1. Introduction

Generally, one would expect a household survey to be an excellent vehicle for the production of estimates of families and their characteristics.[1] However, the principal mandate of most such surveys traditionally has been to produce estimates of individuals, particularly with respect to labour force characteristics; the household has been adopted as the ultimate sampled unit essentially for reasons of cost and convenience. Once sampling and interviewing have been completed, the household as a unit usually vanishes from the scene, with estimation procedures in particular making no allowance for the fact that the individuals in a household are sampled as a unit. Certainly one reason for this is the absence of current and reliable auxiliary information on households or families that could be used in ratio estimation. Characteristics such as family size are subject to sudden change, and the administrative records that are our main source of information on post-censal population change provide an incomplete and imperfect accounting of such change.

It is perhaps not surprising that the point of departure for family estimation has usually been the final weighted survey file of individuals, with a family weight being derived from the weights of one or more individuals within the family. For various reasons, this is a somewhat less than ideal solution. This paper deals with certain problems and issues underlying family estimation and proposes several solutions designed to improve on current methods.

Section 2 presents results on traditional methods for estimating families and discusses some limitations of these methods. It considers as well the question of household size bias and presents one method of compensating for it.

Section 3 discusses more generally desirable features one would hope to see in a family estimator, and introduces an estimation method which possesses several of these features. The section includes an evaluation of the estimator using data from the Canadian Labour Force Survey. Finally Section 4 describes plans for future work.

2. Traditional Methods for Estimating Families

Household surveys almost invariably incorporate as part of their estimation procedures a ratio estimation step carried out by age and sex group. The adjustment ratios calculated at aggregate level are then carried down directly to sampled individuals for micro-level weighting. As a result, individuals within the same household or family generally have different weights, and it is not entirely clear how these weights can be applied or adapted for family estimation. One expedient would be to use the design weight after compensation for non-response (the "subweight"), which is a household-level weight. However, most household surveys are subject to undercoverage, so that the use of this weight would usually result in underestimates of the number of families. Most other strategies involve the choice for each family of an individual family member's final weight (ie. after ratio adjustment) for use as the family weight in tabulations. Typical choices have involved the weight of the head of the family or that of the female spouse. Still another strategy involves the use of the harmonic mean of the weights of all family members (including children)[2], the harmonic mean being the expected value of the family weight if one subsamples family members with equal probability to obtain a family weight. The relative performance of these methods for estimating economic families has been evaluated by comparing results from the May 1981 Canadian Labour Force Survey with June 1st 1981 Census tabulations adjusted for differences in coverage (Levesque 1985). The comparison is summarized in Table 1.

The survey estimates in the table are of course subject to sampling variability so that the differences relative to census figures cannot, strictly speaking, be interpreted as biases. However, it is apparent from the table that estimates of total families using these methods are systematically overestimated, whereas unattached individuals tend to be seriously underestimated. In part this reflects the fact that non-respondents in the Labour Force Survey tend to come from smaller households (Paul and Lawes 1982). Studies of private household undercoverage have shown that missed (ie., non-enumerated) households, also tend to be smaller on average than enumerated households(Statistics Canada 1980). Finally, persons missed from enumerated households would of course reduce the observed household size and contribute to the observed bias as well.

Although the harmonic averaging of all family members' weights results in smaller biases than the choice of either the weight of the head or that of the female spouse, there are some obvious deficiencies in this type of approach to family estimation. In particular, the estimates of unattached individuals are the same for all methods; indeed it appears somewhat anomalous to be claiming bias reductions for families of size 2 and over without corresponding reductions in bias in the estimates of the complement, i.e., unattached persons. Such methods provide no mechanism for transferring to the unattached the bias reductions in the estimates of total families.

A more direct way to achieve bias reductions is through the use of auxiliary information. Although such information is not readily available by size of economic family, post-censal estimates of the Canadian population by age, sex and marital status do exist, and there is obviously a strong correlation between the characteristics "unattached" and "single" that can be used to good account in ratio estimation.

---

[1] A family (i.e., an economic family) is defined as all individuals within a household related to one another by blood, marriage, or adoption. Economic families of size one are referred to as "unattached individuals". In most cases (over 95% of the time) the household and the economic family are one and the same unit. Although most analyses concern themselves with economic families, the sample design often makes it more convenient to work with households. Clearly an improvement in household estimates should have a favourable impact on family estimates as well.

[2] Suggested by G. Feeney, Australian Bureau of Statistics, personal communication.

An evaluation of the use of marital status information in ratio estimation has been carried out with Canadian Labour Force Survey data. The marital status variable collected by the Survey classifies individuals in one of the four following categories:

1. Now married or living common-law
2. Single (never married)
3. Widowed
4. Separated or divorced

Because of sample size limitations, it was not possible to retain the complete breakdown for each age and sex group (24) commonly used in ratio estimation . Automatic Interaction Detection (Sonquist and Morgan 1964) was used to determine the optimal groupings by age, sex, and marital status with respect to economic family size. Although the groupings varied somewhat from province to province, the following groupings emerged as the most common and practicable ones:

Age Groups	Marital Status Grouping
15-19	One group
20-24	Single/Other
25-29, 30-34, 35-44	Married/Single/Other
45-54, 55-59, 60-64, 65+	Married/Other

Sex was of minor importance in the groupings. Children 0 to 14 years of age were included in three five-year age groups. The complete cross-classification consisted of 46 age/sex/marital status groups. Table 2 summarizes the impact of including marital status information on the estimates of families and unattached individuals.

The results show in most cases reductions in bias in the estimates of both total families and unattached individuals, regardless of whether the head's weight, the female spouse's weight, or the harmonic mean is used as the family weight. Results by family size are somewhat uneven. In practice, the gains may not be as great as suggested by the results of this evaluation, since the quality of post-censal populations estimates in general will not be as good as that of census figures. Procedures for generating post-censal population estimates use census figures as a base and project population growth by accounting for the components of change from the date of census to the date of estimation. In the case of marital status, the components of change are obtained from administrative records on deaths, marriages, and divorces. Although common-law unions are included in the census figures, there are no administrative records that cover changes between censuses, so that post-censal figures essentially assume no change in the number of common-law unions since census.

In addition, in the derivation of the post-censal population estimates, separated persons are combined with persons who were married at time of Census and then "aged" with the married population thereafter, which would make retention of the optimal marital status groupings problematical. However, it would be possible to break out separated persons as a distinct category in the post-censal estimates if one assumed that for persons 25 years of age and over, the proportion of separated persons among all divorced and separated persons had remained approximately the same since the last census. In practice, this would likely be the procedure utilized, although it is not yet clear what impact it would have on the efficacy of the marital status adjustment.

3. A Proposed Family Estimator

Although the adjustment procedure described above provides estimates of families that, empirically at least, appear to be only slightly biased, tabulations using any of the various family weights defined by these methods will inevitably run into problems of consistency with estimates of individuals.

The reason is that many characteristics of interest to analysts of family data can be estimated by means of either individual weights or family weights. Analysts working with sample-based family estimates generally expect them to obey the same sort of consistency rules as do census figures for the entire population. Among such rules are the following:

- the number of families of a particular size times the family size, summed over all family sizes, should equal the total population;
- the number of male spouses in families in which both spouses are present should be equal to the number of female spouses in such families;
- the total income of families by size should equal the total income of individuals in families of the corresponding size.

The list could be extended indefinitely. Because of sampling variability, and the presence of non-response and coverage bias, such relationships will rarely hold for sample estimates under the family weighting schemes described in the previous section. However, under ideal sampling conditions (i.e. no non-response or coverage biases), they can be expected to hold approximately, and this may be sufficient for most analytical purposes. On the other hand, under a weighting scheme yielding a single weight per household applicable to all members of the household (i.e producing the appropriate population totals when used as an individual weight) the relationships described above would necessarily hold, as they of course do for the total population or for any subgroup of the population. Achieving this result would require modifying the usual weighting scheme in which the adjustment ratios calculated at aggregate level are applied directly to the design weights of sampled individuals. A reasonable strategy might involve making the household weight depend on the age/sex composition of the household, so that sampled persons belonging to age/sex groups subject to substantial undercoverage, for example, would have their weights adjusted less if they happened to be living with persons who are relatively well represented in the sample.

A regression-based method of weighting due to Bethlehem and Keller (1983) can be adapted somewhat to attain this objective. Let Z be an n by p design matrix of p variables defined on n sampled individuals as follows:

For individual j ($j = 1, \ldots, n_i$) of household i ($i = 1, \ldots, h$),

let $$Z_{ijk} = \frac{X_{ik}}{h_i},$$

where X_{ik} = number of persons with characteristic $k (k = 1,2,\ldots,p)$ in household i

h_i = household size

and $$\sum_{i=1}^{h} n_i = n.$$

We assume that auxiliary population totals are available for the p characteristics above. Note that all members of a household contribute the same row vector to the matrix Z. Now let Y be an n by q matrix of target variables defined on sampled individuals and Π the n by n matrix of first order inclusion probabilities (identical for all members of a household). The auxiliary variables are assumed to be correlated with the target variables. Then for a suitable p by q matrix B of regression coefficients, the elements of E = Y - ZB will vary less than the values of the target variables. An estimator for B based on sample data is given by

$$\hat{B} = (Z'\Pi^{-1}Z)^{-1}Z'\Pi^{-1}Y$$

The regression estimator $\hat{y}_r$ of the population totals y is then defined by $\hat{y}_r = \hat{B}'x$, where x is the vector of auxiliary population totals. But $\hat{y}_r = Y'w$ where $w = \Pi^{-1}Z(Z'\Pi^{-1}Z)^{-1}x$, so that the regression estimator implicitly produces an n-vector of weights which are the same for all members of a household.

Since each household member contributes the same row vector to Z and since each has the same first order inclusion probability, the term within parentheses above is equal to $X'(\tilde{\Pi}H)^{-1}X$ where Π is the diagonal matrix of household inclusion probabilities and H the diagonal matrix of household sizes. Hence w_h, the vector of household weights can be written as

$$w_h = (\tilde{\Pi}H)^{-1}\, X\left[X'\,(\tilde{\Pi}H)^{-1}\, X\right]^{-1}x.$$

The estimate of the auxiliary population totals is then given by $X'w_h$, which is equal to x. Thus the use of the household weight as an individual weight yields the correct auxiliary population totals.

It may be of interest to note that the household weight w_i for household i can be expressed as

$$w_i = \sum_k \frac{X_{ik}\,\alpha_k}{\pi_i h_i}$$

where $(\alpha, \ldots, \alpha_p)' = \left[X'\,(\tilde{\Pi}H)^{-1}\, X\right]^{-1} x$ and π_i is the inclusion probability for household i.

The household weight is thus seen to be an average of ajustment factors $\{\alpha_k: k = 1, 2, \ldots, p\}$ applied to each household member, where the α_k are determined by the constraint that the auxiliary variable population totals must be respected.

Table 3 presents estimates of families for the regression-based estimator described above. The independent variables were the same age/sex/marital status groupings used earlier in ratio estimation, with the exception that only two marital status categories (married & other) were adopted for the three age groups in the 25-44 range. The regrouping was carried out because the finer breakdown resulted in some parameter estimation problems in certain of the smaller provinces. The estimates of families obtained with this one household/one weight approach and using essentially the same auxiliary information as previous methods are comparable to the better results obtained via a simple ratio estimation approach. Again, however, estimates by household size are of somewhat uneven quality.

The benefits of consistency between individual and family estimates yielded by this approach are obtained at a price, however. The attainment of a single weight per household which yields the appropriate population totals when used as an individual weight necessitates some redistribution of weights at the micro level. Table 4 summarizes the percent deviations relative to the subweight of the final weights for the regression-based estimator compared to the post-stratification estimator incorporating the marital status adjustment.

The regression-based weights have a somewhat greater dispersion than those based on standard post-stratification methods and reflect the extent to which the age/sex/household size composition of the sample fails to mirror that existing in the general population. It could perhaps be argued that the imposition of a one household/one weight requirement in weighting an imperfect sample is a somewhat artificial one, particularly when the sample is subject to age/sex or household size biases or when the auxiliary variable categories are differentially represented in the sample. Under such circumstances, it is less than clear what the properties of the resulting estimator will be. In addition, estimating the sampling variance may be more complex under this approach. On the other hand, the age/sex composition of sampled households does provide additional information concerning sampled individuals and to the extent that household composition is associated with non-response or undercoverage, the use of such information in estimation may result in improvements in estimates of individual characteristics.

4. Plans for Further Investigations

Although the use of ancillary information on marital status seems likely to result in some reductions in bias for estimates of total families and unattached persons, the somewhat imperfect association between marital status and family size does not ensure a corresponding reduction in the variability of the resulting estimates. However, recent studies in demography have revealed that post-censal estimates of the total number of census families [3] are of much better quality than was originally thought to be the case and indeed, incorporate a growth factor for increasing the number of common-law families from the census base figure. Since over 95% of economic families are also census families, the use of this source of auxiliary data should considerably stabilize the estimates of total economic families as well as ensure less deterioration in accounting for common-law unions as one moves away from the census base. Plans are to evaluate the impact on economic family estimates of the use of census family data, independently of and in conjunction with marital status information.

In addition to the household size bias mentioned earlier, it is known that the Labour Force Survey is subject also to a non-response bias by month in sample (panel or rotation group bias), with non-response rates being higher for households in the sample for the first time (Paul and Lawes 1982). Not a few of the problems associated with family estimation may be associated with these two biases acting in conjunction. The availability of an independent estimate of total

[3] The term census family refers to a husband and a wife (with or without children who have never married regardless of age), or a lone parent with one or more children who have never married (again regardless of age) living in the same dwelling.

census families would make it possible to adjust the sample to ensure equal representation by census family size for each survey panel.

Finally a Monte Carlo study is planned to evaluate the properties of the one household/one weight approach to weighting. The bias and variance of the regression-based estimator will be calculated for key individual and household characteristics and compared to the bias and variance of the standard post-stratification estimator.

Summary and Conclusions

In this paper we examine traditional methods of weighting families and point out some limitations of these methods. The use of ancillary information on marital status is evaluated as a means of reducing household size bias. A regression-based estimator yielding one weight for all household members is introduced; family estimates are calculated and found to compare favourably to those obtained from traditional methods, although resulting in a greater dispersion in the distribution of final weights.

REFERENCES

Bethlehem, J.C. and W.J. Keller (1983). A Generalized Weighting Procedure Based on Linear Models. Proceedings of the American Statistical Association, Section on Survey Research Methods, 1983, 70-75

Levesque, J-M (1985). Weighting of Family Data in the Labour Force Survey. Internal Report, Labour and Household Surveys Analysis Division, Statistics Canada, Ottawa.

Paul, E.C. and M. Lawes (1982). Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey. Survey Methodology, 8, numbers 1 & 2, 48-85.

Sonquist, J.N. and J.A. Morgan (1964). The Detection of Interaction Effects. Monograph no. 35, Survey Research Center, Institute for Social Research, University of Michigan.

Statistics Canada (1980), Catalogue 99-840, 1976 Census of Canada, Quality of Data (Series I: Sources of Error - Coverage), Ottawa.

TABLE 1
Absolute and Percent Differences Relative to Census, Economic Family Estimates, Canada and Regions, Labour Force Survey, May 1981

Differences in thousands		Total Families[1]		Family Size 2		3		4		5+		Unattached Individuals	
		Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ
Canada	A	116	1.8	21	0.9	21	1.5	57	3.8	16	1.5	-156	-6.0
	B	82	1.3	26	1.1	12	0.9	41	2.7	3	0.3	-156	-6.0
	C	47	0.7	16	0.7	6	0.4	28	1.9	-3	-0.0	-156	-6.0
Atlantic Region	A	12	2.2	9	5.6	-1	-0.8	7	5.5	-3	-2.6	- 6	-3.7
	B	4	0.7	7	4.2	-3	-2.5	5	3.7	-5	-3.9	- 6	-3.7
	C	8	1.5	8	4.8	-2	-1.5	6	4.4	-4	-3.0	- 6	-3.7
Quebec	A	31	1.8	-8	-1.4	10	2.5	18	4.2	12	4.0	-63	-9.8
	B	27	1.6	3	0.6	9	2.4	10	2.4	4	1.4	-63	-9.8
	C	12	0.7	-5	-0.8	5	1.3	7	1.8	4	1.4	-63	-9.8
Ontario	A	34	1.5	19	2.3	3	0.6	12	2.1	-1	-0.1	-37	-4.1
	B	42	1.8	21	2.6	2	0.5	15	2.7	4	0.9	-37	-4.1
	C	18	0.8	18	2.1	-1	-0.3	5	0.9	-3	-0.8	-37	-4.1
Prairie Region	A	20	1.9	-3	-0.8	8	3.5	10	3.9	6	2.9	-15	-2.8
	B	-1	-0.1	-7	-1.7	3	1.4	3	1.1	-1	-0.4	-15	-2.8
	C	1	0.1	-6	-1.4	5	2.1	3	1.2	-1	-0.4	-15	-2.8
British Columbia	A	18	2.5	4	1.2	1	0.9	11	6.4	3	2.7	-35	-10.0
	B	11	1.5	0	0.0	0	0.0	9	5.4	2	1.6	-35	-10.0
	C	8	1.1	1	0.3	-1	-0.7	7	4.2	1	1.0	-35	-10.0

A = weight of family head
B = weight of female spouse
C = harmonic mean

[1] Excluding unattached individuals

TABLE 2
Absolute and Percent Differences Relative to Census, Economic Family Estimates, Canada and Regions, Labour Force Survey, May 1981 (with Adjustment by Marital Status)

Differences in thousands		Total Families[1]		Family Size 2		3		4		5+		Unattached Individuals	
		Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ
Canada	A	51	0.8	16	0.7	8	0.6	32	2.1	- 6	-0.5	-58	-2.2
	B	18	0.3	22	0.9	5	0.4	10	0.7	-19	-1.7	-58	-2.2
	C	12	0.2	4	0.2	-2	-0.1	18	1.2	-9	-0.8	-58	-2.2
Atlantic Region	A	4	0.7	9	5.1	-3	-2.1	4	3.2	- 6	-4.8	2	1.1
	B	-8	-1.5	5	3.0	-6	-4.7	1	0.4	- 8	-6.1	2	1.1
	C	5	0.9	7	4.3	-3	-2.1	5	3.7	- 5	-3.4	2	1.1
Quebec	A	13	0.8	1	0.2	5	1.2	6	1.5	1	0.5	-20	-3.1
	B	12	0.7	4	0.7	9	2.2	3	0.6	- 3	-1.0	-20	-3.1
	C	0	0.0	-5	-0.8	2	0.5	2	0.6	1	0.2	-20	-3.1
Ontario	A	17	0.7	13	1.5	-2	-0.3	8	1.5	- 3	-0.8	-14	-1.5
	B	11	0.5	17	2.1	-3	-0.7	1	0.3	- 5	-1.2	-14	-1.5
	C	8	0.4	12	1.4	-4	-0.7	3	0.6	- 4	-0.9	-14	-1.5
Prairie Region	A	8	0.7	-6	-1.4	5	2.3	6	2.5	2	1.1	-12	-2.3
	B	1	0.1	-3	-0.7	4	1.9	1	0.3	- 1	-0.4	-12	-2.3
	C	-0	-0.0	-5	-1.2	4	1.8	2	1.0	- 2	-0.8	-12	-2.3
British Columbia	A	9	1.2	-1	-0.3	2	1.5	7	4.2	1	0.5	-14	-3.8
	B	3	0.4	-2	-0.6	1	0.8	5	3.1	2	-1.7	-14	-3.8
	C	-1	-0.2	-4	-1.2	-2	-1.2	5	3.1	0	0.3	-14	-3.8

A = weight of family head
B = weight of female spouse
C = harmonic mean

[1] Excluding unattached individuals

TABLE 3
Absolute and Percent Differences Relative to Census, Economic Family Estimates, Canada and Regions, Labour Force Survey, May 1981 (Regression estimator)

Differences in thousands	Total Families[1]		Family Size 2		3		4		5+		Unattached Individuals	
	Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ	Δ	%Δ
Canada	20	0.3	23	1.0	3	0.2	10	0.6	-16	-1.4	-13	-0.5
Atlantic Region	1	0.1	2	1.4	-3	-2.7	4	3.3	3	-2.1	3	1.9
Quebec	2	0.1	2	-0.4	-2	0.4	-0	-0.1	-1	-0.4	1	0.2
Ontario	12	0.5	19	2.3	-0	0.0	-0	-0.0	-7	-1.8	-4	-0.4
Prairie Region	7	0.6	2	0.5	7	3.0	2	0.7	-4	-2.0	-11	-2.1
British Columbia	-2	-0.3	-3	-0.9	-2	-1.4	4	2.1	-1	-0.6	-2	-0.6

[1] Excluding unattached individuals

TABLE 4
Distribution of Percent Deviations Relative to the Subweight of the Final Weight, Regression and Post-Stratification Estimators, Canada, May 1981

Percent Deviation			Percentage of Total Sample	
			Regression	Post-Stratification
<		-30%	0.4	0.0
-30	to	-20%	1.2	0.4
-20	to	-10%	6.4	2.3
-10	to	0%	27.3	32.2
0	to	10%	35.6	44.0
10	to	20%	20.1	14.8
20	to	30%	5.9	4.0
>		30%	3.0	2.2

N = 159014