Discussion Papers No. 206, Desember 1997
Statistics Norway, Research Department

Liv Belsby and Jan F. Bjørnstad

Modeling and Estimation Methods for Household Size in the Presence of Nonresponse

Applied to The Norwegian Consumer Expenditure Survey

Abstract:
This paper considers the problem of estimating 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 while taking into account possible nonresponse biases by modeling the response mechanism conditional on household size. Various models are considered together with regression estimation and imputation-based poststratification. Comparisons are made with the estimation methods used in official statistics for The Norwegian Consumer Expenditure Survey. We conclude that poststratification, response modeling and imputation should be used instead of the current method in official statistics, a modified Horvitz-Thompson method.

Keywords: Household size, nonresponse, imputation, poststratification

JEL classification: C42, C13

Address: 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

Discussion Papers comprises research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborated than a usual article by including intermediate calculation and background material etc.

Abstracts with downloadable postscript files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
P.O. Box 8131 Dep
N-0033 Oslo

Telephone: +47 22 00 44 80
Telefax: +47 22 86 49 76
E-mail: Salg-abonnement@ssb.no

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. The primary aim is to estimate the number of households of various sizes and the total number of households, trying to take into account the fact that nonresponse may produce biased estimates. Estimating household-size totals is an important issue in social planning. It is a difficult problem, as indicated by the results in the 1990 Norwegian Population and Housing Census where the number of one-person households was underestimated by about 100 000, as shown by The Post Enumeration Survey (see Schjalm 1996). Our application is based on the data from the 1992 *CES*. *CES* is a yearly survey and until 1992 used a form of poststratified estimation method. This was not satisfactory 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). Also this estimation method has shown weaknesses in *CES.*

We shall instead 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 and thus improve the estimation method currently in use. 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) and Bjørnstad and Skjold (1992).

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

We have not computed estimated standard errors of the estimates under the different models. The main issue here is a comparison of models with estimation bias as the basic problem. Moreover, the regression estimation and imputation-based poststratification turn out to be identically the same for some of the models that we end up recommending, However, standard errors of the estimates should, of course, be computed in the publication of the estimates in *CES.*

Section 2 describes the data-structure of *CES*, and Section 3 considers modeling issues. Section 3.1 presents the various models to be considered for the 1992 *CES*, and Section 3.2 describes the maximum likelihood method for parameter estimation. Section 3.3 evaluates the various models for household size and response. We consider two different types of models for the household-size distribution. They give different estimates for this distribution. However, the two models lead to similar estimates for the probabilities of nonresponse. This indicates that the model choice for the size-distribution does not seem to influence the estimated nonresponse probabilities strongly. A family size group model for household size and a logistic link for the response probability using household size as a categorical variable gives the best fit of the models under consideration.

Section 4 considers model-based estimation, the imputation method and imputation-based estimators. It is shown that for the chosen model for household size from Section 3.3, the regression estimator and 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 the key ingredients for a satisfactory approach.

2. Norwegian Consumer Expenditure Survey

The main purpose of the Norwegian Consumer Expenditure Surveys (*CES*) is to estimate the average consumer expenditure in private households. In this regard it is important to gain information about the composition of households, which is our goal in this paper. Hence, the variable of interest is the size of the *household* which is observed only in the response sample. A household is defined as persons having a common dwelling and sharing at least one meal each day (having common board). Persons who are temporarily absent due to school, vacation, etc., are included. Servants living in and lodgers receiving board are included in the household, while lodgers not receiving board are considered to be a separate household. For a complete description of *CES* we refer to Statistics Norway (1996). In *CES*, the auxiliary variables known for the total sample, including the non-respondents, 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. 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.

Every person in the sample population (persons in Norway between the ages of 16 and 80) has the same inclusion probability to the total sample. The whole household where the selected person lives is included in the survey for expenditure variables. It follows that the probability for a household to be included in the survey is approximately proportional to the household size. For our purpose (of estimating household size-distribution and the total number of households) we only consider those persons actually selected by the survey design, not the entire corresponding households.

The population of interest consists of all persons less than 80 years old including those under 16 years of age, since we are interested in the sizes of households. Let N be the total number of persons in this population. We let N_y denote the number of persons living in households of size y, and H_y the number of households of size y, $y = 1, \ldots, J$. 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. The total number of households is denoted by H, $H = \Sigma_y H_y$. Then, $N_y = yH_y$ for $j = 1, \ldots, J\text{-}1$ and $N_J \approx JH_J$. In our application we choose $J = 5$ due to the low frequency in the sample of households size greater than five.

Table 1 shows the data for the 1992 *CES*. We base our modeling and estimation on two corresponding tables, one for the households in rural areas and one for the households in urban areas. These data are given in appendix A. Initially we also split these households in two groups, one for the households interviewed during summer and another group the rest of the year. But the time of the survey did not turn out to be significant in any of the models. Thus, appendix A has the data for our analysis.

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

Family size	Household size					Total	Non-response
	1	2	3	4	≥ 5		
1	83	48	20	9	2	162	153
2	9	177	37	4	3	230	160
3	10	25	131	40	6	212	91
4	2	13	37	231	17	300	123
≥ 5	1	4	4	17	181	207	60
Total	105	267	229	301	209	1111	587

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

3. Modeling of household size and nonresponse

Let the variable Y denote household size, and let Y_i denote household size for person i in the population , for $i = 1, \ldots, N$. The main statistical problem is to estimate $H_1, \ldots, H_J$ and H. In terms of the variable Y, $H_y = \sum_{i=1}^{N} I(Y_i = y) / y$ where the indicator function $I(Y_i = y) = 1$ if $Y_i = y$, and 0 otherwise.

We shall assume a population model for Y, given auxiliary variables $\boldsymbol{x}$, i.e., we model the conditional probability $P(Y=y \mid \boldsymbol{x})$. Let nonresponse be indicated by the variable R, where $R=$ 1/0 according to response/nonresponse. To take nonresponse into account in the statistical analysis, we must model the response mechanism, i.e., the distribution of R conditional on y and $\boldsymbol{x}$. The sampling mechanism is ignorable for the survey we consider, i.e., is independent of the population vector $\boldsymbol{y}$ of household sizes. The statistical analysis is therefore done conditional on the total sample s, following the likelihood principle, see (Bjørnstad, 1996).

For *CES*, the auxiliary vector $\boldsymbol{x}^t = (x_1, x_2, x_3)$ where x_1 is family size, x_2 is place of residence with $x_2 =$ 0 if rural area and 1 if urban area, and x_3 is time of survey. The notation $\boldsymbol{x}^t$ symbolizes that the vector is transposed. Let $\boldsymbol{x}_i$ denote the values of the auxiliary variables for person i.

3.1. The Models

We shall consider various models for Y given $\boldsymbol{x}$ and for R given y and $\boldsymbol{x}$. If units i,j belong to the same household then $Y_i = Y_j$. If i,j belong to different households Y_i and Y_j are assumed to be independent. Let us first consider the simple model, where the household size is assumed to depend only on the family size x_{1i} with no additional assumptions, expressed

(3.1) $$P(Y_i = y|\boldsymbol{x_i}) = P(Y_i = y|x_{1i}) = p_{y,x_{1i}} ,$$

where $\sum_y p_{y,x_{1i}} = 1$, *for each possible value of* x_{1i} .

The model (3.1) is flexible in the sense that it does not include any restrictions on the assumed model function of x_{1i}. We will later refer to this model as the "free" model. The drawback is the high number of parameters compared with a model using a link function. If nonresponse is ignored, the estimates in this model would simply be the observed rates.

The household size takes values on an ordinal scale. Thus a natural choice for a model is the cumulative logit model, known as the proportional-odds model, see (McCullagh and Nelder, 1991). We shall denote it by CLM($\boldsymbol{x}$),

(3.2) $$\text{CLM}(\boldsymbol{x}):\ P(Y_i \le y|\boldsymbol{x}) = \begin{cases} \dfrac{1}{1+\exp(-\theta_y + \beta^t \boldsymbol{x})} & \text{for } y = 1,2,3,4 \\ 1 & \text{for } y \ge 5. \end{cases}$$

The parameters θ_y take values increasing in y, and are represented as $\theta_1 = \theta_1, \theta_y = \theta_{y-1} + k_{y-1}$, for $y = 2,3,4$. The vector β consists of the parameters that measure the influence of the auxiliary variables. The auxiliary vector here is $\boldsymbol{x} = (x_1,x_2)^t$ (x_1= family size, and x_2 = place of residence). In Section 3.3 we discuss the estimates and the validity of the assumed link function.

The data are incomplete due to nonresponse. 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. Nonignorable response mechanism is equivalent to

$$P(Y_i = y_i \,|\, \boldsymbol{x_i}, r_i = 0) \ \neq \ P(Y_i = y_i \,|\, \boldsymbol{x_i}, r_i = 1)$$

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

Thus estimating the parameters in the model for $P(Y=y \mid \boldsymbol{x})$ 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 $\boldsymbol{x}$. For example, the observ-ed 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 $\boldsymbol{z}_i$, which includes part of $\boldsymbol{x}_i$, expressed by

(3.3) $$\text{RM1}(y,\boldsymbol{z}):\ P(R_i = 1 \mid y_i, \boldsymbol{z}_i) = \frac{1}{1+\exp(-\alpha - \gamma y_i - \psi^t \boldsymbol{z}_i)} \quad .$$

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 (3.3) as an ordinal variable rather than a class variable. Thus the logit function,

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

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.,

RM2(y,$\boldsymbol{z}$):

(3.4) $$P(R_i = 1 \mid y_i, \boldsymbol{z}_i) = \frac{1}{1+\exp(-\alpha_0 - \alpha_1 I_1(y_i) - \alpha_2 I_2(y_i) - \alpha_3 I_3(y_i) - \alpha_4 I_4(y_i) - \psi^t \boldsymbol{z}_i)} \quad ,$$

where the indicator variable $I_y(y_j)$ equals 1 if $y_i = y$ and 0 otherwise. The drawback with this model is that it includes three parameters more than model (3.3).

3.2. Maximum likelihood parameter estimation

All the selected persons in the sample *s* are from different households (duplicates have been removed), such that all Y_i, $i \in s$ are assumed to be independent.

We 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 (3.3). 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 (3.4) replacing (3.3).

For the respondents let $L_i = P(Y_i = y_i \cap R_i = 1|\boldsymbol{x_i})$. Then, for model (3.1)

(3.5) $$L_i = \frac{1}{1+\exp(-\alpha-\gamma y_i-\psi^t \boldsymbol{z_i})} \cdot p_{y_i,x_{1i}} \ , \ i=1,\ldots,n_r \, .$$

For the nonrespondents let $L_i = P(R_i = 0 \mid \boldsymbol{x_i})$. Then

(3.6) $$L_i = \sum_{y=1}^{5} \frac{1}{1+exp(\alpha+\gamma y \ +\psi^t \boldsymbol{z_i})} \cdot p_{y,x_{1i}} \ , \ i=n_r+1,\ldots,n \, .$$

With model (3.2) instead, the expression for the respondents is given by

(3.7) $$L_i = \frac{1}{1+\exp(-\alpha-\gamma y_i-\psi^t \boldsymbol{z_i})} \cdot (P(Y_i \le y_i|\boldsymbol{x_i}) - P(Y_i \le y_i-1|\boldsymbol{x_i})) \ , \ i=1,\ldots,n_r \, ,$$

where $P(Y_i \le y_i|\boldsymbol{x_i})$ is given by (3.2). The expression for the nonrespondents equals

(3.8) $$L_i = \sum_{y=1}^{5} \frac{1}{1+exp(\alpha+\gamma y \ +\psi^t \boldsymbol{z_i})} \cdot (P(Y_i \le y|\boldsymbol{x_i}) - P(Y_i \le y-1|\boldsymbol{x_i})) \ , \ i=n_r+1,\ldots,n \, .$$

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

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

For $i = 1,..., n_r$, L_i is according to (3.5) or (3.7), and for $i = n_r + 1, \ldots, n$, L_i is given by (3.6) or (3.8).

Estimates are found by maximizing the likelihood function (3.9). The maximization was done numerically using the software TSP(1991). The optimizing algorithm is a standard gradient method, using the analytical first and second derivatives. These derivatives 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 estimated models together with the Pearson goodness-of- fit statistics. The model assumptions are illuminated with plots and we discuss the assumed linear logit function in the models (3.2) -(3.4). The estimates are 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 $\boldsymbol{z} = x_2$, place of residence. 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.

For models with no latent structure it would have been standard procedure to plot the data in order to illuminate our model assumptions. Consider the cumulative logit model. This model is based on the restriction that the cumulative logit, i.e., the function $\log\{ P(Y \le y \mid x_1)/(1- P(Y \le y \mid x_1))$ is approximately linear with respect to family size. The logit functions for household sizes 1,...,4 should also be approximately parallel. *CES* does not have callbacks. Neither does the survey include any

quality control survey. Hence there are no data available to illuminate the model assumption for the nonrespondents, making it difficult to investigate the assumption. What we can do is to plot the logit function for the estimated $P(Y \leq y \mid x_1)$ based on the free model (3.1) together with RM2(y,x_2). This estimate is not restricted by an assumed function for the household size, and the estimate is adjusted according to a rather flexible model for the response mechanism.

Figure 1. The estimated cumulative logit function, $\log\{\hat{P}(Y \leq y \mid x_1) / (1 - \hat{P}(Y \leq y \mid x_1))\}$**, for household sizes 1,.., 4 with respect to family size x_1. The estimates are based on the free model (3.1) in combination with the response model RM2(y,x_2)**

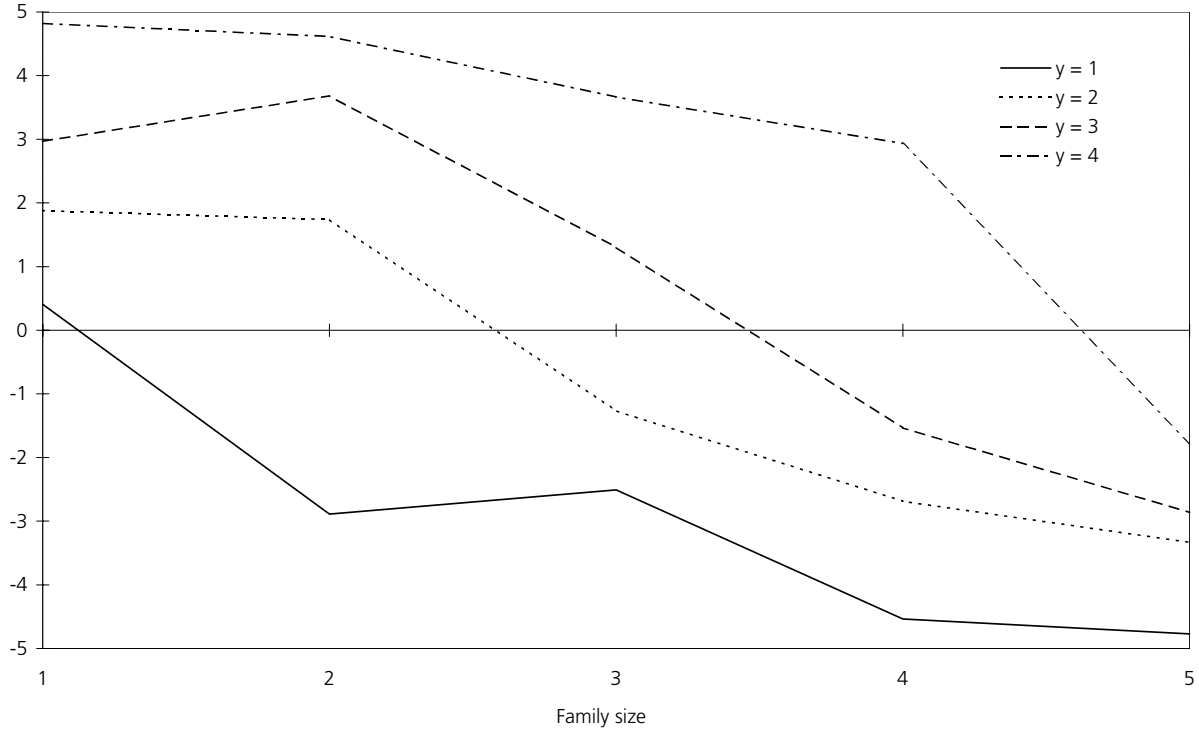
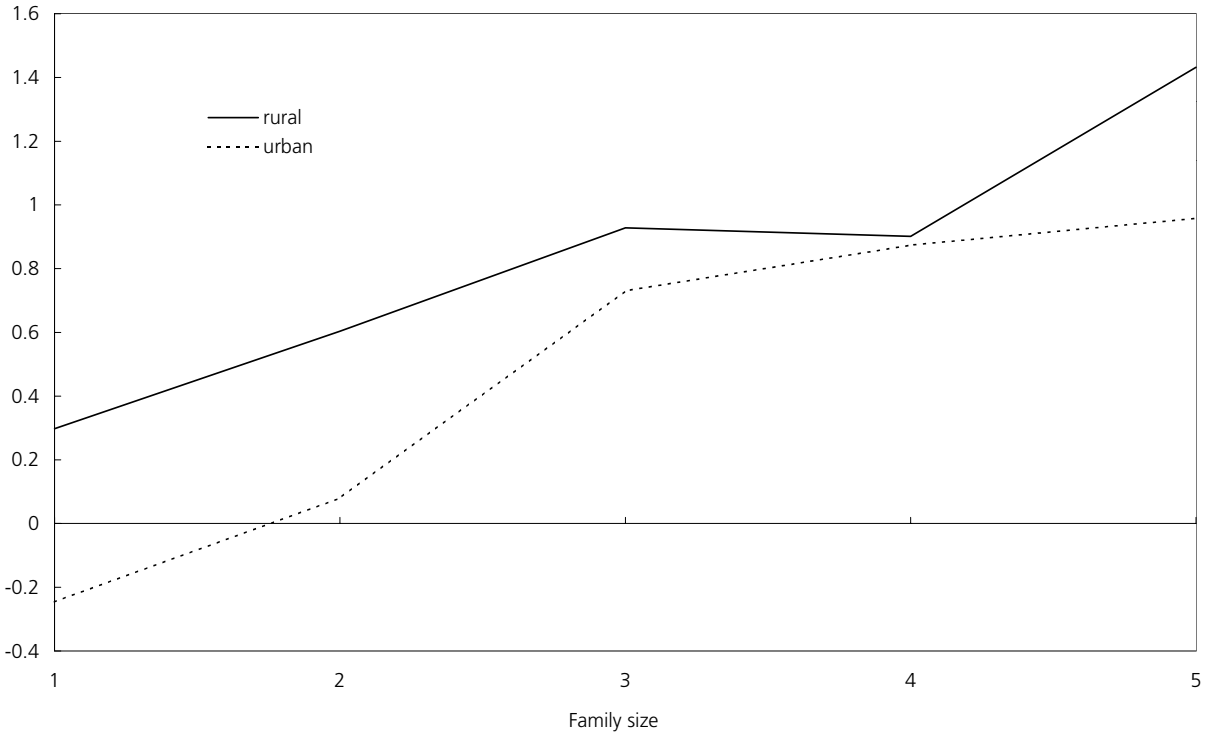


Figure 1 shows that the logit functions for the household sizes are clearly not parallel nor linear, which they should have been, at least approximately, in order to fulfil the model assumptions. Moreover, a goodness of fit test shows that the cumulative logit model fit the data badly. Therefore we choose to reject this model.

We also want to investigate the empirical logit function for response as a function of household size. However, the household size is unavailable for the nonrespondents. Instead we plot the logit-function against family size, see Figure 2. From family size one to two, the rural and urban functions increase fairly parallel. However, for family size three and four the logit functions depart from being linear and parallel. Thus we suspect that coding 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.

Figure 2 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 both the respondents and the nonrespondents from Table A1 in Appendix A



We now proceed with the free model in (3.1) combined with the response models RM1(y,x_2) and RM2(y,x_2), respectively. In order to test the goodness of fit of the models, we consider the Pearson chi-square statistics, conditional on the auxiliary variables family size, x_1 , and place of residence, x_2,. 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,x_2), d.f. = 52 -10-(20+3)=19, and for model (3.1)& RM2(y,x_2), d.f. = 52 -10-(20+6)=16.

The estimates for free model in (3.1) and response model RM1 are displayed in table 2.

Table 2. 1992 CES. Parameter estimates for p_{y,x_1}, the free model (3.1), in percentages combined with the response model RM1(y,x2) in (3.3). In parentheses, the estimates based on the free model (3.1) for respondents only, i.e., ignoring the response mechanism

	Household size				
Family size, x_1	1	2	3	4	5 or more
1	57.61 (51.23)	27.68 (29.63)	9.93 (12.35)	3.98 (5.56)	.80 (1.23)
2	4.71 (3.91)	78.58 (76.96)	14.36 (16.09)	1.39 (1.74)	.96 (1.30)
3	5.97 (4.72)	13.04 (11.79)	61.40 (61.79)	17.15 (18.87)	2.40 (2.83)
4	1.01 (0.67)	5.47 (4.33)	13.46 (12.33)	74.99 (77.00)	5.07 (5.67)
5 or more	0.80 (0.48)	2.66 (1.93)	2.31 (1.93)	8.75 (8.21)	85.48 (87.44)

The Pearson chi-square statistic equals 26.35. With d.f. equal to 19, this corresponds to the p-value 0.121. By studying the standardized residuals, (*observed-expected*)/$\sqrt{\hat{V}ar(observed)}$, it is seen that the model has the best fit in the rural area where the response rate is highest. The model does not quite manage to fit the data for the urban areas, especially predicting the number of respondents of households of size 3 among persons with family size 3. Here the model predicts 46 while the observed count is 57, out of the 123 persons in the urban area with family size 3.

We interpret now some of the values in the household model. The probability that a household size equals one, given that the family size is one, is 0.576. The estimates based on the traditional approach, ignoring the nonresponse is 0.512. The response model adjusts the observed rate among the respondents to a higher value. This seems reasonable since the rate of nonrespondents is higher for small households. The estimated probability of household size five or more, given family size of five or more is 0.855 which differs little from the observed rate among the respondents, 0.874. This indicates that, given family size five or more, the household size distributions are about the same among respondents and nonrespondents.

Table 3 displays the estimates for the model (3.1) together with the response model RM2 in (3.4). That means that the difference is that the estimates in table 3 are not restricted by a linear logit in household size for the response model.

Table 3. 1992 CES. Parameter estimates for the free model p_{y,x_1} in (3.1), in percentages, combined with the response model RM2(y,x_2) in (3.4). In parentheses are the estimates based on the free model p_{y,x_1} in (3.1) for respondents only, i.e., ignoring the response mechanism.

	Household size				
Family size, x_1	1	2	3	4	5 or more
1	60.01 (51.23)	26.75 (29.63)	8.35 (12.35)	4.09 (5.56)	.80 (1.23)
2	5.27 (3.91)	79.80 (76.98)	12.48 (16.09)	1.47 (1.74)	.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	.84 (0.48)	2.60 (1.93)	1.96 (1.93)	9.05 (8.21)	85.55 (87.44)

The Pearson chi-square statistic is 21.77, which with d.f. equal to 16 gives a p-value of 0.151. This p-value indicates that coding the household size as a categorial variable, as in RM2, improves the fit compared to using it as an ordinal variable. Consequently, the data indicate that the free model in (3.1) combined with RM2 is the best of the models we have considered so far. By studying the standardized residuals it is seen that the main reason for the better fit is that the model in table 3 does a better job of predicting the observed counts for the urban area.

The differences between the estimates outside and within the parentheses indicate the effect of the response mechanism on the estimates. As in the previous model, the estimates for the rates of one-person households display the largest differences. This reflects that one-person households have the lowest response probability.

Comparing the estimates in table 2 and 3, we find that the estimated household size model is only slightly altered when using response model RM2 in the place of model RM1. Hence, the estimates give the impression that these two different response models do not give very different estimates for the household size distribution. However, as we shall see in section 5, the estimates of the household size totals are somewhat differently distributed over the five sizes for the two response models.

3.4. Estimated probabilities of response

The model choice may influence the estimates for response. To illuminate this, table 4 displays the response probabilities for the models we have considered so far.

Table 4. Estimated probability of response in percentages, for the models (3.1)& RM1(y,x_2), (3.1)& RM2(y,x_2), CLM(x_1)&RM1(y,x_2) and CLM(x_1,x_1^2)& RM1(y,x_2)

Place of residence	Household size				
	1	2	3	4	5 or more
Model p_{y,x_1} in (3.1) combined with RM1(y,x_2)					
Rural	52.29	61.11	69.26	76.36	82.25
Urban	43.58	52.55	61.36	69.49	76.55
Model p_{y,x_1} in (3.1) combined with RM2(y,x_2)					
Rural	47.77	60.90	79.05	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46
The CLM(x_1) model combined with RM1(y,x_2)					
Rural	52.86	61.33	69.16	76.03	81.77
Urban	44.19	52.83	61.30	69.13	76.00
The CLM(x_1,x_1^2) model combined with RM1(y,x_2)					
Rural	53.10	61.42	69.13	75.90	81.58
Urban	44.39	52.89	61.22	68.95	75.74

For all the models in table 4 the probability of response is higher in rural than in urban areas, and one-person households have the lowest response probability. We also see that the response probabilities based on RM1, but on different models for the household distribution, are almost the same. However, replacing RM1 with RM2 influences the estimates. For example, the estimated response probability for household size one in rural areas decreases from 0.523 to 0.478 . The largest discrepancy is found in the probability of response for household size 3. This is possible because RM2 let the response probability vary freely as a function of household size, while RM1 assumes a linear logit in household size. The "free" model RM2 estimates, surprisingly maybe, that $P(R=1|\, y = 3) > P(R=1|\, y = 4)$. However, we see from table 2 and 3 that more of families with sizes 1,2 and 3 belong to household size 3 using model RM1 than model RM2. Furthermore, these categories have low response probability, and hence we get lower response probability for household size 3. For household size 4 it is the other way around. We recall from earlier discussions that model (3.1) with RM1 has trouble estimating the number of households of size 3. These considerations indicate that the estimates based on RM2 is more reliable, and that the linear logit in RM1 is too restrictive.

We also present estimated response probabilities based on an imputation method, investigated in a later section. The model, defined by (4.10), implicitly 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 the place of residence and the family

size, but with no restriction on the link function. This model is saturated, and will from (4.11) give perfect fit. We note that RM1(y,x_2) and RM2(y,x_2) both satisfy (4.10b), but are more restrictive. Model (4.10) allows more freedom, of course, than model (3.1) with RM1(y,x_2) or RM2(y,x_2). The estimated response probabilities based on model (4.10) are displayed in table 5.

Table 5. Estimated probability of response based on the saturated model (4.10) in percentages

	Household size				
Place of residence	1	2	3	4	5 or more
Rural	50.79	62.37	76.90	70.57	83.07
Urban	35.17	50.85	74.79	70.68	72.89

We see the same tendency as for RM2 with (3.1) in table 4; the probability of response is higher for household size 3 than for households of size 4. It is seen from table 4 that RM2 & (3.1) acts as a smoother of the estimates in table 5, because of the added assumption of parallel logits of the response probabilities for urban and rural areas. Since the estimates in table 5 are based on a saturated model, this reinforces the previous conclusion that estimates based on RM2 are more reliable than the ones based on RM1, and that the linear logit in RM1 is too restrictive.

4. Estimators for household size totals

In this section we present the estimators for household size totals for the population consisting of the people less than 80 years by 1.1.93. The estimators fall into two groups. One group of estimators is based directly on the maximum likelihood parameter estimators while the other group is imputation-based. For comparison we also include the standard poststratification estimator. For the free population model (3.1), some of these estimators turn out to be identical.

4.1. Maximum likelihood regression estimation and poststratification

Section 3 evaluated some models for both household size and response probability. For the household size we decided to use the free model p_{y,x_1} in (3.1), where x_1 is the family size. For the response probability we use both logistic models, RM1 and RM2, with place of living and household size as auxiliary variables. Recall that household size is coded as an ordinal variable in RM1, and as a categorical variable in RM2. The parameters in the models are estimated by maximizing the likelihood, as described in Section 3.2.

The data are assumed to be of the form presented in table 1, given registered family size (also for rural/urban areas separately).

Table 6. Family and household sizes with nonresponse. Number of persons

	Household size					
Family size	1	2	. . . . .	$\geq J$	Total	Nonresponse
1	n_{11}	n_{12}	. . . . .	n_{1J}	m_1	m_{1u}
2	n_{21}	n_{22}	. . . . .	n_{2J}	m_2	m_{2u}
:	:	:	. . . . .	:	:	:
$\geq K$	n_{K1}	n_{K2}	. . . . .	n_{KJ}	m_K	m_{Ku}
Total	n_1	n_2	. . . . .	n_J	n_r	n_u

Here, n_{ky} is the number of respondents belonging to a family of size $x_2 = k$ and household of size y. Furthermore, m_k (m_K) is the number of persons in the response sample belonging to families of size k ($\geq K$), $m_k = \sum_{y=1}^{J} n_{ky}$, and n_y (n_J) is the number of respondents belonging to households of size y($\geq J$), $n_y = \sum_{k=1}^{K} n_{ky}$. In our application we choose $K = 5$. The total size of the response sample is denoted by n_r . The total number of nonrespondents is n_u , and m_{ku} is the number of missing observations for persons with family size k. The total sample size n is given by $n = n_r + n_u$. We shall estimate $H_1, \ldots, H_J$ and H, where $H_y = N_y / y$ and $N_y = \sum_{i=1}^{N} I(Y_i = y)$. Hence,

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

A general model-based estimator for H_y can be obtained by estimating $E(H_y)$, replacing $P(Y_i=y \mid \mathbf{x}_i)$ by an estimate $\hat{P}(Y_i = y | \mathbf{x}_i)$ obtained by estimating the unknown parameters in the population model. This is what is usually called the regression estimator ,

$$\hat{H}_{y,reg} = \frac{1}{y} \sum_{i=1}^{N} \hat{P}(Y_i = y | \mathbf{x}_i) .$$

Since the household size Y is assumed to depend only on the family size x_1, the regression estimator takes the form

(4.1) $$\hat{H}_{y,reg} = \frac{1}{y} \sum_{k=1}^{K} M_k \hat{P}(Y = y | x_1 = k),$$

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

An alternative estimator to (4.1) can be obtained by utilizing that a part of N_y is observed, $N_y = n_y + \sum_{i \notin s_r} I(Y_i = y)$, where s_r is the response sample. There is no need to estimate n_y leading to what is sometimes called the prediction estimator,

$$\hat{H}_{y,pred} = \frac{1}{y}\left(n_y + \sum_{i \notin s_r} \hat{P}(Y_i = y | \mathbf{x}_i) \right). \tag{4.2}$$

The prediction estimator gives practically the same results as the regression estimator for the survey we consider. To the nearest 100, the estimates for this survey are identical. This will typically happen when the sample is a small proportion of the population, as seen from the relationship

$$\hat{H}_{y,pred} = \hat{H}_{y,reg} + \frac{1}{y}\left(n_y - \sum_{k=1}^{K} m_k \hat{P}(Y = y | x_1 = k) \right).$$

A second alternative approach to (4.1) is poststratification. See, for example Holt and Smith (1979) and Särndal, Swensson and Wretham (1992, ch. 7.6). We shall consider the poststratified estimator $\hat{H}_{y,post}$, using family size as the stratifying variable,

$$\hat{H}_{y,post} = \frac{1}{y} \sum_{k=1}^{K} M_k \frac{n_{ky}}{m_k}. \tag{4.3}$$

This estimator corresponds to (4.1) using the model (3.1) for Y, and assuming ignorable response mechanism. In this case the data are only the y- and x_1 -values in the response sample, and the likelihood function is given by $\prod_{i=1}^{n_r} P(Y_i = y_i | x_{1i})$. The maximum likelihood estimate $\hat{P}(Y = y | x_1)$ is simply the observed rate among the respondents with household size y given family size x_1, $\hat{P}(Y = y | x_1 = k) = n_{ky}/m_k$.

4.2. Estimation by imputation, based on the free model p_{y,x_1}

A common approach to correct for nonresponse is by imputation of the missing values in the sample. We then assign imputed values for the nonrespondents, by some estimation method based on the response sample. In this section we consider the three estimators in the previous section, based on the completed sample obtained by filling in the missing values with model-based imputed values. It is seen that the imputation-based regression-, prediction-, and poststratified estimators are all identically the same, when using the free model p_{y,x_1}. Moreover, it turns out that this common imputation-based estimator equals the maximum likelihood regression estimator given by (4.1) for this population model, thereby showing us which imputed values the regression estimator implicitly are using.

For imputation, we shall use the estimated distribution for Y given family size and place of residence for the nonrespondents, $\hat{P}(Y = y|\, x_1, x_2, r = 0)$ for x_1 = 1,...,5 and x_2 = 0,1. We then assign, for a given family size x_1 and place of residence x_2, the nonrespondents to the values 1, ..., 5 in proportions given by $\hat{P}(Y = y|\, x_1, x_2, r = 0)$ for y = 1, ..., 5. Let $n^*_{x_1y}(0)$ ($n^*_{x_1y}(1)$) be the number of imputed values with family size x_1 and household size y, for rural (urban) areas and let $m_{x_1u}(0)$ ($m_{x_1u}(1)$) be the number of missing observations for persons in rural (urban) areas with family size x_1. Then

(4.4) $$n^*_{x_1y}(x_2) = m_{x_1u}(x_2) \cdot \hat{P}(Y = y|\, x_1, x_2, r = 0),\ x_2 = 0,1,$$

and $$n^*_{x_1y} = n^*_{x_1y}(0) + n^*_{x_1y}(1)$$

is the total number of imputed values with family size x_1 and household size y.

The imputed poststratified estimator (estimator (4.3) computed for the completed sample) becomes

(4.5) $$\hat{H}_{y,post}{}^I = \frac{1}{y}\sum_{k=1}^{K} M_k \frac{n_{ky} + n_{ky}{}^*}{m_k + m_{ku}}.$$

From (4.1) and (4.2) , the imputation-based regression- and prediction estimators are given by

(4.6) $$\hat{H}^I_{y,reg} = \frac{1}{y}\sum_{k=1}^{K} M_k \hat{P}^I(Y = y|\, x_1 = k).$$

and

(4.7) $$\hat{H}^I_{y,pred} = \frac{1}{y}\left(n_y + n_y{}^* + \sum_{k=1}^{K}(M_k - m_k - m_{ku})\hat{P}^I(Y_i = y|\, x_1 = k)\right),$$

where $n_y^* = \sum_k n_{ky}^*$. The estimated probabilities $\hat{P}^I$ are obtained from the completed sample, using the model (3.1), i.e., $\hat{P}^I(Y = y|x_1)$ is given by the rate of household size y for family size x_1 in the completed sample. Hence, both $\hat{H}^I_{y,reg}$ and $\hat{H}^I_{y,pred}$ correspond to the poststratified estimator (4.5). With population model (3.1), the imputation estimators (4.6) and (4.7) will always be identical to the imputed poststratified estimator. Furthermore, the following general result holds, showing that with population model (3.1), the imputation-based poststratified estimator (4.5) is identical to the maximum likelihood regression estimator (4.1).

Theorem. Assume model (3.1) for Y. I.e., $P(Y=y|\, x_1, x_2) = p_{y,x_1}$ is independent of x_2, but otherwise the p_{y,x_1}'s are completely unknown with the only restriction $\sum_y p_{y,x_1} = 1$, for all values of x_1. The response mechanism is arbitrarily parametrized, i.e., no assumption is made about $P(R=1|Y=y,x_1,x_2)$. Then the maximum likelihood estimates for p_{y,x_1} are given by, for $x_1 = k = 1,...,K$,

$$\hat{p}_{y,k} = \frac{n_{ky} + n_{ky}^*}{m_k + m_{ku}} .$$

Proof. See Appendix B.

4.3. Imputation-based poststratification with a saturated model

We know proceed to an intuitive metod of imputation that was used to estimate response probabilites for a modified Horvitz-Thompson estimator in the official statistics from the 1992 *CES*, described in (Belsby, 1995). As we shall see in section 5, the Horvitz-Thompson estimator often fail to correct for biased samples, and we will use this imputation method for the poststratified estimator (4.5).

The imputation method consists of distributing, within rural/urban area, the $m_{ku}(x_2)$ 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 areas respectively, is identical for different family sizes. Denote, for $x_2 = h$, the number of nonresponse persons with family size $x_1 = k$ and household size y obtained in this manner by $z_{ky}(h)$. The corresponding number among the respondents is $n_{ky}(h)$. The values of $z_{ky}(h)$ are determined by the equations

(4.8) $$\frac{z_{ky}(h)}{z_{ky}(h)+n_{ky}(h)} = \frac{z_{iy}(h)}{z_{iy}(h)+n_{iy}(h)} \quad , \qquad h = 0,1.$$

When $n_{ky}(h) = 0$, we let $z_{ky}(h) = 0$. The equation (4.8) is solved under the conditions

(4.9) $$\sum_y z_{ky}(h) = m_{ku}(h) \quad ; k = 1, ..., 5 \text{ and } h = 0,1.$$

Solving (4.8) and (4.9) requires, for each value of h, one row $(n_{k1}(h), n_{k2}(h), \ldots, n_{k5}(h))$ of nonzeros, which holds for our case. The imputed values $z_{ky}(h)$ determined by (4.8) and (4.9) correspond to the imputation method described by (4.4) for the following model:

(4.10a) $P(Y=y|x_1, x_2) = p_{y,x_1,x_2}$ with no restrictions

(4.10b) $P(R=1|Y=y, x_1, x_2) = q_{y,x_2}$, independent of x_1

This can be seen as follows:

For the ten multinomial trials determined by the different (x_1, x_2)- values, we have 50 unknown cell probabilities $\pi_{yk,h} = P(Y=y,R=1| x_1=k, x_2=h)$. With no restrictions on cell probabilities, the maximum likelihood estimates (mle) are given by the observed relative frequencies,

$$\hat{\pi}_{yk,h} = \frac{n_{ky}(h)}{m_k(h)+m_{ku}(h)} \quad .$$

This also holds when $n_{ky}(h) = 0$. Now, it can be shown that there is a one-to-one correspondence between $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ and $(\boldsymbol{p}_0,\boldsymbol{q}_0,\boldsymbol{p}_1,\boldsymbol{q}_0)$, where $\boldsymbol{\pi}_h = (\pi_{yk,h} : y=1,...,5 ; k= 1,...,5)$, $\boldsymbol{p}_h = (p_{yk,h}: y=1,...,5 ; k= 1,...,5)$ and $\boldsymbol{q}_h = (q_{1,h}, \ldots, q_{5,h})$. Since $\pi_{yk,h} = p_{y,k,h} \cdot q_{yh}$, the mle of $p_{yk,h}$ and $q_{y,h}$ must satisfy

(4.11) $$\hat{p}_{yk,h} \cdot \hat{q}_{y,h} = \frac{n_{ky}(h)}{m_k(h)+m_{ku}(h)},$$

and are uniquely determined by $\hat{\pi}_{yk,h}$.

Consider $z_{ky}(h)$, given by (4.8)&(4.9). Let $z_y(h) = \sum_k z_{ky}(h)$ and $n_y(h) = \sum_k n_{ky}(h)$. From (4.8),

(4.12) $$\frac{z_j(h)}{z_j(h)+n_j(h)} = \frac{z_{kj}(h)}{z_{kj}(h)+n_{kj}(h)} \quad , \text{ if } n_{kj}(h) > 0 \, .$$

From (4.11) and (4.12) we have that the following intuitive estimates also are mle.

(4.13) $$\hat{q}_{y,x_2} = \frac{n_y(x_2)}{n_y(x_2)+z_y(x_2)}$$

and

(4.14) $$\hat{p}_{y,x_1,x_2} = \frac{n_{x_1y}(x_2)+z_{x_1y}(x_2)}{m_{x_1}(x_2)+m_{x_1u}(x_2)} \quad \text{(also when } n_{x_1y}(x_2) = z_{x_1y}(x_2) = 0\text{).}$$

We could also have shown (4.13) and (4.14) by maximizing the loglikelihood directly. Next, we show that the imputed values (4.4) for the model (4.10) equal $z_{x_1y}(x_2)$. From (4.4), we have $n^*_{x_1y}(x_2) = m_{x_1u}(x_2) \cdot \hat{P}(Y = y | x_1, x_2, r = 0)$. Under model (4.10) and estimates (4.13) and (4.14), we find that

$$\hat{P}(Y = y | x_1 = k, x_2 = h, R = 0) = \frac{\hat{P}(Y = y | x_1 = k, x_2 = h) - \hat{P}(Y = y, R = 1 | x_1 = k, x_2 = h)}{\hat{P}(R = 0 | x_1 = k, x_2 = h)}$$

$$= \frac{\hat{p}_{yk,h} - \hat{\pi}_{yk,h}}{1 - \sum_y \hat{\pi}_{yk,h}} = \frac{n_{ky}(h) + z_{ky}(h) - n_{ky}(h)}{m_{ku}(h)} = \frac{z_{ky}(h)}{m_{ku}(h)} \quad ,$$

and it follows that $n_{ky}^*(h) = z_{ky}(h)$. If $n_{ky}(h) = 0$, then $\hat{p}_{y,k,h} = \hat{\pi}_{yk,h} = 0$, and $n_{ky}^*(h) = 0$. We note that model (4.10) is saturated and will, from (4.11), give perfect fit.

5. Estimated number of households of different sizes for the 1992 Norwegian Consumer Expenditure Survey

In this section we consider, for *CES* 1992, the estimation of the number of households of size 1,...5, i.e., $H_1, \ldots, H_5$ and the total number of households, H for the population, based on the estimators considered in Section 4. For comparison, and to illustrate the effects of nonreponse modeling and poststratification, we also present estimates based on the regular expansion estimator, given by

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

and the imputation-based expansion estimator given by

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

Recall that n_y is the number of respondents belonging to households of size y, n_r is the total number of respondents, and $n_y{}^* = \sum_{x_1} n^*_{x_1 y}$. The estimator (5.1) does not seek to correct for nonresponse nor use the family population distribution as a post-stratifying tool to improve the estimation, while estimator (5.2) tries to take the response mechansim into account, but cannot correct for biased samples.

To compute the estimates we need the number of families of different sizes in the population, i.e., M_k, 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 1.1.93. These are given in table 7.

Table 7. Families and persons with age less than 80 years in Norway at 1.1.93.

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 670528/127653 = 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 regression estimation and poststratification

We give the regression estimates from (4.1) using the free model p_{y,x_1} in (3.1) in combination with the response models RM1(y,x_2) and RM2(y,x_2). The estimated household distributions are presented in table 8 for the different models, using tables 2 and 3. To illustrate the effect of nonresponse modeling versus poststratification we also present the poststratified estimate given by (4.3), using family size as the stratifyer and the the simple expansion estimate, given by (5.1).

Table 8. Estimated household totals for persons aged less than 80 years in Norway at 1.1.93 in units of 100

	Correction for nonresponse and family size				Correction for family size		No correction for nonresponse nor family size	
Household size, y	Regression estimator - model (3.1) & RM1(y,x_2)	%	Regression estimator - model (3.1) & RM2(y,x_2)	%	Post-stratified estimator	%	Regular expansion estimator	%
1	558 800	32	595 400	34	486 000	29	390 500	24
2	520 200	30	525 800	30	507 800	30	496 500	31
3	278 900	16	249 100	14	286 200	17	283 900	18
4	258 900	15	269 000	15	270 600	16	279 900	18
≥ 5	125 800	7	126 000	7	131 300	8	148 000	9
Total	1 742 600	100	1 765 300	100	1 681 900	100	1 598 800	100

The expansion estimates indicate serious bias due to nonresponse, especially the estimates for H_1 and H, with poststratification correcting for some of the bias (probably about 50 per cent for the estimates of H_1 and H). 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 7 we see that these family size distributions are given by (in percentages), for $k = 1, \ldots, 5$:

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. Poststratification 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. This is most likely not the case. It is reasonable to assume, as in our response models, that response rates will vary with the actual household sizes rather than the registered the family sizes. Typically, estimates of the number of one-person households will be biased when the nonrespondents are ignored, as seen in The Post Enumeration Survey for the 1990 Norwegian Population and Housing Census, *PES 1990*. Here it was estimated to be 626 000, while the Census estimate was about a 100 000 less, see (Schjalm, 1996). In the Census 1990, poststratification primarily after family size was used and gave estimates similar to the poststratified estimates in table 8.

The two 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 mostly influenced by taking the response mechanism into account.

For robustness considerations we also present the estimates from model CLM(x_1) and RM1(y,x_2), 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 $H_1, \ldots, H_5$ and H, respectively. Compared to table 8, 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. Imputation and estimation based on the saturated model

From Section 4.2 we have that the imputation-based estimators based on the regression-, prediction- and poststratified estimator are identically equal to the maximum likelihood regression estimator. The completed sample, including the imputed values, for the two models are given in Appendix C.

We shall consider the poststratified estimator (4.5) based on the the imputation method (4.4) for the model (4.10), given by the equations (4.8) and (4.9), i.e.,

$P(Y=y|x_1, x_2) = p_{y,x_1,x_2}$ with no restrictions

$P(R=1|Y=y, x_1, x_2) = q_{y,x_2}$, independent of x_1,

i.e., within rural/urban area and given household size, the rate of nonresponse is the same for all family sizes. This method was used to estimate response probabilities and employed in the official statistics from the 1992 *CES* with the modified Horvitz-Thompson estimator. Model (4.10) is saturated and will, from (4.11), give perfect fit. The completed sample based on the equations (4.8) and (4.9) is given in Appendix D. Table 9 presents the table of ($n_{ky} + n_{ky}$*), necessary to compute the poststratified estimates given by (4.5).

Table 9. The total numbers of family and household sizes for imputed complete sample. Based on model (4.10)

	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	1698

The imputation-based poststratified estimates (4.5) are given in table 10. To illustrate the effect of poststratification to correct for possible biases in the sample, we include the imputation-based expansion estimates (5.2) using the completed total number of households from the last row in table 9.

Table 10. Imputation-based poststratified estimates and expansion estimates of household totals for persons aged less than 80 years in Norway at 1.1.93; model (4.10), in units of 100

Household size	Post-stratified estimator	Per cent	Expansion estimator	Per cent
1	596 600	34	583 900	33
2	523 600	30	567 700	32
3	250 000	14	244 300	14
4	268 900	15	259 300	15
≥ 5	126 200	7	122 400	7
Total	1765 300	100	1 777 600	101

We note that model (3.1) and RM2(y,x_2) gives practically the same poststratified estimates as model (4.10). In units of 1000 households, we have 597, 524, 250, 269, 126 and 1 765 for model (4.10) and 595, 526, 249, 269, 126 and 1 765 for model (3.1) and RM2(y,x_2). Because of the freedom of model (4.10), with perfect fit, it seems that model (3.1) & RM2(y,x_2) works well for estimating the number of households of different sizes.

Comparing the poststratified- and expansion estimates, 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 k = 1, ..., 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 two-persons families is much too high, and even though we have corrected for nonresponse bias, the expansion estimator cannot correct for a biased sample. This bias will necessarily lead to biased estimates of H_2. We need poststratification to correct for a biased sample.

5.3. Comparison with the quality survey for the 1990 Census and a projection study

The quality survey for the Census 1990, *PES 1990,* contains 8280 respondents and uses the same household definition as *CES*. It had a response rate of 95 per cent. The H_y-estimates uses poststratification with respect to household size in the Census. However, no attempt were made to correct for possible nonresponse bias with respect to actual household size. *PES* deals with the whole population. Table 11 presents the estimates for the 0-79 age group with the same poststratification method as in *PES.*

Table 11 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 11. Estimated household size totals for persons aged less than 80 years in Norway at 1.1.93, PES 1990 and projections in units of 100

Household size	PES 1990	Per cent	Projections	Per cent
1	626 000	35	668 300	37
2	494 200	28	549 000	30
3	291 500	16	211 900	12
4	250 000	14	221 500	12
≥ 5	115 300	6	97 500	5
Unknown			78 500	4
Total	1 777 000	99	1 826 700	100

As seen in Section 5.1, it seems that estimates based on modeling of the reponse mechanism leads to less biased estimates than mere poststratification or simple expansion, especially of H_1 and H. This is further supported by the *PES 1990* estimates which are based on a sample with only 5 per cent nonresponse.

5.4. Comparisons with estimates in official statistics

The imputation-based expansion estimates in the second column of table 10 are identical to the modified Horvitz-Thompson estimates with $\hat{q}_{y,x_2} = n_y(x_2)/[n_y(x_2) + n_y^*(x_2)]$ (from (4.13)) 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

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

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

$$\pi_i = \frac{n}{N} \hat{P}(R_i = 1 | x_{1i}, x_{2i}, Y_i = y) = \frac{n}{N} \hat{q}_{y,x_{2i}}$$

and

$$\hat{N}_{y,HT} = \frac{N}{n} \left(\frac{n_y(0)}{\hat{q}_{y,0}} + \frac{n_y(1)}{\hat{q}_{y,1}} \right) . \tag{5.3}$$

Here,

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

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

So this modified Horvitz-Thompson estimator suffers from the same negative feature as the imputation based expansion estimator (5.2); 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, ch.15).

Since 1993, an alternative, computationally simpler, modified Horvitz-Thompson estimator of type (5.3) has been in use in the production of official statistics from the Norwegian *CES*, see (Belsby 1995). Here, the probability of response is estimated using a logistic model similar to RM2(y,x_2), except that when estimating the parameters in the response model, the family size is used in the nonresponse group in place of household size, replacing (3.6). Of course, using (3.6) is possible only when a population model is considered, which *CES* has not done. The estimated response probabilities are given in table 12, with the estimates based on RM2(y,x_2) & (3.1) from table 4.

Table 12. 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_1} in (3.1) combined with RM2(y,x_2)					
Rural	47.77	60.90	79.05	73.26	81.52
Urban	38.92	52.04	72.44	65.62	75.46

Compared to the estimated response probabilities based on model RM2(y,x_2) with (3.1), we see that replacing household size with family size in the nonresponse group is not a satisfactory approximation. 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 biased samples (as has been a problem in *CES* since 1993). In the 1992 *CES*, however, the sample is biased with a too high proportion of families of size 2, and the H_2-estimate will be of the right magnitude, by accident. The H_y - estimates in the 1992 *CES* with this current "official" estimator are 622 900, 518 500, 259 900, 258 500, 124 600 and 1 784 400 for H_1, ..., H_5 and H respectively.

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 nonresponse bias. 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 model-based approaches, a regression estimator and imputation-based poststratification after registered family size. With a population model that corresponds to a group model after family size only, the regression estimator and imputation-based poststratified estimator are identical. This family group model for household size and a logistic

link for the response probability using household size as a categorical variable seems 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 per cent 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* corresponds to imputation-based expansion estimates. Hence, they cannot correct for biased 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.

Our final conclusion is that Horvitz-Thompson-estimation should be replaced by poststratified imputation-based estimation with an response model like RM2 (with population model (3.1)), or a total model like (4.10). Moreover, in any estimation problem of totals in survey sampling, one must be aware of the fact that a Horvitz-Thompson estimator cannot correct for biased samples, even when modified with good response estimates. Poststratification should always be considered as well as imputation based on nonignorable response model.

References

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 Statististical Assosiation* **91**, 791-806.

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

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

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

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

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

Keilman, N. and H. Brunborg (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 Assosiation* **77,** 237-250.

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

Schjalm, A. (1996): The Post Enumeration Survey for the 1990 Norwegian Population and Housing Census, *Proceedings from the 5th IAOS Conference*, Reykjavik, Iceland, 2-5 July 1996.

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

Appendix A

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

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
1 urban	28	24	7	2	0	61	78	139
rural	55	24	13	7	2	101	75	176
2 urban	6	70	12	3	0	91	84	175
rural	3	107	25	1	3	139	76	215
3 urban	4	8	57	11	3	83	40	123
rural	6	17	74	29	3	129	51	180
4 urban	0	3	15	80	5	103	43	146
rural	2	10	22	151	12	197	80	277
≥ 5 urban	0	1	0	6	66	73	28	101
rural	1	3	4	11	115	134	32	166
Total urban	38	106	91	102	74	411	273	684
Total rural	67	161	138	199	135	700	314	1014

Appendix B

Theorem. Assume model (3.1) for *Y*. I.e., $P(Y=y|x_1=k, x_2=h) = p_{yk}$ is independent of *h*, but otherwise the p_{yk}'s are completely unknown with the only restriction being that $\sum_y p_{yk} = 1$, for all *k*. The response mechanism is arbitrarily parametrized, i.e., no assumption is made about $P(R=1|Y=y, x_1=k, x_2=h)$. Then the maximum likelihood estimates for p_{yk} are given by

$$\hat{p}_{yk} = \frac{n_{ky} + n_{ky}^*}{m_k + m_{ku}} .$$

Proof. Let $q_{yk,h} = P(R=1|Y=y, x_1=k, x_2=h)$. The log likelihood is given by

$$\ell = \sum_k \sum_y n_{ky} p_{yk} + \sum_{h=0}^{1} \sum_k \sum_y n_{ky}(h) q_{yk,h} + \sum_{h=0}^{1} \sum_k m_{ku}(h) \log P(R=0|x_1=k, x_2=h)$$

$$= \sum_k \sum_y n_{ky} p_{yk} + \sum_{h=0}^{1} \sum_k \sum_y n_{ky}(h) q_{yk,h} + \sum_{h=0}^{1} \sum_k m_{ku}(h) \log(1 - \sum_{y=1}^{1} p_{yk} q_{yk,h}).$$

We use Lagrange method and maximize $G = \ell + \sum_{k=1}^{5} \lambda_k (\sum_{y=1}^{5} p_{yk} - 1)$.

Let the solutions be $\hat{p}_{yk}(\lambda_k)$, and determine the λ_k's such that $\sum_y \hat{p}_{yk}(\lambda_k) = 1$, for all *k*. No matter how the $q_{yk,h}$'s are parametrized, the mle $\hat{p}_{yk}$ must satisfy, by solving the equations $\partial G/\partial p_{yk} = 0$,

$$\text{(A1)} \qquad \frac{n_{ky}}{\hat{p}_{yk}} - \sum_{h=0}^{1} m_{ku}(h) \frac{\hat{q}_{yk,h}}{\hat{P}(R=0|x_1=k, x_2=h)} + \lambda_k = 0$$

which is equivalent to

$$n_{ky} = \hat{p}_{yk} \sum_{h=0}^{1} \frac{m_{ku}(h)}{\hat{P}(R=0|x_1=k, x_2=h)} - \sum_{h=0}^{1} m_{ku}(h) \frac{\hat{P}(R=0, Y=y|x_1=k, x_2=h)}{\hat{P}(R=0|x_1=k, x_2=h)} - \hat{p}_{yk}\lambda_k .$$

We determine λ_k by summing over *y*

$$m_k = \sum_{h=0}^{1} \frac{m_{ku}(h)}{\hat{P}(R=0|x_1=k, x_2=h)} - \sum_{h=0}^{1} m_{ku}(h) \frac{\hat{P}(R=0|x_1=k, x_2=h)}{\hat{P}(R=0|x_1=k, x_2=h)} - \lambda_k ,$$

hence

$$\lambda_k = \sum_{h=0}^{1} \frac{m_{ku}(h)}{\hat{P}(R=0|x_1=k, x_2=h)} - (m_k + m_{ku}).$$

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

(A2) $$\hat{p}_{yk} = n_{ky} \Big/ \left(m_k + m_{ku} - \sum_{h=0}^{1} m_{ku}(h) \frac{\hat{P}(R=0|Y=y, x_1=k, x_2=h)}{\hat{P}(R=0|x_1=k, x_2=h)} \right).$$

The imputed values are given by , from (4.4),

$$n_{ky}{}^*(h) = m_{ku}(h)\hat{p}_{yk} \frac{\hat{P}(R=0|Y=y, x_1=k, x_2=h)}{\hat{P}(R=0|x_1=k, x_2=h)},$$

and from (A2),

$$\hat{p}_{yk} = n_{ky} \Big/ \left(m_k + m_{ku} - \sum_{h=0}^{1} \frac{n_{ky}{}^*(h)}{\hat{p}_{yk}} \right)$$

$$= n_{ky} \Big/ \left(m_k + m_{ku} - \frac{n_{ky}{}^*}{\hat{p}_{yk}} \right)$$

$\Leftrightarrow$

$$\hat{p}_{yk}(m_k + m_{ku}) - n_{ky}{}^* = n_{ky} \quad \text{, i.e.} \quad \hat{p}_{yk} = \frac{n_{ky} + n_{ky}{}^*}{m_k + m_{ku}}.$$ Q.E.D

Appendix C

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,x_2)

	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	1014

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,x_2)

	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	1014

Appendix D

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 (4.10), i.e imputations determined by (4.8) and (4.9)

	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	1014