Tilastokeskus
Statistikcentralen
Central Statistical
Office of Finland

SVT

Tutkimuksia nro 147
Studies

Katovirheen korjaus kotitalousaineistossa

Correcting for Nonresponse in Household Data

Seppo Laaksonen

Tilastokeskus
Central Statistical
Office of Finland

Tutkimuksia nro 147
Studies

Katovirheen korjaus kotitalousaineistossa

Correcting for Nonresponse in Household Data

Seppo Laaksonen

Lokakuu 1988

Tiedustelut - Förfrågningar

Seppo Laaksonen
(90) 17 341

Helsinki 1988

Hakaniemen Valtimo

Alkusanat

Tilastokeskuksen tilastotuotannosta osa perustuu otosaineistoihin. Näistä taas osa kerätään haastattelemalla. Sekä otantaan yleensä että haastatteluun erityisesti perustuvan tiedon määrä on ollut kasvussa niin Tilastokeskuksessa kuin muuallakin yhteiskunnassa. Koska haastattelun antaminen on vapaaehtoista, ja sellaisena se tulee säilymäänkin, on luonnollista, ettei kaikilta otokseen poimituilta saada vastauksia. Vastaamattomuudesta johtuva aineiston väheneminen, jota kutsutaan kadoksi tai vastauskadoksi, on ollut kasvussa. Tämä on osaltaan vauhdittanut pyrkimystä vastauskadosta aiheutuvien virheiden korjaukseen. Viime vuosina on alan sekä teoreettinen että empiirinen tutkimustyö laajentunut huomattavasti. Myös Tilastokeskuksessa on tämän alueen tutkimukseen panostettu entistä enemmän.

Keskeisimmäksi sovellutuskohteeksi otettiin aluksi kotitaloustiedustelu, jossa kato-ongelma on tiedonkeruun laajuuden takia tavallista suurempi. Käsillä olevassa tutkimusraportissa esitetään menetelmät, joita käytettiin vuoden 1985 tiedustelussa. Lisäksi se tarjoaa teoriaa, analyysiä ja tuloksia, joita voidaan käyttää hyväksi myös muiden vastaavanlaisten tutkimusten otannan suunnittelussa ja tulosten estimoinnissa.

Raportin on laatinut erikoistutkija Seppo Laaksonen, joka toimii tilastotieteellisten menetelmien kehittäjänä Tilastokeskuksessa.

Tilastokeskuksessa elokuussa 1988

Olavi E. Niitamo

Hilkka Vihavainen

KATOVIRHEEN KORJAUS KOTITALOUSAINEISTOSSA

CORRECTING FOR NONRESPONSE IN HOUSEHOLD DATA

SISÄLLYSLUETTELO

Seppo Laaksonen

CORRECTING FOR NONRESPONSE IN HOUSEHOLD DATA

Summary

The Finnish Household Survey is a major study whose primary purpose is to produce annual-level estimates of household consumption. In addition, the survey produces data on the ownership of durable consumer goods, the use of social services, and income. To provide a background, i.e. for classification purposes, data are collected on household characteristics such as size, socio-economic status and residential area.

The data of the Household Survey are collected in three stages, consisting of an initial interview, household budget survey, and final interview. The data collection is thus a heavy burden for both the Central Statistical Office's interviewers and the households, which explains why in 1985 all the central information was obtained from less than 70% of the households selected for the sample. This fall-off of data, nonresponse, grew by about 3 percentage points from the previous survey in 1981, in spite of the fact that the Central Statistical Office had taken great pains to reduce nonresponse.

It goes without saying that interview data will always be saddled with nonresponse. As nonresponse is not random, it causes bias in results. The impact of nonresponse can also be reduced after data collection, with a view to improving the estimates of the data. This report focuses on examining such methods and discussing their application to the Household Survey and other similar data sets.

The methods of correcting for nonresponse can be divided into two main groups, **weighting** and **imputation** methods (see e.g. Kalton and Kasprzyk 1986, Little and Rubin 1987). The former are used to correct the weighting of the original observation units in such a way that the new weights also reflect the nonresponding units. Imputation methods are used to replace the missing data by as accurate estimates

as possible, thus doing away with the problem of nonresponse.

Both methods have been tested in this study. Regarding the weighting methods, the focus is on what is known as the response probability model. Its application is called **Method A.** Regarding the imputation methods, tests have been performed on a method based on a regression model. It is called **Method B.** In the comparisons, use was also made of the original method, uncorrected for nonresponse. It is called **Method 0.**

Correcting for nonresponse also requires the availability of other information than that provided by the interviewed households. To carry out nonresponse corrections in the Household Survey, such **supplementary or auxiliary information was obtained from population, taxation and other similar registers.** Use was also made of information obtained in the initial interviews with households that subsequently failed to respond.

The methods by which corrections for nonresponse are to be carried out must always be adapted to the respective sampling design. In this case, the situation is compounded by the fact that the sampling frame is a register of persons, the Central Population Register, while the data collected relate to households. For this reason, the sample is also selected from among persons, though after a stratification by region. Round the target persons, i.e. the persons selected for the sample, household dwelling units are formed with the help of registers, followed by the formation of households with the help of interviews.

A household dwelling unit and a household are two different concepts: the former refers to persons living in the same dwelling, while the latter also requires that the members of the household dwelling unit make common provision for food or other essentials for living. For this reason, the data of nonresponding households are of a somewhat lower quality than the data of households that have been interviewed, which also hampers the carrying out of corrections for nonresponse.

Under Method A, an inflation coefficient is calculated for each household, i.e. a weight which indicates how many households in the population are represented by the household in question. In the Household Survey of 1985, it was calculated from the following expression:

$$w_k(A) = \frac{M_h}{n_h \cdot m_{hk}} \cdot \frac{1}{P_c(k)} \quad ,$$

where

$w_k(A)$ = the inflation coefficient of household k calculated by Method A (later also to be denoted without the A)

h = a regional stratum for sample selection, one out of a total of 24

M = the number of persons in the sampling frame consisting of persons over 0 year of age, excluding persons in institutions or without an address

n = the number of households at the selection stage, exclusive of the overcoverage observed later

m_k = the number of members of household k belonging to the sampling frame

$p_c(k)$ = the response probability of cell c estimated using a generalized model in which a logistic scale was used for the response variable and which had four regressors:

- Region, classified into four groups
- Municipality, classified into municipalities in central and in outlying areas
- Family structure, classified into eight groups by characteristics such as the size of the household and the number and age of the children
- Income from property, classified into two groups according to whether the household had income from property or not.

The total number of the cells thus came to 128.

The sums of the weights w (A) are estimators of the number of households represented by the respective units. Example: the estimator of the mean consumption of commodity y per household is obtained as follows:

$$\bar{y} = \frac{\sum w_k y_k}{\sum w_k}$$

The model of Method A indicated that, among other things, nonresponse was quite common among one-person households in the central areas of the South. On the other hand, high response rates were shown by young couples without children and by families with children under school age. Better than average response rates were also shown by persons living in the East and by persons who had received income from property.

The correcting effect of Method A on a given cell is the greater the more the response rate deviates from the average. This appeared, for instance, from the considerable increase in the number of one-person households compared with Method 0. Correspondingly, the number of households in the whole country grew substantially, by more than 4%. Under method A, consumption averages per household fell by an average of about 4-5%, but in some cases the result showed an increase and the reduction could correspondingly be close to 10%.

The trend of these corrections has been assessed as being right. The data sources drawn upon for these assessments include income distribution statistics, some data items of which overlap with those of the Household Survey. The results of a series of simulation tests performed to study bias also indicate that the trend of the corrections has been right.

In Method B based on regression imputation, the aim is to build optimal regression models for outcome variables. Only variables for which information is also available on nonrespondents or on the whole population can be used as regressors. Estimation offers several alternatives depending on, for instance, the quantity to be estimated. Example: the mean value of an outcome variable can be estimated using either of the following two methods:

(i) Using the model, predicted values are calculated for the missing observations, i.e. for the data of nonrespondents, after which the mean value is calculated from the body of data thus supplemented.

(ii) The mean values of the regressors obtained from the whole body of selected data or from the whole population are inserted in the estimated model, after which the model will produce the estimated mean value of the outcome variable.

Method B can be applied to a wide range of variables, for the data contain over 1,000 collection-level variables which in turn can be aggregated in a number of different ways. Application to disaggregated variables seldom produced models with reasonable regressive power. The reason was that the values of the variables for individual households varied, from zero in a good number of cases to very high values in some others. The model served better in cases where the outcome variables were aggregated enough. Examples:

- The model of total consumption had a coefficient of determination of about 62%. The most important regressors were taxable net income, the taxable value of transport equipment, and the number of children of school age.

- The model of rent, fuel and light expenditures had a coefficient of determination of about 48%. In this case, too, net income was the best regressor, followed by assets liable to taxation and the number of children of school age.

- The model of transport expenditures had a coefficient of determination of 36%. The taxable value of transport equipment was naturally the best regressor. Other significant regressors included net income and the number of persons of working age in the household.

- The best model in the tests was achieved for disposable income, its coefficient of determination being 84%. The excellent result is due to the fact that taxable income alone accounts for a major proportion of disposable income, over 90% on average.

The analysis of regression estimates, i.e. of mean values and totals, indicated that, in the case of inferior models, the results changed very little compared with Methods 0 and A and thus did not impair the estimates. In the case of superior models, mean values and totals generally changed more. Simulation tests and comparisons with estimates of

other data sets indicated that the trend of corrections was right.

The application of the regression model took place after the weighting corrections had already been carried out using Method A. Thus, the weights of Method A and not of Method O were used as the weights of the model. The effect on mean values and totals by the application of Method A was considerably greater than the corresponding effect of the subsequent application of Method B.

In addition to mean values and totals, information is also needed on the distributions of consumption and income and on the correlations of these. Successful estimation of these is hampered by variations in the reference period: budget data relate to two weeks; interview data to one month, three months or one year; register data to one year. Since the results are estimated to the level of a year, variables relating to a reference period shorter than a year in particular should be treated with care when analysing distributions and correlations. Corrections due to nonresponse may compound these estimation problems. This study shows, however, that weighting corrections do not create more problems; rather, they help to solve them.

On the other hand, regression imputation, i.e. imputation of missing data and compilation of statistics from the body of data thus supplemented, is less suitable for use in statistical production. The main reason is the need to use as authentic data as possible when analysing distributions and correlations. Regression imputation is, however, suitable for the estimation of mean values and totals. Provided that the model was a good one, as it was in the case of disposable income in particular, regression imputation turned out to be suitable for most analytical purposes.

The following overall conclusions can be drawn from the tests on the two methods:

- *Weighting corrections carried out according to Method A are highly suitable for statistical data sets of the Household Survey type. The basic requirement is that, regarding auxiliary variables, proper information is also available on nonrespondents. The quality of this information should be improved in the Household Survey. Another important point is that the correction cells involved in the method are properly*

formed. This can be helped, for instance, by applying a modeling approach in the analysis of response propensity. Criteria can also be set for the quality of cells. It is helpful, for instance, if cells are homogeneous and if the behaviour of outcome variables is the same for the nonrespondents and the respondents of each cell.

- *Regression imputation is not recommended as a general method for as massive a study as the Household Survey. Rather, the method lends itself to correcting the nonresponse error of individual variables central to the survey. Advance production of ready-imputed results for a situation like this is, however, difficult, for the application of the method, i.e. specification of the regression model, is also affected by the objective of the study. For this reason, even information concerning nonrespondents or the whole population should, as necessary, be made available to the person analysing the data, so that he or she can work out an imputation application suitable for the particular purpose in question. It is also possible to furnish the data with several imputation solutions, of which the user can choose the one that is best suited for the situation at hand.*

Nonresponse error is perhaps the most important systematic or **non-sampling error** that occurs in sample data. Sample data also contain other systematic errors. The procedures for eliminating different errors are not mutually exclusive, but one should, of course, be realistic and, before proceeding to correct any error, consider how high a stake would be worth the elimination of the error.

It is, of course, more important to correct large errors than small ones. On the other hand, all known errors that can be corrected should be corrected even though it may be known that some larger errors in the data will remain uncorrected. In the Household Survey, one such major error defying correction is the considerable undercoverage of alcohol consumption.

In analysing sample data, it is also important to measure **sampling errors**, i.e. the variance estimators or root mean square errors of estimators. Development of suitable estimators for the

purpose of correcting nonresponse errors turned out to be difficult. Two alternative estimators were obtained. The first of them was formed as an application of Cochran (1977, pp. 260-261) and the second on the basis of Särndal and Svensson (1987). In the absence of nonresponse, the estimators give the same values, but where nonresponse occurs, the latter gives higher values than the former. The Cochran-based estimator is applied in statistical production. Thus, the estimator of variance of consumption total Y for each stratum is as follows:

$$\hat{V}(\hat{Y}) = \frac{M - r}{M} \cdot r \cdot s^2 (w_k(A) \cdot y_k ; r)$$

where $s^2(.;.)$ denotes the usual sample variance of variable $w_k(A) \cdot y_k$ when the number of observations is r.

By summing up the variances for all the strata, the variance for the whole country is obtained. If every y_k in the expression equals one, the estimator of variance of the number of households $\hat{V}(\hat{N})$ is obtained. The variances of the mean values of consumption again are calculated from the following expression:

$$\hat{V}(\bar{\hat{Y}}) = \frac{1}{\bar{N}^2} \hat{V}(\hat{Y}) + \frac{\bar{Y}^2}{\bar{N}^4} \hat{V}(\hat{N}) - 2 \frac{\bar{Y}}{\bar{N}^3} \widehat{cov}(\hat{Y},\hat{N}),$$

where

$$\widehat{cov}(\hat{Y},\hat{N}) =$$

$$\frac{M - r}{M} \cdot \frac{r}{r - 1} \cdot \left(\Sigma w_k^2 y_k - \frac{\Sigma w_k y_k \Sigma w_k}{r} \right) .$$

0. Johdanto

Kotitaloustiedusteluilla on vuosisadan alkupuolelle asti ulottuva historia. Alkuvaiheissaan ne eivät koskeneet koko maata eivätkä olleet tietosisällöltään kovin laajoja. Vuodesta 1966 lähtien tiedustelut ovat olleet yleistettävissä koko maahan ja käytetyt käsitteet ja laskentatapa ovat perustuneet YK:n suosituksiin. Tiedustelujen otokset, tietosisältö ja keräystavat ovat sen jälkeenkin vaihdelleet.

Kotitaloustiedustelun kohdejoukkona ovat koko maan kotitaloudet. Sama kohdejoukko on myös vuodesta 1977 lähtien tuotetulla tulonjakotilastolla. Näillä kahdella tiedustelulla on yhtäläisyyksiä myös tietosisällöissä: samojen luokittelutietojen lisäksi molemmat sisältävät mm. samalla tavoin määriteltyjä tulotietoja.

Olisi hyvä, jos samasta asiasta saataisiin eri tiedusteluista yhdenmukaisia tuloksia. Tämä toive ei aina ole täyttynyt. Esimerkiksi vuoden 1981 kotitaloustiedustelun mukaan kotitalouksien lukumäärän estimaatti oli 1873000 kotitaloutta, kun vastaavana ajankohtana tulonjakotilasto (joka tulos on julkaisematon) kertoi kotitalouksia olevan yli 40000 enemmän ja seuraavana vuonna kotitalouksien lukumäärä oli kasvanut tulonjakotilaston mukaan 1958000:een.

Jotkut tulosten erot johtuivat käsitteiden erilaisesta määrittelystä ja soveltamisesta. Niiden yhtenäistäminen ei yksinomaan auttanut. Merkittävämmäksi erojen aiheuttajaksi todettiin kato, joka kotitaloustiedustelussa on ollut suurempi kuin tulonjakotilastossa. Tämä myös tilastojen käyttäjiä sekottava tekijä oli yksi vaikutin pyrkimyksille kehittää menetelmiä kadosta aiheutuvien virheiden korjaamiseen. Toisena keskeisenä tekijänä oli pelko kadon kasvusta tulevaisuudessa. Kolmantena muttei vähäisimpänä tekijänä on syytä mainita kansainväliset esikuvat, jotka osoittivat tämänsuuntaisen tutkimustoiminnan sekä lisääntyneen merkittävästi että nostaneen tilastojen laatua.

Ensimmäisenä konkreettisena toimenpiteenä hankittiin konsultti, jonka edellytettiin kadon lisäksi kiinnittävän huomiota otoksen poimintaan. Ehdotusten piti olla myös sellaisia, että niitä olisi voitu hyödyntää muissa vastaavissa otostilastoissa, kuten tulonjakotilastossa. Konsultti (Logit Ky 1984) ehdotti eräitä muutoksia otoksen poimintaan sekä erityisesti sitä, että ryhdyttäisiin käyttämään ns. regressioestimointia tulosten laskennassa.

Regressioestimoinnin soveltamisen tutkimiseksi Tilastokeskus palkkasi allekirjoittaneen 1985, mutta edellytti tutkittavan myös muita vaihtoehtoja. Monien vaiheiden jälkeen on päädytty ratkaisuihin, joita on jo sovellettu vuoden 1985 kotitaloustiedustelun estimoinnissa. Vastaava menetelmä on sisällytetty vuoden 1981 kotitaloustiedustelun aineistoon. Näin saadaan kohtuullisen vertailukelpoisia tuloksia aikavälin 1981-85 muutosten estimoimiseksi.

Konsultin ehdotuksella ja käyttöön otetulla sovellutuksella ei ole juuri samoja ominaisuuksia. Paljon on muutosta tapahtunut myös väliraportin julkaisemisen jälkeen (Laaksonen 1986). Raportti ja siihen saadut kommentit omalta osaltaan vauhdittivat uusien ajatusten esille tuloa. Toki on niin, että tämän raportin valmistumisen jälkeen aihepiirin teoreettinen ja käytännöllinen tutkimus tulee edelleen kehittymään ja uusien raporttien aika tulee pian vastaan.

Raportin rakenne on seuraava:

- Luvuissa 1-3 tarkastellaan kotitaloustiedustelun luonnetta, otoksen poimintaa ja katoongelmaa.

- Luvuissa 4-5 käsitellään kadon korjausmenetelmiä ja pohditaan niiden tarjoamia mahdollisuuksia kotitaloustiedustelun tyyppisissä aineistoissa.

- Luvuissa 6-7 esitetään kaksi menetelmää, joita tämän projektin puitteissa on kokeiltu kadon korjaamiseen. Tilastotuotantoon on näistä menetelmistä valittu luvussa 6 esitetty ratkaisu.

- Luvuissa 8-9 esitetään estimointituloksia kummallakin menetelmällä.

- Luvussa 10 on yhteenveto.

- Liitteenä esitetään keskeisten symbolien lisäksi estimointitilanteisiin liittyviä erityisnäkökohtia sekä mm. ajatuksia kadon korjausedellytysten kehittämiseksi. Kiintoisa on myös liite 7, jossa tarkastellaan vuosien 1981-85 tulomuutoksia eri menetelmillä estimoituna.

Tutkimusraportti käsittelee sellaista otantateorian lohkoa, jota on harrastettu Suomessa vähän. Tämä on hidastanut projektin etenemistä. Myös suomenkielisen raportin kirjoittaminen on eräiltä osin ollut vaikeata, koska monilla käsitteillä ei ole yleisesti hyväk-

syttyä suomennosta. Omat suomennokseni ja tulkintani onkin katsottava keskustelun käynnistämiseksi näistä tärkeistä aiheista.

Raportin ja sen pohjana olevien tutkimustulosten valmistuminen on ollut monimuotoinen prosessi. Kaikkia eri vaiheita on mahdotonta jälkeenpäin jäljittää esimerkiksi siksi, että voisi oikeudenmukaisesti kiitosta jakaa. Elintärkeitä ovat tietysti olleet keskustelut Tilastokeskuksen piirissä, erityisesti kotitaloustiedustelun mutta myös tulonjakotilaston ja koko sosiaalitilastolinjan henkilöstön kanssa. Arvostan myös apua, jota olen saanut tietojenkäsittelylinjasta ja haastattelutoimistosta. Hyödyllisiä ovat edelleen olleet alan kansalliset ja kansainväliset kokoukset, joihin olen ollut tilaisuudessa osallistua.

Tilastokeskuksen ulkopuolisista "neuvonantajistani" haluan erityisesti kiittää Erkki Pahkista Jyväskylän ja Anders Ekholmia Helsingin yliopistosta. VTT Pahkinen on lukenut sekä väliraportin että tämän raportin käsikirjoituksen ja antanut kommentteja jatkotyölle. Apulaisprofessori Ekholm toimi vuoden ajan Tilastokeskuksessa. Tänä aikana kehitettiin keskeiset osat menetelmästä, jota kotitaloustiedustelun tuotannossa on sovellettu (luku 6 kokonaan, osittain luvut 8-9). Myös huomattavaan osaan tästä raportista hän on antanut varteen otettuja kommentteja. Raportissa edelleen olevista puutteellisuuksista ja mahdollisista virheistä vastaan kuitenkin itse.

Tilastokeskuksessa elokuussa 1988

Seppo Laaksonen

1. Kotitaloustiedustelun luonteesta otannan ja estimoinnin kannalta

Kotitaloustiedustelua on tuotettu yleensä 5 vuoden välein. Sen tärkeimmiksi **tulosmuuttujiksi** voidaan katsoa kotitalouksien ja niiden jäsenten kulutusta ja kestokulutushyödykkeiden omistusta ja hankintaa sekä yhteiskunnallisten palveluiden käyttöä koskevat tiedot. Lisäksi tuotetaan tietoja mm. tuloista. Tiedustelujen yhteydessä kerätään myös muita tietoja esimerkiksi ulkopuolisia tilaaajia varten.

Tulosmuuttujien lisäksi tarvitaan tietoja eräistä ns. **taustamuuttujista**, joista tärkeimpiä ovat

- kotitalouden koko
- kotitalouden päämiehen ja kotitalouden jäsenten sosio-ekonominen asema
- asuinalue (lääni, kaupunki vs. muu kunta, mahdolliset muut aluejaot)
- päämiehen ja jäsenten koulutus
- päämiehen ja jäsenten ikä
- taustatietojen yhdistelmät so. niiden ristiinluokittelut.

Aineisto kootaan **rekistereistä ja haastattelujen sekä tilinpidon avulla** (ks. liitteen 1 kuvio 1). Ne tiedot, jotka ovat kohdistettavissa talouden yksittäiselle jäsenelle (esim. tulot rekisteristä tai vaatehankinnat tilinpidosta), kerätään jäsenkohtaisesti. Monet tiedoista ovat kuitenkin sellaisia, joita ei useamman kuin yhden jäsenen talouksissa voida kohdistaa yksittäiselle jäsenelle (esim. asunto- ja kalustomenot) ja ne kerätään vain talouskohtaisesti.

Tuloksia on siis saatavissa sekä jäsenkohtaisesti että talouskohtaisesti. Useimmat summamuuttujat, kuten esimerkiksi kokonaiskulutus tai käytettävissä oleva tulo, ovat tuotettavissa ainoastaan talouskohtaisesti. Otanta- ja estimointimenetelmillä tulee saada myös jäsenkohtaista tietoa. Siitä ei kuitenkaan tässä tapauksessa aiheudu ongelmaa, koska talouskohtaiset

pohjatiedot riittävät myös jäsenkohtaisten tietojen tuottamiseen. Tästä syystä tässä raportissa ei käsitellä laajemmin jäsenkohtaisten tulosten estimointia, vaan pitäydytään kotitalouskohtaisiin tarkasteluihin.

Estimointijärjestelmän tulee olla sellainen, että samasta aineistosta voidaan esim. tilastollisten ohjelmistojen avulla tuottaa kaikki tarpeelliset tiedot. Tärkeimpiä tunnuslukuja ovat

- **kokonaissummat** eli **totaalit** ja
- **keskiarvot,**

jotka tulee tuottaa tarpeellisten taustamuuttujien mukaan.

Myös **muut jakaumatiedot,** kuten esimerkiksi jakauma kokonaisuudessaan ja desiili- ym. osuuspisteet, ovat tärkeitä. Niiden luotettavuus on kuitenkin yksittäisten kulutustietojen osalta huonohko, koska tiedot on yleensä kerätty varsin lyhyeltä aikaväliltä. Jakaumatietojen luotettavuus on parempi rekisteritietojen ja summatietojen osalta. Kotitaloustiedustelua on yhä enemmän käytetty (ks. esim. Sullström 1987 tai Nordberg 1987) riippuvuuksien (korrelaatiot, regressiot yms.) tutkimiseen. Niihin sisältyy samantapaisia ongelmia kuin jakaumatietoihinkin.

Aineiston keskeisistä osista Tilastokeskus tuottaa julkaisuja. Tämän lisäksi aineistosta on tuotettu taulukoita ym. ulkopuolisten asiakkaiden käyttöön. Aineiston supistettuja osia on luovutettu Tilastokeskuksen ulkopuolelle. Nämä kaikki tekijät edellyttävät estimointimenetelmiltä helppokäyttöisyyttä ja sitä, että eri käyttäjät tuottavat aineistosta yhdenmukaisia tuloksia. Edelleen on tärkeätä, että peräkkäiset tiedustelut ovat estimointimenetelmän osalta yhdenmukaiset.

Tietoja käytetään Tilastokeskuksen sisällä muiden tilastojen tarpeisiin. Kuluttajahintaindeksin painojen muodostamisessa pohjana ovat kotitaloustiedustelun kulutustiedot. Myös kansantulotilasto käyttää tietoja hyväkseen, vaikka onkin huomattava, että täydelliseen yhteensopivuuteen näiden kahden tilaston kesken ei päästä mm. keruutapaerojen takia.

Tärkeitä ovat yhteydet tulonjakotilastoon, jonka tulokäsite on periaatteessa sama kuin tässä tiedustelussa. Näiden tiedustelujen pitäisi antaa mahdollisimman yhdenmukaisia tietoja kotitalouksien lukumäärästä ja monista taustatiedoista. Esimerkiksi po. kotitaloustiedusteluvuonna 1985 ei tehty ollenkaan tulonjakotilastoa ja onkin luontevaa odottaa, että kotitaloustiedustelun tietojen avulla "jatketaan" vastaavia aikasarjoja tähän ajankohtaan.

Kappaleen lopuksi muttei tärkeysasteeltaan vähäisimpänä on syytä korostaa tarvetta tuottaa tietoja estimaattien luotettavuudesta eli tulosten laadusta. Tätä varten on hyvin monia eri keinoja käytettävissä (ks. esim. Tilastokeskus 1987). Tässä raportissa tarkastellaan vain kahta laatutekijää:

- ns. **otantavirhettä** ja
- **ei-otantavirheistä** (engl. non-sampling errors) katoa.

Jälkimmäisen virheen ja sen korjaamisen käsittely on raportin keskeinen tavoite, jonka vuoksi sitä ei tässä yhteydessä tarkastella enempää. Otantavirheeksi katsotaan estimaatin varianssin (tai keskihajonnan) estimaatti, joka lasketaan sekä absoluuttisena että suhteessa estimaattiinsa. Tällaiset estimaatit kertovat, sitä paremmin mitä vähemmän harhaisia estimaatit ovat, keskiarvon tai totaalin tarkkuudesta ja ovat tulosten tulkitsijan kannalta tärkeitä. Niitä täytyy voida laskea koko maasta, alueittain ja taustamuuttujittain. Aihetta käsitellään tarkemmin raportin luvuissa 6, 7 ja 9 (ks. myös liite 2).

2. Kotitaloustiedustelun otoksen poiminta ja estimointi 1981 sekä otoksen poiminta 1985

Tiedustelun kohdeperusjoukon muodostavat Suomen kotitaloudet. Kotitalouksien muodostamaa kehikkoa ei kuitenkaan ollut käytettävissä, vaan **henkilöpohjainen** väestön keskusrekisteri (ks. myös Tilastokeskus 1986 ja Logit Ky 1984). Vuoden 1981 tiedustelussa **kehikkoperusjoukko** muodostui yli 15-vuotiaista kotitalouksissa asuneista henkilöistä, so. kaikista muista paitsi laitoksissa asuneista ja osoitteettomista henkilöistä. Vuoden 1985 tiedustelussa kehikko oli muuten sama, mutta mukaan otettiin kaikki yli 0-vuotiaat.

Kehikosta poimittiin ensiksi ns. **kohdehenkilöt**, joille etsittiin ns. viitehenkilöt. Viitehenkilön etsinnässä oli pääkriteerinä asuminen samassa huoneistossa kohdehenkilön kanssa. Näistä viitehenkilöiksi kelpasivat mm. kohdehenkilön lapset, vanhemmat ja puolisot. Muodostetut ryppäät vastaavat likimain ns. **asuntokuntia** (vrt. esim. Tilastokeskus 1983). Asuntokunnan jäsenille kerättiin poimintavaiheessa väestörekisteristä tärkeimmät väestötiedot ja myöhemmin myös muita rekisteritetoja.

Kohdehenkilöiden valinta suoritettiin vuonna 1981 tasaväliotantaa käyttämällä, siis sama poimintaväli yli kehikon. Vuonna 1985 poimintavälit vaihtelivat osittaittain, joita oli 25. Ositteet muodostettiin läänin ja kuntamuodon perusteella ja lisäksi otettiin Uusimaalta omaksi ositteeksi Pääkaupunkiseutu. Ahvenanmaan kahdella alkuperäisellä ositteella olivat poimintavälit likimain samat ja myöhemmin nämä ositteet onkin yhdistetty. Tästä syystä myöhemmin tässä raportissa käsitellään vain 24 ositetta.

Vuoden 1981 aineistolle muodostettiin estimointivaiheessa ns. **jälkiositus**, jolloin käytettiin hyväksi katotietoja sekä kuntien ominaisuuksia kuten sijaintia ja elinkeinorakennetta. Jälkiositteita muodostettiin 35 kpl. Vuoden 1985 aineistoon ei tehty jälkiositusta, mutta poimintaan muodostetun "etukäteisosituksen" suunnittelussa käytettiin hyväksi kokemuksia vuoden 1981 jälkiosituksesta. Vuoden 1985 aineis-

tossa on sovellettu menetelmiä, jotka ilmenevät raportin luvuista 6, 8 ja 9.

Estimoinnin kannalta ovat otosaineistoissa keskeisiä korotuskertoimet. Ne muodostettiin vuoden 1981 aineistossa seuraavalla *Horvitz-Thompson-tyypin* kaavalla (Horvitz ja Thompson 1952, ks. myös Cochran 1977, 259-261 sekä Tilastokeskus 1986, 14):

$$w_k(O) = \frac{M}{r \cdot m_k} \qquad (1)$$

jossa M = otoskehikkoon kuuluvien henkilöiden lukumäärä (siis yli 15-vuotiaat suomalaiset jotka eivät olleet laitoksissa)

r = haastateltujen kotitalouksien lukumäärä

m_k = otoskehikkoon kuuluvien henkilöiden lukumäärä kotitaloudessa k.

Kaavan (1) kerroin laskettiin jokaiselle ositteelle so. jälkiositteelle erikseen. Kukin kerroin ilmaisee, kuinka montaa kotitaloutta po. kotitalous vastaa perusjoukon tasolla. Siten kertoimien summaaminen yhden ositteen yli kertoo estimaatin ositteen kotitalouksien lukumäärälle. Vastaavasti summaaminen voidaan ulottaa yli kaikkien ositteiden, jolloin saadaan koko maan kotitalouksien lukumäärän estimaatti.

Erityisryhmiä, kuten yrittäjiä tai 3 hengen kotitalouksia, koskevat laskelmat tehdään siten, että summaukseen otetaan mukaan vain tämän ehdon mukaiset kotitaloudet. Nämä näkökohdat on kaavojen muodossa esitetty em. raportissa (Tilastokeskus 1986). Estimaattien keskihajontojen estimaattien laskentakaavat vuoden 1981 tiedustelussa on esitetty samassa raportissa. Niistä käytetään tavallisesti lyhyempää nimitystä **"keskivirhe"**.

Lausekkeen (1) mukaista korotuskerrointa on käytetty vuositason tietojen laskemiseen. Tilinpitotietojen estimoinnissa on kuitenkin käytetty ns. tilinpitopohjaista kerrointa, jonka soveltamisesta sekä vuoden 1981 että vuoden 1985 tiedustelun yhteydessä on esitetty selostus **liitteessä 3.**

Vuoden 1985 aineiston käsittelyyn vaikutti myös se, että aineistosta poistettiin sellaiset taloudet, joita oli kolmen edellisen vuoden aikana haastateltu Tilastokeskuksen kahdessa muussa keskeisessä tiedustelussa, so. tulonjakotilastossa ja työvoimatiedustelussa. Tästä hyväksyttävästä toimenpiteestä aiheutui kuitenkin ongelmia, koska tiedustelujen otoskehikot ovat erilaisia. Sen vuoksi korotuskertoimiin tehtiin ns. ikäkorjaus, jonka perusteet on selostettu liitteessä 4.

3. Kadon määrittelyt, mittaaminen ja esimerkkejä

Kato (engl. non-response tai nonresponse)(1) määritellään haastatteluaineistoissa siten kuin sen englanninkielinen versio osoittaa "vastaamattomuutena". Tilastokeskuksen tilastojen laatua koskevan käsikirjan (1987) mukaan katoon kuuluvat "täysin vastaamisesta kieltäytyneet ja tavoittamattomat sekä jokaisen muuttujan suhteen täysin käyttökelvottomia vastauksia antaneet kohdeperusjoukon alkiot."

Käsikirja erittelee edelleen osittaisen kadon (engl. partial non-response), jolloin kyse on vastaamattomuudesta yhteen tai useampaan muuttujaan. Osittaisen kadon ja kadon summaa käsikirjassa nimitetään kokonaiskadoksi.

Käsikirjan määrittelyt ovat vähän horjuvia. Olisihan luonnollisempaa pitää käsitettä "kato" yleisterminä, jolla on "alakäsitteitä". Käsikirjan määrittely kadosta tarkoittaa itse asiassa **"yksikkökatoa" tai "totaalikatoa"** (engl. unit non-response tai total non-response). **Osittaiskadon** käsite sen sijaan on selkeämpi mutta voidaan puhua myös muuttujakohtaisesta kadosta tai **"eräkadosta"** (engl. item non-response).

Katoa, olkoonpa kysymys yksikkö- tai eräkadosta, voidaan luokitella sen syyn mukaan. Luokituksia on tehty myös Tilastokeskuksen aineistoille. Seuraavassa esitetään esimerkkinä vuoden 1985 kotitaloustiedustelun luokitus ja sen frekvenssit alkuhaastatteluvaiheessa:

	kpl	%
- kohde kieltäytyi antamasta haastattelua	2065	81
- hylätty puutteellisten vastausten takia	3	0
- henkilö pysyvästi sairas eikä pystty antamaan haastattelua	68	3

---

(1) Suomen kieleen soveltuu myös nimike vastauskato, jota esimerkiksi Pahkinen (1986) käyttää.

- henkilö tilapäisesti sairas	12	0
- kohdetta ei tavattu, asuinpaikka tiedetään	231	9
- osoite tuntematon, asuinpaikkaa ei löydetä	45	2
- kohde tilapäisesti poissa esim. opiskelun tai työkomennuksen takia	27	1
- kohde matkoilla tai kesälomalla	7	0
- kohde tilapäisesti laitoksessa	3	0
- muu syy	20	1
- hylätty erikseen määritellyn syyn takia	63	2
- haastattelua ei voitu tehdä kielen vuoksi	4	0
- kenttätyöaika loppui kesken	6	0
- kohdetta ei yritetty tavoittaa haastattelijan eroamisen yms. syyn takia	4	0.

Tämä katosyiden luokittelu ei sisällä selkeää hierarkiaa esimerkiksi ottaen huomioon kadon syntymisjärjestyksen. Tällainen voitaisiin muodostaa Mokkenin (1987) esittämällä tavalla. Hänen katoluokituksensa sisältää neljä pääryhmää seuraavassa järjestyksessä (alaryhmiä ei tässä yhteydessä tarkastella, ks. emt.):

1. vastaajan osoitetta ei tiedetä tai se on väärä
2. vastaaja ei ole kotona
3. vastaaja ei ole halukas vastaamaan
4. vastaaja ei ole kyvykäs vastaamaan.

Tämä hierarkinen ryhmittely ilmaisee samalla, kuinka kato määräytyy. Esimerkiksi osoitteen puuttuminen tai sen virheellisyys, jollei korjausta tehdä, aiheuttaa jo automaattisesti kadon, vaikka vastaaja olisi kotona ja halukas sekä kyvykäs vastaamaan. Sen vuoksi ensimmäisen katosyyn poistamiseksi kannattaa nähdä paljon vaivaa eli etsiä oikea osoite selville. Tällöin voi lisäksi osoittautua, että ao. talous on muuttanut maasta tai kuollut, jolloin taas taloutta ei pitäisi lukeakaan katoon vaan ylipeittoon (ks. lisäpohdiskelua tästä aiheesta liitteestä 5).

Estimointivaiheessa voidaan harvoin vaikuttaa kadon määrään tai laatuun. Estimointi joudutaan sopeuttamaan syntyneeseen tilanteeseen ja käytettävissä olevaan aineistoon. Kotitaloustiedusteluun liittyy kuitenkin erikoispiirteenä tietojen keruun kolmivaiheisuus ja vastaavasti kolme eri katoa, so. kato **alkuhaastattelu-, tilinpito- ja loppuhaastatteluvaiheissa.**

Liitteen 1 kuvio 1 kertoo, kuinka kato vuoden 1985 tiedustelussa lisääntyi, kun haastatteluja lisättiin. Koska aineiston eri muuttujaryhmillä on eri suuri kato, on ongelmana myös, mistä aineistosta tulokset lasketaan. Yksittäistä muuttujaa mitattaessa olisi ilmeisesti järkevää käyttää kaikkea informaatiota, joka tästä muuttujasta on saatavissa. Siis esim. alkuhaastattelutietojen laskennassa käytettäisiin 9220 havaintoyksikköä, tilinpitotietojen laskennassa 8500 havaintoyksikköä ja loppuhaastattelutietojen laskennassa 8200 havaintoyksikköä. Tämä aiheuttaisi kuitenkin yhteensopivuusongelmia, ja on päätettävä, menetelläänkö näin vai lasketaanko kaikki tulokset yhdestä aineistosta, jollaiseksi tässä tapauksessa soveltuisi 8200 havaintoyksikön aineisto.

Kotitaloustiedustelun tulosten laskennassa on yhteensopivuus asetettu suurempaan arvoon kuin yksittäisen muuttujan tuloksen tarkkuus, josta syystä perusaineisto on muodostettu 8200 havaintoyksikön perusteella. Tätä päätöstä on puoltanut myös se tieto, että tilinpitovaiheessa katoon tulleen talouden ilmoitukset ovat olleet keskimääräistä puutteellisempia jo alkuhaastatteluvaiheessa ja loppuhaastatteluvaiheessa katoon tulleen talouden ilmoitukset jo tilinpitovaiheessa keskimääräistä puutteellisempia. Osittaiskatotalouksia koskevaa aineistoa ei silti ole kokonaan jätetty hyväksi käyttämättä, vaan sitä on hyödynnetty katokorjauksessa.

Kuvio 1 kertoo, että vastanneita vuonna 1985 oli alkuhaastatteluvaiheessa 78,3%, tilinpitovaiheessa 72,2% ja loppuhaastatteluvaiheessa 69,6%. Kaikkia lukuja voidaan pitää matalina, erityisesti viimeistä, jota käytetään varsinaisten tulosten laskennassa. Kuitenkin aineisto sisältää monia tietoja myös 30,4%:sta katotaloudesta:

- kotitalouden koostumus sekä eräät sosioekonomiset ja demografiset tiedot on tarkistettu alkuhaastattelun antaneiden talouksien osalta
- rekistereistä saatuja tulo-, verotus-, koulutus- ja varallisuustietoja on olemassa myös katoaineistosta (ks. liite 6).

Katoaineistoa voidaan käyttää hyväksi pyrittäessä parantamaan estimaattien tarkkuutta. Mahdollisia menetelmiä tähän tarkoitukseen on runsaasti. Luvussa 4 luodaan yleiskatsaus katokorjausmenetelmiin.

4. Katsaus katokorjausmenetelmiin

Katoa tai laajemmin sanottuna **puuttuvan tai epätäydellisen** datan käsittelyä survey-aineistoissa on viimeaikaisessa kirjallisuudessa käsitelty varsin paljon (ks. esim. Little 1986, Little ja Rubin 1987, Särndal 1986, Särndal ja Svensson 1987, Michaud 1986, Rao 1986, Kalton ja Kasprzyk 1986, Platek ja Grey 1986, Rubin 1986, Giles ja Patrick 1986, Trembley 1986, Williams ja Nisselson 1987, Pahkinen 1986). Syynä tähän on aihepiirin tuleminen entistä tärkeämmäksi käytännössä, so. survey-aineistojen käsittelyssä.

Tässä luvussa esitetään yleiskatsaus käytettävissä oleviin puuttuvan datan korjausmenetelmiin survey-aineistojen käsittelyssä. Menetelmien ryhmittely perustuu olennaisilta osiltaan Kanadan tilastoviraston aikakauskirjassa "Survey Methodology" julkaistuun **Kaltonin ja Kasprzykin (1986)** artikkeliin. Myöhemmin ovat samantyyppisen ryhmittelyn esittäneet Little ja Rubin (1987). Luokittelun pohjalta arvioidaan luvussa 5 mahdollisia lähestymistapoja kotitaloustiedustelun estimoinnissa käytettäväksi.

Kalton ja Kasprzyk (1986) jakavat puuttuvan datan käsittelyyn liittyvät tarkistus- ja korjausmenetelmät KAHTEEN PÄÄRYHMÄÄN:

A. Painotukseen pohjautuviin ja

B. Imputointiin pohjautuviin menetelmiin.

Painotusmenetelmien he katsovat ensisijaisesti soveltuvan käytettäväksi korjaamaan yksikkökadon aiheuttamia virheitä. Vastaavasti imputointi(1) soveltuu parhaiten osittais- eli eräkadon virheiden korjaukseen. Täysin selkeänä tätä käyttöä koskevaa jakoa he eivät kuitenkaan pidä.

---

(1) Sanan suomennoksia olen kuullut kaksi: "sijaistaminen" jota on käytetty Tilastokeskuksessa ja "paikkausmenetelmä" tai "paikkaaminen" jota on ehdottanut E. Pahkinen Jyväskylän yliopistosta. Näistä edellinen kattaa minusta vain pienen osan imputoinnista ja jälkimmäinenkään ei kata kaikkia sen ominaisuuksia. Tästä syystä käytän yleisterminä sanaa "imputointi" tämän raportin jatko-osissa.

Kummassakin menetelmässä käytetään hyväksi varsinaisen aineiston lisäksi ulkopuolista informaatiota.(1) Menetelmien erot perustuvat toisaalta tämän informaation erilaiseen käyttötapaan ja toisaalta aineiston erilaiseen muodostamiseen:

- Painottavia menetelmiä käytettäessä aineisto koostuu vain niistä, joista kaikki tarvittavat tiedot ovat olemassa. Siis kadosta tms. syystä puuttuva datan osa ei ole estimoinnissa mukana. Menetelmien avulla ja siis apuinformaatiota hyväksi käyttämällä tuotetaan kullekin havainnolle parhaat mahdolliset painot, jotka voidaan tulkita korotuskertoimiksi. Tästä painotuksesta käytetään usein nimikettä **uudelleenpainotus**, mikä ilmaisu havainnollistaa sitä, että ilman apuinformaatiota muodostetut painot ovat "vanhoja".

- Imputointimenetelmiin liittyvä aineisto on tai voidaan ajatella sellaiseksi, että se sisältää myös puuttuvat tiedot. Jotta estimointi voitaisiin puuttuvan datan osalta suorittaa, on tilalle tuotettava sopivat sijais-, korvike- tms. arvot. Näiden arvojen muodostamiseksi käytetään hyväksi apuinformaatiota.

4.1. Painottavat menetelmät

Kalton ja Kasprzyk (1986) jakavat painottavat menetelmät neljään alaryhmään:

A1. Painot perustuvat **perusjoukosta** olevaan tietoon (engl. population weighting adjustments)

A2. Painot perustuvat poimitusta **otoksesta** saatavissa olevaan tietoon (engl. sample weighting adjustments)

---

(1)
Muita nimikkeitä ovat lisäinformaatio, oheisinformaatio ja apuinformaatio tai informaatio lisämuuttujista, informaatio apumuuttujista ja informaatio x-muuttujista, jolloin viimeksi mainitussa tapauksessa varsinaiset tutkittavat muuttujat ovat y-muuttujia.

A3. Painot perustuvat **sekä otoksesta että perusjoukosta saatavissa olevaan tietoon** (engl. raking ratio adjustments)(1)

A4. Painot perustuvat **vastaustodennäköisyyksiin** (engl. weighting with response probabilities).

Menetelmän A1 soveltamiseksi täytyy olla käytettävissä apumuuttujien $x_1, x_2, \ldots, x_q$ kautta tietoa sekä vastanneista että perusjoukosta. Apumuuttujat luokitellaan yhteen tai useampaan luokkaan. Jos esimerkiksi muuttuja x_1 on luokiteltu c_1:een, muuttuja x_2 c_2:een ja x_q c_q:een luokkaan, ja halutaan käyttää näitä kaikkia muuttujia uusien painojen muodostamiseen, saadaan ristiinluokittelulla luokkia tai soluja yhteensä $c_1 \cdot c_2 \cdots c_q$ kpl. Kustakin tällaisesta solusta tuotetaan perusjoukon tiedot poimintakehikkoon kuuluvien tilastoyksikköjen lukumääristä (kaavan (1) tapauksessa lukumääristä M). Näitä lukuja käytetään vastaavasti kunkin otossolun uusien painojen muodostamiseen (kaavan (1) tapauksessa luvut m_k ja n määrätään solujen havaintoyksiköitä koskevien tietojen avulla).

Menetelmää A1 nimitetään usein **jälkisitukseksi** (engl. poststratification), jolloin siis edellä esitetyllä tavalla saatuja soluja kutsutaan jälkio-

---

(1)
Käsite on tässä yhteydessä suomennettu samassa hengessä kuin kaksi ensimmäistä menetelmää, koska hyvää suoraa käännöstä sanalle "raking" (rake=haravoida, harata; rake up=haalia kokoon; rake=olla kallellaan; ym.) ei ole keksitty.

sitteiksi. Kalton ja Kasprzyk (1986) katsovat kuitenkin, että menetelmä A1 on yleisempi kuin jälkiositusmenetelmä. Jälkimmäistä nimitystä tulisi käyttää heidän mukaansa vain silloin, kun solujen pohjalle rakentuva uudelleenpainotus pyrkii poistamaan otantavirhettä so. otoksen satunnaisvirheitä. Jos sen sijaan menetelmällä tähdätään myös systemaattisten virheiden, kuten kadon, korjaamiseen, menetelmästä ei tulisi käyttää nimitystä jälkiositus. Oman käsitykseni mukaan tällainen erottelu ei käytännön tilanteissa ole tarpeen, vaan menetelmä A1 ja jälkiositus voidaan vallan hyvin rinnastaa toisiinsa.

Menetelmä A2 perustuu samanlaiseen solujen muodostamisstrategiaan kuin menetelmä A1. Tässä tapauksessa painot otetaan kuitenkin samasta otoksesta, siis poimitusta otoksesta. Kalton ja Kasprzyk (1986) katsovat, että tämä menetelmä muistuttaa kaksivaiheista otantaa (engl. two-phase sampling). Tällöin ajatellaan, että menetelmä sisältää ensimmäisessä vaiheessa totaaliotoksen vastaajista ja ei-vastaajista sekä toisessa vaiheessa osaotoksen vastaajista.

Menetelmillä A1 ja A2 on erilaiset aineistovaatimukset ja tästä johtuen myös erilaiset harhanlähteet. Jos molempien menetelmien käyttöä varten on aineistoa, voidaan myös menetellä niin, että sovelletaan molempia: ensiksi menetelmää A2 ja sitten menetelmää A1. Kummankin menetelmän ongelmana on se, että jos painottavia luokkia tai soluja muodostetaan paljon, voivat havaintomäärät joissakin luokissa jäädä vähäisiksi, josta taas seuraa painojen suuri vaihtelu sekä tulosten luotettavuusongelmia.

Menetelmä A3 voidaan katsoa menetelmien A1 ja A2 välimuodoksi. Siinä käytetään toisaalta hyväksi tietoa perusjoukosta ja toisaalta vastanneista. Olkoon meillä esimerkiksi taustatietoa kahden apumuuttujan so. x_1:n ja x_2:n kautta. Näistä edellinen olkoon luokiteltu luokkiin F1 ja F2 ja jälkimmäinen luokkiin E1, E2 ja E3. Tällöin meillä on käytettävissä seuraavan taulukon mukainen lähtötieto:

		MUUTTUJA x_2			
		E1	E2	E3	Yhteensä
MUUT-TUJA x_1	F1	n_{11}	n_{12}	n_{13}	$N_{1\cdot}$
	F2	n_{21}	n_{22}	n_{23}	$N_{2\cdot}$
	Yhteensä	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	$N_{\cdot\cdot}$

jossa

n_{ij} = otoslukumäärät solussa ij

$N_{i\cdot}$ = perusjoukon lukumäärät muuttujan x_1 luokissa

$N_{\cdot j}$ = perusjoukon lukumäärät muuttujan x_2 luokissa.

Tässä tapauksessa meillä ei siis ole käytettävissä perusjoukkoa koskevia tietoja soluista so. lukuja N_{ij}, mutta sen sijaan niiden reunajakaumista so. luvuista $N_{i\cdot}$ ja $N_{\cdot j}$. Nämä luvut on sen vuoksi estimoitava. Estimointimahdollisuuksia on useita. Tavallisin lienee menetelmä, jonka ovat esittäneet sekä Kalton ja Kasprzyk (1986) että Little ja Rubin (1987). Siinä estimointi perustuu seuraaviin kahteen ehtoon:

(i) Perusjoukon reunajakaumat pysyvät voimassa.
(ii) Otoksessa havaitut yhdysvaikutukset pätevät myös perusjoukossa seuraavaan tapaan:

$$\frac{N_{11} \cdot N_{22}}{N_{12} \cdot N_{21}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

Näitä ehtoja hyväksi käyttämällä voidaan tuntemattomille N_{ij}-luvuille määrätä estimaatit ja muodostaa

vastaavasti uudet painot. Estimaattien määrääminen vaikeutuu apumuuttujien ja ristiinluokittelulla saatujen solujen määrän lisääntyessä.

Menetelmä A4 pyrkii hyödyntämään tietoja vastaajaehdokkaiden vastaustodennäköisyyksistä. Oletetaan, että kaikilla perusjoukon jäsenillä on jokin todennäköisyys vastata survey-tiedusteluun. Nämä todennäköisyydet estimoidaan ja estimaattien käänteislukuja käytetään hyväksi painojen määräämisessä. Menetelmän käytännön soveltamisessa on monia vaihtoehtoja. Kalton ja Kasprzyk (1986) esittävät kaksi olennaisesti erilaista tilannetta:

- Menetelmän varhaisen sovellutuksen vuosilta 1949 ja 1950 mukaan haastattelija tiedustelee vastaajalta myös, kuinka monena 5:stä edellisestä illasta samaan aikaan henkilö oli ollut kotona. Tältä pohjalta saadaan vastaustodennäköisyys estimointia varten. Tämä menetelmä koskee siis vain yhtä kadon syytä. Ongelma on myös se, että niiltä, jotka kuuluivat katoon, ei saatu tätä informaatiota.

- Toisena lähestymistapana esitetään "vastausstatuksen" regressointia jossa selitettävinä on muuttujia, joista on tietoja sekä vastaajista että ei-vastaajista. Mahdollisina malleina esitetään logistista tai probit regressiota. Mallin avulla laskettujen vastaustodennäköisyyksien käänteisluvut olisivat pohjana painotukselle.

4.2. Imputointi

Kaltonin ja Kasprzykin imputointimenetelmien kuvaus on varsin mittava alkaen yksinkertaisista ja edeten monimutkaisempiin. Useimmat näistä menetelmistä voidaan periaatteessa esittää seuraavan yleisen regressiomallin erikoistapauksina:

$$y_{mi} = b_{ro} + \sum b_{rq} x_{miq} + e_{mi}, \qquad (2)$$

jossa

y_{mi} = imputoitu arvo i:nnelle puuttuvalle y-arvolle

x_{miq} = x-muuttujaa koskeva arvo vastaavasti

b_{rq} = regressiokerroin

b_{ro} = regressiovakio

e_{mi} = residuaali.

m = ao. tieto on aineiston puuttuvasta osasta

r = ilmoittaa, että ao. estimaatti on määrätty vastaajien aineistosta

q = kyseessä q:s mallin selittäjä.

Malli antaa sen eri osia koskevin oletuksin seuraavat imputointimenetelmät (ei täysin vastaa Kaltonin ja Kasprzykin esitystä):

B1. **Deduktiivisessa imputoinnissa** on pyrkimyksenä päätellä oheistiedon avulla tai loogisella päättelyllä puuttuvan muuttujan arvo. Asian havainnollistamiseksi otetaan kaksi esimerkkiä:

(i) Jos taloudessa on alle 16-vuotias lapsi, voimme määrätä lähes tarkasti, kuinka paljon lapsilisää talous vuosittain saa.

(ii) Jos tutkintorekisteristä havaitsemme talouden päämiehen olevan akateemisen loppututkinnon suorittanut ja verorekisteristä, että hänen veronalaiset palkkatulonsa ovat 150000 mk ja muut tulonsa 20000 mk, voimme melko varmasti katsoa hänet ylempiin toimihenkilöihin kuuluvaksi.

Menetelmää voidaan soveltaa kaikessa tilastoaineistojen tarkistusvaiheessa, mutta laajasti eli olennaista osaa puuttuvasta aineistosta ei tällä menetelmällä ole mahdollista hyvin korjata. Yleismalli (2) voi kuvata menetelmää B1 eri tavoin: pohjana on joka tapauksessa muuttujista x saatavissa oleva ja niiden tunnettu yhteys muuttujaan y.

B2. **Keskiarvoimputointi** asettaa puuttuviksi arvoiksi vastaajilta saadut keskiarvot. Tätä keskiarvoistamista voidaan soveltaa koko aineistoon kerralla tai erillisesti sopivasti määriteltyihin soluihin tai luokkiin. Solujen muodostamisen ehtona on tieto puuttuvasta muuttujasta. Malli (2) antaa menetelmän B2, jos e=0 ja x on dummy-muuttuja. Regressiomallin yleiskeskiarvo eli vakiotermi on kussakin solussa kyseinen imputoitu arvo. Jos dummy-muuttujia ei olisi, imputoitu arvo olisi koko aineistolle lasketun mallin vakio eli keskiarvo.

B3. **Satunnaisimputointia** voidaan soveltaa myös koko aineistolle tai sopivasti valituille soluille erikseen. Soveltamismahdollisuuksia on useita. Lähtökohtana on, että puuttuva havaintoyksikkö korvataan satunnaisesti valitulla havaintoyksiköllä vastaajista. Jos tätä sovelletaan hyvin valittuihin soluihin, menetelmä voi olla varsin onnistunut. Mallimielessä kysymys on siitä, että menetelmän B2 lisäksi otetaan käyttöön residuaali e.

B4. **Hot-deck imputoinnissa**(1) lähdetään liikkeelle imputointisoluista tai -luokista, kuten menetelmien B2 ja B3 solusovellutuksissa. Kalton ja Kasprzyk (1986) esittävät kaksi hot-deck imputointimenetelmää: jaksottaisen ja hierarkisen.

Jaksottaisessa menetelmässä asetetaan ensin jokaiselle tiedolle kussakin solussa sopiva "varastoarvo", esimerkiksi edellisen tiedustelun perusteella. Kukin tieto käydään tämän jälkeen vuorotellen läpi. Jos vastaus on olemassa, tieto säilyy ennallaan, mutta varastoarvo muutetaan samalla tämän vastauksen mukaisesti. Jos vastausta ei löydy, otetaan tilalle senhetkinen varastoarvo.

Tämä hot-deck menetelmä on siten sama kuin satunnaisimputointi, jos vastaajien ja ei-vastaajien muodostama aineisto on satunnaisessa järjestyksessä. Järjestyksen asettamisella voidaan estimointia muutenkin tehostaa. Siihen tähdätään erityisesti **hierarkisella** hot-deck imputoinnilla, jossa järjestys muodostetaan yhden tai useamman x-muuttujan avulla.

---

(1)
Käytän "englanninkielistä" suomennosta, koska virallista suomennosta ei liene. Yksi mahdollinen suomennos olisi "kuumapakka"-imputointi.

Hot-deck imputoinnin lisäksi voi tulla kysymykseen myös **cold-deck imputointi** ("kylmäpakka" imputointi) (ks. esim. Little ja Rubin 1987, 60). Siinä puuttuva arvo korvataan vakiolla, joka on saatu esimerkiksi edellisestä tiedustelusta. Menetelmä tullee kysymykseen tiedusteluissa, joissa on aikaisempia tietoja samasta aineistosta olemassa, kuten panelitutkimuksessa. Tässä raportissa en ole asettanut cold-deck imputointia omaksi menetelmäkseen, koska sen voi katsoa olevan erikoistapaus jostakin muusta (hot-deck tai deduktiivinen imputointi) ja koska tyydyttävää teoriaa tämän menetelmän taustaksi ei mm. Littlen ja Rubinin mukaan ole olemassa.

B5. **Regressioimputoinnissa** imputoidut arvot muodostetaan regressiomallilla, jossa estimointi perustuu vastaajien aineistoon: x-muuttujien perustiedot täytyy tietää myös vastaamattomien aineistosta tai koko perusjoukosta. Mallista saadut ennustetut arvot asetetaan puuttuvien tietojen paikalle. Arvoihin voidaan lisätä myös residuaalit. Siihen on useita mahdollisuuksia esitetty myös po. artikkelissa. Malli (2) on perusominaisuuksiltaan menetelmän B5 mukainen.

B6. **Etäisyysfunktion käyttöön perustuva imputointi** (engl. distance function matching) on eräänlainen hot-deck menetelmä myös. Siinä etsitään kullekin vastaamattomalle sopiva korvaava havaintoyksikkö käyttäen hyväksi etäisyysfunktiota, toisin sanoen vastaamattomman tilalle asetetaan mahdollisimman samankaltainen havaintoyksikkö vastaajien joukosta. Etäisyysfunktio perustuu x-muuttujien arvojen vertailuun kunkin vastaamattoman ja vastanneiden kesken. Funktioiden muodostamismahdollisuuksia on runsaasti, riippuen käytettävissä olevien x-muuttujien määrästä ja niiden ominaisuuksista.

Esitetyn lisäksi voidaan imputointimenetelmiä jakaa muillakin perusteella. Rubin (1986 ja 1987) esimerkiksi puhuu **single-** (yksinkertaisesta) ja **multiple-** (moninkertainen, moni-, multippeli) imputoinneista. Single viittaa siihen, että sen avulla pyritään korvaamaan jokainen puuttuva havaintoarvo yhdellä arvolla. Multippeli imputoinnissa kukin puuttuva havaintoarvo korvataan vektorilla, siis useiden mahdollisten arvojen "luettelolla". Vektorin eri komponentteja käytetään erityyppisten täydennettyjen aineistojen muodostamiseen.

"Moni-imputoinnilla" tuotettuja vektoreita voidaan käyttää myös apuna määrättäessä estimaateille virheym. epävarmuusmittoja. Tällöin moni-imputointi voitaneen katsoa yhdeksi bootstrap-menetelmän sovellutukseksi (ks. esim. Hinkley 1988).

Imputointisovellutus voi myös olla eri imputointimenetelmien yhdistelmä. Esimerkiksi hot-deck ja regressio- imputointi voidaan yhdistää. Yksi mahdollisuus on laskea ensin regressiomallilla ennustetut arvot ja lisätä näihin satunnaistekijät hot-deck imputoinnilla, jolloin satunnaistekijät valitaan empiiristä residuaaleista. (ks. esim. Little ja Rubin 1987, 61).

4.3. Keskustelua

Kaltonin ja Kasprzykin esittämä puuttuvan datan korjausmenetelmien luokittelu kahteen pääryhmään on varsin onnistunut (vrt. myös Little 1986 sekä Little ja Rubin 1987). On kuitenkin huomattava, että erilaisen menetelmän käyttäminen ei välttämättä anna erilaisia tuloksia. Esimerkiksi Little (1986) toteaa, että sopivasti muodostettu painottava menetelmä ja samoja soluja hyväksi käyttävä keskiarvoimputointi tuottavat samat keskiarvoestimaattorit kullekin solulle ja myös koko perusjoukolle. Sen sijaan estimaattorit eivät aina ole samoja "ristiluokka"-ryhmissä, jollaisia voidaan muodostaa mm. sosioekonomisen ja koulutusaloittaisen luokituksen avulla.

Useat po. menetelmät käyttävät hyväkseen korjaussoluja tai -luokkia. Niiden muodostamiseen on useita mahdollisuuksia. Luvun alussa mainituissa artikkeleissa on esitetty teoreettisia ja käytännöllisiä perusteita solujen muodostamisen pohjaksi. Seuraavassa muutamia käytännön tilastotyön kannalta tärkeitä näkökohtia:

- Solut saisivat olla sisällöllisesti homogeenisia. Tällöin korjattujen estimaattien hajonta vähenee ja harha pienenee (ks. esim. Trembley 1986, 93).

- Käytössä tulee olla varovainen erityisesti, jos solun vastaustodennäköisyys on pieni. Esimerkiksi Chapman ym. (1986) pitivät pienenä alle 50%:n todennäköisyyttä, mutta tätä arviota ei voi pitää ainoana oikeana. Jos todennäköisyys on pieni, voidaan soluja yhdistää tai käyttää painojen muodostamisessa myös harkintaa.

- Solussa täytyy olla riittävästi havaintoja. Riittävä havaintojen määrä on harkinta- ja tilannekysymys. Ongelma tulee vastaan erityisesti silloin, kun solut perustuvat useamman x-muuttujan ristiinluokitteluun.

- Solujen sisällä pitäisi tulosmuuttujan ja vastaamisalttiuden olla toisistaan riippumattomia. Tämä saisi näkyä vähintään siten, että kussakin solussa tulosmuuttujan keskiarvo olisi sama sekä vastaajille että vastaamattomille (ks. Little 1986, 144-145).

5. Keskustelua katokorjausmenetelmien soveltuvuudesta kotitaloustiedusteluun

Tässä luvussa olen valinnut lähtökohdaksi edellisessä luvussa esitetyt korjausmenetelmävaihtoehdot, joiden jokaisen soveltuvuudesta kotitaloustiedustelun yhteydessä keskustellaan. Keskustelun taustaksi on hyvä muistaa myös luvuissa 0-3 esitetyt kotitaloustiedustelun tavoitteet yms. taustatekijät.

5.1. Painotus vs. imputointi

Ensimmäiseksi on syytä pohtia kahden päävaihtoehdon painotuksen ja imputoinnin soveltuvuutta kotitaloustiedusteluun. Tältä osin olen päätynyt seuraavaan johtopäätökseen:

> Imputointi, alkuperäisessä merkityksessään, ei tule kotitaloustiedustelussa perusmenetelmänä kyseeseen.

Tätä johtopäätöstä perustelen seuraavasti:

> Imputoinnin tarkoitus on korvata puuttuva havaintoarvo jollakin toisella. Kotitaloustiedustelussa 1985 tämä merkitsisi laajimmillaan sitä, että 30,4% havaintoyksiköistä pitäisi imputoida eli täydentää haastateltu 8200--9220 havaintoyksikön aineisto 11776 yksikön aineistoksi. Tällainen täydentäminen on periaatteessa mahdollista soveltamalla erityyppisiin muuttujiin erilaista imputointia. Tulokseksi saataisiin yleensä alkuperäisiä parempia estimaatteja totaaleista, keskiarvoista tai vastaavista suureista. Samalla kuitenkin muuttujien väliset yhteydet muuttuisivat, eikä olisi takeita siitä, että ne muuttuisivat oikeaan suuntaan. Tätä ei perustilastoaineistoissa kuitenkaan tulisi sallia, koska monet käyttäjät haluavat tutkia näitä yhteyksiä mahdollisimman todelliselta pohjalta, ei imputoinnin tuottaman keinotekoisuuden "lisämausteella".

Painottaviin menetelmiin ei liity po. ongelmaa, koska alkuperäiset haastattelusta saadut tiedot ovat aineistossa jatkuvasti käytettävissä, joskin niiden painotus muuttuu. Näin ollen painottavat menetelmät soveltuvat yleismenetelmäksi perustilastotuotantoon.

Imputointimenetelmiäkään ei silti ole syytä kokonaan hylätä. Niitä voidaan hyödyntää kahdella tavalla:

(i) Osalle muuttujista voi olla järkevää täydentää aineisto joko kokonaan tai osittain (osittain esim. niihin asti jotka olivat alkuhaastattelussa mukana). Voidaan myös menetellä niin, että imputointi toteutetaan kerran koko aineistolle mutta sovellutuksesta riippuen käytetään joko tätä tai suppeampaa aineistoa hyväksi (esim. keskiarvot, totaalit ja mahdollisesti osa jakaumatarkasteluista tuotetaan imputoidusta ja riippuvuusanalyysit yms. suppeammasta aineistosta).

(ii) Imputointimenetelmällä tuotetaan tarvittavat puuttuvat havaintoarvot, mutta näin saatua aineistoa ei käytetä sellaisenaan, vaan vastaava informaatio siirretään mahdollisimman hyvin vastaajien aineistoon. Tämä siirtäminen voidaan tehdä monin tavoin. Yksi mahdollisuus on se, että jokainen vastaajien aineiston havainto y_k kerrotaan seuraavalla korjaussuhteella:

$$\bar{y}_{oi}/\bar{y}_{o},$$

jossa $\bar{y}_o$ = muuttujan y keskiarvo vastaajien antamista todellisista havaintoarvoista laskettuna

$\bar{y}_{oi}$ = muuttujan y keskiarvo laskettuna vastaajien osalta todellisista ja ei-vastaajien osalta imputoiduista havaintoarvoista.

Tämä menetelmä voidaan katsoa suhde-estimoinnin sovellutukseksi. Se tuottaa alkuperäisiä paremmat keskiarvon ja totaalin estimaatit, mutta ei muuta muuttujien välisiä riippuvuussuhteita. Estimaattien tarkkuus voi parantua, jos korjaussuhteet estimoidaan osajoukoittain tai soluittain. Tällöin muuttuvat myös riippuvuussuhteet sitä enemmän mitä enemmän näitä osajoukkoja on.

5.2. Painottavat menetelmät

Kaltonin ja Kasprzykin (1986) esittelemistä painottavista menetelmistä kaikkia voi periaatteessa käyttää vuoden 1985 tyyppisessä kotitaloustiedustelussa. Menetelmän A1 voidaan katsoa olleen eräänlainen perusmenetelmä kaikissa tiedusteluissa: onhan korotuskertoimen pohjana aina ollut jokin tieto perusjoukosta (vrt. kaavan (1) symboli M).

Menetelmän A1 jälkiositusversiota on myös sovellettu. Esimerkiksi vuoden 1971 tiedustelussa (ks. Tilastokeskus 1977) muodostettiin väestölaskentatietojen perusteella 84 korotuskerrointa alueittain, sosio-ekonomisen aseman ja ruokakunnan koon mukaan. Vuoden 1981 tiedustelussa (Tilastokeskus 1986) taas muodostettiin 35 alueellista jälkiositetta.

Menetelmällä A2 voidaan täydentää menetelmän A1 tulosta käyttämällä poimittua otosta koskevaa tietoa. Vastaavasti voidaan soveltaa myös menetelmää A4, mutta nyt lähtökohtaksi otetaan nimenomaan vastaustodennäköisyydet poimitun aineiston perusteella määritellyissä soluissa.

Menetelmän A3 hyödyntämistä kotitaloustiedustelussa ei ole kovin paljon pohdittu. Sen hyvä piirre on nähdäkseni se, että soveltaminen ei edellytä tietoa aineiston kato-osasta, koska apuinformaatio otetaan suoraan perusjoukosta. Jos siis perusjoukosta ja vastaajien otoksesta olisi yhteensopivaa tietoa useiden sellaisten muuttujien kautta, jotka vaikuttavat katoon, menetelmän soveltaminen kannattaisi. Kotitaloustiedustelussa tätä yhteensopivuutta sotkee se, että perusjoukko koostuu henkilöistä ja otos kotitalouksista. Ongelma on myös solufrekvenssien

N_{ij} estimointi, joka tulee sitä hankalammaksi mitä

enemmän soluja on.

Totesimme edellisessä luvussa, että menetelmä A2 voidaan ajatella myös kaksivaiheisen otanta-asetelman mukaisena, jolloin toisen vaiheen otos poimitaan kato-osasta. Asetelman eräs sovellutus voisi olla kato-osan tutkiminen rekisteritietojen lisäksi erityishaastatteluilla, joka kohdistuisi osaotokseen kato-osasta. Näin saatua tietoa voidaan käyttää kahteen tarkoitukseen:

(i) Kato-osaa koskevien tietojen estimointiin. Tällöin estimaatteja voidaan olennaisesti tarkentaa, ellei synny uutta kato-ongelmaa ja otos on riittävän suuri (ks. esim. Rao 1986).

(ii) Muilla estimointimenetelmillä saatujen estimaattien tarkistuksiin, erityisesti harhan tutkimiseen. Esimerkiksi olisi mahdollista käyttää näitä tietoja soluja koskevien oletusten paikkansapitävyyden selvittämiseen (ks. kappaleet 4.3. ja 6.4.).

Tässä raportissa on painottavista menetelmistä erityistarkastelun kohteeksi valittu menetelmä A4, jota on sovellettu kotitaloustiedustelun 1985 tilastotulosten estimointiin. Sovellutuksen yksityiskohdat on esitetty luvussa 6.

5.3. Imputointi

Deduktiivisen imputoinnin B1 valinta yleismenetelmäksi ei ole järkevää, mutta sen rajoitettu käyttö eräkadon korjaamiseen sen sijaan on. Käytännön tilastotuotannossa tätä käytetään tietojen tarkistusvaiheessa nykyisinkin.

Yleiset, koko aineistoon kerralla sovelletut imputointimenetelmät, kenties regressioimputointia lukuunottamatta, eivät liene käyttökelpoisia kotitaloustiedustelun tapaisissa aineistoissa. Ainakin olisi käytettävä usean taustamuuttujan avulla tuotettua solujen ryhmää eli solukkoa, joka lajittelee havaintoyksiköt mahdollisimman homogeenisiin ryhmiin. Sovellettaessa tällä ehdolla keskiarvoimputointia B2 puuttuvat tiedot saisivat kussakin solussa saman arvon so. solukeskiarvon. Tämä ei kuvaisi oikein todellisuutta, koska jakauman painotus siirtyisi kohti keskiarvoa ja myös hajonta pienenisi. Silti tämä korjaus parantaisi koko aineiston keskiarvon ja totaalin estimaatteja. Menetelmän käyttö kannattaisi siis, jos viimeksi mainittujen tietojen estimoinnille asetettaisiin keskeisin paino.

Jos imputoinnilla saatu informaatio siirretään vastaajien aineistoon suhteuttamalla (ks. sama luku aiemmin), menetelmä vastaa **suhde-estimointia.** Tilastokeskuksessa tätä menetelmää on käytetty mm. tulonjakotilastossa rekisteripohjaisten tietojen korjaamiseen.

Keskiarvoimputointia parempia ja helppokäyttöisempiä jakauman estimoinnin kannalta ovat satunnaisimputointi B3, hot-deck imputointi B4 ja etäisyysfunktioon perustuva imputointi B6. Jos imputoidut havainnot otetaan aineistoon mukaan kokonaisina tietueina eli kaikki po. havaintoyksikköön liittyvät puuttuvat arvot samanaikaisesti, säilyvät muuttujien väliset yhteydet likimain samanlaisina kuin haastateltujen aineistossa. Havaintoyksiköiden painotus tosin muuttuu, mutta sen voi olettaa hyvässä imputoinnissa vievän tilannetta enemmän oikeaan kuin väärään suuntaan.

Mitään näistä periaatteessa käyttökelpoisista menetelmistä B3, B4 ja B6 ei tässä yhteydessä ole voitu testata. Tämä on johtunut ensinnäkin siitä, ettei projektille ole asetettu tämäntyyppisiä tavoitteita, toiseksi siitä, että menetelmien atk-tekninen soveltaminen Tilastokeskuksen valmisohjelmien avulla on aika raskasta ja kolmanneksi siitä, ettei vastaavanlaisista sovellutuksista laajojen tilastoaineistojen yhteydessä ollut käytettävissä esimerkkejä. Myöhemmin on syytä harkita myös näihin menetelmiin liittyvien kokemusten kartuttamista.

Imputointimenetelmistä on tämän projektin yhteydessä tutkittu regressioimputointia B5. Sen soveltamista tarkastellaan yksityiskohtaisemmin luvussa 7.

6. Vastaustodennäköisyysmalli

Tässä luvussa tarkasteltavaa korjausmenetelmää kutsutaan vastaustodennäköisyysmalliksi, vaikka menetelmä sisältää muitakin kuin tämän nimikkeen piirteitä. Menetelmän lyhyt tiivistelmä on aikaisemmin julkaistu Ekholmin ja Laaksosen (1987) artikkelissa ja keskeiset atk-tekniset toiminnot Laaksosen (1987) seminaariesitelmässä.

Aluksi esitetään mallin lähtökohdat, ml. estimointikaavat totaaleille, keskiarvoille ja estimaattien variansseille. Sen jälkeen esitetään vastaustodennäköisyyksien estimointi vuoden 1985 aineistoon ja lopuksi keskustellaan mallin hyvistä ja huonoista puolista. Vastaavanlainen sovellutus on tehty myös vuoden 1981 kotitaloustiedusteluun. Siitä esitetään muutamia piirteitä.

6.1. Totaalin ja keskiarvon estimointi

Olkoot lausekkeen (1) merkinnät koskien symboleita $w_k(O)$, M, r ja m voimassa. Merkitään edelleen

U = perusjoukko so. Suomen kotitalouksien joukko

S = otokseen poimittujen kotitalouksien joukko ilman ylipeittoa; siinä oli vuonna 1985 n = 11776 havaintoyksikköä.

R = haastateltujen kotitalouksien joukko, jossa oli r = 8200 havaintoyksikköä.

Merkintöjen yksinkertaistamiseksi esitetään kaikki symbolit ja lausekkeet aluksi sovellettuna johonkin ositteeseen, h = 1, 2,..., 24. Ositekohtaisten lausekkeiden ulottaminen yli ositteiden esitetään perustarkastelujen jälkeen. Otetaan käyttöön seuraa-

vat ns. sisältymistodennäköisyydet(1) kullekin ositteelle:

$$\pi_k = P(k \in S) = \frac{m_k \cdot n}{M}$$

$$\pi_{kl} = P(k \in S,\ l \in S) = \frac{m_k \cdot m_l \cdot n \cdot (n-1)}{M \cdot (M-1)}$$

$$\pi_{k/S} = P(k \in R \mid k \in S) = \frac{r}{n}$$

$$\pi_k^* = P(k \in R) = \pi_k \cdot \pi_{k/S} = \frac{m_k \cdot n}{M} \cdot \frac{r}{n}$$

Merkitsemme edelleen

y_k = rahamäärä (tai kulutusmäärä), jonka talous k on käyttänyt tietyn erän (kulutus, palvelu tms.) hankintaan.

Tavoitteenamme on estimoida muuttujan y totaali ja keskiarvo, siis

$$Y = y_1 + y_2 + \ldots + y_N = \Sigma_U\, y_k$$

ja

---

(1)
Todennäköisyydet ovat tosiasiassa oikeiden todennäköisyyksien approksimaatioita (paitsi kolmas). Niissä oletetaan, että samasta taloudesta voi tulla vain yksi jäsen otokseen, vaikka poimintamenetelmän perusteella talous voisi tulla useamman jäsenen kautta otokseen. Todennäköisyydet ovat siten oikeita vain yhden hengen talouksille. Mitä suurempi talous on, sitä pienempi on oikea todennäköisyys. Approksimatiivisia todennäköisyyksiä käytetään kahdesta syystä: (i) ne yksinkertaistavat laskutoimituksia huomattavasti ja (ii) virhe on pieni eli keskimäärin 0,5%:n luokkaa.

$$(1) \quad \bar{Y} = \frac{Y}{N} \quad .$$

Kotitaloustiedustelussa 1981 sovellettiin seuraavaa menetelmää, jota kutsumme **menetelmäksi O** (lue "oo"):

$$\hat{Y}_O = \frac{y_1}{\pi_1^*} + \ldots + \frac{y_r}{\pi_r^*} = \sum_R \frac{y_k}{\pi_k^*} \qquad (3)$$

$$\text{jossa } \pi_k^* = \frac{m_k \cdot r}{M} \quad .$$

Näiden todennäköisyyksien käänteislukuja, joita merkitään

$$w_k^* = w_k(O),$$

kutsutaan menetelmän O korotuskertoimiksi (vrt. lauseke (1)).

Menetelmälle A, joka perustuu vastausalttiusositukseen (josta Little 1986 käyttää sanontaa "response propensity stratification"), vastaava estimaattori lasketaan seuraavasti

$$\hat{Y}_A = \sum_C \sum_R \frac{y_k}{\pi_{kc}^*} \quad , \text{ jossa} \qquad (4)$$

c = solu tai luokka, jossa vastaustodennäköisyyksiä tutkitaan

---

(1)
Huom. Totaali tai keskiarvo voivat olla lasketut myös jollekin "osaperusjoukolle", kuten alueelle tai sosioekonomiselle ryhmälle. Periaate on tällöin sama kuin tarkasteltavassa tapauksessa. Silloin summaukseen valitaan perusjoukosta vain mainittuun osaperusjoukkoon kuuluvat taloudet.

C = solujen c muodostama joukko

$$\pi^*_{kc} = \frac{m_k \cdot n}{M} \cdot p_c(k), \quad \text{jossa}$$

$$p_c(k) = P(k \in R | k \in S,\ k \in C).$$

Tässä tapauksessa merkitsemme todennäköisyyksien π^*_{kc} käänteislukuja $w_k(A)$ tai lyhyemmin $w(A)$ ja kutsumme niitäkin korotuskertoimiksi.(1) Lausekkeen (4) erikoistapauksena asettamalla $y_k = 1$ kaikille k saamme kotitalouksien lukumäärän estimaattorin

$$\hat{N}_A = \sum_C \sum_R \frac{1}{\pi^*_{kc}} \qquad (5)$$

ja edelleen keskiarvoestimaattorin

$$\hat{\bar{Y}}_A = \hat{Y}_A / \hat{N}_A \qquad (6)$$

Solujen muodostamiseen ja vastaustodennäköisyyksien estimointiin palataan jäljempänä. Tässä yhteydessä kuvataan tilannetta oheisilla kuvioilla, jotka selventävät menetelmien eroja:

---

(1)
Tarkemmin sanottuna menetelmällä A tuotetuiksi tai vastaustodennäköisyysmallilla tuotetuiksi korotuskertoimiksi. Kuviossa 1 on käytetty myös nimikettä "korotuskertoimet katoaineistoa hyödyntäen".

Menetelmä O

Ositteet			
1	2	...	24

Menetelmä A

		Ositteet			
		1	2	...	24
SOLUT	1	11	12		
	2	21	22		
	.	.			
	.	.			
	.	.			
	c	c1	c2	...	
	.	.			
	.	.			
	.	.			

Solujen sisällä vastaustodennäköisyyksien oletetaan olevan vakioita. Solujen ja poimintaositteiden muodostamisen välillä ei välttämättä ole mitään yhteyksiä. Useaan osiitteeseen h = 1,..., 24 ulottuvat estimaattorit esimerkiksi lausekkeen (4) tapauksessa määrätään seuraavana summana:

$$\hat{Y} = \sum_h \hat{Y}_h = \sum_h \left(\Sigma_C \Sigma_R \frac{y_k}{\pi^*_{kch}} \right) \qquad (7)$$

Muodostetun estimaattorin perusosa ositetasolla on siis

todennäköisyys π^*_{kc} ja vastaava korotuskerroin

$$w_k(A) = \frac{M}{m_k \cdot n \cdot p_c(k)} .$$

Vertaillaksemme tätä korotuskerrointa $w_k(O)$:hon saamme

$$w_k(A) = \frac{M}{m_k \cdot r} \cdot \frac{r}{n} \cdot \frac{1}{p_c(k)} = w_k(O) \cdot \frac{r}{n} \cdot \frac{1}{p_c(k)} .$$

(8)

Näin ollen korjaamattomasta kertoimesta w(O) päädytään korjattuun kertoimeen w(A), jos se osittaittain kerrotaan ositteen vastaustodennäköisyydellä ja sen solun vastaustodennäköisyyden käänteisluvulla johon talous kuuluu. Tätä yhteyttä voidaan käyttää hyväksi, jos korjaamista aloitettaessa on jo käytössä korjaamaton korotuskerroin tyyppiä w(O).

Lausekkeesta (8) havaitsemme myös, että jos korjaussolut ovat samoja kuin poimintasolut, päädytään tulokseen $w_k(O) = w_k(A)$ kertoimet ovat samoja.

6.2. Totaalin keskivirhe

Tässä kappaleessa johdetaan tilastotuotannossa käytetyn totaalin estimaattorin varianssin estimaattori, jonka neliöjuurta kutsutaan totaalin keskivirheeksi. Sen pohjana oleva teoria on lähtöisin lähinnä Cochranilta (1977, 260-261), Oh'lta ja Scheurenilta (1983) ja Särndalilta ja Svenssonilta (1987).

Merkitään
$$S_k = \begin{cases} 1 \text{ jos } k \in S \\ 0 \text{ jos } k \notin S. \end{cases}$$

Tällöin pätee $S_1 + \ldots + S_N = n.$

Jos tästä yhtälöstä otetaan molemmilta puolilta odotusarvot, saadaan

$$\pi_1 + \ldots + \pi_N = n.$$

Jos edellisen yhtälön molemmat puolet kerrotaan nyt S_k :lla ja molemmilta puolilta otetaan uudelleen odotusarvot, saadaan

$$\pi_{1k} + \ldots + \pi_{Nk} = n\pi_k - \pi_k$$

eli toisessa muodossa

$$\sum_{1 \neq k}{}_{U} \pi_{1k} = n\pi_k - \pi_k \qquad (9)$$

Edelleen pätee kun $1 \neq k$

$$\mathrm{cov}(S_1, S_k) = E(S_1 S_k) - E(S_1)\,E(S_k)$$

$$= \pi_{1k} - \pi_1 \pi_k \,.$$

Tästä ja lausekkeesta (9) saadaan kun $k = 1,\ldots,N$

$$\sum_{1 \neq k} \mathrm{cov}(S_1, S_k) = -\pi_k(1-\pi_k) = -\mathrm{var}(S_k) \qquad (10)$$

Toisaalta olkoon

$$R_k = \begin{cases} 1 \text{ jos } k \in R \\ 0 \text{ jos } k \notin R \end{cases}$$

ja

$$E(R_K) = \pi_k^* = \pi_k \, \pi_{k/S}$$

Silloin vastaavalla tavalla kuin lauseketta (10) johdettaessa saadaan

$$\sum_{1 \neq k} \mathrm{cov}(R_1, R_k) = -\mathrm{var}(R_k) \qquad (11)$$

Jatkotarkasteluja varten me voimme olettaa

joko että

(a) $$\pi_{kl/S} = \frac{r}{n} \cdot \frac{r-1}{n-1}$$

tai että

$$\text{(b)} \qquad \pi_{kl/S} = \frac{r}{n} \cdot \frac{r}{n}$$

Toisin sanoen kotitalouden k vastaaminen tiedusteluun kussakin ositteessa voi (a) riippua tai (b) olla riippumatta kotitalouden l vastaamisesta. Lähestymistavasta riippuen saadaan erilaiset varianssiestimaattorit totaaleille. Oletetaan ensiksi, että missään ositteessa ei ole katoa eli r = n. Silloin $\hat{Y}$:n teoreettinen varianssi on

$$V(\hat{Y}) = E\Big(\sum_{k}{}_{U} S_k \, y_k/\pi_k - \sum_{k}{}_{U} y_k \Big)^2 \qquad (12)$$

$$= \sum_{k}{}_{U} \frac{y_k^2}{\pi_k^2} \operatorname{var}(S_k) + \sum_{k}{}_{U} \sum_{l \neq k}{}_{U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \operatorname{cov}(S_l, S_k).$$

Sijoitetaan yhtälöön (12) yhtälö (10):

$$V(\hat{Y}) = \sum_{k}{}_{U} \sum_{l=k}{}_{U} \left(\frac{y_k^2}{\pi_k^2} - \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \right) \left(\pi_k \, \pi_l - \pi_{kl} \right)$$

Varianssin (12) estimaattoriksi saadaan

$$\hat{V}(\hat{Y}) = \sum_{k}{}_{S} \sum_{l \neq k}{}_{S} \left(\frac{y_k^2}{\pi_k^2} - \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \right) \cdot \frac{\pi_k \, \pi_l - \pi_{kl}}{\pi_{kl}} \qquad (13)$$

Sovellettuna kotitaloustiedustelun otanta-asetelmaan lauseke (13) saadaan muotoon

$$\hat{V}(\hat{Y}) = \frac{M - n}{M} \cdot n \cdot s^2\left(\frac{y_k}{\pi_k} \, ; \, n \right) \qquad (14)$$

jossa $s^2\left(\frac{y_k}{\pi_k} \, ; \, n \right)$ tarkoittaa tavallista

muuttujalle $\frac{y_k}{\pi_k}$ laskettua otosvarianssia

havaintojen määrän ollessa n.

Oletetaan, että edellä mainittu riippuvuusehto (a) on voimassa ja $r \leq n$ on vastanneiden määrä. Silloin totaalin estimaattorin varianssiestimaattori voidaan esittää esimerkiksi seuraavassa muodossa

$$\hat{V}(\hat{Y}) = \frac{M - r}{M} \cdot r \cdot s^2\left(\frac{y_k}{\pi^*_{kc}} \; ; \; r\right) \qquad (15)$$

Lauseketta (15) voidaan käyttää sekä menetelmän O että menetelmän A estimaattorina. Menetelmillä saatuja variansseja vertaamalla saadaan käsitys esimerkiksi siitä, miten totaalin estimaattorin tarkkuus muuttuu katokorjausta käytettäessä verrattuna korjaamattomaan estimaattoriin. Laskelmien tekeminen on helppoa: menetelmän O tapauksessa otosvarianssi s^2 lasketaan muuttujalle $w_k(O) \cdot y_k$ ja menetelmän A tapauksessa muuttujalle $w_k(A) \cdot y_k$.

Jos katsotaan, ettei oletus (a) päde po. otantatilanteessa, vaan oletus (b), saadaan varianssiestimaattori, joka tuottaa yleensä suurempia arvoja kuin lauseke (15). Tässä raportissa ei kuitenkaan esitetä oletukseen (b) perustuvaa estimaattoria, koska sen taustana oleva teoria vaatii eräitä tarkennuksia (tullaan esittämään Ekholmin ja Laaksosen tekeillä olevassa artikkelissa).

Toinen syy tämän "paremman" varianssin esittämättömyyteen on käytännöllinen: tuloksen (15) hyvä puoli on se, että se poikkeaa vain vähän aikaisemmin kotitaloustiedustelussa käytetystä varianssista (ks. Tilastokeskus 1986). Näin tiedustelujen 1981 ja 1985 variansseja voidaan varsin hyvin verrata toisiinsa.

Lausekkeet (15) laskettiin kullekin ositteelle erikseen. Jos halutaan laskea vastaava varianssi yli ositteiden h = 1,...,24, esimerkiksi koko perusjoukolle, muodostetaan niiden summa(1)

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{24} \hat{V}_h \left(\hat{Y}_h \right) . \qquad (16)$$

Lausekkeilla (15) ja (16) saadaan varianssiestimaatit muuttujien y totaaleille. Kiinnostavia tietoja ovat myös vastaavat estimaatit **kotitalouksien lukumäärille.** Ne saadaan samoista lausekkeista asettamalla jokainen $y_k = 1$:

$$\hat{V}(\hat{N}) = \frac{M - r}{M} \cdot r \cdot s^2 \left(w_k(A) \; ; \; r \right) . \qquad (17)$$

ja koko maalle

$$\hat{V}(\hat{N}) = \sum_h \hat{V}_h(\hat{N}_h).$$

Varianssien neliönjuuret eli keskivirheet ovat usein tulkinnassa variansseja käyttökelpoisempia. Useita keskivirheitä vertailtaessa on hyödyllistä laskea myös joko variansseja tai keskivirheitä suhteeessa estimaatteihinsa. Jos suhteellista varianssia merkitään RV:llä, saadaan sen estimaattorin lausekkeeksi totaalille seuraava :

$$\hat{RV}(\hat{Y}) = \frac{\hat{V}(\hat{Y})}{\hat{Y}} .$$

---

(1)
Varianssin symbolia en ole muuttanut, vaikka olen siirtynyt ositetasolle. Lukija näkee asiayhteydestä, mistä varianssista on kysymys.

6.3. Keskiarvon keskivirhe

Varianssit (15)-(17) koskivat totaaleja. Niiden lisäksi on tarvetta tuottaa vähintäänkin keskiarvoestimaattorien varianssiestimaattoreita ja näiden neliönjuuria eli keskiarvojen keskivirheitä. Niiden muodostamiseksi lähdetään liikkeelle lausekkeesta (6)

$$\hat{V}(\bar{\hat{Y}}) = \hat{V}\left(\frac{\hat{Y}}{\hat{N}}\right) .$$

Jos tähän lausekkeeseen sovelletaan tunnettua Taylorin kehitelmää, saadaan varianssille approksimatiivisesti riittävän hyvä tulos seuraavasti:

$$\hat{V}(\bar{\hat{Y}}) = \frac{1}{\hat{N}^2}\hat{V}(\hat{Y}) + \frac{\hat{Y}^2}{\hat{N}^4}\hat{V}(\hat{N}) - 2\frac{\hat{Y}}{\hat{N}^3}\hat{\mathrm{cov}}(\hat{Y},\hat{N}) \quad (18)$$

jossa

$$\hat{\mathrm{cov}}(\hat{Y},\hat{N}) = \frac{M-r}{M} \cdot r \cdot \frac{1}{r-1}\left[\sum y_k w_k^2 - \frac{\sum y_k w_k \sum w_k}{r}\right]$$

Vastaava suhteellinen varianssi taas on seuraava:

$$\hat{RV}(\bar{\hat{Y}}) = \frac{\hat{V}(\hat{Y})}{\hat{Y}^2} = \frac{\hat{N}^2}{\hat{Y}^2}\hat{V}(\hat{Y}) = \frac{\hat{V}(\hat{Y})}{\hat{Y}^2} + \frac{\hat{V}(\hat{N})}{\hat{N}^2} - \frac{2\hat{\mathrm{cov}}(\hat{Y},\hat{N})}{\hat{N}\hat{Y}} \quad (19)$$

Jos variansseja tarvitaan useita ositteita sisältävistä joukoista, enimmillään koko maan 24 ositteesta h, menetellään seuraavasti:

$$\hat{RV}(\bar{\hat{Y}}) = \frac{\sum \hat{V}_h(\hat{Y}_h)}{\left(\sum \hat{Y}_h\right)^2} + \frac{\sum \hat{V}_h(\hat{N}_h)}{\left(\sum \hat{N}_h\right)^2} - \frac{2\sum \hat{\mathrm{cov}}_h(\hat{Y}_h,\hat{N}_h)}{\sum \hat{N}_h \sum \hat{Y}_h} \quad (20)$$

Jos varianssi koskee tiettyä kotitalouksien ryhmää D (esim. maanviljelijätaloudet, joiden päähenkilö on alle 40-vuotias), valitaan summaukseen mukaan vain ne jotka kuuluvat ryhmään D.

6.4. Vastaustodennäköisyyksien estimointi

Vastaustodennäköisyyksien $p_{Ci}(k)$ (ks. kappale 6.1.) estimointia varten oli käytettävissä poimittujen talouksien otosaineisto S. Siten vain tähän aineistoon sisältyviä muuttujia voitiin käyttää selittämään vastausalttiutta. Seuraavia muuttujia tutkittiin:

- Kohdehenkilön sukupuoli
- Kohdehenkilön koulutusaste
- Kotitalouden/asuntokunnan varallisuus
- Kotitalouden/asuntokunnan nettotulot
- Kohdehenkilön verovelvollisuusryhmä
- Kotitalouden/asuntokunnan perherakenne
- Kaupungistumisaste
- Asumisalue
- Kotitalouden/asuntokunnan omaisuustulot.

Näiden lisäksi olisi ollut muitakin mahdollisuuksia, kuten muut tulo- ja varallisuuserät tai niiden summat verotuksesta (ks. liite 6). Myös em. muuttujia olisi voitu soveltaa eri tavoin mm. luokituksia vaihtelemalla.

Paras malli saatiin käyttämällä neljää viimeksi mainittua muuttujaa. Tästä ratkaisusta on selostus jäljempänä. Viiden ensinmainitun muuttujan putoaminen pois ei johtunut yksinomaan siitä, ettei niillä olisi ollut yhteyksiä katoon/vastausalttiuksiin, kuten seuraavasta tarkastelusta havaitaan:

- Kohdehenkilön sukupuoli erotteli vastaamista niin vähän, ettei tätä muuttujaa otettu mukaan malliin. Sukupuoli voisi kuitenkin tulla kysymykseen esimerkiksi siten, että se kytketään ikään tai kotitalouden kokoon. Tällaisia malleja ei kokeiltu tässä yhteydessä.

- Ne, joilla tutkintorekisterissä ei ollut tutkintoa, vastasivat selvästi huonommin kuin tutkintoja suorittaneet. Viimeksi mainituilla vastausalttius kasvoi vain lievästi koulutusasteen noustessa. Kokonaisuudessaan koulutusasteen merkitys ei tullut enää riittävästi

esille sen jälkeen, kun malliin lopullisesti valitut neljä muuttujaa jo olivat mallissa mukana.

- Veronalainen varallisuus on sikäli hankala muuttuja, että se jakautuu varsin epätasaisesti. Erityisen paljon on sellaisia, joilla ei tätä varallisuutta ole ollenkaan. Vaikka kato vähenikin varallisuuden kasvaessa, lukuunottamatta jakauman yläpäätä, sellaisen hyvän ja loogisen luokituksen löytäminen ei onnistunut, joka olisi parantanut kokonaistulosta.

- Nettotuloja tutkittiin sekä kotitaloutta että kulutusyksikköä kohti. Edellisen laskentatavan mukaan havaittiin, että jakauman keskiosat vastasivat paremmin kuin reunat. Jälkimmäisen mukaan kato väheni hitaasti tulojen kasvaessa, jakauman huippua lukuunottamatta. Kumpikaan muuttuja ei tullut mukaan lopulliseen malliin, koska vastaava tieto tuli pääosin mukaan muiden selittäjien kautta. Voidaan keskustella myös siitä, voidaanko ylipäänsä tuloja käyttää katokorjauksen tarpeisiin, koska ne ovat merkittävä osatekijä käytettävissä olevista tuloista, jotka taas ovat yksi tutkimuksen tulosmuuttujista.

- Verovelvollisuustunnus sisälsi tässä tapauksessa vain 6 luokkaa. Käytettävissä ollut luokitus oli hyvin heterogeeninen: palkansaajat kattoivat 67% kaikista, eläkeläiset 16% ja maatilaverotuksen piiriin kuuluvat 8%. Kato vaihteli lähinnä siten, että maataloustaloudet vastasivat paremmin ja eläkeläiset huonommin, muut keskinkertaisesti. Samat asiat tulivat esille myös mukaan valittujen muuttujien kautta, joten verovelvollisuustunnus ei tullut lopulliseen malliin mukaan.

Lopulliseen malliin

hyväksyttiin siis neljä selittäjää.(1)
Niiden kuvaus luokitteluineen esitetään seuraavassa:

(i) Muuttujan, jota olemme kutsuneet nimeltä "Perherakenne" (lyhennet. **FAMILY**), pyrkimyksenä on ollut ottaa samanaikaisesti huomioon sekä kotitalouden koko että sen jäsenten ikä. Kumpaakin näistä kokeiltiin myös erikseen, mutta niiden toimivuus ei ollut yhtä hyvä.

---

(1)
Reunahuomautuksena mainittakoon, että kadon selittäjät tai taustatekijät ovat paljolti samanlaisia kuin äänestämättömyyden (ks. esim. Martikainen 1988). Kaksi poikkeusta voidaan mainita:
- Itäsuomalaiset vastasivat paremmin kuin länsisuomalaiset mutta jälkimmäiset taas äänestivät paremmin.
- Äänestäminen lisääntyi suhteellisen tasaisesti tulotason noustessa, mutta vastaamisalttiuden nousu sen sijaan pysähtyi huipputuloisiin tultaessa.

Esimerkiksi päämiehen ikä erotteli lähinnä vain nuorimpia ja vanhimpia: kato oli keskitasoa suurempi alle 30-vuotiaissa ja yli 75-vuotiaissa.

Perherakenne-muuttuja muistuttaa "Elinvaihe"-muuttujaa, joka on eräs tiedustelussa käytetty sovellettu muuttuja. Muuttujan historia pohjautuu kokemuksiin, joita saimme vuoden 1981 tiedustelun soveltamisessa. Nyt luokittelu oli kuitenkin jonkin verran erilainen. Olennaisin ero oli siinä, että silloin luokitus jakoi yksinäiset kahteen ryhmään: eläkeläisiin ja muihin. Tämä johtui siitä, että eläkeläiset vastasivat vuonna 1981 paremmin kuin muut yksinäiset. Tällä kerralla ei yksinäisiä kuitenkaan ollut tarpeen jakaa iän perusteella, mutta on uskottavaa, että jokin muu lisätekijä kuvaten esimerkiksi syrjäytyneisyyttä, tuottaisi parempia tuloksia.

Seuraavassa esitetään perherakenne-muuttujan luokitus koodeineen ja vastanneiden osuudet eli empiiriset vastaustodennäköisyydet kussakin luokassa:

- 10 = Yksinäiset	.568
- 22 = Kahden hengen taloudet, joissa molemmat 16-34 - vuotiaita	.800
- 29 = Muut kahden hengen taloudet	.697
- 35 = Kolmen hengen taloudet, joissa vähintään yksi 0-6 - vuotias	.741
- 45 = Vähintään neljän hengen taloudet, joissa vähintään yksi 0-6 - vuotias	.766
- 37 = Kolmen hengen taloudet, joissa vähintään yksi 7-15 - vuotias muttei kukaan alle 7-vuotias	.734
- 47 = Vähintään neljän hengen taloudet, joissa vähintään yksi 7-15 - vuotias muttei kukaan alle 7-vuotias	.723
- 49 = Muut eli vähintään kolmen hengen taloudet, joissa kaikki yli 15-vuotiaita	.660

(ii) Kaupungistumisaste on aina haastattelutiedusteluissa todettu merkittäväksi kadon selittäjäksi. Sitä on tavallisesti mitattu jaolla kaupungit vs. muut kunnat. Tässä yhteydessä luokitus oli Uusimaan läänin osalta erilainen: siellä muita kuntia on varsin vähän eikä jako muutenkaan kovin hyvin kerro läänin kaupungistumisesta. Sen vuoksi muodostettiin muuttuja (lyhennet. **URBANICITY**), joka muiden läänien osalta on

sama kuin kaupungit vs. muut kunnat, mutta Uusimaalla vastaava jako on Pääkaupunkiseutu vs. muut kunnat. Näin saatua jakoa kutsutaan nimellä "Keskusalueet vs. reuna-alueet". Niiden empiiriset vastaustodennäköisyydet olivat:

.647 ja .757.

(iii) Kolmanneksi muodostettiin alueluokitus (lyhennet. **AREA**), joka pohjautui läänijakoon seuraavasti ml. empiiriset vastaustodennäköisyydet:

- ALUE 1:	
Uusimaan lääni	.553
- ALUE 2:	
Turun ja Porin, Hämeen ja Kymen läänit sekä Ahvenanmaa	.685
- ALUE 3:	
Mikkelin, Pohjois-Karjalan, Kuopion ja Keski-Suomen läänit	.798
- ALUE 4:	
Vaasan, Oulun ja Lapin läänit	.749.

(iv) Neljänneksi selittäjäksi soveltui verotuksen omaisuustulo (lyhennet. **PROPERTY**), joka lopullisessa ratkaisussaan sai kaksi arvoa ml. vastaustodennäköisyydet:

- Omaisuustuloa ei ollut	.675
- Omaisuustuloa oli	.742.

Vastaustodennäköisyydet estimoitiin seuraavalla

additiivisella logistisella mallilla

$$\log \frac{p_c(k)}{1 - p_c(k)} = \text{VAKIO} + \text{FAMILY} + \text{URBANICITY} + \text{AREA} + \text{PROPERTY} \qquad (21)$$

Mallilla on 1+7 + 1 + 3 + 1 = 13 parametriä,
8 x 2 x 4 x 2 = 128 solua ja siis vapausasteiden määrä
df = 128-13 = 115.

Uskottavuusosamäärän testisuureen arvo oli 126. Tämä merkitsee sitä, että malli sopii hyvin aineistoon ja ns. p-arvo oli .22. Selittäjistä paras oli muuttuja AREA ja huonoin PROPERTY. Muut muuttujat FAMILY ja URBANICITY olivat likimain yhtä merkitseviä.

Mallin avulla määrättiin kullekin 128 solulle **estimoidut vastaustodennäköisyydet**. Näitä todennäköisyyksiä käytettiin korotuskerrointa w(A) muodostettaessa havaittujen todennäköisyyksien sijasta, siis $p_c(k)$-lukuina. Kirjallisuus ei anna selkeätä vastausta siihen, kumpia todennäköisyyksiä tulisi käyttää. Voidaan ajatella tästä sovellutuksesta poiketen, että havaittuja todennäköisyyksiä tulisi käyttää, jos ne ovat saatavissa. Silti malli, esimerkiksi logistinen malli, olisi erinomainen apuväline etsittäessä hyvää solukkoratkaisua: parhaassa mallissa käytetty solukko soveltuisi perustellusti korotuskertoimen muodostamiseen.

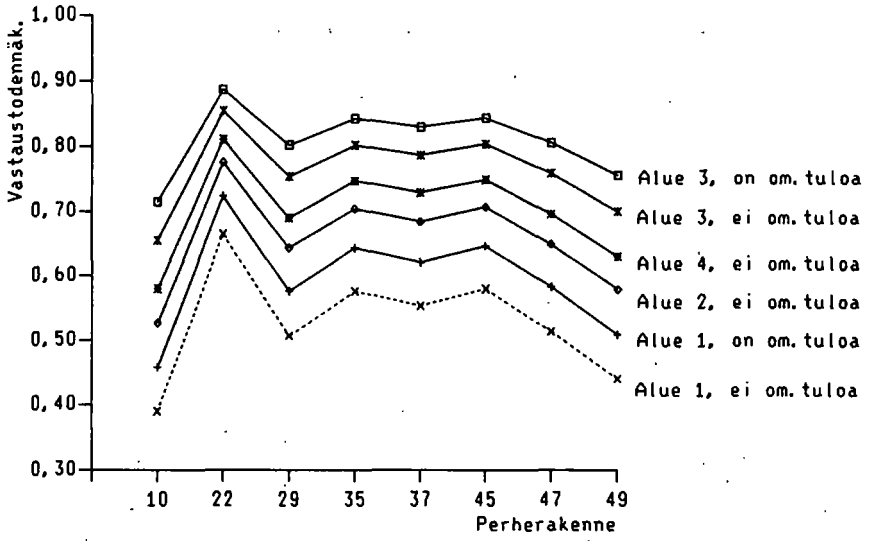
Kuvioissa 2 ja 3 esitetään muutamia tuloksia estimoiduista vastaustodennäköisyyksistä. Havaitaan, että ne vaihtelevat varsin paljon, ääriarvojen ollessa .39 ja .92. Alin tulos saatiin sellaisten Alueen 1 keskusalueiden so. Pääkaupunkiseudun yksinäisten talouksien ryhmässä, joilla ei ollut omaisuustuloja. Vastaavasti korkein tulos saatiin sellaisten Alueen 3 so. Itä-Suomen läänien kaupunkien kahden hengen alle 35-vuotiaiden talouksissa, joilla ei ollut lapsia mutta joilla oli omaisuustuloja. Nämä nuorten parien taloudet vastasivat myös muilla alueilla hyvin (vastaavat likimain muotikäsitettä "dinkki", DINK = double income, no kids).

6.5. Keskustelua

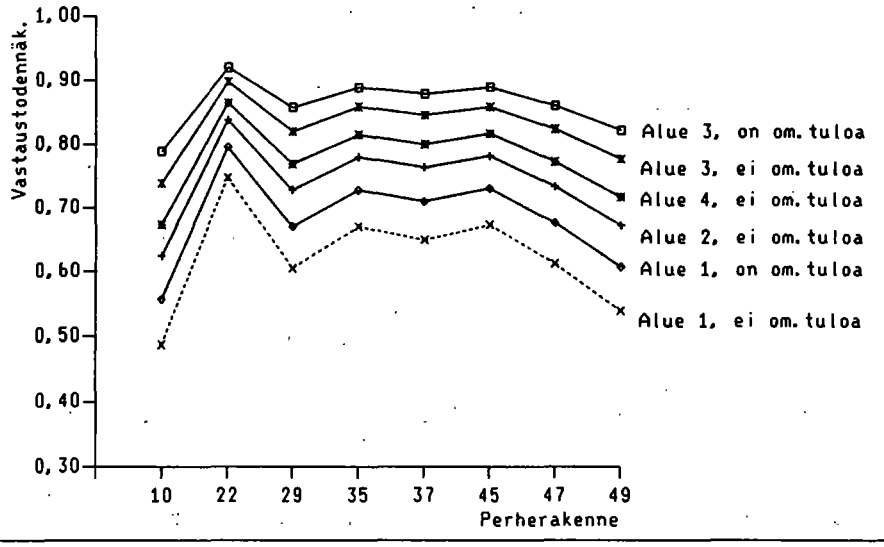
Kappaleessa 6.4. todettiin, että havaitut vastaustodennäköisyydetkin voivat tulla kyseeseen uusia painoja muodostettaessa. Tässä kappaleessa keskustellaan muista malliin ja sen sovellutukseen liittyvistä epävarmuustekijöistä ja vaihtoehtoisista lähestymistavoista.

Ensiksikin on syytä todeta se, että koska empiiriset ratkaisut perustuvat olennaisella tavalla katoosaa koskevan tiedon hyödyntämiseen, tulosten laatutaso riippuu olennaisesti tämän tiedon laadusta. Tässä tapauksessa kato-osan tiedoissa on kahdenlaisia heikkouksia:

Kuvio 2. Menetelmällä A estimoituja vastaustodennäköisyyksiä keskusalueilla (käsitteiden määrittely tekstissä).



Kuvio 3. Menetelmällä A estimoituja vastaustodennäköisyyksiä reuna-alueilla (käsitteiden määrittely tekstissä).



(i) Ylipeittoa ei ole puhdistettu täydellisesti.

(ii) Kotitalouksien koostumusta 11776-9220=2556 taloudessa ei ole määritelty aivan yhtä hyvin kuin alkuhaastatteluun osallistuneiden ryhmässä.

Näiden heikkouksien korjaamiseen tulisi tulevaisuudessa pyrkiä entistä tarmokkaammin (vrt. liite 4, jossa on esitetty joitakin ajatuksia tässä tarkoituksessa). Olen arvioinut, että po. heikkoudet liioittelevat yksinäisten ihmisten katoa ja sitä kautta korjaavat liikaa vastaavia kotitalouksien lukumääräestimaatteja. Kulutukseen tai muihin tulosmuuttujiin virhe vaikuttaa monimutkaisemmin.

Toiseksi on syytä kiinnittää huomiota solujen muodostamiseen. Ilman muuta on selvää, että jotkut muut solukkoratkaisut voivat olla parempia kuin tässä käytetty. Tähän voin todeta vain, että kokeiluja on tehty useita, joista edellä esitetty osoittautui siinä tilanteessa parhaaksi. Nämä kokeilut osoittivat toisaalta myös, että kovin suuria muutoksia keskeisiin estimaatteihin ei tullut, jos solukkoratkaisussa käytetyn mallin selittäjät oli luokiteltu osittain toisella tavalla kuin mallissa (21) tai jos neljäs selittäjä oli joku muu kuin omaisuustulo.

Esitetty malli ja sen luokitukset on pyritty tekemään mahdollisimman todellisiksi. Voitaisiinhan solukot muodostaa myös siten, että nostettaisiin kaikin keinoin mallin selitysvoimaa. Tässä yhteydessä ei näin ole tehty, koska olemme halunneet pyrkiä sekä vastaamattomuuden selittämiseen että estimaattien korjaamiseen.

Solujen ominaisuudet ovat (ks. luku 5 ja liite 2) tärkeitä estimaattien hyvyyden kannalta. Kaikkia ominaisuuksia ei voida täsmällisesti tutkia, koska varsinaisista tulosmuuttujista ei ole tietoja katoosasta. Pohtikaamme kuitenkin kahta Littlen ja Rubinin (1987) esittämää näkökohtaa, jotka tulkitsisin seuraavasti: solujen sisällä ei (i) vastaamisalttius eikä (ii) tulosmuuttuja vaihtele minkään tekijän suhteen.

(i) Vastausalttius siis pitäisi solujen sisällä olla mahdollisimman riippumatonta, jopa vakio. Tässä tapauksessa mallittamisella tähdättiin juuri siihen, että mahdolliset systemaattiset tekijät saataisiin solujen sisältä eriteltyä pois. Silti on varmaa, että puutteita jää jäljelle. Erityisesti kannattaisi tutkia sellaisia soluja, joissa havaintoja on kohtuullisesti mutta vastausalttius pieni (esim. yksinäiset): on mahdollista että niiden sisältä löytyy vielä systemaattisia tekijöitä jotka voitaisiin poistaa.

(ii) Vastaamisen riippuvuutta tulosmuuttujasta solujen sisällä ei voida täsmällisesti tutkia, koska tulosmuuttujasta ei ole tietoja kato-osasta. Joitakin viitteitä saadaan vertaamalla keskeisten x-muuttujien käyttäytymistä toisaalta vastanneiden ja toisaalta kaikkien poimittujen kesken solujen sisällä. Kulutuksen kannalta keskeisimpiä x-muuttujia ovat tulot. Sen vuoksi tarkasteltiin verotuksen nettotuloja em. kahdessa ryhmässä.

Tulos osoittautui **hyväksi**: solujen keskiarvot kaikkien poimittujen ryhmässä olivat toisinaan pienempiä, toisinaan suurempia kuin vastanneiden ryhmässä. Erot olivat tyypillisimmillään 1%:n suuruusluokkaa ja vain harvoin yli 3%:n. Sen sijaan ilman solujakoa laskettu keskiarvo oli vastaajien aineistossa 4,5% suurempi kuin poimittujen aineistossa.

Kolmanneksi on aiheellista keskustella siitä, voitaisiinko esitetty painotuskorjaus korvata jollakin toisella. Menetelmän yksi idea on osituksen muodostaminen, jollaista käytetään myös esimerkiksi jälkiositukseen perustuvassa korjausmallissa A1 (ks. luku 4). Voidaankin kysyä, mitä eroa näillä kahdella menetelmällä on?

Jälkiositus pohjautuu korotuskertoimeen, joka on lausekkeessa (1) esitettyä tyyppiä. Ositus ulotetaan kuitenkin alkuperäisten poimintaositteiden sisälle. Merkitään näin saatuja korotuskertoimia

$$w_{hj}\text{:lla,}$$

jossa h viittaa poimintaositteeseen ja j jälkiositteeseen. Jos jälkiositus olisi tehty täsmälleen samalla tavalla kuin edellä esitetyssä vastaustodennäköisyysmallissa, saataisiin siis

$$w_{hjk} = \frac{M_{hj}}{r_{hj} \cdot m_{hjk}}$$

Tästä lausekkeesta tunnemme aineiston perusteella m_{hjk}:t ja r_{hj}:t, mutta luvut M_{hj} pitäisi tuottaa perusjoukosta. Esitetyn 128 solun tietoja ei kuitenkaan ole kovin luotettavasti saatavissa. Jos ne estimoitaisiin kato-osan tietojen perusteella ja käyttäen hyväksi vastaustodennäköisyysmallilla saatuja estimaatteja, päädyttäisiin samaan kertoimeen kuin menetelmällä A,

jolloin jälkiositus ei siis tuottaisi mitään uutta, aiheuttaisi vain lisätyötä.

Silti en kokonaan luopuisi jälkiosituksesta: ehkä sitä voitaisiin käyttää ennen vastaustodennäköisyysmallia siltä osin kuin perusjoukosta on M-tiedot saatavissa (ongelmana on mm. laitosväestö). Näin otos ja perusjoukko "kiinnitettäisiin useammista kohdista" toisiinsa todellisuutta vastaavasti. Sehän on lukujen M perustehtävä. Asia edellyttäisi lisäselvityksiä.

7. Regressiomalli

Regressiomallin soveltaminen oli tämän projektin alkuperäisiä lähtökohtia. Tässä luvussa ei esitetä kaikkia niitä ajatuksia ja kokeiluja, joita kuluneiden vuosien aikana aihepiiristä on syntynyt (ks. osaa niistä Logit Ky 1984 ja Laaksonen 1986). Pääpaino on menetelmän perusteissa ja sellaisissa sovellutuksissa, jotka ovat antaneet kohtuullisia tuloksia. Niistä on aikaisemmin esitetty tiivistelmä Laaksosen ja Ekholmin (1987) artikkelissa. Estimointitulokset vuosien 1981 ja 1985 aineistoilla esitetään luvussa 8.

7.1. Regressiomallin perusteet katovirheen korjauksessa

Malli perustuu sellaisten muuttujien x hyväksi käyttöön, joista on saatavissa tietoa myös kato-osasta tai koko perusjoukosta. Ensin tarkastelemme edellisen mukaista tilannetta.

Lähtökohtana on seuraava malli

$$y_k = x_k\beta + e_k \qquad k = 1,\ldots,N \qquad (22)$$

Yhtälössä (22) oletamme

(i) y_k, x_k ja e_k ovat kiinteitä suureita, eivät satunnaismuuttujia

(ii) y_k ja e_k ovat tunnettuja vain vastaajille, siis joukolle R

(iii) x_k on tunnettu koko otokselle S

(iv) β on tuntematon, se estimoidaan joukosta R

(v) x_k ja β ovat 1 x (q+1)- ja (q+1) x 1-vektoreita, jossa q+1 = estimoitavien parametrien lukumäärä ja q = selittäjien lukumäärä.

Hypoteettisesta mallista (22) seuraa

$$Y = \Sigma_U y_k = (\Sigma_U x_k)\beta + \Sigma_U e_k \qquad (23)$$

Yhtälön (23) kokonaiskulutuksen Y laskemiseksi estimoidaan erikseen sen osat seuraavasti:

$$(\hat{\Sigma}_U x_k) = \hat{X}_S = \Sigma_S \frac{x_k}{\pi_k} \qquad (24a)$$

$$(\hat{\Sigma}_U e_k) = \Sigma_C \Sigma_R \left(\frac{y_k - \hat{y}_k}{\pi^*_{kc}}\right) \qquad (24b)$$

Vielä on määrättävä parametrin β estimaatti b. Se lasketaan painotetulla lineaarisella pienimmän neliösumman menetelmällä joukosta R olettamalla, että

$$\mathrm{var}(e_k) = \sigma^2 t_k^2 / \pi^*_k ,$$

jossa t_k:t ovat tunnettuja vakioita, kuten esimerkiksi

$$t_k^2 = 1 \quad \text{tai}$$

$$t_k^2 = x_k .$$

Näin meillä on siis regressioestimaattori kokonaiskulutukselle, jota on saatu **menetelmällä B:**

$$\hat{Y}_B = \hat{Y}_A + (\hat{X}_S - \hat{X}_R)b \qquad (25)$$

jossa $\hat{X}_S = \Sigma_S \frac{x_k}{\pi_k}$ ja

$$\hat{X}_R = \Sigma_C \Sigma_{R \cap C} \frac{x_k}{\pi^*_{kc}} .$$

Vastaava estimaattori kulutuksen keskiarvolle saadaan seuraavasti

$$\hat{\bar{Y}}_B = \frac{\hat{Y}_B}{\hat{N}_A} . \qquad (26)$$

jolloin siis kotitalouksien lukumäärä on estimoitu menetelmällä A.

Yhtälön (24) esitystapa kertoo, että menetelmällä A saatua estimaattoria korjataan taustamuuttujien x estimaattien erotuksilla, jonka erotuksen ensimmäinen estimaatti perustuu koko otosaineistosta S laskettuun tietoon ja jälkimmäinen estimaatti vain haastateltua osaa R koskevaan tietoon. Nämä erotukset kerrotaan vastaavilla aineistosta R estimoiduilla regressiokertoimilla b.

Lauseke (25) voidaan esittää myös seuraavassa muodossa

$$\hat{\bar{Y}}_B = \hat{X}_S \, b. \qquad (27)$$

Yhtälön (26) mukaan voidaan regressioestimaatti tulkita mallilla lasketuksi ennusteeksi: aineistosta R estimoituun yhtälöön sijoitetaan x-muuttujien arvot, jotka on estimoitu aineistosta S. Menetelmillä A ja B saadaan siis samat tulokset silloin, kun sijoitettavat x-arvot ovat samoja sekä R:stä että S:stä laskettuina. Tyypillinen tilanne on keskiarvon estimointi, jolloin menetelmät A ja B antavat samat estimaatit, jos x-muuttujien keskiarvot ovat samoja sekä R:ssä että S:ssä.

Yhtälön (25) mukainen estimaattori $\hat{Y}$ on harhaton, jos

$$b = \beta .$$

Tämä merkitsee siis, että mikäli otosaineistosta estimoitu malli on voimassa koko perusjoukossa, estimaattori on harhaton. Ehto tuskin on täysin voimassa missään aineistossa, koska vino kato todennäköisesti vie myös regressiosuoran "väärään paikkaan".

Oheinen Ten Caten (1986) kuvio havainnollistaa tilannetta yhden selittäjän tapauksessa. Siinä todellinen ja estimoitu suora poikkeavat aika selvästi toisistaan erityisesti aineiston reunalla. Keskiarvon kohdalla (jota ei kuvioon tosin ole merkitty) poikkeama ei sen sijaan ole huomattava. Siten harhaisuus ei näytä kovin suurelta ongelmalta käytettäessä regressiomallia keskiarvojen, totaalien yms. tunnuslukujen estimointiin.

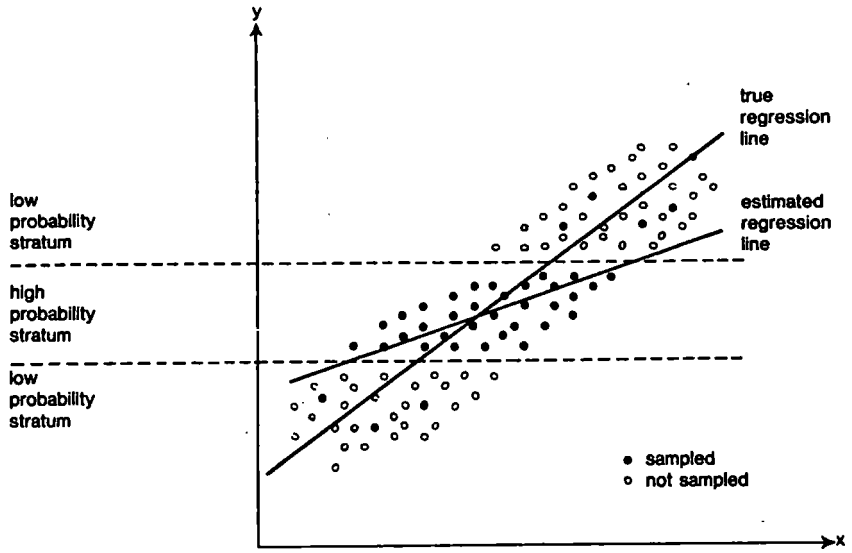


Figure 1. The Effect of Endogenous Stratification on the Estimated Regression Line

Jos estimaattori $\hat{Y}$ on harhaton, voidaan sille määrätä myös harhaton varianssin estimaattori:(1)

---

(1) Lausekkeen (28) estimaattori on muodostettu toisaalta menetelmän A varianssin estimaattorista (ks. lauseke (15)) ja toisaalta Särndalin ja Svenssonin (1987) esittämästä regressioestimaattorin varianssin estimaattorista.

$$\hat{V}(\hat{Y}_B) = \frac{M - r}{M} \cdot r \cdot s^2(e_k w_k; r) \qquad (28)$$

Lauseke (28) osoittaa, että regressioestimaattorin varianssi yleensä pienenee alkuperäisen estimaattorin $\hat{Y}$ varianssista. Muutos on sitä suurempi, mitä parempi regressiomalli aineistosta R on saatu. Regressioestimaattori on, mikäli harha on pieni, siis yleensä tehokkaampi kuin painokorjattu estimaattori.

Regressiomallilla tuotetun **keskiarvon** $\hat{\bar{Y}}$ varianssi riippuu lausekkeen (19) tapaan $\hat{Y}$:n ja kotitalouksien lukumäärän $\hat{N}$ variansseista sekä näiden välisestä kovarianssista. Edellinen varianssi saadaan lausekkeella (28) ja jälkimmäinen lausekkeella (17). Kovarianssitekijän muodostaminen on hankalampaa. En olekaan saanut muodostettua täysin perusteltua kovarianssiestimaattoria, vaan olen käyttänyt seuraavaa ehtoa tulosten laskennassa: $\hat{Y}$:n ja $\hat{N}$:n välinen korrelaatio ei muutu siirryttäessä vastaustodennäköisyysmallilla saadusta estimaattorista regressiomallilla saatuun. Ainoastaan $\hat{Y}$:n varianssi muuttuu.

Regressiomallissa (22) muuttujat x olivat saatavissa poimittujen joukosta S. On myös mahdollista, että **tunnemme x-muuttujat koko perusjoukosta.** Esimerkiksi jos tunnemme selittävien muuttujien summat koko väestölle sekä väestömäärän, voimme määrätä niiden keskiarvot asukasta kohti. Sen sijaan emme voi määrätä perusjoukon kotitalouskohtaisia keskiarvoja, koska kotitalouksien perusjoukkoa ei ole käytettävissä. Tämän vuoksi malli täytyy rakentaa siten, että sekä

selitettävä että selittäjät on muodostettu suhteessa talouden jäsenten lukumäärään. Asetetaan siten

$$Z = \sum_U \frac{y_k}{g_k} \qquad \text{ja} \qquad \bar{Z} = \frac{1}{N} \sum_U \frac{y_k}{g_k}$$

jossa g_k = kotitalouden k jäsenten lukumäärä.

Haluamme nyt estimoida Z:n. Jos asetamme

$$\hat{N}\hat{\bar{Z}} = \hat{Z},$$

voimme muodostaa estimaattorin $\hat{Z}$ vastaavalla tavalla kuin Y:lle menetelmällä A. Regressioestimointi perustuu nyt malliin (vrt. lauseke (22))

$$\frac{y_k}{g_k} = \frac{x_k}{g_k} \cdot \beta + e_k \qquad k = 1, \ldots, N \qquad (29)$$

Tästä saadaan vastaavalla tavalla kuin edellä

$$\hat{Z}_B = \hat{Z}_A + \left(X_S^{(z)} - X_R^{(z)} \right) \cdot b$$

jossa $X_S^{(z)} = \sum_S \frac{x_k/g_k}{\pi_k}$.

Jos $\sum_U \frac{x_k}{g_k}$ on tunnettu, ja merkitsemme sitä $X^{(z)}$:lla, saamme **menetelmän C estimaattorin:**

$$\hat{Z}_C = \hat{Z}_A + (X^{(z)} - X_R^{(z)}) \cdot b \; . \qquad (30)$$

Tälle estimaattorille saadaan varianssin estimaattori vastaavalla tavalla kuin menetelmälle B. Myös harhaongelma on samanlainen kuin edellä.

Menetelmää C voidaan soveltaa myös siten, että yhtälön

perusjoukkoa koskeva tieto $\bar{x}^{(z)}$ estimoidaan otosaineistosta S. Silloin menetelmä C poikkeaa menetelmästä B vain mallin spesifioinnin osalta. On myös mahdollista, että osalle selittäjistä tämä tieto otetaan perusjoukosta, osalle aineistosta S.

Menetelmä C antaa ilmeisestikin parempia tuloksia kuin menetelmä B, jos x-muuttujia koskeva tieto perusjoukosta on luotettavaa. Näin ei välttämättä kotitaloustiedustelussa ole, koska laitosväestöä koskevia tietoja ei voida hyvin puhdistaa aineistosta pois.

7.2. Sovellutuksia

Regressiomallin soveltaminen kotitaloustiedusteluun kappaleessa 7.1. esitetyillä menetelmillä B ja C on periaatteessa helppoa, ainakin silloin kun tulosmuuttuja y on välimatka- tai suhdeasteikon muuttuja. Käytännössä tilanne on hankalampi, koska tällaisia muuttujia on runsas 1000. Hankaluuksia tulee lisää, jos eri osajoukoille, esimerkiksi alueellisille, muodostetaan omat mallinsa. Malleja voisi siten kokonaisuudessaan tulla hyvin paljon, jopa kymmeniä tuhansia. Teknisesti tämä ei olisi ongelma ainakaan silloin, jos useimpien mallien selittäjät olisivat samoja. Tällainen kokeilu tehtiin vuoden 1981 aineistolla. Tulokset eivät olleet huonoja, mutta eivät keskimäärin juuri parempia kuin ilman mallia lasketut estimaatit.

Tilastollisten mallien rakentamisen tulisi olla huolellisempaa kuin mihin esitetynlaisella massatuotannolla päästään. Tästä syystä on lähdetty myöhemmin siitä, että regressiomalleja on tarkoituksenmukaista soveltaa vakavassa tarkoituksessa vain suppeahkoon muuttujajoukkoon, mahdollisesti alueosituksia käyttäen.

Tästä huolimatta sovelluksia voidaan tehdä monin tavoin. Jo yksistään aggregointitason valinta y-muuttujille on hankalaa. Alkuperäisen tiedonkeruutason käyttäminen on harvoin järkevää. Esimerkiksi yksittäisiin tilinpitoeriin malli ei "pure" hyvin siksi, että suuri osa muuttujien arvoista on nollia, ehkä maitoa, leipää tms. tms. kulutuserää lukuunottamatta. Tämä johtuu siitä, että kahden viikon tilinpidon aikana ei

ehditä hankkia kovinkaan monia eri hyödykkeitä. Myös pidempiaikaiseen tiedonkeruuseen perustuvien muuttujien, esimerkiksi kotitalouskaluston, mallittamiseen liittyy samoja ongelmia.

Ensisijaiseksi sovellutusten aggregointitasoksi on otettu ns. pääryhmätaso, joita on 8. Tällöin nollaarvoja on vähän ja on muutenkin paremmat edellytykset regressiomallien rakentamiseen. Tässä kappaleessa esitetään muutamia tuloksia pääryhmätasolta. Niiden tarkoitus on kuvata esimerkinomaisesti regressiomallien soveltuvuutta katokorjaukseen.

Selittäjien valinta on mallituksen ensimmäisiä tehtäviä. Liitteessä 6 luetellaan kaikki ne muuttujat, jotka rekistereistä valittiin selittäjäehdokkaiksi. Taulukossa 1 esitetään tuloksia parhaista pääryhmätason malleista vuodelta 1985. Vuoden 1981 tuloksista on raportoitu aikaisemmin (Laaksonen 1986).

Taulukon estimoinnit perustuvat koko maan aineistoon. Selitysasteet ei ole kovin korkeita, mikä on varsin luonnollista tämäntyyppisissä poikkileikkausaineistoissa. Kulutusta selittävät siis monet muutkin tekijät kuin mitä rekistereistä oli saatavissa. Huonoiten malli sopii kotitalouskalustokulujen selittämiseen, jonka selityksestä jo edellä käytiin keskustelua: tätä hyödykeryhmää varten tarvittaisiin vuottakin pidempään tiedonkeruuseen perustuvia aineistoja.

Kaikissa muissa tapauksissa tulokset ovat kohtuullisia ja uskottaviakin. Nettotulot (NETTU) näyttelevät varsin suurta roolia kaikissa muissa paitsi sairauden- ja terveydenhoidon ryhmässä, jossa taas saadaan varsin odotettu ainoa merkitsevä selittäjä so. sairausvähennykset verotuksessa. Viimeksi mainittu malli ei selitysasteeltaan ollut kovin korkea, mutta vuoden 1981 aineistoille se oli lähes 50%.

Nettotulojen lisäksi havaitaan monia muitakin luonnollisia selityksiä, kuten varallisuuden "vaikutus" asumis- ym. kustannuksiin tai kulkuvälineiden omistuksen "vaikutus" liikennekuluihin. Myös väestömuuttujien käyttäytyminen on loogista. Kaiken kaikkiaan tulokset antavat viitteitä siitä, että regressiokorjaus voi olla käyttökelpoinen katoaineiston korjaamiseen.

Liitteessä 7 on lisäksi tarkasteltu käytettävissä olevien tulojen selittämistä veroaineiston tiedoilla. Nämä tulokset ovat kulutusmuuttujilla saatuja parempia, selitysasteen ollessa yli 80%. Syynä hyvään

TAULUKKO 1. Tuloksia regressioestimoinnin pohjana olevista kulutuksen selitysmalleista vuodelta 1985. Selittäjien selostus liitteessä 6. Kaikki muuttujat havaintoaineistossa tyyppiä per kotitalous.

SELITETTÄVÄ TULOSMUUTTUJA	MERKITSEVIMMÄT SELITTÄJÄT JÄRJESTYKSESSÄ (edessä kertoimen merkki)	SELITYS-ASTE
Pääryhmä 1: Elintarvikkeet	+ IMUKU + NETTU + TYOIKA + MUKU - VELAT	43 %
Pääryhmä 2: Vaatteet ja jalkineet	+ NETTU + IMUKU + MUKU	19 %
Pääryhmä 3: Asunto, lämpö ja valo	+ NETTU + VARAT + IMUKU - VELAT	49 %
Pääryhmä 4: Kotitalouskalusto	+ NETTU + MUKU	8 %
Pääryhmä 5: Sairauden- ja terveydenhoito	+ SAIR	14 %
Pääryhmä 6: Liikenne ja tietoliikenne	+ KULKU + NETTU + TYOIKA	36 %
Pääryhmä 7: Virkistys- ja harrastustoiminta	+ NETTU + MUKU + IMUKU	20 %

tulokseen on se, että verotiedot muodostavat käytettävissä olevista tuloista varsin suuren osuuden. On siten aivan luontevaa käyttää tämä tieto hyödyksi esimerkiksi regressiomallin avulla.

Tässä yhteydessä ei esitetä sovellutuksia menetelmällä C, koska sitä sovellettiin ainoastaan vuoden 1981 aineistoon. Tulokset olivat selitysaste-mielessä yleensä huonompia, koska silloin kotitalouden koon vaikutus mallista on poistettu. Estimaattimielessä se antaa varsin samanlaisia tuloksia kuin menetelmä B, jos x-muuttujista saatu informaatio on sama.

7.3. Keskustelua

Hyvän regressiomallin tulisi täyttää ns. **homoskedastisuusehto** eli residuaalien tulisi pysyä samanlaisina sovitteen arvojen kasvaessa. Tässä tapauksessa kuitenkin residuaalien vaihtelu oli kasvavaa eli mallit olivat jossain määrin heteroskedastisia. Häiriön poistamiseksi kokeiltiin havaintoarvojen logaritmointia ja painotusta siten, että alkuperäiset painot jaettiin tärkeimmällä selittäjällä eli nettotuloilla. Tulokset eivät kummallakaan keinolla olennaisesti parantuneet. Koska nämä spesifioinnit hankaloittivat laskelmia, ei niitä otettu estimointien pohjaksi.

Tässä raportissa esitetyt sovellutukset perustuvat painotettuun lineaariseen regressiomalliin. Olisi mahdollista ottaa selittäjiksi myös kvalitatiivisia muuttujia, kuten koulutus tai alue. Estimoinnissa käytettäisiin tällöin yleistä lineaarista mallia. Tältä pohjalta ei tuloksia ole toistaiseksi tuotettu.

Menetelmien B ja C soveltamista edelsi menetelmä A, joka käytti hyväkseen jo joitakin luokiteltuja x-muuttujia. Regressiomalleissa x-muuttujia on käytössä enemmän kuin menetelmässä A, mutta mukana voi olla myös samoja x-muuttujia joskin eri tavoin käytettynä. Tämä "kaksoiskorjaus"-strategia voi herättää kysymyksen, voidaanko näin tehdä. Särndal ja Svensson (1987) käsittelevät samantyyppisiä tilanteita, tosin teoreettisemmin. Käsitykseni mukaan he eivät sulje tätä strategiaa pois. Myös esim. Littlen ja Rubinin (1987) tai Kaltonin ja Kasprzykin (1986) pohdiskelut olen tulkinnut niin, että useampia eri menetelmiin perustuvia korjauksia voidaan tehdä peräkkäin. Myös estimointitulokset, joita seuraavaksi tarkastellaan, näyttävät myönteisiltä tämän strategian kannalta.

8. Estimaatteja vuosille 1981 JA 1985

Tässä luvussa esitetään keskeisiä estimointituloksia menetelmillä O, A ja B. Kappaleessa 8.1. tarkastellaan kotitalouksien lukumääriä, kulutuseräkeskiarvoja ja -totaaleja sekä jakaumia todellisten aineistojen valossa, pääpainon ollessa vuodessa 1985. Kappaleessa 8.2. esitetään tuotetulla keinotekoisella so. simuloidulla aineistolla tuloksia vuodelta 1981. Ne koskevat vuotta 1981 siksi, ettei vastaavaa koesarjaa katsottu aiheelliseksi tehdä vuodelle 1985.

8.1. Todellisten aineistojen tuloksia

Kotitalouksien lukumäärien estimointi on tiedustelun keskeisimpiä tehtäviä, koska tällä estimaatilla on vaikutuksia moniin muihin, kuten kotitalouksien keskikokoon (joka on sama kuin koko kotitalousväestön lukumäärä jaettuna kotitalouksien lukumäärällä), kokonaiskulutukseen, keskimääräiseen kulutukseen kotitaloutta, kotitalouden yhtä jäsentä tai kulutusyksikköä kohti. Taulukossa 2 esitetään kotitalouksien lukumäärien, kotitalousväestön ja kotitalouksien keskikoon estimaatteja menetelmillä O ja A.

Menetelmillä B ja C voitaisiin laskea myös kotitalouksien lukumäärien estimaatit. Jos kuitenkin, kuten olisi toivottavaa, saman informaation pitäisi näkyä myös kotitalouksien jakaumassa kotitalouden koon mukaan, nämä menetelmät eivät toimi asiallisesti, koska ne voisivat antaa katokotitalouden kooksi desimaaliluvun. On mahdollista, että jollakin muulla mallituksella, muttei perusregressiomallilla, tätä ongelmaa ei syntyisi.

Tulokset osoittavat, että menetelmällä A koko maan kotitalouksien lukumäärä lisääntyi(1) vuonna 1981 2,5%:lla ja vuonna 1985 4,1%:lla. Uusimaan läänissä

---

(1) Vertailuestimaatit on tuotettu menetelmällä O.

TAULUKKO 2. Tuloksia kotitalouksien lukumääristä vuosilta 1981 ja 1985 menetelmillä O ja A.

ESTIMOITU MUUTTUJA	VUOSI	MENETELMÄT	
		O	A
Kotitalouksien lukumäärä (1000) koko maassa	1981 1985	1873 1960	1920 2042
Kotitalouden keskikoko koko maassa	1981 1985	2,54 2,46	2,46 2,36
Kotitalousväestön määrä (1000) koko maassa	1981 1985	4763 4828	4720 4830
Kotitalouksien lukumäärä (1000) Uusimaan läänissä	1981 1985	477 504	494 540
Kotitalouksien lukumäärä (1000) Lapin läänissä	1981 1985	67 73	63 78

lisäykset olivat 3,6% ja 7,1% ja Lapin läänissä -6,0% ja 6,8%. Näitä voidaan pitää oleellisina muutoksina. Kotitalouksien keskikoot muuttuivat vastaavasti päinvastaiseen suuntaan. Tämä näkyy, kuten kuuluukin, kotitalouden kokojakaumissa. Ne osoittavat erityisesti, että yhden hengen kotitalouksien osuus kasvoi selvästi:

> Alkuperäisellä kertoimella yhden hengen talouksia tuli 30,4%, mutta korjatulla kertoimella 35,1%. Tämä korjaus on epäilemättä oikeansuuntainen, mikä näkyy myös alkuhaastatteluun osallistuneiden kertoimen perusteella lasketusta vastaavasta osuudesta 31,8%. On kuitenkin mahdollista, että korjattu kerroin liioittelee yhden hengen talouksien ja myös kaikkien talouksien lukumääriä ehkä runsaalla 10000:lla. Tämä harha johtuu etupäässä käytettävissä olleen katoaineiston epätarkkuudesta (vrt. luku 7 ja liite 4).

Taulukon 2 kolmas tieto, kotitalousväestön määrä vuodelta 1985, on lähes sama kummallakin menetelmällä. Sen sijaan vuoden 1981 luvut poikkeavat selvästi. Menetelmä O yliarvioi noin 25000:lla kotitalousväestön, kun taas menetelmä A aliarvioi sen noin 10000:lla. Tämä johtuu olennaisesti siitä, että otoskehikko vuonna 1981 koostui yli 15-vuotiaista ja vuonna 1985 yli 0-vuotiaista. **Jälkimmäisestä kehikosta voidaan väestötiedot estimoida luotettavammin kuin edellisestä.** Tämä taas parantaa kotitalouksien lukumäärä- ja kulutusestimaattien laatua.

TAULUKKO 3. Kulutustuloksia vuosien 1981 ja 1985 kotitalous-tiedustelusta menetelmillä O, A ja B. Selityksiä myös tekstissä.

ESTIMOITU MUUTTUJA		VUOSI	MENETELMÄT		
			O	A	B
Kulutus yhteensä	mk/talous	1981	55869	54260	54008
	"	1985	80150	77400	76545
	indeksi/talous	1985	100,0	96,6	95,5
	indeksi/summa	1985	100,0	100,6	99,5
Pääryhmä 1: Elintarvikkeet	mk/talous	1981	14056	13654	13621
	"	1985	18092	17527	17315
	indeksi/talous	1985	100,0	96,9	95,7
	indeksi/summa	1985	100,0	101,1	99,9
Pääryhmä 2: Vaatteet ja jalkineet	mk/talous	1981	3678	3549	3509
	"	1985	5282	5064	5007
	indeksi/talous	1985	100,0	95,9	94,8
	indeksi/summa	1985	100,0	100,1	98,8
Pääryhmä 3: Asunto, lämpö, valo ym.	mk/talous	1981	11096	10834	10891
	"	1985	16384	15956	15871
	indeksi/talous	1985	100,0	97,4	96,9
	indeksi/summa	1985	100,0	101,5	101,0
Pääryhmä 4: Koti-talous-kalusto	mk/talous	1981	3253	3159	3142
	"	1985	5543	5277	5233
	indeksi/talous	1985	100,0	95,2	94,4
	indeksi/summa	1985	100,0	99,2	98,3
Pääryhmä 5: Sairauden ja terveydenhoito	mk/talous	1981	1176	1149	1149
	"	1985	2491	2443	2446
	indeksi/talous	1985	100,0	98,1	98,2
	indeksi/summa	1985	100,0	102,2	102,3
Pääryhmä 6: Liikenne ja tietoliikenne	mk/talous	1981	9201	8848	8786
	"	1985	13969	13420	13173
	indeksi/talous	1985	100,0	96,1	94,3
	indeksi/summa	1985	100,0	100,1	98,2
Pääryhmä 7: Virkistys- ja harrastus-toiminta	mk/talous	1981	4743	4587	4516
	"	1985	7066	6778	6691
	indeksi/talous	1985	100,0	95,9	94,7
	indeksi/summa	1985	100,0	99,9	98,7

Taulukkoon 3 on koottu aggregoituja kulutustietoja. Siinä "Kulutus yhteensä" tarkoittaa kaikkien 8 pääryhmän summaa. Menetelmän B tapauksessa se on kuitenkin estimoitu suoraan eikä laskettu pääryhmien estimaattien summana, mikä myös olisi mahdollista. "Indeksi/summa" merkitsee koko maan kulutussumman, totaalin, avulla laskettuja estimaatteja suhteessa menetelmällä O saatuihin estimaatteihin.

Menetelmä A pienensi taulukossa 3 mainittuja kulutuskeskiarvoja, mikä johtui olennaisimmin kotitalouksien keskikoon pienentymisestä. Muutamat disaggregoidut kulutuserät kuitenkin kasvoivat. Esimerkkeinä tällaisista mainittakoot "asumistuki" ja kodin ulkopuolella syödyt "puurot, vellit ja suurimot". Kasvu näissä tapauksissa johtui lähinnä siitä, että yksinäisten osuus (huomaa, että korjaus nostaa yksinäisten painoja) näiden erien "kuluttajina" oli huomattavan suuri. Vastaavantyyppinen tekijä(1) liittyy myös mm. pääryhmään 5 so. "sairauden- ja terveydenhoitomenoihin", jossa korjaus aiheutti varsin vähäisen pienennnyksen.

Regressiokorjaus pienensi useissa tapauksissa keskiarvoja edelleen. Pääryhmä 5 on vähäinen poikkeus molempina ajankohtina, pääryhmä 3 vain vuonna 1981. Regressiokorjaus vaikutti eniten vuonna 1981 pääryhmään 7 ja vuonna 1985 pääryhmään 6.

Totaaliestimaatteihin korjausten vaikutus on selvästi vähäisempi kuin keskiarvoihin. Menetelmä A vaikutti vuonna 1985 keskimäärin kasvattavasti, menetelmä B vähentävästi. Sairauden- ja terveydenhoitomenot on nytkin selvin poikkeus edellisestä suunnasta ja kotitalouskalustomenot jälkimmäisestä. Näitä totaaleja koskevia tuloksia voitaneen pitää odotettuina ja toivottuinakin. Silti voidaan kysyä, onko tuloksissa edelleen harhaa ja mihin suuntaan? Tarkkaa vastausta tähän kysymykseen ei voida antaa. Seuraavassa kappaleessa tilannetta yritetään hahmottaa simuloinnin avulla.

---

(1) Vanhoilla ihmisillä sairauden- ja terveydenhoitomenot ovat keskimääräistä suurempia. He ovat toisaalta tavallista yleisemmin yksinasuvia.

8.2. Simuloituja tuloksia vuoden 1981 aineistolla

Simulointi eli todellisen tilanteen matkiminen keinotekoisesti soveltuu myös otostilanteisiin, tavoitteena testata menetelmien ja sovellutusten hyvyyttä. **Simulointia käytetään toisaalta estimaattien harhan tutkimiseen ja toisaalta estimaattien tarkkuuden so. varianssin empiiriseen mittaamiseen** mm. kun halutaan tarkistaa teoreettisesti kehitetyn varianssin pätevyyttä, tai jos analyyttistä varianssia ei ole saatu muodostetuksi.

Simuloitujen koesarjojen käyttö on varsin yleistä otanteorian tutkimuksessa, vaikkakin asetelmat ovat usein varsin pelkistettyä (ks. esim. Särndal ja Svensson 1987). Tässä yhteydessä olen yrittänyt rakentaa muutaman koesarjan lähes sellaisena kuin se kotitaloustiedustelun käytännössä on tapahtunut. Koesarjat tähtäävät vain estimaattien harhan tutkimiseen. Varianssien tutkimiseen tarvittaisiin huomattavasti suurempi simulointien määrä. Menettely on ollut seuraava:

(i) Haastateltujen aineisto on asetettu perusjoukoksi.

(ii) Perusjoukosta on poimittu satunnaisesti sellainen kotitalouksien joukko, joka vastaa todellisuudessakin haastatteluun saatua otosta. Siten vuoden 1981 aineistosta "simuloitu otos" tähtäsi 72,6%:iin koko haastatellusta aineistosta. Satunnaisuudesta johtuen simuloitu otos ei toki sisältänyt täsmälleen odotusarvoa eli 5349 taloutta (joka on 72,6% 7368:sta).

(iii) Todellisuuden paremmaksi matkimiseksi on poimintaa ositettu eri taustatekijöiden mukaan. Tähän on rajattomat mahdollisuudet. Siten simuloituja aineistoja voidaan muodostaa eri lähtökohdista niin paljon kuin halutaan. Tässä yhteydessä aineistoja on muodostettu vain kolme. Simulointi I perustuu yhden tekijän, so. kotitalouden koon mukaiseen ositukseen. Simuloinnissa II on otettu huomioon lisäksi alue ja nettotulotaso. Simuloinnin III ositus taas perustuu kotitalouden kokoon, alueeseen ja varallisuustasoon.

(iv) Saaduista simuloiduista aineistoista lasketaan eri menetelmillä estimaatteja. Silloin voidaan tuloksia verrata suoraan todelliseen tietoon, so. kohdan (i) perusjoukosta laskettuun tietoon.

Taulukko 4 sisältää tuloksia em. kolmesta simuloinnista. Menetelmä A parantaa kaikissa tapauksissa menetelmällä O laskettua kotitalouksien lukumääräestimaattia, mutta tulokset jäävät kuitenkin vielä pienemmiksi kuin todellinen estimaatti. Vastaavasti kotitalouden keskikoon estimaatit jäävät vielä liian suuriksi.

TAULUKKO 4. Tuloksia kolmesta eri simuloinnista vuoden 1981 kotitaloustiedustelusta menetelmillä O, A ja B. Simuloinnilla tavoiteltiin taulukkojen 2 ja 3 menetelmällä O laskettuja tuloksia, jotka ilmoitetaan myös tässä taulukossa suluissa.
. = ei mielekäs.

ESTIMOITU MUUTTUJA JA SEN ESTIMAATTI	SIMULOINTI	MENETELMÄT		
		O	A	B
Kotitalouksien lukumäärä, 1000 (1873)	I	1817	1864	.
	II	1811	1828	.
	III	1802	1829	.
Kotitalouden keskikoko (2,54)	I	2.67	2.57	.
	II	2.67	2.58	.
	III	2.62	2.58	.
Kokonaiskulutus, mk (55869)	I	58088	56445	56109
	II	58664	57033	56267
	III	58038	57206	56058
Pääryhmä 1: Elintarvikkeiden kulut, mk (14056)	I	14764	14315	14193
	II	14853	14420	14222
	III	14483	14273	14023
Pääryhmä 2: Vaatteiden yms. kulut, mk (3678)	I	3883	3751	3707
	II	3883	3756	3691
	III	3846	3788	3690
Pääryhmä 3: Asunto-, lämpö yms. kulut, mk (11096)	I	11451	11173	11104
	II	11459	11184	11065
	III	11468	11334	11139
Pääryhmä 4: Kotitalouskalusto- yms. kulut, mk (3253)	I	3401	3298	3272
	II	3432	3340	3295
	III	3411	3362	3288
Pääryhmä 5: Sairauden- ja terveydenhoito, mk (1176)	I	1229	1204	1196
	II	1207	1182	1172
	III	1203	1186	1164
Päärymä 6: Liikennekulut, mk (9201)	I	9605	9280	9267
	II	9778	9469	9313
	III	9670	9480	9258
Pääryhmä 7: Virkistys- ja harrastustoiminta, mk (4743)	I	4898	4755	4713
	II	5008	4857	4768
	III	4912	4846	4731

Kuten edellä todettiin, keskikulutusestimaatteihin vaikuttavat paljolti kotitalouden koon estimaatit. Tämä näkyy menetelmällä A lasketuissa tuloksissa selvästi siten, että estimaatit pienenevät, mutta eivät kuitenkaan riittävästi. Jos sen lisäksi sovelletaan menetelmää B, tulokset yleensä edelleen pienenevät ja samalla lähenevät "oikeata" arvoa. Silti estimaatit olivat useimmiten "oikeaa" arvoa suurempia.

Kokonaisuudessaan nämä keinotekoiset todellisuuden kuvaukset vahvistavat käsitystä, jonka mukaan menetelmä A parantaa olennaisesti keskiarvo- ja totaalituloksia. Jos sen lisäksi sovelletaan menetelmää B, tulokset paranevat vielä, mutta eivät enää yhtä merkittävästi. Se, kannattaako molempia menetelmiä käyttää, riippuu monista tekijöistä. Tähän asiaan palataan kappaleessa 8.3. (ks. myös liite 7).

8.3. Keskustelua

Sekä todellisista että keinotekoisista eli simuloiduista aineistoista tuotetut estimointitulokset osoittavat korjausmenetelmät hyödyllisiksi ainakin keskiarvo- ja totaalitasolla. Harhaa toki jää jäljelle, mutta sen suuruuden mittaaminen on hankalaa.

Keskiarvot, totaalit tai vastaavat tiedot eivät kuitenkaan ole ainoat tuotettavat estimaatit kotitalousaineistoista (ks. myös luku 1). Tärkeätä on saada hyviä estimaatteja myös jakaumista ja eri tekijöiden välisistä yhteyksistä. Tämä asettaa ehtoja hyväksyttäville korjausmenetelmille. Ongelmia on vähemmän, jos korjaukset perustuvat ainoastaan painotusmenetelmään A. Tällöin jakaumat muuttuvat siltä osin kuin painotus muuttuu ja vastaavasti yhteydet muuttujien välillä muuttuvat, mutta alkuperäiset piirteet säilyvät sikäli, että alkuperäisiä havaintoarvoja ei muuteta.

Regressiokorjausmenetelmä B (tai C) sen sijaan voi aiheuttaa enemmän muutoksia jakauma- ja yhteystietoihin. Vaikutus riippuu siitä, miten tätä menetelmää sovelletaan. Seuraavat kaksi tapaa ovat mahdollisia:

(i) Määrätään keskiarvoestimaatit siten, että asetetaan selittäjien arvoiksi vastaavat poimitun aineiston S keskiarvot. Tämä voidaan tehdä ositteittain. Korjataan tämän jälkeen haastateltujen aineiston R arvoja kussakin ositteessa lasketun ja alkuperäisen estimaatin suhteella. Korjausten vaikutus ei ulotu tällöin perusaineistoon muuten kuin ositetasolla. Jakaumatkin muuttuvat vain näiltä osin, samoin yhteysestimaatit.

Jos regressioestimointia sovellettaisiin perustilastotuotannossa, olisi turvallisinta menetellä esitetyllä tavalla. Aineiston käyttö eri tarkoituksiin sujuisi ongelmitta ja erityisesti keskiarvot ja totaalit olisivat alkuperäisiä parempia.

(ii) Imputoidaan aineisto regressiomallin avulla eli täydennetään myös kato-osan tiedot. Imputoidut arvot muodostetaan "ennustamalla" (engl. predict) tulosmuuttujien arvot kato-osan havaintoyksiköiden x-arvojen avulla. Haastateltuja koskevat tiedot (a) voidaan pitää ennallaan tai (b) ne korjataan vastaavalla ennusteella. Imputointi asettaa omat rajoituksensa aineiston käytölle. Yhteystiedot "häiriintyvät" erityisesti tapauksessa (b) sitä enemmän, mitä huonompia regressiomallit ovat. Myös kato-osan imputointi aiheuttaa yhteystietoihin muutoksia, jotka eivät vastaa todellisuutta.

Jakaumatarkasteluun imputointi näyttää eräissä tapauksissa soveltuvan varsin hyvin, kuten liitteen 7 tarkastelu osoittaa. On huomattava, että sen esimerkissä mallien selitysasteet olivat korkeita. Keskiarvojen ja totaalien estimaatit ovat imputoiduista aineistoista yhtä suuria kuin kohdan (i) menetelmällä, jos taustaaineistot x-muuttujien määräämiseksi ovat samat.

9. Keskivirheitä vuodelle 1985

Estimaattien keskihajontaestimaatit so. keskivirheet antavat tietoa estimaattien tarkkuudesta. Luvuissa 6 ja 7 on esitetty näiden laskentatavat. Tässä luvussa esitetään muutamia tuloksia eri menetelmillä ja vertaillaan niitä.

TAULUKKO 5. Keskivirheitä (%) eräille muuttujille vuoden 1985 kotitaloustiedustelusta O, A ja B.
Selitykset: T = totaali . = ei mielekäs
K = keskiarvo

ESTIMOITU MUUTTUJA	TASO	MENETELMÄT		
		O	A	B
Kotitalouksien lkm koko maassa	T	0,69	0,86	.
Kotitalouksien lkm Uusimaalla	T	1,59	2,06	.
Kotitalouksien lkm Lapissa	T	2,74	3,19	.
Kokonaiskulutus	T	0,60	0,66	0,44
	K	0,78	0,83	0,59
Pääryhmä 1: Elintarvikkeet	T	0,57	0,65	0,58
	K	0,79	0,86	0,81
Pääryhmä 2: Vaatteet ja jalkineet	T	1,49	1,54	1,48
	K	1,60	1,67	1,61
Pääryhmä 3: Asunto, lämpö ja valo	T	0,62	0,72	0,50
	K	0,68	0,73	0,51
Pääryhmä 4: Kotitalouskalusto	T	2,25	2,24	2,21
	K	2,32	2,31	2,28
Pääryhmä 5: Sairauden ja terveyden hoito	T	2,39	2,52	2,45
	K	2,38	2,49	2,42
Pääryhmä 6: Liikenne ja tietoliikenne	T	1,50	1,56	1,47
	K	1,62	1,70	1,62
Pääryhmä 7: Virkistys- ja harrastustoiminta	T	1,33	1,41	1,36
	K	1,46	1,54	1,50

Taulukosta 5 näemme, että useimmiten keskivirheet kasvavat, jos siirrytään korjaamattomasta menetelmästä O painokorjattuun menetelmään A (vähäinen poikkeus on kotitalouskalusto ja isompi terveyden- ja sairaanhoito). Sen sijaan regressiomenetelmän käyttö pienentää keskivirheitä sitä enemmän mitä parempia mallit ovat. Esillä olevissa tapauksissa menetelmän B keskivirhe on aina pienempi kuin menetelmän A ja useimmiten pienempi kuin menetelmän O. Erot eivät kuitenkaan yleensä ole kovin suuria eli estimaattorin tehokkuus-kriteerin perusteella regressioestimointi ei olisi kovin hyödyllinen. Joissakin tapauksissa keskivirhe kuitenkin pienenee olennaisesti, kuten kokonaiskulutusta tai pääryhmää 3 (asunto- ym menot) mallitettaessa menetelmällä B. Siten vähintään näihin muuttujiin regressioestimointi tehoaisi hyvin.

Tulosten perusteella voidaan myös kysyä, onko menetelmä A sittenkään hyvä menetelmä, koska sillä on saatu suurempia keskivirheitä kuin menetelmällä O. Tähän kysymykseen voidaan vastata kahdesta vasta-argumentista lähtien:

(i) Täytyy selvittää, ovatko estimaattorit harhattomia vai eivät. Jos ovat, tilannetta tarkastellaan kohdassa (ii). Jos ne taas ovat harhaisia, kuten käytännössä aina on mm. kadosta ja kadonkorjausmenetelmästä johtuen, täytyisi tutkia, kuinka merkittäviä harhat ovat.

Tässä tilanteessa emme voi harhaa tarkasti selvittää, mutta joissakin tapauksissa voimme sitä arvioida. Esimerkiksi kotitalouksien lukumäärään sisältyy luvun 8 arvion perusteella noin 70000 talouden eli 3,5%:n harha menetelmällä O ja noin 10000 talouden eli 0,5%:n harha menetelmällä A.

Tällöin näemme, että "kokonaiskeskivirhe" (sekä estimaatin hajonnan että harhan huomioiva, ks. liitteen 2 loppu) menetelmällä O on 3,6% ja menetelmällä A 1,0%, joten ero menetelmän A hyväksi on huomattava.

(ii) Jos menetelmät keskiarvomielessä ja siitä johtuen totaalimielessä ovat harhattomia, ei menetelmää O taulukon tulosten valossa silti tarvitse pitää parempana kuin menetelmää A, jos tärkeitä ovat myös jakaumatiedot. On uskottavaa, että jakauman eri osiin sijoitetut korjaussolut antavat parempia jakaumatietoja kuin alkuperäiset. Sitä osoittavat myös keskivirheet: jakauman reunojen suurempi edustavuus aineistossa aiheuttaa samalla myös keskivirheen kasvua.

10. Yhteenveto

Kotitaloustiedustelu on erittäin mittava tiedustelu, joka tähtää ensisijaisesti kotitalouksien kulutustietojen estimointiin vuositasolla. Kulutustietojen lisäksi tuotetaan tietoja mm. kestokulutushyödykkeiden omistuksesta, yhteiskunnallisisten palveluiden käytöstä ja tuloista. Taustalle, luokitteluja varten, kerätään tietoja kotitalouksien ominaisuuksista, kuten niiden koosta, sosiaaliasemasta ja asuinalueesta.

Kotitaloustiedustelun tiedot kerätään kolmessa vaiheessa: alkuhaastattelulla, tilinpidolla ja loppuhaastattelulla. Tietojen keruu on näin ollen raskasta sekä Tilastokeskuksen haastattelijoille että talouksille. Tästä syystä vuonna 1985 kaikki keskeiset tiedot saatiin alle 70%:lta otokseen poimituista talouksista. Tämä aineiston väheneminen eli ns. kato kasvoi noin 3 %-yksikköä vuoden 1981 tiedustelusta, vaikka Tilastokeskuksessa tehtiin monia toimintoja kadon vähentämiseksi.

On selvää, että katoa tulee aina haastatteluaineistoihin jäämään. Koska kato ei ole satunnaista, se vinouttaa tuloksia. Kadon vaikutusta voidaan vähentää myös tiedonkeruun jälkeen, tarkoituksena parempien tulosten estimointi. Tässä raportissa on keskitytty tällaisten menetelmien kartoittamiseen ja soveltamiseen kotitaloustiedusteluun ja muihin samankaltaisiin aineistoihin.

Katokorjausmenetelmät voidaan jakaa kahteen pääryhmään: **painotus- ja imputointimenetelmiin.**
Edellisillä korjataan alkuperäisten havaintoyksikköjen painotusta siten, että uusiin painoihin sisältyisi myös katoon kuuluvien yksikköjen osuus. Imputointimenetelmien lähtökohtana on puuttuvien tietojen korvaaminen mahdollisimman hyvillä estimoiduilla arvoilla, jolloin katoa ei tavallaan enää ole.

Tässä tutkimuksessa on testattu molempia menetelmiä. Painotusmenetelmistä on keskitytty ns. vastaustodennäköisyysmallin, kutsuttu **menetelmäksi A,** *tutkimiseen. Imputointimenetelmistä on testattu regressiomalliin perustuvaa menetelmää, josta käytetään nimitystä* **menetelmä B.** *Lisäksi on vertailuihin käytetty alkuperäistä katokorjaamatonta menetelmää, jota kutsutaan* **menetelmäksi 0.**

Katokorjaus on mahdollista vain, jos on käytettävissä muutakin informaatiota kuin mitä saadaan haastateltujen aineistosta. Tällaista ns. **lisä-, oheis- tai apuinformaatiota** *saatiin kotitaloustiedustelun katokorjauksia varten väestö-, vero- ym. rekistereistä. Hyödyksi käytettiin myös alkuhaastattelusta saatu tieto sellaisilta kotitalouksilta, joilta sen jälkeen ei enää vastauksia saatu.*

Katokorjausmenetelmät on aina sopeutettava po. otantaasetelmaan. Tässä tapauksessa tilannetta monimutkaistaa se, että otoskehikko on henkilöpohjainen väestön keskusrekisteri, kun taas aineisto kerätään kotitalouskohtaisesti. Tämän vuoksi poimintakin on henkilöpohjainen, kuitenkin alueellisesti ositettu. Poimittujen eli ns. kohdehenkilöiden ympärille muodostetaan aluksi rekisterien avulla ns. asuntokunnat ja myöhemmin haastattelujen avulla kotitaloudet.

Asuntokunta ja kotitalous ovat eri käsitteitä: edellinen sisältää samassa huoneistossa asuvat ja jälkimmäinen edellyttää lisäksi, että asuntokunnan jäsenillä on yhteinen taloudenhoito. Tästä johtuen kato-osan aineisto on jonkin verran huonompitasoisempi kuin haastateltujen aineisto ja vaikeuttaa myös katokorjausmenetelmien käyttöä.

Menetelmää A käytettäessä muodostetaan kullekin kotitaloudelle korotuskerroin so. paino joka ilmoittaa kuinka montaa kotitaloutta ko. kotitalous vastaa koko perusjoukossa. Korotuskertoimen lauseke vuoden 1985 tiedustelussa oli seuraava:

$$w_k(A) = \frac{M_h}{n_h \cdot m_{kh}} \cdot \frac{1}{p_c(k)}$$

Lausekkeen symbolit on määrätty seuraavasti:

h = alueellinen poimintaosite, yhteensä 24 kpl

M = otoskehikon henkilöiden lukumäärä; otoskehikko koostui yli 0-vuotiaista ilman laitoksissa yms. asuvia henkilöitä.

n = kotitalouksien lukumäärä poimintavaiheessa ilman jälkeenpäin havaittua ylipeittoa

m_k = otoskehikkoon kuuluvien henkilöiden lukumäärä kotitaloudessa k

$p_c(k)$ = solun c **yleistetyllä lineaarisella mallilla** estimoitu vastaustodennäköisyys, jossa mallissa selitettävässä muuttujassa käytettiin logistista skaalaa ja jossa oli neljä selittäjää:

- Alue luokiteltuna neljään ryhmään
- Kunnat luokiteltuina keskus- ja reuna-aluekuntiin
- Perherakenne luokiteltuna 8 luokkaan; muuttujan luokittelu sisälsi piirteitä kotitalouden koosta, lasten määristä ja jäsenten iästä
- Omaisuustulot luokiteltuna kahteen ryhmään: ne joilla niitä oli ja joilla niitä ei ollut.

Solujen kokonaismäärä oli siten 128.

Menetelmän A malli osoitti mm., että kato oli huomattavan suurta etelän keskusalueiden yhden hengen talouksissa. Hyvin sen sijaan vastasivat lapsettomat nuoret parit ja perheet joissa oli alle koulukäisiä lapsia. Itä-Suomessa asuvat ja omaisuustuloja saaneet vastasivat myös keskimääräistä paremmin.

Menetelmän A korjaava vaikutus tietyssä solussa on sitä suurempi mitä enemmän kato poikkeaa keskimääräisestä. Tämä näkyi mm. yhden hengen talouksien määrän selvänä kasvuna verrattuna menetelmään 0. Vastaavasti koko maan kotitalouksien määrä kasvoi selvästi eli yli 4%. Kotitalouskohtaisia kulutuskeskiarvoja menetelmä A muutti keskimäärin noin 4-5% alaspäin, mutta joissakin tapauksissa tulos myös kasvoi ja vastaavasti vähennys saattoi lähetä 10%:ia.

Nämä korjaukset on arvioitu oikeansuuntaisiksi. Arvioissa on käytetty hyväksi mm. tulonjakotilaston tietoja, joka tilasto sisältää eräitä samoja tietoja kuin kotitaloustiedustelu. Myös harhan tutkimistarkoituksessa tehty simulointikoesarja osoitti korjaukset oikeansuuntaisiksi.

Regressioimputoinnin *lähtökohtana on rakentaa haastateltujen aineistosta mahdollisimman hyviä selitysmalleja tulosmuuttujille. Selittäjiksi kelpaavat vain sellaiset muuttujat, joista on tietoa myös katoosasta tai koko perusjoukosta. Estimoinnissa on useita mahdollisuuksia riippuen mm. estimoitavasta suureesta. Jos esimerkiksi halutaan estimoida tulosmuuttujan keskiarvo, voidaan menetellä kahdella vaihtoehtoisella tavalla:*

(i) Mallilla lasketaan ennustearvot (engl. predicted values) puuttuville havainnoille, siis katoosalle, jonka jälkeen tästä täydennetystä aineistosta lasketaan keskiarvo.
(ii) Estimoituun malliin sijoitetaan selittävien muuttujien keskiarvot, jotka on saatu joko koko poimitusta aineistosta tai koko perusjoukosta, jonka jälkeen malli tuottaa tulosmuuttujan keskiarvoestimaatin.

Menetelmää B voidaan soveltaa monenlaisiin muuttujiin, onhan aineistossa runsas 1000 keruutason muuttujaa joita taas voidaan aggregoida useilla tavoilla. Soveltaminen disaggregoituihin muuttujiin tuotti vain harvoin kohtuullisen selitysvoiman omaavia malleja. Syynä tähän oli se, että muuttujien arvot yksittäisille talouksille ovat vaihtelevia, sisältäen huomattavan usein nollia ja joskus myös hyvin suuria arvoja. Mallin toimivuus parani, jos tulosmuuttujat olivat riittävän aggregoituja, kuten esimerkiksi seuraavissa tapauksissa:

- Kokonaiskulutusta malli selitti noin 62%:sti. Tärkeimmät selittäjät olivat nettotulot verotuksesta, kulkuvälineiden verotusarvo ja kouluikäisten lukumäärä kotitaloudessa.

- Asunto-, lämpö- ja valomenoja malli selitti noin 48%:sti. Paras selittäjä oli nytkin nettotulot, seuraavaksi parhaita taas veronalaiset varat ja kouluikäisten määrä.

- Liikennemenojen mallin selitysaste oli 36%. On luonnollista, että paras selittäjä oli nyt liikennevälineiden verotusarvo. Muut merkitsevät selittäjät olivat nettotulot ja työikäisten lukumäärä kotitaloudessa.

- Kokeilujen paras malli saatiin käytettävissä oleville tuloille, selitysasteen ollessa 84%. Hyvä tulos perustuu siihen, että jo verotuksesta saatavien tulojen osuus käytettävissä olevista tuloista on merkittävä, keskimäärin yli 90%.

Regressioestimaattien, so. keskiarvojen ja totaalien, selvitys osoitti, että huonojen mallien tapauksessa tulokset muuttuivat menetelmiin O tai A verrattuna varsin vähän, eivät siis myöskään huonontaneet estimaatteja. Hyvien mallien tapauksessa keskiarvot ja totaalit muuttuivat yleensä enemmän. Testaus simuloinnilla ja vertailut muiden aineistojen estimaatteihin osoittivat korjaussuunnan oikeaksi.

Regressiomallin sovellutus tapahtui sen jälkeen, kun painotuskorjaus menetelmällä A oli jo suoritettu. Siis mallin painoiksi otettiin menetelmän A, eivät menetelmän O painot. Keskiarvoihin ja totaaleihin vaikutti

ensin suoritettu menetelmä A selvästi enemmän kuin mitä sen jälkeen vaikutti menetelmä B.

Keskiarvojen ja totaalien lisäksi tarvitaan tietoja myös kulutus- ja tulojakaumista sekä niiden keskinäisistä yhteyksistä. Näiden hyvää estimointia vaikeuttaa tietojen keruuaikojen vaihtelu: tilinpitotiedot kerätään kahdelta viikolta, haastattelut kuukaudelta, kolmelta kuukaudelta tai vuodelta, rekisterit ovat vuositasoisia. Koska tulokset estimoidaan vuositasolle, on erityisesti vuotta lyhyempään tiedon keruuseen perustuvien muuttujien tulkinta jakauma- ja riippuvuustutkimuksissa suoritettava varovaisesti. Katokorjaukset voivat lisätä ongelmia. Tässä tutkimuksessa on arvioitu, että painotuskorjaus ei kuitenkaan aiheuta lisää ongelmia, vaan pikemminkin poistaa niitä.

Regressioimputoinnin käyttö siten, että puuttuvat tiedot täydennetään ja tulokset tuotetaan tästä täydennetystä aineistosta, soveltuu sen sijaan huonommin tilastotuotantoon. Suurin syy tähän on tarve käyttää mahdollisimman alkuperäistä aineistoa jakauma- ja riippuvuustutkimuksiin. Keskiarvojen ja totaalien estimointiin regressioimputointi toki soveltuu. Jos mallit olivat hyviä, kuten erityisesti käytettävissä olevien tulojen tapauksessa, regressioimputointi osoittautui hyväksi muihinkin tutkimustarkoituksiin.

Kokonaisuudessaan olen kahden testatun menetelmän osalta päätynyt seuraaviin johtopäätöksiin:

- *Painotuskorjaus menetelmän A tapaan sovellettuna on erinomaisen soveltuva kotitaloustiedustelun tyyppisiin tilastoaineistoihin. Perusedellytyksenä on se, että apumuuttujista saadaan hyvää tietoa myös kato-osasta. Näiden tietojen laatua olisi pyrittävä parantamaan kotitaloustiedustelussa. Myös on tärkeätä, että menetelmään sisältyvät ns. korjaussolut muodostetaan hyvin. Tätä edesauttaa mielestäni mm. selitysmallin käyttö vastausalttiuden tutkimiseen. Solujen hyvyydelle voidaan asettaa myös kriteereitä. Mm. on hyvä, jos solut ovat homogeenisia ja jos tulosmuuttujat käyttäytyvät kunkin solun kato-osassa ja haastateltujen osassa samalla tavalla.*

- *Regressioimputointia en suosittele yleiseksi menetelmäksi niin massiiviseen tiedusteluun kuin kotitaloustiedustelu. Sitä kannattaa mieluummin käyttää yksittäisten, tutkimuksen kannalta keskeisten muuttujien katovirheen korjaamiseen. Tällaiseen*

tilanteeseen taas on vaikea etukäteen tuottaa valmiiksi imputoituja tuloksia, koska menetelmän soveltamiseen so. mallin spesifiointiin vaikuttaa myös tutkimuksen tavoite. Tästä syystä on aineiston tutkijalle tarvittaessa annettava käyttöön myös kato-osaa tai koko perusjoukkoa koskevaa informaatiota, jotta tämä voisi tehdä po. tarkoitukseen oman imputointisovelluksensa. On myös mahdollista, että aineistoihin sisällytetään useampia imputointiratkaisuja (ns. moni-imputointi), joista käyttäjä voi valita tilanteeseensa soveltuvimman.

Katovirhe on ehkä keskeisin otosaineistoihin sisältyvä systemaattiseksi katsottava virhe eli ns. **ei-otantavirhe.** *Aineistot sisältävät myös muita systemaattisia virheitä. Eri virheiden poistamiseen tähtäävät toimenpiteet eivät ole toisensa poissulkevia, mutta tietysti on aina pidettävä jalat maassa eli harkittava, kuinka suurella panoksella kunkin virheen poistamiseen ryhdytään.*

Suurten virheiden korjaaminen on tietysti tärkeämpää kuin pienten. Toisaalta on niin, että on syytä korjata kaikki virheet, jotka tunnetaan ja osataan korjata, vaikka tiedettäisiin aineistoihin silti jäävän suurempiakin virheitä. Kotitaloustiedusteluun esimerkiksi jäi kummankin katokorjauksen jälkeen huomattavan suuri aliestimaatti alkoholin kulutukseen.

Otosaineistojen tutkimuksessa on tärkeätä mitata myös ns. **otantavirheitä,** *joilla tarkoitetaan estimaattorien varianssiestimaattoreita tai keskivirheitä. Hyvien keskivirhe-estimaattorien kehittäminen katokorjaustilanteisiin osoittautui hankalaksi. Esimerkiksi menetelmälle A saatiin kaksi lauseketta, joista vain ensimmäisellä, ns. Horvitz-Thompson-tyypin estimaattorilla, on tässä vaiheessa tuotettu tuloksia. Regressiokorjauksen keskivirhe saatiin hyvin lasketuksi vain totaalin estimaattoreille, vaikka keskiarvojen keskivirheet olisivat vähintäänkin yhtä tärkeitä.*

Lähdeluettelo

Bethlehen J.G. - Keller W.J. (1987). Linear Weighting of Sample Survey Data. J. of Official Statistics, Vol. 3, No. 2, pp. 141-153. Statistics Sweden.

Ekholm A. - Laaksonen S. (1987). Correcting for Nonresponse by Response Propensity in the Finnish Household Survey. 46th Session of the ISI. Contributed Papers. pp. 101-102. Tokyo.

Chapman D.W. - Bailey L. - Kasprzyk D. (1986). Nonresponse Adjustment Procedures at the U.S. Bureau of the Census. Survey Methodology, Vol. 12, no. 2, pp. 161-180. Statistics Canada.

Cochran W. G. (1977). Sampling Techniques. Third Edition. New York, Wiley&Sons.

Giles P. - Patrick C. (1986). Imputation Options in a Generalized Edit and Imputation System. Survey Methodology, Vol. 12, No. 1, pp. 49-60. Statistics Canada.

Hinkley D. (1988). Lectures on Bootstrap Techniques. 12th Nordic Conference in Statistics. Turku.

Horvitz D.G. - Thompson D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Population. J. Am. Statist. Associ. Vol 47, pp. 663-985.

Kalton G. - Kasprzyk D. (1986). The Treatment of Missing Survey Data. Survey Methodology, Vol. 12, no. 1, pp. 1-16. Statistics Canada.

Laaksonen S. (1986). Katovirheen korjaaminen tilastollisen mallin avulla kotitalousaineistossa. Moniste. 29 sivua. Tilastokeskus.

Laaksonen S. (1987). SAS:n käyttö kotitalouspohjaisen otanta-aineiston katovirheen korjaamisessa. SAS-seminaarin 14.5. seminaariesitelmämoniste.

Laaksonen S. - Ekholm A. (1987). Correcting for Nonresponse by Regression Modelling in the Finnish Household Survey. 46th Session of ISI. Contributed Papers. pp. 251-252. Tokyo.

Liedes M. - Manninen P. (1974). Otantamenetelmät. Helsinki: Gaudeamus.

Little R.J.A. (1986). Survey Nonresponse Adjustments. Int. Stat. Review, 54, 2, pp. 139-157.

Little R.J.A. - Rubin D.B. (1987). Statistical Analysis with Missing Data. New York, Wiley&Sons.

Logit Ky/Vartia Y (1984). Vuoden 1985 kotitaloustiedustelun otantasuunnitelma. Moniste. Tilastokeskuksen sisäinen paperi.

Martikainen T. (1988). Puuttuva punainen viiva. Äänestäminen vuoden 1987 eduskuntavaaleissa. Tilastokeskus. SVT.

Michaud S. (1986). Comparison of Weighting and Imputation Methods for Estimating Unsampled Data. Survey Methodology, Vol. 12, No. 2, pp. 197-206. Statistics Canada.

Mokken R.J. (1987). Dealing with Non-Response. Quarterly Journal of Central Bureau of Statistics. Netherlands Official Statistics. Vol. 2, no. 2. pp. 5-12. Voorburg.

Nordberg L. (1987). On the Statistical Analysis of Income and Consumption Distributions. 46th Session of the ISI. Contributed Papers. pp. 321-322. Tokyo.

Oh H.L. - Scheuren F.S. (1983). Weighting Adjustments for Unit Nonresponse, in Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography (W.G.Madow, I.Olkin and D.B.Rubin, eds.). New York, Academic Press.

Pahkinen E. (1986). Tilastolliset otantamenetelmät. Harjoituskirja itseopiskeluun. Jyväskylän yliopisto. Tilastotieteen laitos. 311.213.2 311(1)V

Platek R. - Gray G.B. (1986). On the Definitions of Response Rates. Survey Methodology, Vol. 12, no. 1, pp. 17-27. Statistics Canada.

Rao J.N.K. - Wu C.F.J. (1987). Methods for Standard Errors and Confidence Intervals from Sample Survey Data: Some Recent Work. Tokyo. September 8-16 1987. Invited Papers. International Assoc. of Survey Statisticians, Booklet. pp. 143-159.

Rao P.S.R.S. (1986). Ratio Estimation with Subsampling in Nonrespondents. Survey Methodology, Vol. 12, No. 2, pp. 217-230. Statistics Canada.

Rubin D.B. (1986). Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology, Vol. 12, No. 1, pp. 37-47. Statistics Canada.

Rubin D.B. (1987). Multiple Imputation for Nonresponse. New York, Wiley&Sons.

Sullström R. (1987). Alimman tuloviidenneksen toimeentulo Suomessa vuonna 1981. Kansantaloudellinen Aikakauskirja, nide 1.

Särndal C-E. (1986). A Regression Approach to Estimation in the Presence of Nonresponse. Survey Methodology, Vol. 12, No. 2, pp. 207-216. Statistics Canada.

Särndal C-E. - Svensson B. (1987). A General View of Estimation for two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. Int. Stat. Rev., 55, 3, pp. 279-294.

Särndal C-E. - Wright R.L. (1984). Cosmetic Form of Estimators in Survey Sampling. Scand J Statist 11: 146-156.

Ten Cate A. (1986). Regression Analysis Using Survey Data with Endogenous Design. Survey Methodology, Vol. 12, No. 2, pp. 121-138. Statistics Canada.

Tilastokeskus (1977). Kotitaloustiedustelu 1971. Osa II. Kotitalouksien tulot ja tulojen jakautuminen. Tilastollisia tiedonantoja. N:o 55. Helsinki.

Tilastokeskus (1983). Väestö- ja asuntolaskenta 1980. Osa IIIB. Tulot; asuntokunnat ja perheet. Suomen virallinen tilasto VI C:106.

Tilastokeskus (1986). Kotitaloustiedustelu 1981, osa III. Laatuseloste. Tilastollisia tiedonantoja nro 71. Helsinki.

Tilastokeskus (1987). Tilastojen laadun kuvaaminen. Ohjeita tuoteselosteiden laatimiseksi. Käsikirjoja nro 23. Helsinki.

Thomsen I (1973). A Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Nonresponse when Analyzing Survey Data. Statistisk Tidskrift, 4, 278-283.

Trempley V. (1986). Practical Criteria for Definition of Weighting Classes. Survey Methodoloy, Vol. 12, No. 1, pp. 85-97. Statistics Canada.

Williams P.D. - Nisselson H. (1987). Methodology for Missing Data in the U. S. National Health and Nutrition Examination Survey. 46th Session of ISI. Contributed paper. Tokyo.

LIITE 1: Luettelo keskeisistä symboleista

Tässä liitteessä esitetään luettelo keskeisistä symboleista. Aihepiiriä ml. symboleja hahmotetaan myös kuviossa 1.

c = solu tai solukko joka on muodostettu sopivan ristiintaulukon tai siihen liittyvän mallin avulla

C = solujen c joukko

COV = teoreettinen kovarianssi; sama "hatulla" (^) merkitsee estimoitua kovarianssia

g = kotitalouden jäsenten lukumäärä

h = osite sekä poiminta-, että yleensä myös jälkiositus- ja laskentavaiheessa

k = kotitalous k joka on muodostettu määrätyn kohdehenkilön ympärille

l = kuten k

M = otoskehikkoon kuuluvien (kohde)henkilöiden lukumäärä

menetelmät O, A, B ja C ovat tutkimuksen keskeiset menetelmät katokorjauksia varten; näistä O tarkoittaa katokorjaamatonta, A painokorjattua sekä B ja C regressioimputoinnilla korjattua menetelmää (ks. luvut 6-10)

N = Suomen kaikkien kotitalouksien lukumäärä; sen vastaava estimaattori sisältää lisäksi "hatun" (^)

n = kotitalouksien lukumäärä hyväksytyssä otoksessa, so. ilman ylipeittoa

P = tavallinen todennäköisyys; pieni p-kirjain merkitsee myös todennäköisyyttä (ks. esim. lauseke (4))

π = ns. sisältymistodennäköisyys; käytössä useita erilaisia todennäköisyyksiä (ks. teksti)

r = haastateltujen kotitalouksien lukumäärä otoksessa; voidaan käyttää myös alaindeksejä jolloin on kysymys määrättyyn muuttujajoukkoon vastanneista (kuvio 1)

R = haastateltujen kotitalouksien joukko

RV = suhteellinen varianssi; sen vastaava estimaattori sisältää "hatun" (^)

s^2 = otosvarianssi, jonka perässä olevien sulkujen sisällä ilmoitetaan muuttuja, jolle se lasketaan ja havaintojen määrä

S = otokseen poimittujen kotitalouksien joukko, ilman ylipeittoa (vrt. symboliin n)

U = Suomen kaikkien kotitalouksien muodostama joukko

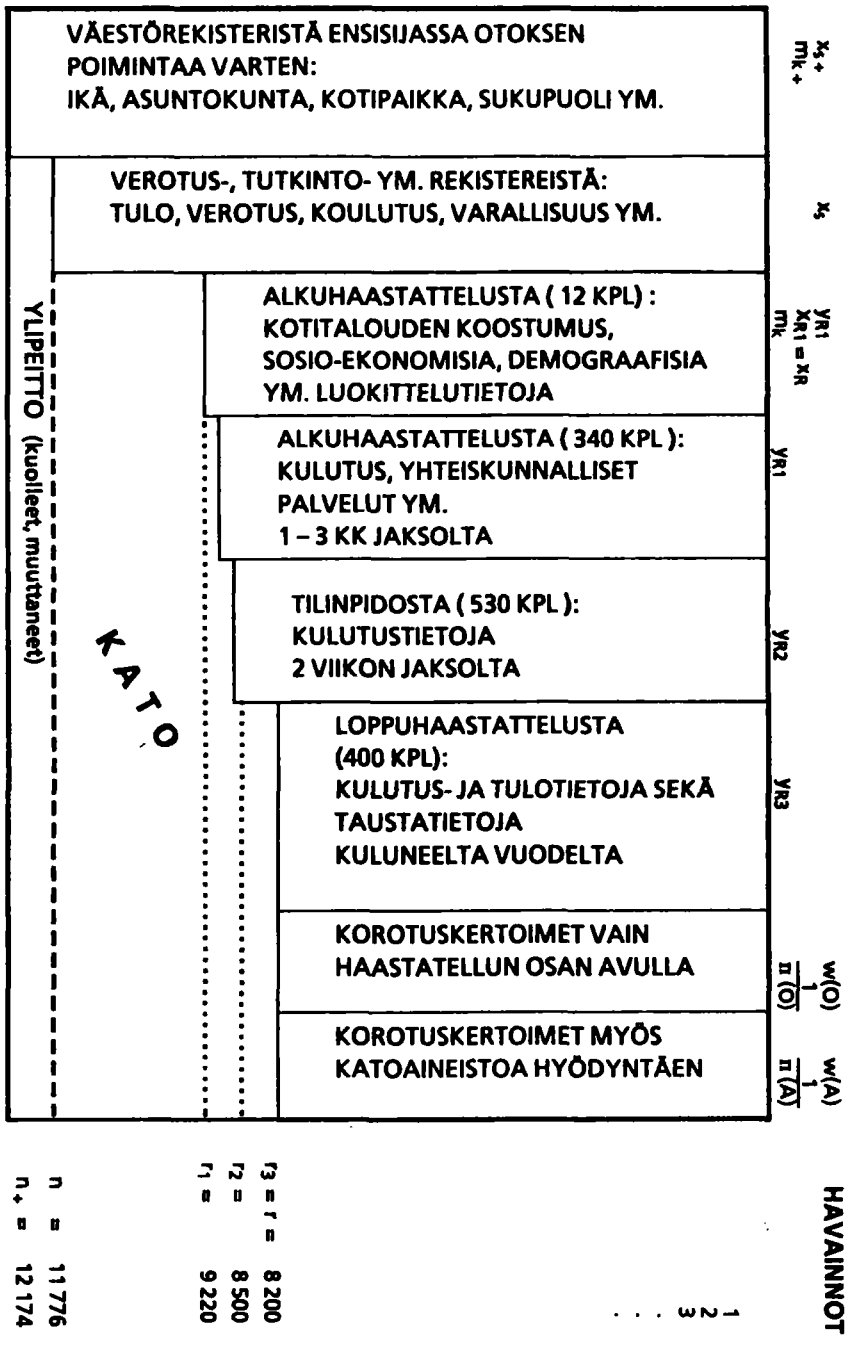
V = teoreettinen varianssi; jos se sisältää lisäksi "hatun", on kysymys vastaavasta estimoidusta varianssista eli varianssiestimaattorista

w = korotuskerroin ($\pi:n$ käänteisluku); erityisesti w(0) tarkoittaa "vanhaa" (0=old) katokorjaamatonta korotuskerrointa ja w(A) menetelmällä A so. luvun 6 menetelmällä saatua korotuskerrointa

x = muuttuja josta on käytettävissä tietoja koko perusjoukosta tai vähintään n:n havainnon otoksesta (engl. auxiliary variable); tyypillisiä x-muuttujia eli apumuuttujia tutkimuksessa ovat nettotulot,omaisuustulo ja muut verorekisterin muuttujat, kaikki laskettuna kotitaloutta kohti; isot X-kirjaimet viittaavat totaalia koskevaan tietoon, pienet yksittäisen kotitalouden tietoon tai kysymys on muuttujaryhmän yleissymbolista

y = tutkittava muuttuja eli tulosmuuttuja (target variable, outcome variable); tyypillisiä tulosmuuttujia tutkimuksessa ovat eri tuotteiden kulutusmäärät tai rahamäärät, kestokulutushyödykkeiden omistus ja käytetyt yhteiskunnalliset palvelut; isot ja pienet kirjaimet tulkittavissa samalla tavalla kuin x-muuttujalle.

z = x-muuttujaa vastaava tulkinta, mutta nyt tiedot ovat tyyppiä "per kotitalouden jäsen"



Kuvio 1. VUODEN 1985 AINEISTON YLEINEN KUVAUS

LIITE 2: Otanta-aineiston estimointiin liittyviä käsitteitä ja näkemyksiä

Otosaineiston (n = otoskoko) avulla pyritään arvioimaan eli estimoimaan perusjoukon (N = perusjoukon koko) eri tunnuslukujen arvoja. Siis jos otos muodostuu havainnoista $Y_1, Y_2, \ldots, Y_n$. pyritään niillä arvioimaan keskeisiä tietoja koko perusjoukon havainnoista $Y_1, Y_2, \ldots, Y_N$.

Mahdollisia estimoitavia tunnuslukuja on useita:

- yksiulotteisessa jakaumassa tyypillisiä ovat totaali eli havaintojen summa, keskiarvo, keskihajonta ja erilaiset jakauman osuuspisteet kuten mediaani ja desiilit.
- kaksiulotteisessa jakaumassa tyypillisiä ovat korrelaatiokerroin, regressiomallin parametrit ja jakaumaa kuvaavat pisteet.

Otostunnusluvun eli estimaattorin "hyvyydelle" voidaan asettaa erilaisia kriteerejä. Tavallisimmin käytetyt kriteerit ovat seuraavat:

- Estimaattori on **tarkentuva eli konsistentti,** jos otoksen ollessa 100%:nen, estimaattorin arvo on sama kuin perusjoukon vastaava parametrin arvo. Tämä ominaisuus ei käytännössä ole kuitenkaan kovin tärkeä, koska vain harvoin on sellainen tilanne, ettei tämä pätisi.

- Estimaattori on **harhaton,** jos sen odotusarvo on sama kuin perusjoukon vastaava arvo. Toisin sanoen estimaattorin arvot ovat keskimäärin samat kuin parametrin arvot. Jos estimaattori ei ole harhaton, se on harhainen. Harhattomuus on varsin tärkeä estimaattorin ominaisuus ja siihen tulee aina pyrkiä. Jos siihen ei päästä, tulee kuitenkin arvioida, missä tilanteissa ja millä ehdoilla harhaisuus ilmenee. Usein päästään harhaisen estimaattorinkin tapauksessa sellaiseen tilanteeseen, että sopivilla ehdoilla havaintomäärän kasvaessa harhaisuus vähenee. Tällaisen estimaattorin muodostamiseen tulee pyrkiä silloin, jos täysin harhattomaan ei päästä. Tällaisen estimaattorin sanotaan olevan asymptoottisesti harhaton.

Harhaa on kahta tyyppiä: (i) harha voi johtua otanta-asetelman ja estimointimenetelmän välisestä "häiriöstä" ja (ii) harha voi johtua aineistoon liittyvistä tekijöistä.

- Voi esiintyä myös tilanteita, jolloin on käytettävissä useampia harhattomia estimaattoreita. Tällöin sovelletaan yleensä ns. minimivarianssin periaatetta eli suositaan sellaista, jonka varianssi on pienin. Tätä estimaattoria sanotaan **tehokkaammaksi** kuin kilpailevia estimaattoreita. (Huom. voi esiintyä tilanteita, joissa harhaisen estimaattorin varianssi on pienempi kuin harhattoman).

Viimeksi mainittu ominaisuus on johtanut otosaineistojen yhteydessä käyttämään estimaattorin varianssia ja sen estimaattia estimaattorin tarkkuuden mittana. Siinä yhteydessä puhutaan myös ns. estimaatin keskivirheestä. Tältä pohjalta voidaan edelleen muodostaa estimaatille luottamusvälejä, jotka siis antavat selkeän kuvan tulosten tarkkuudesta otantamielessä eli perusinformaation otantavirheestä.

Tarkastellaan seuraavaksi esimerkkiä koskien keskiarvon estimointia katotilanteessa.

Olkoon otoskoko=n, joista vastaukset saatiin r:ltä. Vastaavat perusjoukon suureet olkoot N ja R. Kadon suuruus on siis n_2=n-r ja vastaavan "katoperusjoukon" koko N_2=N-R.

Voimme muodostaa tavanomaiset keskiarvot otokselle ja perusjoukolle:

$$\bar{y} = \sum_1^n y_i/n \text{ ja}$$

$$\bar{Y} = \sum_1^N y_i/N.$$

Käytännössä otoskeskiarvoa ei kuitenkaan ole olemassa, vaan ainoastaan vastaava keskiarvo vastanneille

$$\bar{y}_1 = \sum_1^r y_i/r$$

sekä tuntematon keskiarvo vastaamattomille

$$\bar{y}_2 = \sum_{r+1}^{n} y_i/(n-r).$$

Edelleen meillä on näitä vastaavien osaperusjoukkojen keskiarvot

$$\bar{Y}_1 = \sum_{1}^{R} y_i/R \text{ ja}$$

$$\bar{Y}_2 = \sum_{R+1}^{N} y_i/(N-R).$$

Estimaattorien hyvyyden tutkimista varten tarvitsemme edelleen varianssit perusjoukolle

$$s^2 = \sum_{1}^{N} (y_i - \bar{Y})^2/(N-1)$$

ja osaperusjoukoille

$$s_1^2 = \sum_{1}^{R} (y_i - \bar{Y}_1)^2/(R-1)$$

$$s_2^2 = \sum_{R+1}^{N} (y_i - \bar{Y}_2)^2/(N-R-1)$$

Perusjoukon keskiarvo voidaan kirjoittaa sen osajoukkojen keskiarvojen summana

$$\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2,$$

jossa $W_1 = R/N$ ja $W_2 = (N-R)/N$.

Havaitsemme, että otoskeskiarvo $\bar{y}_1$ on vastaavan osaperusjoukon keskiarvon harhaton estimaattori, mutta sillä on estimoitaessa koko perusjoukon keskiarvoa $\bar{Y}$ seuraavan suuruinen harha

$$B = W_2 (\bar{Y}_1 - \bar{Y}_2).$$

Aidossa otantatilanteessa emme tiedä, kuinka suuri harha kulloinkin on. Sitä voidaan kuitenkin arvioida esimerkiksi oheisinformaation avulla. Toinen mahdollisuus on suorittaa sopivia simulointikokeita. Toisin sanoen pyritään matkimaan todellista tilannetta "tekoaineiston" avulla. "Tekoaineiston" ja siihen liittyvän simuloinnin tulisi olla harhan tutkimisen mielessä mahdollisimman oikea.

Kuten edellä todettiin estimaatin tarkkuus muodostuu otantavirheen lisäksi harhasta (ja mahdollisista muista virheistä kuten mittausvirheistä, joita ei tässä yhteydessä kuitenkaan tarkastella). Näiden avulla voidaan muodostaa uusi tarkkuuden mittari (kutsuttu mm. Liedes-Mannisen (1974) kirjassa bruttovarianssiksi), joka on harhan neliö lisättynä estimaattorin varianssilla ehdolla ettei se sisällä harhaa. Esitetyssä esimerkissä tällainen tulos on seuraava

$$[(N-r)/rN]\ s_1^2 + B^2 \ .$$

LIITE 3: Tilinpitopohjainen korotuskerroin

Lausekkeessa (1) (luku 2) on esitetty korotuskerroin $w_k(0)$. Sen muodostamisessa on otettu huomioon alueosite ja talouden koko, muttei sen sijaan, miltä ja kuinka pitkältä ajanjaksolta kukin tieto on kerätty. Tässä suhteessa aineiston laatu vaihtelee: vuositason tietojen lisäksi on käytössä kolmen kuukauden, kuukauden, neljän viikon ja kahden viikon tiedonkeruujaksoja.

Jokaiselle erilaiselle tiedonkeruujaksolle voitaisiin muodostaa oma korotuskerroin. Tätä ei ole pidetty tarpeellisena eikä järkevänäkään. Sen sijaan tilinpitotietojen tarkempi estimointi on katsottu tarpeelliseksi jo vuoden 1981 tiedustelussa. Silloin meneteltiin niin, että ositus ulotettiin alueiden lisäksi 26 tilinpitojaksoon so. ositteita tuli yhteensä 35 x 26 = 910. Näin tiheän osituksen takia moniin ositteisiin tuli havaintoja vain muutamia. Vastaavasti yksittäisen kotitalouden korotuskerroin voi nousta satunnaissyistä liiankin suureksi vääristäen tuloksia.

Vuoden 1985 tiedusteluun muodostettiin saman idean mukaan oma korotuskerroin tilinpitotietojen laskentaan. Tällä tavoin voidaan poistaa ongelmia, joita koituu kadon vaihteluista vuodenaikojen mukaan. Katohan on ollut suurinta kesällä ja pienintä tiedustelun alkuvaiheissa talvella. Tästä seuraa, että tilinpitopohjaisella kertoimella laskettu tulos poikkeaa selvimmin vuositason kertoimella lasketusta ns. kausihyödykkeiden kohdalla (esim. joulukinkut tai mämmi).

Vuoden 1985 tiedustelun tilinpitopohjaisen kertoimen laskemiseksi on em. "harvojen havaintojen ongelmaa" vähennetty siten, että laskentaositteet on muodostettu yhdistämällä peräkkäiset kaksi tilinpitojaksoa. Näin saatiin 13 ajanjaksoa ja koko laskennassa oli siis 24 x 13 = 312 ositetta. Tällöin pienin havaintomäärä oli 10, joka saavutettiin Ahvenanmaalla heinäkuussa.

Lausekkeella (1) saatiin määrättyä menetelmän 0 mukaiset korotuskertoimet. Tilinpitonäkökulman huomioiminen menetelmässä A voidaan sen sijaan tehdä useilla eri tavoilla. Yksi mahdollisuus on se, että solujen muodostamisessa olisivat mukana myös tilinpitojaksot. Tässä yhteydessä ei tällaisia sovellutuksia ole tehty. Sen sijaan on oletettu, että menetelmän A pohjana oleva vastaustodennäköisyysmalli on voimassa kussakin 13 tilinpitopohjaisessa ositteessa. Tältä pohjalta saadaan lähtien menetelmän 0 mukaisesta katokorjaamattomasta tilinpitokorotuskertoimesta $wt_k(0)$ katokorjat-

tu tilinpitokorotuskerroin $wt_k(A)$ seuraavalla tavalla:

$$wt_k(0) = \frac{M}{r \cdot m_k \cdot 13} = \frac{M/13}{r \cdot m_k} = \frac{M/13}{n \cdot m_k \cdot r/n} = \frac{M/13}{n \cdot m_k \cdot p_t}$$

jossa

p_t = vastaustodennäköisyys tilinpitojaksossa t (t=1,...,13)
ja muut merkinnät samoja kuin lausekkeessa (1).

Jos edelleen katokorjaamaton vuosikorotuskerroin on

$$w_k(0) = \frac{M}{r \cdot m_k} = \frac{M}{n \cdot m_k \cdot r/n} = \frac{M}{n \cdot m_k \cdot p_h}$$

jossa

p_h = vastaustodennäköisyys ositteessa h (h=1,...,24)

ja katokorjattu vuosikorotuskerroin

$$w_k(A) = \frac{M}{n \cdot m_k \cdot p_c}$$

jossa

p_c = vastaustodennäköisyys solussa c (c=1,...128)

niin katokorjattu tilinpitokorotuskerroin on

$$wt_k(A) = \frac{w_k(A)}{w_k(0)} \cdot wt_k(0)$$

$$= \frac{M/13}{n \cdot m_k \cdot p_t} \cdot \frac{p_h}{p_c} .$$

LIITE 4: Ikäkorjaus vuoden 1985 tiedustelussa

Vuoden 1985 tiedusteluun tehtiin katokorjauksen lisäksi ns. ikäkorjaus siksi, että alun perin poimitusta aineistosta poistettiin sellaiset kotitaloudet, jotka olivat olleet 3 edellisen vuoden aikana joko tulonjakotilaston tai työvoimatiedustelun otoksissa. Näin vähennettiin osallistumisrasitusta ja uskottiin myös kadon vähentyvän.

Voidaan olettaa, että talouksien poistaminen tapahtui satunnaisesti, joten tästä ei synny ongelmaa. Sen sijaan ongelmana on näiden kolmen tiedustelun erilainen otoskehikko. Kotitaloustiedustelun otoskehikkoon kuuluivat yli 0-vuotiaat, tulonjakotilaston yli 15-vuotiaat ja työvoimatiedustelun 15-74 - vuotiaat. Poistetut taloudet ovat siten sellaisia, joiden kohdehenkilöt olivat vähintään 15-vuotiaita. Tämä aiheuttaa vinoutumaa perusaineistoon. Sen vuoksi tehtiin aineistoon ennen muita korjauksia ikäkorjaus.

Korjaus perustui kaavaan (a), joka vähentää alle 15 - vuotiaiden kohdehenkilöiden ja lisää muiden painotusta. Kaavassa oletetaan esitetyn mukaisesti, että poistetut taloudet ovat jakautuneet samalla tavoin kuin muut samaan ikäryhmään kuuluvat taloudet ja poistetut "palautetaan aineistoon" vastaavalla tavalla:

$$w(\text{ikäkorjattu}) = w(0)\cdot\frac{q\cdot r}{rq} \qquad (a)$$

jossa q = 1, jos kohdehenkilö alle 15 - vuotias,
>1 muissa ikäryhmissä siten kuin poistettuja kotitalouksia oli; tosiasiassa käytettiin vain kahta muuta ikäryhmää so. 15-74 - vuotiaat ja yli 74 - vuotiaat, koska tarkempaa tietoa ei ollut käytettävissä;
r = haastateltujen määrä ositteessa;
rq = laskennallinen haastateltujen määrä ositteessa ehdolla, ettei poistettuja talouksia olisi ollut.

Korjauksen vaikutus tuloksiin ei ollut kovin suuri, koska ao. tavalla poistettuja talouksia oli vain 5% (605 kpl): kotitalouksien lukumäärän estimaattia korjaus kasvatti noin 6000:lla. Lukumäärän kasvu johtuu olennaisesti siitä, että talouksien, joiden kohdehenkilö oli alle 15 - vuotias, alkuperäiset korotuskertoimet w(0) olivat keskimäärin muita pienempiä.

LIITE 5: Ajatuksia katoaineistoja koskevan tiedon parantamiseksi

Katokorjausmenetelmät toimivat sitä paremmin, mitä luotettavampia tietoja aineistojen kato-osista on käytettävissä. Tämä edellyttää myös aineistojen "putsaamista" ali- ja ylipeitosta eli aineistot eivät saisi sisältää peittävyysvirheitä. Tässä liitteessä esitetään muutamia ajatuksia näiden tietojen tason kohentamiseksi ajatellen nimenomaisesti kotitaloustiedustelun tyyppistä aineistoa. Ensiksi esitetään eräitä yleisiä, sitten peittävyyteen ja katoon liittyviä näkemyksiä ja lopuksi pohditaan katotalouden korvaamista toisella taloudella.

Yleisiä näkemyksiä

Katokorjausmenetelmiin liittyy aina hyvin tärkeänä tekijänä se, että kotitalouden koostumus eli tiedot kuhunkin kotitalouteen kuuluvista henkilöistä on määritelty hyvin. Näiden tietojen avulla saadaan sen jälkeen selville heitä koskevia taustatietoja eri rekistereistä. Jos siis kotitalous on huonosti määritelty, ovat myös taustatiedot huonosti määriteltyjä. Haastateltujen henkilöiden osalta nämä tiedot saadaan inhimillisiä tms. virheitä lukuunottamatta hyviksi, mutta muilta osin on vaikeuksia. Kuitenkin tilannetta voidaan parantaa nykyisestään.

Ylipeitto, alipeitto ja kato

Jos kotitalous tavoitetaan ja haastatellaan, voidaan aineistoon helposti merkitä tiedot kotitalouden jäsenistä, jotka otoksen poiminnan jälkeen kuuluvat ylipeittoon (kuolleet, maasta muuttaneet) tai alipeittoon (syntyneet, adoptoidut, talouteen avioliiton tms. syyn takia siirtyneet). Katoa ei tällöin näiden perustietojen osalta synny, paitsi jos myöhemmin havaitaan tiedot niin huonoiksi, ettei niitä katsota aiheelliseksi käyttää.

Jos taloutta ei tavoiteta, kysymys on ongelmallisempi. Talous voi olla edelleen olemassa (samalla voi olla muuttanut koostumustaan) tai poistunut. Olisi tärkeätä

edelleen pyrkiä saamaan joitakin olennaisia piirteitä taloudesta selville. Tähän on uuden osoitteen hankinnan lisäksi mahdollisuus käyttää rekisteritietoja: esim. poistaa rekisteritietojen avulla kuolleiden taloudet tai kuolleet talouden jäsenet tai lisätä syntyneet tiedusteluvuoden lopulla. Tällaisia toimenpiteitä ei ole yleensä riittävästi voitu tähän mennessä tehdä, mutta tilannetta voitaisiin varmasti parantaa vähin kustannuksin.

Jos vastaaja ei useista yrityksistä huolimatta ole tavoitettavissa, voidaan vähintään talouden koostumuksesta yrittää saada ajankohtainen tieto naapurin, talonmiehen tai isännöitsijän avulla. Tämä koostumuksen tarkistus auttaisi jälleen jatkoanalyysissä.

Jos taas talous ei suostu haastatteluun, on mahdollista yrittää saada selville ainakin talouden koostumus henkilötunnuksineen ja taas voitaisiin rekisteritietojen avulla tuottaa entistä parempi aineisto.

Jos vastaaja taas ei ole kyvykäs vastaamaan sairauden, kielivaikeuksien tms. syyn takia, voidaan kuitenkin suorittaa vastaajan keskeisten taustatietojen tarkistus (vähintään kotitalouden koostumus ja mahdollisesti joitakin toimeentuloon liittyviä tekijöitä myös).

Näkökohta edellisen pohjalta: voidaanko katotalous korvata toisella?

Laajoissa tiedusteluissa kato muodostuu varsin isoksi. Kadon voidaan havaita myös olevan pikemminkin kasvussa kuin päinvastoin. Usein on tullutkin esille kysymys, voitaisiinko katoon joutuneen tilalle ottaa jokin toinen?

Yksiselitteistä vastausta esitettyyn kysymykseen en kykene antamaan. Voidaan kuitenkin sanoa, että korvaamisen ehtona on se, että korvaava talous on samanlainen kuin katotalous. Tällainen ei täydellisesti, siis yksittäisen havaintoyksikön tasolla, yleensä ole mahdollista. Tilastollisesti kohtuullisiin tuloksiin voidaan päätyä käsitykseni mukaan seuraavissa tilanteissa:

- Jos katoon joutuminen on satunnaista tutkittavan muuttujan suhteen. Tällaista tilannetta lähellä ollaan, kun haastateltavan osoite on tuntematon tai myös jos haastateltavaa ei tavoiteta. Korvaavat taloudet voidaan heidän tilalleen

valita vastaavasta ryhmästä esim. ottaen huomioon koon ja alueen. Vuoden 1985 tiedustelussa tällä tavoin olisi voitu korvata kuitenkin vain alle 10% kadosta.

- Kieltäytyminen on kuitenkin ollut yleisin kadon syy. Tällaisen talouden korvaamiseen voitaneen jossakin määrin pyrkiä soveltamalla hot-deck tai etäisyysmittaan perustuvaa imputointia. Toisin sanoen x-muuttujien avulla määritellään mieluimmin jo etukäteen kullekin taloudelle korvaava talous, toisin sanoen sen kanssa mahdollisimman samanlainen talous. Periaatteessa voitaisiin korvaamista jatkaa korvaavan talouden kieltäytymisen varalle, mutta tulos epäilemättä tätä kautta edelleen heikkenisi. Tässä vaiheessa en niin kuitenkaan tekisi, ennenkuin olisi saatu kokemuksia yhden vaiheen korvaamisesta.

LIITE 6: Selittäjäehdokkaita vuoden 1985 aineistosta regressiomallia varten

Muuttujat on otettu joko väestörekisteristä tai verorekisteristä, joiden käsikirjoista ilmenee niiden mittaustapa.

Symboli	Seloste
KOKO	Kotitalouden jäsenten lukumäärä
MUKU	alle 7-vuotiaiden lukumäärä
IMUKU	7-17 - vuotiaiden lukumäärä
TYOIKA	18-64 - vuotiaidden lukumäärä
KOULU	Päämiehen koulutusaste
ALUE	Useita mahdollisuuksia lääni- ja kuntapohjalta
VLUOKKA	Verovelvollisuusluokka
NETTU	Verotuksen tietojen perusteella lasketut nettotulot
VTULO	Valtionveronalaiset tulot
VARAT	Valtionveronalaiset varat
VELAT	Velat
KULKU	Kulkuvälineiden verotusarvo
SAIR	Sairausvähennykset verotuksessa
OMA	Omaisuustulot verotuksessa

LIITE 7: Käytettävissä olevien tulojen estimointi eri menetelmillä vuosien 1981 JA 1985 aineistoista

Tässä liitteessä esitetään vertailutuloksia käytettävissä olevien tulojen estimoinnista raportin perusmenetelmillä. Käytettävissä olevat tulot on valittu testimuuttujaksi ensiksikin siksi, että se on kotitaloustiedustelun keskeisiä jollei keskeisin sivutuote. Toisena syynä on se, että sen "sisartilastossa" so. tulonjakotilastossa tämä muuttuja on päätuote.

Aihe on myös sisällöllisesti kiinnostava, ei vähiten siksi, että aikavälin 1981-85 ja ylipäänsä 80-luvun tulojen ja tulojakauman muutoksista esitetään monenlaisia tietoja ja arvailuja. Jotkut väittävät tuloerojen kasvaneen, toiset pysyneen ennallaan, kolmannet tasoittuneen.

Käytettävissä olevat tulot estimoitiin menetelmillä O ja A sillä tavoin kuin luvuissa 6-8 on esitetty. Kutsun tuloksia, jotka molempina vuosina 1981 ja 1985 perustuvat menetelmään O, **korjaamattomalla** menetelmällä saaduiksi. Vastaavasti tulokset, jotka on molemmat saatu menetelmällä A, ovat **painokorjatulla** menetelmällä saatuja. Perustilastoon vuoden 1981 tulokset estimoitiin kuitenkin menetelmällä O ja vuoden 1985 tulokset menetelmällä A. Näin laskettuja tuloksia kutsun ko. aikavälin muutoksia esittäessäni **perustilastotuloksiksi.**

Em. kahden eri estimaatin ja kolmen eri aikasarjan lisäksi esitetään tässä liitteessä tuloksia, jotka on laskettu "aidolla" regressioimputoinnilla, so. puuttuvat havainnot on korvattu regressiomallista saaduilla estimaateilla. Näitä kutsutaan **imputoiduiksi** tuloksiksi. Malleissa käytettiin selittäjinä rekisteritietoja (ks. liite 6). Lopulliseen

- vuoden 1981 malliin tulivat selittäjiksi t-testin mukaisesti paremmuusjärjestyksessä nettotulot, kotitalouden koko, veronalaiset varat ja kulkuvälineiden verotusarvo ja

- vuoden 1985 malliin vastaavasti kotitalouden koko, nettotulot, valtionveronalaiset tulot, veronalaiset varat ja kulkuvälineiden verotusarvo.

Erityisesti parhaat selittäjät ovat varsin odotettuja. Onhan esimerkiksi nettotulot jo merkittävä osa käytettävissä olevista tuloista ja toisaalta kotitalouden koko vaikuttaa tulon määrään koska isommassa taloudessa tulonsaajia on yleensä enemmän. Näistä syistä mallien selitysasteet ovat korkeat: vuoden 1981 aineistossa 80% ja vuoden 1985 aineistossa 84%. Tämä osoittaa sen, että käytettävissä olevat tulot voidaan imputoida varsin hyvin myös katoaineistolle.

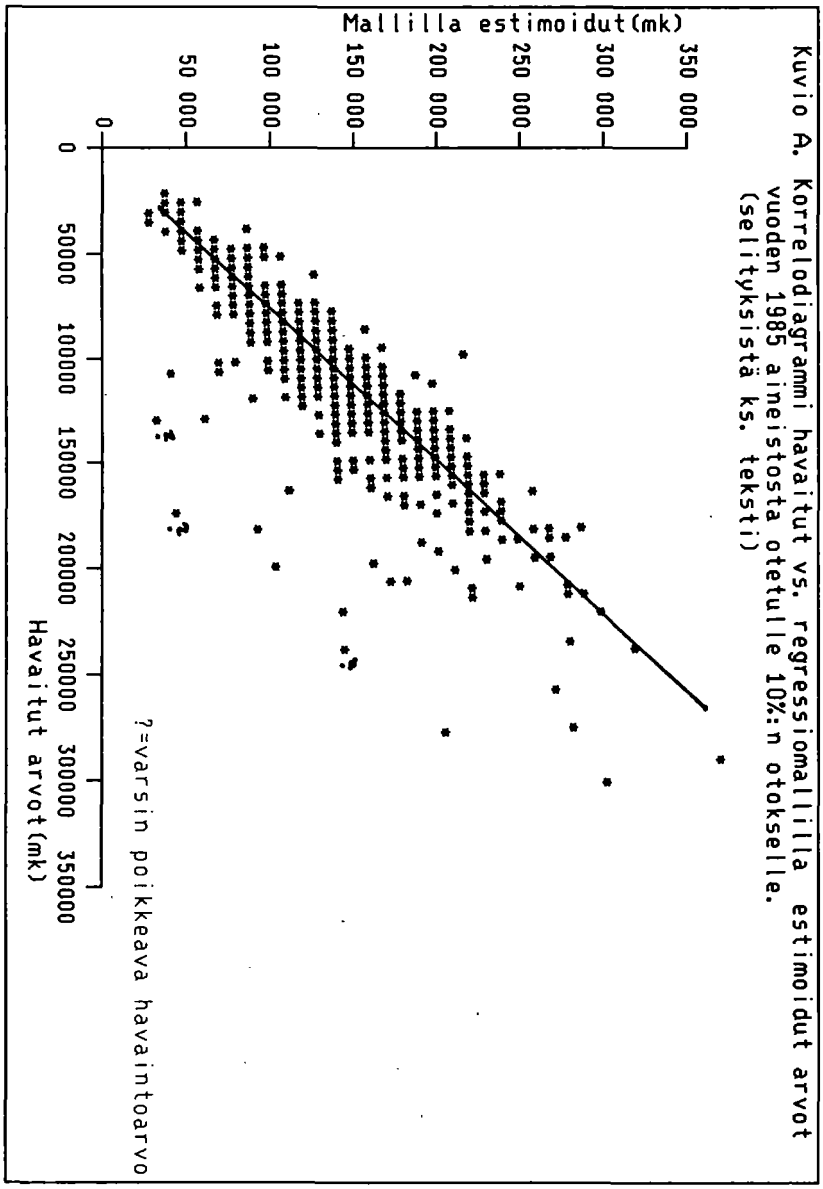
Kuviossa A tätä yhteensopivuutta havainnollistetaan korrelodiagrammilla, jonka havainnot ovat 10%:n satunnaisotos koko aineistosta. Havaintoyksiköt sijoittuvat varsin hyvin samalle suoralle (yksi piste voi sisältää useita yksiköitä), mutta muutamat havaintoyksiköt ovat kuitenkin varsin etäällä "yhteensopivuussuorasta". Yleensä huomattavat poikkeamat ovat suoran alapuolella. Tämä kuuluu luonnostaan mallituksen ominaisuuksiin: se on konservatiivinen (regressiivinen) eli ei välttämättä "usko" erikoishavaintoihin jollei niiden taakse löydy evidenssiä kaikkien malliin mukaan tulleiden selittäjien kautta.

Yksittäisten havaintojen huono sopivuus ei tämän vuoksi välttämättä merkitse sitä, että malli olisi huono. On mahdollista, että osa näistä havainnoista on tullut perusaineistoon virheellisenä (joko haastattelu- tai rekisteriteto tai molemmat). Tästä seuraa myös, että regressiomallia voitaisiin käyttää perusaineiston laatutestaukseen.

Tuloksia

Keskiarvoestimaatit kotitaloutta kohti laskettuina osoittavat samanlaista tendenssiä kuin luvussa 8 esitetyt tulokset kulutusmuuttujille. Siten korjaamaton estimaatti oli kumpanakin vuonna suurempi kuin katokorjattu estimaatti. Vuonna 1981 painokorjattu keskiarvo oli 2,9% ja vuonna 1985 3,5% katokorjaamatonta pienempi. Imputoiduille keskiarvoille luvut olivat 6,6% ja 4,1%.

Kotitalouskohtaisten estimaattien lisäksi tutkittiin kulutusyksikkökohtaisia estimaatteja. Kulutusyksikkökohtaisia estimaatteja käytetään erityisesti silloin, kun halutaan tutkia tulojen tasaisuutta tai epätasaisuutta. Muuten kotitalouden tulotaso riippuu liiaksi ja väärällä tavalla kotitalouden koosta ja sen tulonhankkijoiden määrästä. Tässä yhteydessä kulutusyksiköt olivat ns. OECD:n yksikköjä (ks. esim. Tilastokeskus 1984, 35).



Kuvio A. Korrelodiagrammi havaitut vs. regressiomallilla estimoidut arvot vuoden 1985 aineistosta otetulle 10%:n otokselle. (selityksistä ks. teksti)

Kulutusyksikköä kohti lasketut keskiarvot eivät muuttuneet läheskään niin paljon kuin kotitalouskohtaiset. Painokorjatut tulokset olivat vuonna 1981 0,3% ja vuonna 1985 0,1% ja imputoidut 0,8% ja 0,7% pienempiä kuin korjaamattomat. Vähäisemmät erot johtuvat olennaisesti siitä, että puuttuvaan dataan liittyvä tärkein tekijä kytkeytyy paljolti kotitalouden kokoon, so. kato-osan taloudet ovat keskimääräistä pienempiä. Kulutusyksikkökohtaisissa tuloksissa talouden koko ei enää vaikuta yhtä paljon.

Keskiarvojen muutokset aikavälillä 1981-85 näillä eri tavoilla laskettuina antoivat seuraavat tuloksia (%):

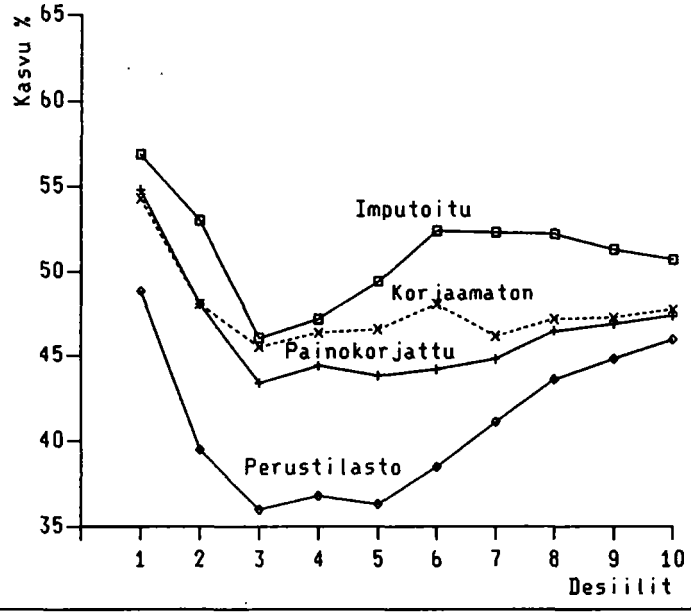
Menetelmä	Per kotitalous	Per kulutusyksikkö
Perustilasto	41,9	50,6
Korjaamaton	47,1	50,8
Painokorjattu	46,0	51,0
Imputoitu	51,0	51,0.

Tulokset osoittavat yllättävän suuria eroja kotitalouskohtaisissa muutoksissa. Erityisesti perustilastoista otetut tulokset antavat pienen kasvuluvun. Syynä on tietenkin se, että korjaus, joka tehtiin vain vuoden 1985 aineistoon, pienentää keskiarvoa. Korjaamattoman ja painokorjatun ero sen sijaan on pieni. Yllättävän suuri puolestaan on imputoiduilla estimaateilla laskettu tulomuutos. Selkeätä selitystä tähän en ole keksinyt. Kulutusyksikkökohtaiset kasvuluvut ovat hyvin samanlaisia, vaikka perustilastotulos on nytkin muita pienempi.

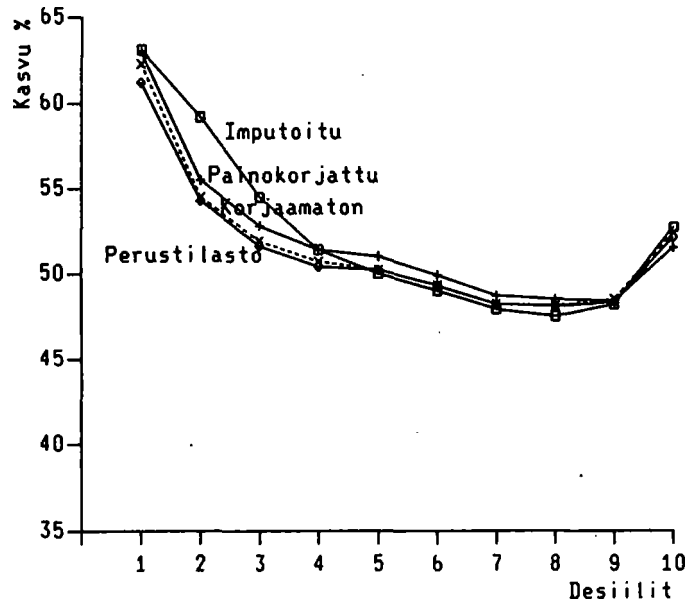
Keskeinen kiinnostuksemme liittyy kuitenkin tulojakaumien muutoksiin, tässä tapauksessa aikavälillä 1981-85. Kuvioihin B ja C on jakaumamuutoksia kuvattu desiilitasolla. Huomattakoon, että aineistoista on poistettu havaintoyksiköt, joiden käytettävissä olevat tulot olivat negatiiviset ja ylisuuret (vuonna 1981 500000 mk ja vuonna 1985 600000 mk). Tämä siksi, etteivät ne sotkisi 1. ja 10. desiilin tuloksia. Näitä havaintoja oli vain muutama.

Kuvion B mukaan eri menetelmät antavat ensimmäisten 4 desiilin osalta varsin samat tulokset tasoeroa lukuunottamatta. Ylimmissä desiileissä on selviä eroja. Perustilaston mukaan tulojen kasvu voimistui siirryttäessä ylimpiin desiileihin. Korjaamaton ja painokorjattu

Kuvio B. Kotitaloutta kohti laskettujen käytettävissä olevien tulojen kasvu 1981-85 desiileittäin neljällä menetelmällä



Kuvio C. Kulutusyksikköä kohti laskettujen käytettävissä olevien tulojen kasvu 1981-85 desiileittäin neljällä menetelmällä.



tulos osoittavat samaa, mutta varsin lievästi. Imputoitujen tulosten mukaan kehitys olisi ollut lähes sama 6-10. desiilissä. Kaikkien menetelmien mukaan huonointa tulokehitys oli 3-5. desiilissä.

Kokonaisuudessaan näemme, että kotitalouksien väliset tuloerot ovat lievästi lisääntyneet. Perustilaston mukaan tuloerot ovat lisääntyneet jopa selvästi. Kehitys ei kuitenkaan ole ollut selkeä, koska erityisesti jakauman alin desiili ja korjatuilla menetelmillä myös 2. desiili ovat parantaneet tulotasoaan muita desiilejä, myös ylimpiä desiilejä paremmin.

Kuviossa C kiinnittää huomiota se, että eri menetelmät antavat varsin samanlaisen kuvan tuloerojen muutoksista. Alkudesiileissä menetelmien erot ovat vielä jossain määrin havaittavissa, mutta 4. desiilin jälkeen erot ovat varsin "kosmeettisia".

Kokonaisuudessaan näemme, että kulutusyksikkökohtaisissa tuloissa on tapahtunut lievää tasoittumista. Tämäkään tasoittuminen ei ole ollut tasaista. Köyhimpien, so. kahden alimman desiilin, tulotaso on noussut eniten, mutta seuraavaksi paras kehitys onkin ollut ylimmällä ja 3. desiilillä. Tähän väliin sijoittuvien noin 60% kattavan "tavallisen kansalaisen" tulokehitys on sen sijaan ollut varsin samanlaista eli suhteellisesti mitaten huonoa.

Tulosten perusteella emme voi siis täysin selkeästi vastata kysymykseen, ovatko tuloerot kasvaneet vai tasoittuneet 80-luvun ensi puoliskolla. Kotitalouskohtaisten tulosten mukaan ne ovat kasvaneet, mutta kulutusyksikkökohtaisten tulosten mukaan taas eivät. Jos jälkimmäistä mittaustapaa pidetään parempana, kuten nykyisin yleensä tehdään, olisi vastaus siis: ne ovat lievästi tasoittuneet. Tällöinkään tämä yksi vastaus ei kerro koko kuvaa, koska tasoittuminen ei ole lähimainkaan suoraviivaista. Tasoittuminenhan johtuu alimpiin desiileihin kuuluneiden hyvästä tulokehityksestä, jota ylimmän desiilin myös hyvä kehitys ei aivan riitä kompensoimaan.

T I L A S T O K E S K U S

TUTKIMUKSIA

Tilastokeskus on julkaissut Tutkimuksia v. 1966 alkaen, v. 1985 lähtien ovat ilmestyneet seuraavat:

115. Aku Alanen, Yritystoiminnan julkinen rahoitus 1982. Helmikuu 1985. 48 s.

116. Naiset ja miehet työelämässä. Maaliskuu 1985. 47 s.

116. Women and men at work. May 1985. 47 p.

117. Iiris Niemi, Harrastusmittareiden luotettavuus, haastattelu- ja päiväkirjamenetelmällä saatujen tulosten vertailua. Maaliskuu 1985. 64 s.

118. Mikko Aaltonen, Jätetilastojen kehittäminen. Toukokuu 1985. 94 s.

119. Juha Kurmela – Eero Tanskanen, Käyttäjän rooli energian kulutuksessa. Kesäkuu 1985. 97 s.

120. Timo Mikander, Muuttoliike 1975 – 82. Kesäkuu 1985. 45 s.

121. Veli-Matti Lehtonen, Talonrakennusten perusparannus ja sen mittaus Suomessa. Elokuu 1985. 64 s.

122. Taru Sandström, Kansantalouden tilinpito, Valtio kansantalouden tilinpidossa. Lokakuu 1985. 88 s.

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 p.

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

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

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

129. Ritva Marin, Ammattikuolleisuus 1977 - 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. 1988:1 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 p.

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.

Tilastokeskus
Statistikcentralen
Central Statistical
Office of Finland

SVT

Tutkimuksia nro 147
Studies

Katovirheen korjaus kotitalousaineistossa

Correcting for Nonresponse in Household Data

Seppo Laaksonen

Kansikuva: Mikko Nurmi

Kato on lisääntyvä ongelma haastatteluaineistossa. Siitä aiheutuvia virheitä voidaan onneksi korjata. Hyväksi käytetään mm. rekistereistä saatavaa tietoa. Tutkimusraportin päätarkoituksena on esitellä menetelmä, jota on käytetty kotitaloustiedustelun 1985 katovirheen korjaamiseen. Raportti sisältää myös katsauksen alan viimeaikaisiin tieteellisiin tuloksiin. Tulosten soveltuvuudesta kotitalouspohjaisiin aineistoihin käydään laajaa keskustelua. Liitteenä on mm. vuosien 1981 - 85 tulonjaon muutoksia koskeva analyysi.

Julkaisujen myynti:

Tilastokeskus
PL 504
00101 Helsinki
(90) 17 341

Försäljning:

Statistikcentralen
PB 504
00101 Helsingfors
(90) 17 341

Hinta — Pris

65 mk

ISSN 0355-2071
ISBN 951-47-1992-1