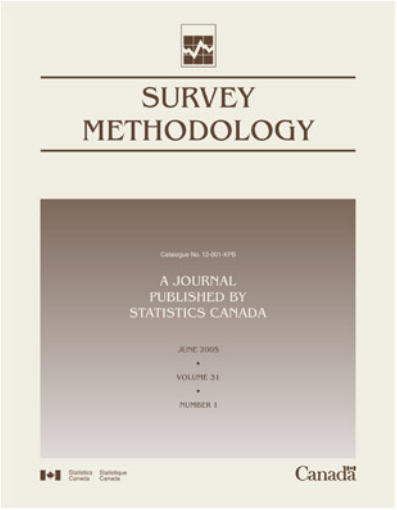
Catalogue no. 12-001-XIE

Survey Methodology

December 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 2, pp. 197-212
Statistics Canada, Catalogue No. 12-001

Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse Applied to the Norwegian Consumer Expenditure Survey

Liv Belsby, Jan Bjørnstad and Li-Chun Zhang [1]

Abstract

This paper considers the problem of estimating, in the presence of considerable nonignorable nonresponse, the number of private households of various sizes and the total number of households in Norway. The approach is model-based with a population model for household size given registered family size. We account for possible nonresponse biases by modeling the response mechanism conditional on household size. Various models are evaluated together with a maximum likelihood estimator and imputation-based poststratification. Comparisons are made with pure poststratification using registered family size as stratifier and estimation methods used in official statistics for The Norwegian Consumer Expenditure Survey. The study indicates that a modeling approach, including response modeling, poststratification and imputation are important ingredients for a satisfactory approach.

Key Words: Household size; Nonresponse; Imputation; Poststratification.

1. Introduction

This work is motivated by the considerable nonresponse rate in the Norwegian Consumer Expenditure Surveys (CES) for private households, for example 32% in the 1992 survey. Nonresponse involves both noncontact and refusal. We focus on the problem of nonignorable nonresponse that occurs when estimating the number of households of various sizes and the total number of households.

We shall consider a completely model-based approach; modeling and estimating the distribution of household size given registered family size and the response mechanism conditional on the household size. This model takes into account that the nonresponse mechanism may be nonignorable, in the sense that the probability of response is allowed to depend on the size of the household. The response model is used to correct for nonresponse. Model-based approaches with nonresponse included, sometimes called the prediction approach, have been considered by, among others, Little (1982), Greenlees, Reece and Zieschang (1982), Baker and Laird (1988), Bjørnstad and Walsøe (1991), Bjørnstad and Skjold (1992) and Forster and Smith (1998).

For various models of household size and response we consider mainly two model-based approaches, a maximum likelihood estimator and imputation-based poststratification after registered family size. These methods are compared to pure poststratification and the methods in current use in CES.

The main issue here is a comparison of models and methods with estimation bias as the basic problem. In addition, standard errors of the estimates and differences of the estimates, conditional on the sizes of post-strata determined by family size, are estimated using a bootstrap approach. In addition to assessing the statistical uncertainty of the estimators, this is done to help evaluate the extent to which differences between the proposed estimators are attributable to sampling error, nonresponse bias or both. However, in this evaluation we keep in mind the following quote from Little and Rubin (1987, page 67): "It is important to emphasize that in many applications the issue of nonresponse bias is often more crucial than that of variance. In fact, it has been argued that providing a valid estimate of sampling variance is worse than providing no estimate if the estimator has a large bias, which dominates the mean squared error."

Section 2 describes the data-structure and the sample design of CES, and Section 3 considers modeling issues. Section 3.1 presents the various models for household size and response to be considered for the 1992 CES, Section 3.2 describes the maximum likelihood method for parameter estimation, and in Section 3.3 the models are evaluated. A family size group model for household size and a logistic link for the response probability using household size as a categorical variable give the best fit of the models under consideration. Section 3.4 gives the estimated household size distributions for different family sizes and estimated response probabilities for different household sizes.

1. Liv Belsby, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: lbe@ssb.no; Jan F. Bjørnstad, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: jab@ssb.no and Li-Chun Zhang, Statistics Norway, Division of Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo. E-mail: lcz@ssb.no.

Section 4 considers model-based estimation, the imputation method, imputation-based estimators and the variance estimation method. It is shown that for the chosen model for household size from Section 3.3, the maximum likelihood estimator and the imputation-based poststratified estimator are identical.

Section 5 deals with the main goal of estimating the total number of household of various sizes based on the 1992 CES, using the estimators in Section 4. The model that gave the best fit seems to work well for our estimation problem. We conclude that poststratification, response modeling and imputation are key ingredients for a satisfactory approach.

2. Norwegian Consumer Expenditure Survey

The population totals within household-size categories provide a more correct number of dwellings than the totals within family-size categories from the Norwegian Family Register. Furthermore, the authorities for evaluating eventual policy intervention aimed at housing construction use the estimated number of households. Estimating household-size totals is therefore an important issue in social planning. It is invariably affected by nonignorable nonresponse, no matter what kind of survey one uses. Hence, it is a good illustration for how to handle nonresponse bias. We shall base our estimation on the Norwegian Consumer Expenditure Surveys (CES), where it is important to gain information about the composition of households, since household size influences consumption.

The actual CES, the survey for expenditure variables, is a sample of private households from all private households in Norway. This is done by selecting a sample of persons and including the whole households these persons belong to. Persons older than 80 years old are excluded since they often live in institutions. For our purpose, the units of interest in the survey are *persons* between the ages of 16 and 80 living in private households, and the variable of interest is the size of the *household* the person belongs to, which is observed only in the response sample of the persons selected.

The sample design is a three-stage self-weighting sample of persons. That is, every person in the population has the same inclusion probability to the total sample. The first two stages select geographical areas in a stratified way, while at the third stage persons are selected randomly from the chosen geographical areas. The primary sampling units (PSU) at stage 1 consists of the municipalities in Norway. Municipalities with less than 3,000 inhabitants are grouped together such that each PSU consists of at least 3,000 persons. The PSUs are first grouped into 10 regions and within each region stratified according to size (number of inhabitants) and type of municipality (*i.e.*, industrial structure and centrality). Totally, we have 102 strata. Towns of more than 30,000 inhabitants are their own strata and therefore selected with certainty at stage 1. For the other strata, one PSU is selected with probability proportional to size. At the second stage, the selected PSUs are divided into three smaller areas (secondary sampling units, SSU) and one of these is selected at random. Finally, at the third stage, for each of the selected SSU, a random sample of persons is selected. The sample sizes for each selected SSU are determined such that the resulting total sample of persons is self-weighting.

Our application is based on the data from the 1992 CES. CES is a yearly survey and since 1992 a modified Horvitz-Thompson estimator, including a correction for nonresponse by estimating response probabilities given household size, has been employed (see Belsby 1995). The weights equal the inverse of the probability of being selected multiplied with the conditional probability of response given selected. Since 1993 the probability of response is estimated with a logistic model with auxiliary variables being place of residence (rural/urban), and household size. For most of the nonrespondents the family size is used as a substitute for the household size.

A household is defined as persons having a common dwelling and sharing at least one meal each day (having common board). For a complete description of CES we refer to Statistics Norway (1996). In CES, the auxiliary variables known for the total sample, including the nonrespondents, are the family size, the time of the survey (summer/not summer), and the place of residence (urban/rural). *Families* are registered in Norwegian Family Register, (*NFR*), and may differ from the household the persons in the family belong to, both by definition and because of changes not yet registered. Hence, the registered *family* size from *NFR* differs to some extent from the household size. Initially, based on experience from previous surveys, all the auxiliary variables and household size are assumed to affect the response rate.

Table 1 shows the data for the 1992 CES with a total sample of 1,698 persons. The households with size five and greater are collapsed due to the low frequency in the sample of households. We base our modeling and estimation on two corresponding tables, one for the persons in rural areas and one for the persons in urban areas. These data are given in table A1 in appendix A1.

For example, the number 48 in cell (1,2) means that of the 162 persons registered to live alone in the response sample, 48 are actually living in a two-persons household. This is explained mostly by young people's tendency to cohabitate without being married; see Keilman and Brunborg (1995).

Table 1
Family and household sizes for the 1992 Norwegian Consumer Expenditure Survey

	Household size							
Family size	1	2	3	4	≥ 5	Total	Nonresponse	Response rate
1	83	48	20	9	2	162	153	0.514
2	9	177	37	4	3	230	160	0.590
3	10	25	131	40	6	212	91	0.700
4	2	13	37	231	17	300	123	0.709
≥ 5	1	4	4	17	181	207	60	0.775
Total	105	267	229	301	209	1,111	587	0.654

3. Modeling of Household Size and Nonresponse

We shall assume a population model for the household size, given auxiliary variables, *i.e.*, we model the conditional probability. To take nonresponse into account in the statistical analysis, we must model the response mechanism, *i.e.*, the distribution of response conditional on the household size and auxiliary variables. The sampling mechanism for persons is ignorable for the survey we consider, *i.e.*, is independent of the population vector of household sizes. The statistical analysis is therefore done *conditional* on the total sample, following the likelihood principle (see Bjørnstad 1996). Hence, probability considerations based on the sampling design is irrelevant in the statistical analysis. This is the so-called prediction approach. However, when evaluating the estimation methods with regard to statistical uncertainty, we do this from a common randomization perspective as described in Section 4.3.

For CES, the auxiliary vector consists of the family size, place of residence divided into rural and urban areas, and time of the data collection.

3.1 The Models

Let us first consider a simple model for the household size, denoted by *Y*. Let $\mathbf{x}$ denote all auxiliary variables. The household size is assumed to depend only on the family size *x*, and as such is a model with a restricted parametric link function, but with no additional assumptions,

$$P(Y_i = y \mid \mathbf{x}_i) = P(Y_i = y \mid x_i) = p_{y,x_i}, \qquad (3.1)$$

where

$$\sum_y p_{y,x_i} = 1, \text{ for each possible value of } x_i.$$

The model (3.1) is flexible in the sense that it does not include any restrictions on the assumed model function of x_i. The drawback is the high number of parameters compared with a model using a logistic type model with a linear, in $\mathbf{x}$, link function (the function linking $P(Y = y)$ with $\mathbf{x}$). If nonresponse is ignored the estimates in this model would simply be the observed rates.

Household size defines ordered categories. Thus a natural choice for a model is the cumulative logit model, known as the proportional-odds model (see McCullagh and Nelder 1991), assuming (with θ_y increasing in y)

$$P(Y_i \leq y \mid \mathbf{x}) = \begin{cases} \dfrac{1}{1+\exp(-\theta_y + \beta' \mathbf{x})} & \text{for } y = 1, 2, 3, 4 \\ 1 & \text{for } y \geq 5. \end{cases}$$

However, a goodness of fit test, with $\mathbf{x}$ consisting of family size and place of residence, indicated that this model fits the data badly. Thus we choose to reject it.

It is assumed that the probability of nonresponse may depend on the household size. For example, one-person households are less likely to respond than households of larger size since larger households are easier to "find at home". Nonresponse is indicated by the variable R, where $R_i = 1$ if person i responds and 0 otherwise. Let R_s be the vector of these indicators in the total sample. From Bjørnstad (1996), the response mechanism (RM), *i.e.*, the conditional distribution of R_s given the $\mathbf{x}-$values in the population and the $y-$values in the total sample, is defined to be ignorable if it can be discarded in a likelihood-based analysis. This means that RM is ignorable if this conditional distribution of R_s does not depend on the unobserved $y-$values, coinciding with the definition used by Little and Rubin (1987, pages 90, 218). For our case it is assumed that all pairs (Y_i, R_i) are independent. Then RM is ignorable if Y_i and R_i are independent. Hence, nonignorable response mechanism is equivalent to

$$P(Y_i = y_i \mid \mathbf{x}_i, r_i = 0) \neq P(Y_i = y_i \mid \mathbf{x}_i, r_i = 1)$$

and then both are different from $P(Y_i = y_i \mid \mathbf{x}_i)$.

Thus estimating the parameters in the model for $P(Y = y \mid \mathbf{x})$ using only the response sample, ignoring that the probability of response depends on the household size, would most likely give biased estimates for the unknown parameters. Also the poststratification estimator would give

biased estimates because it assumes that the distribution of R only depends on the auxiliary $\mathbf{x}$. *E.g.*, the observed lower response rate among one-person families indicates that the same may hold for one-person households. If so, the estimated probability of household size 1, based on respondents only, would be too small. Poststratification with respect to family size will most likely correct only some of this bias.

The model for the probability of response, given auxiliary variables and household size y_i, is assumed to be logistic. It depends on the auxiliary variables $\mathbf{z}_i$, which includes part of $\mathbf{x}_i$, expressed by

RM1$(y, \mathbf{z})$:

$$P(R_i = 1 \mid y_i, \mathbf{z}_i) = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi^t \mathbf{z}_i)}. \quad (3.2)$$

Here, α and γ are scalar parameters and ψ is a vector. The variable y_i has an order. Motivated by this fact, and to avoid introducing many parameters, y_i is used in (3.2) as an ordinal variable rather than a class variable. Thus the logit function,

$$\log\{P(R_i = 1 \mid y_i, \mathbf{z}_i)/P(R_i = 0 \mid y_i, \mathbf{z}_i)\} = \alpha + \gamma y_i + \psi^t \mathbf{z}_i,$$

is linear in y_i. To avoid the assumption of linear logit in y_i, we also consider a model with y_i as a categorical variable, *i.e.*,

$$\text{RM2}(y, \mathbf{z}): P(R_i = 1 \mid y_i, \mathbf{z}_i) = \frac{1}{1 + \exp\begin{pmatrix} -\alpha_0 - \alpha_1 I_1(y_i) - \alpha_2 I_2(y_i) \\ -\alpha_3 I_3(y_i) - \alpha_4 I_4(y_i) - \psi^t \mathbf{z}_i \end{pmatrix}}, \quad (3.3)$$

where the indicator variable $I_y(y_i)$ equals 1 if $y_i = y$ and 0 otherwise. The drawback with this model is that it includes three parameters more than model (3.2).

3.2 Maximum Likelihood Parameter Estimation

All the selected persons in the sample are from different households (duplicates have been removed), The population model then assumes that the household sizes Y_i are statistically independent. For *this* variable, interviewer- or cluster- effect plays no role.

Let us consider the likelihood function for estimating the unknown parameters, assuming that all pairs (Y_i, R_i) are independent and response model RM1 given by (3.2). To simplify notation we relabel the observations such that observations 1 to n_r are the respondents and observations $n_r + 1$ to n are the nonrespondents. With response model RM2 the expression for the likelihood is of the same form with (3.3) replacing (3.2).

For the respondents let $L_i = P(Y_i = y_i \cap R_i = 1 \mid \mathbf{x}_i)$. Then, for model (3.1)

$$L_i = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi^t \mathbf{z}_i)} \cdot p_{y_i, x_i}, \; i = 1, ..., n_r \quad (3.4)$$

For the nonrespondents let $L_i = P(R_i = 0 \mid \mathbf{x}_i)$. Then

$$L_i = \sum_{y=1}^{5} \frac{1}{1 + \exp(\alpha + \gamma y + \psi^t \mathbf{z}_i)} \cdot p_{y, x_i}, i = n_r + 1, ..., n. \quad (3.5)$$

The likelihood function for the entire sample of persons from different households is given by

$$L(\theta, \beta, \alpha, \gamma, \psi) = \prod_{i=1}^{n} L_i. \quad (3.6)$$

For $i = 1, ..., n_r$, L_i is according to (3.4) and for $i = n_r + 1, ..., n$, L_i is given by (3.5).

Estimates are found by maximizing the likelihood function (3.6). The maximization was done numerically using the software TSP (1991) see Hall, Cummins and Schnake (1991). The optimizing algorithm is a standard gradient method, using the analytical first and second derivatives. These are obtained by the program, saving us a substantial piece of programming. The model fitting is based on the chi-square statistic and on the $t-$values, provided by TSP, where the standard errors are derived from the analytical second derivatives. The $t-$values have to be interpreted with some care, since the unbiasedness of the estimated standard errors depends on how well the model is specified as well as the number of observations compared with the number of parameters.

3.3 Evaluation of the Models for Household Size and Response

We present the fit of the models with the Pearson goodness-of-fit statistics. The model study is based on the 1992 CES. The parameters are considered to be significant when the absolute $t-$values are greater than 2. However, we do not want a model that is too restrictive, and therefore some variables are kept even though their absolute $t-$values are less than 2.

In the response models RM1 and RM2 we use the variable $\mathbf{z} = z$, place of residence. We let $z = 0$ if rural area and $z = 1$ if urban area. It was observed in the CES $1986-88$ and CES $1992-94$, see Statistics Norway (1990, 1996), that there is more nonresponse during the summer. Therefore, the time of the survey was also included in the model, that is whether or not the data were collected in the period May $21-$August 12. However, the time of the survey was found to be nonsignificant, with $t-$value clearly less than 2. Also the family size was found to be nonsignificant. But if the household size is omitted in the response model then the family size turns out to be significant.

Ideally, we want to take a look at the empirical logit function for response with respect to the household size. However, household size is unavailable for the non-respondents. As a replacement we plot the logit-function against the family size; see figure 1. From family size one to two the two functions for rural and urban families increase in a fairly parallel way . However, for family size three and four the logit functions depart from being linear and parallel. Thus we suspect that coding the household size as a categorial variable, as in model RM2, will give better fit than restricting the logit functions to be parallel for rural and urban and linear with respect to the household size, as in model RM1.

In order to test the goodness of fit of the models, we consider the Pearson chi-square statistic, conditional on the auxiliary variables x, z. Given rural or urban type of residence and registered family size, there are six possible outcomes; household sizes 1, ..., 5 and nonresponse. Altogether there are ten multinomial trials and sixty cells. For family sizes (1,2) and (4,5), the extreme household sizes (4,5) and (1,2), respectively, are combined because the expected sizes under the models are too small. This reduces the number of cells to 52. The degrees of freedom (d.f.) is calculated as: number of cells – number of trials – number of parameters. For model (3.1) & RM1(y, z), d.f. = $52 - 10 - (20 + 3) = 19$, and for (3.1) & RM2(y, z), d.f. = $52 - 10 - (20 + 6) = 16$. For model (3.1) & RM1(y, z) the Pearson statistic χ^2 is 26.35 and the $p-$value is 0.121. And for model (3.1) & RM2(y, z) χ^2 is 21.77 and the $p-$ value is 0.151.

By studying the standardized residuals, $(observed-expected)/\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{observed})}$, we find that the main reason for the better fit is that model (3.1) & RM2(y, z) does a better job of predicting the observed counts for the urban area where the response rate is lowest (see appendix A1). Thus the data indicates that coding the household size as a categorial variable, as in RM2, improves the fit compared to using it as an ordinal variable. The model (3.1), with the restricted parametric link function, combined with RM2 is the best of the models we have considered so far.

3.4 Estimated Household Size Distribution and Response Probabilities

Table 2 displays the estimates for the population model (3.1) together with the logistic response model RM2 in (3.3).

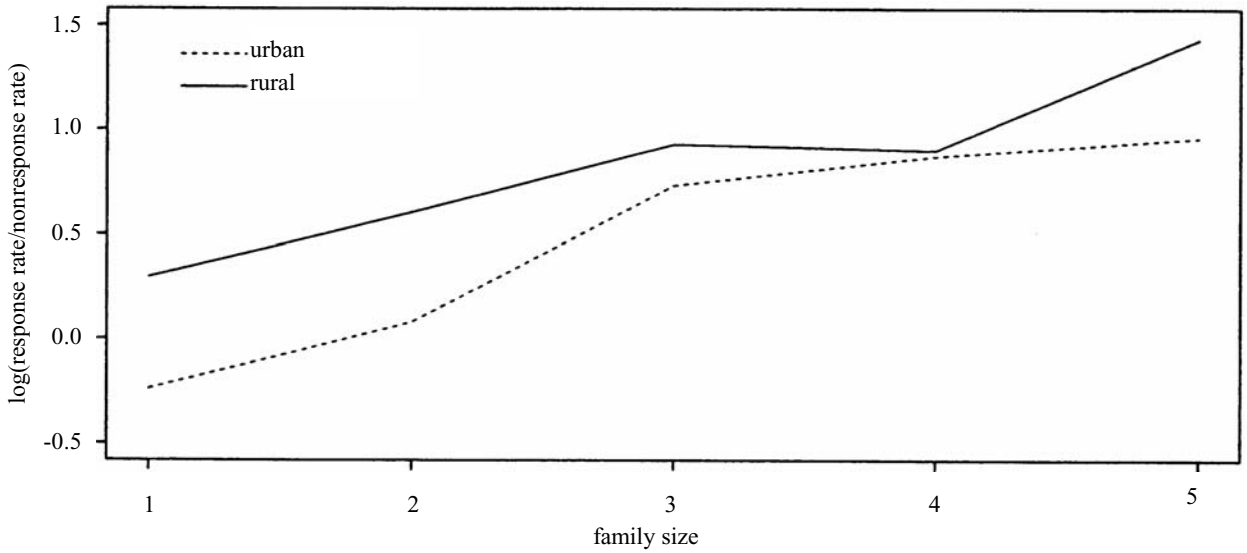


Figure 1. The logit function for the empirical response rate with respect to family size 1, ..., 5 in urban and rural areas, respectively. The computation is based on respondents and nonrespondents from Table 1 in Appendix A1.

Table 2
1992 CES. Parameter Estimates, in Percentages, for the Population Model with a Restricted Parametric Link Function, $p_{y,x}$, Combined with the Logistic Response Model RM2(y, z). In Parentheses are the Estimates for the Population Model, Ignoring the Response Mechanism

	Household size				
Family size, x	1	2	3	4	5 or more
1	60.01 (51.23)	26.75 (29.63)	8.35 (12.35)	4.09 (5.56)	0.80 (1.23)
2	5.27 (3.91)	79.80 (76.98)	12.48 (16.09)	1.47 (1.74)	0.98 (1.30)
3	7.53 (4.72)	14.45 (11.79)	56.67 (61.79)	18.85 (18.87)	2.50 (2.83)
4	1.06 (0.67)	5.31 (4.33)	11.38 (12.33)	77.20 (77.00)	5.05 (5.67)
5 or more	0.84 (0.48)	2.60 (1.93)	1.96 (1.93)	9.05 (8.21)	85.55 (87.44)

Let us interpret some of the values in the household model. Taking the response mechanism into account has largest effect on the estimated household distribution for one-person families. The probability that a household size equals one, given that the family size is one, is estimated as 60.01%. The estimate based on the traditional approach, ignoring the nonresponse, is 51.23%. The response model "adjusts" the observed rate among the respondents to a higher value. This seems reasonable since the rate of non-respondents is higher for small households. The estimated probability of household size five or more, given family size of five or more is 85.55%, which differs little from the observed rate among the respondents, 87.44%. This indicates that, given family size five or more, the household size distribution is about the same among respondents and nonrespondents.

Table 3 presents the estimated response probabilities based on RM2 in combination with the population model (3.1). Furthermore, we present estimated response probabilities based on a saturated model, with perfect fit, presented in Section 4.2. The model, defined by (4.9), assumes that the response probability for persons with the same household size within rural/urban area, respectively, is identical for different family sizes. Moreover, the model for household size depends on place of residence and family size, but with no restriction on the link function. We note that RM2(y, z) satisfies (4.9b), but is more restrictive. Model (4.9) allows for more freedom than model (3.1) with RM2(y, z).

Table 3
Estimated Probability of Response Based on the Logistic Model RM2 in Combination with (3.1), and the Saturated Model (4.9). The Estimates are Given in Percentages

	Household size				
Place of residence	1	2	3	4	5 or more
	Estimated response probabilities for model RM2				
Rural	47.77	60.90	79.16	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46
	Estimated response probabilities for the saturated model				
Rural	50.79	62.37	76.90	70.57	83.07
Urban	35.17	50.85	74.79	70.68	72.89

The estimated response probabilities reflect the lower response rate among one-person households, and the lower response rate in urban areas. Households of size five and higher have the highest response rate. The models estimate, surprisingly maybe, that the the probability of response is higher for households of size three than for households of size four. This may be explained by the fact that women often choose to have two children, and that three-person-households mostly consist of mother, father and a *small* child. Such a family will tend to stay at home and thus be more accessible than a typical four-persons-family with two older children.

The higher estimated response rate for households of size three compared to size four is equivalent to the ratio $P(Y=3\,|\,R=1)/P(Y=3\,|\,R=0)$ being greater than the ratio $P(Y=4\,|\,R=1)/P(Y=4\,|\,R=0)$. This is consistent with the household distribution in table 2, where we estimate that $P(Y=4)\approx P(Y=4\,|\,R=1)$, *i.e.*, $P(Y=4\,|\,R=0)\approx P(Y=4\,|\,R=1)$. On the other hand, the estimates in table 2 indicate that $P(Y=3\,|\,R=1)>P(Y=3)$ which means that $P(Y=3\,|\,R=1)>P(Y=3\,|\,R=0)$.

We see that the logistic model RM2 combined with the population model with the restricted parametric link $p_{y,x}$ acts as a smoother of the estimates based on the saturated model in (4.9), because of the added assumption of parallel logits of the response probabilities for urban and rural areas.

4. Estimators for Household Size Totals

In this section we present the estimators for household size totals and the method for variance estimation. We use a maximum likelihood estimator with the restricted parametric link function in (3.1) as population model. It is shown that this estimator is identical to an imputation-based poststratified estimator, which again turns out as a standard poststratification when the response mechanism is ignored. Furthermore, we present an imputed poststratified estimator, based on a saturated model for household size and response probability.

4.1 Estimators Based on a Restricted Parametric Link Function as Population Model

With N_y denoting the total number of persons living in households of size y, the number of households of size y equals $H_y = N_y/y$. The total number of households is denoted by H, $H=\sum_y H_y$.

The statistical problem is to estimate H_y for $y=1, ..., J$ and H. The largest size J is chosen such that there are few households of size greater than J. Strictly speaking, H_J is the number of households of size J or more, and likewise for N_J. In our application we choose $J=5$ due to the low frequency in the sample of households of size greater than five. We can write $N_y=\sum_{i=1}^{N} I(Y_i=y)$, where the indicator function $I(Y_i=y)=1$ if $Y_i=y$, and 0 otherwise. Hence, with $\mathbf{x}=(\mathbf{x}_1, ..., \mathbf{x}_N)$,

$$E(H_y\,|\,\mathbf{x})=\frac{1}{y}\sum_{i=1}^{N}P(Y_i=y\,|\,\mathbf{x}_i).$$

A maximum likelihood based estimator for H_y can be obtained by estimating $E(H_y\,|\,\mathbf{x})$, *i.e.*, replacing $P(Y_i=y\,|\,\mathbf{x}_i)$ by the maximum likelihood estimator

$\hat{P}(Y_i = y \mid \mathbf{x}_i)$. The data is stratified according to family sizes 1, ..., K, where the last category contains persons belonging to families of sizes $\geq K$. Using the model with the restricted parametric link function, defined in (3.1), Y is assumed to depend only on the family size x, and the estimator takes the form

$$\hat{H}_y = \frac{1}{y}\sum_{x=1}^{K} M_x \hat{P}(Y = y \mid x) \qquad (4.1)$$

where M_x (M_K) denotes the number of persons in the population with registered family size $x(\geq K)$. The M_x's are known auxiliary information from the Norwegian Family Register.

A common approach to correct for nonresponse is by imputation of the missing values in the sample. Based on the estimated distribution for Y for a given family size and place of residence for the nonrespondents, $\hat{P}(Y = y \mid x, z, r = 0)$, we assign the nonrespondents to the values 1, ..., 5 in proportions given by $\hat{P}(Y = y \mid x, z, r = 0)$ for $y = 1, ..., 5$. Let $n^*_{xy}(0)$ ($n^*_{xy}(1)$) be the number of imputed values with family size x and household size y, for rural (urban) areas and let $m_{xu}(0)$ ($m_{xu}(1)$) be the number of missing observations for persons in rural (urban) areas with family size x. Then

$$n^*_{xy}(z) = m_{xu}(z) \cdot \hat{P}(Y = y \mid x, z, r = 0), z = 0, 1. \quad (4.2)$$

and

$$n^*_{xy} = n^*_{xy}(0) + n^*_{xy}(1)$$

is the total number of imputed values with family size x and household size y, *i.e.*, n^*_{xy} is the estimated expected number of households of size y, *given* family size x and $r = 0$.

The following general result holds, showing that with population model (3.1), the maximum likelihood estimator (4.1) is identical to an imputation-based poststratified estimator.

Theorem. Assume model (3.1) for Y. That is, $P(Y = y \mid x, z) = p_{y,x}$ is independent of z, but otherwise the $p_{y,x}$'s are completely unknown with the only restriction $\sum_y p_{y,x} = 1$, for all values of x. The response mechanism is arbitrarily parametrized, *i.e.*, no assumption is made about $P(R = 1 \mid Y = y, x, z)$. Then the maximum likelihood estimates for $p_{y,x}$ are given by, for $x = 1, ..., K$,

$$\hat{p}_{y,x} = \frac{n_{xy} + n^*_{xy}}{m_x + m_{xu}},$$

where n_{xy} is the number of respondents belonging to a family of size x and household size y, m_x (m_K) is the number of respondents belonging to families of size $x(\geq K)$, and $m_{xu} = m_{xu}(0) + m_{xu}(1)$.

Proof. See Appendix A2.

The theorem implies that the estimator can be written as the imputation-based poststratified estimator, using family size as the stratifying variable ,

$$\hat{H}^I_{y,\text{post}} = \frac{1}{y}\sum_{x=1}^{K} M_x \frac{n_{xy} + n^*_{xy}}{m_x + m_{xu}}. \qquad (4.3)$$

Assuming ignorable response mechanism and using the model (3.1), the likelihood function is given by $\prod_{i=1}^{n_r} P(Y_i = y_i \mid x_i)$. Then the maximum likelihood estimate $\hat{P}(Y = y \mid x)$ is simply the observed rate among the respondents with household size y, given family size x. Thus the maximum likelihood estimator turns out to be identical to the standard poststratified estimator, with family size as the stratifying variable,

$$\hat{H}_{y,\text{post}} = \frac{1}{y}\sum_{x=1}^{K} M_x \frac{n_{xy}}{m_x}. \qquad (4.4)$$

For a general study of poststratification see, for example Holt and Smith (1979) and Särndal, Swensson and Wretman (1992, chapter 7.6).

To illustrate the effects of nonresponse modeling and poststratification, we also present estimates based on the regular expansion estimator, given by

$$\hat{H}_{y,e} = \frac{1}{y} \cdot N \frac{n_y}{n_r} \qquad (4.5)$$

and the imputation-based expansion estimator given by

$$\hat{H}^I_{y,e} = \frac{1}{y} \cdot N \frac{n_y + n^*_y}{n}. \qquad (4.6)$$

Here, n_y is the number of respondents in households of size y, n_r is the total number of respondents, and $n^*_y = \sum_x n^*_{xy}$. The estimator (4.5) does not seek to correct for nonresponse nor use the family population distribution as a post-stratifying tool to improve the estimation, while estimator (4.6) tries to take the response mechanism into account, but cannot correct for nonrepresentative samples.

4.2 Imputation-based Poststratification with a Saturated Model

We now proceed to an intuitive method of imputation that was used to estimate response probabilities for a modified Horvitz-Thompson estimator in the official statistics from the 1992 CES (described in Belsby 1995). We will use this imputation method for the poststratified estimator (4.3).

The imputation method consists of distributing, within rural/urban area, the $m_{xu}(z)$ nonresponse units over the household sizes 1, ..., 5 in such a way that, given

household size, the rate of nonresponse is the same for all family sizes. It implicitly assumes that the response probability for persons with the same household size within rural/urban area is identical for different family sizes. Denote the number of nonresponse persons with family size x and household size y and place of residence z obtained in this manner by $h_{xy}(z)$. The corresponding number among the respondents is $n_{xy}(z)$. The values of $h_{xy}(z)$ are determined by the equations

$$\frac{h_{xy}(z)}{h_{xy}(z)+n_{xy}(z)} = \frac{h_{iy}(z)}{h_{iy}(z)+n_{iy}(z)}, \quad z=0,1. \qquad (4.7)$$

When $n_{xy}(z)=0$, we let $h_{xy}(z)=0$. The equation (4.7) is solved under the conditions

$$\sum_y h_{xy}(z) = m_{xu}(z);\ x=1,2,3,4,5 \text{ and } z=0,1. \qquad (4.8)$$

Solving (4.7) and (4.8) requires, for each value of z, one row $(n_{x1}(z), n_{x2}(z), ..., n_{x5}(z))$ of nonzeros, which holds for our case. The imputed values $h_{xy}(z)$ determined by (4.7) and (4.8) correspond to the imputation method described by (4.2) for the following model:

$$P(Y=y\,|\,x,z) = p_{y,x,z} \text{ with no restrictions} \qquad (4.9a)$$

$$P(R=1\,|\,Y=y,x,z) = q_{y,z}, \text{ independent of } x. \qquad (4.9b)$$

This can be seen as follows:

For the ten multinomial trials determined by the different $(x,z)-$values, we have 50 unknown cell probabilities $\pi_{yx,z} = P(Y=y,\ R=1\,|\,x,\ z)$. With no restrictions on cell probabilities, the maximum likelihood estimates (mle) are given by observed relative frequencies,

$$\hat{\pi}_{yx,z} = \frac{n_{xy}(z)}{m_x(z)+m_{xu}(z)}.$$

This also holds when $n_{xy}(z)=0$. Now, it can be shown that there is a one-to-one correspondence between $\boldsymbol{\pi} = (\boldsymbol{\pi}_0,\ \boldsymbol{\pi}_1)$ and $(\boldsymbol{p}_0,\ \boldsymbol{q}_0,\ \boldsymbol{p}_1,\ \boldsymbol{q}_1)$, where $\boldsymbol{\pi}_z = (\pi_{yx,z} : y = 1, ..., 5\ ;\ x = 1, ..., 5)$, $\boldsymbol{p}_z = (p_{yx,z} : y = 1, ..., 5\ ;\ x = 1,\ ...,\ 5)$ and $\boldsymbol{q}_z = (q_{1,z}, ..., q_{5,z})$. Since $\pi_{yx,z} = p_{y,x,z}\cdot q_{yz}$, the mle of $p_{yx,z}$ and $q_{y,z}$ must satisfy

$$\hat{p}_{yx,z}\cdot\hat{q}_{y,z} = \frac{n_{xy}(z)}{m_x(z)+m_{xu}(z)} \qquad (4.10)$$

and are uniquely determined by $\hat{\pi}_{yx,z}$.

Consider $h_{xy}(z)$, given by (4.5) & (4.6). Let $h_y(z) = \sum_x h_{xy}(z)$ and $n_y(z) = \sum_x n_{xy}(z)$. From (4.7),

$$\frac{h_j(z)}{h_j(z)+n_j(z)} = \frac{h_{xj}(z)}{h_{xj}(z)+n_{xj}(z)}, \text{ if } n_{xj}(z)>0. \qquad (4.11)$$

From (4.10) and (4.11) we have that the following intuitive estimates also are mle.

$$\hat{q}_{y,z} = \frac{n_y(z)}{n_y(z)+h_y(z)} \qquad (4.12)$$

$$\hat{p}_{y,x,z} = \frac{n_{xy}(z)+h_{xy}(z)}{m_x(z)+m_{xu}(z)} \qquad (4.13)$$

(also when $n_{xy}(z)=h_{xy}(z)=0$).

(We can also show (4.12) and (4.13) by maximizing the loglikelihood directly.) Next, we show that the imputed values (4.2) for the model (4.9) equal $h_{xy}(z)$. From (4.2), we have $n^*_{xy}(z) = m_{xu}(z)\cdot\hat{P}(Y=y\,|\,x,z,r=0)$. Under model (4.9) and estimates (4.12) and (4.13), we find that

$$\hat{P}(Y=y\,|\,x,z,R=0) = \frac{\hat{P}(Y=y\,|\,x,z)-\hat{P}(Y=y,R=1\,|\,x,z)}{\hat{P}(R=0\,|\,x,z)}$$

$$= \frac{\hat{p}_{yx,z}-\hat{\pi}_{yx,z}}{1-\sum_y \hat{\pi}_{yx,z}}$$

$$= \frac{n_{xy}(z)+h_{xy}(z)-n_{xy}(z)}{m_{xu}(z)} = \frac{h_{xy}(z)}{m_{xu}(z)},$$

and it follows that $n^*_{xy}(z) = h_{xy}(z)$. If $n_{xy}(z)=0$, then $\hat{p}_{y,x,z} = \hat{\pi}_{yx,z} = 0$, and $n^*_{xy}(z)=0$. We note that model (4.9) is saturated and will, from (4.10), give perfect fit.

The imputation-based expansion estimates (4.6), with model (4.9), are identical to the modified Horvitz-Thompson estimates with $\hat{q}_{y,z} = n_y(z)/[n_y(z)+n^*_y(z)]$ (from (4.12)) as the estimated response probabilities, used in the official statistics from the 1992 CES. This follows from the fact that the modified Horvitz-Thompson estimator of N_y is given by

$$\hat{N}_{y,\mathrm{HT}} = \sum_{i\in s_r}\frac{I(Y_i=y)}{\pi_i},$$

where $\pi_i = P(\text{ person } i \text{ is selected to the sample and responds})$. Hence,

$$\pi_i = \frac{n}{N}\hat{P}(R_i=1\,|\,x_i,z_i,Y_i=y) = \frac{n}{N}\hat{q}_{y,z_i}$$

and

$$\hat{N}_{y,\mathrm{HT}} = \frac{N}{n}\left(\frac{n_y(0)}{\hat{q}_{y,0}}+\frac{n_y(1)}{\hat{q}_{y,1}}\right). \qquad (4.14)$$

Here,

$$\hat{N}_{y,\mathrm{HT}} = \frac{N}{n}\left(\frac{n_y(0)}{n_y(0)/(n_y(0)+n^*_y(0))}+\frac{n_y(1)}{n_y(1)/(n_y(1)+n^*_y(1))}\right)$$

$$= N\frac{n_y+n^*_y}{n}.$$

So this modified Horvitz-Thompson estimator suffers from the same negative feature as the imputation-based expansion estimator (4.6); it cannot correct for the bias in an unrepresentative sample. For a general description of the modified Horvitz-Thompson method see, *e.g.*, Särndal *et al.* (1992, chapter 15).

4.3 Variance Estimation

Variance estimation of the various estimates are obtained by bootstrapping. It can be carried out under the modeling or quasi-randomization framework (Little and Rubin 1987). For instance, to estimate the variance under model (3.1) and RM1 (3.2), we may apply the parametric bootstrap with the estimated parameters (Efron and Tibshirani 1993). However, it is not clear how to compare the variances estimated under the alternative models. We have therefore chosen to estimate the variances of the different estimators under a common quasi-randomization framework. We assume simple random sampling conditional to the family size, which is the only assumption we make for variance estimation. Unconditionally we have a self-weighting, but not simple random, sample, and therefore this is a rather crude approximation to the actual conditional sampling design. However, for a comparative study of the estimators the approximation will serve this purpose well. The nonresponse indicator r_i is considered to be a constant associated with person i. We draw the bootstrap sample, resampling $(y_i, z_i, r_i = 1)$, $(z_i, r_i = 0)$ randomly with replacement, as described by Shao and Sitter (1996, Section 5), within each post-stratum of $\{i; x_i = x\}$. While the sizes of the sample post-strata are fixed, both the number of nonrespondents and the number of persons from urban or rural areas vary from one bootstrap sample to another. We calculate the bootstrap estimates in the same way as based on the observed data. In particular, the bootstrap data are imputed in the same way as the original data if the estimator is imputation-based. Finally, the estimated variances and standard errors are obtained by the usual Monte Carlo approximation based on 500 independent bootstrap samples.

5. Estimated Number of Households of Different Sizes Based on the 1992 Norwegian Consumer Expenditure Survey

In this section we present the estimated number of households of sizes one to five and more, and the total number of households for the population in Norway aged less than eighty years old. The estimation uses the data from CES 1992, and is based on the estimators considered in Section 4. To compute the estimates we need the number of families of different sizes in the population, *i.e.*, M_x, at the time of the 1992 survey. The actual number at the time of the survey is not recorded. As an approximation we use the numbers at January 1, 1993. These are given in table 4.

Table 4
Families and Persons with Age Less than 80 Years in Norway at January, 1993

Number of persons in family	Families	Persons
1 person	793,869	793,869
2 persons	408,440	816,880
3 persons	261,527	784,581
4 persons	266,504	1,066,016
5 or more persons	127,653	670,528
Total	1,857,993	4,131,874

Note that the average family size for families with 5 or more persons is 670,528/127,653 = 5.25. We use 5.25 as an estimate of the average household size for households of size 5 or more, and divide by 5.25 instead of 5 in all estimates of H_5.

5.1 Maximum Likelihood Estimation and Poststratification

The estimated household distributions are presented in table 5. The estimates are based on the maximum likelihood (m.l.) estimator (4.1) using the population model with the restricted parametric link function $p_{y,x}$ in combination with the response models RM1(y, z) and RM2(y, z). To illustrate the effect of nonresponse modeling versus poststratification we also present the standard poststratified estimator (4.4). We recall that this is the maximum likelyhood estimator when ignoring the response mechanism. Furthermore, we present the estimated household size distribution based on the imputation-based poststratification (4.3) with the saturated model (4.9). For assessing the sampling variability of the different estimators, the estimated standard errors are also included.

The three models that take the response mechanism into account give higher total number of households. They also give considerable higher numbers of one-person-households. This seems sensible since we expect the one-person households to have the highest nonresponse rate. And thus, these estimates are most influenced by taking the response mechanism into account. We note that the restricted parametric link model (3.1) together with the logistic response model RM2(y, z) gives practically the same poststratified estimates as model (4.9), with also approximately the same standard errors. Because of the freedom of model (4.9), with perfect fit, it seems that model (3.1) & RM2(y, z) works well for estimating the number of households of different sizes. Regarding the uncertainty of the estimates, we see as one might expect that the standard errors typically seem to increase with the number of unknown parameters in the underlying model. Also, the total number of households is rather accurately estimated, not counting possible bias, while it's clearly most difficult to estimate the number of one-person households.

In order to evaluate the extent to which the differences between the estimates are due to sampling error or non-response bias, we consider the estimated standard errors of the differences of the point estimates. Some of these are given in table 6, using mostly the imputation-based post-stratification with the saturated model as a reference. For short, we use the terms Est1 – Est4 for the estimates defined as they appear in table 5:

Est1: M.l. estimator based on population model $p_{y,x}$ and response model RM1

Est2: M.l. estimator based on population model $p_{y,x}$ and response model RM2

Est3: Imputation-based poststratification based on the saturated model (4.9)

Est4: Poststratified estimator without imputation.

Based on tables 5 and 6 we can conclude that Est4 and Est3 have different expected values in estimating H_1, H_3, H_5 and H. Regarding the other comparisons, we see that in estimating H_3 there is a significant difference between Est1 and Est2/Est3, and note from earlier discussions in Section 3.3 that RM2 gives a better fit to the data than RM1.

The estimates based on the expansion estimator $\hat{H}_{y,e}$, given by (4.5), in 100's, are 390,500, 496,500, 283,900, 279,900, 148,000 and 1,598,800 with estimated standard errors equal to 33,100, 21,700, 14,600, 11,600, 6,100 and 23,700 for H_1, ..., H_5 and H, respectively. The standard errors for the differences between these estimates and the Est3-estimates are 52,800, 30,900, 19,100, 10,800, 5,400 and 32,000 for H_1, ..., H_5 and H respectively. These expansion estimates indicate serious bias due to non-response, especially the estimates for H_1, H_5 and H, with poststratification correcting for some of the bias (probably about 50% for the estimates of H_1 and H). We also note that the standard errors for the poststratified estimator and this simple expansion estimator are about the same. So by reducing the bias with poststratification one reduces the total error as well.

Poststratification corrects for the bias caused by the discrepancy between the family size distributions in the response sample and the population. From table 1 and table 4 we see that these family size distributions are given by (in percentages), for $x = 1, \ldots, 5$:

Response sample: 14.6 – 20.7 – 19.1 – 27.0 – 18.6
Population: 19.2 – 19.8 – 19.0 – 25.8 – 16.2.

Since the number of one-person families is much too low in the response sample, so will the expansion estimate of H_1 be. With post strata determined by family size, post-stratification corrects for the family size bias in the response sample, but does implicitly assume that nonrespondents and respondents have the same household size distribution, for a fixed family size. Or, in other words, the respondents are treated as a random subsample of sampled units with the same family size, as mentioned by Little (1993). This is most likely not the case. We recall that the family size variable was not significant when the household variable was included in the response models. Thus it seems reasonable to assume, as in our response models, that response rates will vary with the actual household sizes rather than the registered family sizes. Typically, estimates of the number of one-person households will be biased when the nonrespondents are ignored.

Table 5
Estimated Household Totals for Persons Aged Less than 80 Years in Norway at January 1, 1993, in Units of 100. In Parentheses, the Estimated Standard Error of the Estimates

	Maximum likelihood estimator with nonignorable response mechanism				**Imputation-based poststratification**		**Ignoring the response mechanism**	
Household size, y	Population model $p_{y,x}$ and response model RM1 (y, z)	%	Population model $p_{y,x}$ and response model RM2 (y, z)	%	Saturated population and response model	%	Poststratified estimator	%
1	558,800 (38,900)	32	595,400 (48,000)	34	596,600 (53,500)	34	486,000 (35,800)	29
2	520,200 (20,600)	30	525,800 (27,400)	30	523,600 (29,800)	30	507,800 (20,000)	30
3	278,900 (13,800)	16	249,100 (20,300)	14	250,000 (19,800)	14	286,200 (14,100)	17
4	258,900 (9,800)	15	269,000 (11,600)	15	268,900 (11,500)	15	270,600 (10,100)	16
≥ 5	125,800 (4,700)	7	126,000 (5,100)	7	126,200 (5,000)	7	131,300 (4,700)	8
Total	1,742,600 (25,600)	100	1,765,300 (29,700)	100	1,765,300 (31,900)	100	1,681,900 (23,300)	100

Table 6
Estimated Standard Errors of the Differences of the Point Estimates in Table 5

Household size	Est1−Est2	Est1−Est3	Est2−Est3	Est4−Est3
1	29,700	37,000	16,600	42,400
2	19,300	22,200	8,800	23,100
3	15,400	15,200	5,300	15,500
4	6,700	6,500	1,800	6,600
≥ 5	1,700	1,700	500	1,900
Total	15,300	18,800	8,900	23,300

After having corrected for nonresponse bias by completing the sample with imputed values, the sample itself may be skewed compared to the population. To illustrate the effect of poststratification to correct for this, we shall compare, using the saturated model (4.9), the imputation-based poststratified estimates Est3 with the imputation-based expansion estimates given by (4.6): 583,900, 567,700, 244,300, 259,300, 122,400 and 1,777,600 for H_1, ..., H_5 and H, respectively. As noted in Section 4.2, see (4.14), these estimates are identical to the modified Horvitz-Thompson estimates. The standard errors for these estimates are practically the same as for Est3. Hence, the alternative poststratified estimation methods based on nonignorable response models have standard errors at least no worse than the modified Horvitz-Thompson estimator. So if one reduces the bias with the alternative methods, one reduces the total error too. The standard errors of the differences between Est3 and this modified Horvitz-Thompson estimator in the estimates of H_1, ..., H_5 and H are 3,500, 2,200, 1,100, 600, 200 and 2,100 respectively. Clearly these two methods give significantly different estimates for all household size totals. In this comparison, one feature stands out. The expansion estimate of the number of two-persons households, 567,700, is clearly too high, as seen by comparing the family size distributions in the total sample and the population (in percentages), for $x = 1, \ldots, 5$:

Population: 19.2 – 19.8 – 19.0 – 25.8 – 16.2
Sample: 18.6 – 23.0 – 17.8 – 24.9 – 15.7.

The sample proportion of persons in two-persons families is much too high, and even though we have corrected for nonresponse bias, the expansion estimator, and then also the modified Horvitz-Thompson estimator cannot correct for a nonrepresentative sample. This will necessarily lead to biased estimates of H_2. We need poststratification to correct for a skewed sample. One can regard the difference in expected values for these estimators of H_2 as being close to the bias for the modified Horvitz-Thompson estimator, and note that an approximate 95% confidence interval for this difference is (39,800, 48,400).

For robustness considerations we also present the estimates from the cumulative logit model mentioned in Section 3.1 together with RM1 (y, z), which we know fits the data poorly. They are, in 100's: 591,800, 501,000, 265,200, 267,300, 128,200 and 1,753,500 for H_1, ..., H_5 and H, respectively. Compared to table 5, this seems to indicate that a reasonable model for response plays a more important role than a good population model. It is also evident that nonresponse modeling makes a difference, as seen when compared to poststratification and simple expansion.

5.2 Comparison with the Currently Used Estimates in CES, the Quality Survey for the 1990 Census and a Projection Study

Since 1993, an alternative, computationally simpler, modified Horvitz-Thompson estimator of type (4.14) has been in use in the production of official statistics from CES, see (Belsby 1995). We recall from Section 2 that the weights are the inverse sampling probabilities of the households, multiplied with the estimated probability of response. The response probabilities are estimated using a logistic model similar to RM2 (y, z) with place of residence and household size as explanatory variables. For the nonrespondents with unknown household size the registered family size is used instead, replacing (3.5). Thus, the weights may be regarded as an approximation to using (3.5). Of course, (3.5) is possible only when a population model is considered, which CES has not done. Table 7 presents estimated household distribution based on this CES-modified Horvitz-Thompson estimator.

The quality survey for the Census 1990, PES *1990*, contains 8,280 respondents and uses practically the same household definition as CES. The response rate was 95%. The H_y – estimates uses poststratification with respect to household size in the Census. However, no attempts were made to correct for possible nonresponse bias with respect to actual household size. PES deals with the whole population. Table 7 has the estimates for the 0 – 79 age group with the same poststratification method as in *PES*.

Table 7 also presents estimates based on the Household Projections study by Keilman and Brunborg (1995). This study simulates household structure for the period 1990 to 2020. The data sources are 28,384 individuals from the 1990 Population and Housing Census and 1988 Family and Occupation Survey. Keilman and Brunborg project for the whole population in 1992. We adjust their estimates to the 0 – 79 age group.

Table 7
Estimated Household Size Totals for Persons Less than 80 Years in Norway at January 1, 1993 with CES-modified Horvitz-Thompson, PES 1990 and Projections, in Units of 100

Household size	CES-Modified Horvitz-Thompson	%	PES 1990	%	Projections	%
1	622,900	35	626,000	35	668,300	37
2	518,500	29	494,200	28	549,000	30
3	259,900	15	291,500	16	211,900	12
4	258,500	15	250,000	14	221,500	12
≥ 5	124,600	7	115,300	6	97,500	5
Unknown					78,500	4
Total	1,784,400	1	1,777,000	99	1,826,700	100

Table 8
Estimated Probability of Response Based on the Method Used in CES Since 1993, in Percentages

	Household size				
Place of residence	1	2	3	4	5 or more
	CES-method				
Rural	44.53	66.24	74.55	73.54	80.07
Urban	36.01	57.90	67.25	66.09	73.80
	Model $p_{y,x}$ in (3.1) combined with RM2(y, z)				
Rural	47.77	60.90	79.05	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46

The estimates in table 7 support our impression that the estimates based on modeling the response mechanism leads to less biased estimates compared with ignoring the response mechanism as in mere poststratification or simple expansion. This is especially true for the one-person households and the total. The current "official estimator", the modified Horvitz-Thompson seems to give estimates of the right magnitude and in fact is closer to the results of PES 1990 than the modelbased estimates. However, this is more by accident. As a *method* it has some problems even in a representative sample. We can study this by estimating the response probabilities. Table 8 presents the results together with the estimates based on RM2 (y, z) & (3.1) from table 3.

Compared to the estimated response probabilities based on model RM2 (y, z) with (3.1), we see that replacing household size with family size in the nonresponse group is not a satisfactory approximation. Hence, if compared with the modified Horvitz-Thompson estimator in Section 5.1 based on the saturated model (4.9), the latter one would be preferred. For this particular survey, the CES approach overestimates the probability of response for household of size 2, which in a representative sample would lead to underestimating of H_2. The estimated response probabilities will most likely be biased when we are using family size in place of household size in the nonresponse group when estimating the parameters in the response model. This bias is an additional problem to the previously mentioned one, that the modified Horvitz-Thompson estimates will be similar to the imputation-based expansion estimates and cannot correct for nonrepresentative samples (as has been a problem in CES since 1993). In the 1992 CES, however, the sample is skewed with a too high proportion of families of size 2, and the H_2 – estimate will be of the right magnitude, by accident.

6. Conclusions

We have investigated modeling and methodological issues for estimating the total number of households of different sizes in Norway, based on the Norwegian Consumer Expenditure Survey (CES). The main issue is how to correct for bias due to nonignorable nonresponse. The existing estimation method in CES is a modified Horvitz-Thompson estimator that includes a correction for nonresponse by estimating response probabilities. We have considered basically two modelbased approaches, a maximum-likelihood estimator and imputation-based post-stratification after registered family size. With a population model that corresponds to a group model after family size only, these two estimators are identical. This family group model for household size and a logistic link for the response probability using household size as a categorical variable seem to work well for our estimation problem.

In analyzing the 1992 CES, we find serious bias due to nonresponse, especially the estimates for H_1 and H, with pure poststratification (without imputation) correcting for

some of the bias (probably about 50% for the estimates of H_1 and H). Poststratification does not, however, take into account possible nonresponse bias dependent on household size. Our response models assume that the response rates will vary with the actual household sizes rather than the registered the family sizes, and it is quite evident that such nonresponse modeling makes a difference, leading to less biased estimates than mere poststratification or simple expansion, especially of H_1 and H.

The modified Horvitz-Thompson estimates used in the official statistics from CES correspond to imputation-based expansion estimates. Hence, they cannot correct for nonrepresentative samples. The study in this paper shows that, in addition to a nonignorable response model it is also necessary to poststratify according to family size, *i.e.*, using a population model given family size. Hence poststratification, response modeling and imputation are key ingredients for a satisfactory approach.

In any estimation problem of totals in survey sampling, one must be aware of the fact that a Horvitz-Thompson estimator cannot correct for skewed samples, even when modified with good response estimates. Poststratification should always be considered as well as imputation based on a response model, nonignorable when needed.

Appendix A1

The data for rural and urban areas separately are given in table A1.

Appendix A2

Theorem. Assume model (3.1) for Y. *i.e.*, $P(Y=y\,|\,x,\ z)=p_{y,x}$ is independent of z, but otherwise the $p_{y,x}$'s are completely unknown with the only restriction being that $\sum_y p_{y,x}=1$, for all values of x, for all k. The response mechanism is arbitrarily parametrized, *i.e.*, no assumption is made about $P(R=1\,|\,Y=y,\ x,\ z)$. Then the maximum likelihood estimates for $p_{y,x}$ are given by

$$\hat{p}_{y,x}=\frac{n_{xy}+n^{*}_{xy}}{m_x+m_{xu}}.$$

Proof. Let $q_{yx,z}=P(R=1\,|\,Y=y,\ x,\ z)$. The log likelihood is given by

$$\begin{aligned}\ell &= \sum_x\sum_y n_{xy}p_{y,x}+\sum_{z=0}^{1}\sum_x\sum_y n_{xy}(z)q_{yx,z}\\ &\quad+\sum_{z=0}^{1}\sum_x m_{xu}(z)\log P(R=0\,|\,x,z)\\ &=\sum_x\sum_y n_{xy}p_{y,x}+\sum_{z=0}^{1}\sum_x\sum_y n_{xy}(z)q_{yx,z}\\ &\quad+\sum_{z=0}^{1}\sum_x m_{xu}(z)\log(1-\sum_{y=1}^{5}p_{y,x}q_{yx,z}).\end{aligned}$$

We use the Lagrange method and maximize $G=\ell+\sum_{x=1}^{5}\lambda_x(\sum_{y=1}^{5}p_{y,x}-1)$.

Let the solutions be $\hat{p}_{y,x}(\lambda_x)$, and determine the λ_x's such that $\sum_y\hat{p}_{y,x}(\lambda_x)=1$, for all x. No matter how the $q_{yx,z}$'s are parametrized, the mle $\hat{p}_{y,x}$ must satisfy, by solving the equations $\partial G/\partial p_{y,x}=0$,

$$\frac{n_{xy}}{\hat{p}_{y,x}}-\sum_{z=0}^{1}m_{xu}(z)\frac{\hat{q}_{yx,z}}{\hat{P}(R=0\,|\,x,z)}+\lambda_x=0 \qquad \text{(A1)}$$

which is equivalent to:

$$\begin{aligned}n_{xy}&=\hat{p}_{y,x}\sum_{z=0}^{1}\frac{m_{xu}(z)}{\hat{P}(R=0\,|\,x,z)}\\ &\quad-\sum_{z=0}^{1}m_{xu}(z)\frac{\hat{P}(R=0,Y=y\,|\,x,z)}{\hat{P}(R=0\,|\,x,z)}-\hat{p}_{y,x}\lambda_x.\end{aligned}$$

Table A1
Family and Household Sizes for the 1992 Norwegian Consumer Expenditure Survey, Split into Rural and Urban Areas. The Upper Entry is for the Urban Group

		Household size								
Family size		1	2	3	4	≥ 5	Total response	Non-response	Total	Response rate
1	urban	28	24	7	2	0	61	78	139	0.439
	rural	55	24	13	7	2	101	75	176	0.574
2	urban	6	70	12	3	0	91	84	175	0.520
	rural	3	107	25	1	3	139	76	215	0.647
3	urban	4	8	57	11	3	83	40	123	0.675
	rural	6	17	74	29	3	129	51	180	0.717
4	urban	0	3	15	80	5	103	43	146	0.705
	rural	2	10	22	151	12	197	80	277	0.711
≥ 5	urban	0	1	0	6	66	73	28	101	0.723
	rural	1	3	4	11	115	134	32	166	0.807
Total urban		38	106	91	102	74	411	273	684	0.601
Total rural		67	161	138	199	135	700	314	1014	0.690

We determine λ_x by summing over y:

$$m_x = \sum_{z=0}^{1} \frac{m_{xu}(z)}{\hat{P}(R=0\,|\,x,z)} - \sum_{z=0}^{1} m_{xu}(z)\frac{\hat{P}(R=0\,|\,x,z)}{\hat{P}(R=0\,|\,x,z)} - \lambda_x,$$

hence

$$\lambda_x = \sum_{z=0}^{1} \frac{m_{xu}(z)}{\hat{P}(R=0\,|\,x,z)} - (m_x + m_{xu}).$$

It follows from (A1) that $\hat{p}_{y,x}$ satisfies the following relation:

$$\hat{p}_{y,x} = \frac{n_{xy}}{\left(m_x + m_{xu} - \sum_{z=0}^{1} m_{xu}(z)\frac{\hat{P}(R=0\,|\,Y=y,x,z)}{\hat{P}(R=0\,|\,x,z)}\right)}. \quad \text{(A2)}$$

The imputed values are given by , from (4.2),

$$n^*_{xy}(z) = m_{xu}(z)\,\hat{p}_{y,x}\frac{\hat{P}(R=0\,|\,Y=y,x,z)}{\hat{P}(R=0\,|\,x,z)}$$

and, from (A2),

$$\hat{p}_{y,x} = n_{xy}\Big/\left(m_x + m_{xu} - \sum_{z=0}^{1}\frac{n^*_{xy}(z)}{\hat{p}_{y,x}}\right) = n_{xy}\Big/\left(m_x + m_{xu} - \frac{n^*_{xy}}{\hat{p}_{y,x}}\right)$$

or equivalently,

$$\hat{p}_{y,x}(m_x + m_{xu}) - n^*_{xy} = n_{xy},$$

$$\textit{i.e.,}\ \hat{p}_{y,x} = \frac{n_{xy} + n^*_{xy}}{m_x + m_{xu}}. \quad \text{Q.E.D}$$

Appendix A3

Table A2

The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban. The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group. Based on Model (3.1) and RM1(y, z)

		Household size					
Family size		1	2	3	4	≥ 5	Total
1	urban	77.8	44.1	12.9	3.9	0.3	139
	rural	103.6	43.1	18.4	8.7	2.3	176
2	urban	10.8	137.9	22.1	3.8	0.4	175
	rural	7.5	168.6	33.9	1.7	3.3	215
3	urban	7.5	14.3	81.3	16.4	3.6	123
	rural	10.7	25.3	104.8	35.6	3.7	180
4	urban	0.8	6.4	21.9	110.3	6.6	146
	rural	3.5	16.7	35.1	206.9	14.8	277
≥ 5	urban	0.5	2.4	1.0	9.0	88.2	101
	rural	1.6	4.7	5.2	14.4	140.1	166
Total /urban		97.4	205.1	139.2	143.4	99.1	684
rural		126.9	258.4	197.4	267.3	164.2	1,014

Table A3

The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban. The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group. Based on Model (3.1) and RM2 (y, z)

		Household size					
Family size		1	2	3	4	≥ 5	Total
1	urban	81.6	42.7	10.4	4.0	0.3	139
	rural	107.5	41.5	15.9	8.8	2.3	176
2	urban	11.9	140.4	18.3	3.9	0.5	175
	rural	8.6	170.9	30.3	1.8	3.4	215
3	urban	9.4	16.1	75.2	18.6	3.7	123
	rural	13.4	27.7	96.5	38.5	3.9	180
4	urban	0.8	6.2	18.9	113.5	6.6	146
	rural	3.7	16.2	29.2	213.1	14.8	277
≥ 5	urban	0.5	2.3	0.6	9.3	88.3	101
	rural	1.7	4.6	4.6	14.9	140.2	166
Total /urban		104.2	207.7	123.4	149.3	99.4	684
rural		134.9	260.9	176.5	277.1	164.6	1,014

Appendix A4

Table A4
The Completed Sample Including the Imputed Values, Split Into Two Groups, Rural and Urban.
The Upper Entry is for the Urban Group and the Lower Entry is for the Rural Group.
Based on Model (4.9), *i.e*, Imputations Determined by (4.7) and (4.8)

	Household size					
Family size	1	2	3	4	≥ 5	Total
1 urban	79.6	47.2	9.4	2.8	0.0	139
rural	108.3	38.5	16.9	9.9	2.4	176
2 urban	17.1	137.7	16.0	4.2	0.0	175
rural	5.9	171.6	32.5	1.4	3.6	215
3 urban	11.4	15.7	76.2	15.6	4.1	123
rural	11.8	27.3	96.2	41.1	3.6	180
4 urban	0.0	5.9	20.0	113.2	6.9	146
rural	3.9	16.0	28.6	214.0	14.5	277
≥ 5 urban	0.0	2.0	0.0	8.5	90.5	101
rural	2.0	4.8	5.2	15.6	138.4	166
Total /urban	108.1	208.5	121.6	144.3	101.5	684
rural	131.9	258.2	179.4	282.0	162.5	1,014

Table A5
The Total Numbers of Family and Household Sizes for Imputed Complete Sample. Based on Model (4.9)

	Household size					
Family size	1	2	3	4	≥ 5	Total
1	187.9	85.7	26.3	12.7	2.4	315
2	23.0	309.2	48.6	5.7	3.6	390
3	23.2	43.0	172.4	56.7	7.7	303
4	3.9	21.9	48.7	327.2	21.3	423
≥ 5	2.0	6.8	5.2	24.1	229.0	267
Total	240.0	466.6	301.1	426.3	264.0	1,698

References

Baker, S.G., and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

Belsby, L. (1995). Forbruksundersøkelsen. Vektmetoder, frafallskorrigering og intervjuer-effekt. (The consumer survey. Weight methods, nonresponse correction and interviewer effect), Notater 95/18 Statistics Norway.

Bjørnstad, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.

Bjørnstad, J.F., and Skjold, F. (1992). Interval estimation in the presence of nonresponse. *The American Statistical Association 1992 Proceedings of the Section on Survey Research Methods*. 233-238.

Bjørnstad, J.F., and Walsøe, H.K. (1991). Predictive likelihood in nonresponse problems. *The American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*, 152-156.

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Forster, J.J., and Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse (with discussion). *Journal of the Royal Statistical Society B*, 60, 57-70.

Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

Hall, B.H., Cummins, C. and Schnake, R. (1991). *TSP Reference Manual, Version 4.2A*, Palo Alto California: TSP International.

Holt, D., and Smith, T.M.F. (1979). Post-stratification, *Journal of the Royal Statistical Society A,* 142, 33-46.

Keilman, N., and Brunborg, H. (1995). *Household Projections for Norway, 1990-2020, Part 1: Macrosimulation*, Reports 95/21, Statistics Norway.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

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

McCullagh, P., and Nelder, J.A. (1991). *Generalized Linear Models*, *2nd ed*. London: Chapman & Hall.

Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Statistics Norway (1990). *Survey of Consumer Expenditure 1986-88.* Official Statistics of Norway NOS B919.

Statistics Norway (1996). *Survey of Consumer Expenditure 1992-1994*. Official Statistics of Norway NOS C317.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.