Survey Methodology, June 1994
Vol. 20, No. 1, pp. 23–31
Statistics Canada

Small Domain Estimation for Unequal Probability Survey Designs

D. HOLT and D.J. HOLMES[1]

ABSTRACT

The problem of estimating domain totals and means from sample survey data is common. When the domain is large, the observed sample is generally large enough that direct, design-based estimators are sufficiently accurate. But when the domain is small, the observed sample size is small and direct estimators are inadequate. Small area estimation is a particular case in point and alternative methods such as synthetic estimation or model-based estimators have been developed. The two usual facets of such methods are that information is 'borrowed' from other small domains (or areas) so as to obtain more precise estimators of certain parameters and these are then combined with auxiliary information, such as population means or totals, from each small area in turn to obtain a more precise estimate of the domain (or area) mean or total. This paper describes a case involving unequal probability sampling in which no auxiliary population means or totals are available and borrowing strength from other domains is not allowed and yet simple model-based estimators are developed which appear to offer substantial efficiency gains. The approach is motivated by an application to market research but the methods are more widely applicable.

KEY WORDS: Synthetic estimation; Design-based estimation; Small area estimation; Model-based estimation; Market shares.

1. INTRODUCTION

This paper is concerned with the common problem of estimating domain totals and means from a disproportionately allocated sample survey. Some domains may be large, in which case the achieved sample size may be large too and design-based (or direct) estimators will be satisfactory. Some domains may be small, in which case the achieved sample size may be small too and design-based (or direct) estimators will be too imprecise for practical use. The methods proposed will be motivated through the example of estimating sales, market shares and market penetrations for products in a market research survey. The domains are particular auto manufacturers or models. However, the general approach is applicable to other disproportionately allocated surveys of businesses or institutions.

The problem is analogous to that of using synthetic estimation for small area estimation (Gonzales 1973; Gonzales and Hoza 1978; Platek *et al.* 1987). Synthetic estimation usually depends on two factors: (i) the use of auxiliary variables in conjunction with population means or totals for each small area (or domain) to improve estimates through poststratification or regression estimation, and (ii) the improvement of estimates by pooling data across the small areas (or domains). In our situation no auxiliary population means or totals are available and, since the essential objective is to compare domains (*i.e.*, manufacturers and particular products), the idea of borrowing strength between these is inadmissible. A class of synthetic estimators is proposed which uses neither of these two approaches and yet is preferred to the direct survey estimators. The proposed estimators have a simple structure, an interesting interpretation and can be justified under a set of model assumptions which are testable under the general assumption of non-informative survey design.

2. THE MARKET RESEARCH EXAMPLE

Market researchers often estimate the total volume of sales and market shares for each manufacturer of a particular product. We consider the case of autos purchased for company fleet use in a single year. Estimates of totals and market shares are required for each auto manufacturer and for specific models which are widely purchased for fleet use.

The terms 'fleet' and 'company' are each interpreted widely. A fleet car is taken to mean any auto purchased on a commercial as opposed to a private basis, and used in conjunction with a business in the broadest sense. This includes autos purchased for sales representatives which may be purchased in large numbers. It also includes single purchases of luxury cars for company directors and other senior staff of large companies, as well as purchases by small 'companies' such as groups of doctors, or self-employed people such as shop owners. Thus the population of purchasing companies – termed consumers – includes a large number of small companies that purchase only one or two autos every few years.

[1] D. Holt and D.J. Holmes, Department of Social Statistics, University of Southampton, Highfield, Southampton, UK, SO95NH.

In the reference period of one year we define Y_{ki} to be the number of autos of product type k purchased by consumer i. The product type k (the domain) may refer to a specific model of a particular manufacturer, or to all models produced by a manufacturer. Thus, $Y_k = \sum_i Y_{ki}$ is the total number of autos of type k purchased by all consumers. Let Z_i be the total number of autos of any kind purchased by consumer i, and $Z = \sum_i Z_i$ be the total number of auto sales. The market share for product type k is defined as $R_k = Y_k/Z$.

We further define

$$Y'_{ki} = 1 \quad \text{if} \quad Y_{ki} > 0$$
$$= 0 \quad \text{if} \quad Y_{ki} = 0$$

and

$$Z'_i = 1 \quad \text{if} \quad Z_i > 0$$
$$= 0 \quad \text{if} \quad Z_i = 0.$$

Thus, Y'_{ki} and Z'_i are indicator variables for consumers who purchase product type k and at least one auto of any kind, respectively, in the reference period. The number of consumers that purchase product k is thus given by $Y'_k = \sum_i Y'_{ki}$ and the total number of consumers purchasing at least one auto of any kind is given by $Z' = \sum_i Z'_i$. The market penetration for product k, in terms of the proportion of consumers buying a car of any type in the reference period who buy type k, is given by $R'_k = Y'_k/Z'$.

The four parameters Y_k, R_k, Y'_k and R'_k are all legitimate targets of inference in market research and are defined as finite population parameters; namely, domain totals or ratios of domain totals.

3. THE SURVEY DESIGN AND DIRECT ESTIMATORS

The survey design was based upon two mutually exclusive frames and may be regarded as a simple stratified design with ten strata. The first frame was a register (Dun and Bradstreet) of 35,000 companies, stratified into eight strata on the basis of the number of employees and whether the company was classified as 'manufacturing' or 'distributing'. The second frame was a large register of 1.4 million British Telecom business subscribers, stratified into 'private' and 'commercial' numbers. Note that both private and commercial numbers were business subscribers but commercial numbers were allocated if separate commercial premises were occupied.

Using previous survey data the sample was optimally allocated using Neyman allocation to minimize the variance of the estimator of the total number of autos purchased (Z). Data on auto purchases were collected immediately after the end of the reference year. The strata sizes $\{N_h\}$ and sample allocations $\{n_h\}$ for strata $h = 1, \ldots, 10$ are given in Table 1.

Table 1

Sampling Frame: Sample Size and Weight by Stratum

Stratum (h)	Stratum Size N_h	Sample Size n_h	Weight $\pi_h^{-1} = N_h/n_h$
British Telecom:			
Private	389,445	1,150	338.65
Commercial	1,007,399	7,406	136.02
Dun and Bradstreet:			
Manufacturing			
50-99 employees	6,646	235	28.28
100-499	6,826	1,113	6.13
500-999	992	520	1.91
1,000+	1,110	849	1.31
Distributing			
50-99 employees	8,703	472	18.44
100-499	7,625	1,437	5.31
500-999	1,133	484	2.34
1,000+	1,523	1,117	1.36
Overall	1,431,402	14,783	96.83

The sample is a simple, disproportionately allocated stratified design and the direct estimators and their variances are well known. The stratification results in large differences in sampling weights (1.31 to 338.65) and is useful but far from ideal. Many consumers do not purchase any autos at all in the reference year so that each stratum contains a mixture of zero and non-zero responses. For any particular product k the proportion of zero responses in each stratum is obviously larger.

Table 2 contains the direct survey estimates, estimated standard errors (see Holt and Holmes (1993) for derivation), and coefficients of variation for a selection of products from different auto manufacturers. Products A and B represent all models for two major auto manufacturers. Product C is a single model with a substantial share of the fleet market from manufacturer A. The remaining products have small market shares. Products F and G cater for the executive part of the fleet market. The list is incomplete so that the market shares do not sum to one. Also note that the product categories are not mutually exclusive. In general the survey was judged to perform satisfactorily but it was observed over a period of years that estimates for manufacturers or models with small market shares were unstable. This is best seen in terms of the coefficient of variation which is greater than 0.1 for products with small market shares and can be greater than 0.15 or 0.2 in some cases. This instability also affects the estimates of variance as well as the estimates of total sales or market shares of the products.

Table 2
Direct Survey Estimates, Standard Errors and Coefficients of Variation for Selected Products

Product (k)	Estimating Consumers		Estimating Autos	
	Total $\hat{Y}'_k$	Penetration $\hat{R}'_k$	Total $\hat{Y}_k$	Share $\hat{R}_k$
A	59,890	.3843	270,051	.3781
	(2,651)	(.0144)	(35,704)	(.0315)
	(.044)	(.037)	(.132)	(.083)
B	34,282	.2200	153,518	.2149
	(1,960)	(.0117)	(8,653)	(.0131)
	(.057)	(.053)	(.056)	(.061)
C	23,363	.1499	81,381	.1139
	(1,602)	(.0098)	(17,559)	(.0194)
	(.069)	(.065)	(.216)	(.170)
D	13,857	.0889	25,312	.0354
	(1,311)	(.0081)	(2,906)	(.0039)
	(.095)	(.091)	(.115)	(.110)
E	9,025	.0579	24,370	.0341
	(1,146)	(.0072)	(7,336)	(.0101)
	(.127)	(.124)	(.301)	(.296)
F	5,125	.0329	13,724	.0192
	(676)	(.0043)	(2,369)	(.0030)
	(.132)	(.131)	(.173)	(.156)
G	7,518	.0482	11,031	.0154
	(1,015)	(.0064)	(1,456)	(.0022)
	(.135)	(.133)	(.132)	(.143)

Row 1: estimate Row 2: s.e. Row 3: c.v.

4. A MODEL-BASED APPROACH

Given the sample design there is no prospect of improving the efficiency of the direct survey estimators within the conventional sample survey framework. The usual approaches are through the use of auxiliary information for poststratification, ratio or regression estimation but all of these require knowledge of population means or totals. No such information is available. We turn instead to a model-based approach to provide alternative estimators for the whole range of products.

4.1 Estimating Y'_k: the Number of Consumers Purchasing Product Type k

We consider, initially, the number of consumers who buy product type k. We extend the notation from Y'_{ki} to Y'_{khi} in the obvious way to define the indicator random variable of purchase for product k for consumer i in stratum h. We treat each consumer's decision as the outcome of a Bernoulli trial. Let $P_{k|h}$ be the probability that a consumer in stratum h buys an auto of type k [$P_{k|h}$ = Prob($Y'_{khi} = 1$)]. We define the model-based equivalent of Y'_k, the total number of consumers of product k, as

$$\Theta'_k = \sum_h N_h P_{k|h}. \quad (1)$$

Assuming that each consumer's decision is independent the likelihood may be written as the usual product of binomial terms. The maximum likelihood estimators are given by $\hat{P}_{k|h} = n_{kh}/n_h$, and the maximum likelihood estimator of Θ'_k is the familiar stratified sampling estimator

$$\hat{\Theta}'_k(1) = \sum_h \frac{N_h}{n_h} n_{kh} = \sum_h N_h \bar{y}'_{kh}, \quad (2)$$

where n_{kh} is the sample count of consumers in stratum h that buy product k, n_h is the stratum sample size and $\bar{y}'_{kh} = n_{kh}/n_h$ is the sample mean for consumers in stratum h (*i.e.*, the sample proportion of consumers in stratum h who buy product k). This estimator is generally unsatisfactory when the sample size for product k is too small.

Suppose we introduce an additional conditioning factor such that every consumer may be categorized into one of its categories f, $f = 1, \ldots, F$, and further extend the definition of the indicator random variable to Y'_{khfi}. These categories f will cut across the strata h and the idea is to define f so that, within any particular category, whether a consumer buys product type k or not is independent of the stratum membership h. In the case of fleet purchases we define a categorization based on the total number of autos owned and operated by each consumer (*i.e.*, the fleet size). A more detailed discussion of the choice of f is given in Section 5.

If N_{hf}, the population counts of consumers in stratum h and fleet size category f, are known then (1) may be extended in the obvious way and the target parameter can now be expressed as

$$\Theta'_k = \sum_h \sum_f N_{hf} P_{k|hf}. \quad (3)$$

Equation (3) is the case of poststratification if $\{N_{hf}\}$ are known, and in this case the additional information will lead to a gain in efficiency (Holt and Smith 1979). When $\{N_{hf}\}$ are unknown we may rewrite the model in terms of two sets of probabilities:

$Q_{f|h}$ = Prob {consumer has fleet size f | stratum h},

$P_{k|hf}$ = Prob {consumer buys product type k | stratum h and fleet size f}.

The target parameter may now be expressed as

$$\Theta'_k = \sum_h \sum_f N_h Q_{f|h} P_{k|hf}. \quad (4)$$

To obtain an alternative model-based estimator we make further assumptions about the model parameters. Suppose now that

$$P_{k|hf} = P_{k|f} \quad \text{for all} \quad h. \tag{5}$$

This implies that conditional on the categorization f (the size of the fleet operated by a consumer), the probability of buying product type k is *independent* of the original stratum membership h. Algebraically, the assumption is analogous to that used in synthetic estimation for small area estimation but in that case information is pooled across areas. That form of the assumption is inadmissible in our case. We choose instead pooling across strata within the domain of study. The idea is to choose a conditioning variable which accounts for the marginal association between choice of product and stratum membership.

Using assumption (5) and with the obvious extension of the notation ($n_{kf} = \sum_h n_{khf}$, *etc.*) it may be shown that

$$\hat{Q}_{f|h} = \frac{n_{hf}}{n_h}, \qquad \hat{P}_{k|f} = \frac{n_{kf}}{n_f}$$

and the maximum likelihood estimator of Θ'_k becomes

$$\hat{\Theta}'_k(2) = \sum_h \sum_f N_h \frac{n_{hf}}{n_h} \frac{n_{kf}}{n_f} = \sum_f \hat{N}_f \frac{n_{kf}}{n_f}$$

$$= \sum_h \hat{N}_f \bar{y}'_{kf}, \tag{6}$$

where $\hat{N}_f = \sum_h N_h \, n_{hf}/n_h$, and $\bar{y}'_{kf} = n_{kf}/n_f$ is the unweighted sample mean for consumers in category f (*i.e.* the sample proportion of consumers in category f who buy product k).

Thus (6) has the form of a stratified estimator based on the categorization f but with the population sizes in each stratum $\{N_f\}$ unknown. Note that an estimator of this form, but with known $\{N_f\}$, would arise naturally if a stratified sample based on f had been selected. In fact this is **not** so: the sample members of category f are **not** selected with equal probability. However, the parameter assumptions lead to treating the sample in each category f as if it was an equal probability sample since under assumption (5) the sample weights are uninformative and simply lead to efficiency loss when estimating $P_{k|f}$. Hence, although the sampling fractions n_h/N_h are used to estimate $\{N_f\}$ they are not used explicitly in $\hat{P}_{k|f} = n_{kf}/n_f = \bar{y}'_{kf}$. Note that the estimator pools information across strata h, within domain k but **not** between domains (*i.e.* products).

Note that if n_h/N_h is constant, equation (6) reduces to the usual expansion estimator given by (2), and assumption (5) has not yielded a new estimator. If the sample is disproportionately allocated the assumption leads to the use of the sampling weights for $\hat{N}_f$ (where they are needed) but not for estimating $P_{k|f}$ (where they are uninformative given f and assumption (5)).

Equation (5) is a strong set of assumptions, requiring $P_{k|hf}$ to be exactly equal to a common value $P_{k|f}$ for all h. In practice, random assumptions such as $P_{k|hf} = P_{k|f} + \epsilon_{k|hf}$ may be introduced, where $E[\epsilon_{k|hf}] = 0$ and $V[\epsilon_{k|hf}] = \sigma^2_\epsilon$. These assumptions will lead to hierarchical Bayes or empirical Bayes analysis as described in Ghosh and Rao (1994) or Fay and Herriot (1979). These methods are not developed here since the simple form of the model-based estimator would be lost, together with the insight that this provides. In a similar vein the approach of Särndal and Hidiriglou (1989) or Drew, Singh and Choudhry (1982) may be applied to yield sample size dependent estimators without violating the requirement that no information is pooled across domains (products).

We can compare the estimators in (2) and (6) when assumption (5) holds since it may be shown that

$$V_\xi(\hat{\Theta}'_k(1)) = \sum_h \frac{N_h^2}{n_h} P_{k|h}(1 - P_{k|h})$$

$$= \sum_h \sum_f \frac{N_h^2}{n_h} Q_{f|h} P_{k|f}$$

$$- \sum_h \sum_f \sum_{f'} \frac{N_h^2}{n_h} Q_{f|h} Q_{f'|h} P_{k|f} P_{k|f'}, \tag{7}$$

where the notation $V_\xi(\cdot)$ is used to emphasize that the variance is evaluated with respect to the model-based distribution.

It may also be shown that under assumption (5)

$$V_\xi(\hat{\Theta}'_k(2)) = \sum_h \sum_f \frac{N_h^2}{n_h} P^2_{k|f} \, Q_{f|h} \, (1 - Q_{f|h})$$

$$- \sum_h \sum_f \sum_{\substack{f' \\ f \neq f'}} \frac{N_h^2}{n_h} P_{k|f} \, P_{k|f'} \, Q_{f|h} \, Q_{f'|h}$$

$$+ \sum_h \sum_f \frac{N_h^2}{n_h} \frac{P_{k|f}(1 - P_{k|f}) Q_{f|h}}{\sum_h n_h Q_{f|h}}$$

$$\left\{ (1 - Q_{f|h}) + n_h Q_{f|h} \right.$$

$$\left. + \frac{[1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q^2_{f|h}]}{\sum_h n_h Q_{f|h}} \right\} \tag{8}$$

and that $V_\xi(\hat{\Theta}'_k(1)) - V_\xi(\hat{\Theta}'_k(2)) \geq 0$.

Thus under the additional model assumptions $\hat{\Theta}'_k(2)$ has smaller variance as would be expected. These expressions are model-based variances and no finite population corrections arise. A predictive approach to the unobserved elements in each poststratum would give rise to finite population correction factors.

The maximum likelihood estimator of the market penetration for product type k, R'_k, under assumption (5) is simply given by

$$\hat{\Omega}'_k(2) = \frac{\sum_f \hat{N}_f \frac{n_{kf}}{n_f}}{\sum_f \hat{N}_f \frac{n_{af}}{n_f}} = \frac{\sum_f \hat{N}_f \bar{y}'_{kf}}{\sum_f \hat{N}_f \bar{z}'_f}, \qquad (9)$$

where n_{af} is the sample count of consumers in fleet category f that buy an auto of any kind, and $\bar{z}'_f = n_{af}/n_f$ is the sample proportion of consumers in category f who buy an auto of any kind.

4.2 Efficiency of the Model-Based Estimator of Y'_k

To investigate the gain in efficiency of $\hat{\Theta}'_k(2)$ over $\hat{\Theta}'_k(1)$ we consider the efficiency of the model-based estimator, defined by

$$e[\hat{\Theta}'_k(2)] = \frac{V_\xi(\hat{\Theta}'_k(1)) - V_\xi(\hat{\Theta}'_k(2))}{V_\xi(\hat{\Theta}'_k(1))}, \qquad (10)$$

for various population structures in which assumption (5) holds.

We consider a population with strata $\{h\}$, stratum sizes $\{N_h\}$ and sample allocations $\{n_h\}$ as given in Table 1, and a conditioning factor with ten categories f ($f = 1, \ldots, 10$) of increasing fleet size. We compute the efficiency factor $e[\hat{\Theta}'_k(2)]$ for various combinations of parameter values of $\{Q_{f|h}\}$ and $\{P_{k|f}\}$.

We consider five different structures for $\{Q_{f|h}\}$:

(a) $Q_{f|h} = \begin{cases} 1 & f = h \\ 0 & f \neq h \end{cases}$ for $h = 1, \ldots, 10$.

(b) $Q_{f|h} = \begin{cases} 0.95 & f = h \quad \text{for } h = 1, \ldots, 10 \\ 0.025 & f = h-1 \quad \text{for } h = 2, \ldots, 10 \\ 0.025 & f = h+1 \quad \text{for } h = 1, \ldots, 9 \\ 0.05 & h = 1, f = 2 \text{ and } h = 10, f = 9 \\ 0 & \text{otherwise} \end{cases}$

= Band Matrix (0.025, 0.95, 0.025).

(c) $Q_{f|h}$ = Band Matrix (0.05, 0.90, 0.05).

(d) $Q_{f|h}$ = Band Matrix (0.05, 0.10, 0.70, 0.10, 0.05).

(e) $Q_{f|h} = 0.1$ for $h = 1, \ldots, 10$ and $f = 1, \ldots, 10$.

We consider four different structures for $\{P_{k|f}\}$:

(i) $P_{k|f} = \begin{cases} 0.1 & f = 1, 2 \\ 0 & \text{otherwise.} \end{cases}$

(ii) $P_{k|f} = 0.1 - 0.01\,(f-1)$ for $f = 1, \ldots, 10$.

(iii) $P_{k|f} = 0.1f$ for $f = 1, \ldots, 10$.

(iv) $P_{k|f} = 0.5$ for $f = 1, \ldots, 10$.

Structure (a) is one where the categorization f coincides with the stratification. In structures (b), (c) and (d), in any particular stratum h the majority of consumers fall into one fleet category ($f = h$) with a few consumers in neighbouring categories (*e.g.*, for (b) and (c) $f = h-1$, $h+1$). Finally, structure (e) implies that, in any stratum h, consumers will be equally likely to fall into any one of the fleet categories $f = 1, \ldots, 10$.

Structure (i) for $P_{k|f}$ implies a type of auto that is purchased with a small probability by consumers with small fleet sizes (*i.e.* that fall in categories $f = 1$ or 2), but not purchased by consumers with large (r) fleet sizes. Structure (ii) suggests a type of auto purchased with small probability which decreases as fleet size increases, whilst structure (iii) implies the reverse. In structure (iv) a popular model is bought with probability 0.5 *regardless* of the consumer's fleet size.

Table 3 gives the efficiency factor defined in (10) for each combination of structures for $Q_{f|h}$ and $P_{k|f}$ under the disproportionate allocation given in Table 1. Column (a) of the table is the special case where the stratification and the categorization f coincide, and the two estimators $\hat{\Theta}'_k(1)$ and $\hat{\Theta}'_k(2)$ are the same. The table shows that large gains in efficiency (*e.g.*, 70%) can be attained for certain parameter combinations: the weaker the association

Table 3
Efficiency Factors, $e[\hat{\Theta}'_k(2)]$, for Various Combinations of $Q_{f|h}$ and $P_{k|f}$

		Structure for $Q_{f\|h}$				
		(a)	(b)	(c)	(d)	(e)
Structure for $P_{k\|f}$	(i)	0	0.108	0.196	0.355	0.648
	(ii)	0	0.116	0.206	0.391	0.695
	(iii)	0	0.103	0.181	0.387	0.695
	(iv)	0	0.115	0.203	0.391	0.706

between f and h the greater the efficiency gain. Even for structures (c) and (d) where the association between f and h is strong, substantial efficiency gains can be achieved. The structure $Q_{f|h}$ is much more important than $P_{k|f}$ in determining efficiency gain.

In the special case (e) where $Q_{f|h}$ is a constant for all f and h it can be shown that the efficiency factor can be expressed as

$$e[\hat{\Theta}_k'(2)] = \left(1 - \frac{\delta^2}{\bar{P}_{k|f}\,(1 - \bar{P}_{k|f})}\right) \frac{\sum_h \tau_h N_h^2/n_h}{\sum_h N_h^2/n_h}, \quad (11)$$

where

$$\bar{P}_{k|f} = \frac{1}{F}\sum_{f=1}^{F} P_{k|f} \quad \text{and} \quad \delta^2 = \frac{1}{F}\sum_{f=1}^{F} (P_{k|f} - \bar{P}_{k|f})^2$$

are the mean and variance of $\{P_{k|f}\}$ over the categories f, and $\tau_h = 1 - n_h/n + O(n^{-1})$. The term in parentheses in (11) lies between 0 and 1 and it's value depends on how the $\{P_{k|f}\}$ vary over the categories f. In case (iv) $P_{k|f}$ is constant and so this term is unity. The second term of (11) depends solely on the design, and its value for the sample allocation specified in Table 1 is 0.706.

4.3 Estimating Y_k: the Number of Autos Purchased of Product Type k

The previous approach in Section 4.1 may be extended to the number of purchases. We introduce a further conditioning factor which represents the total number of autos purchased, m, regardless of product type, and we extend the notation in the obvious manner to Y_{khfmi}, the random variable representing the number of autos of product type k purchased by consumer i in stratum h, fleet size f, and buying m autos of any kind. The idea is that the number of purchases of product k is likely to vary depending on the total number of autos purchased. Let

$$S_{m|hf} = \text{Prob}\{\text{consumer buys } m \text{ autos of any kind} \mid h,f\}, \quad m = 0, 1, 2, \ldots,$$

$$T_{\ell|hfm} = \text{Prob}\{\text{consumer buys } \ell \text{ autos of type } k \mid h,f,m\}, \quad \ell = 0, 1, \ldots, m.$$

The model-based target parameter, equivalent to the total purchases of product k, Y_k, is extended from (4) and may now be expressed as

$$\Theta_k = \sum_h \sum_f \sum_m \sum_\ell N_h Q_{f|h} S_{m|hf} T_{\ell|hfm}\, \ell. \quad (12)$$

We consider two sets of additional assumptions, the first of which is

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{for all} \quad h. \quad (13)$$

These assumptions imply that conditional on fleet size category, f, and the total number of new autos purchased, m, the distribution of the number of autos purchased of product type k is independent of stratum h.

The maximum likelihood estimator of Θ_k under assumptions (13) is

$$\hat{\Theta}_k(2) = \sum_f \sum_m \hat{N}_{fm}\, \bar{y}_{kfm}, \quad (14)$$

where $\hat{N}_{fm} = \sum_h N_h n_{hfm}/n_h$, and $\bar{y}_{kfm} = \sum_\ell \ell\, n_{fm\ell}/n_{fm}$ is the unweighted sample mean of the number of autos of product type k purchased by consumers of fleet size f that purchased a total of m autos of any kind.

The selection probabilities are used here to provide a weighted estimator of N_{fm}, the total number of consumers of fleet size f that buy m cars of any kind. The form of the estimator is analogous to that in equation (6). Under the model assumption (13) it may be shown that

$$\begin{aligned} V_\xi(\hat{\Theta}_k(2)) = {} & \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \mu_{fm}^2 Q_{fm|h}\,(1 - Q_{fm|h}) \\ & - \sum_h \sum_f \sum_m \sum_{f'} \sum_{m'} \frac{N_h^2}{n_h} \mu_{fm}\mu_{f'm'} Q_{fm|h} Q_{f'm'|h} \\ & \qquad (f,m) \neq (f',m') \\ & + \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \frac{\sigma_{fm}^2 Q_{fm|h}}{\sum_h n_h Q_{fm|h}} \\ & \left\{ (1 - Q_{fm|h}) + n_h Q_{fm|h} \right. \\ & \left. + \frac{[1 + (2n_h - 3)Q_{fm|h} - 2(n_h - 1)Q_{fm|h}^2]}{\sum_h n_h Q_{fm|h}} \right\}, \end{aligned} \quad (15)$$

where $Q_{fm|h} = Q_{f|h}\, S_{m|hf}$, $\mu_{fm} = E_\xi\{Y_{khfmi}\}$, and $\sigma_{fm}^2 = V_\xi\{Y_{khfmi}\}$.

In practice, $\bar{y}_{kfm}$ will be based on very few observations if few customers in fleet size category f purchase exactly m cars. For more stability m may be defined as an ordinal variable by grouping the total number of autos purchased into a small number of categories. In this case assumption (13) implies that the distribution of purchases for product type k is the same within fleet size category f and total

purchase category m. Also, ℓ may be treated as a continuous random variable and distributional assumptions made about ℓ leading to ratio or regression estimators.

A second and even stronger set of parameter assumptions is

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{for all} \quad h,$$

$$S_{m|hf} = S_{m|f} \quad \text{for all} \quad h. \tag{16}$$

These assumptions imply that conditional on fleet size category, f, the joint distribution of the number of autos purchased of type k and the total number of autos purchased of any kind, m, is independent of the stratum h. In this case the maximum likelihood estimator of Θ_k is given by

$$\hat{\Theta}_k(3) = \sum_f \hat{N}_f \bar{y}_{kf}, \tag{17}$$

where $\bar{y}_{kf} = \sum_\ell \ell\, n_{f\ell}/n_f$ is the unweighted sample mean of the number of autos of product type k purchased by consumers in fleet size f regardless of how many autos the consumer bought in total, and $\hat{N}_f = \sum_h N_h\, n_{hf}/n_h$ is a weighted estimator of the number of consumers of fleet size f overall. It may be shown that under assumptions (16)

$$V_\xi(\hat{\Theta}_k(3)) = \sum_h \sum_f \frac{N_h^2}{n_h} \mu_f^2 Q_{f|h} (1 - Q_{f|h})$$

$$- \sum_h \sum_{\substack{f \\ f \neq f'}} \sum_{f'} \frac{N_h^2}{n_h} \mu_f \mu_{f'} Q_{f|h} Q_{f'|h}$$

$$+ \sum_h \sum_f \frac{N_h^2}{n_h} \frac{\sigma_f^2 Q_{f|h}}{\sum_h n_h Q_{f|h}}$$

$$\left\{ (1 - Q_{f|h}) + n_h Q_{f|h} \right.$$

$$\left. + \left[\frac{1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q_{f|h}^2}{\sum_h n_h Q_{f|h}} \right] \right\}. \tag{18}$$

If assumptions (16) were plausible then $\bar{y}_{kf}$ would be based on larger sample sizes than $\bar{y}_{kfm}$ in (14) and hence $\hat{\Theta}_k(3)$ would be more stable.

The maximum likelihood estimator of the market share for product type k, R_k, under assumption (16), is given by

$$\hat{\Omega}_k(3) = \frac{\sum_f \hat{N}_f \bar{y}_{kf}}{\sum_f \hat{N}_f \bar{z}_f}, \tag{19}$$

where $\bar{z}_f$, defined analogously to $\bar{y}_{kf}$, is the unweighted sample mean number of autos of any kind purchased by consumers in fleet category f.

5. EMPIRICAL RESULTS

5.1 Estimating Consumers

In Section 4.2 the efficiency of $\hat{\Theta}_k'(2)$ was investigated for various population structures when assumption (5) held. Readers may find this measure unconvincing since (5) will not hold in practice. We now use the actual survey data to compute $\hat{\Theta}_k'(2)$ for a particular categorization of the conditioning factor that is defined by a combination of the fleet size *and* whether or not the consumer purchased any autos of any kind for fleet use (see Table 4). Empirical evaluations of synthetic estimators have been carried out by Schaible, Brock and Schnack (1977) and Drew, Singh and Choudhry (1982) in different contexts.

For each of the products A-G listed in Table 2 a χ^2 test was used to test the hypothesis that, conditional on the category of the conditioning factor (f), whether or not a consumer purchases that product is independent of stratum (h). Note that for our example the design is stratified random sampling and standard multinomial assumptions apply. For multistage designs, the standard χ^2 analysis would have to be adjusted by using Rao-Scott adjustments for example. In practice it is difficult to find a categorization f such that conditional independence assumptions (5) hold for every product type. However, for the categorization defined in Table 4 it was found that

Table 4

Definition of the Categories, f, of the Conditioning Factor

Categories f	Definition of f	
	Fleet Size	Fleet Purchases
1	Any	0
2	1-4	> 0
3	5-8	> 0
4	9-15	> 0
5	16-25	> 0
6	26-50	> 0
7	51-100	> 0
8	101-200	> 0
9	201-550	> 0
10	> 550	> 0

most of the variability in the probability of purchasing a particular product type was explained by the category f of the conditioning factor and very little of the residual variation was due to differences in strata.

The model-based estimates for consumers, $\hat{\Theta}_k'(2)$ and $\hat{\Omega}_k'(2)$, obtained from (6) and (9) respectively, are given in Table 5. The model-based variances may give an optimistic view of the precision of the estimators since they depend on the conditional independence assumptions in the model which may be untrue in practice. Alternatively the usual survey estimate of the p-based variance of the model-based estimator may be derived (see Holt and Holmes 1993). This requires no distributional or conditional independence assumptions of any kind and might be considered a more objective measure. These estimates of standard errors are given in Table 5. Since the estimated standard errors are design-based, they include finite population corrections. [We note here that the model-based standard errors for $\hat{\Theta}_k'(2)$ (not shown in Table 5) were consistently around 10% smaller than the p-based standard errors].

Table 5

Model-Based Estimates with p-Based Standard Errors for Selected Products

Product (k)	Estimating Consumers		Estimating Autos	
	Total $\hat{\Theta}_k'(2)$	Penetration $\hat{\Omega}_k'(2)$	Total $\hat{\Theta}_k(3)$	Share $\hat{\Omega}_k(3)$
A	63,433	.4070	263,511	.3722
	(2,230)	(.0105)	(13,007)	(.0048)
B	39,673	.2546	177,067	.2501
	(1,587)	(.0086)	(9,530)	(.0046)
C	21,930	.1407	65,357	.0923
	(1,142)	(.0066)	(3,836)	(.0027)
D	13,422	.0861	22,146	.0313
	(868)	(.0052)	(1,351)	(.0016)
E	7,366	.0473	15,798	.0223
	(675)	(.0041)	(1,223)	(.0014)
F	5,826	.0374	14,398	.0203
	(492)	(.0031)	(1,113)	(.0012)
G	7,686	.0493	11,207	.0158
	(633)	(.0039)	(813)	(.0011)

Row 1: estimate Row 2: p-based s.e.

Comparing these results with the usual survey results given in Table 2 we find that the standard errors for estimating totals are considerably smaller – around 30-40% smaller for all products except A and B (the major manufacturers) where the reduction is about 15-20%. This pattern is expected since the original survey design was optimal for the total sales of autos and therefore relatively efficient for products with a large market share. We expect the products with smaller market shares to benefit most from the model-based approach.

For estimating market penetration the reduction in standard error is again about 30-40% with slightly smaller reductions for products A and B.

5.2 Estimating Autos

Table 5 also contains model-based estimates for the total number of autos purchased of type k and the corresponding market share, $\hat{\Theta}_k(3)$ and $\hat{\Omega}_k(3)$ as defined by (17) and (19) respectively, for the *same* categorization f of the conditioning factor as given in Table 4. P-based standard errors for these estimates are also presented in Table 5.

Comparing with the standard survey estimates given in Table 2 large reductions in standard errors for estimating totals are obtained (40-80%) apart from product type B. Similarly, for estimating the market shares the reduction in standard error is again substantial.

6. DISCUSSION

The model-based estimators are derived using conditional independence assumptions to partition the estimation problem into two components. The first, an estimate of N_f (the number of consumers of fleet size f), makes use of the unequal selection probabilities, whereas the second, an estimate of the proportion of consumers of fleet size f buying product type k (or the average number of autos of product type k purchased by consumers of fleet size f) does not. This can result in a substantial efficiency gain.

If the conditional independence assumptions are invalid then in ordinary design-based terms the estimators will have a residual bias but this may be an acceptable risk to achieve stability of the estimators over the whole product range. For the numerical results in previous sections, only the model-based estimates for product B are outside of the 95% confidence interval based on the direct survey estimator. The conditional independence assumptions will depend on the choice of the categories f, and can be tested using chi-square tests for contingency tables.

Whilst the results in Table 5 show that the design-based standard errors for the model-based estimates are generally smaller than for the direct estimates shown in Table 2, it may be argued that the model-based estimators may be biased and hence provide no gain in terms of mean-squared error (MSE). The bias will arise from the inappropriateness of the conditional independence assumptions (*e.g.*, equation (5)). This is not testable, but a comparison of Tables 2 and 5 can give some insight into the size of bias that would be required to cause the MSE to be the same

for both the direct and the model-based estimators. Consider the estimate of total consumers for product E which is strongly affected by the procedure and hence perhaps most susceptible to bias. The variance (and hence MSE) of the direct estimator is $1{,}146^2 = 1{,}313{,}316$ whereas for the model-based estimator the variance is $675^2 = 455{,}625$. Hence, the model-based estimate of 7,366 would need a bias of 926 in order for the MSEs to be the same.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments.

REFERENCES

DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 19-47.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. To appear in *Statistical Science*.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GONZALEZ, M.E., and HOZA, C. (1978). Small area estimation with application to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, 73, 7-15.

HOLT, D., and HOLMES, D.J. (1993). Small domain estimation for unequal probability survey designs. Working Paper Series, No. 2, Department of Social Statistics, University of Southampton, UK.

HOLT, D., and SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, Ser. A, 142, 33-46.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (1987). *Small Area Statistics*. New York: John Wiley and Sons.

SÄRNDAL, C.-E., and HIDIRIGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L., BROCK, D.B., and SCHNACK, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. *Proceedings of the Social Statistics Section, American Statistical Association*, 1017-1021.