Journal of Official Statistics
Vol. 7. No. 3, 1991, pp. 325–337
© Statistics Sweden

Weighting via Response Modeling in the Finnish Household Budget Survey

Anders Ekholm[1] *and Seppo Laaksonen*[2]

Abstract: The purpose of the Finnish Household Budget Survey is to monitor the consumption of the total population and of segments of the population. The data are collected by interviews. The sampling frame is a person-based register, but the unit of observation is the household. Nonresponse has been increasing in recent rounds. For the estimation of totals and means from the 1985 data, we implemented a "model based Horvitz–Thompson estimator." We model the response propensity by logistic regression on explanatory factors, most important of them being the household structure. The estimator weights by the inverses of the estimated response probabilities from the fitted model. We state the assumption under which the estimator is unbiased and derive its appropriately conditioned variance. We compare the new estimator with the one previously used, theoretically and by selected empirical results. We discuss some alternatives to weighting by response propensity modeling.

Key words: adjustment cell; estimation of mean; estimation of total; model based Horvitz–Thompson estimator; logistic regression; response probability; selection probability; unit nonresponse.

1. Introduction

The Central Statistical Office (CSO) of Finland conducts a Household Budget Survey (HBS) every fifth year. An HBS is a major enterprise whose primary purpose is to produce estimates of annual household consumption in different segments of the population. The data are collected in three stages, consisting of an initial personal interview, followed by two weeks of bookkeeping, and a final interview, as illustrated in Figure 1. In the 1985 round the overall response rate dropped to 70%, although great effort had been made to turn the trend of declining response rates. This called for an estimator less liable to suffer from nonresponse bias, than the estimator used in previous rounds. In this paper we report on the computational and theoretical properties of a "model based Horvitz–Thompson" estimator, which we devised for the 1985 round.

In Section 2 we give the relevant information on the sampling procedure, and in Section 3 we define the old and the new estimators. The new method is built on a response propensity model. In Section 4 we report the model for the log odds ratio of responding that we fitted to the 1985 data. In Section 5 we state the assumptions under which the old and the new estimators are unbiased and report their conditional variances given the appropriate statistics. Section 6 presents empirical comparisons in

[1] Department of Statistics, University of Helsinki, Aleksi 7, SF-00100 Helsinki, Finland.
[2] Central Statistical Office, P.O. Box 770, SF-00101 Helsinki, Finland.

Acknowledgement: Financial support from the Academy of Finland and Suomen Kulttuurirahasto is gratefully acknowledged.

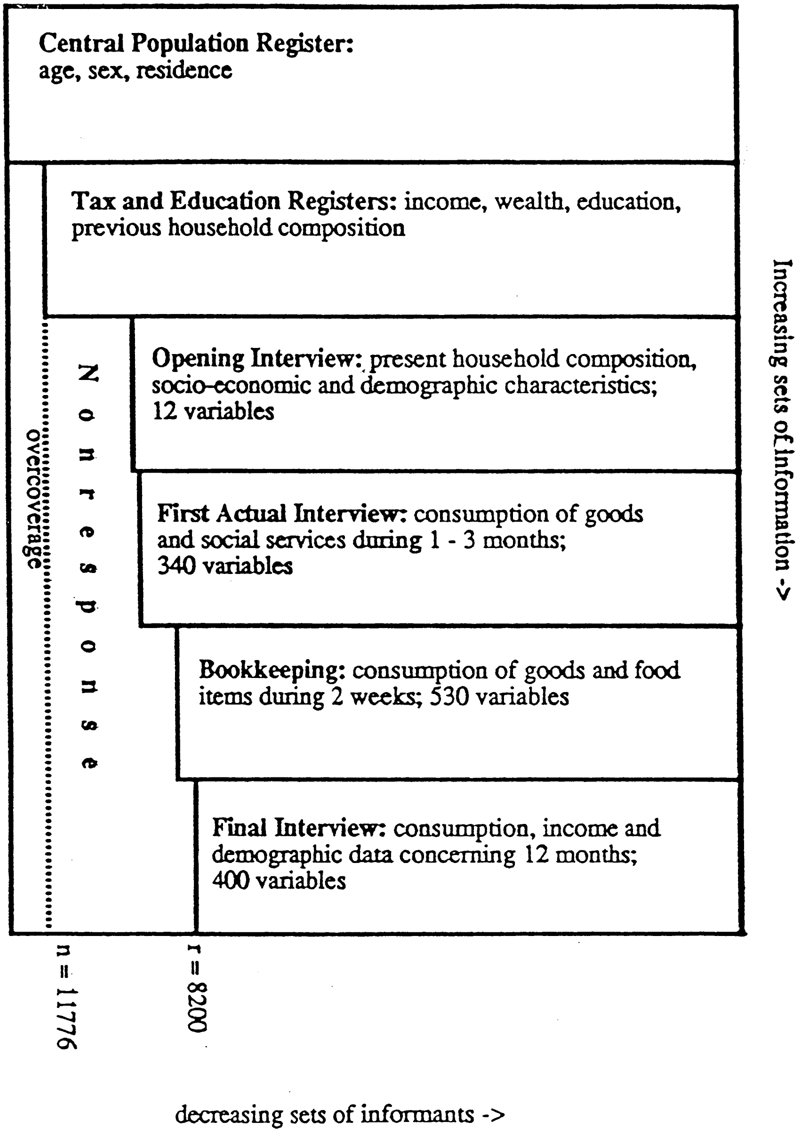
Central Population Register:
age, sex, residence
Tax and Education Registers: income, wealth, education, previous household composition
Opening Interview: present household composition, socio-economic and demographic characteristics; 12 variables
First Actual Interview: consumption of goods and social services during 1 - 3 months; 340 variables
Bookkeeping: consumption of goods and food items during 2 weeks; 530 variables
Final Interview: consumption, income and demographic data concerning 12 months; 400 variables
overcoverage
Nonresponse
n = 11776
r = 8200
Increasing sets of information ->
decreasing sets of informants ->

Fig. 1. The sources and the scope of information in the 1985 Finnish Household Budget Survey

terms of estimates and their standard errors between the methods used in 1981 and in 1985. Thus, Section 5 gives our theoretical and Section 6 our empirical reasons for believing that the new estimator has less bias than the old one. We close in Section 7 with a short discussion of two alternative approaches, regression estimation, and post-stratification. All mathematical derivations are omitted from this paper and are included in Ekholm and Laaksonen (1990), a self-contained corresponding technical report obtainable on request. Laaksonen (1988) offers Finnish readers many more facts about the HBS and a general discussion on different ways of adjusting for nonresponse.

2. Selection and Response Probabilities

There is no register of households in Finland, but there is a well organized person-based official register. The sampling frame for the HBS is therefore this Central Population Register. Persons living in institutions and persons without a permanent address were removed from the sampling frame. The Finnish population so defined numbers 4.8×10^6. This population was divided into 24 regional strata. We denote the number of persons in the sampling frame for stratum h by M_h. The median of the M_h for $h = 1, \ldots, 24$ is 1.6×10^5. For stratum h CSO chose n_h persons from the sampling frame by systematic random sampling. The sampling fractions varied between strata, in order to give the strata with less population a higher sampling fraction. The median of the n_h is 396. The total number of sampled persons giving rise to sampled households was $\Sigma_h n_h = 11776$. The Central Population Register does not list individuals by households. The order is, in fact, best described as haphazard. We assume that the order is randomized and take the sampling of persons to be simple random inside each stratum separately, with the number of sampled households fixed in advance.

Using the population register a preliminary list was composed of the members in the households of each sampled person. Information about the address, the composition and size of the sampled households was thus available already before the households were approached. These data were, of course, checked and updated at the opening interview. Using tax and education registers, the income, wealth, and education of the sampled households were acquired. Thus, data of relatively high quality, although of limited scope, were available on all sampled households, responding and nonresponding alike. Figure 1 describes the pattern of increasing scope of information for a decreasing set of informants when we move from the construction of the sampling frame towards the conclusion of the final interview.

We denote the number of households who responded in stratum h by r_h. The median of the r_h is 290, and $\Sigma_h r_h = 8200$. The overall response rate is $\Sigma_h r_h / \Sigma_h n_h = 0.696$. Let the script letters $\mathscr{P}$, $\mathscr{S}$, and $\mathscr{R}$ denote, respectively, the *population*, the *sample*, and the *responding set* of households, so that $\mathscr{P} \supset \mathscr{S} \supseteq \mathscr{R}$. For stratum h of the population, of the sample, and of the responding set we write, respectively, $\mathscr{P}_h$, $\mathscr{S}_h$, and $\mathscr{R}_h$. The letters k and l index households. The number of persons that both belong to household k and appear in the sampling frame is denoted m_k. The number of households in the population is denoted N, and N_h in stratum h. These are unknown numbers to be estimated.

We write p_k for the probability that household k is selected to the sample. Define a selection indicator by $S_h(k) = 1$ if $k \in \mathscr{S}_h$, and $S_h(k) = 0$ otherwise. Let two households k and l both belong to $\mathscr{P}_h$ with $l \neq k$. Ekholm and Laaksonen (1990) indicate that the following formulae are very good approximations to the single and pairwise selection probabilities

$$p_k = pr(S_h(k) = 1) = (m_k \cdot n_h)/M_h \quad (1)$$

$$p_{k,l} = pr(S_h(k) = S_h(l) = 1) = \frac{m_k \cdot m_l \cdot n_h \cdot (n_h - 1)}{M_h \cdot (M_h - 1)}. \quad (2)$$

Since the sampling was performed independently for each stratum, we have for $h \neq h'$ that $\text{cov}(S_h(k), S_{h'}(l)) = 0$. The fact that the number of sampled households is fixed separately for each stratum implies that for $k, l \in \mathscr{P}_h$, with $h = 1, \ldots, 24$

$$-\sum_{l \neq k}^{\mathscr{P}_h} \text{cov}\,(S_h(k), S_h(l)) = \text{var}\,(S_h(k)). \quad (3)$$

Here the superscript on Σ indicates that the summation is over households in stratum h of the population.

Our estimation procedure assumes that any household k in the population has a fixed but unknown conditional probability $\theta_k = pr(k \in \mathcal{R} | k \in \mathcal{S})$ of responding, given that it is selected to the sample. We presume that θ_k is a function of the characteristics of the household, such as, its composition and the age and education of its members. We assume that different households respond, or do not respond, independently of each other. This implies for $k \in \mathcal{P}$ and $l \neq k$ that $\theta_{k,l} = pr(k \in \mathcal{R},\ l \in \mathcal{R} | k \in \mathcal{S},\ l \in \mathcal{S}) = \theta_k \cdot \theta_l$. This assumption is violated if some interviewers are much more skillful than others in persuading interviewees to respond. We disregard such subtle effects.

We denote the unconditional response probabilities as π_k, and $\pi_{k,l}$, respectively. For household k belonging to $\mathcal{P}_h$ we define a response indicator $R_h(k)$ by $R_h(k) = 1$ if $k \in \mathcal{R}_h$, and $R_h(k) = 0$ otherwise. It follows from the definitions of the conditional response probabilities and equations (1) and (2) that

$$\pi_k = pr(R_h(k) = 1) = p_k \cdot \theta_k$$

$$\pi_{k,l} = pr(R_h(k) = R_h(l) = 1)$$

$$= p_{k,l} \cdot \theta_k\, \theta_l.$$

The response indicators too are uncorrelated across strata, and inside strata the average absolute size of their correlations is smaller than for the selection indicators. Equation (3) is replaced by the following inequality, derived in Ekholm and Laaksonen (1990). Unless θ_l is unity for all $l \neq k$, then

$$-\sum_{l \neq k}^{\mathcal{P}_h} \operatorname{cov}(R_h(k), R_h(l)) < \operatorname{var}(R_h(k)). \quad (4)$$

Inequality (4) is a result of assuming a further randomness for responding, in addition to that introduced by the sampling procedure, and of assuming that households respond independently. This further randomness will affect the variance of our estimator. We shall then need a notation for the random number of responders in stratum h and denote it by R_h, defined as $R_h = \Sigma_{\mathcal{P}_h} R_h(k)$. The realized value of the random variable R_h has already been introduced and is denoted r_h above.

3. The New and the Old Estimator

We denote the amount of money in finnmarks that household k spent in 1985 on some particular consumption item, e.g., travel, by y_k. The task is to estimate the *total*, denoted Y, and the *mean per household*, denoted $\bar{Y}$, defined as follows

$$Y = y_1 + \ldots + y_N = \sum_{\mathcal{P}} y_k$$

$$\bar{Y} = Y/N.$$

We denote the estimates which we implemented for these quantities by $\hat{Y}$ and $\hat{\bar{Y}}$, respectively. The estimate for the number of households is written $\hat{N}$. To be useful in practice these three estimates should fulfill the equation

$$\hat{Y} = \hat{N} \cdot \hat{\bar{Y}}. \quad (5)$$

This reduces the task, essentially, to finding the expression for $\hat{Y}$, since $\hat{N}$ is obtained as a special case of $\hat{Y}$. We introduce the following "model based Horvitz–Thompson" estimator $\hat{Y}$

$$\hat{Y} = \sum_{\mathcal{R}} y_k / \hat{\pi}_k \quad (6)$$

where $\hat{\pi}_k = p_k \cdot \hat{\theta}_k$, with the unknown conditional response probabilities θ_k substituted by maximum likelihood estimates $\hat{\theta}_k$ from a logistic regression model. By substituting 1 for y_k we obtain $\hat{N}$ as

$$\hat{N} = \sum_{\mathcal{R}} (1/\hat{\pi}_k). \quad (7)$$

The selection probabilities p_k are calculated according to formula (1) with n_h and M_h varying between strata. Once the estimates $\hat{\pi}_k$ of the unconditional response probabilities are determined, the stratum division is not relevant, and the sum over $\mathscr{R}$ extends over all responding households. Särndal and Swensson (1987) would refer to estimator (6) as the "$\hat{\pi}$-expanded sum." Little (1986) suggests estimating response probabilities by logistic regression.

In Section 5 we shall contrast estimator (6) with the estimator used in reporting the 1981 HBS. The sampling frame was then the same, but the sampling was done without any stratification. After the interviews were completed, a poststratification was effected. The country was divided into 35 regional poststrata. An aim was to combine neighbouring municipalities with similar response rates to form poststrata. The estimator $\hat{Y}_o$, where the subscript stands for "old," was of familiar Horvitz–Thompson type

$$\hat{Y}_o = \sum_{\mathscr{R}} y_k/\pi_k^o \tag{8}$$

where

$$\begin{aligned} \pi_k^o &= (m_k \cdot r_h)/M_h \\ &= ((m_k \cdot n_h)/M_h) \cdot (r_h/n_h). \end{aligned} \tag{9}$$

The subscript $h = 1, \ldots, 35$ refers to the poststrata when we discuss the estimator $\hat{Y}_o$. The first form for the probabilities π_k^o in (9) is the one used for the actual calculations, while the second is our rewriting which shows that π_k^o can be interpreted as the product of the selection probability (1) and a conditional response probability r_h/n_h. The estimated probabilities π_k^o and $\hat{\pi}_k$ thus have a closely similar interpretation. The difference is that for the old estimator 35 regional response poststrata were used, while for the new estimator 128 different response propensity cells were established. Since sampling is from a person-based register, the numbers M_h of formulae (1) and (2) are known only for regional units, while any segmentation of a regional unit by factors like household structure or income gives rise to M_h values that are unknown. This seriously restricts the usefulness of poststratification for the present problem.

4. The Model for the Response Probabilities

The list of household characteristics available for modeling of responding was short, because these characteristics have to be known for all households in $\mathscr{S}$ and not only for those in $\mathscr{R}$ (see Section 1). Ekholm and Laaksonen (1990) describe the model search and motivate many of the choices made. Here we report only the essentials of the final model. Let θ_{abcd} denote the probability of response from a household in cell (a, b, c, d) of a four dimensional cross-classification. The model expresses the logit of the response probability as an additive function of parameters in the following way

$$\log \frac{\theta_{abcd}}{1 - \theta_{abcd}} = \mu + \alpha_a + \beta_b + \gamma_c + \delta_d. \tag{10}$$

The four explanatory factors are, with the corresponding generic index in parentheses, *Household structure (a)*, *Urbanism (b)*, *Region (c)*, and *Property income (d)*. The number of levels are, respectively, 8, 2, 4, 2. The categories are listed in Table 1, which also gives the parameter estimates and their standard errors. The parameter μ is the logit of the response probability for households belonging to the first level on a four factors. The parameters α_a, β_b, γ_c, and δ_d indicate the contrasts compared to μ for the other categories of the explanatory factors.

The fit of the model to the data is very

Table 1. The estimates of the parameters cum standard errors for the model defined by equation (10)

No	Factor	Code	Parameter	Estimate	Stand. err.
1	Baseline		μ	−0.4263	0.061
2	*Structure*	22	α_2	1.227	0.138
3	*Structure*	29	α_3	0.4732	0.072
4	*Structure*	35	α_4	0.7533	0.106
5	*Structure*	37	α_5	0.6616	0.105
6	*Structure*	45	α_6	0.7674	0.071
7	*Structure*	47	α_7	0.5020	0.074
8	*Structure*	49	α_8	0.2042	0.069
9	*Urbanism*	P	β_2	0.3768	0.044
10	*Region*	S	γ_2	0.5220	0.055
11	*Region*	M	γ_3	1.0700	0.065
12	*Region*	N	γ_4	0.7539	0.063
13	*Property*	I	δ_2	0.3118	0.048

The categories, and the codes, for the factors are: *Household structure:* **10** = One solitary person, **22** = Two persons, both 16–34 years of age, **29** = All other two person households, **35** = Three persons of which at least one is 0–6 years of age, **37** = Three persons of which at least one is 7–15 years of age, but none is under seven, **45** = At least four persons of which at least one is 0–6 years of age, **47** = At least four persons of which at least one is 7–15 years of age, but none is under seven, **49** = At least three persons of which none is under 16 years of age. *Urbanism:* **C** = Center, **P** = Periphery. *Region:* **U** = Uusimaa, **S** = South, except Uusimaa, **M** = Middle Finland, **N** = North. *Property income:* **NI** = No income from property, **I** = Income from property. The parameter for the first category is zero for all four factors, and the parameters for the other categories are contrasts against this baseline. The baseline thus comprises the cell with the following codes: **10**, **C**, **U**, **NI**.

good, except for three cells. The three deviant cells are (2,2,2,1), (3,2,2,1) and (3,1,1,2). When these cells are excluded from the estimation, and the fit is assessed, the log likelihood statistic for testing the model against the saturated model is 116.43 on 112 = 128 − 13 − 3 degrees of freedom. This corresponds to a p-value of 0.37. If these cells are included, then the p-value drops to 0.06. The fitted probabilities for these three cells are computed from the model in exactly the same way as for the other 125 cells, that is, by inserting the parameter estimates at the right hand side of equation (10) and solving for $\hat{\theta}_{abcd}$. The empirical response rates and the fitted probabilities for these three cells are, respectively, (0.67, 0.84), (0.82, 0.73), and (0.44, 0.58). The effect on the final estimates of these three cells is small. We consider it more appropriate to smooth the response probabilities in accordance with a model that fits perfectly for 125 cells, than in accordance with a model estimated from all the 128 cells but with moderate evidence against it. We shall see in Section 5 that it is important for us to use the "right" model. We firmly believe that the simple additive model is a much better approximation to the "right" model than any model which more closely estimates the observed response rates for these three cells. Inclusion of interaction terms does not significantly improve the fit of the model.

We report the structure of the estimated response probabilities in Figure 2. The

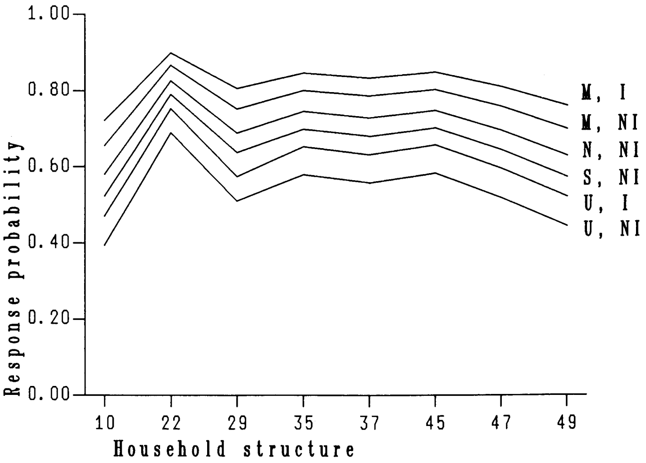

Fig. 2. Estimated conditional response probabilities for center part of regions (symbols in Table 1)

graph is restricted to the Center part of each region. To avoid overlapping, two lines are omitted. The lowest fitted response probability, 0.39, is for solitary persons, without property income, in the Center part of Uusimaa. The highest, 0.93, is scored by households of two young persons with property income in Middle Finland. The explanatory factor with the largest effect is household structure. The full distribution of the 128 different $\hat{\theta}$ values is depicted in Figure 3.

The use of a parsimonious model implies that the estimated probabilities have smaller standard errors. The estimates $\hat{\theta}_{abcd}$ borrow strength across cells. We illustrate the degree of smoothing of the θ's, and the size of their standard errors in Table 2. The table compares the empirical relative frequencies and the estimates from the model for five cells, which all are extreme cases in some respect. The standard errors for the estimates $\hat{\theta}_k$ are with little variation of the order 0.02.

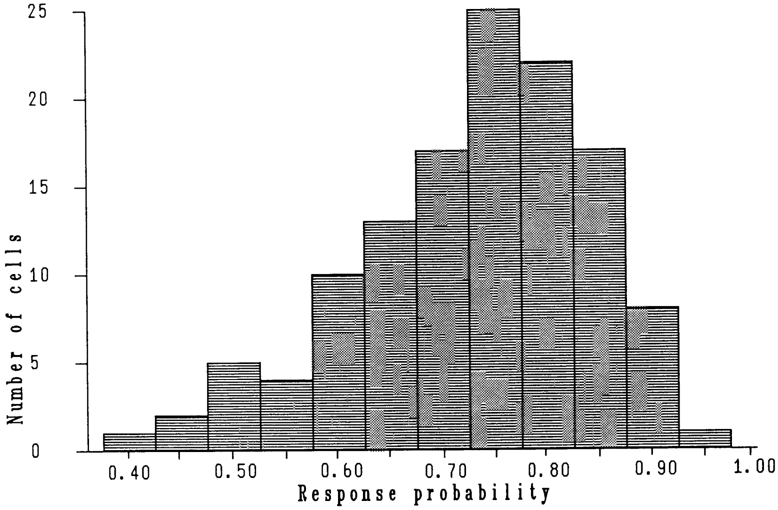

Fig. 3. Distribution of estimated response probabilities

Table 2. Empirical and estimated response probabilities, cum standard errors, for five extreme response propensity cells

Cell no	Raw Empirical				Estimated from Model		Characteristic
	r	n	r/n	s.e.	$\hat{\theta}$	s.e.	
(1,1,1,1)	116	294	0.39	0.029	0.39	0.015	min (r/n)
(1,1,2,1)	184	374	0.49	0.026	0.52	0.015	max (n)
(2,2,2,1)	24	36	0.67	0.079	0.85	0.018	max $(\|(r/n) - \hat{\theta}\|)$
(2,2,1,2)	4	6	0.67	0.192	0.82	0.021	min (n)
(2,2,3,2)	6	6	1.00	–	0.93	0.010	max (r/n)

In Section 5 we invoke the sufficient statistic of the model reported and commented upon in this section. It is well known that a logistic regression model has a minimal sufficient statistic of the same dimension as the vector of parameters, which here is 13 (see Cox and Snell 1989, sec. A1.11). Using any other transformation than the logit in (10) would result in a minimal sufficient statistic of dimension 128. We denote by $\mathbf{T}$ the minimal sufficient statistic under the present model. The observed value of $\mathbf{T}$ is $\mathbf{t}_{13}$. The elements of $\mathbf{t}_{13}$ are the sum and twelve different subsums of the elements of the 128×1 vector of numbers of responding households in the 128 adjustment cells.

5. Means and Variances

We consider first the old estimator $\hat{Y}_o$ and the setup as it was in 1981 with the index h referring to poststrata. We make

Assumption A_o: $pr(R_h(k) = 1 | S_h(k) = 1)$ depends on h but not on k.

The probability that household k responds is then independent of y_k, given the stratum division. Using Rubin's (1976) terminology, *nonresponse is ignorable within strata* (see also Rosenbaum 1987). Intuitively, it is clear that under ignorability the estimator $\hat{Y}_o$ is unbiased. Ekholm and Laaksonen (1990) give a formal proof.

In the HBS report for 1981 the conditional variance of $\hat{Y}_o$, given the values of R_h, $h = 1, \ldots, 35$, was used. This follows Holt and Smith's (1979) general recommendation for poststratification. Let $\mathbf{r}_{35}$ denote the vector of these 35 observed numbers. The estimated conditional variance, neglecting the finite population correction, is

$$\hat{V}(\hat{Y}_o | \mathbf{r}_{35}) = \sum_{h=1}^{35} r_h \cdot s_h^2(y/\pi^o, r_h) \qquad (11)$$

where $s^2(u, n) = (n - 1)^{-1} \Sigma_{k=1}^{n} (u_k - \bar{u})^2$ is the usual unbiased variance estimate from n independent observations $u_k = y_k/\pi_k^o$ of the random variable U. Assumption A_o and conditioning on $\mathbf{r}_{35}$ combine to give the response indicators the same negative correlation structure as the selection indicators have

$$- \sum_{l \neq k}^{\mathscr{P}_h} \mathrm{cov}\,(R_h(k), R_h(l) | \mathbf{r}_{35})$$
$$= \mathrm{var}\,(R_h(k) | \mathbf{r}_{35}). \qquad (12)$$

Compare equation (12) with equation (3) and inequality (4). It is unrealistic to assume a negative correlation structure for the response indicators, and we regard this as an unfortunate side effect of conditioning for $\mathbf{r}_{35}$.

We turn to the new estimator $\hat{Y}$ and consider the 1985 setup, with 24 strata at the sampling stage, and 128 adjustment cells constructed from the response model. We make

Assumption A:
(i) $pr(R_h(k) = 1 | S_h(k) = 1)$ *does not depend on either h nor* y_k*, but only on the adjustment cell to which household k belongs.*
(ii) *The model (10) for the response probabilities is the correct model for the data at hand.*

Part (i) means that *nonresponse is ignorable within adjustment cells.* We are aware that neither (i) nor (ii) strictly hold, but explore first the implications of A, and then return to a comparison of A_o and A.

If one conditions for the statistic **T** which is sufficient under the response probability model (10) and assumes part (ii) of A, then the conditional probabilities of responding in the adjustment cells are asymptotically equal to the probabilities estimated from the model. This remarkable asymptotic equivalence, between probabilities conditioning for a sufficient statistic and maximum likelihood estimates from the corresponding model, enlarges the scope of the Horvitz–Thompson approach. We shall now state, without detailed reference, some theoretical results concerning the model based Horvitz–Thompson estimator. Proofs can be found in Ekholm and Laaksonen (1990).

The data set at hand is large enough for the asymptotic equivalence to be valid, and we thus have

$$pr(R_h(k) = 1 | S_h(k) = 1, \mathbf{t}_{13}) \approx \hat\theta_k. \quad (13)$$

It follows from (13) that the new estimator $\hat Y$ is asymptotically unbiased under assumption A. We use the conditional variance of $\hat Y$ given the observed value of the sufficient statistic. We have not been able to find a usable expression for the covariances of the response indicators under the condition $\mathbf{t}_{13}$. Instead, we use the realistic assumption that households k and l respond independently of each other for all pairs (k, l). In particular we assume that

$$pr(R_h(k) = R_h(l) = 1 | S_h(k) = S_h(l) = 1, \mathbf{t}_{13}) = \hat\theta_k \cdot \hat\theta_l. \quad (14)$$

A slight upward approximation of the estimated conditional variance of the new estimator then produces the following useful expression

$$\hat V(\hat Y | \mathbf{t}_{13}) = \sum_h^{24} [r_h \cdot s_h^2(y/\hat\pi, r_h) + r_h(1 - r_h/n_h)\overline{(y/\hat\pi)}_h^2]. \quad (15)$$

The first term of expression (15) and expression (11) have the single difference that for $\hat V(\hat Y_o)$ one computes the variance estimate of the observations $(y_k/p_k) \cdot (n_h/r_h)$, while for $\hat V(\hat Y)$ one uses the observations $(y_k/p_k) \cdot (1/\hat\theta_k)$. The factor (n_h/r_h) stays fixed for a fixed h, but the factor $1/\hat\theta_k$ varies inside strata. It follows that $s_h^2(y/\hat\pi, r_h) > s_h^2(y/\pi^o, r_h)$. In addition, the estimated variance of the new estimator has the second term in equation (15). This derives from assumption (14). Let $\hat Y_h$ denote the component to $\hat Y$ from stratum $h = 1$. One finds that

$$V(\hat Y_h | \mathbf{t}_{13}) = E(R_h)V(U_h) + V(R_h)(E(U_h))^2 \quad (16)$$

where R_h denotes the random number of responders in stratum h, and U_h is a random variable with estimated mean $\overline{(y/\hat\pi)}_h$ and estimated variance $s_h^2(y/\hat\pi, r_h)$ from r_h independent observations $u_k = y_k/\hat\pi_k$. The second term in expression (15) is, accordingly, due solely to the fact that r_h is interpreted as the observed value of the random variable R_h. Conditioning on $\mathbf{t}_{13}$ and assuming equation (14) leads to a larger estimated variance, but is based on an assumption likely to lead to a reduced bias.

Interestingly, the estimated variance of the estimator of the mean per household is independent of whether we condition for the values of the R_h's or not. The usual first order Taylor expansion of $\hat{Y}/\hat{N}$ leads to the following expression for the estimated variance of the mean per household, conditioning for the sufficient statistic, and assuming (14)

$$\hat{V}(\hat{\bar{Y}}|\mathbf{t}_{13}) = (\hat{Y}/\hat{N})^2 \cdot \sum_h^{24} [\hat{V}(\hat{Y}_h|\mathbf{t}_{13})/\hat{Y}^2 + \hat{V}(\hat{N}_h|\mathbf{t}_{13})/\hat{N}^2 - 2\hat{C}(\hat{Y}_h, \hat{N}_h|\mathbf{t}_{13})/(\hat{Y}\hat{N})] \quad (17)$$

where $\hat{C}(\hat{Y}_h, \hat{N}_h|\mathbf{t}_{13})$ is the estimated covariance of $\hat{Y}_h$ and $\hat{N}_h$. It turns out that

$$\hat{C}(\hat{Y}_h, \hat{N}_h|\mathbf{t}_{13}) = r_h \cdot c(y/\hat{\pi}, 1/\hat{\pi}, r_h) + r_h(1 - r_h/n_h) \times \overline{(y/\hat{\pi})}_h \cdot \overline{(1/\hat{\pi})}_h \quad (18)$$

where

$$c(u, v, r) = (r - 1)^{-1}\Sigma_1^r(u_k - \bar{u})(v_k - \bar{v}).$$

The second term in expression (18) is again due to the randomness of R_h. There are analogous terms for $\hat{V}(\hat{Y}_h|\mathbf{t}_{13})$ and $\hat{V}(\hat{N}_h|\mathbf{t}_{13})$. When we substitute into equation (17) these three terms cancel. The heuristic reason for this result is that the numerator and the denominator of the estimator $\hat{Y}/\hat{N}$ depend on the very same realized values r_h for $h = 1, \ldots 24$.

The comparison of $\hat{Y}_o$ and $\hat{Y}$ shows: (i) that they are unbiased under assumptions A_o and A, respectively, (ii) that their variances differ, in the main because it is relevant to calculate variances under different conditions, and (iii) that the variances of the mean per household estimators have the same structure, despite different conditions. That shifts the focus of the comparison to the assumptions A_o and A. There are two components in both of these assumptions: the ignorability assumptions and the underlying statistical models for response propensity. We refer to the models, too, with the symbols A_o and A, respectively. We believe that A removes more of the nonresponse bias, for the following reasons:

1. Model A is based on four explanatory factors, while A_o is based only on region and urbanism. *Region* and *Urbanism*, with a more systematic classification, are two of the four factors in model A.
2. In constructing A we found a number of factors that did not provide any further explanation of the probability of responding. This does not, of course, prove that responding inside adjustment cells is independent of the y-values, but it does show that it is independent of some important determinants of the y-values. For the variable "taxable income" we report an empirical result in the next section.
3. Model A results in using 128 different adjustment cells, while model A_o uses only 35. The optimal number is, of course, unknown but we believe that the 35 strata are much too internally heterogenous.

Which of the estimators $\hat{Y}_o$ and $\hat{Y}$ should be preferred is finally an empirical question and could be settled definitely only by knowing the y-values for both responding and nonresponding households, from at least a random subsample of selected households. No such data were collected. Fortunately, a parallel study by the CSO can be used for comparison of the two estimators. We turn to that and some other empirical results.

6. Empirical Results

The most important single estimate provided by the HBS is the number of households in the country. In this section we denote this estimate $\hat{N}_o$, when calculated from the 1985 data by the old method, and $\hat{N}$ when calculated from the same data by the new method. It is particularly appropriate to compare the performance of the old and the new estimators by comparing $\hat{N}_o$ and $\hat{N}$, since they do not depend on the y-values but only on the response probabilities π_k^o and $\hat{\pi}_k$, respectively. The definition of $\hat{N}$ is given in equation (7), and $\hat{N}_o$ is defined in the same way using π_k^o, taking the selection strata as response rate strata. If the 1981 estimator had been used for 1985 data, then the selection strata would have been used as the strata in the sense of assumption A_o. The results for the whole country and for the southernmost and easternmost provinces are given in Table 3. We report the relative standard errors of the estimates, that is, the standard error divided by the estimate. The corresponding variances were computed from formulae (11) and (15), respectively.

The estimate produced by the new method for the number of households in the whole country is 4.2% greater than that produced by the old method. The difference is 3.5 times the standard error of the new estimate. The relative increase in the estimated number of households is even more dramatic for the provinces.

Table 3. The old and the new estimates in thousands of the number of households, cum relative standard errors (r.s.e.), for the whole country and two provinces

	$\hat{N}_o$	r.s.e.	$\hat{N}$	r.s.e.
Whole country	1960	0.0069	2042	0.0110
Uusimaa	504	0.0159	540	0.0276
North Carelia	70	0.0203	74	0.0351

In 1984 CSO collected data for the Income Distribution Survey, which also uses the household as the statistical unit. The sampling frame was the Central Population Register. Nonresponse was only 17%, presumably because there is only one short interview. The variations in response probabilities between population segments were considerably smaller than in the HBS. Therefore, the use of the old estimation method was better justified for the Income Distribution Survey than for the HBS. The estimated number of households in the whole country came to 2.01 × 10^6. Our estimate $\hat{N}$ is in much better agreement than the old estimate $\hat{N}_o$ with this independent source of information. We regard this as an empirical confirmation that $\hat{N}$ removes enough of the nonresponse bias to justify its larger standard error.

The relative standard errors of the new estimates are, indeed, larger than those for the old ones. The rates of increase for the three estimates in Table 3 are 1.6, 1.7, and 1.7. Much of this increase derives from the different conditioning. To assess the increase caused by using more variable weights, we computed the ratios $s_h(1/\hat{\pi}, r_h)/s_h(1/\pi^o, r_h)$ for $h = 1, \ldots, 24$. They range from 1.06 to 1.41, the median being 1.22. Still, both effects added, the relative standard errors of the new estimates are of a well tolerated size for the purpose at hand. This is true even for North Carelia which is the smallest province for which the CSO publishes detailed regional data.

In Table 4 we compare by index numbers the old and the new estimate, of the total and of the mean for four y-items. The first of these is "Final Consumption" which means the grand total. Table 4 also reports the relative standard errors of the new estimate.

The new and the old estimates differ in both directions. Obviously, households with high response probabilities, for instance,

Table 4. Index numbers for comparing the old and the new estimate for totals and means, for four groups of expenses in 1985; cum relative standard errors (r.s.e.) of the new estimate

	$100 \times \hat{Y}/\hat{Y}_0$	r.s.e.	$100 \times \hat{\bar{y}}/\hat{\bar{y}}_0$	r.s.e.
Final	100.6	0.0097	96.6	0.0083
Food	101.1	0.0094	96.9	0.0087
Furniture	99.2	0.0237	95.2	0.0232
Medical care	102.2	0.0262	98.1	0.0250

two young persons, spend much on furniture. Households with low response probabilities, for instance, old solitary persons, spend much on medical care. The new estimate of the mean per household is in all four cases smaller than the old estimate. This reflects the fact that $\hat{N}$ is considerably greater than $\hat{N}_o$.

The relative standard errors are much larger for furniture and medical care than for final and food consumption. This holds also for the mean per household. The standard error for the mean does not have a term deriving from the randomness of the r_h's. Therefore, we can conclude that the main source of variation for furniture and medical care is in the y_k-values and not in the weights, $1/\hat{\pi}_k$. Obviously, many households spend almost nothing on furniture and medical care, while others spend large sums.

A final empirical result concerns the difference in "taxable income" between responding and all households inside adjustment cells (see Section 5). Laaksonen (1988) reports, on the basis of register data, that the differences in "taxable income" inside the 128 cells are typically of size 1%, and differences larger than 3% are rare. In contrast, the global difference is 4.5%.

7. Discussion

The fundamental difference between the old estimator $\hat{Y}_o$ and the new $\hat{Y}$ is that the new one employs the auxiliary information available for both responding and nonresponding households (see Section 2 and Figure 1). It is conceivable that a regression of the y-values on the auxiliary variables would remove even more of the bias. Särndal and Swensson (1987) strongly recommend regression on an auxiliary variable when nonresponse is the second stage of a two-phase sampling procedure. Laaksonen (1988) reports part of his work on building regression models for the 1985 HBS data. He concludes that regression methods are useful for constructing estimates of the mean per household for some selected, highly aggregated consumptions. We decided not to implement regression methods for the main report on the following five interrelated grounds: (i) There are many y-items, even aggregated ones, for which no auxiliary variable with high, or even medium, explanatory power could be found. (ii) For disaggregated items, say finnmarks spent on fish, there are a considerable number of $y_k = 0$ observations at all levels of, for instance, income. These observations distort the use of regression models. (iii) If regression models are restricted to highly aggregated groups of expenses, then a problem of inconsistency between the reported group total and the reported subtotals arises. This would be confusing for the users of the CSO reports. (iv) No regression model constructed for one or several y-items is as such applicable for the estimation of N. Estimation of N for different segments is a crucial task, however. (v) It is conceptually clear and technically simple to use a single estimation procedure for all y-items and N alike. This advantage is highlighted by the fact that both totals and means are reported.

In the terminology of Little and Rubin (1987, sec. 4.4.3) the estimation we have used is a "weighting cell procedure." They warn that it may yield estimators with extremely high variances. As a remedy they suggest the use of response propensity cell means, but weighted by the exact population proportion, rather than by the inverse of the response probability. This is a form of poststratification. As pointed out in Section 4, the HBS regional poststrata cannot be subdivided by, for instance, household structure. Poststratification, therefore, is not a realistic alternative.

There are deficiencies in the register information. The most serious of these is in the data concerning the household composition. A promising suggestion for coming rounds of the HBS is to totally separate the opening interview from the actual interview. That could secure the information from the third box in Figure 1 from a very high percentage of the sampled households, perhaps as high as 95%. It would then be possible to model nonresponse on a much broader selection of explanatory factors.

7. References

Cox, D.R. and Snell, E.J. (1989). Analysis of Binary Data, Second Edition. London: Chapman and Hall.

Ekholm, A. and Laaksonen, S. (1990). Reweighting by Nonresponse Modeling in the Finnish Household Survey, Second Edition. Department of Statistics, University of Helsinki, Report No. 68.

Holt, D. and Smith, T.M.F. (1979). Post Stratification. Journal of The Royal Statistical Society, Ser. A, 142, 33–46.

Laaksonen, S. (1988). Correcting for Nonresponse in Household Survey Data. (In Finnish with Summary in English). Central Statistical Office of Finland, Studies No. 147, Helsinki.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157,

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley.

Rosenbaum, P.R. (1987). Model-Based Direct Adjustment. Journal of the American Statistical Association, 82, 387–394.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–592.

Särndal, C-E. and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. International Statistical Review, 55, 279–294.

Received February 1989
Revised March 1991