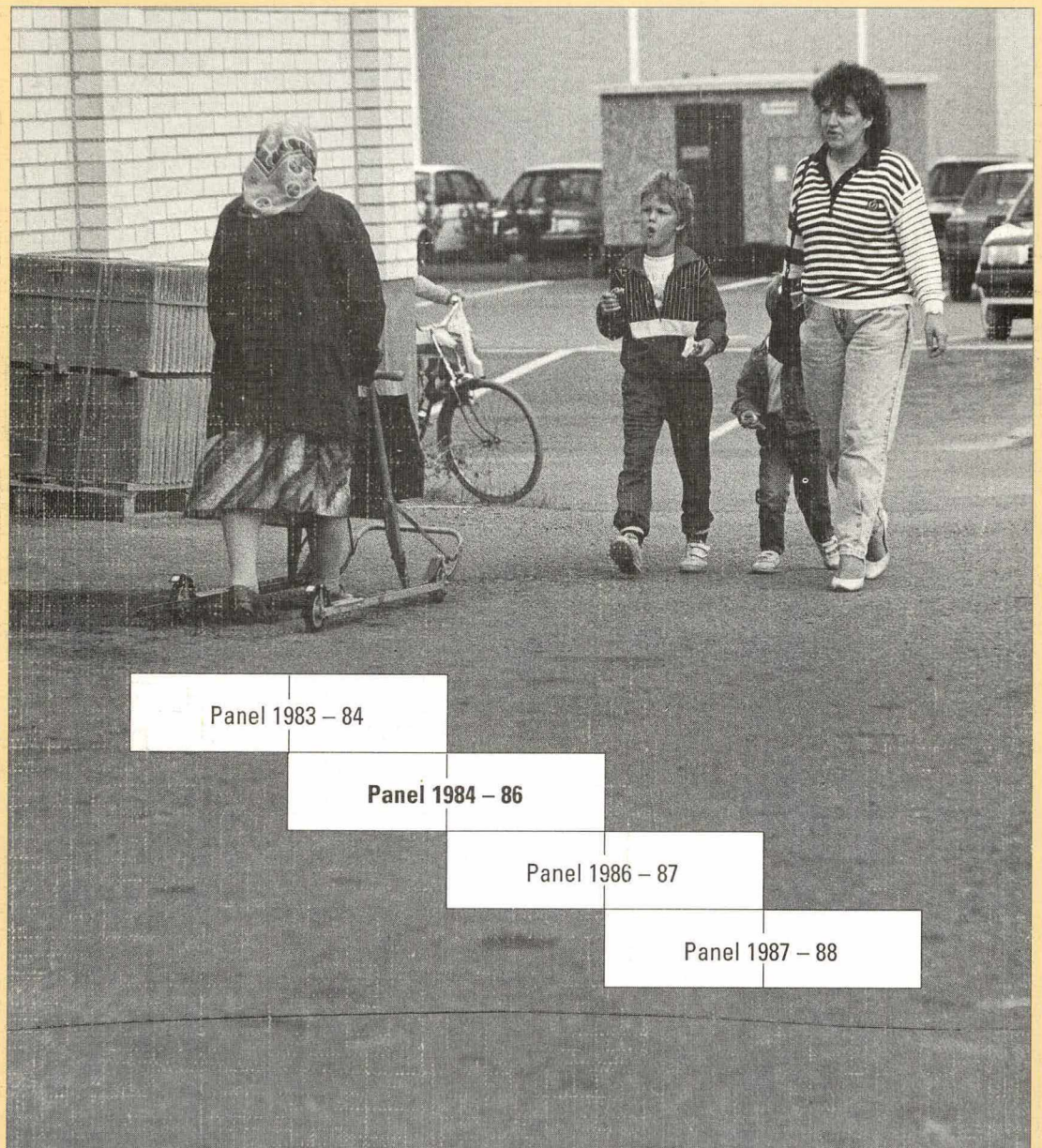
Tilastokeskus
Statistikcentralen
Central Statistical
Office of Finland

Tutkimuksia 179
Studies

Comparative Adjustments for Missingness in Short-term Panels

Applications to Questions of Household Income Distribution

Seppo Laaksonen



Tilastokeskus
Statistikcentralen
Central Statistical
Office of Finland

Tutkimuksia 179
Studies

Comparative Adjustments for Missingness in Short-term Panels

Applications to Questions of Household Income Distribution

Seppo Laaksonen

April 1991

Helsinki 1991

Inquiries:
Seppo Laaksonen

VAPK

Cover: Mikko Nurmi

SVT *Suomen Virallinen Tilasto*
Finlands Officiella Statistik
Official Statistics of Finland

Central Statistical Office of Finland, Studies No. 179, 1991.

COMPARATIVE ADJUSTMENTS FOR MISSINGNESS IN SHORT-TERM PANELS. APPLICATIONS TO QUESTIONS OF HOUSEHOLD INCOME DISTRIBUTION.

BY SEPPO LAAKSONEN

Abstract

The material studied consists of short-term household panel data applying to two points in time. The variable of interest is disposable income, but the methods are also applicable to many other quantitative variables. An unusually large number of requirements is placed on the measurement of this variable in both cross-sectional and panel analysis, objects of interest in addition to the mean being the evenness of the distribution, quantiles and changes in the panel with duration. Importance is also attached to explanation of the reasons for the changes.

Since there are problems with response and overcoverage, is it necessary to carry out adjustments for nonresponse and to handle overcoverage correctly. Both weighting and imputation methods are used for nonresponse adjustment, employing data on both respondents and nonrespondents. The data on the nonrespondents are compiled from registers and from the previous year's interviews in the case of the second year of the panel.

The weighting method used is essentially based on the modeling of response probabilities, and its purpose is adjustment for unit nonresponse. Single and multiple imputation methods based on a regression model are used, and also hot deck imputation based on the results of both the regression model and the weighting method. Imputation is by nature an item nonresponse adjustment method, but it can also be used for unit nonresponse, at least in methodological comparisons, as in the present case.

The results suggest that nonresponse cannot be ignored when studying the distribution of incomes and changes in these if it is as great and as skewed in its distribution as it is in the Finnish Income Distribution Survey (IDS) for the 1980's.

Key Words: Disposable Income, Hot Deck Imputation, Multiple Imputation, Nonresponse, Regression Imputation, Response probability, Single Imputation.

CONTENTS

1. INTRODUCTION

The principal source of data on the distribution of incomes in Finland is the official Income Distribution Survey (IDS) compiled annually since 1977, with the exception of 1985, information for which is obtainable from the Finnish Household Budget Survey. The latter also serves as a source of more long-term income data, since it has been carried out regularly at approximately five year intervals since 1966.

The broadest income concept listed in the IDS is disposable income per household, i.e. the sum of the incomes of all the members of a household. The structure of this measure means that large households are more frequently credited with greater incomes than small ones, just as their consumption needs are also greater. When comparing incomes between different groups it is thus necessary to standardize household sizes by means of 'consumption units'. The most commonly used unit of this kind is one in which the first adult in the household is assigned the value 1.0, any subsequent adult 0.7 and each child 0.5. Summation of these figures gives the number of consumption units in the household, for use as a denominator in the comparison of incomes.

Apart from enabling incomes to be compared between groups, the IDS provides data on the distribution of incomes between households and changes taking place in this distribution. Numerous alternatives exist for describing these changes. When measuring the evenness of the distribution and its changes, it is customary to use the Gini coefficient, Theil measures or a coefficient of variation (see Cowell 1977, Nygård & Sandström 1985). Since indices reduced to single figures seldom give a full picture of what is going on, it is also necessary to examine different parts of the distribution, for which purpose Lorenz curves and deciles may be used, the highest and lowest deciles being those of greatest interest.

Reliable evaluation of the extremes in a distribution is a substantive matter in income surveys based on sampling techniques, because these can have a sensitive effect on variances and other

deviation measures. Some of these deviant observations or outliers may be errors which have remained uncorrected at the compilation stage. Other outliers are in effect correct values but they should not be accepted into the sample with the weighting that they would normally be given because their representativeness is not great.

It is therefore necessary to improve the robustness of the estimates, particularly by handling the margins of the income distribution. This is done here in the same manner as in the author's earlier work (see Laaksonen 1988 and 1989; cf. Curtin et al 1988, p. 27), by limiting the outliers to a certain area which can be deemed acceptable. This method is also known as 'Winsors' principle' or 'winsorization'. Since the same limits are employed in all the examples, the results obtained with different methods are comparable.

The IDS compiled up to 1982 were based on independent annual samples of households, and thus represented cross-sectional data. A procedure change was made at this point, however, to allow half of the 1982 sample to be retained into 1983, and correspondingly in 1984, 1986, etc. This implied the compilation of panel data (for the concept of panel, see Kish 1987 and van de Pol 1990). The disadvantage with this approach is, however, that households can drop out into the void or alter in composition as the panel is followed. The Finnish IDS adopts a simple solution to this problem, taking the one key individual around whom the sample household was originally constituted and following the household to whom this person belongs from one year to the next. Perhaps the best-known panel study, the Survey of Income and Program Participation (SIPP), implemented by the U.S. Bureau of the Census, adopts the approach of following all the members of an original household no matter to what household they may belong later (see Singh et al 1989).

The addition of the panel feature to the IDS opens up new research opportunities, as discussed in Laaksonen (1989) in the context of the 1982-83 and 1983-84 panels. Panel studies nevertheless entail problems of their own, as mentioned in the same paper. The present report is concerned with two of these problems, nonresponse and overcoverage. The material used here consists of the corresponding panel for 1984-86, which offers even better opportunities for examining nonresponse, although it still does not meet all the criteria for a good nonresponse study since it was not

possible to update the information on households failing to reply in either year in 1986. Consequently the 1986 results are not valid for use as such, although they do give an adequate impression of the appropriateness of the various methods tested.

Compensation or adjustment for nonresponse can in principle be performed in two ways (see Kalton & Kasprzyk 1986, Laaksonen 1988), by adjusting the weighting in accordance with the nonresponse data, or by imputation, i.e. replacement with suitably estimated values. The weighting method is particularly appropriate in cases of unit nonresponse and imputation for item and partial nonresponse. Both methods are employed here. The weighting method is quite similar in its basic features to that used in the Finnish Household Budget Survey of 1985, being based on nonresponse probability modeling (see Laaksonen 1988, Ekholm and Laaksonen 1990). Some modifications are made on account of differences in the manner of sampling between the two surveys, and some new features are achieved by the panel approach itself (see also Waterton and Lievesley 1987 who discuss more general panel attrition problems). The principles of this solution are discussed in Section 3.

Of the various imputation methods available, single regression imputation has been used earlier in a similar situation (see Laaksonen 1988). The idea is to define auxiliary variables from a register on the basis of the data for the respondents and to use these as explanatory variables in a model which gives a good fit with the outcome variable y, in this case with disposable income. The model will be assumed to hold for nonrespondents as well. Under this assumption it may be used to predict the missing values, starting out from known explanatory variables.

The multiple imputation method, which has been developed intensively in recent times (see Herzog & Rubin 1983, Rubin 1987, Rubin & Schenker 1987), is also tested here. In this respect, too, the panel approach opens up additional possibilities, and these will be discussed along with regression imputation in general in Section 4. Results obtained with hot deck imputation will also be presented in that section. These are based to a substantial extent on the other findings, i.e. the weighting method, the regression model and a model for the explanation of income changes.

Section 2 is devoted to describing the data to be corrected for nonresponse and the testing of the method on simulated data,

Section 5 summarizes the results and Section 6 contains a discussion of the experiences gained with respect to both the evaluation of income distributions and panel analysis (cf. Laaksonen 1989), and also as far as the practical generation of statistics is concerned.

This report has three main aims:

(a) to develop and apply methods which are useful for adjusting for errors, biases and other confusions of nonresponse and other forms of attrition in large-scale panel surveys in official statistics, especially in relation to studies of household income distribution,

(b) to compare the applicability of these methods of adjustment in certain real cases, and

(c) to describe mechanisms which are crucial when studying short-term panels. We shall call these mechanisms '**panel characteristics**'. Some of these offer advantages, others introduce disadvantages, and others are neutral.

Large-scale sampling surveys are highly complex objects of investigation, and it is not possible here to give full consideration to all the estimates or measures needed. The following five types of estimator have therefore been selected as specific examples:

(1) **Totals**; particularly number of households in the whole country and its regions and in domains such as socio-economic groups. These are highly important from the point of view of other estimates, too, whereas income totals, for example, are not of interest in income distribution analyses.

(2) **Means** or averages; particularly mean size of household and mean disposable income per household or per consumption unit.

(3) **Deciles, other quantiles** and **coefficient of variation**. These estimates describe the variation and evenness in income distribution at different points of time. The first decile points are chosen as they show the greatest relative standard error in our surveys (see Appendix).

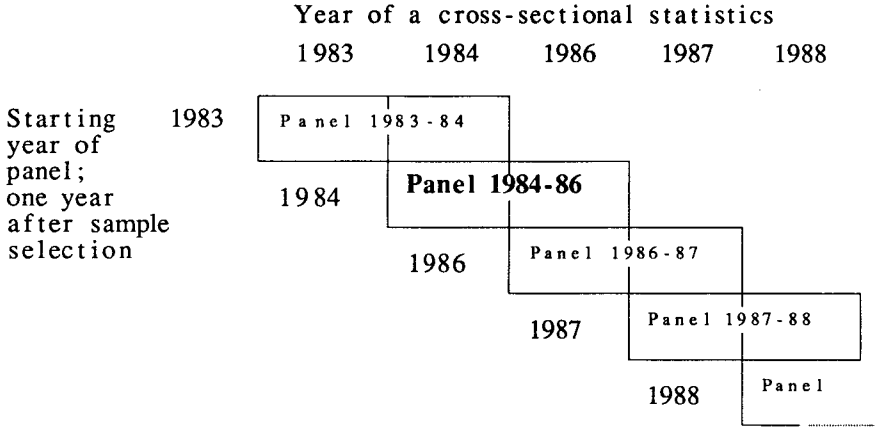
Coefficients of variation have two roles in this report. On the one hand they provide for evenness or unevenness in distributions, and on the other hand they indicate the variance attached to the means and totals and thus serve as indicators of their accuracy. These indicators are reasonably good for our purposes as we are comparing the variances or standard errors of different methods, and as a larger coefficient of variation will in general indicate a greater variance estimate, given the same data material and sampling design (see the formula of Laaksonen 1988, p. 12, in which the majority of the standard error of the total consists of ordinary sampling variance such as weighting with sampling weights, and consequently for the standard error of the means).

(4) **Individual changes** in income between two points in time in the panel. Our examples provide only mean values for the changes, but their distribution is also interesting (see Laaksonen 1989).

(5) **Explanation of individual changes** in income as predictors e.g. previous income, age, changes of states in domains. We present only a few results from these standpoints; see also Laaksonen 1989.

2. THE 1984-86 PANEL DATA AND PRINCIPAL OBJECTS OF EVALUATION

The nature of the 1984-86 panel in the IDS as a part of other panels may be appreciated from the following diagram:



Handling of panel and cross-sectional data

When operating with data files, we benefit in practice by sampling weights. Since we have different data sets, there must be different weights depending on the material needed. Denote sampling weights in general by $w_k^{a(b,c)}$ (for their derivation, see Section 2), where

k refers to a household
a is the year of the statistics
b is the starting year of the panel
c is the last year used for information in the panel.

The weight $w_k^t(t,t)$, for instance, will then refer to a phase when only the first panel year has gone by, whereas the weight $w_k^t(t,t+1)$ indicates that we have information on the second year,

too, both weights being as estimated for the statistical year t. If there is no attrition, the weights will be equal.

Thus panel studies, concerning the same observation units allow two alternatives for forming the weights, using:

(1) weights from the first year,

$$w_k^t(t,t+1), \text{ or}$$

(2) weights from the second year,

$$w_k^{t+1}(t,t+1).$$

A weight of type (1) corresponds in principle to a case of Laspeyres' index and weight (2) to Paasches' index. These give different results when measuring panel changes. The approach of Laspeyres may often be more natural because it illustrates changes which have affected the population of the first panel year. Section 5 will discuss results based on both approaches. In addition, we can form a combination of both weights. Usable combinations are based on appropriate averages, and: in many cases these are recommendable, cf. the results of Stadt and Wansbeek (1990).

For cross-sectional studies, e.g. for t, special weights will be needed. Combined sampling weights of this kind are formed in the IDS in the manner

$$w_k^t = \frac{w_k^t(t,t) + w_k^t(t-1,t)}{2}, \qquad (1.1)$$

in which case both parts of the panel are assigned the same weighting. This represents a simple application of the 'composite estimator'. The estimator (1.1) obviously does not take advantage of any panel characteristic, except that when we generate two different weights based on independent samples for the same year we can compare them and evaluate their quality. Instead, the case can be improved when measuring changes between cross-sectional years, if a positive correlation exists between observations, as the following formula indicates for the variance in average changes (see Duncan and Kalton 1985):

$$V(\bar{y}_{t+1}-\bar{y}_t)=V(\bar{y}_t)+V(\bar{y}_{t+1})-2\mathrm{cor}(y_t,y_{t+1})\sqrt{V(\bar{y}_t)V(\bar{y}_{t+1})}, \qquad (1.2)$$

in which $\bar{y}_t$ is the mean of the values of y in t and cor(.,.) the coefficient of correlation between successive panel observations. The accuracy of the changes will improve if the correlations are large. These were approximately 0.8 for the two-year panels for disposable income and other similar quantities in the IDS and 0.9 for the one-year panels.

The advantages are not so significant for cross-sectional changes because these always consist of two independent parts, i.e. cor=0. The **proportion of overlap** (see Kish 1987), which is now approximately 0.5, is also substantial. The final advantage is reduced if the overlapping decreases, of course. Furthermore the advantages are dependent on other things, e.g. nonresponse and other missingness and changes of variance in duration.

An approximation for the advantages of the panel approach for measuring cross-sectional changes may be presented following Nieuwenbroek (1990), who derives the factor 'deff' which interprets the number by which the sample size of the rotating panel has to be multiplied in order to obtain the same variance as with the repeated cross-sectional design:

$$\text{deff} = \frac{1-p+pv(1-r)}{(1-p+pv)^2} \qquad (1.3)$$

in which p = proportion of overlap, 1-v = attrition rate in the second year of the panel, cor = correlation as in (1.2). The deff values for disposable income in the IDS panel were approximately 0.6-0.7, so that the advantages of panels are significant in this sense.

Quite similar approaches to panel estimation can be found in the theory behind more complicated composite estimators as above (see e.g. Cochran 1977). SIPP of the U.S. Bureau of the Census (see Section 1) uses an estimator which would be analogously applicable to the Finnish IDS. The estimator, named the Ernst-Breau **composite estimator** (Chakrabarty 1989), can be expressed in a recursive form as

$$y_t^c = (1-A_t)\, y_t(t,t+1) + A_t\, y_t(t-1,t) + Z\,(y_{t-1}^c - y_{t-1}(t,t+1)),$$

(1.4)

in which

y_t^c = composite estimator of y in t (of the type average)

$Z = \left(1-(1-r^2)^{1/2}\right)/r$

$A_t = 1/(2-rZ)$

$y_t(t,t+1)$ = simple estimate for t based on 50% of the sample when the panel starts at t and ends at (t+1); the weights $w_k^t(t,t+1)$ are used to estimate this.

This estimator presumes that there is no nonresponse or that its effects have been eliminated. Chakrabarty (pp. 6-7) also presents the variances of y_t and $(y_{t+1} - y_t)$. In the IDS panels, where the correlation coefficient for disposable income is 0.8, the former variance obtains the value $0.58s^2$ where s^2 is the simple or uncomposite variance, and the latter $1.22s^2$. In that case the simple variance is $2s^2$, by formula (1.2), if r=0 and the composite variance is 39% lower, and, thus the results are better.

An undesirable problem in composite estimates is their inconsistency if also applied to subpopulations: the sum or average of the subpopulation composite estimates does not have to be equal to the overall composite estimate. Moreover, other values may again be obtained for another classification criterion.

Panels in the IDS

Although it is thus evident that compensation for nonresponse can also be of value for the cross-sectional analysis of data, we are concerned here principally with the panel part, and more precisely with the panel for 1984-86. The data structure for this panel is presented in the following diagram (sizes of subsamples in the right margin):

1984	1986	Size
Overcoverage		70
Nonresponse	Overcoverage	95
Nonresponse		1379
RESPONSE	Nonresponse	221
RESPONSE		5745
RESPONSE	Overcoverage	98
	ALL	7608

The initial sample of 7608 developed a problem of overcoverage at once, in that 70 key persons had died or emigrated in the first year. By the second data year this overcoverage had increased to 193, so that the viable sample in 1986 comprised in practice 7345 households.

It is possible to form three bodies of data for examining the 1984-86 panel:

A. Those who replied in both years - sample size 5745.

B. The respondents in 1984, excluding those not contained in the sampling frame for 1986, i.e. without any overcoverage - sample size 5966.

C. All those included in the sampling frame in both years, regardless of whether they replied or not - sample size 7345.

Set A may also be called the set of **complete cases**, set B the set of **first-order completed cases** and set C the set of **second-order completed cases**. In computing cross-sectional results by Formula (1.1), for example, the weights and data for the statistics are derived from the **available cases**; where all the respondents are included in the files independently of forthcoming events.

Separate panel-type surveys may be developed further from the set which responded in the first year but belonged to the overcoverage cases in the second year as: it is interesting, for instance, to consider their income before death or emigration. These considerations are excluded from the present report, but that adjustments for nonresponse as in Section 3 would also be necessary in this sense, because future overcoverage cases already show a tendency to move over to nonresponse cases in the previous year.

Any one of the data sets A, B or C could be used for panel analysis. The easiest and most typical method is that also used in an earlier analysis of the similar material, aimed at describing changes in household structure, income and the effects of structural changes on income, which was based on a material of complete cases, type A (Laaksonen 1989). Note also the discussion at the beginning of this section concerning the use of sampling weights in real panel analyses.

It is interesting here to look into the suitability of data sets B and C for panel analysis, but this requires adjustment for the nonresponse bias, as the necessary income figures are missing in these cases. Use can be made of information from population, taxation and education registers etc. for this purpose.

Use of such registers requires prior determination of the composition of the households concerned, in order to obtain register data for the right persons in each household (cf. Laaksonen 1988). This always succeeds excellently for interviewed households, but for nonresponse cases we can only partly specify the household members. All those who have died can be found in registers, but grown up children with households of their own are a serious problem, for example. Panel material presents greater problems in this respect, especially in the latter year, and no high standard of excellence could be achieved in the present case, as nothing had been done to take these needs into account in good

time, i.e. when gathering the data. These experiences will enable us to allow for such problems in future, and thus obtain more useful results.

In addition the actual data were formed into a set of simulated data based on the initial 5745 households interviewed. This was then extended to form a sampling frame of just under 18000 individual persons. Nonresponse and overcoverage data corresponding as closely as possible to the real situation were included in this material, and this basic set was then used in a realistic manner, taking samples from it for the calculation of given indices, etc. The results obtained in this way can thus be compared with the actual data. As the material is relatively restricted, it could not be made to contain the same number of classifications as the real data, and it is thus of limited value for testing purposes.

Nonresponse in the first year of the survey amounted to 1474 households, or 19.3%, and this increased by 221 in the second year, although as 95 of the previous nonresponse cases had been transferred to the overcoverage category, the actual nonresponse was 1600, corresponding to 21.8% of that year's usable sample, or 21.0% of the original sample. With total overcoverage amounting to 3.5%, the proportion of households answering in both years was 75.4%.

General approaches to handling data in cases of nonresponse

Following Rubin (1987) and Little and Rubin (1987) the whole sampling process in question can be presented more exactly using the terms sampling mechanism and response mechanism, and their ignorability or nonignorability. Denote (the points of time can also be stated for each variable):

Y = a matrix or a group of outcome variables.
Some observations for these variables are missing.
Let Y_{obs} be the observed values and Y_{mis} the missing values. Then $Y_{inc} = (Y_{obs}, Y_{mis})$. Further Y_{exc} are unobserved values and $Y_{nob} = (Y_{exc}, Y_{mis})$.

X = covariates or auxiliary variables. These variables are completely observed and can also be used in sample

selection.

S = sampling selection indicators, such that $S_k = 1$ if a household k is included in the sample or $S_k = 0$ if it is excluded. Further $S_{inc} = \{k \mid S_k = 1\}$.

R = response indicators, such that $R_{kj} = 1$ if a household k responds to a variable j $(j=1,\ldots,p)$ or $R_{kj} = 0$ if it does not respond. Further $R_{obs} = \{(k,j) \mid R_{kj} = 1\}$.

The specification for $Pr(S \mid X,Y,R)$ is then the sampling mechanism, so that $Pr(. \mid .)$ refers to conditional probability density in context, while the specification for $Pr(R \mid X,Y)$ is the response mechanism. The sampling mechanism is now said to be ignorable if $Pr(S \mid X,Y,R)$ depends only on observed values (Y_{obs},R_{obs},X); otherwise it is nonignorable.

A sufficient condition for ignorability of the response mechanism, when the sampling mechanism is ignorable at (X,Y_{obs},R_{obs},S) is that

$$Pr(R_{obs} \mid X,Y) = Pr(R_{obs} \mid X,Y_{obs}) \qquad (1.3)$$

Formula (1.3) indicates that the distribution of R, given Y and X, does not depend on the unobserved values Y_{exc} and Y_{mis}. This result can be extended to provide a joined definition of ignorable sampling and response mechanisms as follows (Rubin 1987, p. 53-54):

$$Pr(Y_{nob} \mid X,Y_{obs},R_{obs},S) = Pr(Y_{nob} \mid X,Y_{obs}). \qquad (1.4)$$

Such being the case, the conditional probability distribution of unobserved values of Y does not depend on either response and sample selection.

In the cases in question our aim was to select samples so that the sampling mechanism was ignorable to a reasonable extent, but only in the first panel year, of course. The second year is different because in our case all the nonrespondents in the first year are also nonrespondents in the second, i.e. R=0 in t implies R=0 in (t+1). We will consider these matters in Section 3.

Ignorability of the response mechanism is a larger problem, however, and a major one to be discussed here.

It may be deduced that adjustment for nonresponse becomes more essential, the more nonignorable is the response mechanism. Thus it is appropriate to employ methods of adjustment to try to develop mechanisms or filters which attempt to change nonignorable mechanisms to ignorable ones as far as possible. This principle is a major aim of this report, and Section 3 discusses methods aimed at achieving changes of this kind within 'adjusted' subclasses or cells, leaving out to specific variables of Y. Section 4 is concerned with methods which have similar aims with respect to the characteristics of these specific variables.

3. Sampling and Adjustments by Reweighting

We are concerned here with differences in estimation procedures by means of sampling weights. Weights of these kinds are discussed in the present section. The first is based on the assumption that the selection mechanism is ignorable and all the households respond, or else the response mechanism is also ignorable. In this case the coefficients belong to the family of Horvitz-Thompson (HT) estimators (originally Horvitz and Thompson 1952). When a coefficient is revised, we obtain two adjusted coefficients, one based only on post-stratification and the other on modeling of response probability as well. The latter is described as 'a model based Horvitz-Thompson estimator' by Ekholm and Laaksonen (1990).

'Basic' sampling weights

The process of defining a basic (HT) sampling weight starts out from the sample selection method employed for the IDS, which differs from that used in the Household Budget Survey of 1985 (Laaksonen 1988) on one major point: that all the members of the household were assigned to the same regional stratum in the sampling frame for the Household Budget Survey, whereas in the IDS they could be in two or more successive strata. We denote these as 'pre-strata' as distinct from 'post-strata', which are also used in this report.

The 1983 tax register data were used to form 12 pre-strata for the 1984 sample, each household member aged over 15 years being placed in a pre-stratum in accordance with her or his taxable income and tax classification. This was done in order to improve the accuracy of the data regarding persons with a high income and self-employed persons, as these were assigned higher selection probabilities. The probabilities varied in the range 0.01 - 0.001.

Formation of the sampling weights proceeds as follows.

Denote

k = indicator of a specific key individual (person) or household. Households are determined by key individuals.

a_k = key individual in household k

b_{ki} = ith member of a household k belonging to the sampling frame

h = pre-stratum

$S = \{k \mid S_k(83) = S_k(84) = S_k(86) = 1;$ in short $S_k=1\}$, where $S_k(83)$ illustrates units selected in the sample in 1983, $S_k(84)$ units selected in 1983 belonging to the sampling frame in 1984, $S_k(86)$ correspondingly

n_h = number of key persons and households selected in pre-stratum h

$R = \{(k,j) \mid S_k = 1, R_k = 1\}$ is set of respondents in which R_k is a response indicator. By necessity, if $R_k = O$ in 1984 then $R_k = 0$ in 1986. Since the system lays down that the whole household must answer, each of its members must belong to either the set R or the set S-R, the nonrespondents.

r_h = number of key persons and households responding out of pre-stratum h

M_h = number of persons in the sampling frame at the end of the data year belonging to pre-stratum h. As only persons over 15 years of age are assigned to a pre-stratum, households contain members who do not belong to the sampling frame

m_{kh} = number of frame persons belonging to pre-stratum h in a household k. It follows that $\sum_k m_{kh} = M_h$.

The principle of sampling selection is that the first to be chosen are the key persons, and then the other members, who live in the same dwelling. Selection thus operates in terms of individuals, although the results are stated at the household level. Sampling proceeds without replacement, of course. The mechanism may be considered analogous to simple random sampling within pre-strata, although it is performed equidistantly, because the order of

individuals in the frame is approximately random.

The selection probabilities are determined according to the probabilities assigned to the key individuals in the households. If there is no nonresponse, i.e. $r_h = n_h$, then assuming that sampling takes place with replacement,

$$\Pr(a_k \in S_h) = \Pr(a_k \in R_h) = n_h/M_h = p_h. \qquad (3.1)$$

In the case of sampling without replacement, as now, the result (3.1) is approximative if the number of frame members in that household is more than one. The same values are also given to the other members in the same pre-stratum who fall into the sampling frame

$$\Pr(b_{ki} \in S_h) = p_h.$$

The selection probability of the ith member of a household k in pre-stratum h is more generally as follows:

$$p_{khi} = \begin{cases} p_1 & \text{if } b_{ki} \in S_1 \\ \vdots & \\ p_h & \text{if } b_{ki} \in S_h \\ \vdots & \\ p_{12} & \text{if } b_{ki} \in S_{12}. \end{cases}$$

Note that persons not belonging to the sampling frame (small children in the IDS) in a household k have a probability of zero.

If all the members of a household k in the sampling frame are in the same pre-stratum h, the probability k of being selected is the same as in the Household Budget Survey of 1985, i.e.

$$\pi_{kh} = \frac{m_{kh} n_h}{M_h} = m_{kh} p_h = \sum_{i=1}^{m_{kh}} p_{khi} \qquad (3.2)$$

The last form of Formula (3.1) indicates that the selection probability attached to a household k is the sum of the probabilities attached to those of its members included in the

sampling frame. This result does not hold good when the members are in different pre-strata, however, when the summation principle is the following. If the members of a household fall into two pre-strata, e.g. h_1 and h_2, then

$$\pi_k = \pi_{kh_1} + \pi_{kh_2} - \pi_{kh_1}\pi_{kh_2} \qquad (3.3)$$

Thus the probability of selection is the sum of the probabilities attached to the members less the product of the probabilities attached to their pre-strata, because the pre-strata probabilities are mutually independent.

Correspondingly, if the members fall into a number of pre-strata, the equation consists of the sum of the selection probabilities of the members corrected by a series of subtracted or added terms, of which the first are the subtracted products of the paired stratum probabilities, followed by addition of the products of triplets of these probabilities, and so on. In practice, however, the members of one household rarely belong to more than three pre-strata as the aim is that a married couple should always belong to the same group.

The product of two probabilities in the IDS is of the order of $0.01 \bullet 0.01 m_{kh} = 0.0001 m_{kh}$ at most, i.e. small by comparison with the pre-stratum probabilities, and the products of triplets of these will be even smaller. Consequently it is unnecessary to take these factors into account when determining selection probabilities for whole households, which in turn means that the selection probabilities used are slightly too large when the number of household members belonging to the sampling frame is greater than one. This error is more significant in the case of large households, but the sample contains very few of these.

The selection probability attached to a given household may thus be estimated reasonably accurately by means of the simplified formula

$$\pi_k = \sum_h \sum_i p_{khi} \qquad (3.4)$$

which has been used in producing the statistics.

The approximate sampling weight is the reciprocal of the selection probability for the household k

$$w_k = \frac{1}{\pi_k} = \frac{1}{\sum_h \sum_i p_{khi}} = \frac{1}{\sum_h \frac{n_h \, m_{kh}}{M_h}} . \qquad (3.5)$$

Since the selection probability is slightly too high when $m_{kh}>1$ or the frame members in a household k belong to different pre-strata, the sampling weight is correspondingly slightly too low. If nonresponse exists but can be ignored, the result is obtained by setting $n_h = r_h$

$$w_k^r = \frac{1}{\sum_h \frac{r_h \, m_{kh}}{M_h}} = \frac{1}{\sum_h \frac{n_h \, m_{kh}}{M_h} \bullet \frac{r_h}{n_h}}$$

Further, it holds true within each pre-stratum that

$$w_k^r = w_k \cdot \frac{n_h}{r_h} \qquad (3.6)$$

The basic HT sampling weights for both years of the panel are determined in the same manner, and can be denoted $w_k^{t}{}_{(t,t+1)}$ for the first year and $w_k^{t+1}{}_{(t,t+1)}$ for the second year (cf. the notations in Formula (1.1)). Each individual item M_h, r_h and m_{kh} (see Formula (3.5)) then has to be specified differently for each year, of course. Quite a number of changes do in fact occur, as indicated by the results of Laaksonen (1989), the composition of the household being found to have altered in 10-20% of cases. Note also that the population data for the sampling frame have to be redefined from the register, and the minimum age has to be raised, from 15 to 16 years in a panel of duration one year, and from 15 to 17 years in one of two years' duration.

Adjustment for nonresponse by reweighting

It is stated in Section 2 that if a response mechanism is nonignorable there is a need for its adjustment. Since, post-stratification would be applicable in this case, we will try to develop a new level of classification, i.e. of post-strata, for Formula (3.5). Let denote the post-strata by d ($_{d=1,...D}$). The sampling weights in cases of post-stratification can then be presented within each post-stratum d as follows

$$w_k^d = \frac{1}{\sum_h \frac{r_{dh}\, m_{kdh}}{M_{dh}}} \qquad (3.7)$$

in which M_{dh} is the number of frame persons belonging to the post-stratum d and the pre-stratum h, and correspondingly for r_{dh} and m_{kdh}. The following scheme illustrates this case:

Post-strata													
1					2				...	d			
Pre-strata					Pre-strata					Pre-strata			
1	...	h	...	12	1	...	h	...	12	1	...	h	
M_{11}		M_{1h}					M_{2h}					M_{dh}	
r_{11}		r_{1h}					r_{2h}					r_{dh}	
m_{k11}		m_{k1h}					m_{k2h}					m_{kdh}	

Since the whole household lies within the same post-stratum, $m_{kdh} = m_{kh}$. In order for the nonadjusted and post-stratified weights to be the same, the ratios r_{dh}/M_{dh} and r_h/M_h must be the same for any d. Developing Formula (3.7) this standpoint may be presented within each pre-stratum of each post-stratum as follows:

$$w_k^d = w_k^r \frac{r_h}{M_h} \cdot \frac{M_{dh}}{r_{dh}} = w_k \frac{n_h}{M_h} \cdot \frac{M_{dh}}{r_{dh}}. \qquad (3.8)$$

We will prefer post-stratification if a response mechanism is better ignorable within pre-strata of post-strata than within initial pre-strata. More generally, response mechanisms are better ignorable if the strata, e.g. the post-strata, are as homogeneous as possible in terms of respondents and outcome variables. In the present case this is not due to information always being available to the maximum extent possible, because the terms M_{dh} have to be from the population level and data of this kind are available to a limited extent, the best information being areal.

Post-stratification was also used at the provincial level (24 regions from provinces and their division by level of urbanization) with regard to the 1984 and 1986 data sets, but the method did not give adequate results because the response within post-strata is still fairly nonignorable, as we have seen from the Household Budget Survey of 1985 (Laaksonen 1988, Ekholm and Laaksonen 1990). It follows that we must look for other methods, other classifications within which the response mechanism will be more ignorable. For this purpose we tested a method with similar ideas to those in the paper mentioned but concerning different data and a different sampling method, and tried to exploit the panel characteristics.

This method also starts to yield a new type of division in the data, called cells and denoted by a subscript *c*, forming the set C. The aim is to find a composition of cells such that the response mechanism related to the response probability is ignorable. This method uses modeling for response, the appropriate register variables being used as explanatory variables. The modeling is based on a method of logistic regression with categorial variables. We obtain as our result the estimated response probabilities, denoted by $\hat{p}_c$, which are introduced into new weights w_k^{dc} as follows:

$$w_k^{dc} = w_k^d \cdot \frac{r_{dh}}{n_{dh}} \cdot \frac{1}{\hat{p}_c}, \qquad (3.8)$$

which may be further developed in each pre-stratum of each post-stratum, because the cells are independent of pre-strata and post-strata (cf. Ekholm and Laaksonen 1990), as follows:

$$w_k^{dc} = w_k^r \frac{r_h}{M_h} \cdot \frac{M_{dh}}{n_{dh}} \cdot \frac{1}{\hat{p}_c}, \text{ and} \qquad (3.9)$$

$$= w_k \frac{n_h}{M_h} \cdot \frac{M_{dh}}{n_{dh}} \cdot \frac{1}{\hat{p}_c}. \qquad (3.10)$$

The first ratio in Formula (3.9) illustrates the proportions of respondents in each pre-stratum, the second the inverses of the selection probabilities for key persons in the pre-strata of post-strata, and the third is the inverses of estimated response probabilities in the cells c.

The first two ratios of Formula (3.10) illustrate the effects of post-stratification and the last one is the same as in Formula (3.9). We see, for instance, that if the first ones are inverses of each other, Formula (3.10) is analogous to Formula (3.5). If there are no nonrespondents, it holds that

$$w_k^{dc} = w_k.$$

Correspondingly, if the probabilities $\hat{p}_c$ are the same as the proportions of the respondents in the post-strata, then

$$w_k^{dc} = w_k^d.$$

The goodness of an estimator, say Q, depends on its bias. In this case the structure of the bias B can be presented, after Little (1986), for instance, in terms of two factors

$$B(Q) = \sum_c P_c(Q_{cP}-Q_c) + \sum_c (w_k^{dc} - P_c)Q_{cP}, \qquad (3.11)$$

where

P_c = Pr($k \in C$) in the whole population

Q_{cP} = value of estimator in a cell c of the whole population

Q_c = value of estimator in a cell of the set of respondents.

The first factor of Equation (3.11) indicates that the result becomes better the closer to one other the values for respondents and other households within cells are. The second factor illustrates the same ideas concerning adjusted sampling weights on the one hand and population weights on the other. If these factors are valid, we can also say that the response mechanism is ignorable with respect to (i) the distribution of outcome variables and (ii) the distribution of respondents. The goodness of these factors cannot be proved explicitly unless we have complete information on the levels of population and respondents, which is not the case in practice.

Something can be seen by comparing sampled and responded households, assuming that the sampling mechanism was reasonably ignorable. The ignorability of the second factor can be improved by modeling the response probabilities as well as possible, aiming at the case in which there are no systematic components within the cells, so that the response probabilities are constants. It is possible that the bias of the first factor will similarly be reduced, but this is not certain, because many outcome variables have specific properties which cannot be introduced in a common adjustment with weights. One can then try to find specific weighting adjustments or use similar techniques to those presented in Section 4 concerning imputation methods.

Models for response probabilities

In the present panel case (see Section 2) we have three opportunities to build models for response probabilities, concerning response vs. nonresponse for

(1) the first year,
(2) the second year, and
(3) the first and second years cumulatively.

Choice of the explanatory variables was based on the one hand on the results of the Household Budget Survey, and on the other hand on exploiting the panel characteristics. The data and some of their qualities were different, of course. The following explanatory variables were included in the eventual models (the same symbols are also used in the tables and figures):

Region, divided into four classes:

Region 1: province of Uusimaa (Southern Finland, incl. the Helsinki Metropolitan Area)
Region 2: provinces of Turku and Pori, Häme and Kymi and the Aland Islands (Western Finland)
Region 3: provinces of Mikkeli, Northern Karelia, Kuopio and Central Finland (Eastern Finland)
Region 4: provinces of Vaasa, Oulu and Lapland (Northern Finland)

Status, comprising three classes derived from the taxation data:

1: wage and salary earners
2: farmers
3: others, incl. self-employed, pensioners and students.

Family structure, comprising 7 classes:

10: single person households
20: two adults with no children
21: one or two adults with at least one child
30: three adults with no children
31: three adults with at least one child
40: four or more adults with no children
41: four or more adults and also some children.

Change in **family structure**, comprising 2 classes:

0: unchanged from 1984 to 1986
1: changed.

Change in family structure was not, a useful variable as such, however, because it was interactive with family structure. Several trials were conducted with no very outstanding results, the most suitable combination being one which took account of both variables except for family structures 40 and 41, but the changes in structure were insignificant in these cases. We call this variable CHANGE. The results obtained from the models as defined for the sampling years are presented in Table 1.

Table 1. Basic results obtained from the response probability models using logistic regression for the panel 1984-86.
The values of the test statistics are based on entering the present variable into the model last.
(*) in parenthesis the years of the data for the explanatory variables in the order in which they appear in the columns.
For further explanations, see the text.

Subsample of response vs. nonresponse (*)	Explanatory variables and test statistics								
	REGION		STATUS		FAMILY STR.		CHANGE		Prob. of likelihood ratio
	Chi-square	Prob	Chi-square	Prob	Chi-square	Prob	Chi-square	Prob	
a) 1st year (84,84,84)	44	.0001	16	.0003	676	.0001	-		.128
b) 2nd year (84,84,84)	18	.0005	6	.0451	10	.1180	-		.235
c) 2nd year (84,86,86)	18	.0044	9	.0218	45	.0001	-		.284
d) 2nd year (84,84,84-86)	16	.0010	3	.1154			160	.0001	.988
e) 1st and 2nd year (84,84,84)	59	.0001	21	.0001	625	.0001	-		.225
f) 1st and 2nd year (84,84,84-86)	51	.0001	20	.0001	-		600	.0001	.115

The models for the first year and the two years combined give similar results, as nonresponse in the second year was small compared with that in the first. In both cases family structure is

the main explanatory variable, region the second most important and status the third. Some results are presented in more detail in Figures 1 and 2, which show that the response was very much poorer in Southern Finland, for example, than in the country as a whole, that the farmers answered better than did the wage and salary earners and that the 'other' status group had the lowest response rate of all. Correspondingly, persons living alone answered much worse than the others, while the presence of children in a household was associated with response.

Responses in the second year were modeled on the basis of family structure and status for both 1984 and 1986. Region was clearly a significant factor given the 1984 data, status marginally significant but family structure not significant, whereas the model based on the 1986 data has family structure as a clear explanatory variable. Some results are presented in more detail in Figures 3 and 4. The results based on the 1986 data resemble the profiles in Figures 1 and 2 as far as the least frequent respondents are concerned, i.e. those living on their own and the large households with no children, whereas the large households with no children do not stand out markedly from the others in terms of the 1984 data. It thus seems that the enthusiasm for answering questionnaires declines noticeably at the point where the youngest children in the household reach adult age, perhaps because the loosening of family ties makes it more difficult to gather information regarding the household, with the children beginning to leave home and establish households of their own.

Table 1 also contains two models, change in family structure being one explanatory variable. These express the point more explicitly. Unfortunately, the data are not perfect concerning nonrespondents in both years, and the results are therefore slightly uncertain as regards practical situations. They nevertheless indicate interesting tendencies, in that a change in family structure is fatal from the response point of view. The worst nonresponse rates appear for original single households which had combined and for households with two adults in 1984 who had separated by 1986.

The variable STATUS, on the contrary, is no longer essential for nonresponse. We tried to look for other explanatory variables, but we did not find any, although change of dwelling had an

influence on the nonresponse rate to a limited extent, mostly concerning combined singles.

The likelihood ratio for each of the models in Table 1 is above 0.05, i.e. it is possible to employ an additive model, which means that it is not necessary to include interaction terms for the explanatory variables in these models. This ratio is in a certain sense too high for Model d), which is partly derived from the uncertain data mentioned. Because of this we did not use this model in the applications in Section 5, in which we return to the estimates for the numbers of households and their income data as given by the response probability model.

No discussion of variance estimates using adjustment weights is entered into in this paper, but analytical variances for the totals and means can be obtained in a manner analogous to that presented earlier (see Laaksonen 1988: 49-51; Ekholm & Laaksonen 1990). Variances for specialized quantities such as deciles can be estimated by bootstrap techniques, an example of which is given in Appendix 1.

Fig. 1. Estimated response probabilities in the second year (symbols in the text)

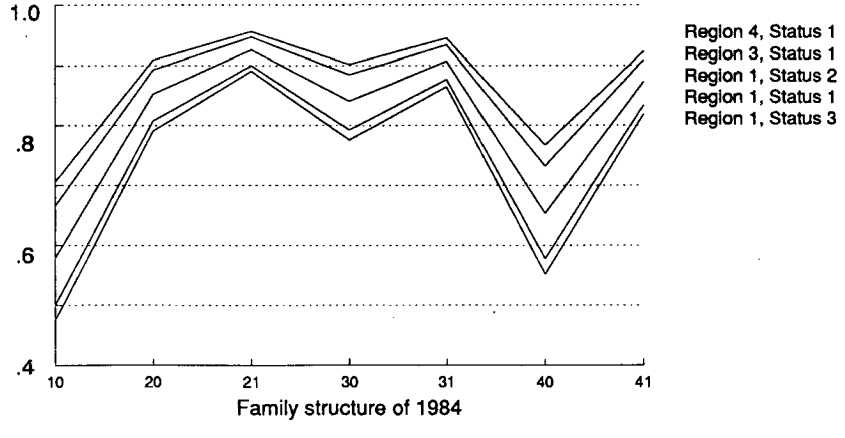


Fig. 2. Estimated response probabilities in both years

(symbols in the text)

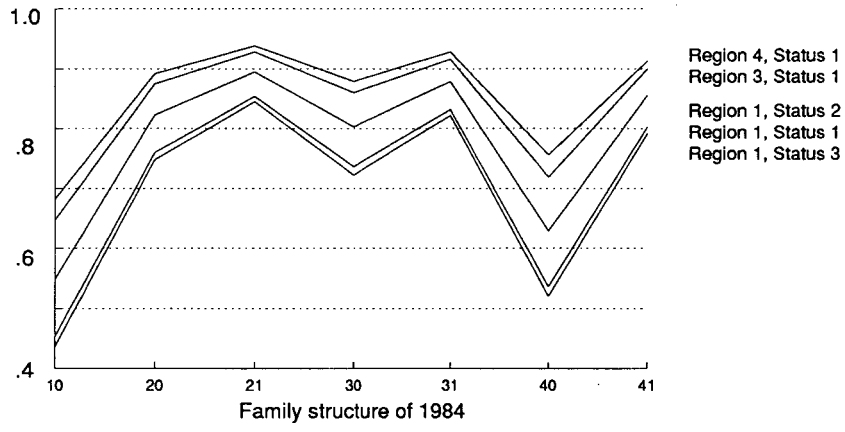


Fig. 3. Estimated response probablities in the second year

(symbols in the text)

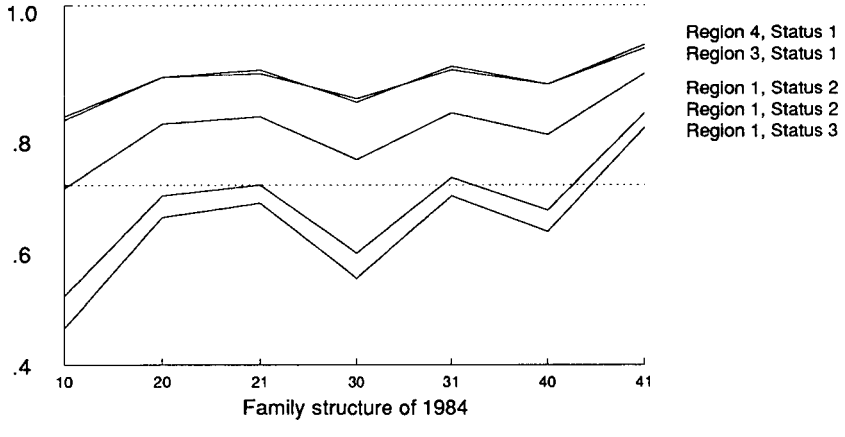


Fig. 4. Estimated response probabilities in the second year (symbols in the text)

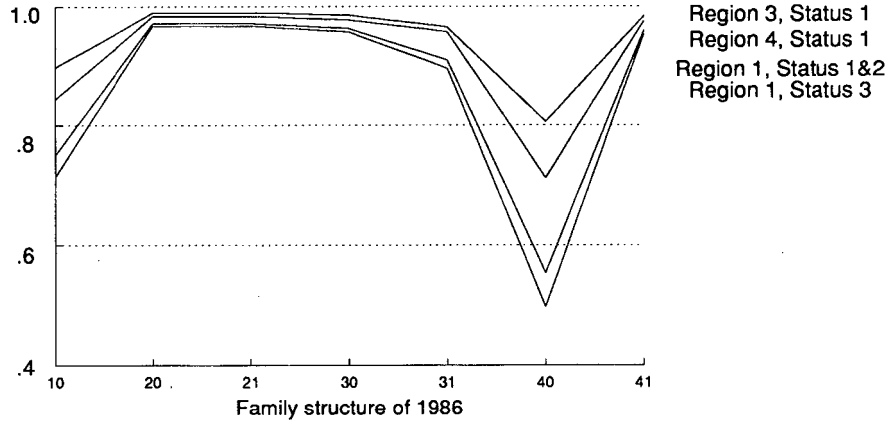
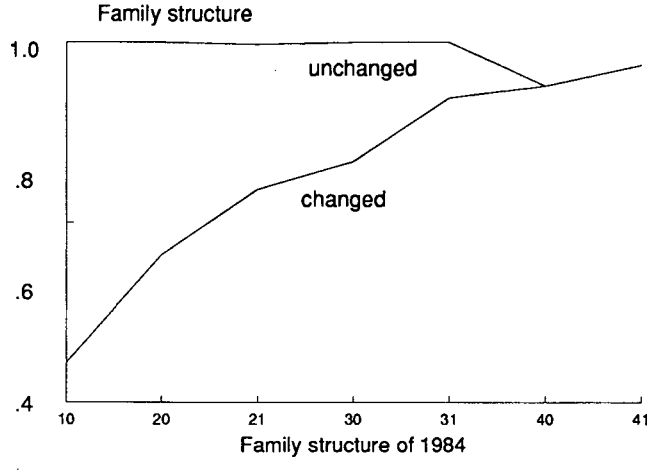


Fig. 5. Estimated response probabilities in the second year in Status 1 of Region 1 when family structure can be changed (symbols in the text)



4. SINGLE AND MULTIPLE IMPUTATION

A standard technique for handling item nonresponse in a survey is imputation. Its aim is to replace missing values with other suitable ones. Where each missing value is replaced with one imputed value we speak of single imputation, while a technique in which each is replaced with more than one value is known as multiple imputation.

Numerous imputation methods have been proposed in the literature (see Kalton & Kasprzyk 1986, Little & Rubin 1987, Laaksonen 1988). Methods can be divided into the following five groups: (1) deductive imputation, (2) mean imputation, (3) random imputation, (4) hot deck imputation, also including gold decking, and (5) imputation using modeling, e.g. ordinary, weighted and censored regression, MANOVA and log-linear models. All imputation methods, except sometimes deductive imputation, will fabricate data to some extent. The extent of fabrication depends on how well the imputation model predicts the missing values.

When choosing the appropriate method it is necessary to take account of the nature of the variable to be imputed. We will discuss here a situation in which the variable for which some observations are missing is expressed on a relative scale and in which even the smallest observations are clearly positive. Depending on the mode of generation of the statistics and possible errors, the figures for disposable income can also be negative, but such eventualities are excluded from the present report.

Delimitation of the type of variable in this way both facilitates and complicates the choice of an appropriate method. The situation is different in a material such as that obtained in the Household Budget Survey, in which there are many variables for which the majority of the values are zero. Since methods have to be chosen with a view to their usefulness for the generation of statistics, it would seem that those to be considered here are either imputations based on regression or some other linear model and hot deck imputation.

Single regression imputation

Imputations based on regression or other corresponding models operate by constructing explanatory models for the variable to be imputed which provide the best possible fit with the response material and have as their predictors variables for which values are also available with respect to the nonrespondents (for more details, see the application to the 1985 Household Budget Survey in Laaksonen (1988: 61-71)). The weightings used in the modeling are either nonadjusted or better still reweighted sampling weights.

In the case of single imputation, the predicted values obtained from the model are inserted in place of the missing ones. In view of the nature of regression models, this can be expected to bring the results closer to the average values and reduce the variance, provided that the fit of the model is good. If a bad fit is obtained, the whole imputation in the context of that model will be suspect. On the other hand, the consequence of a good fit may be that the interval estimates become too sharp. The main reason for this is that the missing values are treated during the analysis as if they were known, and thus the extra variablity due to imputing the missing values is ignored.

In order to avoid these problems, Rubin and his associates (see Herzog & Rubin 1983, Rubin 1987, Rubin & Schenker 1987) propose the use of multiple imputation, in which each missing observation is replaced by two or more estimates (about 5 new observations is usually a suitable number). The outcome is a more extensive body of data which can be processed in the same manner as the basic set.

Multiple regression imputation

Consider the variables $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} is observed and Y_{mis} is missing due to nonresponse. Let X denote the other variables in the survey, which are observed for all units (in this

case, see Section 2).

Suppose that Q is a scalar or vector-valued quantity to be estimated and that if there were no nonresponse, then

$$Q - \hat{Q} \cong N(0,U),$$

where $\hat{Q}= \hat{Q}(X,Y)$ and $U=U(X,Y)$ are complete-data statistics giving an estimate for Q and for the variance of $(Q-\hat{Q})$. Rubin (1987, Ch. 3) shows that in the presence of nonresponse, it holds good approximately that Q is normally distributed with a mean

$$E(Q|X,Y_{obs}) = E(\hat{Q}|X,Y_{obs}) \qquad (4.1)$$

and a variance

$$V(Q|X,Y_{obs}) = E(U|X,Y_{obs}) + V(\hat{Q}|X,Y_{obs}) \qquad (4.2)$$

Equation (4.1) implies that Q is estimated by means of the expected value of the complete-data estimate over the distribution of Y_{mis}. The first term of (4.2) is the expected value for the complete-data variance estimate of U and the second term the variance of the complete-data estimate, the former representing the sampling variance and the latter the imputation variance.

The basic idea of multiple imputation is to draw several, say l=1,...,L, independent sets of imputations for the missing data Y_{mis}, from their posterior distribution, i.e. the imputation model. The result is L completed data sets and hence L sets of complete-data statistics, say $(\hat{Q}_{*1},U_{*1}),...,(\hat{Q}_{*L},U_{*L})$. The L sets of statistics are then combined as follows. Let

$$\bar{Q}_L = \sum_l \hat{Q}_{*l}/L \qquad (4.3)$$

be the average of the L complete-data estimates,

$$\bar{U}_L = \sum_l U_{*l}/L \qquad (4.4)$$

be the variance of the L complete-data variance, i.e. the within-variance of the imputations, and

$$B_L = \sum_1 (\hat{Q}_{*l}-Q_L)^2/(L-1) \qquad (4.5)$$

be the variance of the L complete-data estimates, i.e. the between-variance of the imputations. The total variance of $(Q-\hat{Q}_L)$ is given by

$$T_L = \bar{U}_L + (1+L^{-1})B_L \qquad (4.6)$$

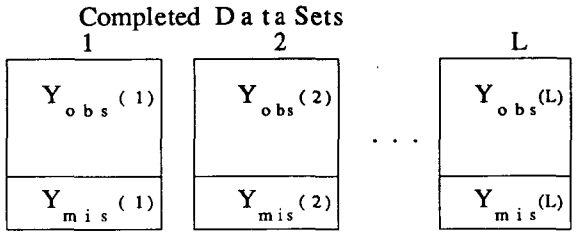
As L approaches infinity, $\hat{Q}_L$ approaches the first term of equation (4.2), $\bar{U}_L$ its second term, and B_L their sum. Thus T_L is the approximative variance of $(Q-\hat{Q})$. More than a few multiple imputations of this kind would become unwieldy, so it is most useful to keep L small, say L<10. To adjust for small L, the factor $(1+L^{-1})$ is included in (4.6). We see that if the number of imputations L=1, the between-variance will be 0, and the interval estimates will be narrower than in the case of multiple imputation.

Rubin (1987, p. 122) points out that each of the L draws from the posterior distribution of Y_{mis} can be simulated in a two-step process. First, the parameter vector of the imputation model is drawn from its posterior distribution; then Y_{mis} is drawn conditionally upon the value drawn for the parameter vector. If this two-step process is followed using appropriate models for the data and the nonresponse, also reflecting parameter uncertainty, the imputation method is **proper**. If the parameter vector is fixed at an estimated value across the L imputations, rather than being drawn from its posterior distribution for each imputation, the variability due to estimating the parameter vector is not reflected and the imputation method is **improper**. Rubin naturally prefers proper methods.

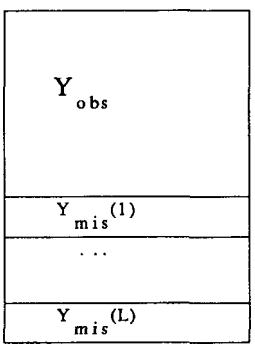
The most natural way of performing multiple imputation is to take all imputed values from the **same imputation procedure**. Kalton and Kasprzyk (1986), however, present a second potential application of multiple imputations in which they are generated by

different imputation procedures, making different assumptions about the nonrespondents.

The basic idea of multiple imputation was thus to yield several independent completed data sets. The following figure illustrates this case:



The other possibility, mostly used in the present study, is to add the L imputed values to the one data set. The following figure illustrates this:

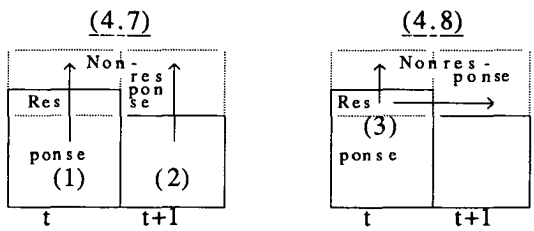


In order for the final weighting to be now valid, we must use for each of the missing units a weight which is the *L*th part of its original weight. This way is obviously simpler in practice, because a user can analyse only one data set as if it had no nonresponse. It is also easier to handle the evaluation of evenness vs. uneveness of distribution, but one disadvantage is of course the larger size.

Three formal tasks can be defined that are needed in order to

create imputed values that simulate the posterior distribution in an explicit Bayesian model. The **modeling** task, the **estimation** task, and the **imputation** task. The modeling task chooses a specific model for the data. The estimation task formulates the posterior distribution of the parameters of that model so that a random draw can be made from it. The imputation task makes L random draws from the posterior distribution of Y_{mis}. There must then be some prior information on the background which we believe to be of use for managing the missing data.

The performing of imputations calls for simulation experiments. In the case of a regression model there are a variety of possibilities, particular when bearing in mind the panel nature of the data. The prior data in the regression-based single and multiple imputations examined here are defined in two ways:



(4.7) The regression models are constructed from the data on those answering in both years, i.e. using the same households for both years in the panel (r=5745). Use is also made of the panel nature of the data when selecting the explanatory variables for the regression model and in the multiple imputation itself, in order to ensure that the changes in incomes between the two years of the panel remain as rational as possible (see below).

(4.8) The regression models required to define predicted values for the nonresponse cases are constructed from the data for those answering in the first year only (r=221), the idea being that this group shows greater homology with the nonresponse cases than does the set of those replying in both years.

Our application of the regression model-based multiple imputation method relied to a substantial extent on use of the mean

square error (MSE) for the model. For alternative (4.7) the procedure was to denote the mean square error obtained from the 1984 model by MSE(t) and that from the 1986 model by MSE(t+1) and assume that MSE is also the **same** in the missing data. In alternative (4.8) the same relative MSE was used for both years in the panel. The second prior assumption is that the same regression model is valid for both respondents and nonrespondents.

There are several opportunities to perform the estimation and imputation tasks, as Rubin's book (1987) points out, for example. In this case, our main application which follows Rubin to some extent (see p. 41 or 219-220) starts out from the standpoint that the posterior distribution of σ^2 is $(r-1)\bullet MSE\bullet\chi^2_{r-1}$, where σ^2 is the population variance of the outcome variable and χ^2_{r-1} is the inverted chi-squared distribution with r-1 degrees of freedom. When we also take into account some of the points made above and the panel-type case, our imputation task for disposable incomes in t and (t+1) is as follows (Formulation (4.9)):

(1) Draw a χ^2_{r-1} random variable, say x_i, and set

$$\sigma^2_*(t) = MSE(t)\bullet(L-1)/x_i,$$

$$\sigma^2_*(t+1) = MSE(t+1)\bullet(L-1)/x_i.$$

(2) Draw n-r independent N(0,1) random numbers, say z_i, $i \in S-R$, and set replacement values for the missing y_i in t:

$$y_i(t) = \hat{y}_i(t) + \sigma_*(t)\bullet z_i,$$

in which $\hat{y}_i = \hat{\beta}X$ is the estimated value for a missing y_i obtained by the weighted regression model, the weights being the adjusted sampling weights; $\hat{\beta}$ is an estimate for the regression coefficient and X a matrix of auxiliary variables.

Correspondingly, define the replacement values in (t+1) using the same values z_i for all i.

Add these new values to the data material.

(3) Repeat stage (2) L times using different random numbers each time.

(4) Divide the sampling weights for the missing part of the data by L, so that their combined weighting will remain the same in spite of the increased number of data points.

(5) Construct estimates for the results from the completed data in the normal way.

This simulation procedure assumes that the mean square error remains the same throughout the material. In actual fact, MSE was taken as being constant relative to the corresponding estimate, so that it increased in absolute terms as the estimates increased, i.e. $MSE/\hat{y}$ is assumed to be constant.

Stage (1) of the simulation defines a random factor for each observation unit, i.e. each household is thought of as being unique, while the same random variables z_i are used for the two years of the panel at stage (2). This is done in order to ensure that the changes in incomes for individual households should not be random. On the other hand, different random variables are used in each simulation. The results are dependent on the random numbers, of course, but the effects of these were minor.

The simulation procedure above contained one point of uncertainty, concerning the parameter σ^2. We can thus consider this imputation method to some extent a proper one, and it is also possible to extend this characteristic, for instance, by adding uncertainty to the regression coefficients β or their estimates with regarding to their estimated variances. This was done but the results are passed over as they gave only small differences compared with the results of method (4.9), although the variances were naturally slightly higher.

Large numbers of regression models were tested for each year, the best of which are set out in Table 2. Many specifications for the models were also tested, including expression of the variables in logarithmic form and the use of qualitative variables, but no essential improvement in the results was achieved, and thus the simplest specification was chosen. See also the note on outliers in Section 1.

Table 2. Models for explaining disposable income, determination coefficients and best explanatory variables in order of significance

A. Model for 1984 based on those replying in both years (r=5745) (case (1) in (4.6))

B. Model for 1986 based on those replying in both years (r=5745) (case (2) in (4.7))

C. Model for 1984 based on those replying in the first year only (r=221) (case (3) in (4.8))

	Estimates of models (above)		
	A	B	C
Intercept	1002	4586	3130
Taxable income 1984 or 1986*)	.360	.378	.380
Size of household 1984 or 1986*)	8299	7701	7902
Number of adults 1984 or 1986*)	11994	10853	10548
Property income 1984 or 1986*)	.087	.088	.091
Disposable income 1984 (in prediction for 1986: missing values by model A)	-	.114	-
R^2 (%)	71	83	73

*) For model B only

The order of the explanatory variables remains the same in these models, but their estimates differ. Also, model B contains an additional explanatory variable, the value obtained in the previous survey, and a prediction based on this. This variable derived from the panel nature of the data is not a powerful explanatory factor, but it does contribute to the fact that the determination coefficient of this model is higher than that of its 1984 counterpart. The two estimates for 1984 are similar in kind, although, e.g., the weighting on the principal variable, taxable income, is greater in the latter case. It is difficult to find particular reasons for the fact that the fit for 1986 is better than that for 1984. Some influence may be exerted by changes in taxation, and by nonresponse itself, as the rest of households in a panel will be more stabile than the original sample.

Single and multiple hot deck imputation

There are a number of approaches to hot deck imputation, the principle of which is to place the data for the respondents and nonrespondents in order and then replace each of the nonresponse cases with the nearest actual observation. It is obvious that the resulting ordering process is of crucial importance for the success of the procedure.

In the present instance, when the primary aim was to estimate disposable income and changes in these, the ordering process was accomplished using results obtained from the response probability estimations of Section 3 and the panel analysis of changes in income (Laaksonen 1989), the latter attributing these changes predominantly to changes in family structure and employment conditions. Since the data on the second of these were not sufficiently comprehensive, the present material was ordered on the following two criteria, concerning the first application and the second-order completed data; see Section 2 and Description (4.7).

The first criterion was the family structure situation in two consecutive years (see Section 3). Since 7 categories were recognized, this allowed 7 x 7 = 49 combinations or cells for the two years, 7 of which imply an unchanged family structure and the other 42 a change of some kind. Seven change cells, however, were collapsed in an attempt to increase the number of observations in them and thus improve their stability.

The second criterion consisted of the disposable incomes of the respondents in 1984 and the predicted values for the nonrespondents. The regression estimates were defined in the same way as in the above regression imputation models (model (4.10)).

The nearest substitute was taken only from the same cell as the missing unit.

Hot deck imputation implemented in the above manner still entails the same problem as the other nonresponse adjustment methods, namely that the number of observations in each cell diminishes as the number of cells increases, to the extent that a cell may contain nothing more than missing data. In such a case no

suitable observation is forthcoming and the cell structure has to be altered. No serious problems emerged in the present experiment due to aggregation of cells, but there are some cells implying changes which contained fairly few observations, so that it was impossible to find a good replacement value.

This technique of replacing missing data with the best possible substitutes from the real material has the advantage that the mutual connections between the variables continue to conform to the real situation. This was verified here by transferring both values from the same respondent household even though that for the first year actually existed.

Hot deck imputation may also be applied by producing several replacement values, all lying close to the basic figures. One possibility in this case would be to select the five nearest substitutes from the same cell as the missing unit, for example. This was not tried here, however, as the 'duplication' of valid replacement observations would have caused difficulties in the cells which contained large numbers of missing values compared with real ones, i.e. particularly in those representing changes in family structure between the two panel years.

Instead another application of multiple imputation was performed concerning the first-order completed data (see Section 2, or (4.8)), which contain relatively few missing values. It would then have been possible to introduce the method described above, but in this context we wished to test the method by using different imputation procedures (see this section), although the differences were still quite minor ones.

We chose three variables to obtain the differences, involving different classifications of: (1) family structure, (2) status or residence (region), (3) income, or their predictions. In case (1) we had three specifications: 42 cells as above, family structure in 1984 and its change (cf. Section 3) yielding 7x2 = 14 cells, and the number of adults in both years forming 4x4=16 cells. In case (2) we had three classes for status and four for residence, as in the response probability models in Section 3, and the oone of these was chosen. In case (3) we had two specifications: known disposable income for 1984, and disposable income or its prediction for 1986.

For each multiple imputation we took one of the specifications (1), (2) and (3). The first two formed the combination within which we searched for the nearest substitute for the missing value by specification (3), altogether five substitutes.

These imputation procedures are close to each other, but nevertheless differ, and it is possible to obtain the same or slightly different substitutes in different imputations. The method is clearly implicit, as hot deck imputation always is, and differentation makes it still more implicit. The results of other adjustment procedures are useful for its application, as also is a good knowledge of the data material and its exploitation.

5. EMPIRICAL FINDINGS

Comparative estimations were made based on mainly real and some simulated data. The simulated data were constructed by using random numbers to form independent samples picked from a population very similar to the original sample. The estimates for comparison were the means of these samples, and the number of simulations was 30, which was reasonable for evaluating a quantity of biases. (*Only few results based on the simulated data are presented here, as they were largely similar to those achieved with the real data. Simulated data were also used to test the programs.).*

The main results obtained from the real data are presented in Tables 3, 4a and 4b. In addition, Table 5 and Figures 6 and 7 deal with some specific considerations. The symbols used in this connection call for some explanations:

Table 3 contains certain household structure statistics which provide a major source for improving the accuracy of the actual substantive data, so that one should always aim at correcting these at the first stage. It also contains a presentation of the methods used to give estimates for these. This post-stratification method is described in formulae (3.7) and (3.8) of Section 3. The same section described the model-based weighting methods, denoted by "weighting" in the tables. These are based on both post-stratification and techniques of response probability modeling. A further method, the "direct Horvitz-Thompson" method, was operated using formula (3.5) and assuming the whole data material to be perfect.

Unfortunately we did not have any opportunity to form weights of this kind for comparison with post-stratification methods. Previous results obtained under quite similar conditions proved that the post-stratification method provides only marginal improvements (Laaksonen 1988, Ekholm and Laaksonen 1990), because it can be used with relatively aggregated classes of post-strata, based on areas, for example.

Since the data were been slightly uncertain concerning nonrespondents in both years, the comparisons between the methods are not all fully consistent. This uncertainty is not serious as long as it is borne in mind when evaluating the results, especially those regarding disposable income in Tables 4 and 5. More generally, the following standpoints give three main lines of approach for evaluating the results:

(1) We can usually compare methods used to adjust for unit nonresponse, or methods used to adjust for item nonresponse. This being the case, we have no problems in comparing (i) post-stratification vs. other weighting results, and (ii) different imputation results. Tables 4a and 4b include both. To a certain extent we can define parallel weighting and imputation methods, too, in which case we have two alternatives: (i) to use only weighting adjustments, or (ii) if the results for these complete cases are not reliable, to produce one or more imputed substitutes for nonresponse items.

(2) The different points of time used for weighting do not entitle us to compare results obtained using different weights, although we can evaluate the substance of these results, of course, if the methods used are comparable. We see, among other things, that income changes, measured as arithmetic means of the logarithms of incomes calculated for individual households, are larger with the weights for the first year than with those for the second year.

This is derived particularly from the higher average incomes for the first year obtained when using the latter weights. Thus the latter weights are, on average, larger for groups with a larger income than the former weights. Because the weights depend on three factors regarding the sample households, i.e. their size, their place of residence and the taxation groups of their members (Section 3), this indicates that households very often move to states with higher rather than lower incomes. This can be considered an indication of regression effect towards the mean (see Laaksonen 1989, Stadt and Wansbeek 1990).

(3) The results concerning the same symbols A to I, and the same statistics are always comparable. Thus we can compare the results obtained by different imputation methods if the data area is the same as for methods H1 to H6, methods I1 to I3, and methods A1 and A2. To some extent we can also compare methods A, B, D, E

and H, methods C, F, and I, and methods C, G, A1 and A2.

Methods G, A1 and A2 concern adjustments using auxiliary information about nonrespondents in the second year. Their effects are naturally very much more minor than those of the other adjustments, and we will therefore evaluate the trends rather than the magnitudes of these results. This may be justified on two scores: (i) the data are excellent compared with other adjustments, and (ii) this nonresponse has special qualities.

Table 3. Estimates for number and mean size of households in the IDS panel 1984-86, obtained by different methods; weighting = adjustments by response probability modeling (Formula (3.8)). . = impossible or irrelevant

Number of respondents (r)	Sample size used in estimation (n)	Mean size of households		Number of households (1000s)		Method Name	Symbol for result (see Table 4)
		1984	1986	1984	1986		
5966	5966	2.52	.	1970	.	Post-stratification	A
(5966)	5966	.	2.46	.	2000	Post-stratification	A1 or A2
5745	5745	2.53	2.53	1970	.	Post-stratification	B
5745	5745	.	2.47	.	1999	Post-stratification	C
5966	7345	2.43	.	2027	.	Weighting Model a) in Table 1	D
5745	7345	2.41	2.41	2033	.	Weighting Model e) in Table 1	E
5745	7345	.	2.36	.	2051	Weighting Model e) in Table 1	F1
5745	7345	.	2.34		2064	Weighting Model f) in Table 1	F2
5745	5966	.	2.45	.	2006	Weighting Model c) in Table 1	G
7345	(7345)	2.40	2.39	2032	.	Direct Horvitz-Thompson	H
(7345)	7345	.	2.35	.	2049	Direct Horvitz-Thompson	I

Note: Results H and I, and partly also D, E and F are estimated from slightly uncertain data

Table 4a. Estimates for means, first decile points and coefficients of variation in disposable income per household and mean changes in income in the IDS panel 1984-86, obtained by different adjustment methods. . =impossible

Method Symbol in Table 3	Description	Mean		First decile point		Coefficient of variation		Mean income change
		1984	1986	1984	1986	1984	1986	1984-86
	a) Weighted with sampling weights for 1984							
A	Post-stratification	80597	.	29407	.	.545	.	.
B	Post-stratification	80892	96165	29416	36958	.540	.520	.190
D	Weighting (n-r=7345-5966=1378), Model a) in Table 1	77797	92837	27637	35074	.561	.538	.197
E	Weighting (n-r=7345-5745=1600), Model e) in Table 1	77528	92576	27602	35074	.563	.539	.197
H1	Single regression imputation; data on respondents in both years	78256	93031	28837	36268	.546	.524	.190
H2	Multiple regression imputation; data and model as H 1	78388	93155	28578	36119	.556	.533	.190
H3	Single hot deck imputation; data as H1; nearness by family structure, its changes, and by disposable income	77736	92953	28065	35118	.557	.546	.194
H4	Single regression imputation; data on respondents in first year	78156	92356	27793	34831	.555	.535	.181
H5	Multiple regresssion imputation; data and model as H 4	78278	92531	27553	34732	.564	.547	.181
H6	Single hot deck imputation as H 3; data and model as H 4 and H 5	78314	93660	27634	35128	.559	.546	.193
	(b) Weighted with sampling weights for 1986							
C	Post-stratification	85156	94175	31099	35191	.529	.535	.107
F1	Weighting (n-r=7345-5745=1600), Model e) in Table 1	81968	90845	29165	34317	.552	.552	.110
F2	Weighting, Model f) in Table 1	82645	90519	29416	33991	.550	.558	.095
G	Weighting (n-r=5966-5745=221), Model c) in Table 1	84907	93919	31024	35128	.531	.536	.104
A1	Single hot deck imputation, only for nonrespondents of 1986	84763	94711	30210	36094	.534	.531	.119
A2	Multiple hot deck imputation, by different imputation procedures	84764	94985	30458	36021	.534	.531	.121
I1	Single regression imputation as H 1	82218	91961	30230	35128	.536	.533	.119
I2	Multiple regression imputation as H 2	82398	92137	30147	35074	.546	.542	.119
I3	Single hot deck imputation as H 3	82908	92123	29710	34751	.545	.553	.111

Table 4b. Estimates for means, first decile points and coefficients of variation in disposable income per consumption unit and mean changes in income in the IDS panel 1984-86, obtained by different adjustment methods. . =impossible

Method Symbol in Table 3	Description	Mean		First decile point		Coefficient of variation		Mean income change
		1984	1986	1984	1986	1984	1986	1984-86
	(a) Weighted with sampling weights for 1984							
A	Post-stratification	41675	.	24717	.	.385	.	.
B	Post-stratification	41638	49602	24554	30169	.389	.380	.181
D	Weighting (n-r=7345-5966=1378), Model a) in Table 1	41477	49460	24423	29987	.392	.383	.183
E	Weighting (n-r=7345-5745=1600), Model e) in Table 1	41488	49477	24423	30033	.392	.382	.183
H1	Single regression imputation; data on respondents in both years	41900	49779	25102	30502	.376	.377	.175
H2	Multiple regression imputation; data and model as H 1	41971	49847	24938	30338	.394	.390	.176
H3	Single hot deck imputation: data as H1; nearness by family structure, its changes, and by disposable income	41506	49473	24717	29878	.393	.391	.179
H4	Single regression imputation; data on respondents in first year	41848	49437	24695	29604	.386	.385	.170
H5	Multiple regresssion imputation; data and model as H 4	41913	49530	24601	29435	.402	.403	.170
H6	Single hot deck imputation as H 3; data and model as H 4 and H 5	41798	49831	24506	30169	.392	.401	.180
	(b) Weighted with sampling weights for 1986							
C	Post-stratification	41973	49422	24716	29914	.388	.380	.169
F1	Weighting (n-r=7345-5745=1600), Model e) in Table 1	41830	49308	24554	29810	.390	.382	.170
F2	Weighting, Model f) in Table 1	41958	49411	24694	29957	.389	.383	.169
G	Weighting (n-r=5966-5745=221), Model c) in Table 1	41982	49436	24716	29916	.388	.380	.169
A1	Single hot deck imputation, only for nonrespondents of 1986	41973	49238	24554	30171	.388	.376	.168
A2	Multiple hot deck imputation, by different imputation procedures	41845	49308	24484	29914	.391	.377	.171
I1	Single regression imputation as H 1	42157	49488	25562	29781	.370	.383	.158
I2	Multiple regression imputation as H 2	42249	49587	25528	29716	.387	.403	.158
I3	Single hot deck imputation as H 3	42122	49425	25087	29309	.379	.395	.158

The results are presented phase by phase using the classification given at the end of Section 1:

(1) The numbers of households are underestimated and their mean sizes overestimated when weightings adjusted only by areal post-stratification are used. The differences can be regarded as appreciable in size if nonresponse is high, but nonresponse in the second year of the panel is small, and the adjustment needed is thus correspondingly small (cf. results A, C and G). Results using "perfect" data, as in methods H and I, are approximately similar to those obtained by the comparable weighting methods.

On the whole Table 3 gives reasonably good information on the number of households and their mean sizes. Correspondingly, these adjustments provide improvements for the same statistics by subgroups as by regions or socio-economic groups. These results are not included in the present report. The number of households was 2.03-2.04 million in 1984 and 2.04-2.06 million in 1986. These figures are fairly consistent with the results obtained from the Household Budget Survey of 1985 (Laaksonen 1988).

(2) Figures for mean disposable incomes are presented in Tables 4a and 4b. The results regarding disposable income 'per *household*'corresponded to expectations, because the post-stratified means were larger than those adjusted by this weighting method or obtained by imputation except by methods A1 and A2 for 1986. This exception can be explained, however, since both single and multiple hot deck imputation for nonrespondents in the second panel year proved that their income increased very much more than those of the nonrespondents in both years and the respondents. The post-stratification C and weighting method H6 do not show any corresponding tendency. Although the differences are small, due to low nonresponse, results A1 and A2 argue in favour of the use of these hot deck imputation methods.

The weighting adjustments with second-order compeleted data yield means which vary surprisingly. In the case of results D, E and F the means per household, especially for 1984, are smaller than with the imputation methods, although the trends are similar. The explanation for this lies in the uncertain data for the nonrespondents. The weighting method is more sensitive to special observations in applications of this kind than are the relatively

"conservative" imputation methods used. Result F2 gives confirmation of this explanation, since its model used change in family structure as one explanatory variable. This reasonably small modification to the model gave substantially higher averages for 1984.

The sensitivity of the model-based weighting methods appeared more clearly in some excluded applications. A detailed consideration of these models showed that the number of respondents was very low in some cells, and thus also the estimated response probabilities were small, even close to zero. Such cases are not acceptable, as some other researchers have also pointed out, referring to other weighting methods (see e.g. Chapman et al 1986). In these cases we then had too many cells, and our solution was to collapse them, and correspondingly to increase the number of observations.

The means calculated *per consumption unit* point to markedly smaller differences between the methods in the various cases. This is caused by the distribution of nonresponse, in which household size and composition are significant factors.

(3) Measured in terms of the coefficient of variation and calculated per household and using the 1984 weights, the differences in income between households were reduced between 1984 and 1986 in all cases in which comparable results existed. The weights for 1986 showed smaller changes and also the opposite trend from those obtained using post-stratification C, weightings F2 and G, and the hot deck method I3.

Comparison of the estimates for the two years using the weights for the statistical year, i.e. for cross-sectional studies, gives the following comparable results (same method and same data):

Methods	Coefficient of variation	
	1984	1986
Post-stratification B and C	.540	.535
Model-based weighting E and F 1	.563	.552
Single regression H1 and I1	.546	.533
Multiple regression H2 and I2	.556	.542
Single hot deck H3 and I3	.557	.553

All the results point to a levelling out of income differences, although the post-stratified and hot deck ones do so to a lesser extent. This observation also shows clearly that the

adjustments by means of weighting, multiple regression and single hot deck imputation increase the variances, which without doubt is the right trend. The highest variance was obtained with the weighting, one reason being the sensitivity of the method to the data. Imputations H4 to H6 give higher variances than imputations H1 to H3 because of the regression model used, the former model being based on only the first year's respondents. If our evaluation in Section 4 is valid that these households are closer to the group of all nonrespondents, then we should prefer these results.

Multiple regression imputations always provide higher variances than the corresponding single ones. Thus one aim of this method, higher and more realistic variances and interval estimates, has obviously been achieved. This result is also valid for statistics expressed per consumption unit, and the other adjustment methods give parallel results. The variances in the hot deck method are certainly slightly more cautious, except for results A1 and A2, as discussed above. Some particular results are also presented in Table 5, which shows clearly how powerful the changes in income for this group were in the hot deck method. While the predicted income of the nonrespondents increased in 1986, the variance in their income decreased. The results of Table 5 obtained using methods H2 and H5 as examples show remarkably even differences between imputed values and all values.

Table 5. Some comparisons of disposable income in multiple imputation cases

	Data and Method					
	Second-order completed data				First-order completed data	
	Regression model					
	first year		both years		Multiple hot deck	
	H 5		H 2		A 2	
	1984	1986	1984	1986	1984	1986
Ratio (%) between imputed and all values for						
- means	87	85	87	87	89	122
- st. deviations	116	118	109	107	120	77
Ratio (%) between between-variance and within-variance (see Formulae (4.4) and (4.5))	1.1	1.3	1.2	1.2	3.4	17.0

Figure 6 illustrates differences in the distributions for 1984. The first peak refers to nonrespondents in both years, and the second to nonrespondents on the second year, whereas the third curve, for respondents in both years, does not have any clear peak at all. It shows clearly how the focus for the nonrespondents lies on groups with lower incomes.

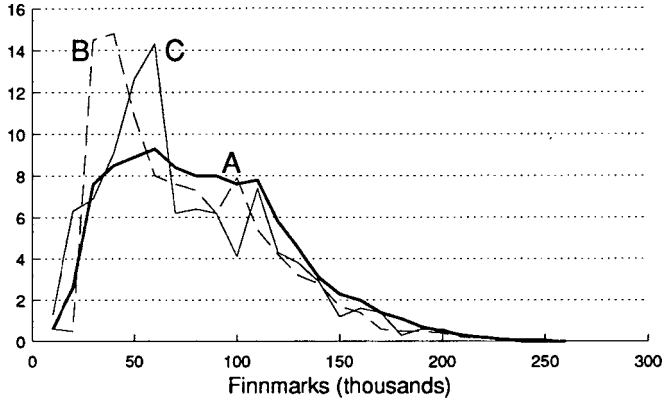
Figure 7 concerns corresponding estimates for 1986. The curves are not similar. Those of nonrespondents are closer to one another, and further away from the origo than that for the respondents in both years. The former result implies that these two types of nonrespondent have not yet merged into the same group, but moved in that direction. The latter result shows that many of nonrespondents increased their income, presumably because their living conditions improved. Thus the exclusion of nonrespondents from a sample and the failure to use imputation will cause more serious problems in longer panel surveys.

Table 5 also gives some results obtained using the original notations of Rubin et al (see formulae (4.4) and (4.5)), which divide variance into between and within variances. The results for the second-order completed data show that the proportion of between-variance of imputations is very minor compared with the within-variance, whereas it is moderate for result A2, using the multiple hot deck methods, and fairly high for 1986. This is attributable to the data rather than the method used.

The results obtained using the single and multiple hot deck imputations are very similar in our cases. This is due to the relatively minor differences between the imputation procedures and the low nonresponse rates. The trend is the same as for the regression imputations in that the variances also increased for the results per household although the results in Table 4a appear to be the same (Table 5 shows differences, too).

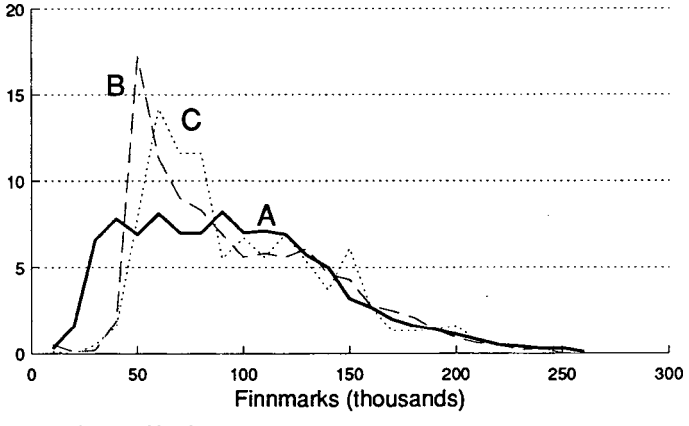
The first decile points vary in the same way as the coefficients of variance to some extent. Larger variance often implies a lower first decile, and vice versa, so that the adjustments in general also give more reliable information about these specific points in the distributions. Detailed consideration of these results is excluded from this report, however, because it is difficult to evaluate exactly the quality of these statistics.

Fig. 6. Distributions of disposable income per household in 1984, obtained by different methods and using different data sets (sampling weights from 1984); nonrespondents by multiple regression imputation



A=respondents of of both years
B=nonrespondents of both years
C=respondents of 1984 only

Fig. 7. Distributions of disposable income per household in 1986, obtained by different methods and using different data sets (sampling weights from 1984); nonrespondents by multiple regression imputation



A=respondents of both years
B=nonrespondents of both years
C=nonrespondents of 1986

It is nevertheless the case that these quantiles are very important in income studies, and there is reason to insist that adjustments for nonresponse should also function realiably in this connection.

(4) The mean changes in income calculated by the various methods differ partly in the same way as the previous statistics indicated, but the single and multiple regression imputations now provide approximately the same results, of course. The estimates given by the model-based weighting methods are fairly variable, which points out to the sensitivity of these methods, but mainly they reflect same trends as the post-stratified results. The hot deck method gives quite similar results as the weighting methods except in cases of methods A1 and A2, which can be explained in the same way as the means for 1986 above.

(5) The uncertainty increases when one turns to examining the reasons for the changes in incomes. This was observed with the models in which changes calculated per consumption unit were explained in terms of incomes in the first year of the panel given a family structure index of four in both years, a simplification of the applications reported in Laaksonen (1989).

The post-stratified case gave similar results to the model-based weighting adjustment method, their differences lying within the confidence limits, whereas those obtained by single and multiple regression imputation departed markedly from these in an obviously erroneous direction, as indicated by calculations made on the simulated data. This is understandable, since both the dependent variable and the independent variables had been imputed and fabricated data had been generated. The results obtained with the hot deck imputations were better, as the method had been adopted precisely with the intention of preserving this connection.

6. SUMMARY AND CONCLUSIONS

The purpose of the research here was to examine nonresponse and other forms of missingness and develop methods for adjustments for the statistical biases introduced by it. It is observed that even a nonresponse of 5% can be too large to be ignored. Reweighting by reference to data in the initial sample and the nonresponse cases is a suitable means of adjusting for unit nonresponse, one approach in this method being the identification of subsets of cells in which the variables being examined behave in as similar a manner as possible.

If the cells were formed at the population level, this would be a case of post-stratification, but so far it has not proved possible to achieve very homogeneous cells or to develop the response mechanism to a state being reasonably ignorable by this means, as fairly rough data have to be used, usually amounting to an areal post- stratification. This is due to the fact that no data on households as such are available from registers, only on the members of households. More disaggregated cells can be formed, however, in selected samples which are updated at each time of inspection.

A more efficient adjustment for nonresponse can be achieved using response probabilities for the cells in addition to areal post-stratification, whereupon modelling can be employed to search for a good cell structure and to estimate the response probabilities in the same manner as in the Finnish Household Budget Survey of 1985 (see Laaksonen 1988, Ekholm & Laaksonen 1990). This method is also applicable to the panel study situation, but then adjustments have to be made to a larger number of subsamples if the scheme of the research permits, and it is also necessary to take account of changes in household structure over the duration of the panel.

Imputation is most relevant as an adjustment method in cases of item nonresponse where it is known which values are missing,

provided that good explanatory variables can be found for the variable to be imputed. In the case of the present two year panel, imputation could be focused on the nonresponse data for either the first or the second year or both.

Where nonrespondents in both years were concerned, all the data had to be extracted from registers, but the second year had the advantage when the first year interview results were available, allowing greater scope for the use of models of the regression type to explain and predict the variable to be imputed.

The best explanatory variable for disposable income proved to be taxable income in the same year, while other good variables were ones connected with the size of the household. Disposable income in the first year could also be included as a variable in the model for the second year, where it achieved statistical significance even though it was not the best explanatory factor.

If the values predicted by the regression model are inserted directly in place of the missing values we have a case of single imputation, but experiments were also performed here with multiple imputation, in which the missing item is replaced with several values by exploiting the error terms in the regression model. The assumption was that these terms would be the same for both the actual and the missing values, and that the models themselves would also be the same for both.

If better data were available, the assumptions could be altered, additional information for this purpose being obtained by taking a sample of the missing observations. This was not possible in the present instance, but an analogous method was tested with a subsample of cases which responded only in the first year and these proved to be more similar to the nonrespondents in terms of their register data than did those who replied in both years.

Multiple imputation proved superior to single imputation for defining the distribution of incomes and the interval estimate of the mean. On the other hand, regression imputation does not succeed in describing changes in income between two successive panel years in a way which allows them to be explained in a rational manner (e.g. in the manner of the model described in Laaksonen 1989).

Better results in this respect are obtained with well planned hierarchical hot deck imputations. In this case the hierarchy was based on stability and changes in family structure on the one hand

and on disposable income (for respondents) and its predicted values (for nonrespondents) on the other hand. It might well be safer in ordinary attempts to explain changes in income, however, to use the weighting method and to ignore the estimated changes in the nonrespondent group.

The use of multiple imputation with permanent data intended for use in official statistics entails problems of its own. Rubin and Schenker (1987) insist that the imputations should be carried out by those responsible for compiling the data, as they are best aware of the deficiencies, including nonresponse. This is true, but it would seem inadvisable to go as far as to include the extensions called for by multiple imputation in microdata intended for the use of external clients of the CSO of Finland.

Single imputation does not cause the same difficulties, but it would still seem dangerous to make adjustments of this kind to sum variables denoting disposable incomes, as they would confuse the relations between the variables and reduce the variance. Care should also be taken to ensure that rational relations continue to exist between the factors contributing to the sum variable. On the other hand, a well constructed hot deck imputation would be less trouble to use for item nonresponse correction in this case, since greater freedom exists with variables that are of less importance for the total material.

A need for nonresponse ajustment by means of imputation has arisen at the CSO from time to time and is likely to do so in the future. For instance, a few specialized variables in the 1987 income distribution survey were defined for one part of the material from a regression model constructed from the 1986 data, e.g. the value of firewood obtained from the respondent's own forests. This enabled the corresponding item to be left out of the 1987 questionnaire altogether and reduced the interview time. The explanatory variables used for this purpose included the amount of forest owned by the household and possession of wood-fired heating in the house.

Another example is the survey of household savings and debts, in which the item nonresponse rate on some variables was of the order of 10-20%, and various routes exist for imputing the missing values from register and interview data; see preliminary results of Laaksonen (1990) using both regression and hot deck methods.

The use of multiple imputation is also justified in special research instances, particularly where the variables to be imputed and the changes in them are of sufficient importance that they furnish a motive for carrying out the additional work involved. Disposable income, as discussed here, and other crucial income meausures, may well be regarded as a variable of this degree of importance, because the changes in income distribution from year to year are highly sensitive to the methods and data used in a survey. On the other hand, politicians, journalists and ordinary citizens are very interested in this information, which places considerable demands on its quality.

ACKNOWLEDGEMENTS

I am grateful to the following persons who have read some parts of the first version of this report and made valuable comments: Dr. Anders Ekholm (University of Helsinki), Mr. Nico Nieuwenbroek (Netherlands Central Bureau of Statistics), Dr. Leif Nordberg (Åbo Akademi), Dr. Erkki Pahkinen (University of Jyväskylä) and Dr. Frank van de Pol (Netherlands Central Bureau of Statistics). The English was revised by Malcolm Hicks, Oulu.

REFERENCES

Chakrabarty, R.P. (1989). Composite Estimation for SIPP Annual Estimates. Working Papers of SIPP. U.S. Bureau of the Census.

Chapman, D.W., Bailey, L. and Kasprzyk, D. (1986). Nonresponse Adjustment Procedures at the U. S. Bureau of the Census. *Survey Methodology*, 12,2, pp. 161-180.

Cochran, W.G. (1977). Sampling Techniques. Third Edition. John Wiley & Sons. U.S.A.

Cowell, F.A. (1977). Measuring Inequality. Philip Allen. Great Britain.

Curtin, R.T. and Juster, F.T. and Morgan, J.N. (1988). Survey Estimates of Wealth: An Assessment of Quality. Institute for Social Research. University of Michigan. Working Paper Series.

Duncan, G.J. and Kalton, G. (1986). Issues of Design and Analysis of Surveys Across Time. International Association of Survey Statisticians. Invited Paper, 45th Session of the ISI, Amsterdam August 12-22 1985. Booklet, 1, pp. 167-182.

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

Herzog, T.N. and Rubin, D.B. (1983). Using Multiple Imputation to Handle Nonresponse in Surveys. In Incomplete Data in Surveys, Vol. 2: Theory and Bibliographics. W.G. Madow, I. Olkin, and D.B. Rubin (eds.), Academic Press, New York, pp. 209-245.

Holt, D. And Skinner, C.J. (1987). Components of Change in Surveys. International Association of Survey Statisticians. Invited Paper, 46th Session of the ISI, Tokyo September 8-16. Booklet, pp. 65-85.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Population. *Journal of the American Statistical Association,* 47, pp. 663-685.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology,* 12,1. pp. 1-16.

Kish, L. (1987). Statistical Design for Research. John Wiley & Sons.

Laaksonen, S. (1988). Correcting for Nonresponse in Household Data (In Finnish with English Summary). Central Statistical Office of Finland. Studies 147.

Laaksonen, S. (1989). Use of Panel Data in Applications of Income Dynamics. *Finnish Economic Papers,* 2,1, pp. 55-64.

Laaksonen, S. (1990). Handling of Measurement Problems for Complicated Quantitative Variables in a Two-Year Panel. Paper presented for the International Conference of Measurement Errors in Surveys, Tucson, November 11th-14th.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. John Wiley & Sons. U.S.A.

Nieuwenbroek, N.J. (1990). Precision of Net Change in a Rotating Panel Survey. Department of Statistical Methods. Netherlands Central Bureau of Statistics. First Version.

Nygård, F. and Sandström, A. (1985). The Estimation of the Gini and the Entropy Inequality Parameters in Finite Populations. *Journal of Official Statistics*, 1,4, pp. 399-412.

van de Pol, F. (1990). Issues of Design and Analysis of Panels. Sociometric Research Foundation. Amsterdam.

Rubin, D.B. (1978). Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 20-34.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse. John Wiley & Sons. U.S.A.

Rubin, D.B. and Schenker, N. (1987). Interval Estimation from Multiply-Imputed Data: A Case Study Using Census Agriculture Industry Codes. *Journal of Official Statistics*, 3,4, pp. 375-387.

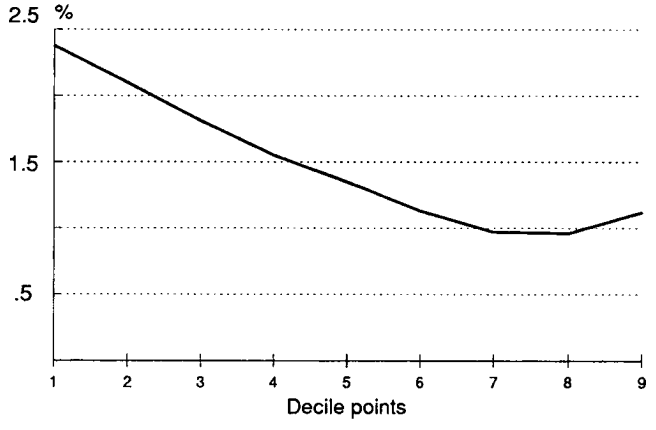
Singh, R. and Weidman, L. and Shapiro, G.M. (1989). Quality of the Survey of Income and Program Participation (SIPP) Estimates. Working Papers of SIPP. U.S. Bureau of the Census.

Smith, T.M.F. and Holt, D. (1989). Some Inferential Problems in the Analysis of Surveys over Time. International Association of Survey Statisticians. Invited Paper, 47th Session of the ISI, Paris, Booklet II, pp. 363-382.

van de Stadt, H. and Wansbeek, T. (1990). Regression Effects in Tabulating from Panel Data. *Journal of Official Statist i cs*, 6,3, pp. 311-318.

Waterton, J. and Lievesley, D. (1987). Attrition in a Panel Study of Attitudes. *Journal of Official Statistics*, 3,3, pp. 267-282.

Appendix: Relative standard errors in disposable income per household by decile points in 1984, obtained using 30 bootstrap simulations (for more details, see the text)



SUMMARY IN FINNISH

Laaksonen, S. (1991). Vastauskadon adjustointimenetelmien vertailu lyhyen aikavälin paneliaineistossa, sovellutuksia kotitalouksien tulojakaumiin. Tilastokeskus, Tutkimuksia 179. Helsinki.

Avainsanat: esiositus, imputointi eli paikkaus, jälkiositus, käytettävissä oleva tulo, logistinen regressio, moni-imputointi, paneliominaisuus, regressioimputointi, sijaistaminen eli hot deck, vastaustodennäköisyys, yksinkertainen imputointi.

YHTEENVETO

Tilastokeskus on vuodesta 1977 lähtien tuottanut kotitalous-pohjaista *tulonjakotilastoa*, *jonka* tärkein muuttuja on käytettävissä oleva tulo. Monia tilaston tietoja tuotetaan myös henkilöittäin, siis kotitalouksien jäsenistä. Aluksi tilasto perustui poikkileikkaustietoihin eli vuosittaiset otokset olivat toisistaan *riippumattomia*. Vuodesta 1982 lähtien tiedot on kerätty panelimaisesti siten, että puolet otoksesta on jatkanut myös seuraavaan tilastovuoteen. Kyseessä on siis ollut **rotatoiva paneli.**

Kaikki otokseen poimitut eivät vastaa vapaaehtoisuuteen perustuviin *tiedusteluihin*. *Tulonjakotilaston* paneleissa vastauskato, ns. **yksikkökato**, on ensimmäisenä vuonna ollut 15-20% ja toisena vuonna 4-6%. Jos samassa yhteydessä, kuten vuonna 1987, on kerätty muitakin tietoja, luvut ovat nousseet korkeammiksi. Vastauskadon lisäksi poistumaa ja häiriötä aiheuttaa ylipeitto, mutta sen määrä näin lyhyissä paneleissa on vain muutaman prosentin luokkaa, joskin painottuu selvästi vanhempiin ihmisiin. Yksikkökadon lisäksi esiintyy **eräkatoa.**

Jos aineisto sisältää vastauskatoa ja muuta samantyyppistä puuttuneisuutta, on useita käsittelymahdollisuuksia. Helpoin ja huonoin vaihtoehto on sen unohtaminen, jota tilastovirastoissa harvoin käytetään. Parempi, ja varsin yleisesti käytetty vaihtoehto on tutkia puuttuneisuuden laatu, mutta unohtaa se itse aineiston käsittelyssä. Sen sijaan tämän laatuanalyysin tuloksia voidaan käyttää laatuselosteissa, käyttäjäohjeissa sekä uusia tiedusteluja suunniteltaessa ja toteutettaessa. Paras vaihtoehto olisi tietysti korjata häiriö hankkimalla todellista tietoa vastaamattomista, jollei kaikista, niin jostakin osasta hyvällä otosmenetelmällä. Tähän on käytännössä hyvin harvoin mahdollisuuksia. Sen vuoksi joudutaan etsimään muita menetelmiä, keräyksen jälkeisiä menetelmiä, joilla puuttuneisuuden harhaa voitaisiin lieventää, tasoittaa ja kompensoida mahdollisimman paljon. Tällaisia menetelmiä on tässä tutkimuksessa kutsuttu **adjustointimenetelmiksi** (aitoa suomenkielistä nimeä tälle ei ole keksitty).

Kolme vastauskadon tai muun puuttuneisuuden adjustointi-menetelmää on käytettävissä: (i) **uudelleenpainotus**, (ii) **imputointi** eli puuttuvien tietojen **paikkaaminen** 'korvikearvoilla' ja (iii) **estimointimenetelmän tai mallin sisäiset menetelmät,** jotka eivät muuta itse havaintoaineistoa. Viimeksi mainitusta lähestymistavasta ei Tilastokeskuksessa eikä juuri muissakaan tilastovirastoissa ole kokemusta, koska tavoitteena on ollut sisällyttää kaikki adjustoinnit itse havaintoaineistoon, havaintomatriisiin.

Tämä tutkimus esittää vertailuja neljän adjustointimenetelmän välillä. Tärkeimpänä tulosmuuttujana on käytettävissä olevaa tulo, laskettuna joko kotitaloutta tai kulutusyksikköä kohti. Uudelleenpainotusmenetelmistä on sovellettu **jälkiositusta** ja **malli-adjustoitua uudelleenpainotusta,** jossa on mallitettu kotitalouksien vastaustodennäköisyyksiä. Tätä ns. Ekholmin ja Laaksosen (1990) menetelmää on käytetty aikaisemmin kotitaloustiedustelussa, siis poikkileikkausaineistossa. Tässä tutkimuksessa taas kysymyksessä on paneliaineisto ja monimutkaisempi otanta-asetelma. Imputointi-menetelmistä on sovellettu **yleiseen lineaariseen malliin (regressioon tms.) perustuvaa imputointia** ja **hot deck- eli sijaistamismenetelmää**. Myös **moni-imputointeja** on tuotettu soveltaen ja kehittäen Rubinin (1987) varsin tuoretta teoriaa.

Tulokset osoittavat, ettei yhtä ja ainoaa suositeltavaa adjustointimenetelmää hankaliin tilanteisiin ole olemassa. Ratkaisut riippuvat useista eri seikoista, mm. siitä, mitkä estimaatit ovat tärkeitä, koska eri adjustointimenetelmillä ja niiden empirialla eli saatavissa olevalla datalla on erilaisia vaikutuksia estimaattien harhaan ja tarkkuuteen. Totaalien, esimerkiksi kotitalouksien lukumäärien kannalta, mahdollisimman hienojakoinen ja huolellisesti muodostettu uudelleenpainotus on paras vaihtoehto. Tämä tutkimus vahvisti aikaisempia tuloksia siitä, että malli-adjustoidulla menetelmällä hyvin helposti päästään parempiin tuloksiin kuin jälkiosituksella, koska jälkimmäistä varten ei ole riittävän yksityiskohtaista perus-joukkotasoista tietoa käytettävissä.

Pelkät malliperusteiset imputoinnit, jos mallien selitysvoima on korkea, ovat yleensä riittäviä keskiarvojen ja totaalien estimointiin. Sen sijaan jakaumasta tulee silloin liian keskittynyt ja vastaavasti luottamusvälit ovat liian kapeita. Tämä häiritsee myös tulojakaumatyyppisiä sovelluksia, vaikka käytettävissä oleva tulo voidaankin mallittaa hyvin rekisteritietojen ja edeltävän panelivuoden tietojen perusteella. Parempia tuloksia jakauman, keskivirheiden ja luottamusvälien estimoinnin kannalta saadaan lisäämällä satunnaistekijä mallien antamiin arvoihin ja sijaistamisella (hot deck), jos ne osataan kohdentaa hyvin tässä mielessä. Molempien monentaminen, moni-imputointi, tuo vielä lisää parannusta ja robustisuutta. Myös uudelleenpainotus parantaa tuloksia, mutta se ei sovellu aidon eräkadon adjustointiin, jolloin imputointi on väistämätön vaihtoehto.

Hankalampi tilanne syntyy, jos tavoitteena on mitata hyvin myös yksilöllisiä paneliaikaisia muutoksia ja yhteyksiä eri muuttujien välillä. Tässä tutkimuksessa ei löydetty hyvää pelkkään regressiomalliin perustuvaa moni-imputoinnin identifikaatiota, jolla nämä yhteydet olisi saatu riittävän todellisiksi. Sijaistaminen sen sijaan voidaan helpommin suunnitella niin, että tällaiset yhteydet säilyvät, koska ne perustuvat aitoihin havaintoihin. Tässä tapauksessa tämä taattiin siten, että molempien panelivuosien tiedot korvattiin, vaikka ensimmäisen vuoden tieto olisikin ollut olemassa. Toisaalta sijaistamisessa käytettiin

merkittävästi hyväksi mallittamista sekä muita samasta aihepiiristä saatuja tuloksia.

Siten regressio- tai muu malli osoittautui välttämättömäksi apuvälineeksi kaikkia imputointeja rakennettaessa. Tämä tuo sovellukseen **eksplisiittisyyttä**, joka on hyvän imputoinnin keskeisiä tavoitteita. Yleensä pelkällä mallituksella ei kuitenkaan selvitä, vaan tarvitaan myös **implisiittisiä** elementtejä. Yksi tällainen on myös ***priori-informaation*** valinta imputoinnin edellyttämien mallien pohjaksi. Tässä tutkimuksessa käytettiin sekä (i) saman vuoden vastaajista että (ii) vain ensimmäisen vuoden vastanneista saatua tietoa. Tehdyt empiiriset sovellukset antoivat suosituksia jälkimmäisen puolesta, mutta haittana voi olla pienehkö aineiston koko, kuten oli tämän tutkimuksen aineiston, vuosien 1984-86 tulonjakotilaston panelin tapauksessa.

Huolellisen tulotutkimuksen kannalta vastaamattomat ovat ilmeinen ongelma. Tätä osoittaa heidän painottumisensa tulojakauman alapäähän molempina panelivuosina ja molemmille vastaamattomien ryhmille, siis sekä vain toisena vuonna että molempina vuosina vastaamattomiin. Viimeksi mainitut olivat kuitenkin edellisiä selkeämmin jakauman alapään ryhmä. Muutosta tapahtui kuitenkin vuodesta 1984 vuoteen 1986 sikäli, että molemmat vastaamattomat siirtyivät keskituloisiin päin ja toisaalta ryhmät lähentyivät hieman toisiaan. Nämä tulokset antavat viitteitä siitä, että vastauskato voi vaihtelevasti vaikuttaa panelituloksiin ja siis sen huomiotta jättäminen voi tuottaa ristiriitaisia tuloksia panelin kestäessä.

Adjustointien pohjana olevan 'filosofian' ymmärtämiseksi tarvitaan muutama käsite. Otoshan kerätään tietyllä poiminta-menetelmällä, sitä päivitetään ja lopulta kysymyksiin vastataan. Rubinin (1987) mukaan kysymys on **mekanismeista**: otanta-, päivitys- ja vastausmekanismeista, jotka viimeksi mainitut tapahtuvat toisena panelivuonna uudelleen. Nämä mekanismit ovat jokseenkin **ignoraabeleja** (sellaisia jotka tiettyjen todennäköisyysjakauman kriteerien perusteella voidaan jättää huomioon ottamatta eli eivät aiheuta harhaa estimaatteihin) tai **ei-ignoraabeleja**. Jälkimmäisessä tapauksessa tulee ryhtyä adjustointeihin, joiden tavoitteena on

muuntaa mekanismit mahdollisimman ignoraabeleiksi. Tämä vastaa tilastollisen mallin yleistä filosofiaa: pyritään siihen, että residuaali ei sisällä mitään systemaattisia tekijöitä, on siis satunnainen tai 'valkoista kohinaa.'

***

Adjustointivertailujen lisäksi tutkimusraportissa on käsitelty itse panelin käsitettä, sen etuja, mahdollisuuksia ja hankaluuksia. Tässä mielessä keskustellaan erityisesti useiden otospainojen ongelmasta ja esitetään tuloksia keskivirheiden vähentymisen tuottamista eduista, jos tulosmuuttujien arvot peräkkäisinä panelivuosina korreloivat keskenään. Tässä yhteydessä tulee esille mm. 'composite'-estimaattori.

***

Paneliaineistoilla ja niiden käsittelyn teorialla on lyhyt todellinen historia, vaikka niiden juuret ulottuvat jopa 1800-luvulle. Jossain määrin nykyisessä muodossa sen on van de Polin (1990) mukaan esittänyt Paul Lazarsfeld 1930-luvun lopulla. Viime vuosina on alan kehitys tai ainakin kokeilu ripeytynyt olennaisesti, kiitos sen, että sopivia aineistoja on luotu. Samalla on tullut välttämättömäksi kehittää alan metodista ja sisällöllistä tasoa, muutakin kuin vastauskatoon ja muuhun poistumaan liittyvää. Tässä raportissa on useita tällaisia kysymyksiä tuotu esiin, mutta moniin on vain viitattu. Lähdeluettelo antaa siten hyödyllisiä virikkeitä aihepiirin muistakin kysymyksistä kiinnostuneille. Niidenkin perusteella voidaan päätellä, että panelimaisilla tai muilla pitkittäistutkimuksilla on edessään valoisa ja tutkijoita työllistävä tulevaisuus. Tämän raportin merkitys pidemmällä aikavälillä jää nähtäväksi, mutta tekijälleen se on joka tapauksessa tuottanut nyös innostuksen ja oivaltamisen hetkiä.

Documentation page

Published by

Central Statistical Office of Finland

Date of publication

10.4.1991

Authors

Seppo Laaksonen

Type of publication

Research, Statistical Methods

Commissioned by

Title of publication

Comparative Adjustment for Missingness in Short-term Panels.
Applications to Questions of Household Income Distribution

Parts of publication

See page 2

Abstract

See page 1

Keywords

Disposable Income, Hot Deck Imputation, Multiple Imputation, Nonresponse, Regression Imputation, Response Probability, Single Imputation

Other information

Series (key title and no.)		ISSN	ISBN
Studies No. 179		0355-2071	951-47-4580-9
Pages: 66	Language: English	Price: 65 FIM	Confidentiality:
Distributed by		Publisher	

155. **Sirkka-Liisa Kärkkäinen – Timo Matala – Virpi Tiitinen – Ari Tyrkkö,** Asunto-olot ja asumisen tuki. Heinäkuu 1989. 295 s.

156. **Jorma Huttunen,** Asuntovarauma 1985. Heinäkuu 1989. 168 s.

157. **Christian Starck,** Vuoden 1985 väestölaskennan luotettavuus. Elokuu 1989. 136 s.

158. **Pekka Rytkönen,** Tekninen palvelu 1970-1980 - luvulla. Heinäkuu 1989. 55 s.

159. **Ari Luukinen,** Tietojenkäsittelypalvelu 1970-1980-luvulla. Elokuu 1989. 72 s.

160. **Risto Kolari,** Ammatillinen liikkuvuus Suomessa 1975/1980/1985. 192 s.

161. **Pekka Rytkönen,** Liikkeenjohdon, kirjanpito- ja lakiasiain palvelu 1980-luvulla. Lokakuu 1989. 71 s.

162. **Ari Luukinen,** Markkinointipalvelu 1970 - 1980-luvulla. Marraskuu 1989. 72 s.

163. **Anna-Maija Lehto,** Tietotekniikka työssä. Muutoksista 1980-luvulla. Marraskuu 1989. 56 s.

164. **Henry Takala,** Kunnat ja kuntainliitot kansantalouden tilinpidossa. Tammikuu 1990. 60 s.

165. **Jarmo Hyrkkö,** Palkansaajien ansiotasoindeksi 1985=100. Tammikuu 1990. 66 s.

166. **Pekka Rytkönen,** Siivouspalvelu, ympäristöhuolto ja pesulapalvelu 1980-luvulla. Tammikuu 1990. 70 s.

167. **Jukka Muukkonen,** Luonnonvaratilinpito kestävän kehityksen kuvaajana. 119 s.

168. **Juha-Pekka Ollila,** Tieliikenteen tavarankuljetus 1980-luvulla. Helmikuu 1990. 45 s.

169. **Tuovi Alen – Seppo Laaksonen – Päivi Keinänen – Seija Ilmakunnas,** Palkkaa työstä ja sukupuolesta. Huhtikuu 1990. 90 s.

170. **Ari Tyrkkö,** Asuinolotiedot väestölaskennassa ja kotitaloustiedustelussa. Huhtikuu 1990. 63 s.

171. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Nuorten koulutus ja kotitausta. Toukokuu 1990. 115 s.

171b. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Education and the family background of the young in Finland. 1990. 115 pp.

172. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Eriarvoisuus kuoleman edessä. Sosioekonomiset kuolleisuuserot Suomessa 1971–85. Kesäkuu 1990. 145 s.

173. **Jukka Muukkonen,** Sustainable development and natural resource accounting. August 1990. 96 pp.

174. **Iiris Niemi – Hannu Pääkkönen,** Time use changes in Finland in the 1980s. August 1990. 118 pp.

175. **Väinö Kannisto,** Mortality of the elderly in late 19th and early 20th century Finland. August 1990. 50 pp.

176. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Socio-economic mortality differences in Finland 1971-85. December 1990. 108 pp.

177. **Jaana Lähteenmaa – Lasse Siurala,** Nuoret ja muutos. Tammikuu 1991. 211 s.

178. **Tuomo Martikainen – Risto Yrjönen,** Vaalit, puolueet ja yhteiskunnan muutos. Maaliskuu 1991. 120 s.

179. **Seppo Laaksonen,** Comparative Adjustments for Missingness in Short-term Panels. April 1991. 74 pp.

180. **Ágnes Babarczy – István Harcsa – Hannu Pääkkönen,** Time use trends in Finland and in Hungary, April 1991. 72 pp.

D:\udvt\painatus\tuluvp.chp/ElRi

TILASTOKESKUS

TUTKIMUKSIA

Tilastokeskus on julkaissut Tutkimuksia v. 1966 alkaen, v. 1986 lähtien ovat ilmestyneet seuraavat:

123. **Pellervo Marja-Aho.** Kansantalouden tilinpito. Yksityinen palvelutoiminta kansantalouden tilinpidossa. Tammikuu 1986. 60 s.

124. Palkansaajien ansiotasoindeksi 1980=100. Helmikuu 1986. 68 s.

125. **Matti Kortteinen – Anna-Maija Lehto – Pekka Ylöstalo,** Tietotekniikka ja suomalainen työ. Huhtikuu 1986. 164 s.

125. **Matti Kortteinen – Anna-Maija Lehto – Pekka Ylöstalo,** Information Technology and Work in Finland. January 1987. 131 pp.

126. **Väinö Kannisto,** Geographic differentials in infant mortality in Finland 1871-1983. April 1986. 82 pp.

127. **Kaj-Erik Isaksson – Simo Vahvelainen,** Muoviteollisuuden jätteet. Kesäkuu 1986. 93 s.

128. Time Use Studies: Dimensions and Applications. October 1986. 192 pp.

129. **Ritva Marin,** Ammattikuolleisuus 1971 - 80. Joulukuu 1986. 265 s.

130. **Maija Sandström,** Tukku- ja vähittäiskaupan aikasarjat 1968 - 85. Tammikuu 1987.

131. **Eeva-Sisko Veikkola – Riitta Tolonen,** Elinkeinoelämän tuki taiteille 1984. Tammikuu 1987. 34 s.

132. **Eero Tanskanen,** Asuintaloyhtiöiden energiankulutus ja kuluttajakäyttäytyminen. Maaliskuu 1987. 106 s.

133. **Heidi Melasniemi-Uutela – Eero Tanskanen,** Asuintaloyhtiöiden kaukolämpöenergian ja veden kulutus 1984. Maaliskuu 1987. 82 s.

134. Perusparannuksen panoshintaindeksi 1985=100. Huhtikuu 1987. 52 s.

135. **Reijo Kurkela,** Tupakka tupakkalain jälkeen. Toukokuu 1987. 81 s.

136. Tie- ja maarakennuskustannusindeksit 1985=100. Joulukuu 1987. 25 s.

137. 1988: **Aila Repo,** Väestön tutkinto- ja koulutusrakenne-ennuste 1985 - 2000. Tammikuu 1988. 62 s.

138. **Anna-Maija Lehto,** Naisten ja miesten työolot. Maaliskuu 1988. 222 s.

139. **Johanna Korhonen,** Teollisuustilaston ennakkotietojen estimointimenetelmä. Maaliskuu 1988. 46 s.

140. **Markku Tahvanainen,** Asuntolainojen korot ja verot. Huhtikuu 1988. 90 s.

141. **Leo Kolttola – Marja Tammilehto-Luode – Erkki Niemi,** Luonnonvaratilinpito, Esitutkimusraportti. Toukokuu 1988. 93 s.

142. **István Harcsa, Iiris Niemi & Agnes Babarczy,** Use of Time in Hungary and in Finland II, The effects of life cycle and education. May 1988. 55 pp.

143. **Heidi Melasniemi-Uutela,** Kiinteistönhoitotavat ja energian kulutus taloyhtiöissä. Kesäkuu 1988. 112 s.

144. **Ilkka Lehtinen – Tuula Koskenkylä,** Kuluttajahintaindeksi 1985=100. Kesäkuu 1988. 50 s.

145. **Elli Paakkolanvaara,** Informaatioyhteiskunta ja informaatioammatit. Heinäkuu 1988. 160 s.

146. **Ilkka Lehtinen – Jarmo Ranki,** Tuottajahintaindeksi 1985=100. Lokakuu 1988. 80 s.

147. **Seppo Laaksonen,** Katovirheen korjaus kotitalousaineistossa. Lokakuu 1988. 110 s.

148. **Hannu Uusitalo,** Muuttuva tulonjako. Lokakuu 1988. 137 s.

148. **Hannu Uusitalo,** Income Distribution in Finland. July 1989. 123 pp.

149. **Pekka Rytkönen,** Palvelusten ulkomaankauppa 1987. Marraskuu 1988. 66 s.

150. **Seppo Varjonen,** Kansainvälinen BKT- ja hintavertailu. Joulukuu 1988. 92 s.

151. **Erkki Niemi – Päivi Väisänen,** Energiatilinpito 1985, Tutkimusraportti. Maaliskuu 1989. 136 s.

152. **Helena Korpi,** Pääasiallinen toiminta ja ammattiasema vuoden 1985 väestölaskennassa: rekisteripohjaiset rinnakkaistiedot. Huhtikuu 1989. 154 s.

153. **Iiris Niemi – Hannu Pääkkönen,** Ajankäytön muutokset 1980-luvulla. Toukokuu 1989. 120 s.

154. **Kari Lindström – Anna-Maija Lehto – Irja Kandolin,** Ikä ja työ, Toukokuu 1989. 92 s.

Tilastokeskus
Statistikcentralen
Central Statistical
Office of Finland

Tutkimuksia 179
Studies

Comparative Adjustments for Missingness in Short-term Panels

Applications to Questions of Household Income Distribution

Seppo Laaksonen

This report has three main aims: (i) to develop and apply methods which are useful for adjusting for errors and biases of nonresponse and other forms of attrition in large-scale panel surveys in official statistics, especially in relation to studies of household income distribution, (ii) to compare the applicability of these methods in real cases, and (iii) to describe mechanisms and characteristics which are crucial when studying short-term panels.

Orders:

Central Statistical
Office of Finland
P.O.B. 504
SF-00101 Helsinki
Tel. +358 0 173 41

Price

65 mk

ISSN 0355-2071
ISBN 951-47-4580-9

Studies 179
Comparative Adjustments for Missingness in Short-term Panels. Central Statistical Office of Finland