Survey Methodology, December 1987
Vol. 13, No. 2, pp. 199–207
Statistics Canada

An Integrated Method for Weighting Persons and Families

G. LEMAÎTRE and J. DUFOUR[1]

ABSTRACT

Household surveys generally use separate procedures for estimating characteristics of persons and those of families. An integrated procedure is proposed and a least-squares estimator introduced to achieve this end. The estimator is shown to be unbiased under certain general conditions. Using data from the Canadian Labour Force Survey, variances for the estimator are calculated and shown to compare favourably to those from current procedures.

KEY WORDS: Family estimation; Family weighting; Least-squares weighting.

1. INTRODUCTION

It is customary for many household surveys to incorporate in their estimation procedures a post-stratification step in which the design-based estimates of the population, generally by age and sex group, are benchmarked to independent totals obtained from demographic sources. In practice, for ease of tabulation, a weight is normally associated with each responding person, equal to the product of the inverse sampling rate, an adjustment for non-response, and an age/sex ratio adjustment factor. Estimates for a particular characteristic are then obtained by summing up the weights of all responding persons in the sample bearing that characteristic. Because of the age/sex adjustment factors, the weight so assigned will usually differ from person to person within the same household. When estimating characteristics of persons, this may not pose any particular problem; in producing estimates of households or families, however, it is not entirely clear which weight is the appropriate one to use, if any.

To estimate family characteristics, one might well elect to carry out a ratio estimation step using auxiliary information on families as well as persons. However, reliable and timely auxiliary counts of families that could be used in ratio estimation are in general not available. As a result of events such as births, deaths, marriages, divorces and persons leaving or entering a household, characteristics such as family size change from one census to the next, in ways that are less predictable than a characteristic such as age. The administrative records that are the main source of information on post-censal population change (i.e. birth, death and migration records), do not provide information on household-related change. Birth records, for example, do not provide information on the size of a family into which a child is born. Tax records can compensate in part for this deficiency (see Auger 1987); however, such records do not cover the entire population nor are they available in a timely enough fashion to be used in producing current estimates. In the absence of auxiliary counts of families, household surveys generally have adapted the weights obtained from "person-weighting" for use in estimating characteristics of families. For various reasons this is a somewhat less than ideal solution. The present paper proposes a method of estimation that results in a single uniquely defined weight per household which would be appropriate for both individual and family estimation.

[1] G. Lemaître and J. Dufour, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Techniques to achieve a single household weight have been proposed in the past, with an emphasis on using auxiliary information on persons to improve estimates of families. Oh and Scheuren (1978) proposed a method of "multivariate raking" which consists of successively ratio adjusting population estimates by post-stratum by means of the ratio adjustments calculated for each post-stratum in turn, and then iterating to convergence. The adjustments at each stage are applied to households containing persons in the particular post-stratum being adjusted for. Zieschang (1986) adopted a Generalized Least Squares (GLS) approach in which the sum of weighted squared adjustments to the design weights were minimized, subject to a set of linear constraints. Alexander (1987) examines several constrained minimum distance weighting methods, including the GLS method, and evaluates them in the context of survey undercoverage. Although the above methods were originally proposed as ways of improving estimates of families, the survey weights derived from the various estimators can clearly be used to estimate characteristics of persons as well. This paper argues in favour of adopting such an integrated approach to individual and family estimation. Section 2 discusses the limitations of the current approaches to estimating characteristics of persons and families. Section 3 introduces a model-based estimator adapted from a generalized weighting procedure due to Bethlehem and Keller (1987). Section 4 presents some empirical results taken from the Canadian Labour Force Survey. Section 5 discusses plans for further study.

2. CURRENT ESTIMATION PROCEDURES

The principal mandate of most household surveys traditionally has been to produce estimates for characteristics of persons, particularly of labour force characteristics. Such surveys adopt the household as the ultimate sampled unit essentially for reasons of cost and convenience. Although the household unit is normally respected in preliminary weighting steps (non-response adjustments, rural/urban adjustments, etc.), it is generally ignored in the final weighting step, i.e. no allowance is made for the fact that the members of a household are sampled as a unit. In particular, any coverage biases associated with the sampled unit are not directly taken into account or compensated for in estimation. Undercoverage is thus assumed to be ignorable in the sense of Rubin (1976); every person in an age/sex post-stratum is treated the same in estimation whether he/she is living alone or comes from a multi-person household. One study of non-response in the Labour Force Survey (Paul and Lawes 1983), however, has demonstrated that smaller households, particularly households without children, tend to be underrepresented in the sample. Although no comparable studies exist for missed households in the Labour Force Survey, studies of private household undercoverage in the census have shown that non-enumerated households are indeed smaller on average than enumerated households (Gosselin and Théroux 1980). A missing-at-random type procedure can lead to biases in labour force estimates for persons, particularly if the labour force distribution of persons in smaller households is different from that of persons in larger ones, all things being equal. Intuitively, an estimation procedure which takes into account (even if only indirectly) the fact that smaller households are more subject to non-response and undercoverage than larger ones could correct in part for this deficiency in the sample.

In the absence of auxiliary information on households or families that could be incorporated into an appropriate weighting procedure to produce a well-defined family weight, many current methods adopt as the family weight the weight of a "principal person" in the family. In the Canadian Labour Force Survey, this person is the female spouse if present, otherwise the head. Since such methods do not take household composition into account,

family estimates generated using this weight tend to overestimate larger families and to underestimate unattached persons. In addition many characteristics (e.g., population, income) can be estimated using either the individual weight or the family weight, and the estimates will in general disagree, sometimes substantially. Of course even under ideal sampling and interviewing conditions, with no differential non-response or undercoverage, family and individual-based estimates of the same characteristic will disagree somewhat. With a large enough sample, however, the discrepancies should be small. Under actual, i.e., less than ideal conditions, differences may be too large to explain away by a facile appeal to sampling variability. An estimation procedure that yields a single household weight which, when used as an individual weight, respects the auxiliary population totals will eliminate the awkwardness of having two estimation systems. It is these deficiencies that the estimator described in the following section was designed to deal with.

3. A PROPOSED ESTIMATOR

We begin by introducing a generalized weighting procedure based on linear models due to Bethlehem and Keller (1987) and applying it first to person-based estimation as was done in their paper. A modification of the procedure is introduced which leads to household weights appropriate for estimating characteristics of persons. We will borrow freely from their original presentation in what follows.

Assume a survey target population consisting of N units, an N-vector Y of values of a target variable, and an N by p matrix X of auxiliary variables defined for each unit of the target population. The population totals for each auxiliary variable are assumed to be known and will be denoted collectively by the p-vector x. In our application x will consist of age-sex totals. If the auxiliary variables are correlated with the target variable, then for an appropriate p-vector B, the values of $E = Y - XB$ will vary less than the values of the target variable Y. Ordinary least squares on all units of the target population yields

$$B = (X'X)^{-1}X'Y, \tag{3.1}$$

provided X is of full rank. A sample-based estimate for B is given by

$$\hat{B} = (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}TY, \tag{3.2}$$

where T is a diagonal matrix whose i-th element is 1 if the i-th unit of the population is in the sample, 0 otherwise, and $E(T) = \pi$.

It can be shown that for large samples $\hat{B}$ will be approximately unbiased. The parameter of interest, however, is not B but the population total y. If we define $\hat{y} = \hat{B}'x$, $\hat{y}$ will be an approximately unbiased estimator of y provided that $B'x = y$, or equivalently, provided the sum of the residuals for the population model $Y = XB + E$ is equal to zero. This will hold if the N-vector whose elements consist of ones is in the space spanned by the columns of X, and in particular, if the auxiliary variables X include an exhaustive and mutually exclusive set of indicator variables (for age/sex groups, for example).

If we write $\hat{y} = \hat{B}'x = Y'\Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x$, we see that the estimator implicitly defines an N-vector of weights given by

$$W = \Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x,$$

that do not depend on the particular target variable being estimated. If these weights are used to produce sample estimates for the auxiliary variable characteristics, we have that $X'W = x$, so that the weights do indeed yield the appropriate population totals. Furthermore if X consists exclusively of an exhaustive and mutually exclusive set of indicator variables, then the regression estimator $\hat{y}$ will be equivalent to the ordinary post-stratification estimator. For further details, see Bethlehem and Keller (1987).

The weight of an arbitrary sample person i under this procedure can be expressed generally as

$$W_i = \sum_j \frac{x_{ij} b_j}{\pi_i}, \tag{3.3}$$

where $(b_1, \ldots, b_p) = (X'\Pi^{-1}TX)^{-1}x$ and π_i is the inclusion probability for person i. This suggests that the estimation method described above can be adapted to yield the desired weights by defining the auxiliary variables in the same way for all household members. An obvious way to do this is to define auxiliary variables at the household level, for example by replacing the corresponding variables defined at the person level by the household mean. More formally let Z be an N by p matrix defined for person i $(i = 1, \ldots, N)$ belonging to household h $(h = 1, \ldots, H)$ by

$$Z_{ij} = \frac{U_{hj}}{n_h},$$

where U_{hj} is the total for characteristic j in household h, i.e. $U_{hj} = \Sigma_k X_{kj}$, with the summation being over all members k of household h, n_h = size of household h, and $\Sigma_h n_h = N$. Let Y again be an N-vector of values for an arbitrary target variable defined on *persons*. As in person-level estimation, we work with the population model $Y = ZC + E$ and apply least squares to the sample data to obtain an estimate

$$\hat{C} = (Z'\Pi^{-1}TZ)^{-1}Z'\Pi^{-1}TY. \tag{3.4}$$

We define $\hat{y} = \hat{C}'x$ where x is again the vector of population totals for the auxiliary variables. $\hat{y}$ will be an approximately unbiased estimator of y provided the N-vector of ones is in the space spanned by the columns of Z. In a manner analogous to (3.3), the weight for an arbitrary sampled person in household h will be given by

$$W_h = \sum_j \frac{U_{hj} c_j}{\pi_h n_h}. \tag{3.5}$$

Since each household member contributes the same row vector to Z and since each has the same first order inclusion probability, each person within a household will have the same weight. Furthermore the use of the household weight as a person weight yields the correct auxiliary population totals. Although it is possible to obtain negative weights under this procedure (if some of the c_j's are less than zero), for well-behaved samples (i.e., not subject to serious non-response or undercoverage) households whose weights are changed substantially by this procedure tend to be households of unusual composition that are uncommon in the sample and in the population at large. Recently in weighting twenty-four months of Labour Force

Survey data under this procedure, only one household had a (small) negative weight attributed to it. Negative weights are problematic because it is difficult to attach the usual meaning one assigns to weights, that is, the number of persons/households in the population at large represented by a particular sampled person/household. However, under the formulation described above, the final weights are defined only implicitly and indeed could be viewed as merely a convenient means of generating estimates. In practice even with some negative weights, it is unlikely that a meaningful estimate of level for a characteristic of interest would turn out negative. The problem of explaining a negative weight to a mystified user is of course a different question.

The variance of the estimator $\hat{y} = \hat{C}'x$ described in this paper can be obtained using methods described in Fuller (1975). In addition the estimator can be shown to be equivalent to the GLS estimators proposed by Zieschang (1986) and Alexander (1987) when the space spanned by the auxiliary variables Z contains a vector of ones. Further properties of this type of estimator can be found in Wright (1983).

4. EMPIRICAL RESULTS

The Canadian Labour Force Survey is a monthly rotating panel survey of approximately 48,000 households across Canada (see Platek and Singh 1976 and Singh, Drew, and Choudhry 1984). Households once selected remain in the sample for six consecutive months before being replaced. The primary geographic strata are the ten provinces. Sample sizes vary from a low of 1500 households in Prince Edward Island, the smallest province, to about 9000 households in Ontario, the most populous one. The survey collects data concerning the labour market situation of respondents during a reference week each month and publishes a wide variety of estimates related to the nation's labour supply.

A preliminary evaluation of the estimator described above was carried out using data from one of the monthly surveys. May 1981 was chosen to permit comparisons to results from the 1981 census held at about that time. Although we have been using the terms "household" and "family" interchangeably up to now, user interest is often focused on estimates of "economic families", which consist of all persons in a household related by blood, marriage, or adoption. For weighting purposes it is conceptually more appealing to deal with the actual sampled unit, i.e. the household. However, the empirical results presented here will be based on estimates for economic families. The evaluation carried out focused on both characteristics of persons (labour force status) and of families (number of economic families and number of unattached persons). The least-squares weighting was carried out for two sets of five-year age/sex groups, with persons seventy and over being grouped according to sex. The first set of (twenty-four) age/sex groups excluded children 0 to 14 years of age from the weighting, to permit a comparison to a standard person-based post-stratification estimator using the same auxiliary information. The second set included children grouped into six age/sex groups and was used only for least-squares weighting, since under standard post-stratification the weighting of children would have no effect on the weighting of persons 15 and over.

Although all estimators considered are approximately unbiased for estimates of characteristics of persons, each makes different assumptions about the nature of undercoverage and non-response. (The Labour Force Survey's non-response adjustment procedure assumes that non-responding households are missing at random within geographic area). The post-stratification estimator implicitly assumes that any differential non-response and undercoverage depends only on age and sex and is therefore adequately compensated for by

person-based estimation using auxiliary information on these characteristics. Under least-squares weighting, the weight of a person will depend on the age/sex composition of the household (without children in one case, with children in the other). Thus, all things being equal, one would expect the design weight of a person belonging to an age/sex group subject to substantial undercoverage to be adjusted less if that person is living with persons belonging to age/sex groups well covered by the sample than if he/she is living alone.

Since the auxiliary population totals by age and sex are available by province, estimation was carried out separately for each province. However, the smaller provinces have been collapsed into two groups in the following tables.

In general the three estimators do not yield substantially different estimates, particularly A and B. The inclusion of children in the weighting does appear to lead to slightly higher estimates of employment and of unattached persons and slightly lower estimates of economic families nationally and in the larger provinces (compare results from Scheuren *et al.* 1981). This is in line with expectations, although there is still some ground to cover vis-a-vis census results, which show (rounded to thousands) 6,369,000 economic families and 2,583,000 unattached persons at the national level. The moral of the tale is that, although the least-squares estimator does take us part of the way home (when the presence of children is taken into account), it will require accurate and timely auxiliary information to eliminate the residual bias.

Table 1

Number of Persons Employed and Unemployed, Number of Economic Families and Unattached Persons, Labour Force Survey, May 1981 (In Thousands)

	Estimator[a]	Employed	Unemployed	Economic Families	Unattached Persons
Canada	A	11,094	850	6,424	2,432
	B	11,090	850	6,446	2,442
	C	11,120	851	6,410	2,495
Atlantic Region	A	819	102	563	156
	B	819	102	570	154
	C	821	102	569	156
Quebec	A	2,725	304	1,723	587
	B	2,724	304	1,725	596
	C	2,735	305	1,714	614
Ontario	A	4,198	274	2,325	863
	B	4,200	273	2,325	861
	C	4,211	273	2,310	881
Prairie Region	A	2,074	83	1,078	506
	B	2,072	84	1,089	510
	C	2,074	83	1,085	517
British Columbia	A	1,277	88	735	319
	B	1,276	88	738	321
	C	1,280	88	734	327

[a] A = post-stratification/principal person, B = least squares with children excluded from weighting and C = least squares with children included in weighting.

The expected performance of the least-squares estimator with regard to efficiency is not altogether obvious. Certainly, if one were to base a prediction on the results observed above, then the similarity of the estimates to those produced by the post-stratification estimator would lead one to expect it to perform as well as the latter. On the other hand, one might expect efficiency gains for estimates of economic families, because of the fact that the least-squares estimator makes use of the auxiliary population totals in determining the household weight. However, a single weight per household is not achieved without some redistribution of weights at the micro level.

Table 2

Distribution of Percent Deviations of Final Weights Relative to the Design Weights, Labour Force Survey, May 1981

Percent Deviation	Percentage of Total Sample		
	Post-Stratification	Least-Squares	Least-Squares (With Children)
> −30%	0.0	0.1	0.2
−30 to −20%	0.0	0.5	0.9
−20 to −10%	0.6	3.0	5.3
−10 to 0%	23.9	20.4	27.1
0 to 10%	53.9	44.6	37.3
10 to 20%	20.6	26.3	21.6
20 to 30%	0.6	4.4	6.2
30 to 40%	0.1	0.4	0.9
40 to 50%	0.0	0.0	0.2
< 50%	0.0	0.0	0.2

Note: Sample size is N = 159014.

Table 3

Estimated Efficiencies of Least-Squares Estimators Relative to Post-Stratification Estimator, Labour Force Survey, May 1981

Estimator[a]		Employed	Unemployed	Economic Families	Unattached Persons
Canada	B	1.044	0.999	1.565	1.038
	C	1.066	0.999	1.616	1.036
Atlantic Region	B	1.110	0.977	1.266	0.998
	C	1.193	0.992	1.567	1.070
Quebec	B	1.059	1.005	1.553	1.020
	C	1.063	0.992	1.582	0.992
Ontario	B	1.028	1.011	1.825	1.064
	C	1.059	1.010	1.828	1.037
Prairie Region	B	1.001	1.009	1.205	1.009
	C	1.072	1.066	1.420	1.134
British Columbia	B	1.038	0.964	1.248	1.048
	C	1.053	0.978	1.203	1.045

[a] B = least squares with children excluded from weighting and C = least squares with children included in weighting.

As Table 2 illustrates, the least-squares weights have a somewhat greater dispersion than those based on standard post- stratification methods. Including children in the weighting results in an even greater dispersion. The movement in the weights essentially reflects the extent to which the age/sex household size composition of the sample fails to mirror that existing in the general population. Since the objective of a single weight per household imposes an additional constraint on the estimation procedure, one might expect variances to suffer somewhat, particularly if no additional auxiliary information is brought to bear in estimation.

Variances for the post-stratification estimator were estimated using the Keyfitz method (1957) with PSU's (primary sampling units) or collapsed PSU's as replicates. The least-squares variances were estimated using the method described in Fuller (1975). To ensure comparability, variances for several characteristics estimated by means of post-stratification were calculated using the Fuller technique and compared to those from the Keyfitz approach. In all cases the two sets of variance estimates were very close (within one or two percent).

Table 3 summarizes the estimated efficiencies of the least-squares estimators relative to post-stratification for the characteristics considered in Table 1. The efficiency gains for estimates of economic families are substantial. Estimates of persons employed and of unattached persons also appear to gain somewhat; however, the variance reductions for these characteristics are small, with the exception of employed in the Atlantic Region, particularly when children are included in the weighting. Interestingly average family sizes in the Atlantic Region are higher than in the rest of the country, although it is not clear how this would affect estimates of employed persons. The variances for the characteristic unemployed are essentially unaffected by the least-squares procedure. One can probably expect these results to hold in general, i.e. for arbitrary characteristics. Although the one-weight-per-household criterion is a restrictive one for estimates of characteristics of persons, the least-squares estimators appear to compensate through the additional "explanatory" variables of the linear model, i.e. the household means of all auxiliary variables. The above preliminary results suggest that individual and family estimation could be integrated at little or no loss in efficiency for estimates of persons.

5. PLANS FOR FURTHER STUDY

The results presented in this paper are preliminary, and a more extensive empirical evaluation of the properties of the least-squares estimator is currently under way, with particular attention being given to the behaviour of estimates over time and to efficiencies for a larger group of characteristics relative to estimates produced with the Labour Force Survey's current raking ratio estimator. The foregoing results have suggested that at least for some characteristics of persons, the "explanatory power" of the age-sex composition of a household is at least as great as that of the age-sex group alone. It will be instructive to see if the relative efficiencies will be as favourable for characteristics more strongly correlated with age-sex. In addition although in practice negative weights have been uncommon, it is likely that some procedure must be developed to deal with them when they occur. Among the possibilities one might consider would be to accord them outlier treatment or perhaps to forestall their occurrence by imposing some bound on changes to the weights (Zieschang 1987). Finally it would be useful to make explicit the undercoverage model underlying the least-squares estimator to permit an evaluation of the model on its own merits.

ACKNOWLEDGEMENTS

The author would like to thank F. Scheuren for his comments and suggestions regarding this paper.

REFERENCES

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

AUGER, E. (1987). Family data from the Canadian personal income tax file. In *Statistics of Income and Related Administrative Record Research: 1986-1987,* (Eds. W. Alvey and B. Kilss), Washington, D.C.: Internal Revenue Service, 177-184.

BETHLEHEM, J.C., and KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics,* 3, 141-153.

FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā,* C37, 117-132.

GOSSELIN, J.-F., and THÉROUX, G. (1980). 1976 Census of Canada Quality of Data Series I: Sources of Error - Coverage. Catalogue No. 99-840, Statistics Canada.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association,* 52, 503-510.

OH, H.L., and SCHEUREN, F. (1978). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

PAUL, E.C., and LAWES, M. (1982). Characteristics of respondent and non-respondent households in the Canadian Labour Force Survey. *Survey Methodology,* 8, 48-85.

PLATEK, R., and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.

RUBIN, D.B. (1976). Inference on missing data. *Biometrika,* 63, 581-592.

SINGH, M.P., DREW, J.D., and CHOUDHRY, G.H. (1984). Post '81 censal redesign of the Canadian Labour Force Survey. *Survey Methodology,* 10, 127-140.

SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Studies from Interagency Data Linkages, Report No. 10: Methods of Estimation for the 1973 Exact Match Study. U.S. Department of Health and Human Services, Social Security Administration, SSA Publication No. 13-11750.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association,* 78, 879-884.

ZIESCHANG, K.D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

ZIESCHANG, K.D. (1987). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. Unpublished manuscript, U.S. Bureau of Labor Statistics.