and some alternatives are suggested. A simulation study using the Canadian Labor Force Survey data is presented for comparing the two methods with respect to estimation bias and variance assuming no coverage error.

KEY WORDS: Generalized regression; modified regression; optimal regression; weight calibration; working covariance.

RÉSUMÉ

Dans les enquêtes-ménages, les poids d'échantillonnage initiaux sont généralement communs pour toutes les personnes d'un même ménage. Ces poids sont habituellement ajustés pour la non-réponse, donnant lieu à un nouvel ensemble de poids, appelé sous-poids. Lorsque l'information auxiliaire sous forme de totaux de population est disponible, des poids de calibration sont introduits pour satisfaire les contrôles auxiliaires. La calibration ajuste les sous-poids pour améliorer l'efficacité des estimateurs et réduire l'erreur de recouvrement. Les poids calibrés n'ont pas besoin d'être les mêmes pour les personnes d'un même ménage puisque les contrôles auxiliaires peuvent inclure des caractéristiques au niveau de la personne telles que l'âge et le sexe. Si des estimations basées sur les ménages sont requises, chaque ménage peut avoir un seul poids. Pour réaliser ceci, deux méthodes sont utilisées habituellement: (i) chaque personne dans le ménage possède la valeur moyenne de la valeur auxiliaire dans le ménage et la calibration est effectuée au niveau de la personne; (ii) chaque ménage se voit attribuer la valeur combinée pour la variable auxiliaire et la calibration est effectuée au niveau de ménage. Dans cet article, les deux méthodes ci-dessus sont comparées dans le contexte d'estimation par la régression en remarquant qu'elles sont des cas spéciaux d'estimateurs de régression généralisés. Aussi, en utilisant la théorie de la régression optimale, on évalue les propriétés de ces estimateurs pour des plans spéciaux, et quelques alternatives sont suggérées. Une simulation, effectuée à l'aide des données de l'Enquête sur la population active est présentée pour comparer les deux méthodes selon le biais et la variance de l'estimation en supposant aucune erreur de couverture.

MOTS CLÉS: Régression généralisée, régression modifiée, régression optimale, poids de calibration, covariance intermédiaire.

¹ Methodology Research Advisory Group, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6, wushiyi@statcan.ca, singavi@statcan.ca.

² Survey and Analysis Methods Development Section, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6, kennbri@statcan.ca.

1. INTRODUCTION

In a typical household survey, the initial design weights are common for all persons in the same household. These initial weights are usually adjusted for nonresponse to form a set of new weights, called subweights. The subweights are then calibrated to known auxiliary controls to improve efficiency as well as to adjust for coverage bias. The calibrated weights need not be common for persons in the same household because auxiliary controls may include person-level characteristics such as population counts for age and sex groups. When an estimate of a characteristic of household (e.g. household income) is needed, each household must have a single weight. Several methods have been proposed for producing this single weight. This paper discusses some of those methods and suggests a class of alternative methods.

For many household surveys, the person or household subweights are calibrated by the external demography controls using post-stratification. Since the post-stratification is applied to the persons, the resulting weights are typically not the same for all the persons in the same household. To obtain estimates based on households, a method known as "principal person" was traditionally used. In this method, a "principal person" is designated for each household. The post-stratified weight of the principal person is then assigned to the household. One problem of this method is that no matter how the principal person is designated, the resulting household weight based estimates in general will not be in agreement with the corresponding estimates based on persons. For example, the estimated total income based on households will be in general not equal to the estimated total income based on persons. To overcome this problem, Alexander (1987) considered constrained minimum distance methods in which a distance between the subweights and a set of new weights is minimized subject to external controls. The estimators are then produced using these new weights as expansion estimators. The distance can be applied to the weights at the household level or the person level. Note that the general class of constrained minimum distance methods was suggested for household weighting by Luery (1986). Also, different methods based on the generalized least squares (GLS) or chi-square distance were proposed by Zieschang (1986) for the weighting of U.S. Consumer Expenditure Surveys, Lemaître and Dufour (1987) for the Canadian Labour Force Survey, and Bankier *et al.* (1992) for the Canadian Census.

The two estimators (household-level and person-level) for different distance functions were compared with

respect to the coverage bias in Alexander (1987). In this paper, however, it is assumed that there is no coverage bias and the two estimators produced by minimizing the GLS distance are compared with respect to estimation bias and variance.

The above two methods are also compared using the theory of optimal regression. To do this, it is first observed that the two GLS-methods are special cases of the generalized regression (GREG) estimation. Deville and Särndal (1992) used the term calibration estimation for various constrained minimum distance methods, and established their asymptotic equivalence to the GREG estimation.

In this paper, Section 2 gives a brief review of the GLS methods. In Section 3, GLS-estimators are compared using the theory of optimal regression. A limited empirical study is provided in Section 4. Section 5 contains some concluding remarks.

2. CONSTRAINED MINIMUM GLS DISTANCE FOR HOUSEHOLD AND PERSON LEVEL WEIGHTING

The GLS distance is defined by $\sum_k (w_k - d_k)^2 / d_k$, where d_k and w_k are, respectively, the subweight and calibrated weight of unit k . The household-level generalized least squares (GLS-H) method treats households as units. This method selects a set of $\{w_k\}$ such that the GLS distance is minimized subject to $\sum_k x_{kj} w_k = \tau_j$, where $j=1, \dots, p$ and x_{kj} is the total of the j^{th} control variable for the household k , and τ_j is the corresponding known population total. Alternatively, the person-level generalized least squares (GLS-P) method starts with persons as units. Hence the GLS distance can be written as $\sum_k \sum_i (w_{ki} - d_{ki})^2 / d_{ki}$, where d_{ki} and w_{ki} are, respectively, the subweight and calibrated weight of the person i in the household k . However, minimizing this distance subject to $\sum_{ki} x_{kij} w_{ki} = \tau_j$, $j=1, \dots, p$ will generally produce different weights for persons in the same household. To ensure one weight per household, w_{ki} is replaced by w_k . (This is equivalent to assigning the average household value of x -variables to all persons in the same household, as in Lemaître and Dufour, 1987.) Note that $d_{ki} = d_i$ and the distance reduces to $\sum_k m_k (w_k - d_k)^2 / d_k$ where m_k is the size of the k^{th} household. As a result, GLS-P minimizes this distance subject to $\sum_k x_{kj} w_k = \tau_j$. Both methods have a closed-form solution. Let $W = (w_1, \dots, w_K)'$ be the adjusted weights. For the GLS-H,

$$W_H = D + \Gamma_H X(X' \Gamma_H X)^{-1} (\tau - X'D) \quad (2.1)$$

where $D=(d_1, \dots, d_k)'$, $\Gamma_H=diag(d_1, \dots, d_k)$, $X=(x_1, \dots, x_K)'$, $x_k=(x_{k1}, \dots, x_{kp})'$, and $\tau=(\tau_1, \dots, \tau_p)'$ is the vector of external controls. K is the total number of sampled households. For GLS-P, simply replace \underline{W}_H and Γ_H with \underline{W}_p and $\Gamma_p=diag(d_1/m_1, \dots, d_K/m_K)$ respectively. The resulting estimators are $\hat{Y}_H=Y'\underline{W}_H$ and $\hat{Y}_p=Y'\underline{W}_p$ respectively, where $\underline{Y}=(y_1, \dots, y_k)'$, and y_k is the observed total of the variable of interest y for the k^{th} household.

3. COMPARISON OF GLS-H AND GLS-P BASED ESTIMATORS

The two estimators \hat{Y}_H and \hat{Y}_p are first shown to be special cases of the GREG estimation. We then consider their design-based properties (bias and variance) in the light of optimal regression. In particular, it is shown that the estimators can be viewed as modified regression estimators (as defined by Singh, 1996) where the regression coefficients of the optimal regression is modified under a *working* covariance. Using this connection, some alternative estimators for household surveys are also suggested.

The GLS-H estimator \hat{Y}_H can be written as

$$\hat{Y}_H = Y'\underline{W}_H = Y'D + \hat{\beta}'_H(\tau - X'D) = \hat{Y} + \hat{\beta}'_H(\tau - \hat{x}), \quad (3.1)$$

where $\hat{\beta}_H = (X'\Gamma_H X)^{-1} X'\Gamma_H Y$ is the estimated regression coefficient, \hat{Y} and \hat{x} are the Horvitz-Thompson estimates of totals of the variables y and x respectively. The GLS-P estimator \hat{Y}_p can be expressed similarly. Clearly, both estimators are special cases of GREG (as defined by Särndal, 1980) and are in the class of difference estimators $\hat{Y}(\beta) = \hat{Y} + \beta'(\tau - \hat{x})$ where β is a constant vector. Minimizing the variance of $\hat{Y} + \beta'(\tau - \hat{x})$ with respect to β leads to $\beta_{opt} = Var(\hat{x})^{-1} Cov(\hat{x}, \hat{Y})$. Since β is typically unknown, by replacing $Var(\hat{x})$ and $Cov(\hat{x}, \hat{Y})$ with their estimators $\hat{V}ar(\hat{x})$ and $\hat{C}ov(\hat{x}, \hat{Y})$ respectively, we get $\hat{Y}_{opt} = \hat{Y} + \hat{\beta}'_{opt}(\tau - \hat{x})$ where $\hat{\beta}_{opt} = \hat{V}ar(\hat{x})^{-1} \hat{C}ov(\hat{x}, \hat{Y})$. This is the optimal regression estimator discussed in the literature, see, e.g., Cochran (1977, Chap. 7), Fuller and Isaki (1981), Montanari (1987) and Rao (1994), among others. If $\hat{V}ar(\hat{x})$ and $\hat{C}ov(\hat{x}, \hat{Y})$ are consistent for $Var(\hat{x})$ and $Cov(\hat{x}, \hat{Y})$, $\hat{\beta}_{opt}$ is a consistent estimator of β_{opt} . Further, \hat{Y}_{opt} is asymptotically unbiased and consistent, with minimum asymptotic variance in the class of difference estimators.

We now compare \hat{Y}_H and \hat{Y}_p from the optimal regression point of view. It follows from GREG that both are asymptotically unbiased and consistent. However, they are not optimal in general for complex designs. In particular, for a stratified household survey, they fail to take into account the aspect of stratification. Note that in

household surveys, the ultimate sampling unit is a cluster of individuals forming the household. The GLS-H treats the household as the sampling unit while GLS-P treats the individual as the sampling unit and hence misses the cluster effect in the estimation of β_p . Moreover, unlike \hat{Y}_p , the estimator \hat{Y}_H is sub-optimal in that for simple random sampling of households, it reduces to \hat{Y}_{opt} , see e.g., Singh (1996). In light of these considerations, \hat{Y}_H seems preferable to \hat{Y}_p in terms of asymptotic efficiency.

There is another useful interpretation of \hat{Y}_H and \hat{Y}_p . They can be viewed as modified optimal regression estimators (see Singh, 1996) in that the regression coefficients in \hat{Y}_{opt} are computed under suitable choices of the working covariance. Use of a working covariance is beneficial when it is cumbersome or infeasible to calculate $\hat{\beta}_{opt}$, or when there is insufficient degrees of freedom in estimating β_{opt} as in the case of a few PSUs per stratum. The resulting estimators would be asymptotically unbiased, consistent, but only sub-optimal in that they would be optimal for simpler designs corresponding to working covariances. The working covariance is chosen such that the design features are taken into account as much as possible while the $\hat{\beta}$ remains stable and its computation still manageable. One simple choice of a working covariance is the one calculated under the assumption of simple random sampling. This would result in $\hat{\beta}_H$ if households are the sampling units. More elaborate working covariances can be worked out to account for stratification and clustering effect, while maintaining the computation manageable and the degrees of freedom sufficient. The resulting estimators of Y will have different efficiencies for different choices of the working covariance. In general, a working covariance that accounts for more aspects of the design is expected to lead to a more efficient estimator of Y .

It may be noted that optimal estimators corresponding to different working covariances can be generally expressed as weight calibration estimators. Therefore, some methods as alternatives to GLS-H and GLS-P for household weight calibration can be developed.

4. EMPIRICAL COMPARISON

It has been argued that the GLS-H is expected to perform better than GLS-P. In this section we compare the two methods via simulation. As mentioned in the introduction, it is assumed that there is no coverage error.

The design and the data used in this study are adapted from Stukel, Hidioglou and Särndal (1996). A detailed description of the simulation setup can be found in their paper. For readers' convenience, we give below a brief account of the background information and the design of the simulation.

The population of the simulation study is taken to be the December 1990 sample of the Canadian Labour Force Survey from the province of Newfoundland. In the LFS, monthly sample data is collected using a multi-stage sampling design with several levels of stratification. The province of Newfoundland was first stratified into four Economic Regions (ERs) of similar economic structure. These ERs were further stratified into smaller strata. There were a total of 45 strata at the lowest level of stratification and each of these strata contained 6 or less primary sampling units (PSUs). For the purpose of simulation, a larger number of PSUs in each strata was preferred. Thus, the 45 strata were collapsed into 18 each containing 6 to 18 PSUs. While collapsing, the boundaries of ERs and the Census Metropolitan Areas (CMAs) were kept intact. A detailed description of the LFS design prior to 1991 can be found in Singh, Drew, Gambino and Mayda (1990).

The population (i.e., the LFS sample) for the simulation study contains 9152 persons. The sample design is a two-stage design. For strata in larger cities (population exceeds 15,000), three PSUs were selected at the first stage using Probability Proportional to Size with replacement (PPS-WR) sampling, where the size was the number of dwellings in the PSU. Since there were not enough dwellings per PSU to subsample, all the dwellings in the sampled PSUs were selected at the second stage. Hence this part of the design was actually a one-stage cluster design. For strata in the rest of the province, two PSUs were selected with PPS-WR at the first stage. Then one-fifth of dwellings were selected in each sampled PSU using Simple Random Sample without replacement (SRS-WOR) at the second stage. In all, 47 PSUs were selected and all the dwelling members in the sampled dwelling were included in the sample. The resulting sample size was approximately 1000 (persons). In the simulation study, 1000 samples were

drawn according to this sample design.

Three parameters of interest were chosen for evaluation: The total number of unemployed, the total number of employed in single person households, and the total number of employed in households of size four or more. Total unemployed was chosen because it is considered an important parameter. The other two were chosen to show the impact of the two methods on the household size related measures. It would also have been interesting to see the estimates of total unemployed in single person households and in households of size four and more. In this study, however, they are not considered because these would be small domain estimates and hence unstable.

The control variables contain 14 indicator variables for the 4 ERs and 10 age-sex groups. The 10 age-sex groups are 2 sex categories by 5 age categories: less than 15, 15 to 24, 25 to 44, and 65 and more. The known population totals for the 14 variables were calculated from the population of this simulation. Note that the total of the 10 age-sex groups add up the total of the 4 ERs. So a redundant variable was dropped in the actual computation.

Let Y be the population total of a variable of interest, \hat{Y}_j be the estimated total from the j^{th} sample, $j=1, \dots, R$. Define $E_s(\hat{Y})=1/R\sum_{j=1}^R \hat{Y}_j$, $V_s(\hat{Y})=1/R\sum_{j=1}^R (\hat{Y}_j - E_s(\hat{Y}))^2$. Then the percent relative bias of \hat{Y} is given by $RB(\hat{Y})\%=(E_s(\hat{Y})-Y)/Y*100$, and the percent Coefficient of Variation of \hat{Y} is given by $CV(\hat{Y})\%=\sqrt{V_s(\hat{Y})}/E_s(\hat{Y})*100$. These two quantities were used to measure the performance of the two weighting methods: GLS-H and GLS-P.

Let \hat{Y}^H and \hat{Y}^P denote the estimate given by GLS-H and GLS-P respectively. Table 1 shows the simulation results based on $R=1000$ samples. The following are observed from the table:

1. The two methods have negligible bias of similar magnitude: 1% or less.
2. The GLS-H estimates may have slightly smaller variance, more so for domains of households of size one.

Table 1. Percent Relative Bias of the estimated Y and Percent CV

	$RB(\hat{Y}_p)\%$	$RB(\hat{Y}_H)\%$	$CV(\hat{Y}_p)\%$	$CV(\hat{Y}_H)\%$
Unemployed	-0.74	-0.79	17.3	16.6
Emp. Single	-1.03	-0.55	24.7	23.4
Emp. (Size 4+)	0.51	0.46	9.9	9.9

5. CONCLUDING REMARKS

The problem of household weight calibration is discussed in the light of optimal regression. Optimal regression leads to minimum asymptotic variance. However, in general it may be infeasible or cumbersome to implement. In addition, the estimator may not be stable when there are too few of degrees of freedom. As a compromise, a class of Modified Regression estimators can be suggested. These estimators use working covariances in the calculation of the regression coefficient in an effort to (i) stabilize the results by increasing the degrees of freedom, and (ii) make the calculation feasible or easier. It is shown that both GLS-H and GLS-P are special cases of optimal regression estimation with suitable working covariances. Both theoretical considerations and empirical study suggested that while both estimators have negligible bias, GLS-H may have slightly smaller variance, and may be more so for domains of smaller household size. Sautory (1993) also performed a similar simulation study. He found that the adjustment factors for the weights from GLS-P were more variable; and no significant difference in the variances. Our findings seem to be consistent with his. It may be noted that throughout the paper, it was assumed that there is no coverage error. Ironically, Alexander (1987) found GLS-P preferable under certain models for undercoverage. This point should be investigated in future. It is also planned to compare empirically the optimal regression and a few more modified regression methods under more elaborate working covariances than that of the GLS-method.

ACKNOWLEDGEMENT

The authors are grateful to P. Lavallée for bringing their attention to this problem through his initial internal paper dated 1994.

REFERENCES

- Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Methodology Branch Working Paper, SSMD-92-007E, Statistics Canada.
- Cochran, W.G. (1977). *Sampling Techniques* (3rd Edition). New York: John Wiley.
- Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Fuller, W.A. and Isaki, C.T. (1981). Survey design under superpopulation models. In Current topics in Survey Sampling (D. Krewski, R. Platek, J.N.K. Rao, and M.P. Singh eds.), New York: Academic Press, 196-226.
- Lavallée, P. (1994). Integrated household weighting: Two different approaches. Internal memorandum, Social Survey Methods Division, Statistics Canada.
- 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. *Proceedings of the Social Statistics Section, American Statistical Association*, 325-330.
- Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.
- Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- Särndal, C.-E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- Sautory, O. (1993). Méthodes de pondération des ménages et des individus dans les enquêtes.

- Document presented at the <<XXVème Journées de statistique>>, Vannes (France), May 24-28, 1993.
- Singh, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 120-129.
- Singh, M.P., Drew, J.D., Gambino, J.D. and Mayda F. (1990). Methodology of the Canadian Labour Force Survey: 1984-1990. Catalogue No. 71-526, Statistics Canada.
- Stukel, D.M., Hidioglou, M.A. and Särndal C.-E. (1996). Variance estimation for calibration estimators: A comparison of Jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.
- Zieschang, K.D. (1986). Generalized least squares: an alternative to principal person weighting. In *Population Controls in Weighting Sample Units*, Section 2. Washington, D.C., U.S. Bureau of Labor Statistics, 1-41.