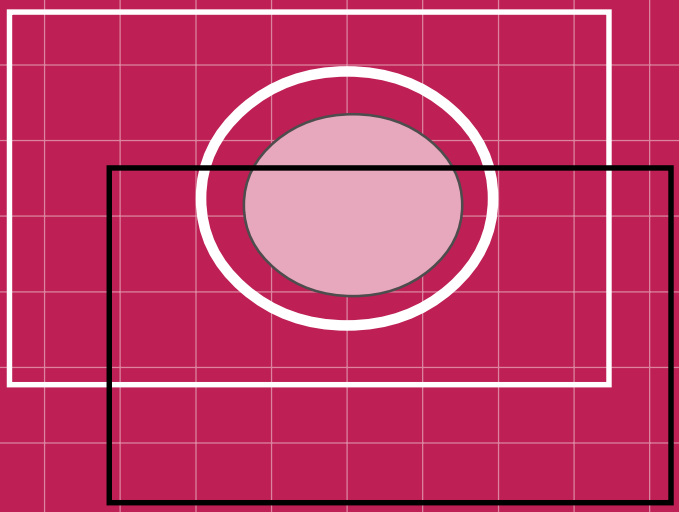
Estimation
in the presence of
Nonresponse
and
Frame Imperfections



by
Sixten Lundström and
Carl-Erik Särndal

SCB
Statistiska centralbyrån Statistics Sweden

ISBN 91-618-1107-6

Statistikpublikationer kan beställas från SCB, Publikationstjänsten, 701 89 ÖREBRO, e-post: publ@scb.se, telefon: 019-17 68 00, fax: 019-17 64 44. De kan också köpas genom bokhandeln eller direkt hos SCB, Karlavägen 100 i Stockholm och Klostergatan 23 i Örebro.

Aktuell publicering redovisas i SCB:s publikationskatalog och på vår webbplats (www.scb.se). Ytterligare hjälp ges av SCB:s Informationsservice, e-post: infoservice@scb.se, telefon: 08-506 948 01 eller 019-17 62 00, fax: 08-506 948 99.

SCB-Tryck, Örebro 2001.11

For further information, please contact:
Sixten Lundström, phone +46 19 17 64 96,
e-mail sixten.lundstrom@scb.se
Carl-Erik Särndal, phone +1 613 739 8836,
e-mail carl.sarndal@home.com

ISBN:

Printed in Sweden
SCB-tryck, Örebro 2001.

Director General's preface

Estimation in the Presence of Nonresponse and Frame Imperfections has been prepared as a Current Best Methods (CBM) manual, within the framework of quality improvement work at Statistics Sweden. It offers a review of effective methods to reduce the influence of nonresponse in statistical surveys, together with recommendations for their use. The manual will be used in production processes at Statistics Sweden and also has evident applications throughout the whole system of Swedish official statistics.

Svante Öberg

Authors' preface

Nonresponse has been a matter of concern for several decades in the relatively short history of survey theory and practice. As is especially apparent in the recent literature, the problem is viewed from two different, but complementary, angles: the prevention or avoidance of nonresponse before it occurs, and the special estimation techniques required once nonresponse has occurred.

The first angle is covered in Statistics Sweden (1997), a CBM manual entitled *Minska Bortfallet* (Reduce the Nonresponse). The second angle is examined in the present CBM manual. The two principal approaches for nonresponse adjustment, reweighting and imputation, are explained and illustrated. Also covered are some guidelines for dealing with a related set of problems, frame imperfections and coverage errors.

In writing this CBM we benefited greatly from a reading group, whose members commented on two preliminary versions of this document. The group consisted of experienced statisticians at Statistics Sweden: Claes Andersson, Stefan Berg, Claes Cassel, Jan Hörngren, Pär Lindholm, Peter Lundqvist, Lennart Nordberg, Jan Selén, Sara Tångdahl, Peter Vorverk. We gratefully acknowledge their contributions.

In Sweden as in many other countries, the practice of imputation is linked to legal aspects of the use of constructed variable values. Restrictions apply to the inclusion of such values in data files, particularly files on individuals. The material on imputation, especially Chapter 7, reflects these concerns relating to respondents' rights and the protection of privacy. In particular, Section 7.4 was written in consultation with Statistics Sweden's chief legal advisor, Birgitta Pettersson. We gratefully acknowledge this cooperation.

Örebro and Ottawa,
July, 2001,
Sixten Lundström
Carl-Erik Särndal

Contents

1. Introduction

This document is one in a series of Current Best Methods (CBM) manuals produced in recent years at Statistics Sweden. Their objective is to present in easily accessible form those techniques that are viewed as "best" for a given aspect of the statistical production process. They are intended as guides for survey statisticians in survey design, redesign and maintenance. Although produced mainly for statisticians at Statistics Sweden, they can also provide useful information for many other readers.

Nonresponse has long been a matter of concern in surveys. In the recent literature, the problem of nonresponse is viewed from two different (but complementary) angles: the prevention or avoidance of nonresponse before it has occurred, and the special techniques required in estimation when nonresponse has occurred.

The methods for prevention draw on knowledge from the behavioural sciences. This is natural because the data collection involves establishing contact with respondents, overcoming respondent scepticism and promoting a positive attitude to the survey. Motivational factors play an important role. Adjustment for nonresponse once it has occurred, on the other hand, draws on knowledge primarily from estimation theory. The reasoning is of a mathematical/statistical nature.

Both producers and users of statistics are well aware that nonresponse can have a negative impact on the quality of statistics. Considerable resources are therefore spent on improving data collection procedures, so as to prevent nonresponse from occurring. Research in this area is intensive and is reported in numerous articles. Some statistical agencies have their own guidelines for effective data collection, see, for example, Statistics Sweden (1997). Nevertheless, once data collection is concluded, one has to accept some, perhaps even considerable, nonresponse.

Nonresponse error is the most publicized of the "nonsampling errors", that is, those errors attributed to causes other than the limitation of the investigation to a sample only, rather than the entire population. There exists an enormous literature on the topic.

In a perfect world, a survey has no nonresponse; all selected elements will participate and provide all of the requested data. However, today's reality is very different. Missing data due to nonresponse is a normal although undesirable feature of any survey.

The objective of this CBM is to give an up-to-date account of methods of estimation for use when data collection has been “disturbed” by nonresponse. Nonresponse adjustment is not treated as an isolated issue. We embed nonresponse adjustment in the broader context of estimation. The issue is how to make the best possible estimates based on the data collected from those who respond to the survey, and on any relevant auxiliary information that one may have about the population and its elements, whether respondents or nonrespondents.

The title of this CBM reads “Estimation in the presence of nonresponse ... ” and continues “ ... and frame imperfections”. These last few words refer to issues closely related to nonresponse and missing data, namely, frame errors (or coverage errors). Frame undercoverage, in particular, is discussed in the concluding Chapter 11, though we do not provide an exhaustive treatment of this difficult problem on which the literature has remained comparatively silent.

This CBM is written for employees (“handläggare”) at Statistics Sweden, and in particular for its survey methodologists. The background required to understand the contents is not uniform throughout the document. The degree of technical difficulty can be described as follows.

Chapters 2 and 3 are written as a wholly non-technical overview. They present the main issues and outline general approaches to the treatment of nonresponse without going into technical arguments or specific solutions. Chapters 2 and 3 can be read with only a rudimentary background in statistical science.

More specific practical advice is presented in Chapters 4 to 11, which can be described as moderately technical. In order to well assimilate the material in Chapters 4 to 11, it should be sufficient to have a level of preparation corresponding to the C-level in statistics or mathematical statistics in the Swedish university system.

Appendices A to D contain mathematical and other technical material, including derivations and proofs of certain results stated in the preceding chapters. The appendices need not be read or understood in order to apply the methods explained in Chapters 4 to 11.

A statistician who is not actively involved in survey methodology work can easily read Chapters 2 and 3 as a general orientation, and should be able to follow parts of the subsequent chapters. An active survey methodologist should be able to read the whole document.

How should this CBM be used? It is not a handbook in the strict sense of the term. In other words, one should perhaps not hope to open this CBM at a certain page and find a tailor-made answer to a specific problem encountered in a survey. What the CBM does provide is a range of techniques commonly used for resolving issues arising from nonresponse. The survey statistician should always be prepared to adapt the techniques in this CBM to fit the environment of his/her own survey. This often requires a certain familiarity with statistical derivation techniques, including probabilistic evaluations of expected values and variances, and other tasks that survey methodologists are accustomed to performing.

This document is not the first on nonresponse produced at Statistics Sweden. A predecessor is the 1980 publication “Räkna med bortfall”, a title that can be translated in two different ways: “You can count on some nonresponse” and “Computing in the presence of nonresponse”. This publication was one of the products resulting from a large project undertaken at Statistics Sweden in the late 1970s, as a reaction of senior management to what were then considered “alarmingly high nonresponse rates” – rates that would be modest or almost insignificant in today's hardened survey climate. Nonresponse methodology has undergone considerable technical development since 1980, and the present document reflects these developments.

One important practical issue not discussed in this CBM is the following budgetary aspect: at a statistical agency, available resources must be split between (i) efforts to avoid nonresponse, and (ii) efforts to get high quality estimates despite nonresponse. The literature does not offer very concrete advice on this issue, nor is it covered in this CBM. The negative effects of

nonresponse on the quality of estimates can vary considerably from one survey to another. Costly attempts to reach a token minimum survey response rate, such as 80%, may not be the best use of the available resources.

2. The survey and its errors

2.1. Terminology

This CBM discusses issues arising in surveys carried out by national statistical institutes, such as Statistics Sweden. The objective of a survey is to provide information about unknown characteristics, called parameters, of a finite collection of elements, called a population (for example, a population of individuals, of households, or of enterprises). A typical survey involves many study variables and produces estimates of different types of parameters, such as the total or the mean of a study variable, or the ratio of the totals of two study variables. Sometimes different kinds of elements are measured in the same survey, as when both individuals and households are observed. Many surveys are conducted periodically, for example, monthly or yearly, and one of the objectives is then to get accurate measures of the change in a variable between two survey occasions.

The origin of a survey is usually that a government or some other users express a need for information about a social or economic issue, and that existing data sources are insufficient to meet this need. The first step in the planning process is to determine the survey objectives as clearly and unambiguously as possible. The next step, referred to as *survey design*, is to develop the methodology for the survey.

Survey design involves making decisions on a number of future survey operations. The data collection method must be decided upon, a questionnaire must be designed and pretested, procedures must be set out for minimizing or controlling response errors, the sampling method must be decided on, interviewers must be selected and trained (unless self-administered questionnaires are used), the techniques for handling nonresponse must be decided on, and procedures for tabulation and analysis must be thought out.

A survey will usually encounter various technical difficulties. No survey is perfect in all regards. The statistics that result from the survey are not error-free. The *frame* from which the sample is drawn is hardly ever perfect, so there will be *coverage errors*. There will be *sampling error* whenever

observation is limited to a sample of elements, rather than to the entire population. Also, no matter how carefully the survey is designed and conducted, some of the desired data will be missing, because of refusal to provide information or because contact cannot be established with a selected element. Since nonresponding elements may be systematically different (have larger or smaller variable values, on average) than responding elements, there will be *nonresponse error*.

These three types of error – sampling error, nonresponse error and coverage error – are discussed at length in this CBM. It is true that a survey will usually also have other imperfections, such as measurement error and coding error. These errors are not discussed.

Sub-populations of interest are called *domains*. If the survey is required to give accurate information about many domains, a complete enumeration of these domains may become necessary, especially if they are small.

The survey planner will probably first consider whether statistics derived from available *administrative registers* could satisfy the need for information. If not, a *census* (a complete enumeration of the population) may have to be conducted. If all domains of interest are at least moderately large, a *sample survey* may give statistics of sufficient accuracy.

These three different types of survey (survey based on administrative registers, census survey, sample survey) differ not only in the extent to which they can produce accurate information for domains, but also in other important respects. For example, sample surveys have the advantage of yielding diverse and timely data on specified variables, whereas statistics derived from administrative registers, although perhaps less expensive, may give information of limited relevance, because except in very fortunate cases, available registers are not designed to meet the specific information needs. On the other hand, a census might provide the desired information with great accuracy, but is very expensive to conduct. For a discussion of these issues, see Kish (1979).

Most of the issues raised in the following apply to all three types of survey. But often, we will have in mind a sample survey. Therefore, the term "survey" will usually refer to "a sample survey". We will now review some frequently used survey terminology.

A survey aims at obtaining information about a *target population*. The delimitation of the target population must be clearly stated at the planning stage of the survey. The statistician's interest does not lie in publishing information about individual elements of the target population (such disclosure is often ruled out by law), but in providing descriptive measures (totals or functions of totals) for various domains, that is, for various aggregates of population elements.

These unknown quantities are called *parameters* or *parameters of interest*.
For example, three important objectives of a Labour Force Survey (as conducted in most industrialized countries) are to get information about (i) the number of unemployed, (ii) the number of employed, (iii) the unemployment rate. These are examples of parameters. The first two parameters are *population totals*. The third is a *ratio of population totals*, namely, the number of unemployed persons divided by the total number of persons in the labour force.

Examples of other population parameters are *population means* – for example, mean household income – and *regression coefficients* – say, the regression coefficient of income (dependent variable) regressed on number of years of formal education (independent variable), for a population of individuals.

We can estimate any of these parameters with the aid of data on the elements of a probability sample from the population. We then assume that all sampled elements are measured for the variables whose totals define the parameter of interest.

The sample is drawn from the *frame population*, that is, the set of all elements that could possibly be drawn. The frame population and the target population are not always identical.

Sampling design is used as a generic term for the (usually probabilistic) rule that governs the sample selection. Commonly used sampling designs are: simple random sampling (SRS), stratified simple random sampling (STSRS), cluster sampling, two-stage sampling, and Poisson sampling. With the possible exception of SRS, these designs require some planning before sampling is carried out. STSRS requires well-defined strata composition. Cluster sampling requires a decision on what clusters to use. Two-stage

sampling requires that we define the first stage sampling units (the psu's) and the second stage elements (the ssu's).

Every sampling design involves two other important general concepts: (i) *inclusion probabilities* and (ii) *design weights*. The inclusion probability of an element is the probability with which it is selected under the given sampling design. The design weight of an element is the inverse of its inclusion probability. The sampling design may generate different probabilities of selection for different elements. In SRS and STSRS with proportional allocation, all inclusion probabilities are equal, but this is not the case in general.

The inclusion probability can never exceed one. Consequently, a design weight is greater than or equal to one. The inclusion probability (and the design weight) is equal to one for an element that is selected with certainty. Many business surveys include a number of elements (usually very large elements) that are "certainty elements". These form a sub-group often called a *take-all stratum*.

A majority of the elements have inclusion probabilities strictly less than one. For example, in an STSRS design, an element belonging to a stratum from which 200 elements are selected out of a total of 1600 has an inclusion probability equal to the sampling rate in the stratum, 200/1600 = 0.125, and its design weight is then 1/0.125 = 8. One interpretation often heard is that "an element with a design weight equal to 8 represents itself and seven other (non-sampled, non-observed) population elements as well". When it comes to estimation, the observed value for this element is given the weight 8. Another stratum in the same survey may have 100 sampled elements out of a total of 200. Each element in this stratum has the inclusion probability 100/200 = 0.5, and its design weight is then 1/0.5 = 2.

At Statistics Sweden, STSRS is used in a number of surveys. In particular, it is used in *surveys of individuals and households*, because the Swedish Total Population Register (see Example 2.2.1) contains a number of variables suitable for forming strata, such as age, sex and geographical area. It is often of interest to measure households as well as individuals in the same survey. We obtain a random sample of households from the random sample of individuals by identifying the households to which the selected individuals belong.

In *business surveys*, the distribution of the variables of interest is often highly skewed. The "giants" in an industry account for a major share of the total for a typical study variable such as industrial production. The largest elements (enterprises) must be given a high inclusion probability (probability one or very near to one). Many business surveys use coordinated sampling for small enterprises to distribute the response burden. This entails some control over the frequency with which an enterprise is asked to provide information over a period of time, say a year. In Sweden the JALES technique (see Atmer, Thulin and Bäcklund, 1975) is used for coordinated sampling.

The JALES technique and similar coordinated sampling techniques are based on permanent random numbers. "Permanent" means that a uniformly distributed random number is attached at birth to a statistical element (an enterprise), and remains with that element for the duration of its life.

2.2. A discussion of sources of error

In this section we discuss frame imperfections, sampling and nonresponse. Figure 2.2.1 is designed to support the discussion.

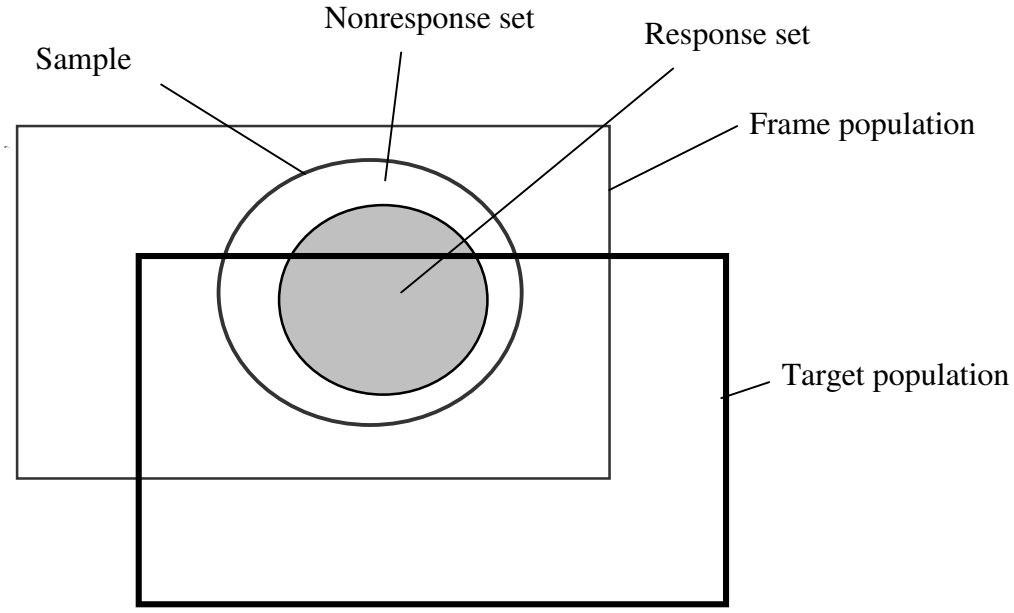


Figure 2.2.1.

Frame imperfections

We define the *target population* as the set of elements that the survey aims to encompass at the time when the questionnaire is filled in. This point in time is called the *reference time point for the target population*. The sampling frame is usually constructed at an earlier date, sometimes as much as twelve months earlier; this time point is referred to as the *reference time point for the frame population*. The lag between the two time points should be as short as possible, because the risk of coverage errors increases with the time lag. Three types of coverage error are commonly distinguished: *undercoverage*, *overcoverage*, and *duplicate listings*. We will now comment in particular on the first two of these. As the name suggests, duplicate listings refer to the type of errors occurring when a target population element is listed more than once in the frame.

Elements that are in the target population but not in the frame population constitute undercoverage. Especially in business surveys, a significant part of the undercoverage is made up of elements that are new to the target population but are not present in the frame population. These are commonly referred to as “births”. Undercoverage may, of course, also have other causes.

Elements that are in the frame population but not in the target population constitute overcoverage. Elements that have ceased to exist somewhere between the two reference time points can be a significant source of overcoverage. These elements are often referred to as “deaths”.

It follows that undercoverage elements have zero probability of being selected for any sample drawn from the frame population. This is an undesirable feature, because if the study variable values differ systematically for undercoverage elements and other population elements, there is a risk of biased estimates. Bias from overcoverage can usually be avoided if it is possible to identify the sample elements that belong to the overcoverage. One procedure is to treat these elements as a special domain. However, it is usually impossible to correctly classify all sample elements as belonging either to the target population or to the overcoverage. The problem becomes particularly acute for nonresponding elements, and biased estimates can be the result.

Attempts are usually made to keep the lag between the frame population reference time point and the target population reference time point as short as possible, but nevertheless, for practical reasons, the time lag is sometimes considerable. One reason may be slowness in the updating of the frame. The result of events that motivate a change or update of frame information is sometimes recorded only after a considerable delay. Births and deaths are examples of events that need to be recorded. Such events entail a change in the set of elements in the frame. Another example is a change in a variable value for an element existing in the frame, e.g. when updated information is received about the number of employees or the gross business income of an enterprise. It follows that the values recorded for a given frame variable may refer to different time points for different frame elements. This is not ideal, but it is a reality that has to be accepted.

The frame population for a given survey is sometimes created from a larger, more extensive set of elements, each having recorded values for a number of variables. An appropriate frame population for the survey is then constructed from this larger frame, using some of the variables as a tool for the delimitation process. Errors in the recorded variable values due to different reference times or other causes may detract from the effectiveness of the delimitation. Also, frame variables are often used before sampling for stratifying the population and/or after sampling for poststratifying the sample. Again, imperfect frame variable values may impede the efficiency of these important practices.

EXAMPLE 2.2.1. *The Total Population Register.*

The Total Population Register (TPR) aims to achieve a complete listing of the Swedish population. Register variables recorded for every person include the unique Personal Identity Number (PIN), name and address. This makes it possible to access every person for a broad range of surveys. The addresses are classified by Swedish administrative regions, such as counties ("län") and municipalities. Every piece of real estate is identified in the TPR by co-ordinates, which makes it possible to construct regions other than counties and municipalities. Other important register variables are date of birth, sex, civil status, country of birth and taxable income. Information about births, deaths, immigration, emigration and changes of other register variables is received by Statistics Sweden continuously, so in principle the register can be kept almost perfectly up-to-date. Persons arriving from abroad and intending to stay at least one year are entered in the register after

the necessary permission has been granted. Since this may take some time, the TPR at any given point in time has some undercoverage, i.e., persons who properly belong to the Swedish population but are not yet entered in the register. The information about births is almost error-free. There is also some overcoverage, i.e., persons in the TPR but not (or no longer) in the Swedish population. This overcoverage, estimated to be around 0.4% of the entire population, is made up essentially of persons who have emigrated from Sweden without notification for removal from the TPR. Since the PIN is unique, duplicates do not occur in the TPR.

□

EXAMPLE 2.2.2. *The Business Register.*

Every second week, Statistics Sweden's Business Register (BR) Programme receives information from the National Tax Board about births and deaths of enterprises. The births can be divided into three categories, namely, (i) pure births, that is, enterprises generated by new business activity; (ii) births occurring because of reorganisation, as when an existing enterprise is split into several entities; (iii) births arising because of a registration of a new legal form. SOS (1998) shows that out of the enterprises that were births in 1997, only 54% belonged to group (i). For the remaining 46%, in particular, there may be identification difficulties, which can give rise to duplicates. The BR contains two important variables for every enterprise, namely, the Standard Industrial Classification (SIC) code and the number of employees. The information for updating these variables comes from several different sources, relating to different subsets of the BR, so at any given point in time, the most recent variable values do not refer to the same point in time.

The BR is an ideal source for establishing sampling frames only a few times a year. Many changes in business structure are registered at the turn of a calendar year. Furthermore, in November/December every year, all enterprises with two or more local units are updated in a specific survey (the "PAYE survey"). All this information is processed between January and the middle of March. For this reason, March is a suitable time to establish a sampling frame. The end of May is another suitable time because fresh data from the PAYE register arrive at that time. By the middle of August, results arrive from the "BR survey", which is a large survey carried out by Statistics Sweden's Business Register unit. Its purpose is to update information on SIC code and size for all enterprises with two or more local units. Finally, November is the traditional time to establish frames for most of the annual

surveys. By this time the BR also contains new information on SIC code for units in manufacturing. The annual commodity survey is the source used for this update. Hence, new sampling frames are produced every March, May, August and November, based on the BR as it exists at those different points in time.

□

Sampling

The basic procedure for estimating a population total consists in summing weighted variable values for the elements that happened to be in the sample. Under the assumption of 100% response, this gives an unbiased estimator of the population total in question. This point estimator is called the Horvitz-Thompson (HT) estimator; see (4.2.1).

A more advanced point estimator is the generalised regression (GREG) estimator; see (4.3.1). This uses a more sophisticated weighting. GREG point estimates are computed as the sum (again for the elements in the sample) of the weighted observed values, where the weight of an observed value is the product of two sub-weights, the design weight and the *g-weight*; see (4.3.4). The latter weight is computed with the aid of the available auxiliary information. A simple form of *g*-weights familiar to many statisticians is *poststratification* weights.

We have already stated that many parameters of interest in a survey are more complex than simply one population total. They can often be expressed as a function of two or more population totals. There is a simple principle for estimating a function of totals: to replace each unknown population total by its HT estimator or by the GREG estimator. For example, to estimate the population mean for a characteristic y, we compute the estimate of the population total for y and divide it by the estimate of the population size (which is known in some surveys).

The *variance* of an estimator is the average of the square of the deviation of the estimator from its central value (its mean). This average is with respect of all possible samples that can be drawn using the given sampling design. Since each of these samples has a known probability, determined by the sampling design, we can derive the variance. It is important to note that variance is measured as "variability over all possible samples". But in practice we never draw all possible samples; we draw just one single sample. Variance, therefore, is an unknown quantity. But it is one that we

would very much like to quantify, by performing a computation based on the data we have. This is what *variance estimation* does.

When statisticians speak about *sampling error* they mean the error caused by the fact that values of a study variable are recorded only for a sample of elements, not for all elements of the population. If the whole population were indeed observed, the sampling error would be zero. This situation is exceptional. (There could be other errors, for example, measurement error and nonresponse error, but the sampling error would be zero.) Statisticians usually measure "error" by a variance. Hence, the sampling error is measured by the variance of the estimator in use, *assuming that there are no other errors*.

The variance of an estimator cannot be computed because it depends on data for the whole population. We must *estimate* the variance with the aid of the data available, namely, the observed values of the sampled elements. (When this is possible the sampling design is said to be *measurable*.) We attempt to do this so that the variance estimator is (almost) unbiased.

The estimated variance is used in confidence interval calculation. The familiar procedure for obtaining a confidence interval at (roughly) the 95% level is to compute the end points of the interval as: point estimate plus or minus 1.96 times the estimated standard deviation. The estimated standard deviation is defined as the square root of the estimated variance of the estimator.

CLAN97 is a computer software constructed at Statistics Sweden. Its most current version is described in Andersson and Nordberg (1998). It is designed to compute point and standard error estimates in sample surveys. It can be adapted to most sampling designs in current use at Statistics Sweden, and is focused on the HT and GREG estimators; see Chapter 4. This gives CLAN97 wide flexibility. It is also possible to build nonresponse adjustments into the variance calculation in CLAN97; see Section 6.5. We discuss this process in more detail in the following sections of this CBM.

Another computer software is Statistics Canada's Generalized Estimation System (GES). The theoretical underpinnings are very similar to those of CLAN97; see Estevao, Hidiroglou and Särndal (1995).

Nonresponse

Considerable resources are spent on improving data collection procedures, so as to prevent nonresponse from occurring. Nevertheless, once data collection is concluded, one has to accept some, perhaps even considerable, nonresponse. A 20% nonresponse rate is common, and in many surveys it is much higher, as Table 2.2.1 illustrates. One alarming fact is that nonresponse rates are on the increase in many surveys and many countries.

Table 2.2.1. Nonresponse rates (unweighted) in per cent for some surveys at Statistics Sweden in the year 2000.

a) *Business surveys*	Nonresponse rate
Credit Market Statistics	5.5
Quarterly Survey of Income and Expenditures for municipalities	9.0
Foreign Trade Credits	8.5
Energy Statistics for Non-Residential Premises	19.0
Business Investments	17.0
Swedish National and International Road Goods Transport	27.2
Turnover in Domestic Trade and Certain Service Activities	19.8
b) *Surveys on individuals*	
Energy Statistics for One- and Two-Dwelling Buildings	20.0
Income Distribution Survey	25.5
Labour Force Survey	15.1
Party Preference Survey	24.5
Swedish Survey of Living Conditions	22.6
Consumer Buying Expectations	32.9
Surveys of Receipts and Costs of Multi-Dwelling Buildings	21.0
Energy Statistics for Multi-Dwelling Buildings	22.0
Transition from Upper Secondary School to Higher Education	28.0

In surveys on individuals there is a vast literature illustrating the distribution of nonresponse with respect to basic variables such as age, sex, and region. Experience gathered from these *nonresponse analyses* shows that, for surveys on individuals, lower response rates are usually expected for metropolitan residents, single persons, members of childless households,

older persons, divorced or widowed persons, persons with lower educational attainment, and self-employed persons; see Holt and Elliot (1991) and Lindström (1983). Similar studies have been reported for business and establishment surveys, showing how the response rate varies between different sub-groups of the population; see Groves and Couper (1993). We return to this topic in Chapter 10.

Since variables such as age, sex and region often co-vary with many social survey study variables, the nonresponding elements are likely to be atypical with respect to these variables, leading to nonresponse bias in the estimates. Another effect of nonresponse is an increase in the variance of estimates, because the effective sample size is reduced. This can be counteracted by some degree of "oversampling", so that the sample size is fixed at the design stage at an appropriately "higher-than-normal" rate. A slight drawback may then be some increase in administrative burden, postage fees, and so on. Another problem, although relatively minor, is that if the desired sample is allocated to strata in an optimal fashion, the resulting allocation of responding elements may not be optimal.

As we have noted, a survey may contain many study variables. In some surveys, it is possible to obtain data on some of these variables from available registers, called *register variables* in what follows; these sources will almost always show data on all variables and all elements, so no data will be missing. For the other study variables, called *questionnaire variables* in what follows, one may have to rely on data collected by questionnaire or by other means, and these data will be affected by some nonresponse. It is customary to distinguish two types of nonresponse, *unit nonresponse* and *item nonresponse*. A unit nonresponse element is one for which information is missing on all the questionnaire variables. An item nonresponse element is one for which information is missing on at least one, but not all, of the questionnaire variables. The set of elements with a recorded response on at least one questionnaire item will be called the *response set.* These concepts are illustrated by the following example.

EXAMPLE 2.2.3. *Unit nonresponse, item nonresponse and response set.*

The following table illustrates the result of a (hypothetical) data collection in a survey with 8 sampled elements. The symbol x indicates a presence of data, *nr* indicates that data are missing.

Identity	Register variables		Questionnaire variables		
	1	2	1	2	3
1	*x*	*x*	*x*	*x*	*x*
2	*x*	*x*	*x*	*x*	*nr*
3	*x*	*x*	*x*	*nr*	*x*
4	*x*	*x*	*x*	*x*	*x*
5	*x*	*x*	*x*	*x*	*x*
6	*x*	*x*	*nr*	*x*	*nr*
7	*x*	*x*	*nr*	*nr*	*nr*
8	*x*	*x*	*nr*	*nr*	*nr*

Although all 8 sample elements have data for the two register variables, we shall say that elements 7 and 8 constitute the unit nonresponse, because neither of these has any response in the questionnaire part of the survey. Elements 1 to 6, which have values recorded for at least one questionnaire item, form the response set in this example.

□

3. Nonresponse adjustment

3.1. Introduction

Nonresponse adjustment is a collective term for the various attempts made by statisticians to deal with nonresponse once it has occurred, that is, after an acceptance of the fact that some desired data will be missing. As the word "adjustment" suggests, changes are made to an original or "ideal" estimation procedure, namely, the one intended for use in the ideal case of 100% response. The principal methods for nonresponse adjustment are *reweighting* and *imputation*.

Reweighting entails altering the weights of the respondents, compared to the weights that would have been used in the case of 100% response. Since observations are lost by nonresponse, reweighting will imply increased weights for all, or almost all, of the responding elements. In this CBM, reweighting is treated by a general approach, the *calibration approach*, which has the favourable property of incorporating most "standard" methods found in different places in the literature; see Section 6.6.

Imputation entails replacing missing values by proxy values. The statistician can choose to use imputation for item nonresponse only and then treat unit nonresponse by reweighting. We call this alternative the *ITIMP-approach*. The other alternative is to impute values for the item nonresponse as well as for the unit nonresponse. We call this the *UNIMP-approach*. Different imputation methods, and the "imputed estimators" that they lead to, are discussed in Section 7.2.

Remark 3.1.1. Imputation may be disallowed for legal reasons. Some countries prohibit imputation, at least for some categories of observed elements. Information on legal aspects in Sweden is given in Section 7.4.

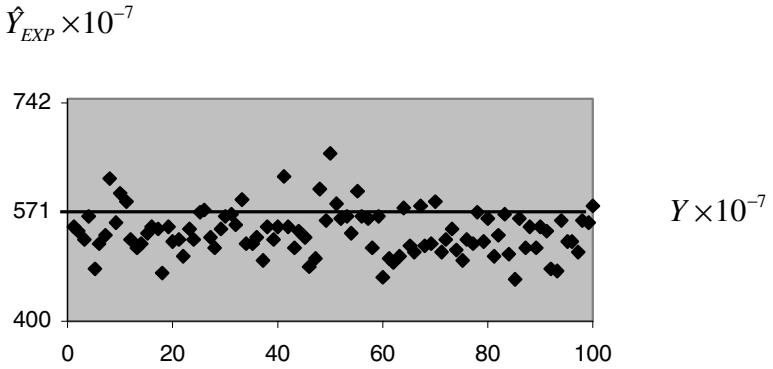
□

EXAMPLE 3.1.1. *Illustration of sampling error and nonresponse bias, assuming a perfect frame.*

In order to illustrate certain nonresponse adjustment procedures we constructed an artificial population of size N = 34,478 composed of 17,062

men and 17,416 women. The data came from the Income Distribution Survey 1999 and the study variable value y_k represents the sum of total earned and capital income of person *k*. The mean value was 196,592 for men and 135,689 for women.

Suppose we want to estimate the total $Y = \sum_U y_k$ under the following survey conditions. An SRS of size 400 is drawn, and a response mechanism then operates in such a way that all males respond with probability 0.5 and all females with probability 0.9. The response set has a size of around 281 and will tend to strongly overrepresent women, compared to the more natural male/female distribution found in the desired SRS sample. Assume that we use the expansion estimator, given by (6.6.1), to estimate the total income. This estimator formula implies essentially that we treat the response set as a simple random selection from the population, which is, of course, wrong under these circumstances. Consequently, the estimates derived with this formula will be biased. We drew 100 SRS samples, and for each of these, a response set was realised by the response mechanism mentioned. The results are given in the following figure, which illustrates both the sampling error and the nonresponse bias. The horizontal axis represents the numbering, from 1 to 100, of the response sets, and the vertical axis represents the estimate of total income, using (6.6.1). The total Y to be estimated – the target parameter value – is indicated by the horizontal straight line.



The majority (but not all) of the 100 estimates, and their mean, fall clearly below the target. The bias is very distinctly negative. This bias cannot be

estimated from a single sample, which is all we have in a real survey. The variance, which is made up of sampling variance and nonresponse variance (see Section 5.2), is illustrated by the fluctuation of the 100 estimates around their average. The nonresponse causes not only a bias but also an increase in variance. The sampling variance in itself is therefore smaller than the variance indicated by the figure.

□

3.2. The importance of auxiliary information

The key to successful nonresponse adjustment lies in the use of "strong" auxiliary information. Such use will reduce both the nonresponse bias and the variance.

Register variables play an important role in many of Statistics Sweden's surveys. They are used in creating an appropriate sampling design and/or in the computation of the survey estimates. In both uses, the register variables can be called *auxiliary variables*, because they assist and improve the procedures. Most often, as usually in this CBM, the term "auxiliary variable" refers to a variable used at the estimation stage to create better alternatives to the simplest estimators.

Register variables are frequently used to construct the strata for stratified sampling designs. Such designs aim at achieving a targeted precision for estimates made for the whole population and/or for particularly important domains (subpopulations). It is then important to designate each important domain as a separate stratum. In other surveys, particularly in business surveys, a register variable may be used as the "size variable" necessary for constructing a probability-proportional-to-size design (a pps or a πps design), in the manner discussed, for example, in Särndal, Swensson and Wretman (1992), Section 3.6.

Terms frequently used in the following are *auxiliary variable*, *auxiliary vector, auxiliary information* and *auxiliary population total(s)*. We now explain our use of these terms. The minimum requirement to qualify as an auxiliary variable is that the values of the variable are available for every *sampled* element (that is, for both responding and nonresponding elements). For many surveys at Statistics Sweden, such variable values can be found in available registers, and are then usually known not only for the sampled elements but, more extensively, for all elements in the population.

An auxiliary vector is made up of one or more auxiliary variables. There are two important steps in the process leading to the form of the auxiliary vector that will be ultimately used in the estimation. These are:

(i) Making an inventory of potential auxiliary variables;
(ii) Selecting and preparing the most suitable of these variables for entry into the auxiliary vector.

The auxiliary variables deemed potentially useful for the estimation may come from several registers allowing the possibility of linking of elements. A rather long list of potential variables may result from this scrutiny. The next important step is the procedure by which we arrive at the final form of the auxiliary vector to be used in the estimation. This process requires considerable reflection and study. The decisions to be taken include the selection of variables from the available larger set, the setting of appropriate group boundaries for converting a quantitative variable into a categorical variable, and fixing rules for collapsing very small groups into larger groups.

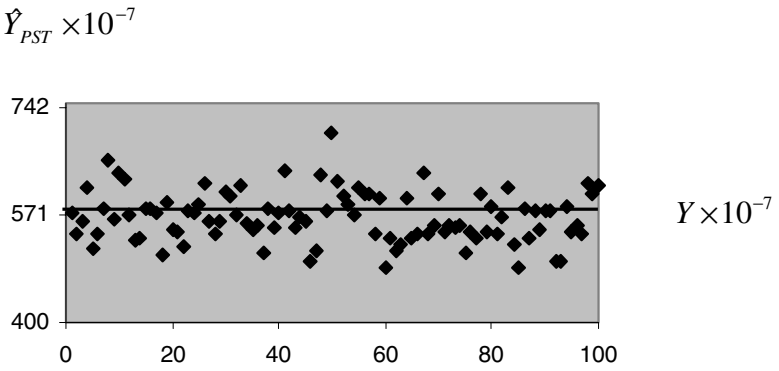
The estimator scheduled for use in the survey will usually require a known population total for each variable in the auxiliary vector. We use the term "auxiliary information" with reference both to the auxiliary vector itself, and to the known totals for the variables in the vector. Imputation can usually be carried out with auxiliary information limited to the sample elements.

Note that when register variables are used in the construction of the sampling design, their values must be known for every element in the population, as when strata are constructed for a stratified design. When auxiliary variables are used at the estimation stage, such detailed information may not be necessary. It may suffice to know the population total for each auxiliary variable, while knowledge of individual variable values may be limited to the sampled elements only.

The following simple example illustrates how nonresponse bias can be reduced by incorporating relevant auxiliary information in the estimation procedure.

EXAMPLE 3.2.1. *Reducing the nonresponse bias through the use of auxiliary information in the estimator.*

We return to Example 3.1.1, where the expansion estimator was found to have an unacceptably large bias. Its poor performance is explained in part by the absence of auxiliary information. Suppose now that we can use sex as an auxiliary variable and that the frequency of males and of females in the population is known. With this information we can instead use the poststratified estimator (6.6.4) with $P = 2$ poststrata, men and women. We computed this alternative estimator for each of the 100 response sets realised in Example 3.1.1. The results are shown in the following figure. As in Example 3.1.1, the horizontal axis represents the numbering, from 1 to 100, of the response sets, and the vertical axis represents the estimated total income. The target value Y to be estimated is indicated by the horizontal straight line.



Compared with the figure in Example 3.1.1, we see a striking improvement. By visual inspection alone, the mean of the 100 poststratified estimates is now seen to be very close to the target value. (It can be shown that the bias is actually zero.)

□

The main topic of this CBM is estimation rather than sampling design. However, let us consider one example of how nonresponse bias can be reduced by the use of auxiliary information in the sampling design.

EXAMPLE 3.2.2. *Reduction of nonresponse bias by use of auxiliary information in the sampling design.*

A very common procedure at Statistics Sweden has the following two features: (i) stratified simple random sampling, and (ii) nonresponse adjustment by straight expansion within each stratum, using the inverse of the stratum response fraction. The Swedish term for the method is "rak uppräkning inom strata". The underlying assumption is that every element within a given stratum responds with the same probability. It is a very convenient procedure in routine statistics production; however, there is usually little or no attempt to verify if the assumption is satisfied or nearly satisfied.

Suppose we use this procedure in sampling from the population discussed in Example 3.1.1 and Example 3.2.1: Let there be two strata, men and women, and nonresponse adjustment by straight expansion in each stratum. We then obtain the same reduction of the nonresponse bias as in Example 3.2.1, illustrating that the use of auxiliary information used in the sampling design can also serve the purpose of reducing nonresponse bias.

As mentioned previously, important domains of interest are often designated as strata for stratified sampling. This permits allocating the total available sample resources in such a way that every important domain is sufficiently represented to realise a desired precision. As a result, one may decide to overrepresent smallish domains, compared to larger ones. For instance, if geographic areas are important domains for the population, then the strata should be based on these areas.

It should be noted that although the procedure defined by (i) and (ii) is very common at Statistics Sweden, it amounts to a very restricted use of auxiliary information. If applied in a mechanical fashion, the procedure is oblivious to the more effective options that could be realised after an inventory and constructive use of other auxiliary information.

□

Almost all techniques described in this CBM require some form of auxiliary information. The availability of strong auxiliary information is particularly important in treating nonresponse, because such information lends strength to reweighting and imputation procedures, thereby reducing several errors: sampling error, nonresponse error and coverage error.

EXAMPLE 3.2.3. *Examples of reduced nonresponse bias as a result of identifying new powerful auxiliary variables.*

That a search for more powerful auxiliary information can significantly improve the estimates is illustrated by recent developments in the Labour Force Survey (LFS) in Finland and in Sweden. The old survey design, in both countries, involved the use of the poststratified estimator, with a poststratification carried out on two or all of the dimensions age group, sex and region. In both countries it was found that the inclusion of a new, dichotomous register variable into the auxiliary vector improved the estimation considerably. In its simplest form, this dichotomous variable has the value "1" for a person registered in the country's Job Seekers' register, and "0" otherwise. For the estimation, the register is matched with the LFS sample with the aid of the unique PIN. The number of unemployed, most likely underestimated under the old design, was estimated to be significantly higher after inclusion of the job seeker variable into the auxiliary vector, as revealed by a comparison of the old estimation method with the new. It is highly likely that the change in the level of the estimates corresponds to a considerable reduction of the nonresponse bias. These developments are reported in Djerf (1997), Djerf (2000), Hörngren (1992).

□

In this section we mention the main principles for selecting auxiliary variables when the calibration approach to reweighting is used. These principles are illustrated by Example 3.2.4, set in the context of an actual survey at Statistics Sweden. The procedure for selecting such information is discussed in more detail in Chapter 10.

To reduce the nonresponse bias and the variance of the calibration estimator, one should select an auxiliary vector that satisfies as far as possible one or both of the following principles:

(i) explains the variation of the response probabilities

(ii) explains the variation of the main study variables.

A third principle to take into account is that the auxiliary vector should

(iii) identify the most important domains.

When principle (i) is fulfilled the nonresponse bias is reduced in the estimates for all study variables. However, if only principle (ii) is fulfilled the nonresponse bias is reduced only in the estimates for the main study variables. Then the variance of these estimates will also be reduced. When principle (iii) is fulfilled the effect is mainly a reduction of the variance for the domain estimates. Example 3.2.4 below illustrates how one may reason in a practical situation to fulfil the principles.

For Statistics Sweden's surveys on individuals, several available registers provide a rich source of auxiliary information. We use the Survey on Life and Health to illustrate the extent of the potential auxiliary information and the steps taken in developing a suitable auxiliary vector for the calibration approach.

EXAMPLE 3.2.4. *The Survey on Life and Health* (*Liv och Hälsa*).

The population consists of persons aged 18-79 in the county of Södermanland. As the name suggests, the study variables concern different conditions of life and health. Some of these variables are considered more important than others and can be designated as “main study variables”, as we discuss later. The frame population, as determined by the TPR (see Example 2.2.1), was stratified by municipality. The total sample was allocated to the strata so as to meet specified precision requirements for each municipality. Within each stratum an SRS was drawn. At the end of the data collection stage, the nonresponse rate was found to be 34.4%. This very high rate would probably have caused a substantial nonresponse bias if no nonresponse adjustment had been attempted. Fortunately, as we now explain, quite strong auxiliary information was available. Therefore, a significant reduction of the bias is a likely result of the calibration approach to reweighting.

Two sources of auxiliary information were used: the TPR and the Register of Education. As a result of an inventory, six prospective auxiliary variables, all of them categorical, were retained: Sex (male; female), Age group (4 classes), Country of birth (the Nordic countries; other), Income group (3 classes), Civil status (married; other) and Education level (3 classes).

Two different analyses were carried out. The objective was to see which, if any, of the six variables were particularly strong (a) in explaining the

variation of the response probabilities, and (b) in explaining the variation of the main study variables.

The first analysis is a typical *nonresponse analysis*, consisting in a computation of the response rates in the different classes for each of the six variables. The results are given in Tables 3.2.1-3.2.6.

Table 3.2.1. Response rate (%) by Sex.

Sex	Male	Female
Response rate (%)	60.1	71.2

Table 3.2.2. Response rate (%) by Age group.

Age group	18-34	35-49	50-64	65-79
Response rate (%)	54.9	61.0	72.5	78.2

Table 3.2.3. Response rate (%) by Country of birth.

Country of birth	Nordic countries	Other
Response rate (%)	66.7	50.8

Table 3.2.4. Response rate (%) by Income group.

Income class (in thousands of SEK)	0-149	150-299	300-
Response rate (%)	60.8	70.0	70.2

Table 3.2.5. Response rate (%) by Civil status.

Civil status	Married	Other
Response rate (%)	72.7	58.7

Table 3.2.6. Response rate (%) by Educational level.

Educational level	Level 1	Level 2	Level 3
Response rate (%)	63.7	65.4	75.6

The response rates differ considerably for different categories of a variable. Thus, we expect that all six prospective auxiliary variables may be important for explaining the variation of response probabilities. The response rates are very similar in the last two income groups and the first two educational groups, suggesting that they should perhaps be collapsed. However, maintaining all the groups may contribute to satisfying principle (ii), so no collapsing is undertaken at this stage.

The client was asked to identify the most important study variables. Several variables were mentioned, some of them dichotomous. (A dichotomous variable is one whose value is 1 for a person who has the attribute and 0 otherwise.) Four of the identified dichotomous variables were: (a) poor health, (b) avoidance of walking outdoors after dark for fear of attack, (c) housing problems, (d) poor personal finances. For each of these an analysis was carried out to see how well principle (ii) was satisfied by the prospective auxiliary variables.

The analysis relied on the estimates given in Tables 3.2.7 to 3.2.12. The estimates were derived by the method "nonresponse adjustment by straight expansion within each stratum" (in Swedish "rak uppräkning inom strata").

Table 3.2.7. Estimated proportion (%) of individuals with property (a)-(d), by Sex.

Property	Male	Female
(a)	7.5	8.9
(b)	7.8	21.1
(c)	2.6	2.4
(d)	19.6	19.8

Table 3.2.8. Estimated proportion (%) of individuals with property (a)-(d), by Age group.

Property	18-34	35-49	50-64	65-79
(a)	4.3	6.6	10.6	10.9
(b)	11.8	11.4	14.3	23.4
(c)	5.9	2.8	1.0	0.8
(d)	31.0	26.6	12.5	9.6

Table 3.2.9. Estimated proportion (%) of individuals with property (a)-(d), by Country of birth.

Property	Nordic countries	Other
(a)	8.0	11.7
(b)	14.7	18.3
(c)	2.4	4.2
(d)	19.2	28.5

Table 3.2.10. Estimated proportion (%) of individuals with property (a)-(d), by Income group (in thousands of SEK).

Property	0-149	35-49	300-
(a)	10.0	7.2	4.0
(b)	18.6	12.6	8.1
(c)	3.8	1.5	1.0
(d)	25.3	16.5	6.9

Table 3.2.11. Estimated proportion (%) of individuals with property (a)-(d), by Civil status.

Property	Married	Other
(a)	8.2	8.2
(b)	13.8	16.3
(c)	1.1	4.3
(d)	14.1	26.5

Table 3.2.12. Estimated proportion (%) of individuals with property (a)-(d), by Educational level.

Property	Level 1	Level 2	Level 3
(a)	10.5	7.3	4.6
(b)	19.1	12.6	12.9
(c)	1.7	3.2	1.8
(d)	17.5	21.6	16.8

Tables 3.2.7 to 3.2.12 show that the prospective auxiliary variables explain the study variables to different extents. Most appear to be strong explanatory variables, although Sex and Civil status seem weaker, at least for some of the four study variables.

The prospective auxiliary variables are to some extent intercorrelated, so if all of them were entered together as input into the calibration, some of the information would be redundant. For some responding persons, this could lead to a few weights that are abnormally high or too low, even negative. An increase in variance may be an undesirable consequence of this. Groups that are too small may also give this effect. Therefore, in any application of calibration, it is recommended to analyse the distribution of the weights.

In this survey the decision was finally made to use an auxiliary vector composed of five categorical variables: Municipality, Sex, Age group, Country of birth and Educational level. Out of these, Municipality is also used as a stratification variable. This does not prevent its inclusion into the auxiliary vector; in fact, in order to ensure consistency as defined by (6.3.4), it must be included.

The principal domains of interest in this survey are determined by the cross-classification Municipality by Sex by Age group, represented by the expression *municipality*sex*age*. In order to satisfy principle (iii), these three variables should be present in the auxiliary vector. If no other auxiliary information were used, the calibration estimator would take the form of a poststratified estimator with $M \times 2 \times 4$ poststrata, where M is the number of municipalities. However, the decision was taken to also include Country of birth and Educational level. The most detailed use of the information would then be to completely cross-classify all five variables. However, a consequence might be that some of the five-dimensional cells contain

extremely few observations, or that they are completely empty. This might cause an increase in variance. Therefore, Country of birth and Educational level were treated as separate variables. The final auxiliary vector can then be represented as *municipality*sex*age +country of birth+educational level.* Its dimension is $(M \times 2 \times 4) + 2 + 3$. The calibrated weights resulting from this formulation of the auxiliary vector have the following properties: When applied to the auxiliary vector, they will give "exact estimates" for: (a) the known population counts for the cells determined by Municipality by Sex by Age group, (b) the known marginal counts in the population for Country of birth; (c) the known marginal counts in the population for Educational level. In this survey, the calibration estimator and the corresponding variance estimator were calculated, for all domains of interest, by CLAN97.

□

Essentially all estimation methods reviewed in this CBM revolve around different uses of auxiliary information. Reweighting by the calibration approach is discussed in Chapter 6. The technique is very general, and most of the "conventional" methods are covered as special cases. A number of imputation techniques are discussed in Chapter 7. These methods differ in their auxiliary information requirements. Chapter 10 examines the question of how to select the "best" auxiliary information from an available larger pool of information.

4. Estimation under ideal conditions

4.1. Introduction

Both nonresponse and frame imperfections are normal features of any survey. But they are undesirable, because without them the quality of the statistics (the accuracy of the estimates) would generally be better. Neither of the two can be completely treated at the *design stage* of the survey, so we need a procedure for dealing with these nuisance factors at the *estimation stage*.

It is easier to develop the principles for estimation under the assumption that the two nuisance factors are absent. This is what we do in this chapter. Then in Chapters 5 to 10 we discuss the modifications of the estimators made necessary by nonresponse and in Chapter 11 we extend the discussion to also deal with frame coverage problems.

To fix ideas we introduce some notation. Consider the finite population of N elements $U = \{1,...,k,...,N\}$, called the *target population*. We wish to estimate the total

$$Y = \sum_U y_k \qquad (4.1.1)$$

where y_k is the value of the study variable, y, for the kth element.

We assume that s is a probability sample of size n, drawn from the target population U (see Figure 4.1.1) with the probability $p(s)$. The inclusion probabilities, known for all $k \in U$, are then $\pi_k = \sum_{s \ni k} p(s)$. We assume that the design is such that $\pi_k > 0$ for all elements k. Let $d_k = 1/\pi_k$ denote the *design weight* of element k. The design weights are very important for computing point estimators.

For computation of variance estimates we also need to consider second order inclusion probabilities. A typical probability of this kind is denoted π_{kl} and it represents the known probability that both k and l are included in

the sample, that is, $\pi_{kl} = \sum_{s \supset \{k,l\}} p(s)$. The corresponding weight is denoted $d_{kl} = 1/\pi_{kl}$. These are defined for all $k \in U$ and $l \in U$. Note that if $k = l$, then $\pi_{kk} = \pi_k$ and $d_{kk} = d_k$.

Commonly used designs are *simple random sampling* (SRS) and *stratified simple random sampling* (STSRS). In rarer cases at Statistics Sweden, *probability-proportional-to-size* (pps or πps) sampling is used. The weights d_k and d_{kl} depend on the sampling design in use. The following example illustrates the form of the weights for one particular design, namely, STSRS.

EXAMPLE 4.1.1. *Weights under the STSRS design.*

Consider a STSRS design with H strata and such that n_h elements are selected from N_h in stratum h, $h = 1,...,H$. Then the design weights needed for the point estimation are $d_k = N_h / n_h$ for all k in stratum h. The weights d_{kl} needed for the variance estimation are of three types: $d_{kl} = d_{kk} = d_k = N_h / n_h$ if $k = l$ is in stratum h, $d_{kl} = \frac{N_h}{n_h} \frac{(N_h - 1)}{(n_h - 1)}$ if $k \neq l$ and both k and l are in stratum h and $d_{kl} = \frac{N_h}{n_h} \frac{N_{h'}}{n_{h'}}$ if k and l are in different strata, h and h'.

□

Both the sampling design and the estimators are usually constructed with the aid of auxiliary information about the elements k = 1, ..., N. The information used for the sampling design is usually different from that used for the estimators, but nothing prevents the same information being used repeatedly, at both stages.

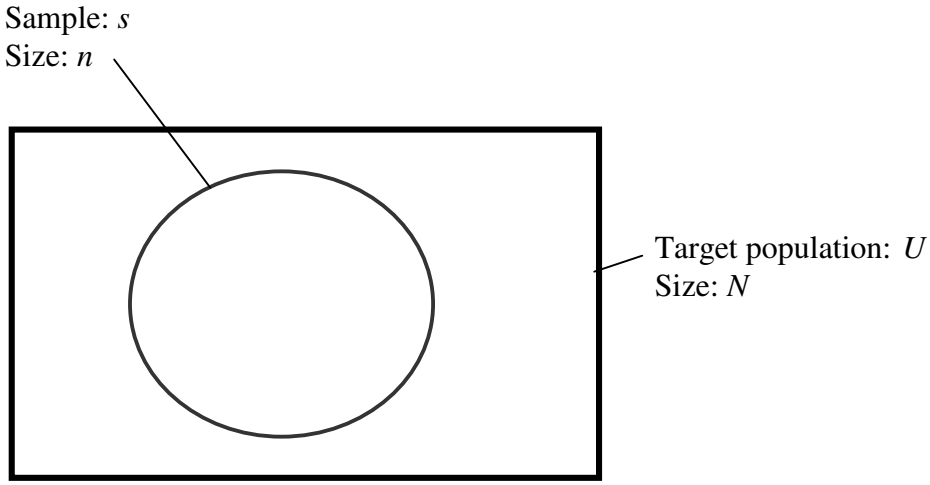


Figure 4.1.1.

The computational load is increased by the fact that most surveys require estimation not only for the whole population but for a perhaps considerable number of subpopulations as well. They are called *domains of study* or *domains of interest* or simply *domains*. A domain of interest can be any subpopulation. Some domains may be very small in the sense that very few observed y-values fall into it. The precision of any estimate made for such a domain will be questionable.

A special case arises when the domains form a set of mutually exclusive and exhaustive subpopulations. The domains are then said to form a *partition* of the population U. For example, a set of domains for a population of individuals may based on a cross-classification of sex (male, female) with 8 age groups covering all ages. Then the resulting 16 domains form a partition of the whole population of individuals.

We denote the domains of interest by $U_1,...,U_d,...,U_D$. Suppose that we want to estimate the total of the variable y for each domain separately. The targets of estimation are then the D quantities $Y_1,...,Y_d,...,Y_D$, where $Y_d = \sum_{U_d} y_k$, $d = 1,...,D$.

We can also express the domain total $Y_d = \sum_{U_d} y_k$ with the aid of a "new" study variable, y_d, derived as a transformation of the original y-variable, but specific for the domain U_D. We denote this new variable y_d, and its value for element k is defined by

$$y_{dk} = \begin{cases} y_k & \text{for} \quad k \in U_d \\ 0 & \text{for} \quad k \notin U_d \end{cases} \tag{4.1.2}$$

Then Y_d can be expressed as the total over the entire population total of the new variable y_d, that is,

$$Y_d = \sum_U y_{dk} \tag{4.1.3}$$

To derive the domain variable y_d for the sample elements, we must be able to observe the domain membership for every sample element. (However, we do not usually know the domain membership for every population element. An exception is when domain membership is indicated in the population frame from which we are sampling.)

As mentioned in Section 2.1 we may be interested in estimating different types of parameters, such as a population total, a ratio of population totals, a population mean. More complex parameters, such as a regression coefficient or a correlation coefficient, are sometimes of interest.

In general such parameters can be expressed as a function of totals, that is, the parameter has the form $\psi = f(Y_1,...,Y_q,...,Y_Q)$, where $Y_1,\ldots,Y_q,\ldots,Y_Q$ are the totals involved and f is a given function. The principle for estimating ψ used in this CBM is that each total is replaced by its estimate, that is, $\hat{\psi} = f(\hat{Y}_1,...,\hat{Y}_q,...,\hat{Y}_Q)$. Thus, when we know how to estimate a population total, estimating other types of parameters is straightforward. For certain types of functions *f*, CLAN97 can be used for calculating the point estimates $\hat{\psi}$ and their variance estimates. We return to this topic in Section 6.5.

4.2. The Horvitz-Thompson estimator

Let us take an example to show how information may be used to construct a sampling design. The construction of an STSRS design begins with the assignment of the frame elements to a set of well-defined strata, for example, a set of age/sex groups, if the elements are individuals. We must thus have information about age and sex allowing every element in the frame to be assigned to one and only one of the strata.

Building a probability-proportional-to-size design requires information in the form of a positive size measure, z_k, known for every element k in the population. For example, z_k can be the number of employees of an enterprise, if we are dealing with a business survey. The design is constructed so that the inclusion probability for element k is proportional to the known value z_k.

When the sampling design has been fixed, the inclusion probabilities π_k and the sampling design weights $d_k = 1/\pi_k$ are fixed, known quantities. We can then construct an unbiased estimator of Y, namely, the *Horvitz-Thompson estimator* (HT estimator). It is given by

$$\hat{Y}_{HT} = \sum_s d_k y_k \qquad (4.2.1)$$

This estimator is unbiased for Y, under any sampling design satisfying $\pi_k > 0$ for all elements k. Note that once the sampling design has been fixed, the variance and other statistical properties of $\hat{Y}_{HT}$ are also fixed. In other words, after sampling and data collection, we cannot change the variance of $\hat{Y}_{HT}$; it is determined entirely by the choice of sampling design. Consequently, if the plan is to use the HT estimator, the sampling design should be chosen so as to obtain a small variance for this estimator. The sampling design must of course also be practical in other respects.

4.3. The generalised regression estimator

A wider and more efficient class of estimators are those that use auxiliary information explicitly at the estimation stage. Some information may already have been used at the design stage. Denote the auxiliary vector by $\mathbf{x}$,

and its value for element k by $\mathbf{x}_k = (x_{1k},...,x_{jk},...,x_{Jk})'$, a column vector with J components, where x_{jk} is the value, for element k, of the j:th auxiliary variable. We assume that the population total, $\sum_U \mathbf{x}_k$, is accurately known.

An estimator that uses this information is the *generalised regression estimator* (GREG estimator). This estimator is explained and illustrated by several examples in Särndal, Swensson and Wretman (1992), Chapters 6 and 7. It is given by

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + (\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)'\hat{\mathbf{B}} \qquad (4.3.1)$$

where

$$\hat{\mathbf{B}} = (\sum_s d_k c_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k c_k \mathbf{x}_k y_k) \qquad (4.3.2)$$

is a vector of regression coefficients, obtained by fitting the regression of y on $\mathbf{x}$, using the data $(y_k, \mathbf{x}_k)$ for the elements $k \in s$. The data are weighted by $d_k c_k$, where the factor c_k is specified by the statistician (Section 4.5 gives some examples). A simple choice is to take $c_k = 1$ for all k.

The GREG estimator is "almost unbiased". The bias, although not exactly zero, tends to zero with increasing sample size, and even for modest sample sizes it is normally so small that we do not need to consider it.

The term $(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)'\hat{\mathbf{B}}$ in the formula for $\hat{Y}_{GREG}$ can be viewed as a *regression adjustment* applied to the HT estimator, $\hat{Y}_{HT} = \sum_s d_k y_k$. The effect is an important reduction of the variance of $\hat{Y}_{HT}$, especially when there is a strong regression relationship between y and $\mathbf{x}$.

Although we call $\hat{Y}_{GREG}$ the GREG estimator – in singular form – it is in reality a whole set of estimators, corresponding to the different specifications that we can give to the auxiliary vector $\mathbf{x}_k$ and to the factor c_k. If a number of auxiliary variables, or x-variables, each with a known

population total, are available at the estimation stage, we may include in $\mathbf{x}_k$ those x-variables that promise to be the most efficient ones for reducing the variance. That is, we select some or all of the available x-variables for inclusion in the auxiliary vector $\mathbf{x}_k$. Consequently, the vector $\mathbf{x}_k$ to be used in $\hat{Y}_{GREG}$ can take a variety of forms, given that we have at our disposal a certain quantity of auxiliary information. This is illustrated by several examples in Section 4.5.

Note that we can wait until after sampling and data collection to specify which of the possible GREG estimators we are going to use, because the decision on the x-variables to include in $\mathbf{x}_k$ need not be made until after these survey operations have been completed.

The presentation in this CBM is facilitated by expressing the estimator $\hat{Y}_{GREG}$ as a linearly weighted sum of the observed values y_k. When we do this, we get

$$\hat{Y}_{GREG} = \sum_s d_k g_k y_k \qquad (4.3.3)$$

where the total weight given to the value y_k is the product of two weights, the design weight $d_k = 1/\pi_k$, and the weight, g_k, which depends both on the element k and on the whole sample s of which k is a member. It is given by

$$g_k = 1 + c_k \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)' \left(\sum_s d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \qquad (4.3.4)$$

The value of g_k is near unity for a majority of the elements $k \in s$, and the greater the size of the sample s, the stronger is the tendency for the g_k to hover close to unity. It is rare to find elements with a weight g_k that is greater than 4 or less than 0. Negative weights are allowed; such weights do not invalidate the theory, but some users would like all weights to be positive. In Section 6.5 we discuss methods to ensure that all weights are strictly positive.

As is easily verified, the HT estimator is a special case of $\hat{Y}_{GREG}$, obtained when (i) $\mathbf{x}_k = c_k = 1$ for all $k \in s$, and (ii) the design satisfies $\sum_s d_k = N$. The condition (ii) holds, for example, for the SRS and STSRS designs.

When we apply the weight system $d_k g_k$ to the auxiliary vector $\mathbf{x}_k$, and sum over the elements $k \in s$, we obtain an estimate of the population total of $\mathbf{x}_k$. This estimate turns out to be exactly equal to the known value of that total, that is, we have

$$\sum_s d_k g_k \mathbf{x}_k = \sum_U \mathbf{x}_k \qquad (4.3.5)$$

This makes good sense, because the weight system would not seem reasonable if it led us to estimate the total of $\mathbf{x}_k$ by anything other than the known value that we have for this total. The weight system is called *calibrated* or, sometimes, *consistent*. More specifically, it is calibrated to the known population total $\sum_U \mathbf{x}_k$.

Estimation for a domain is straightforward. We calculate the GREG estimator of $Y_d = \sum_U y_{dk}$ recognizing that the study variables the new variable y_d (rather than y itself), given by (4.1.2). This means that the weights are kept the same as when y is the variable of interest. In other words, using (4.3.3) and (4.3.4), we have the point estimator

$$\hat{Y}_{dGREG} = \sum_s d_k g_k y_{dk} \qquad (4.3.6)$$

4.4. Variance and variance estimation

With every estimator $\hat{Y}$ is associated an unknown variance (over repeated samples). The variance, $V(\hat{Y})$, always remains an unknown quantity, a function of the whole population, which we do not observe in totality. Nevertheless, the variance is a theoretical quantity of considerable interest. An important survey objective is to estimate the variance $V(\hat{Y})$. The usual procedure is to start with the formula for the variance, and to transform it into an estimated variance. Once computed from the sample data, the estimated variance, denoted $\hat{V}(\hat{Y})$, opens up the possibility of assessing the

precision of $\hat{Y}$. We can for example use $\sqrt{\hat{V}(\hat{Y})}$, and the point estimate $\hat{Y}$, to compute a confidence interval for the unknown parameter Y.

Here and in the following sections we assume that an approximate 95% confidence interval is computed according to the formula

point estimate $\pm$ 1.96 (variance estimate)$^{1/2}$

To close approximation, the expression for the variance of $\hat{Y}_{GREG}$ is

$$V(\hat{Y}_{GREG}) = \sum\sum_U \left(\frac{d_k d_l}{d_{kl}} - 1 \right) E_k E_l \qquad (4.4.1)$$

where the residuals are those arising from the "population regression fit". (For any set of elements A, $A \subseteq U$, we write for simplicity the double sum $\sum_{k \in A} \sum_{l \in A}$ as $\sum\sum_A$.) This fit, which cannot be carried out in practice, has the residuals

$$E_k = y_k - \mathbf{x}_k' \mathbf{B}$$

where

$$\mathbf{B} = (\sum_U c_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U c_k \mathbf{x}_k y_k) \qquad (4.4.2)$$

Here we focus on the estimated variance of the GREG estimator $\hat{Y}_{GREG}$. The computational procedure is relatively simple and can be carried out, for some important sampling designs, by the software CLAN97. Not surprisingly, $\hat{V}(\hat{Y}_{GREG})$ is a function of the regression residuals arising from the regression of y_k on the auxiliary vector $\mathbf{x}_k$, and is such that the smaller these residuals, the smaller the estimated variance of estimator $\hat{Y}_{GREG}$, which makes good intuitive sense. We have

$$\hat{V}(\hat{Y}_{GREG}) = \sum\sum_s (d_k d_l - d_{kl})(g_k e_k)(g_l e_l) \qquad (4.4.3)$$

where $e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$, with $\hat{\mathbf{B}}$ determined by (4.3.2), and g_k is given by (4.3.4). This formula requires that all first and second order inclusion probabilities be strictly positive. For a derivation of $\hat{V}(\hat{Y}_{GREG})$, see, for example, Särndal, Swensson and Wretman (1992). Because of the double sum, there are $n(n - 1)/2$ different terms, a very large number for most surveys. For example, there are around 0.5×10^6 terms when the sample size is $n = 1000$. To compute them all would be very tedious. In practice, computation therefore proceeds via the usually much simpler expression that $\hat{V}(\hat{Y}_{GREG})$ reduces to after an algebraic manipulation taking into account the particular form taken by the weights $d_k = 1/\pi_k$ and $d_{kl} = 1/\pi_{kl}$ under the sampling design used in the survey.

Remark 4.4.1. There are some examples in the literature showing that the variance estimator (4.4.1) of the general regression estimator suffers from a negative bias when the sample is small. A variance estimator better suited for such a situation is $\hat{V}_{adj}(\hat{Y}_{GREG})$, which has the same form as $\hat{V}(\hat{Y}_{GREG})$, given by (4.4.3), but where e_k is replaced by an adjusted residual, namely $e_{adj,k} = f_k e_k$. In principle, f_k, adjusts for the number of degrees of freedom lost when further parameters are estimated. Lundström (1997), Section 2.3.1, explains this procedure in more detail.

□

EXAMPLE 4.4.1. *Variance and variance estimator under STSRS.*

Example 4.1.1 gives the weights d_k under the STSRS design. An algebraic manipulation of (4.4.1), using these weights, gives the factor $N_h^2(1-f_h)/n_h$ characteristic of stratified sampling variances, and we obtain

$$V(\hat{Y}_{GREG}) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{EU_h}^2$$

where $S_{EU_h}^2$ is the stratum variance of the residuals E_k.

The variance estimator (4.4.3) becomes

$$\hat{V}(\hat{Y}_{GREG}) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} \frac{\sum_{s_h}(g_k e_k - \frac{1}{n_h}\sum_{s_h} g_k e_k)^2}{n_h - 1}$$

□

Finally, we need to discuss the variance (and its estimation) for the domain estimator $\hat{Y}_{dGREG}$ given by (4.3.6). The expressions follow by a simple modification of (4.4.1), where we simply replace y_k by y_{dk}, given by (4.1.2). This implies that the residual e_k in (4.4.3) is replaced by the new residual

$$e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_d \qquad (4.4.4)$$

where $\hat{\mathbf{B}}_d$ is given by (4.3.2) if we replace y_k by y_{dk}. The theoretical justification for the procedure is that the GREG estimator can be applied to any study variable. If the GREG estimator works for the study variable y, it will also work for the new variable y_d, even though it is "unusual" in that many of its values y_{dk} may be zero.

It is important to identify auxiliary information that comes as close as possible to identifying the domains. Otherwise the residuals (4.4.4) can be relatively large and the effect of the auxiliary information might be slight, especially for small domains. We pursue this issue further in Example 4.5.3.

4.5. Examples of the generalised regression estimator

Consider, as a simple first example, a survey of individuals for which it is possible to use sex as an auxiliary variable; see also Example 3.2.1.

EXAMPLE 4.5.1. *One-way classification.*

For a population of individuals, we assume that the number of males and females, N_1 and N_2 respectively, are known. In this case, the vector $\mathbf{x}_k$ has only two possible values, namely, $\mathbf{x}_k = (1,0)'$ for all males, and $\mathbf{x}_k = (0,1)'$ for all females. The population total of the $\mathbf{x}_k$ is thus $(N_1, N_2)'$ which is known. A derivation of the *g*-weights, given by (4.3.4),

shows that $g_k = \dfrac{N_1}{\sum_{s_1} d_k}$ when k is male, where s_1 denotes the male part of the whole sample s. Analogously, we get $g_k = \dfrac{N_2}{\sum_{s_2} d_k}$ when k is female, where s_2 is the female part of s. As is easily verified, the weights $d_k g_k$ satisfy the calibration property (4.3.5). The GREG estimator for this simple case of auxiliary information is therefore $\hat{Y}_{GREG} = N_1 \bar{y}_{s_1} + N_2 \bar{y}_{s_2}$ with $\bar{y}_{s_j} = \sum_{s_j} d_k y_k / \sum_{s_j} d_k$ for j = 1,2. The GREG estimator resulting from this structure of the $\mathbf{x}_k$-vector is called a *poststratified estimator* (with two poststrata).

□

The generalisation to an arbitrary number of categories or poststrata is obvious. The categories may be defined by a cross-classification, as in the following example.

EXAMPLE 4.5.2. *Two-way classification.*

Consider a register listing a population of individuals distributed according to sex and three different regions. Then all population counts in the following table can be derived.

Sex		Region 1	2	3	Total
Male	1	N_{11}	N_{12}	N_{13}	$N_{1.}$
Female	2	N_{21}	N_{22}	N_{23}	$N_{2.}$
Total		$N_{.1}$	$N_{.2}$	$N_{.3}$	$N_{..}$

Here the most detailed auxiliary information consists of the six cell counts $N_{11},\ldots,N_{23}$. The $\mathbf{x}_k$-vector that expresses this information consists of six components, one of which is "1" while the others are zeros. For example, for every population element in cell (1,2), males in Region 2, the vector has the value $\mathbf{x}_k = (0,1,0,0,0,0)'$. The sum of these vectors over all population elements is the known vector $(N_{11}, N_{12}, N_{13}, N_{21}, N_{22}, N_{23})'$. Using this

auxiliary information, the resulting GREG estimator $\hat{Y}_{GREG}$ is a poststratified estimator, though now with six terms, corresponding to six poststrata.

There are situations where complete cross-classification of the variables is impractical or inconvenient, for instance, when (i) the variables come from different registers, or (ii) some cell counts are small. In the first situation the use of cell counts may be costly, since the registers have to be matched. In the second situation small cell counts may make the estimator unstable. This can sometimes be avoided by collapsing cells; see Chapter 10. However, an alternative is to use only information defined by the marginal counts. The auxiliary vector for this case is of dimension five and is such that the first two positions indicate sex and the final three positions indicate region. For example, the auxiliary vector for each population individual in cell (1,2) has the form $\mathbf{x}_k = (\underbrace{1,0}_{sex},\underbrace{0,1,0}_{region})'$. The required population sum of all these $\mathbf{x}_k$-vectors is $\left(N_{1.}, N_{2.}, N_{.1}, N_{.2}, N_{.3}\right)'$ which is known. We return to the *Two-way classification* in Section 6.6.

□

A guiding principle in estimation for domains is to use an auxiliary vector that as closely as possible identifies the domains. This will reduce the absolute size of the residuals, which in turn results in a lower variance. The following example illustrates this.

EXAMPLE 4.5.3. *Domain estimation.*

For a population of individuals assume that we want separate estimates for males and females. They define two domains of the population. Further, assume that the sampling design is SRS and that we know the number of males and females in the population. We have decided to use the GREG estimator for the domain total Y_d, $d = 1, 2$, and we choose between two alternative formulations of the auxiliary vector:

Alternative (i): the GREG estimator based on the simplest possible specification, that is, $\mathbf{x}_k = c_k = 1$ for all elements.

Alternative (ii): the GREG estimator where $\mathbf{x}_k = (1,0)'$ for all males, and $\mathbf{x}_k = (0,1)'$ for all females and $c_k = 1$ for all k; see Example 4.5.1.

We find, using the procedure stated at the end of the previous section, that the variance is

$$V(\hat{Y}_{dGREG}) = N^2 \frac{1-n/N}{n} \frac{1}{N-1} \sum_U E_{dk}^2$$

where the only difference between the two alternatives lies in the residuals E_{dk}.

In Alternative (i), the residuals are

$$E_{dk} = \begin{cases} y_k - Y_d / N & \text{for} \quad k \in U_d \\ -Y_d / N & \text{for} \quad k \in U - U_d \end{cases}$$

and in Alternative (ii), the residuals are

$$E_{dk} = \begin{cases} y_k - Y_d / N_d & \text{for} \quad k \in U_d \\ 0 & \text{for} \quad k \in U - U_d \end{cases}$$

It is easily seen that $\sum_U E_{dk}^2$ (and therefore the variance) is considerably greater for Alternative (i) than for Alternative (ii).

In Alternative (ii), the situation for domain estimation is highly favourable in that the auxiliary vector coincides exactly with the domain indicator. The reduction of the variance will be significant, compared to Alternative (i).

□

The examples discussed so far in this section involve categorical auxiliary information. Let us see what can be accomplished if there is also a quantitative auxiliary variable. This variable, x_k, can be used alone or in different combinations with the categorical variables. Some possibilities are studied in Example 4.5.4.

EXAMPLE 4.5.4. *A one-way classification combined with a quantitative variable.*

Assume that the frame specifies sex and region as in Example 4.5.2 and also the value x_k of a quantitative auxiliary variable, such as income. Let us construct some auxiliary vectors that lie within the limits set by this auxiliary information. The population cells are denoted $U_{11},...,U_{23}$ and the regions are $U_{.1}, U_{.2}$ and $U_{.3}$.

Case	Auxiliary vector $\mathbf{x}_k$	Auxiliary population total $\sum_U \mathbf{x}_k$
i	x_k	$\sum_U x_k$
ii	$(1, x_k)'$	$(N, \sum_U x_k)'$
iii	$\left(0, x_k, 0, 0, 0, 0\right)'$	$\left(\sum_{U_{11}} x_k, ..., \sum_{U_{23}} x_k\right)'$
iv	$(\underbrace{0,1,0,0,0,0}_{counts}, \underbrace{0, x_k, 0,0,0,0}_{x-variable})'$	$\left(N_{11},...,N_{23}, \sum_{U_{11}} x_k, ..., \sum_{U_{23}} x_k\right)'$
v	$(\underbrace{1,0}_{sex}, \underbrace{0, x_k, 0}_{region})'$	$\left(N_{1.}, N_{2.}, \sum_{U_{.1}} x_k, \sum_{U_{.2}} x_k, \sum_{U_{.3}} x_k\right)'$

Some well-known estimators arise from these five cases. Let us consider two of them, for the SRS sampling design. When $\mathbf{x}_k = x_k$ and $c_k = 1/x_k$, the *ratio estimator* is obtained from the general formula (4.3.3), that is,

$$\hat{Y}_{GREG} = \sum_U x_k \frac{\bar{y}_s}{\bar{x}_s} \qquad (4.5.1)$$

where $\bar{y}_s = \frac{1}{n}\sum_s y_k$ and $\bar{x}_s = \frac{1}{n}\sum_s x_k$.

When $\mathbf{x}_k = (1, x_k)'$ and $c_k = 1$ for all k, the *regression estimator* is obtained, that is,

$$\hat{Y}_{GREG} = N\left\{\bar{y}_s + \left(\bar{X} - \bar{x}_s\right)\hat{B}\right\} \qquad (4.5.2)$$

where $\bar{X} = \sum_U x_k / N$ and $\hat{B} = \frac{Cov_{xys}}{S_{xs}^2}$

with $Cov_{xys} = \frac{1}{n-1}\sum_s \left(x_k - \bar{x}_s\right)\left(y_k - \bar{y}_s\right)$ and $S_{xs}^2 = \frac{1}{n-1}\sum_s \left(x_k - \bar{x}_s\right)^2$.

At this point, we only wish to emphasise that a given quantity of auxiliary information may lead to several different formulations of the auxiliary vector $\mathbf{x}_k$. Note that (iv) represents the most complete use of the existing information.

□

5. Introduction to estimation in the presence of nonresponse

5.1. General background

As in Chapter 4, our objective is to estimate the target population total of the study variable y, that is, $Y = \sum_U y_k$ or the domain totals $Y_d = \sum_{U_d} y_k$, $d = 1,...,D$. To this end, a sample s is drawn from the frame according to a given sampling design. As in Chapter 4 we assume in this chapter that the frame population agrees exactly with the target population, U. In practice this condition is often not met, because of frame errors. These imperfections are discussed in Chapter 11. A difference in this chapter compared with Chapter 4 is that we no longer assume full response.

The given sampling design determines the design weight $d_k = 1/\pi_k$ for every element $k \in U$, where π_k is the inclusion probability of k. We assume that response is obtained for the elements in a set denoted r. (The concept of "response set " is illustrated in Example 2.2.3.) Full response implies that $r = s$. Nonresponse implies that r is a proper subset of s. The nonresponse set is denoted $o = s - r$. The situation is illustrated in Figure 5.1.1.

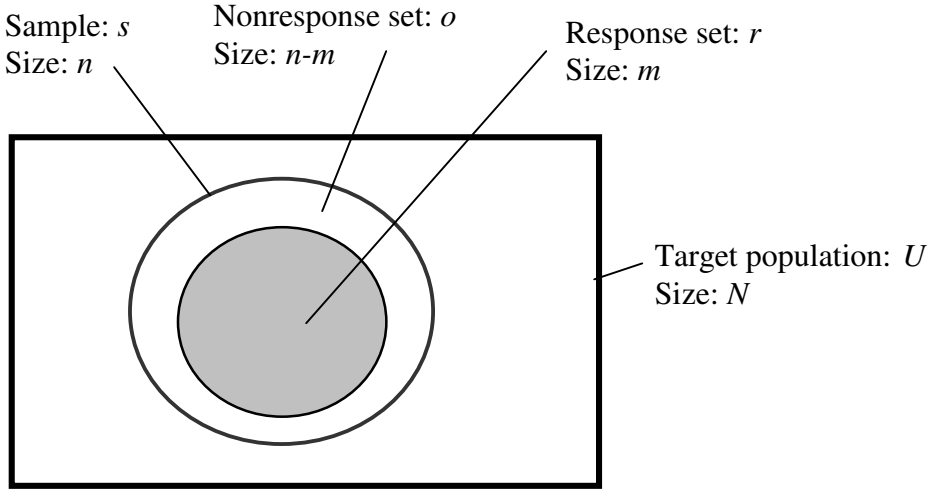


Figure 5.1.1.

Most surveys involve more than one study variable, and consequently we will ordinarily have both *unit nonresponse* and *item nonresponse* as explained in Section 2.2. In this chapter and in Chapters 6 and 7, we limit the discussion to the case of a single study variable y. Consequently, we can proceed in this chapter as if only unit nonresponse existed. If the survey has item nonresponse as well as unit nonresponse, important decisions must be made about how item nonresponse is to be treated in the estimation process. We defer this discussion until Chapter 9.

As mentioned earlier, the estimation methodology currently used by statistical agencies distinguishes two main approaches for dealing with nonresponse. These are *reweighting* and *imputation*. When reweighting is used, new weights are determined, with the aid of the available auxiliary information, and applied to the y-values for the responding elements $k \in r$. In this CBM, we use a *calibration approach* to compute these new weights. Consequently, the estimator of the parameter of interest, $Y = \sum_U y_k$, will be of the form $\hat{Y}_W = \sum_r w_k y_k$, where the new weights w_k are, at least for most elements, greater than the weights that would have been applied in the case of full response. This is in order to compensate for elements lost due to nonresponse. A detailed discussion of reweighting using the calibration approach is given in Chapter 6.

The other principal approach for nonresponse treatment, imputation, implies that proxy values are created for the values y_k that are missing because of nonresponse. The proxy value for element $k \in o$, often called the *imputed value* for k, is denoted $\hat{y}_k$. The superimposed "hat" serves as a reminder that an imputed value is in some sense an estimated value, rather than one that has been observed. For element k, the "estimation" is carried out by a given *imputation method*. There are many imputation methods in current use, and more than one imputation method is often used in one and the same survey. In other words, all $\hat{y}_k$ may not be constructed by the same method. Chapter 7 contains a detailed discussion of different imputation procedures. The *completed data set* will contain the same number of values as the originally intended sample s, that is, n values, and they are given by

$$y_{\bullet k} = \begin{cases} y_k & \text{for } k \in r \\ \hat{y}_k & \text{for } k \in o \end{cases} \qquad (5.1.1)$$

The estimator of the parameter of interest, $Y = \sum_U y_k$, will now be of the form $\hat{Y}_I = \sum_s d_k g_k y_{\bullet k}$, assuming that the GREG estimator (4.3.3) is used as a starting point. We call this estimator the imputed GREG estimator.

The imputed HT estimator is of the form $\hat{Y}_I = \sum_s d_k y_{\bullet k}$.

5.2. Error caused by sampling and nonresponse

This CBM proposes techniques for simultaneously reducing the sampling error and the nonresponse error *after the data collection stage*, that is, after some nonresponse has occurred. The principal avenue for reducing these errors is an effective use of auxiliary information. The success of this operation is contingent upon access to "strong" or "powerful" auxiliary information. As mentioned, two main approaches have evolved in this regard, namely, *reweighting* and *imputation*. Our general notation will be $\hat{Y}_W$ for a reweighting estimator and $\hat{Y}_I$ for an estimator created by imputation.

Within each approach, there are many possible uses of auxiliary information. Even when the available auxiliary information is "strong", we have to accept that both $\hat{Y}_W$ and $\hat{Y}_I$ are affected by sampling error and nonresponse error. However, generally speaking, the stronger the auxiliary information, the smaller the two errors.

The two types of error were briefly mentioned in Section 2.2. Our objective in this section is to arrive at a deeper understanding of these errors and the procedures used to reduce them.

In what follows the *nonresponse estimator*, $\hat{Y}_{NR}$, represents both $\hat{Y}_W$ and $\hat{Y}_I$. Further, we denote by $\hat{Y}$ the expression taken by $\hat{Y}_{NR}$ for the case of full response, when $r = s$. We assume that $\hat{Y}$, called the *full response estimator*, is either the GREG estimator or the HT estimator; see Sections 4.3 and 4.2. The total error of the $\hat{Y}_{NR}$ estimator (its deviation from the

target parameter value Y) can be decomposed into a sum of two error components,

$$\hat{Y}_{NR} - Y = (\hat{Y} - Y) + (\hat{Y}_{NR} - \hat{Y}) \qquad (5.2.1)$$

The first term on the right hand side, $\hat{Y} - Y$, is the *sampling error* (the error caused by selecting a sample only, rather than the whole population) and the second term, $\hat{Y}_{NR} - \hat{Y}$, is the *nonresponse error*.

We consider first the expected value, or average, of the estimator $\hat{Y}_{NR}$. It measures the *central tendency* of the estimator $\hat{Y}_{NR}$. The average (over all possible samples s) of the sampling error is zero or almost zero, since the full-response estimator is unbiased or nearly unbiased. The average (over all possible samples s *and* all possible response sets r) of the nonresponse error is likely to be different from zero. That is, nonresponse introduces bias into the estimation.

To analyse the *accuracy* of the estimator $\hat{Y}_{NR}$, we need to analyse its Mean Squared Error, MSE($\hat{Y}_{NR}$), which is the average of the squared total error, $(\hat{Y}_{NR} - Y)^2$, over all samples s and all response sets r.

The notions of expected value, unbiasedness and MSE are thus tied to a two-fold averaging process: over all possible response sets r, realised by the (unknown) response mechanism denoted $q(r|s)$, for a fixed sample s, and over all possible samples s, drawn by the known sampling design $p(s)$. We denote the expectation operators with respect to these two distributions by E_q and E_p, respectively. Operators with respect to both distributions jointly will be given the index pq.

The nonresponse bias can be expressed as

$$B_{pq}(\hat{Y}_{NR}) = E_p(B_c) \qquad (5.2.2)$$

where $B_c = E_q(\hat{Y}_{NR}|s) - \hat{Y}$ is the conditional nonresponse bias, given the realised sample s.

In practice, it is virtually impossible to tell whether the *condition for unbiasedness*, $B_{pq}(\hat{Y}_{NR}) = E_p(B_c) = 0$, is fulfilled, because the response mechanism $q(r|s)$ is unknown. However, in practice one often makes the subjective assumption that the nonresponse bias is "sufficiently" small. The assumption is sometimes justified. Much of the knowledge about the nonresponse bias comes from simulation studies where different population and response mechanisms are used. These problems are discussed in more detail in Chapter 10.

If we assume that the conditional bias B_c is zero or negligible for any realised samples s, then the variance is shown in Appendix A to be given by

$$V_{pq}(\hat{Y}_{NR}) = V_{SAM} + V_{NR} \qquad (5.2.3)$$

where $V_{SAM} = V_p(\hat{Y})$ and $V_{NR} = E_p V_q(\hat{Y}_{NR}|s)$. The component V_{SAM} is called the *sampling variance*. This is the variance over all possible samples that can be drawn with the given sampling design; it does not depend on the nonresponse or the response mechanism. The component V_{NR} is called the *nonresponse variance*. This is an average over all samples s as well as over all response sets r.

In order to assess the probable error of $\hat{Y}_{NR}$, we need an estimate of the total variance, represented here by the sum of the two terms in (5.2.3). (A survey normally has other significant errors, but they are not considered here.)

There is considerable interest also in evaluation of each of the two components individually. Practitioners usually have a very vague idea of how much the total variance is accounted for by the nonresponse variance component V_{NR}. This is because the routine measurement of the two components in (5.2.3) is seldom or never attempted. However, it is of practical interest to know the relative size of these two components. For example, if V_{NR} in a regularly repeated survey is found to account for a major proportion of the total variance, it may be an important signal to try to devote more of the survey resources to reducing the nonresponse on the next survey occasion.

What can be done to estimate the two components of (5.2.3)? Section 6.4 provides a general answer for the reweighting estimator $\hat{Y}_W$. This method is also applicable to the imputed estimator $\hat{Y}_I$, when it coincides with $\hat{Y}_W$, as is the case for the GREG-conformable multiple regression method; see Remark 7.2.1. For other imputation methods reviewed in Chapter 7 there is no general variance estimation method but attempts have been made to estimate the two components separately. The estimate of V_{NR} depends on the imputation method in use, because the methods are more or less accurate. For the sampling variance component V_{SAM}, one may use an appropriate modification of the formula intended for 100% response, given, in the case of GREG estimator, by (4.4.3). These questions are examined in Sections 7.3.4 and 7.3.5.

The nonresponse bias cannot be estimated, but some analysis is possible. Appendix C gives a general expression for the bias of the reweighting estimator $\hat{Y}_W$. The expression is a function of the study variable y_k, the auxiliary vector $\mathbf{x}_k$, the factor c_k and the response probability θ_k, given by (6.1.1). Some specific estimators and their bias expressions are analysed in Chapter 10. For most of the imputation methods there is no general expression for the nonresponse bias. One exception is the GREG-conformable multiple regression method, because then $\hat{Y}_I$ is identical to $\hat{Y}_W$ (see Remark 7.2.1 and Appendix D) and consequently, the general bias expression for $\hat{Y}_W$ applies.

6. Reweighting for nonresponse

6.1. Background and conventional methods for reweighting

At the time when Statistics Sweden's (1980) handbook "Räkna med bortfall" was written, the predominating view of nonresponse was centred on a *deterministic model* of survey response: The population was assumed to consist of two non-overlapping parts, a response stratum and a nonresponse stratum. Every element in the former was assumed to respond with certainty if selected for the sample, and every element in the latter stratum had probability zero to respond. An obvious criticism that could be levied against that model is that it is simplistic and unrealistic. Moreover, the sizes of the two strata could usually not be assumed to be known. An approach that was sometimes used was to estimate the total for the response stratum, and then to add a term to compensate for the nonresponse stratum.

In the 1980's, a more satisfactory *two-phase approach to reweighting for nonresponse* became popular. The name refers to a view of the selection process as one in which a desired sample s is first selected from the population U, whereupon a set of respondents, r, is realized as a subset of s. The approach is more realistic than the deterministic one in that it allows every element k to have its own individual response probability θ_k where $0 \leq \theta_k \leq 1$ for all k. This generality is not without a price: the response probabilities θ_k are usually unknown, and progress with this approach requires that the θ_k be replaced with estimates, constructed with the aid of auxiliary information.

The two-phase approach is discussed in the literature, for example, in Särndal, Swensson and Wretman (1992). Chapter 9 of that book develops the theory of two-phase sampling in the presence of auxiliary information. In the traditional formulation of two-phase sampling, a first sample is selected from U, certain variables (although not the study variable(s)) are observed, then a smaller subsample is realized from the first sample, and the study variable(s) are observed for the elements of the subsample. All inclusion probabilities are known by design, those for the first phase as well as those for the second phase.

Chapter 12 of the book adapts the two-phase theory to the case where sampling is followed by nonresponse. Assume for a moment that the response distribution $q(r|s)$ is known. (In practice this is not the case.) This implies that the first and second order response probabilities,

$$\Pr(k \in r|s) = \theta_k$$
and (6.1.1)
$$\Pr(k \& l \in r|s) = \theta_{kl}$$

are known. Let $\mathbf{x}_k$ be the auxiliary vector to be used in the estimator. Under these conditions, the two-phase GREG estimator of the population total $Y = \sum_U y_k$, as obtained from Chapter 9 of Särndal, Swensson and Wretman (1992), is given by

$$\hat{Y}_{SSW} = \sum_r d_k g_{k\theta} y_k / \theta_k \qquad (6.1.2)$$

where $d_k = 1/\pi_k$ and

$$g_{k\theta} = 1 + c_k (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k / \theta_k)' \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k' / \theta_k\right)^{-1} \mathbf{x}_k \qquad (6.1.3)$$

The transformation of this estimator into one that is useful for a sample survey with nonresponse requires replacing the unknown θ_k by estimates $\hat{\theta}_k$. This step entails: (a) the formulation of a realistic model for the response mechanism with the response probabilities θ_k as unknown parameters, and (b) the estimation of these response probabilities, using any relevant auxiliary variables and the fact that some sample elements were observed to respond whereas the others did not.

An often used model states that the population consists of nonoverlapping groups with the property that all elements within one and the same group respond with the same probability, and in an independent manner. Such groups are known as *response homogeneity groups* (RHGs). In a survey of individuals, the groups may be based on age by sex categories, for example. The auxiliary information required is that we can uniquely classify every sampled element, respondent or nonrespondent, into one of the groups. The

point estimator obtained from (6.1.2) when the unknown θ_k are replaced by the estimates $\hat{\theta}_k$ flowing from this RHG model is discussed in detail in Chapter 12 of Särndal, Swensson and Wretman (1992). These authors also give an appropriate variance estimator, composed as a sum of two components, one measuring the sampling variance, the other the nonresponse variance. The point estimator is essentially unbiased if the assumed RHG model is a true representation of the response pattern in the survey; the difficulty in practice is, of course, that it virtually impossible to foresee the true response pattern. Other attempts at modelling the response mechanism have been made, including logistic regression modelling, as in Ekholm and Laaksonen (1991).

The two-phase approach to reweighting has the following characteristics:

(i) the modelling of the response mechanism constitutes a separate step;

(ii) if a set of auxiliary variables is available, one subset of these variables is used in the estimation of the response mechanism, another subset (which may have some overlap with the first subset) is used to formulate the auxiliary vector $\mathbf{x}_k$ required for the estimator (6.1.2) of Y, where θ_k is replaced by $\hat{\theta}_k$.

In practice, the two-phase approach to reweighting requires analysis and decision making. The statistician must decide on the best use of the total set of available auxiliary variables for each of the two tasks in (ii). If nothing else, these selection tasks will take time. A simpler (but usually not any less efficient) alternative is the calibration approach to reweighting described in the following sections.

6.2. Introduction to the calibration approach

In this CBM, *calibration* is the main tool for reweighting for nonresponse. This *calibration approach* requires the formulation of a suitable auxiliary vector, through a selection from a possible larger set of available auxiliary variables. This step follows a few basic and simple principles. They are explained in Section 10.2.1**.** The next step is computational. A set of *calibrated weights* is produced, using the selected auxiliary information as an input. One of several existing computer programs can be used, for

example, CLAN97; see Section 6.5. This software can handle any auxiliary vector, and a number of important sampling designs.

This calibration approach leads to a *calibration estimator* of Y, denoted $\hat{Y}_W$, and a corresponding *variance estimator*, denoted $\hat{V}(\hat{Y}_W)$. The index W was chosen to suggest the term "weighting". The calibration approach provides a unified treatment of the use of auxiliary information in surveys with nonresponse. In the presence of powerful auxiliary information, the approach meets the objective of reducing both the sampling error and the nonresponse error. The approach is general in that it can be applied for most of the common sampling designs and with any number of variables present in the auxiliary vector. In the following sections we highlight the practical aspects of the approach and illustrate it by a number of examples. The theoretical aspects are only briefly outlined; for more detail on these, the reader is referred to Lundström and Särndal (1999) and Lundström (1997).

The calibration approach has only a single computational step, in which the calibrated weights are produced. It is thereby more direct than the two-phase approach, in that it requires no separate modelling of a nonresponse mechanism. For these reasons, the calibration approach is better suited for a routine treatment of nonresponse in a organization such as Statistics Sweden. It is in many cases as efficient as the two-phase approach.

6.3. Point estimation under the calibration approach

For the survey that interests us, we assume that the GREG estimator (4.3.3), with a specified vector $\mathbf{x}_k$, would be chosen if the survey had full response, so that $r = s$. A required input is the population total of the $\mathbf{x}_k$-vector, $\sum_U \mathbf{x}_k$. This estimator would be a good choice, because (a) it is unbiased, (b) its variance is small when $\mathbf{x}_k$ is a good explanatory vector for the study variable y_k, and (c) it is consistent in the sense mentioned in Section 4.3: the weights satisfy the calibration equation

$$\sum_s d_k g_k \mathbf{x}_k = \sum_U \mathbf{x}_k \qquad (6.3.1)$$

However, we are concerned here with surveys with nonresponse, so y_k-values are available only for the elements k in the response set r, a subset

of the sample s. Then, whatever the estimation technique, there will be some bias. Desirable properties of the chosen estimator are now: (i) a small nonresponse bias, (ii) a small total variance, and (iii) agreement with the GREG estimator (4.3.3) when $r = s$. The total variance is now the sum of the sampling variance and the nonresponse variance. Property (i) is particularly important.

The calibration estimator is, like the GREG estimator, formed as a linearly weighted sum of the observed y_k-values. It is defined by

$$\hat{Y}_W = \sum_r w_k y_k \qquad (6.3.2)$$

where $w_k = d_k v_k$ with

$$v_k = 1 + c_k \left(\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k \right)' \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k \quad \text{for } k \in r \qquad (6.3.3)$$

We omit the derivation that leads to the calibrated weights $w_k = d_k v_k$ in (6.3.2). Details are given in Lundström (1997). The principle behind the derivation is to minimize a function measuring the distance between the "old" weights, d_k, and the "new" weights, w_k, subject to the calibration equation

$$\sum_r d_k v_k \mathbf{x}_k = \sum_U \mathbf{x}_k \qquad (6.3.4)$$

The calibrated weights are therefore "as close as possible" (with respect to the given distance measure) to the design weights d_k, and they ensure consistency with the known auxiliary variable totals.

The degree to which $\hat{Y}_W$ succeeds in realising the desired properties (i) and (ii) depends on the quality of the auxiliary vector $\mathbf{x}_k$. Some $\mathbf{x}_k$-vectors succeed better than others, as the examples in the following sections will show. The desired property (iii) is easily verified: when $r = s$, v_k reduces to g_k given by (4.3.4), so for full response, $\hat{Y}_W$ is identical to the GREG estimator (4.3.3).

The following example is an over-simplification of any situation arising in practice, but it suggests very convincingly that in the presence of powerful auxiliary information, the calibration approach can produce highly accurate estimates.

EXAMPLE 6.3.1. *Perfect linear relationship in the population.*

Assume that a perfect linear relationship exists between the study variable y_k and the auxiliary vector $\mathbf{x}_k$, so that

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} \quad \text{for every} \quad k \in U \tag{6.3.5}$$

where $\boldsymbol{\beta}$ is a (column) vector of constants. Then $\hat{Y}_W$ provides an exact estimate of the total Y that we seek to estimate, that is, $\hat{Y}_W = Y$, which is easily shown by

$$\hat{Y}_W = \sum_r w_k y_k = (\sum_r w_k \mathbf{x}_k')\boldsymbol{\beta} = (\sum_U \mathbf{x}_k')\boldsymbol{\beta} = \sum_U \mathbf{x}_k'\boldsymbol{\beta} = \sum_U y_k = Y$$

□

This example is unrealistic, because in practice, one can never count on having the perfect linear relationship that (6.3.5) expresses; if it were known to hold, there would be no need for a survey. But the example does suggest that when the relationship, or the correlation, between y_k and $\mathbf{x}_k$ is strong, then the calibration estimator $\hat{Y}_W$ should come very near the "truth", *Y*, in other words, both the sampling error and the nonresponse error would be essentially eliminated.

We can also produce a calibration estimator for a survey in which the auxiliary vector values $\mathbf{x}_k$ are known up to the level of the sample *s*, whereas the population total $\sum_U \mathbf{x}_k$ is unknown. We still know enough to form the sample-based HT estimator of that total, namely, $\sum_s d_k \mathbf{x}_k$. As shown in Lundström (1997), and Lundström and Särndal (1999), calibration on this estimated total produces the weights $d_k v_{sk}$ in the following calibration estimator:

$$\hat{Y}_{Ws} = \sum_r d_k v_{sk} y_k \tag{6.3.6}$$

with

$$v_{sk} = 1 + c_k \left(\sum_s d_k \mathbf{x}_k - \sum_r d_k \mathbf{x}_k \right)' \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k \qquad (6.3.7)$$

The calibration approach is very flexible. Also, it is convincing to find that many "conventional" techniques for nonresponse reweighting are special cases of (6.3.2) or (6.3.6). Several such techniques are described in Statistics Sweden's handbook on nonresponse, Statistics Sweden (1980), and are familiar to methodologists at Statistics Sweden. All of them are derivable from the calibration approach, for simple formulations of the $\mathbf{x}_k$-vector, as illustrated in Section 6.6.

But it should be emphasized that when the calibration approach is applied in survey practice, there is no need to derive formulas for specific applications. All necessary computation is carried out by CLAN97, or a similar software, once the $\mathbf{x}_k$-vector and the factor c_k have been specified.

As pointed out earlier, most surveys require estimation not only for the whole population but also for various domains of the population. When the survey has nonresponse, and reweighting is carried out by the calibration approach, then the estimation of the domain total Y_d proceeds as follows:

If the auxiliary information consists of the known vector total $\sum_U \mathbf{x}_k$, a set of calibrated weights are given by (6.3.3). They were used in (6.3.2) to produce an estimator of the whole population total Y. Now, for the domain total Y_d, we keep the same weights and change only the study variable from y into y_d, defined by (4.1.2). The resulting calibration estimator of the domain total Y_d is therefore

$$\hat{Y}_{dW} = \sum_r w_k y_{dk} \qquad (6.3.8)$$

with $w_k = d_k v_k$ where v_k is given by (6.3.3).

In some applications the domains of interest $U_1, \ldots, U_d, \ldots, U_D$ form a partition of U, as when the domains are D regions making up a country. The D domain estimates $\hat{Y}_{1W}, \ldots, \hat{Y}_{dW}, \ldots, \hat{Y}_{DW}$ then have the appealing

property that they add up to the calibration estimate made for the whole population, that is, $\hat{Y}_W$ given by (6.3.2). This property follows from

$$\sum_{d=1}^{D} \hat{Y}_{dW} = \sum_{d=1}^{D} \sum_r w_k y_{dk} = \sum_r w_k \sum_{d=1}^{D} y_{dk} = \sum_r w_k y_k = \hat{Y}_W$$

Similarly, we can adapt the calibration estimator (6.3.6), which has auxiliary information up to the sample level only. The calibrated weights are then $d_k v_{sk}$, as in (6.3.6). To arrive at an estimator of the domain total Y_d, we again preserve the weights and substitute y_k for y_{dk}. The result is

$$\hat{Y}_{dWs} = \sum_r d_k v_{sk} y_{dk} \qquad (6.3.9)$$

6.4. Variance estimation under the calibration approach

For statements of precision and confidence intervals, we need to estimate the variance of the different calibration estimators introduced in the preceding section. We rely on an analogy with the estimator (6.1.2) for two-phase sampling. An appropriate variance estimator for (6.1.2) is given by formula (9.7.22) in Särndal, Swensson and Wretman (1992). It assumes that the first and second order response probabilities, θ_k and the θ_{kl} are known. In the calibration approach, inclusion probabilities do not even enter the picture. Nevertheless, proxies for the inclusion probabilities are needed for variance estimation as we now explain.

To estimate the variance of $\hat{Y}_W$, we propose to use formula (9.7.22) of Särndal, Swensson and Wretman (1992) as follows: (i) replace $\pi_{k|s_a}$ by θ_k and then θ_k by the proxy value $\hat{\theta}_k = 1/v_{sk}$, where v_{sk} is given by (6.3.7), and (ii) assume elements to respond independently, so that $\theta_{kl} = \theta_k \theta_l$. The rationale for (i) is given in Appendix B; for further detail, see Lundström and Särndal (1999) and Lundström (1997). We arrive at the following variance estimator:

$$\hat{V}(\hat{Y}_W) = \hat{V}_{SAM} + \hat{V}_{NR} \qquad (6.4.1)$$

where

$$\hat{V}_{SAM} = \sum\sum_r (d_k d_l - d_{kl})(g_k v_{sk} e_k)(g_l v_{sl} e_l) -$$

$$- \sum_r d_k (d_k - 1) v_{sk} (v_{sk} - 1)(g_k e_k)^2 \qquad (6.4.2)$$

and

$$\hat{V}_{NR} = \sum_r d_k^2 v_{sk} (v_{sk} - 1)\, e_k^2 \qquad (6.4.3)$$

where v_{sk} is given by (6.3.7),

$$e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_v ; \qquad (6.4.4)$$

$$\hat{\mathbf{B}}_v = \left(\sum_r d_k v_{sk} c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k v_{sk} c_k \mathbf{x}_k y_k \qquad (6.4.5)$$

and g_k is given by (4.3.4).

The variance estimator (6.4.1) has two components, an estimate of the sampling variance, V_{SAM}, and an estimate of the nonresponse variance, V_{NR}; see Section 5.2. When (6.4.1) is used for computing confidence intervals, there is an implicit assumption that the conditional nonresponse bias, B_c, is small. If the bias is considerable, the true confidence level of an interval centred on $\hat{Y}_W$ and computed with the aid of (6.4.1) may be rather far from the desired 1 - α level. For a nearly correct confidence level, it is important that the bias is near zero, or is at least only modest.

Remark 6.4.1. Remark 4.4.1 stated that the variance estimator (4.4.3) of the GREG estimator often suffers from some negative bias in not-so-large samples and suggested the use of an alternative denoted $\hat{V}_{adj}(\hat{Y}_{GREG})$. The variance estimator (6.4.1) has the same weakness as (4.4.3). As proposed in Remark 4.4.1, the underestimation is attenuated by using instead the residuals $e_{adj,k} = f_k e_k$, where f_k adjusts for a loss of degrees of freedom.

□

Next, consider the calibration estimator $\hat{Y}_{Ws}$, given by (6.3.6). Appendix B provides a rationale for the following variance estimator:

$$\hat{V}(\hat{Y}_{Ws}) = \hat{V}_{SAM} + \hat{V}_{NR} \tag{6.4.6}$$

where

$$\hat{V}_{SAM} = \sum\sum_r (d_k d_l - d_{kl})(v_{sk} y_k)(v_{sl} y_l) -$$

$$- \sum_r d_k (d_k - 1) v_{sk} (v_{sk} - 1) y_k^2 \tag{6.4.7}$$

and

$$\hat{V}_{NR} = \sum_r d_k^2 v_{sk} (v_{sk} - 1) e_k^2 \tag{6.4.8}$$

Finally, we need to address the domain estimators, $\hat{Y}_{dW}$ given by (6.3.8) and $\hat{Y}_{dWs}$ given by (6.3.9). An appropriate variance estimator for $\hat{Y}_{dW}$ follows easily by replacing y_k by y_{dk} throughout the calculations defined by (6.4.1) to (6.4.5). That is, in (6.4.2) and (6.4.3) we replace e_k by

$$e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{dv} \tag{6.4.9}$$

where

$$\hat{\mathbf{B}}_{dv} = \left(\sum_r d_k v_{sk} c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k v_{sk} c_k \mathbf{x}_k y_{dk} \tag{6.4.10}$$

A similar argument produces a variance estimator for $\hat{Y}_{dWs}$ given by (6.3.9).

6.5. Software for computing point estimates and variance estimates

Two steps need to be considered: computing the point estimator and computing the corresponding variance estimator.

When the auxiliary information consists of the known total $\sum_U \mathbf{x}_k$, the point estimator of the population y-total is $\hat{Y}_W$, given by (6.3.2), and that of the domain total is $\hat{Y}_{dW}$, given by (6.3.8). Both are easily computed by CLAN97, for the sampling designs in common use at Statistics Sweden.

The corresponding variance estimates are also computed by CLAN97 according to the two-component formula (6.4.1), in the case of $\hat{Y}_W$. This step requires two additional sets of weights, the v_{sk} given by (6.3.7) and the g_k given by (4.3.4).

CLAN97 can also compute point estimates and variance estimates for more complex parameters, built as certain types of functions of totals. Consider the parameter $\psi = f(Y_1,...,Y_q,...Y_Q)$, where f is a specified function of the Q population totals $Y_1,\ldots,Y_q,\ldots Y_Q$. In particular, rational functions are of interest in many surveys. (A rational function is one that is limited to use of the four basic algebraic rules, addition, subtraction, multiplication and division.) An example of such a parameter is a difference of ratios, $\psi = Y_1/Y_2 - Y_3/Y_4$. For any rational function of totals, CLAN97 can compute (i) the point estimate determined as $\hat{\psi} = f(\hat{Y}_1,...,\hat{Y}_q,...\hat{Y}_Q)$, where $\hat{Y}_1,\ldots,\hat{Y}_q,\ldots\hat{Y}_Q$ are the respective calibration estimates of the Q totals (population totals or domain totals), and (ii) the corresponding variance estimate.

Several sampling designs are implemented in CLAN97. They include SRS and STSRS (of elements or of clusters), probability-proportional-to-size sampling, and two-phase sampling for stratification. The variance computation for pps sampling uses an approximate formula. CLAN97 also handles network sampling of the type arising when individuals are sampled from the TPR (see Example 2.2.1) and the observational elements are the households formed around the selected individuals. Sampling designs in two or more stages are not yet implemented in CLAN97, because such designs presently find little or no application at Statistics Sweden.

6.6. Examples of calibration estimators

In survey practice, we can always, for any specified $\mathbf{x}_k$-vector, use CLAN97 to compute the calibration estimators defined in Section 6.3. There is just one general approach. Specific formulas need never enter the picture. But many methodologists are accustomed to specific formulas corresponding to particular "methods" for nonresponse reweighting. Therefore, the objective with this section is to show that the calibration approach reproduces formulas that many readers are familiar with. Thus, we derive the explicit form of (6.3.2) and (6.3.6) for some simple specifications of the auxiliary information and show that commonly used estimators are obtained. We start with the simplest forms of $\mathbf{x}_k$, then gradually increase the auxiliary information content and thereby also the complexity of the formulas. We examine only a very limited subset of all the different possibilities covered by (6.3.2) and (6.3.6) when $\mathbf{x}_k$ is allowed to vary. For simplicity, the following example assume the SRS design, so that $d_k = N/n$ for all k, where n is the sample size. We start with the simplest formulation of $\mathbf{x}_k$.

The simplest auxiliary vector

The simplest formulation of the auxiliary vector is $\mathbf{x}_k = 1$ for all k. This vector recognises no differences among elements. Specifying also $c_k = 1$ for all k, (6.3.3) gives the weight $v_k = n/m$ for all k, so the calibration estimator (6.3.2) becomes

$$\hat{Y}_W = \frac{N}{m}\sum_r y_k = \hat{Y}_{EXP} \qquad (6.6.1)$$

The subscript EXP reflects the often-used term *expansion estimator*. Clearly, $\hat{Y}_{EXP}$ is a primitive estimator, often misleading because of a large bias. Nevertheless, it may occasionally find use in practice, namely, if no useful auxiliary information is available and nonresponse is, for good reasons, considered as occurring at random. Also, it is sometimes computed in a survey as a benchmark estimator to which better alternatives can be compared. In Statistics Sweden (1980), the technique is called "straight expansion" ("rak uppräkning").

One-way classification

In this formulation the target population U is divided into non-overlapping and exhaustive groups, U_p, $p = 1,\ldots,P$, based on a specified classification criterion, for example, age by sex groups. The auxiliary vector for element k is the group identifier $\mathbf{x}_k = (\gamma_{1k},\ldots,\gamma_{pk},\ldots,\gamma_{Pk})'$ where, for $p = 1,\ldots,P$,

$$\gamma_{pk} = \begin{cases} 1 & \text{if } k \in U_p \\ 0 & \text{otherwise} \end{cases} \qquad (6.6.2)$$

We have $\sum_U \mathbf{x}_k = \left(N_1,\ldots,N_p,\ldots,N_P\right)'$, where N_p is the size of U_p. Thus the requirement that the population auxiliary total be known is tantamount to requiring that the P group sizes be known. Letting $c_k = 1$ for all k, we obtain from (6.3.3)

$$v_k = N_p n / N m_p \quad \text{for } k \in r_p \qquad (6.6.3)$$

and the calibration estimator (6.3.2) becomes

$$\hat{Y}_W = \sum_{p=1}^{P} N_p \bar{y}_{r_p} = \hat{Y}_{PST} \qquad (6.6.4)$$

where $\bar{y}_{r_p} = \dfrac{1}{m_p}\sum_{r_p} y_k$ and m_p being the number of respondents in group p.

This estimator is commonly called the *poststratified estimator*, so we denote it by $\hat{Y}_{PST}$. The term is mildly misleading in that the traditional poststratified estimator is defined with reference to a single phase of sampling. Recognising this, Kalton and Kasprzyk (1986) make a distinction between the poststratified estimator, as used for the case of full response in single-phase sampling, and $\hat{Y}_{PST}$ given by (6.6.4), which they call the *population weighting adjustment estimator*. In the latter, more accurate term lies a recognition of a sampling phase followed by a nonresponse phase. Several authors have discussed this estimator, including Jagers (1986), Bethlehem and Kersten (1985) and Thomsen (1973, 1978).

When knowledge of the auxiliary vector $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{pk},...,\gamma_{Pk})'$ is limited to the elements of the sample *s*, the estimator $\hat{Y}_{Ws}$, given by (6.3.6), can be used. Letting $c_k = 1$ for all *k*, we obtain

$$\hat{Y}_{Ws} = \sum_{p=1}^{P} \hat{N}_p \bar{y}_{r_p} = \hat{Y}_{WCE} \qquad (6.6.5)$$

with $\hat{N}_p = \frac{N}{n} n_p$, where n_p is the number of sampled elements in group *p*. Known as *the weighting class estimator*, and therefore denoted by $\hat{Y}_{WCE}$, it is also an often discussed estimator; see, for example, Oh and Scheuren (1983), Kalton and Kasprzyk (1986), Little (1986), Statistics Sweden (1980). It can be described as "expansion by groups, using the response rate", the term used in Statistics Sweden (1980) being "gruppvis uppräkning med andel svar".

Remark 6.6.1. The concept of Response Homogeneity Groups (RHGs) was defined in Section 6.1. Estimators derived from the RHG model for response can be computed by CLAN97 in the following situations. Suppose first that an SRS is drawn and partitioned into a set of predefined RHGs. Define $\mathbf{x}_k$ to be the RHG identifier vector, observed for the sample elements only, let $c_k = 1$ for all *k*, and assume no other auxiliary information. The resulting estimator is then $\hat{Y}_{WCE}$ as given by (6.6.5), where p now denotes the RHG index. Now, in a typical survey at Statistics Sweden, group counts are usually known at the population level, and profiting from this more extensive information we obtain instead $\hat{Y}_{PST}$ as given by (6.6.4), where p is again the RHG index. Both (6.6.4) and (6.6.5) can be computed by CLAN97. They have the same nonresponse bias, but (6.6.4) usually has the smaller variance, since the information is at a higher level. CLAN97 also computes estimates for the STSRS design, allowing a set of RHGs to be defined within each stratum.

□

A single quantitative variable

Assume that a quantitative auxiliary variable x_k is available, for example, the number of employees of enterprise k in a business survey, $k = 1,...,N$. Its population total, $\sum_U x_k$, is assumed known. If this is the only auxiliary variable, the auxiliary vector is uni-dimensional, $\mathbf{x}_k = x_k$. If we also specify $c_k = x_k^{-1}$, the estimator obtained from (6.3.2) is

$$\hat{Y}_W = (\sum_U x_k)\frac{\bar{y}_r}{\bar{x}_r} = \hat{Y}_{RA} \qquad (6.6.6)$$

where $\bar{y}_r = \frac{1}{m}\sum_r y_k$, and $\bar{x}_r$ is defined analogously. It has the well-known form of a *ratio estimator*, hence the notation $\hat{Y}_{RA}$. Note, however, that the ratio estimator usually discussed in textbooks is (4.5.1), which is the full-response version of (6.6.6), obtained when $r = s$. Under SRS, (4.5.1) is unbiased, but such a property cannot be claimed for (6.6.6), because of the nonresponse.

With the same information, we can alternatively formulate the auxiliary vector as $\mathbf{x}_k = (1, x_k)'$. This option exists, because the auxiliary information required, in addition to $\sum_U x_k$, is the population size, $N = \sum_U 1$, which is known. When $c_k = 1$ for all k, (6.3.2) gives

$$\hat{Y}_W = N\{\bar{y}_r + (\bar{X} - \bar{x}_r)\hat{B}\} = \hat{Y}_{REG} \qquad (6.6.7)$$

where $\bar{X} = \frac{1}{N}\sum_U x_k$ and

$$\hat{B} = \left[\sum_r y_k x_k - \frac{1}{m}\sum_r y_k \sum_r x_k\right] / \left[\sum_r x_k^2 - \frac{1}{m}(\sum_r x_k)^2\right]$$

The notation $\hat{Y}_{REG}$ is used to indicate the *regression estimator* form. The estimator is discussed, for example, in Bethlehem (1988). The classic simple regression estimator found in standard textbooks is (4.5.2), which is the special case of (6.6.7) obtained for $r = s$.

One-way classification combined with a quantitative variable

In this application, the auxiliary information concerns a P-valued categorical variable and a quantitative variable, x, which may be an indicator of the size of an element. Assume that we can place every sampled element k into the appropriate group, that we know its value x_k, and that for each group, $p = 1,...,P$, we know the size, N_p, and the x-total, $\sum_{U_p} y_k$. There are more than one way to use this information. One option is to define the auxiliary vector as

$$\mathbf{x}_k = (\gamma_{1k} x_k, ..., \gamma_{pk} x_k, ..., \gamma_{Pk} x_k)'$$

where γ_{pk} is defined by (6.6.2). The population total of $\mathbf{x}_k$ is then the vector composed of the P known group sums $\sum_{U_p} x_k$. It is true that this formulation of $\mathbf{x}_k$ ignores the information about the group sizes N_p, and this may amount to a nonnegligable waste of information. Nevertheless, it leads to a well-known estimator, because if we let $c_k = x_k^{-1}$, (6.3.2) becomes

$$\hat{Y}_W = \sum_{p=1}^{P} (\sum_{U_p} x_k) \frac{\bar{y}_{r_p}}{\bar{x}_{r_p}} = \hat{Y}_{SEPRA} \qquad (6.6.8)$$

where $\bar{y}_{r_p} = \frac{1}{m_p} \sum_{r_p} y_k$ and $\bar{x}_{r_p}$ is analogously defined. Thus, $\hat{Y}_{SEPRA}$ has the well known form of a *separate ratio estimator*, that is, one that is built as a sum of ratio estimators, one for each group.

If the auxiliary information goes only up to the level of the sample s, we get from (6.3.6) $\hat{Y}_{Ws} = \frac{N}{n} \sum_{p=1}^{P} (\sum_{s_p} x_k) \frac{\bar{y}_{r_p}}{\bar{x}_{r_p}}$, which is also commonly used and can be described as "weighting by groups using a size variable"; see Statistics Sweden (1980), p. 3:12.

To take advantage of the complete information, the sizes N_p as well as the x-totals $\sum_{U_p} x_k$, we should instead formulate the auxiliary vector as

$$\mathbf{x}_k = (\gamma_{1k},\ldots,\gamma_{pk},\ldots,\gamma_{Pk},\gamma_{1k}x_k,\ldots,\gamma_{pk}x_k,\ldots,\gamma_{Pk}x_k)'$$

Then, if we let $c_k = 1$ for all k, the estimator (6.3.2) becomes

$$\hat{Y}_W = \sum_{p=1}^{P} N_p \left\{ \bar{y}_{r_p} + \left(\bar{X}_p - \bar{x}_{r_p} \right) \hat{B}_p \right\} = \hat{Y}_{SEPREG} \qquad (6.6.9)$$

with $\bar{X}_p = \frac{1}{N_p} \sum_{U_p} x_k$ and $\hat{B}_p = \frac{Cov_{xyr_p}}{S^2_{xr_p}}$

where

$$Cov_{xyr_p} = \frac{1}{m_p - 1} \left[\sum_{r_p} y_k x_k - \frac{1}{m_p} \sum_{r_p} y_k \sum_{r_p} x_k \right]$$

and

$$S^2_{xr_p} = \frac{1}{m_p - 1} \left[\sum_{r_p} x_k^2 - \frac{1}{m_p} \left(\sum_{r_p} x_k \right)^2 \right]$$

The estimator (6.6.9) is another well-known form, namely that of the *separate regression estimator*; consequently, the notation is $\hat{Y}_{SEPREG}$.

Two-way classification

In practice it is common to have information on two or more categorical auxiliary variables. We discuss the case of two categorical variables. The reasoning can be extended to a multi-way classification.

Suppose there are P categories of the first factor, for example, a geographical classification, and H categories of the second, for example, a socio-economic classification. We can think of the population U as partitioned into to $P \times H$ subsets or cells, U_{ph}; $p = 1,\ldots,P$; $h = 1,\ldots,H$. Depending on the information available about the cells, several formulations of the $\mathbf{x}_k$-vector are possible.

Consider the auxiliary vector formulation

$$\mathbf{x}_k = (\gamma_{1k},\ldots,\gamma_{pk},\ldots,\gamma_{Pk},\delta_{1k},\ldots,\delta_{hk},\ldots,\delta_{Hk})' \qquad (6.6.10)$$

where the γ's indicate the first classification with P groups and the δ's indicate the second classification with H groups. Specifically, γ_{pk} is defined by (6.6.2), and, for, $h = 1,\ldots,H$,

$$\delta_{hk} = \begin{cases} 1 & \text{if element } k \in \text{group } h \\ 0 & \text{otherwise} \end{cases} \qquad (6.6.11)$$

It is easily seen that this formulation of $\mathbf{x}_k$ requires knowledge of the $P+H$ marginal group counts, $N_{p.}\ (= \sum_{h=1}^{H} N_{ph})$, $p = 1,...,P$, and $N_{.h}\ (= \sum_{p=1}^{P} N_{ph})$, $h = 1,\ldots,H$. With this formulation, we can treat three commonly occurring situations (see also Example 4.5.2):

(i) The $P \times H$ cell counts, N_{ph}, $p = 1,\ldots,P$; $h = 1,\ldots,H$, are known, but it is considered that the set of $P + H$ marginal counts, $N_{p.}$, $p = 1,\ldots,P$, and $N_{.h}$, $h = 1,\ldots,H$, contain almost as much information.

(ii) The $P \times H$ cell counts, N_{ph}, $p = 1,\ldots,P$; $h = 1,\ldots,H$, are known, but a number of them are extremely small or zero, a situation frequently arising in practice. Collapsing of cells, although a commonly used remedy for this problem, could cause a non-negligible loss of auxiliary information. It may then be preferable to simply use the margin totals.

(iii) The marginal counts, the $N_{p.}$ and the $N_{.h}$, are known, but not the cell counts N_{ph}. An example of this happening in practice is when the $N_{p.}$ and the $N_{.h}$ are taken from two different registers.

Under the formulation (6.6.10) of the $\mathbf{x}_k$-vector, the calibration approach estimator, $\hat{Y}_W$ arising from (6.3.2), has no simple form. Computationally, it however easy to obtain this estimator in any given application, using existing software such as CLAN97.

A general discussion of calibration for cross-classifications in the full-response case is found in Deville, Särndal and Sautory (1993).

An alternative treatment of the same auxiliary information as that required by (6.6.10) (that is, the $P + H$ marginal counts) is the *raking ratio method*, which is also a well-known procedure; see, for example, Oh and Scheuren (1983). The raking ratio method does not give identically the same point estimator as the calibration approach estimator (6.3.2), but they differ by very little, in most situations.

7. Imputation

7.1. Introduction

7.1.1. Types of imputed values

Imputation is the procedure whereby missing values for one or more study variables are "filled in" with substitutes. These substitutes can be constructed according to some rule, or they can be observed values but for elements other than the nonrespondents. Thus imputed values are artificial; they contain error. *Imputation error* is similar to measurement error (as when a respondent provides an erroneous value for an item on a questionnaire) in that the true value is not recorded. But unlike measurement error, imputation error occurs "by construction", since the statistician knowingly inserts a value that is more or less wrong.

Another term used in connection with constructed values is *derived variable*. For legal reasons, imputation is not always allowed at Statistics Sweden, but derived variables are allowed. Therefore it is important to know the difference between an imputed value and a derived value. In Section 7.4 we discuss the Swedish legislation and define the terms in more detail.

Another type of artificial value construction practiced by some statistical agencies is *mass imputation*. In this procedure, values are imputed not only for the sampled elements, but for all non-observed elements in the population. Mass imputation is not discussed here.

Imputed values can be classified into three major categories:

(i) values constructed with the aid of a statistical prediction rule;
(ii) values observed not for the nonresponding elements themselves, but for (similar) responding elements;
(iii) values constructed by expert opinion or "best possible judgment".

Categories (i) and (ii) can be termed *statistical rules*, because they use a statistical technique to produce a reasonably close substitute value. Category (i) is often based on regression prediction. Category (ii) methods can also be described as *donor-based*, in that the value of another observed element is

imputed. Category (iii) methods are more subjective and often rely heavily on expert skill.

By another often-used distinction, imputed values are either *deterministic* (when repeating the imputation procedure would yield exactly the same imputed values) or *random* (when repeating the procedure would, barring pure chance, yield different imputed values). Regression imputation is an example of a deterministic rule, whereas an example of a random method is to impute the value of a randomly selected observed element (this is called "hot deck imputation"; see below).

Categories (i), (ii) and (iii) are presented in more detail in the following.

Imputation is regarded by many – both statisticians and subject-matter specialists – with some suspicion. This is because it goes against common statistical sense to use values known at the outset to be more or less wrong.

Nevertheless, there is no convincing evidence that careful imputation does any more harm to the quality of survey estimates than the reweighting methods described in Chapter 6. Both procedures lead to estimates with some - hopefully small - bias. Imputation may sometimes give better results, as when, for a highly skewed population (such as those occurring in many business surveys), expert judgment can be relied on to get a "close" imputed value for a large, influential nonresponse element.

The choice between reweighting and imputation is perhaps not so much the central issue as rather the threat posed to the quality of the survey estimates by two undesirable circumstances, namely (i) that a nonnegligible nonresponse has occurred and (ii) that a less than perfect approach (reweighting or imputation) is used to "correct" for the nonresponse. Thus, the quality of the estimates is at stake if survey managers become complacent and uncritical about the amount of nonresponse they tolerate in their survey and about the methods (reweighting or imputation) they use to treat the nonresponse that has occurred.

It goes without saying that the construction of imputed values should be carried out with professional care. The imputed values must come as close as possible to the true unobserved values for which they are substitutes.

7.1.2. The objective of imputation

An often-mentioned justification for imputation (rather than reweighting) is that it produces a *rectangular data set.* There are usually several (or many) y-variables in a survey. Each record (each element) defines a row in a data matrix with J columns, where J is the number of y-variables. If there are n records, we have an n by J data matrix, which, before imputation, contains a number of "holes" caused by missing y-values.

It is an advantage if all variables can be treated in a uniform manner in the production of statistics and if the same weighting can be applied to all variables when estimates are made. With imputation, this becomes possible.

There are two frequently used approaches for imputation, both leading to rectangular data matrices, namely the ITIMP-approach and the UNIMP-approach. (These approaches, and the concepts of item nonresponse and unit nonresponse, were defined in Section 2.2):

ITIMP-approach: Imputation is used to treat the item nonresponse only. In this procedure, we impute values for the m elements for which at least one but not all y-values are missing. The resulting rectangular data matrix has the dimensions m by J. Reweighting is then applied to compensate for the unit nonresponse.

UNIMP-approach: Imputation is used for both item nonresponse and unit nonresponse. In this procedure, we impute for all elements having at least one y-value missing. The resulting completed rectangular data matrix has the dimensions n by J, where n is the sample size. There is no nonresponse weight adjustment.

In the UNIMP-approach, when estimates are produced from the completed data matrix, other weights also enter into consideration:

(i) sample weighting generated by the sampling design, and, if applicable,
(ii) g-weighting (see (4.3.4)) to take into account any available auxiliary information at the population level.

In the rest of this chapter, we discuss, without loss of generality, in terms of a single y-variable.

EXAMPLE 7.1.1. *Illustration of the terminology.*

The following simple example illustrates some of the terminology. Suppose that there is only one y-variable and that every missing value is imputed by the average of the respondent y-values. The data set after imputation (the completed data set) will then consist of m actually observed values, y_k, $k \in r$, and $n - m$ imputed values, all of which are equal to $\bar{y}_r = \sum_r y_k / m$. As is intuitively clear, the method is not very efficient and is not recommended in a survey with high requirements for quality. It does realise one objective, namely, to obtain a completed data set. It is easy to identify several shortcomings of this data set. In most cases neither the central tendency nor the variance of these data will agree with what is expected of a data set with 100% response. The variance will be unnaturally small, because $n - m$ missing values have been imputed by one and the same value, the respondent mean $\bar{y}_r$. Also, the central tendency of these data will often not reflect the true central tendency of the y-variable: If, for example, large y-value elements respond less often than small y-value elements, then the average for the completed data set is likely to fall short of the mean of a data set with 100% response. If instead we impute the value of a randomly selected responding element (the "donor element"), then the variance of the completed data set will look more "normal". But the central tendency will have the same shortcoming as in imputation by the respondent mean. Also, since the donor is randomly selected, we run the risk of imputing the value of a very large element by that of a very small element, so the "closeness" of any individual imputed value is often compromised by this procedure.

□

7.1.3. The completed data set

There exist many imputation methods. Some of the more common ones are reviewed and discussed in the following. Each method has a number of variations. A minor variation of a well-established method may be required to meet the particular needs and requirements of a given survey.

We shall first discuss imputation and its consequences within a framework that is sufficiently general to cover the various imputation techniques that are in common use.

As before let U, s and r denote, respectively, the target population, the probability sample drawn from U, and the response set realised from s. The nonresponse set is denoted $o = s - r$. Denote by y_k the value of the variable y for element k. If element k is a nonrespondent and imputation is used for this element, we denote the imputed value as $\hat{y}_k$. More than one imputation method may be used in the same survey, so not all $\hat{y}_k$ may result from the same method.

The *completed data set* is defined as the set of values $\{y_{\bullet k} : k \in s\}$, where

$$y_{\bullet k} = \begin{cases} y_k & \text{for} \quad k \in r \\ \hat{y}_k & \text{for} \quad k \in o \end{cases} \tag{7.1.1}$$

That is, the value $y_{\bullet k}$ equals the observed value y_k when k is a respondent, or the imputed value $\hat{y}_k$ when k is a nonrespondent. Traditional descriptive statistics (mean, variance, and so on) can be computed from the completed data set. For example, the mean of the completed data set, $\bar{y}_{\bullet s} = \sum_s y_{\bullet k} / n$, can be computed. A different mean, not computable, is the one that would have been computed in the case of 100% response, namely, $\bar{y}_s = \sum_s y_k / n$. Both means are based on n values, but they will (barring pure chance) differ, to an unknown extent. Similarly, we can compute a variance and other standard statistics from the completed data set. They will differ from their counterparts for a hypothetical data set of 100% observed values.

7.2. Point estimation when imputation is used

7.2.1. The estimator

We discuss imputation in the context of estimating the population total for the variable y, $Y = \sum_U y_k$. Suppose that we consider that imputed values are as "good" as true observations. Such a belief is a justification for using exactly the same estimation method as in the ideal case of 100% response. We will employ the "standard estimator formula" and simply apply it to the completed data set. Consequently, when an estimate is computed, element *k* will receive the same weight whether its recorded *y*-value is a true observation, y_k, or an imputed value, $\hat{y}_k$. This is worth pointing out,

because some would argue that imputed values should be weighted according to some principle other than that used for truly observed values.

Current practice for point estimation is in fact that survey statisticians treat imputed data as real, observed data. That is, the procedure is: determine an estimator suitable for 100% response, then, after imputation, compute this estimator for the completed data set.

The estimator intended for use in the case of 100% response will be called the *full response estimator*. Here we assume that this estimator is the GREG estimator discussed in Section 4.3. It is of the form $\hat{Y} = \sum_s d_k g_k y_k$ as described by formulas (4.3.3) and (4.3.4). A special case is the Horwitz-Thompson estimator, $\hat{Y} = \sum_s d_k y_k$, discussed in Section 4.2.

Now suppose there is nonresponse treated by imputation. We then have a completed data set, given by (7.1.1). It replaces the desired (but not realised) data set composed of 100% real observations. We apply the weighting, $d_k g_k$, of the full response estimator (4.3.3). This gives the imputed GREG estimator

$$\hat{Y}_I = \sum_s d_k g_k y_{\bullet k} \tag{7.2.1}$$

which can also be written as

$$\hat{Y}_I = \sum_r d_k g_k y_k + \sum_o d_k g_k \hat{y}_k \tag{7.2.2}$$

In current practice, point estimation in the presence of imputation is thus very simple, since the weights are not changed. By contrast, variance estimation becomes a complex issue, as discussed in Section 7.3. The imputed HT estimator is

$$\hat{Y}_I = \sum_s d_k y_{\bullet k} \tag{7.2.3}$$

7.2.2. Statistical rules versus expert judgment

At statistical agencies, imputation is usually motivated by a desire to provide the "best possible imputed value" on an *element by element basis*. In other

words, imputation is practiced to provide good data at the micro level, rather than at some aggregated level.

Three categories of imputation methods were mentioned in Section 7.1.1: (i) imputation by applying a statistical prediction rule; (ii) imputation by inserting the value of a donor element; (iii) imputation by expert opinion or "best possible judgment". For most elements, in particular small to medium-sized elements, imputation is usually carried out by one or more of the methods in categories (i) and (ii), using a computerised routine. Category (iii) is usually reserved for a small number of large and influential elements, paying special attention to the characteristics of these elements. It is always debatable what constitutes the "best possible" imputation for any given element. There is usually more than one possibility.

The following example illustrates the difference between statistical imputation (categories (i) and (ii)) and "special imputation" (category (iii)).

EXAMPLE 7.2.1. *The difference between statistical imputation and "special imputation".*

Large elements are influential in that their impact on published survey estimates can be considerable. Consider a business survey in which a very large enterprise happens to be a nonrespondent and requires imputation. Since this enterprise is large compared to other enterprises in the same industry group (for example, an SIC code category), a simple imputation based on the respondent mean for this group would yield a large negative imputation error, $\hat{y}_k - y_k$, for this element. Similarly, the respondent mean for a group of enterprises considered "large" might also be misleading, because an enterprise that is "large" in one industry may have very different characteristics (Gross Business Income, for example) than one that is "large" in another industry. Even imputing the respondent mean of an industry-by-size group may be considered unsatisfactory. The value of a donor element identified as the "closest neighbour" may also be an imprecise imputation, because in the upper tail of the distribution, even the closest element may be numerically very different. A better approach may be a combination of historical examination and subjective ("expert") judgment. Thus, one might start by examining the series of earlier reported values, in particular the most recently reported value, and then adjust this value in the light of the best available judgment about trends in the industry and in the economy in general. The justification for this procedure is that truly large enterprises are

often so unique that none of the statistical rules are likely to "come close". This practice will often require excellent skill and judgment.

□

7.2.3. Imputation practices based on a statistical rule

General comments

Some of the more commonly used statistical rules are:

- ratio imputation;
- (multiple) regression imputation;
- nearest neighbour imputation;
- hot deck imputation;
- respondent mean imputation.

For each of these there are several minor variations, as explained later.

The first three rules require auxiliary information. We refer to the auxiliary vector as the *imputation vector* in order to distinguish it from the possibly different auxiliary vector $\mathbf{x}_k$ appearing in the GREG full response estimator. We denote by $\mathbf{z}$ the imputation vector and by $\mathbf{z}_k$ its value for element k. The vector $\mathbf{z}_k$ is composed of one or more *imputation variable values*. When $\mathbf{z}$ is univariate, the notation will be z and z_k, respectively. The imputation vector is instrumental in producing the imputed values $\hat{y}_k$. If the imputation variable(s) are strong predictors for the imputed variable y, we can expect "close" imputations, that is, the imputation errors, $\hat{y}_k - y_k$, should be small. All three rules give *deterministic* imputations.

Nearest neighbour and hot deck are *donor-based* methods, which means that we impute a value that was actually observed, but for a different element. The advantage that this brings – perhaps not a very great advantage – is the assurance that the imputed value is one that can occur; it is not an impossible value. Hot deck, as described below, is a *random* imputation method, while nearest neighbour is *deterministic*.

Imputation for qualitative variables merits a special comment. Consider the case of a dichotomous study variable that indicates the presence or absence of a given property, such as "employed" or "unemployed", with values 1 and

0, respectively. In order to meet the requirement of being a value that actually occurs, the imputed value should be either a 1 or a 0. Hot deck and nearest neighbour have the advantage that they satisfy this requirement. In contrast, multiple regression imputation and its special cases will normally impute values other than 0 or 1. For example, a simple procedure is to impute values using the response rate within groups, so that the imputed value is $\hat{y}_k = m_g / n_g$ for all nonresponding elements in group g. These imputed values may in some sense be "good on average", but for any one particular element, the rule produces an "impossible" value. The same holds true when imputation is based on the often used logistic regression model; the imputed value for element k is then of the form $\hat{y}_k = \exp(-\mathbf{z}_k'\hat{\boldsymbol{\beta}})\left[1+\exp(-\mathbf{z}_k'\hat{\boldsymbol{\beta}})\right]^{-1}$, where $\hat{\boldsymbol{\beta}}$ is a parameter estimate based on data available for the elements in the sample. As long as imputation is only used to produce statistics for aggregates, there is, however, no clear-cut disadvantage in imputing "impossible" values.

Imputation by a statistical rule is often carried out mechanically using computer software. An example is Statistics Canada's Generalized Editing and Imputation System (GEIS) software. Such mechanical imputation is often performed within *imputation groups*. These groups have to be identified at the outset. An imputation group is one deemed to consist of "similar elements".

Practitioners often impute according to a *hierarchy of methods*, such that a stronger method (likely to produce "closer" imputations) is first applied within one group of nonrespondents, then, if the auxiliary information required for this preferred imputation is not at hand for all elements, the second strongest method is applied within the next group, and so on.

It is clear that imputation is often motivated by the statistician's perception of a (regression) relationship between the study variable, *y,* and the imputation vector, **z**, used to construct "close" imputed values. From this perspective we now examine the five imputation methods listed previously.

The formulas assume that all elements $k \in o = s - r$ are imputed by the same method, which amounts to saying that there is only one imputation group, the whole sample set s. The case of two or more imputation groups will be discussed later in the section.

(Multiple) regression imputation

The imputed value for element k is

$$\hat{y}_k = \mathbf{z}_k' \hat{\boldsymbol{\beta}} \tag{7.2.4}$$

where $\mathbf{z}_k$ is the value of the imputation vector for element k, and

$$\hat{\boldsymbol{\beta}} = \left(\sum_r q_k \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \sum_r q_k \mathbf{z}_k y_k \tag{7.2.5}$$

Here $\hat{\boldsymbol{\beta}}$ is a vector of regression coefficients, resulting from the fit of a multiple regression using the data $(y_k, \mathbf{z}_k)$ available for $k \in r$, and weighted with suitably specified q_k.

In the special case where $\mathbf{z}_k = (1, z_k)'$, the imputed value takes the form $\hat{y}_k = \hat{\alpha} + \hat{\beta} z_k$, corresponding to the fit of a simple linear regression with an intercept.

Two other important special cases are *ratio imputation* and *respondent mean imputation*.

Remark 7.2.1. An interesting version of multiple regression imputation is GREG-*conformable multiple regression imputation*. We mean by this term that the multiple regression imputation (7.2.4) is specified to agree with the full response GREG estimator (4.3.3) in regard to the weighting and the auxiliary vector (the imputation vector). That is, in formula (7.2.4) for $\hat{y}_k$, the weighting is chosen as $q_k = d_k c_k$ and the imputation vector as $\mathbf{z}_k = \mathbf{x}_k$. When this imputation is used in (7.2.1), the resulting imputed estimator $\hat{Y}_I$ has an interesting property: it is identical to the calibration estimator $\hat{Y}_W$ determined by (6.3.2). This result is rather exceptional, as in general reweighting and imputation do not give identical results. The property is formulated more explicitly as follows:

Consider the imputed GREG estimator (7.2.1) where we use regression imputation according to (7.2.4) with $q_k = d_k c_k$ and $\mathbf{z}_k = \mathbf{x}_k$. Then the resulting imputed GREG estimator $\hat{Y}_I$ is identical to the calibration

estimator $\hat{Y}_W$ given by (6.3.2), that is, $\hat{Y}_I = \hat{Y}_W$ for every possible response set r. (We assume that the same c_k and $\mathbf{x}_k$ are used in both $\hat{Y}_I$ and $\hat{Y}_W$.) A proof of this property is given in Appendix D.

Note that while the GREG estimator is unbiased for full response, the imputed GREG estimator will have some (unknown) bias, even if the regression imputation is GREG- conformable.

GREG-conformable multiple regression imputation is of interest for another important reason: it offers a convenient method for variance estimation. The argument is as follows. For this type of imputation, we have $\hat{Y}_I = \hat{Y}_W$, where $\hat{Y}_W$ is the weighting estimator (6.3.2). It follows that for the imputed estimator $\hat{Y}_I$ we can use the variance estimator presented in Section 6.4 for the reweighting approach. That is, a variance estimator appropriate for $\hat{Y}_I = \hat{Y}_W$ is $\hat{V}(\hat{Y}_W)$ given by (6.4.1) to (6.4.5).

By a similar argument, when the full response estimator is the HT estimator and the imputed values are given by (7.2.4), with $q_k = d_k c_k$ and $\mathbf{z}_k = \mathbf{x}_k$, the imputed estimator is $\hat{Y}_I = \hat{Y}_{Ws}$, where $\hat{Y}_{Ws}$ is given by (6.3.6). It then follows that for this imputed estimator we can use the variance estimator given by (6.4.6) to (6.4.8).

□

Ratio imputation

Assuming that $\mathbf{z}_k = z_k$ is an always positive, unidimensional imputation variable, and that $q_k = 1/z_k$, the imputed value (7.2.4) becomes $\hat{y}_k = z_k\hat{\beta}$, with $\hat{\beta} = \sum_r y_k / \sum_r z_k$. This ratio imputation is often used when the same variable is measured on two different occasions in a repeated survey; y is then the variable on the present survey occasion, and z is the same variable on the preceding occasion. To illustrate, if y and z represent "Gross Business Income" on the present occasion and the preceding occasion respectively, then the "current ratio" $\hat{\beta}$ measures the change in the business income level between the two occasions.

Nearest neighbour imputation

The imputed value for element k is $\hat{y}_k = y_{l(k)}$, where $l(k)$ is the *donor element* for the nonresponding element k, that is, the element that provides its y-value as an imputed value for element k. The statistical idea that motivates this method is that two elements whose z-values are close should also have y-values that are close. The donor for element k is identified by distance minimisation. Assuming a unidimensional imputation variable, z, define the distance from element l to element k as $D_{lk} = |z_\ell - z_k|$. The donor $l(k)$ is the element belonging to the set r such that $\min_{\ell \in r} D_{lk}$ is obtained precisely for $l = l(k)$. That is, the distances D_{lk} are computed for all elements $l \in r$, and the donor element for k will be the one with the minimum distance D_{lk}. For element k, we impute the donor's y-value, that is, we let $\hat{y}_k = y_{l(k)}$. Since $l(k)$ is the closest element (measured by this distance), it is fitting to call it the *nearest neighbour* of k. If the imputation vector is multivariate, we can instead minimise a multivariate distance measure, for example, $D_{lk} = \left(\sum_{j=1}^{J} h_j (z_{jl} - z_{jk})^2 \right)^{1/2}$, where the h_j are specified to give a suitable weighting of the J components of the vector difference $\mathbf{z}_l - \mathbf{z}_k$.

In multiple regression imputation and nearest neighbour imputation, the hope for "close" imputed values rests on a strong relationship between the study variable, y, and the imputation vector $\mathbf{z}$. The next two methods, *respondent mean imputation* and *hot deck imputation*, do not use an imputation variable (or vector) and are therefore "weaker" and less likely to produce close imputations. Neither method is recommended when better alternatives exist. They may be used as "methods of last resort", in the absence of a reasonable imputation variable. They will accomplish at least one of the objectives of imputation, that of creating a completed rectangular data matrix.

Respondent mean imputation

The imputed value for element k is $\hat{y}_k = \bar{y}_r$ for all elements $k \in o$, where $\bar{y}_r = \sum_r y_k / m$. Since all imputed elements will thus receive the same imputed value, the distribution of the completed data will have a rather unnatural appearance, with a spike at $\bar{y}_r$. It is easily seen that respondent

mean imputation is the special case of (7.2.4) obtained when $\mathbf{z}_k = 1$ and $q_k = 1$ for all k.

Hot deck imputation

The imputed value for element k is $\hat{y}_k = y_{l(k)}$, where $l(k)$ is a randomly selected donor from among all potential donor elements $l \in r$. This is a donor-based, random imputation method. The distribution of the values of the resulting completed data set will look rather natural, but may still differ from the visual image obtained from the distribution of a complete sample of actually observed data, $\{y_k : k \in s\}$, if these data had been available. This is because in hot deck imputation, every donor must necessarily be a respondent, and respondents and nonrespondents may be significantly different in regard to characteristics such as mean, variance, etc.

Imputation groups

Imputation is often performed within non-overlapping *imputation groups*, $s_g, g = 1,...,G$, whose union is the entire sample s. Within each imputation group, the imputation is done by one and the same method. When imputation is performed within the group s_g, then s, r and $o = s - r$ are replaced by, respectively, s_g, r_g and $o_g = s_g - r_g$ in the description of the methods above.

We can distinguish two reasons, (a) and (b) in the following, for using more than one imputation group:

(a) Different relationships in different subgroups of the sample. The relationship between y and the imputation vector $\mathbf{z}$ may be deemed different in different subgroups of the sample. For example, if ratio imputation is used, the ratio "sum of y_k" divided by "sum of z_k" may differ appreciably in different parts of the sample, which would suggest ratio imputation within groups. The groups may be formed on the basis of subject matter knowledge, corresponding, for example, to industry category (in a business survey) or age/sex categories (in a social survey).

(b) Limited availability of auxiliary variables for imputation. The imputation variable(s) needed for a certain imputation method may not be available for the entire sample s. This may lead to the use of *several different imputation methods for the same data set* and a *hierarchy of imputation methods*, as we will now explain. Suppose that a strongly related imputation variable z is available, but only for a subset of the sample elements. Then one of the stronger methods, such as regression imputation or nearest neighbour imputation, can be carried out for the group consisting of these elements. If no imputation variable is available for some elements, they may have to be collected into an imputation group treated with a weaker kind of imputation, such as respondent mean or hot deck. In such a hierarchy of imputation methods, the stronger imputation methods are applied first, in one or more groups and for as many nonresponding elements as possible, and the weaker methods are applied to the remaining groups.

Adding a randomly selected residual

Multiple regression imputation and its special case ratio imputation are deterministic methods in that they give the same imputed value if repeated. They can be made stochastic through the addition of a *randomly selected residual*. There may be good reasons for doing this. Then, in the case of regression imputation, the imputed value for element k is $\hat{y}_k = \mathbf{z}_k'\hat{\boldsymbol{\beta}} + e_k^*$, where $\hat{\boldsymbol{\beta}} = \left(\sum_r q_k \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \sum_r q_k \mathbf{z}_k y_k$ as before, and e_k^* is a randomly selected residual from the set of computed residuals $\{e_k : k \in r\}$, where $e_k = y_k - \mathbf{z}_k'\hat{\boldsymbol{\beta}}$. Adding such a residual has the effect of making the completed data set more realistic. A completed data set containing regression imputed values $\hat{y}_k = \mathbf{z}_k'\hat{\boldsymbol{\beta}}$ tends to have less variability than a set of truly observed values y_k. Adding a residual will alleviate this problem.

The technique of adding a randomly selected residual has several potential uses: It can be done (a) for point estimation only, (b) for variance estimation only, or (c) for both. The consequence of (a) is to add variance to the imputed estimator. Case (b) represents a perhaps more important use of the technique. Suppose that the technique is practiced only for variance estimation. The resulting advantage for variance estimation is that the completed data set tends to contain "the right amount of variation". A

"standard formula" may then be used as part of the variance estimation procedure. We return to this issue in Section 7.3.

Remark 7.2.2. *Multiple imputation* has been proposed as a technique for treating nonresponse. As the name suggests, several imputations are made in the same survey data set. So far, we have discussed single-value imputation, that is, every missing value is replaced by one and only one proxy value. By contrast, in multiple imputation, two or more imputations are made for a missing value. It leads to several different completed data sets. Suppose that three such sets are produced. The y_k -values for respondents are the same in all three, but the imputed values (for item nonresponse and/or for unit nonresponse) are different in the three sets. This assumes that a random imputation technique is used, for example, hot-deck imputation. (Deterministic methods such as nearest neighbour and regression imputation do not qualify, because they give one and the same value in repeated attempts, unless the method is suitably modified.) The multiple imputation technique was proposed by Rubin (1978) and is described in considerable detail in Rubin (1987). Multiple imputation is intended both for point estimation and for variance estimation. One of the principal advantages lies in the variance estimation, which becomes very simple, given the existence of several completed data sets. However, in national statistical agencies, multiple imputation has, so far at least, found little use. One reason may be that the method makes heavy demands on data storage space (even though only the imputed values differ from one set to the next). In some countries, notably the United States, multiple imputation is used in "secondary analysis" carried out on survey data by, for example, analysts situated outside the national statistical agency.

□

7.3. Variance estimation when imputation is used

7.3.1. Why the "standard variance formula" is misleading when imputation is used

The survey statistician's responsibilities include both (i) point estimation and (ii) the corresponding variance estimation. The latter task is often more demanding than the first, both in terms of computing time and in the statistical reasoning required to do it correctly. Variance estimation in the presence of imputation is a complex statistical problem. In recent years, considerable research has been put into "correct" variance computation

when imputation has been used. Different "solutions" have been proposed, none of them necessarily optimal. These developments seem to be far from a conclusion, and one can expect the next few years to bring new results. The recommendations in this section are therefore preliminary.

We noted previously that the usual practice in regard to point estimation is to treat the completed data set (7.1.1) as a set of n actually observed values and, consequently, to compute the standard estimator formula (the full response estimator) on these data. When it comes to variance estimation, it is not easy to pinpoint a "good" procedure.

By the *standard formula for variance calculation* we mean the variance estimator formula that accompanies the full response estimator. It is correct to use this formula for variance estimation and confidence interval calculation in the ideal case of 100% response, according to the recipe:

$$\text{point estimate} \pm 1.96 \ (\text{variance estimate})^{1/2}$$

In repeated samples drawn by the given sampling design, this interval will cover the parameter value for roughly 95% of all samples. It is called a *design-based confidence interval,* because it refers to repeated samples drawn with the given design.

Nonresponse brings additional variance over and above the sampling variance. When imputation is used to treat the nonresponse, this is seldom recognised in practice. Some users may argue that the variance increase is small and can be ignored. This may be true for a modest imputation rate – say, 3% – but not for a 30% imputation rate.

At many statistical agencies, the current practice is to treat imputed values as observed values for purposes of variance calculation also. That is, the n values of the completed data set are inserted into the standard formula for variance estimation, in the belief that this will give a sufficiently good indication of the variance of the imputed estimator. This approach of "acting as if the imputed values were perfect substitutes" leads to incorrect variance estimates, for two reasons:

(i) The standard formula for variance, computed on the completed data set, gives a biased estimate of the sampling variance, the bias being usually negative;

(ii) No attempt is then made to estimate the additional variance caused by nonresponse.

Consequently, the computed confidence intervals will be wrong, on average, usually too short. The interval based on the normal score 1.96 will *not* cover the parameter value in roughly 95% of all cases, as is the intent.

The fact that the standard formula gives a misleading indication of the variance of the imputed estimator $\hat{Y}_I$ is illustrated by the following example.

EXAMPLE 7.3.1. "*Right amount of variation" in the completed data?*

Consider the sampling design SRS with the sampling fraction n/N. Suppose first that there is no nonresponse and that we estimate the population total Y by $\hat{Y} = N\bar{y}_s$, where $\bar{y}_s$ is the arithmetic mean of the n sample values y_k. The well-known expression for the sampling variance is $V_p(\hat{Y}) = N^2(1/n - 1/N)S_{yU}^2$, where S_{yU}^2 is the population variance of the N values y_k. An unbiased variance estimator (based on the sample) is $\hat{V}_p(\hat{Y}) = N^2(1/n - 1/N)S_{ys}^2$, where S_{ys}^2 is the variance of the n observed values y_k. Now turn to the situation with nonresponse and imputation (by any suitable method). The full response estimator $\hat{Y} = N\bar{y}_s$ yields the imputed estimator $\hat{Y}_I = N\bar{y}_{\bullet s}$, where $\bar{y}_{\bullet s}$ is the arithmetic mean of the n values $y_{\bullet k}$ of the completed data set. The standard formula for the estimated variance, $N^2(1/n - 1/N)S_{ys}^2$, when computed on the completed data set, gives the result $N^2(1/n - 1/N)S_{y\bullet s}^2$, where $S_{y\bullet s}^2$ represents the variance of the n values $y_{\bullet k}$ of the completed data set. What can be said about $S_{y\bullet s}^2$? It depends on the imputation method whether or not $S_{y\bullet s}^2$ is close to the desired value S_{ys}^2. The completed data set may not contain the "right amount of variation" for this to happen. A case in point is respondent mean imputation (this imputation method is not recommended except as a last resort!), in which case $\hat{y}_k = \bar{y}_r$ for every $k \in o = s - r$, and a simple analysis shows that computation of the standard formula on the completed

data set gives $N^2(\frac{1}{n}-\frac{1}{N})\frac{m-1}{n-1}S_{yr}^2$, where m is the number of respondents and $S_{yr}^2=\sum_r(y_k-\bar{y}_r)^2/(m-1)$. This looks "too small" as an indicator of the sampling variance; $N^2(\frac{1}{n}-\frac{1}{N})S_{yr}^2$ would be more acceptable. If the response mechanism is uniform, then, on average, S_{yr}^2 equals S_{ys}^2, and the standard formula will thus underestimate the sampling variance by a factor of roughly m/n. For example, if the response rate m/n is 70%, the standard formula will underestimate the sampling variance by as much as 30%. Note that we still have not made any attempt to estimate the nonresponse variance; this requires a separate estimation, as we discuss later.

□

7.3.2. The framework for evaluating bias and variance

For a correct assessment of the variance, we need to examine the statistical properties (bias, variance, mean squared error) of the imputed estimator $\hat{Y}_I$. This is done in detail in Appendix A; see also Section 5.2. The total variance of $\hat{Y}_I$ is given by (5.2.3) with $\hat{Y}_{NR}=\hat{Y}_I$, that is,

$$V_{pq}(\hat{Y}_I)=V_{SAM}+V_{NR} \qquad (7.3.1)$$

where $V_{SAM}=V_p(\hat{Y})$ is the sampling variance, discussed in Section 7.3.4, and $V_{NR}=E_pV_q(\hat{Y}_I|s)$ is the nonresponse variance, discussed in Section 7.3.5. Recall that $\hat{Y}$ denotes the full response estimator.

The variance estimator of $\hat{Y}_I$ is given by

$$\hat{V}_{pq}\left(\hat{Y}_I\right)=\hat{V}_{SAM}+\hat{V}_{NR} \qquad (7.3.2)$$

In survey practice, it is important to evaluate each of the two components individually. The information about the relative size of the nonresponse variance may lead to important modifications in the survey design, for example, regarding the amount of resources allocated to nonresponse prevention and treatment.

To illustrate, suppose it is found in a regularly repeated survey that the nonresponse variance accounts for 40% of the total variance. If this proportion seems high, it may be a signal that more resources should be devoted to trying to reduce the nonresponse on the next survey occasion.

How do we compute the two components of (7.3.2)? This question is examined in Sections 7.3.3 to 7.3.6.

For the *sampling variance*, we have noted that when the "standard formula" is computed on the completed data set, the result is often an incorrect indicator of the sampling variance component $V_p(\hat{Y})$. This problem is examined in Sections 7.3.3 and 7.3.4.

For the *nonresponse variance*, most statistical agencies presently lack a well-developed practice. We consider methods for estimating the nonresponse variance in Sections 7.3.5 and 7.3.6.

The approaches to variance estimation in the presence of imputed values are reviewed in Lee, Rancourt and Särndal (2000, 2001). For further detail, the reader is referred to these two sources and to the many references given there.

7.3.3. The use of standard software for variance calculation

The total estimated variance of the imputed GREG estimator $\hat{Y}_I$ given by (7.3.2) is a sum of two components, one for the sampling variance and one for the nonresponse variance. In this section and in Section 7.3.4 we comment on the sampling variance component. A guiding principle is the desire to profit as far as possible from already existing software for variance estimation. A number of national statistical agencies now use specially designed software to carry out point estimation and variance estimation, for example, Statistics Sweden's CLAN97 and Statistics Canada's GES. Such software is equipped to handle estimation in the case of 100% response, but cannot always be directly applied to situations with survey nonresponse.

Now, if the nonresponse is treated by imputation, and if no special arrangements are made, these software packages will act as if imputed values were as good as truly observed values. They will compute a point estimate and a variance estimate from the completed data set, using the ready-programmed formulas, which we refer to as "standard variance formulas".

The standard variance formula computed on the completed data set is usually an invalid estimate of the *sampling variance* of the imputed estimator. Furthermore, neither CLAN97 nor GES is presently equipped to calculate the *nonresponse variance* component generated by imputation.

7.3.4. Estimating the sampling variance component

The objective in this section is to carry out a "correct" computation of the estimated sampling variance component V_{SAM} in (7.3.2).The procedure should, on average, yield a correct level for V_{SAM} in (7.3.1). A simple but often incorrect approach is to calculate the standard variance estimation formula on the completed data set, using the readymade programme in a software package such as CLAN97 or GES. For some imputation methods, this will indicate the wrong level for the sampling variance. (Consequently, it is even more in error as an indicator of the total variance, if no attempt is made to account for the nonresponse variance.)

Two approaches to estimating the sampling variance V_{SAM} are:

(i) use only the available observed y-values y_k for $k \in r$,
(ii) use the completed data set $y_{\bullet k}$ for $k \in s$, consisting of real observations and imputed values.

Here (ii) is more attractive than (i), because the $n - m$ imputed values, although artificial and non-observed, will often contain important additional information, supplementing the information contained in the m truly observed y-values y_k, especially when the imputed values are created with the aid of a powerful predictor variable $\mathbf{z}_k$. Although the imputed values are not perfect substitutes, they are usually better than a total absence of data on the $n - m$ nonresponding elements.

Another important reason for focusing on alternative (ii) is the desire to benefit from existing software. As mentioned, CLAN97, GES and similar software contain a "standard variance formula" for computing the sampling variance calculation of the GREG full response estimator. Considerable effort may be saved if we can directly insert data into such a standard variance formula and obtain a correct indication of the sampling variance component. It depends on the imputation method (or methods, if more than one is used in the same survey) whether the standard variance formula will yield a correct level when the completed data set $\{y_{\bullet k} : k \in s\}$ is inserted into it. For some of the usual imputation rules this is not the case. The variability in the completed data set may be insufficient, with underestimation as a result.

The literature suggests two directions for correcting this problem. Both use the standard variance formula, but in different ways. They are:

(a) to amend the standard variance formula by adding a suitable correction term;

or

(b) to amend the completed data set (7.1.1) used to produce the point estimate $\hat{Y}_I$, so that the amended completed data give a "correct" level for the sampling variance when inserted into the standard variance formula.

We focus here on alternative (b), because of the desirability of using available software. Note that in (b), the amended completed data set is used for variance calculation only (not for point estimation).

The exact nature of the amendment depends on the imputation method. The following procedures can be applied with some confidence to obtain an approximately correct level for the sampling variance when the GREG is used as the full response estimator. The recommendations are tentative and are likely to be subject to further study in the near future.

(Multiple) regression imputation, including ratio imputation

For these types of imputation, the completed data set contains too little variability for the standard variance formula to correctly reflect the sampling variance. Therefore, for every element $k \in o = s - r$, carry out the following

procedure. To the imputed value $\hat{y}_k$ given by (7.2.4), add a regression residual, randomly selected from the computed set of residuals $\{e_k;\ k \in r\}$, where $e_k = y_k - \mathbf{z}_k'\hat{\boldsymbol{\beta}}$. Let the randomly chosen residual for element k be e_k^*. The amended completed data set is thus made up of the values y_k for $k \in r$ and $\hat{y}_k + e_k^*$ for $k \in o$. Then compute the standard variance formula on this amended data set composed of n values.

Nearest neighbour imputation

The situation for nearest neighbour imputation is the opposite, compared to regression imputation, in that the completed data set tends to contain too much variability. No amendment is needed. Insertion into the standard variance formula will instead overestimate the sampling variance somewhat. The overestimation is usually modest, unless the nonresponse rate is high. The standard variance formula will thus give a variance estimate that is somewhat on the conservative side.

7.3.5. Approaches to estimating the nonresponse variance

The objective is to construct a (computable) measure, $\hat{V}_{NR}$, of the nonresponse variance component $V_{NR} = E_p V_q(\hat{Y}_I \mid s)$ of the total variance $V_{pq}(\hat{Y}_I)$ given by (7.3.1). The formula for $\hat{V}_{NR}$ will depend on the imputation method used. This is easy to understand, since imputation methods are more or less accurate. The size of the nonresponse set, n - m, will influence the size of $\hat{V}_{NR}$: the more we impute, the greater, normally, the nonresponse variance.

The dependence of $\hat{V}_{NR}$ on the imputation method is an inconvenience in that a new formula must be worked out for every imputation method. The formula also depends on the sampling design in use. These inconveniences are particularly pronounced when more than one imputation method is used in the same survey.

Two principal methods exist for constructing the measure $\hat{V}_{NR}$ of the nonresponse variance:

(i) the two-phase approach;
(ii) the model-assisted approach.

The two-phase approach relies on the two distributions that we associate with "sampling followed by nonresponse", namely, the (known) sampling design $p(s)$, and the (unknown) response mechanism $q(r|s)$, as discussed in Section 5.2 and Section 6.1.

The model-assisted approach, in turn, also relies on two distributions: the sampling design distribution $p(s)$, and what is known as the *imputation model* distribution. To explain the latter, we note that when the statistician imputes by a specified method, he/she refers, implicitly or explicitly, to a hypothetical relationship between the study variable y and a vector of predictor variables $\mathbf{z}$. We refer to this relationship as the imputation model. It underlies the construction of the imputed values $\hat{y}_k$, and it plays a crucial role in deriving the nonresponse variance component in the model-assisted approach. On the other hand, the assumptions regarding the response mechanism $q(r|s)$ are minimal in the model-assisted approach.

The underlying imputation model ξ is particularly evident in the case of regression imputation, as defined in Section 7.2. The simple regression model $y_k = \alpha + \beta z_k + \varepsilon_k$ lies behind the univariate regression imputation $\hat{y}_k = \hat{\alpha} + \hat{\beta} z_k$. Ratio imputation corresponds to the special case of this model obtained for $\alpha = 0$. The multiple regression model $y_k = \mathbf{z}_k' \boldsymbol{\beta} + \varepsilon_k$ lies behind the multiple regression imputation $\hat{y}_k = \mathbf{z}_k' \hat{\boldsymbol{\beta}}$. In these models, ε_k is a random error term, about which assumptions are made: $E_\xi(\varepsilon_k) = 0$ for every k and $E_\xi(\varepsilon_k^2) = \sigma^2 f(\mathbf{z}_k)$, where $f(\cdot)$ is a specified function. The unknown model parameter σ^2 must be estimated, and this is done by seeing how well the imputation method succeeds in predicting the study variable values y_k for the *responding* elements. For these elements, y_k is observed, and by comparing observed and imputed values, we can measure how prone the imputation method is to correctly predict the study variable.

Nearest neighbour imputation, defined and discussed in Section 7.2, is also motivated by an underlying model, though this may be less apparent because there is no explicit model fitting, as in regression imputation. Nevertheless, when the statistician argues, as in nearest neighbour imputation, that the donor element $l(k)$ has a y-value "close" to the missing value y_k, he is clearly thinking in terms of a relationship between y and the imputation variable z used to identify the donor such that if $z_{l(k)}$ is close to z_k, then $y_{l(k)}$ should also be close to the missing value y_k. When the imputation vector is multivariate, an appropriate model for nearest neighbour imputation is $y_k = \mathbf{z}_k'\boldsymbol{\beta} + \varepsilon_k$. In the univariate case, either $y_k = \alpha + \beta z_k + \varepsilon_k$ or $y_k = \beta z_k + \varepsilon_k$ may be appropriate.

7.3.6. Expressions for the nonresponse variance estimate in some special cases

Here we consider one example to illustrate the model-assisted derivation of the estimated nonresponse variance, $\hat{V}_{NR}$. For other cases, the reader can develop the appropriate formula to suit the particular conditions of his own survey. The formula $\hat{V}_{NR}$ will depend on the following factors: (i) the survey specifications (the sampling design, the full response estimator that underlies the imputed estimator), (ii) the imputation method in use, and (iii) the assumed imputation model.

We consider the following survey conditions: *Sampling design*: SRS with n elements drawn from N; *Estimator*: the imputed GREG estimator $\hat{Y}_I = \sum_s d_k g_k y_{\bullet k}$ given by (7.2.1), with the weights $d_k = N/n$ and where g_k is given by (4.3.4); *Imputation model*: $y_k = \beta z_k + \varepsilon_k$ with $E_\xi(\varepsilon_k) = 0$; $E_\xi(\varepsilon_k^2) = \sigma^2 z_k$; $E_\xi(\varepsilon_k \varepsilon_l) = 0$ for all $k \neq l$, where the subscript ξ denotes expectation under the imputation model. Note that the imputed GREG estimator uses an auxiliary vector, **x**, which may or may not include the variable z used in imputation. Under these conditions, we examine: (i) Ratio imputation and (ii) Nearest neighbour imputation.

(i) Ratio imputation with the imputed value $\hat{y}_k = z_k\hat{\beta}$ for $k \in o$, where $\hat{\beta} = \sum_r y_k / \sum_r z_k$: The model-assisted approach, based on the imputation model, leads to

$$\hat{V}_{NR} = \left(\frac{N}{n}\right)^2 \left\{ \frac{(\sum_o g_k z_k)^2}{\sum_r z_k} + \sum_o g_k^2 z_k \right\} \hat{\sigma}^2 \qquad (7.3.3)$$

where $\hat{\sigma}^2$ is a model unbiased estimator of the imputation model parameter σ^2 given by

$$\hat{\sigma}^2 = C_r \frac{\sum_r e_k^2}{\sum_r z_k} \qquad (7.3.4)$$

with $C_r = (\frac{m}{m-1})\left(\frac{1}{1-(cv_{zr})^2/m}\right)$ where cv_{zr} equals the standard deviation of z in r, divided by the mean of z in r, and $e_k = y_k - z_k\hat{\beta}$. A special case is the HT-estimator, which becomes $N\bar{y}_s$ under SRS. Its estimated nonresponse variance, obtained by setting $g_k = 1$ in (7.3.3), is given by

$$\hat{V}_{NR} = N^2(1/m - 1/n)(\bar{z}_o\bar{z}_s/\bar{z}_r)\hat{\sigma}^2$$

where $\bar{z}_o$, $\bar{z}_s$ and $\bar{z}_r$ are means over the indicated sets, and where $\hat{\sigma}^2$ is given by (7.3.4).

(ii) Nearest neighbour imputation: The imputed value is $\hat{y}_k = y_{l(k)}$, where $l(k)$ is the donor element such that the minimum of the distance $|z_l - z_k|$, over the potential donors $l \in r$, occurs for $l = l(k)$. The model assisted approach produces the following estimator of the nonresponse variance component:

$$\hat{V}_{NR} = [\sum_o (d_k g_k)^2 z_k + \sum_r S_l^2 z_l]\hat{\sigma}^2 \qquad (7.3.5)$$

where $S_l = \sum_{o_l} d_k g_k$ with $o_\ell = \{k: k \in o \text{ and } k \text{ uses } l \text{ as donor}\}$ and $\hat{\sigma}^2$ is given by (7.3.4). It is seen from this expression that the multiple utilisation of the same donor has a tendency to increase the nonresponse variance. One can show that the nonresponse variance is about twice as large for nearest neighbour imputation as for ratio imputation.

Remark 7.3.1. As Remark 7.2.1 points out, when the GREG-conformable multiple regression imputation is used, $\hat{Y}_I$ is identical to $\hat{Y}_W$ and the variance estimator $\hat{V}\left(\hat{Y}_W\right)$, given by (6.4.1), can be used.

□

7.4. When is imputation allowed?

Imputation may not always be possible because of legal restrictions. In some countries imputation is prohibited by law, at least for certain types of population elements. The present situation in Sweden, as in all Member States of the European Union, is as follows. In a data file on individuals, the insertion of any value that is not a "true observation" is disallowed. However, if the Personal Identity Number (see Example 2.2.1), or other unique person identifier, is suppressed in the file, then imputation is permitted.

For business enterprises, however, the legal restriction applies only to those enterprises that are identified by the Personal Identity Number. For most business enterprises in a typical business register, the identifier is not tied to the Personal Identity Number. Consequently, for such elements imputation is allowed and used in many Swedish business surveys.

The law relates to "personal data", a term defined in the Directive 95/46/EC issued in 1995 by the European Community: "… 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."

The same Directive states that personal data must be: "...(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified."

The Swedish Data Inspection Board (1974), abbreviated DI for "Datainspektionen", declares that imputation is contrary to the law since imputed values are known at the outset to be proxies rather than true values. DI is also of the opinion that imputation impinges on personal integrity; the more sensitive the variable, the more integrity is affected.

In this CBM we make a distinction between *imputed value* and *derived value*. Following the wording of the law, we mean by imputed value one that is inserted in a data file, is not observed, and is therefore unlikely to be exact or "true". A derived value, on the other hand, may also be inserted, but it is a true value. An example arises when the true sum of the values of a set of variables is available, and when individual, true variable values are available for all but one variable, whose true value is obtained by subtraction and then inserted into the data file. This situation may occur in economic statistics, with data derived from financial statements.

Sometimes, derived values are determined for every element in a realised sample, and the result is referred to as a *derived variable*. There are two types of derived variable: (i) a variable constructed by the statistician and such that there is no need to explain its meaning to users outside the agency, and (ii) other derived variables. To the first type belong, for example, the design weight, a calibrated weight, a constructed index. All of these arise from a specified formula whose value is computed for each sample element; hence they are derived variables. The second type includes, for example, a taxable income derived according to a formula that produces a value of taxable income, as a function of other variables with known values.

Statistics Sweden's interpretation of the law is that the possibility of imputing for identified individuals still exists, namely, if this is done in such a way that no imputed values appear on the final output data file and if imputation and tabulation occur in the same production step (within a few seconds and using the same software). Sometimes new tabulations from the file may be required later, and the need may then arise for imputations for the missing values in the file. In order to reproduce the original imputed values we must then assume that a deterministic imputation technique is used.

In most surveys, several preliminary versions of the observation data file are created in the production process. All these files are scrapped when the final

output file is produced. The interpretation of Statistics Sweden is that the law allows the use of imputed values in these preliminary data files.

It must be emphasised that, in any case, imputation is not carried out for the purpose of infringing the rights of whatever elements are involved. Imputation is a tool used by the statistical agency to enhance the quality of the estimates, which is in the interest of the nation and its citizens.

8. Comparing reweighting and imputation: Which is preferable?

8.1. Introduction

In earlier chapters we presented two approaches to the treatment of nonresponse, reweighting and imputation. So far little has been said about which approach is preferable. Guidelines for choosing between them are proposed in this section.

Both practical and statistical considerations influence the choice. Examples of statistical considerations are: which of the two approaches is better for reducing the nonresponse bias? Which is better for reducing the variance? Variance estimates are usually also required in the survey, so a practical consideration is whether computer software is available for the computation of variance estimates. Another practical matter is that legal restrictions may have a bearing on the choice, as discussed in Section 7.4.

The current situation at Statistics Sweden is that imputation is mainly limited to business surveys. In some of these surveys, imputation is limited to item nonresponse (ITIMP-approach), in others, imputation is used both for unit nonresponse and item nonresponse (UNIMP-approach). The prevailing practice for imputation is that of a hierarchy of methods, as discussed under the heading of Imputation Groups in Section 7.2.3.

Reweighting has long been the dominant approach in Statistics Sweden's surveys on individuals and households. One reason for this is that imputation is not allowed for data on individuals, as discussed in Section 7.4. Consequently, if imputation were to be freely used for a data set on individuals, then the person identifier would first have to be suppressed, which may severely reduce the usefulness and quality of the resulting completed data set.

Within each of the two main approaches, reweighting and imputation, there is considerable variety of choice. Thus, even after a choice between reweighting and imputation has been made, the question remains of how to choose within the set of available alternatives. In reweighting, the question

arises as to which auxiliary variable(s) should be chosen for the calibration. In imputation, the question is which variables should be used as imputation variables, and, secondly, which of the several available imputation methods should be used.

8.2. Practical considerations

The recommended approach for Statistics Sweden's surveys, in particular for surveys on individuals and households, continues to be reweighting. A recent development is that reweighting can now be carried out in a standardised fashion through the calibration technique described in Chapter 6. Another strong reason for reweighting is that calibration is easily carried out using Statistics Sweden's software CLAN97. This computes the calibrated point estimate for a population total or for the total of any specified domain. It also permits the estimation of any parameter that can be expressed as a rational function of such totals. The corresponding variance estimates are also computable with CLAN97. This software can handle the majority of the sampling designs in current use at Statistics Sweden. A significant portion of these sampling designs are variations of STSRS (see Section 2.1). It is true that from a more international perspective, this range of designs may appear limited.

A prerequisite for a "good" calibration estimator is the availability of powerful auxiliary variables. In some surveys at least, there may be a good supply, even an abundance, of auxiliary variables from which to choose. Some guidelines for this choice are given in Chapter 10, where the main objective is to reduce as far as possible the nonresponse bias, as given by the general expression (10.2.1).

At Statistics Sweden the software support for imputation is less developed than that for reweighting. In many cases, the user will have to construct his own program for the computation of imputed values. However, this task is often not very demanding. But the computation of a variance estimate for an imputed estimator is not a trivial matter, as Section 7.3.6 indicated. The formulas depend on the imputation method(s) in use in the survey. Explicit formulas for such variance estimates were given in Section 7.3.6 for a few cases only. This means that, as a preliminary step, the user will often be required to perform the mathematical derivation of the appropriate formula. Furthermore, the variance estimation becomes highly complex when several

statistical imputation methods are used in the same survey, perhaps complemented with expert judgment imputation as well, for instance, for large elements.

It should be noted that variance estimation for imputed estimators starts from the notion that a set of elements are in some respect similar. For example, if elements in a specified group are considered similar, the imputation for nonresponding elements in this group can be justified on the assumption that these nonrespondents are "typical members" of the group in question. The similarity pattern, expressed for example as a linear regression through the origin, is estimated from a perhaps considerable number of responding elements in the group. The situation is very different when one single element, say a very large element, is imputed by expert judgment. Then there is no reference group of "similar elements", and therefore no basis for variance estimation. An "easy alternative" is to treat some or all imputed values as "true values", or sufficiently close to true values, but as Section 7.3.1 warned, this may lead to a considerable underestimation of the variance.

The difficulty with variance estimation disappears in one notable case, namely, when GREG-conformable multiple regression imputation is used. The imputed GREG estimator under this imputation method is, as we noted in Remark 7.2.1, identical to the calibration estimator based on similar conditions. Thus the variance estimation procedure for the latter estimator applies; it is simple and is given in Section 6.4. The identity in question is, however, limited to the estimate of the entire population total. For the estimation of the total for a domain, the problem persists; no easy variance estimator is available.

Our discussion of variance estimation under imputation was limited to one illustration of the mathematical technique leading to the variance estimator. The state of the art in regard to variance estimation for imputed estimators is, at this point in time, rather fragmentary. We recommend using the sum of the two terms $\hat{V}_{SAM}$ and $\hat{V}_{NR}$ as given in (7.3.2). The expressions for these two components would have to be derived by the user for each particular imputation method.

8.3. Statistical considerations

To reduce the bias of the estimates as far as possible is the principal objective of nonresponse treatment methods. Compared to this, the objective of realising a small variance comes second only. The basis of this reasoning is the Mean Squared Error. When there is nonresponse, the Mean Squared Error is, at least for large samples, usually dominated by the squared bias term, and the variance term is often small by comparison. This holds for both reweighting and imputation.

In the course of writing this CBM, we carried out some simulations in order to compare reweighting (using the calibration approach) with imputation (using several alternative imputation methods). The simulations involved a single auxiliary variable, x, which was used both for calibration and for imputation. We drew repeated SRS samples of size $n = 200$ from a population of size $N = 1100$. Here we present only a brief summary of the findings.

(i) In estimating of the total for the entire population, calibration and deterministic imputation (see Section 7.1.1) gave very similar results for both the bias and the variance. The squared bias was considerably greater than the variance. The nearest neighbour imputation estimator had a larger variance compared both to other imputation methods and to the calibration estimator. This was expected.

(ii) In estimating the total for a domain, a clear contrast emerges between calibration and imputation. In most cases, the nonresponse bias is smaller for imputation than for calibration. The variance is also usually smaller with the imputation techniques, especially for the smaller domains. Therefore, in estimation for domains, imputation appears to have clear advantages over calibration, as far as the Mean Squared Error is concerned.

(iii) Among the imputation techniques compared, the nearest neighbour method appears to be the one that yields the smallest bias.

(iv) Grouping is an efficient means of reducing the nonresponse bias. Also, it is efficient to include an intercept (a constant "1") in the definition of the imputation vector. In other words, (simple) regression imputation is preferable to ratio imputation.

In practice, there is often more than one x-variable available for use as auxiliary variable(s) for calibration or imputation. Choosing the best one(s) then becomes a question for both calibration and imputation. Some recommendations for this choice are given in the case of calibration in Chapter 10.

A disadvantage attaching to some imputation methods is that they may distort the distribution of a study variable or the relationship between two or more study variables. When the procedure is to impute by the overall respondent mean, it is strikingly obvious that the resulting completed data set will have an "unnatural" distribution, since a perhaps considerable number of elements will be assigned one and the same value, namely, the respondent mean. This disadvantage is somewhat reduced when imputation is instead by the respondent mean within groups, or by multiple regression imputation. However, even a multiple regression imputation based on several x-variables tends to yield an completed data set with unnaturally low variability, compared to a data set with 100% response. Of the methods that we have discussed, nearest neighbour imputation seems to be the least susceptible to this drawback. That is, it comes closest to rendering a natural distribution and a natural variability in the completed data set.

The relationships between variables are also likely to be more or less perturbed by imputation. As a result, a regression analysis or other type of analysis can give misleading results, compared to the same analysis carried out for the ideal case with 100% response on all variables involved. Again, nearest neighbour imputation is likely to be the method offering the best protection against this disadvantage.

Finally, it should be pointed out that, whatever the method used, the data set after imputation should be subjected to the usual checks for internal consistency. That is, all imputed values should be subjected to the editing checks normally carried out for the survey in question.

9. The treatment of item nonresponse

Most surveys are affected by both item nonresponse and unit nonresponse. In the preceding chapters we have examined the calibration approach and the imputation approach as two possible ways of treating unit nonresponse. The question now arises how the item nonresponse should be dealt with. We discuss several scenarios. All of these are motivated by the desire to create a rectangular data matrix. A unique set of weights can then be applied to all study variables.

A. Reweighting (by the calibration approach) as described in Chapter 6 is used following a treatment of item nonresponse by one of the following methods:

A1. Values missing because of item nonresponse are imputed, using one or more of the methods reviewed in Chapter 7. This becomes the ITIMP-approach, in the terminology of Section 3.1. The result is a rectangular data matrix, to which we can apply the calibration approach for reweighting. Every y-variable in the survey will have values recorded (by observation or by imputation) for every element k in the response set r. The difference between the imputed value and the true value is here regarded as a measurement error. The extent of the item nonresponse is assumed to be relatively small.

A2. Information is discarded in such a way that all study variable values observed for item nonresponse elements are ignored (sometimes called "amputation"). No imputation occurs. In this case, too, the result is a rectangular data set. This voluntary sacrifice of data can in some cases cause a substantial loss of information. The technique cannot be recommended unless the item nonresponse set is small. As noted in Section 8.3, imputation may distort the distribution of a study variable or the relationship between two or more study variables. This will cause particular problems when a regression or some other model is fitted. Amputation may then be an alternative to imputation.

A3. For a categorical study variable a special "data missing" (DM) category is created, in addition to the "real" categories. The DM category is then

treated as any “real” category. This procedure will tend to underestimate the population frequencies for the “real” categories. Usually, in surveys at Statistics Sweden the parameters of interest are proportions rather than frequencies. The usual procedure is then to treat the estimation of the proportion of elements in a given category as a procedure for estimating a domain mean. The item nonresponse elements (that is, the set of elements assigned to the DM category) are regarded as not belonging to the domain. The estimated proportion will have the structure of “weighted estimate of the number of responding elements belonging to a given 'real' category” divided by “weighted estimate of the total number of responding elements”.

B. Imputation is used for the item nonresponse as well as for the unit nonresponse. One or more of the methods in Chapter 7 can be applied. This creates a completed data matrix with n completed records, where n is the size of the whole sample s.

10. Selecting the most relevant auxiliary information

10.1. Discussion

In many surveys, extensive auxiliary information may be readily available from registers and other reliable sources. For example, Statistics Sweden's registers on individuals contain variables such as address, sex, age, income, occupation and education. Additional information may come from matching with other registers. Altogether, these variables provide a valuable source of information. The question then arises as to how one should go about selecting the most relevant part of the total available information, since all of it may not necessarily be used.

Auxiliary information can be used both at the design stage (in constructing the sampling design) and at the estimation stage (in constructing the estimators). This CBM is about estimation, so we concentrate on the second type of usage.

Remark 10.1.1. All available register variables, except for sensitive variables, may be used in the nonresponse analysis, in the computation of weights and in the construction of imputed values. *Sensitive variables* are defined by Swedish law as those which relate to the following conditions: (i) race or ethnic extraction; (ii) political orientation; (iii) religious or philosophical orientation; (iv) trade union membership; (v) health and sexual orientation. Sensitive variables are relatively rare in registers, so no serious restrictions arise for the applicability of the methods advocated in this CBM. In the following we use the term "available auxiliary variable" to mean a variable that is *both* available *and* allowable for use under Swedish law.

□

Chapter 6 has shown that the calibration approach to reweighting is highly flexible in its use of auxiliary information. The other approach, imputation, is in some ways even more flexible in that it allows more than one imputation method to be used in the same survey; see Section 7.2.3. However, even if technically feasible, the use of all available auxiliary

information is not necessarily the preferred solution. Professional judgment must be exercised in selecting the auxiliary information that will finally be used.

The MSE measures the accuracy of an estimate. It consists of the sum of the variance and the squared nonresponse bias. The latter term is likely to dominate the MSE, at least when the sample size is fairly large. A sizeable nonresponse bias has several negative consequences. Neither a good variance estimate nor a valid confidence interval can be derived. These are serious drawbacks, because the survey results lose some of their value if they cannot be accompanied by valid confidence statements.

The overriding objective is therefore to reduce the nonresponse bias as far as possible. One should try to identify variables that meet this objective. Usually, however, it is impossible to assess the nonresponse bias, and subjective judgments become necessary. Some approaches are suggested in Section 10.2.

Reweighting sometimes implies a trade-off between nonresponse bias and variance: for the reduction of nonresponse bias, it is desirable to use as many auxiliary variables as possible, yet this course of action is not necessarily the best for realising the smallest possible variance.

There is ample evidence in the literature that the choice of auxiliary information has a considerable impact on both the sampling variance and the nonresponse bias. A literature search on the choice of auxiliary information reveals two different types of articles: on the one hand, those that emphasise sampling error reduction, and on the other hand, those that emphasise nonresponse bias reduction. Relatively few articles address both aspects jointly.

There is a need for guidelines and methods for selecting “the best” from a larger set of potential auxiliary information. We discuss such guidelines in Section 10.2. Section 10.3 offers a review of the literature on the importance of different kinds of auxiliary information.

10.2. Guidelines

10.2.1. Introduction

Example 6.3.1 showed that in the unlikely situation where a perfect linear relationship exists between the study variable and the auxiliary vector, then the calibration estimator $\hat{Y}_W$ provides an exact estimate of Y. Thus we can expect that if powerful, although less than perfect, auxiliary information can be identified and used, then both the sampling error and the nonresponse bias will be small. More specifically, one should select an auxiliary vector that satisfies, as far as possible, one or both of the following principles:

(i) the auxiliary vector explains the variation of the response probabilities

(ii) the auxiliary vector explains the variation of the main study variables.

A third principle to take into account is that

(iii) the auxiliary vector should identify the most important domains.

When principle (i) is fulfilled the nonresponse bias is reduced in the estimates for all study variables. However, if only principle (ii) is fulfilled the nonresponse bias is reduced only in the estimates for the main study variables. Then the variance of these estimates will also be reduced. Example 4.5.3 showed that the residuals are likely to be considerably smaller if the auxiliary vector can be formulated to identify the principal domains. This is the motivation behind (iii). In the following we examine how the first two principles help in reducing the nonresponse bias of the calibration estimator $\hat{Y}_W$.

Appendix C gives the following general expression for the nonresponse bias, valid for large response sets:

$$B_{pq}(\hat{Y}_W) \approx -\sum_U (1-\theta_k) E_{\theta k} \qquad (10.2.1)$$

where $E_{\theta k} = y_k - \mathbf{x}_k' \mathbf{B}_\theta$ and $\mathbf{B}_\theta = \left(\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k c_k \mathbf{x}_k y_k$.

It is shown in Appendix C, Proposition C.2, that the nonresponse bias given by (10.2.1) will be zero if a certain relation exists between the response probability θ_k and the auxiliary vector $\mathbf{x}_k$, namely,

$$\theta_k^{-1} = 1 + c_k \boldsymbol{\lambda}' \mathbf{x}_k \quad \text{for } k \in U \tag{10.2.2}$$

where $\boldsymbol{\lambda}$ is a column vector independent of k. (If (10.2.2) holds, the right hand side of (10.2.1) is zero, and the bias $B_{pq}(\hat{Y}_W)$ is thus approximately zero. For simplicity, we will use the phrases "the bias is zero" and "the bias is eliminated", although they may hold only in an approximate sense.)

Formula (10.2.1) simplifies when c_k are chosen to be of the form $c_k = 1/\boldsymbol{\mu}'\mathbf{x}_k$. The expression is then

$$B_{pq}(\hat{Y}_W) \approx \sum_U \mathbf{x}_k' \mathbf{B}_{\theta E} \tag{10.2.3}$$

where

$$\mathbf{B}_{\theta E} = \left(\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k c_k \mathbf{x}_k E_k \tag{10.2.4}$$

and $E_k = y_k - \mathbf{x}_k' \left(\sum_U c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U c_k \mathbf{x}_k y_k$

It follows from (10.2.1) and (10.2.3) that $B_{pq}(\hat{Y}_W) \approx 0$ when the ideal conditions of Example 6.3.1 hold, because then $E_{\theta k} = 0$ for all k and $E_k = 0$ for all k. It can also be shown that the variance of $\hat{Y}_W$ is a function of the residuals E_k and that for many sampling designs, a reduction of the residuals will reduce the variance; see Section 4.4. Consequently, an auxiliary vector that explains the variation of the study variable is effective in reducing the MSE.

10.2.2. Analysis of the nonresponse bias for some well-known estimators

In Section 6.6 we discussed several special cases of the general calibration estimator, corresponding to different formulations of the auxiliary vector $\mathbf{x}_k$ and the factor c_k. Here we revisit these examples with the purpose of

showing how the theoretical results (10.2.1) to (10.2.3) can guide the selection of relevant auxiliary information. For simplicity, these examples assume the SRS design and a single quantitative variable x_k.

Six well-known estimators were discussed in Section 6.6: the expansion estimator (EXP), the poststratified estimator (PST), the weighting class estimator (WCE), the ratio estimator (RA), the regression estimator (REG), the separate ratio estimator (SEPRA) and the separate regression estimator (SEPREG). They are obtained from the general calibration estimator (6.3.2) under the specifications of $\mathbf{x}_k$ and c_k given in Table 10.2.1.

Table 10.2.1. The specifications of the auxiliary vector $\mathbf{x}_k$ and the factor c_k leading to well-known estimators. The notation is explained in Section 6.6.

Estimator	Auxiliary vector $\mathbf{x}_k$	Factor c_k
EXP	1	1
PST and WCE	$(\gamma_{1k},...,\gamma_{pk},...,\gamma_{Pk})'$	1
RA	x_k	x_k^{-1}
REG	$(1, x_k)'$	1
SEPRA	$(x_k\gamma_{1k},...,x_k\gamma_{pk},...,x_k\gamma_{Pk})'$	x_k^{-1}
SEPREG	$(\gamma_{1k},...,\gamma_{pk},...,\gamma_{Pk}, x_k\gamma_{1k},...,x_k\gamma_{pk},...,x_k\gamma_{Pk})'$	1

Let us examine how well these estimators satisfy the first two principles given in Section 10.2.1. We start with the first principle, that is, *(i) the auxiliary vector should, as far as possible, explain the variation of the response probabilities*. For each of the six cases in Table 10.2.1, we insert the specifications of $\mathbf{x}_k$ and c_k into (10.2.2) and we obtain the results in Table 10.2.2.

The thought process is then as follows: assume that we know the value of θ_k for every k, and that we examine the set of N points $\{(\theta_k^{-1}, u_k) : k = 1,...,N\}$, where $u_k = 1 + c_k\boldsymbol{\lambda}'\mathbf{x}_k$, in order to see how closely θ_k^{-1} agrees with u_k. (In practice the θ_k are unknown, so the procedure is purely hypothetical.) If the relationship is perfect, so that u_k equals θ_k^{-1}

for every k, then the nonresponse bias is totally eliminated. This perfect relationship is stated in the second column of Table 10.2.2, and is further described in the third column.

Table 10.2.2. The relationship between θ_k^{-1} and u_k needed to eliminate the nonresponse bias for six well-known estimators. a, a_p, b, b_p denote constants.

Estimator	Form of the θ_k^{-1} needed to eliminate bias	Description of the θ_k^{-1} needed to eliminate bias
EXP	$\theta_k^{-1} = a$ for all $k \in U$	constant throughout
PST and WCE	$\theta_k^{-1} = a_p$ for all $k \in U_p$	constant within groups
RA	$\theta_k^{-1} = a$ for all $k \in U$	constant throughout
REG	$\theta_k^{-1} = a + bx_k$	linear in x_k
SEPRA	$\theta_k^{-1} = a_p$ for all $k \in U_p$	constant within groups
SEPREG	$\theta_k^{-1} = a_p + b_p x_k$	linear in x_k within groups

We can now ask: which of the six estimators in Table 10.2.2 is likely to succeed best in coming close to a zero nonresponse bias? The table shows that the nonresponse bias for the EXP and the RA estimator is eliminated if the response probabilities are constant throughout the whole population. This is highly unlikely to happen. Many studies have shown that the response probability varies with observable factors such as age, sex and others. The situation is much more favourable when grouping is involved, as for the PST, WCE and SEPRA estimators. The response probabilities need then "only" be constant for all elements within a group.

In Statistics Sweden's surveys, the auxiliary variables are almost always derived from registers that comprise the entire target population. Thus, it is realistic to assume that x_k is a known value for all $k \in U$. Consequently, we can calculate both the auxiliary population total $\sum_U x_k$, required for the RA estimator, and the population count N. The REG estimator requires knowledge of both $\sum_U x_k$ and N. Table 10.2.2 shows that the REG estimator has a near-zero nonresponse bias if $\theta_k^{-1} \approx a + bx_k$ for some

constants a and b. This condition is more likely to hold than $\theta_k^{-1} \approx a$, which is the condition needed for the RA estimator (and for the EXP estimator) to have a near-zero bias. This favours use of the REG estimator rather than the RA estimator.

Finally, the relation $\theta_k^{-1} \approx a_p + b_p x_k$ has an even stronger potential for holding true, because each of the p groups may then have its own linear relationship between θ_k^{-1} and x_k. Therefore, of the six estimators examined, the SEPREG estimator has the best potential to realise the objective of a zero nonresponse bias.

The six examples of calibration estimators in Tables 10.2.1 and 10.2.2 illustrate the following important principle: the more we succeed in incorporating important auxiliary information into the auxiliary vector, the better are the chances that the nonreponse bias will be reduced to near-zero levels. In practice, we are of course not limited to the six cases in those tables. The auxiliary information is more extensive in many surveys.

Remark 10.2.1. A quantitative variable x is sometimes used to establish the grouping of the population (or of the sample) that underlies the estimators PST, WCE, SEPRA and SEPREG. In a business survey, x may measure a size-related concept such as "number of employees". The known auxiliary variable values x_k, $k = 1, \ldots, N$, can then be used to create size groups, for example, small, medium, or large elements.

□

We turn to the second principle, that is, *(ii) the auxiliary vector should, as far as possible, explain the variation of the most important study variables.*

Example 6.3.1 has shown that if the perfect linear relationship

$$y_k = \mathbf{x}_k' \boldsymbol{\beta} \qquad (10.2.5)$$

holds for all $k \in U$, then $\hat{Y}_W = Y$. That is, $\hat{Y}_W$ gives a "perfect estimate" of the target parameter Y if (10.2.5) holds. All population residuals E_k in (10.2.3) are then zero, and so is the nonresponse bias. Let us therefore take (10.2.5) as a starting point for analysing whether principle (ii) is likely to be satisfied for the six estimators in Table 10.2.1.

The thought process is now: assume that we could examine the N points $\{(y_k, y_k^0) : k = 1,...,N\}$, where $y_k^0 = \mathbf{x}_k'\boldsymbol{\beta}$. If the relationship is perfect, so that y_k^0 equals y_k for every k, then the nonresponse bias is totally eliminated. This perfect relationship is stated in the second column of Table 10.2.3 and further explained in the third column.

Table 10.2.3. The linear relationship between y_k and the auxiliary vector that eliminates the nonresponse bias for six well-known estimators. α, α_p, β, β_p denote constants.

Estimator	Form of the y_k needed to eliminate bias	Description of the y_k needed to eliminate bias
EXP	$y_k = \alpha$ for all $k \in U$	constant throughout
PST and WCE	$y_k = \beta_p$ for all $k \in U_p$	constant within groups
RA	$y_k = \alpha x_k$ for all $k \in U$	linear in x_k through the origin
REG	$y_k = \alpha + \beta x_k$	linear in x_k
SEPRA	$y_k = \alpha_p x_k$ for all $k \in U_p$	linear in x_k through the origin within groups
SEPREG	$y_k = \alpha_p + \beta_p x_k$ for all $k \in U_p$	linear in x_k within groups

We can now ask: which of the six estimators in Table 10.2.3 is most likely to succeed best in yielding small differences $y_k - y_k^0$ for all k, and thereby a small bias? Table 10.2.3 states that the nonresponse bias for the EXP estimator will be small if all population y-values are essentially identical. This is a far-fetched possibility. The form $y_k = \beta_p$ for the PST and WCE estimators stands a somewhat better chance to approximate the truth. It implies that we should attempt to identify groups U_p, $p = 1,...,P$, which are as far as possible homogeneous with respect to the y-variable, a motivation similar to the one that lies behind the construction of efficient strata for a stratified sampling design. The RA estimator is seen to be in a better position than the EXP estimator to realise small residuals, and even better placed is the REG estimator, since the latter also "allows" an

intercept. Grouping is also involved for the SEPRA and SEPREG estimators, and of the six possibilities in the table, the latter shows the best promise of achieving small residuals. In the interest of further reducing the nonresponse bias, we should try to construct $\mathbf{x}_k$-vectors that give even better chances of small residuals than the six cases in Table 10.2.3.

We have now discussed principles (i) and (ii) in the light of expressions (10.2.2) and (10.2.5). However, it should be noted that the nonresponse bias can be small for other reasons as well. For example, if there are P groups, then the nonresponse bias is a sum of P terms, and it can happen, fortuitously, that these terms are of different signs with the effect of essentially cancelling each other; see Section 10.2.3. Also, for the REG estimator, an analysis shows that the nonresponse bias is small if the E_k are nearly uncorrelated with the θ_k *and* with the quantities $\theta_k x_k$. However, it appears difficult to suggest a concrete action that will meet both of these conditions.

Grouping is a particularly promising avenue for attempts to reduce the nonresponse bias. In the next section we further analyse two of the estimators that involve a grouping, PST and WCE. In particular, we consider principles for the construction of the groups.

10.2.3. Which grouping is optimal?

An effective type of auxiliary information is one that permits a grouping of the elements of the population into poststrata (or the elements of the sample into weighting classes). As Section 10.2.2 shows, the groups should, ideally, be homogeneous with respect to response probabilities and/or the study variable values. To establish such a grouping is not a trivial task. A number of issues enter into consideration, as we will now discuss.

The simple EXP estimator (6.6.1) involves no groups, but it provides a benchmark with which we can compare the usually better alternatives that use groups. As Table 10.2.1 states, EXP is obtained from the general estimator (6.3.2) under the simplest possible formulation of the auxiliary vector, $\mathbf{x}_k = 1$ for all k, and $c_k = 1$ for all k. With these specifications the general expression (10.2.3) for the nonresponse bias becomes:

$$B_{pq}(\hat{Y}_{EXP}) \approx N\left(\frac{\sum_U \theta_k y_k}{\sum_U \theta_k} - \bar{Y}\right) = \frac{N}{\sum_U \theta_k} \sum_U \theta_k E_k \qquad (10.2.6)$$

where $\bar{Y} = \frac{1}{N}\sum_U y_k$ and $E_k = y_k - \bar{Y}$. Another way of writing (10.2.6) is

$$B_{pq}(\hat{Y}_{EXP}) \approx (\sum_U y_k) R_{y\theta U} cv_{yU} cv_{\theta U} \qquad (10.2.7)$$

where

$$R_{y\theta U} = \frac{\sum_U E_k (\theta_k - \bar{\theta})}{[\sum_U E_k^2 \sum_U (\theta_k - \bar{\theta})^2]^{1/2}} \qquad (10.2.8)$$

with $\bar{\theta} = \frac{1}{N}\sum_U \theta_k$, is the finite population correlation coefficient between y and θ,

$$cv_{yU} = \frac{\left(\sum_U E_k^2 / [N-1]\right)^{1/2}}{\bar{Y}} \qquad (10.2.9)$$

is the coefficient of variation of y, and

$$cv_{\theta U} = \frac{(\sum_U (\theta_k - \bar{\theta})^2 / [N-1])^{1/2}}{\bar{\theta}} \qquad (10.2.10)$$

is the coefficient of variation of θ. Thus, the relative bias satisfies

$$B_{pq}(\hat{Y}_{EXP}) / (\sum_U y_k) \approx R_{y\theta U} cv_{yU} cv_{\theta U} .$$

This relative bias will often be large. One reason is the rudimentary form of the residuals $E_k = y_k - \bar{Y}$. They "correct" only in the simplest possible way, namely, by subtracting the overall mean of y. They can be very large (even though their average is zero). Another possibility for a near-zero nonresponse bias is that θ_k is constant for all $k \in U$. Neither of these two conditions is likely to hold. We conclude that $\hat{Y}_{EXP}$ will not give efficient

protection against nonresponse bias. Most survey statisticians are well aware of this, and they seek more powerful auxiliary information. Grouping is one way to improve the situation.

The *One way classification* discussed in Section 6.6 involves the auxiliary vector $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{pk},...,\gamma_{Pk})'$, where γ_{pk} is defined by (6.6.2). If we also specify $c_k = 1$ for all k, we get the poststratified estimator $\hat{Y}_{PST}$, given by (6.6.4), or the weighting class estimator $\hat{Y}_{WCE}$, given by (6.6.5). Both give rise to the same expression for the nonresponse bias. The question is now how to choose the P population groups U_p, $p = 1,...,P$, so as to reduce the nonresponse bias as far as possible. The specification $c_k = 1$ for all $k \in U$ is of the form $c_k = 1/\boldsymbol{\mu}'\mathbf{x}_k$, so the nonresponse bias expression (10.2.3) applies. After some algebra we get:

$$B_{pq}(\hat{Y}_{PST}) \approx B_{pq}(\hat{Y}_{WCE}) \approx \sum_{p=1}^{P} N_p \left(\frac{\sum_{U_p} \theta_k y_k}{\sum_{U_p} \theta_k} - \bar{Y}_p \right) =$$

$$= \sum_{p=1}^{P} a_p \sum_{U_p} \theta_k E_k \qquad (10.2.11)$$

where $\bar{Y}_p = \frac{1}{N_p} \sum_{U_p} y_k$; $a_p = N_p / \sum_{U_p} \theta_k$ and $E_k = y_k - \bar{Y}_p$.

An equivalent expression is

$$B_{pq}(\hat{Y}_{PST}) \approx B_{pq}(\hat{Y}_{WCE}) \approx \sum_{p=1}^{P} (\sum_{U_p} y_k) R_{y\theta U_p} cv_{yU_p} cv_{\theta U_p} \qquad (10.2.12)$$

where $R_{y\theta U_p}$, cv_{yU_p} and $cv_{\theta U_p}$ are defined in analogy with (10.2.8) to (10.2.10), with U_p replacing U. Formulas (10.2.11) and (10.2.12) confirm the message in Tables 10.2.1 and 10.2.2, namely, that the nonresponse bias of PST and WCE is eliminated if (i) the response probabilities are constant within every group (because $cv_{\theta U_p}$ is then zero) or (ii) the y_k-values are constant within every group (because cv_{yU_p} is then zero). In practice, one would most likely have to settle for groups in which some variability remains both in the y_k-values and in the response probabilities. We would

like to identify groups that come close to fulfilling one or both of conditions (i) and (ii). Let us consider some guidelines for this endeavour.

A grouping that fulfils condition (i) will result in a zero nonresponse bias for *all* study variables in the survey. Attempts to meet condition (i) are therefore particularly important. The individual response probabilities θ_k are unknown in essentially all applications. Hence, $cv_{\theta U_p}$ in (10.2.12) is unknown for $p = 1,...,P$. However, an indicator of a good grouping is given by the between groups component of the total variation of the response probabilities, that is, by the second term on the right hand side of

$$\sum_U (\theta_k - \bar{\theta})^2 = \sum_{p=1}^{P} \sum_{U_p} (\theta_k - \bar{\theta}_p)^2 + \sum_{p=1}^{P} N_p (\bar{\theta}_p - \bar{\theta})^2 \qquad (10.2.13)$$

where $\bar{\theta}_p = \frac{1}{N_p} \sum_{U_p} \theta_k$.

The left hand side of (10.2.13) is independent of the grouping, so a grouping that increases the between groups component $\sum_{p=1}^{P} N_p (\bar{\theta}_p - \bar{\theta})^2$ will decrease the within groups component $\sum_{p=1}^{P} \sum_{U_p} (\theta_k - \bar{\theta}_p)^2$. Thus, the numerator of $cv_{\theta U_p}$ will decrease for most or all groups. So if several groupings are compared, the best alternative, in a certain sense, is the one that gives the largest between groups component. This component can be estimated from the realised sample, as we will now indicate. Let I_k, $k \in s$, be defined by

$$I_k = \begin{cases} 1 & \text{if element } k \text{ in sample } s \text{ responds} \\ 0 & \text{otherwise} \end{cases}$$

Then $\hat{\bar{\theta}}_p = \frac{1}{N_p} \sum_{s_p} d_k I_k$ is an unbiased estimator of $\bar{\theta}_p$, $p = 1,\ldots,P$, since

$$E_{pq}(\hat{\bar{\theta}}_p) = E_p\left[\frac{1}{N_p}\sum_{s_p} d_k E_q(I_k)\right] = E_p\left[\frac{1}{N_p}\sum_{s_p} d_k \theta_k\right] = \left[\frac{1}{N_p}\sum_{U_p} \theta_k\right] = \bar{\theta}_p$$

(As earlier in this CBM, we assume that the response probabilities are independent of the realised sample s.) Similarly, $\hat{\bar{\theta}} = \frac{1}{N}\sum_s d_k I_k$ can be shown to be an unbiased estimator of $\bar{\theta}$. Using the realised sample, with any given grouping, we can then compute

$$\sum_{p=1}^{P} N_p (\hat{\bar{\theta}}_p - \hat{\bar{\theta}})^2 \qquad (10.2.14)$$

This estimate of the between groups component is a tool that can be used, but with some caution, in the search for an effective grouping. It would be misleading to believe that the optimal grouping is the one that maximises (10.2.14), because the realised sample s is random, and over- or under-representation of groups will occur by chance. Therefore, information from other sources should be used whenever available, for example, evidence drawn from other surveys about population groups having atypical response rates.

The residuals $E_k = y_k - \bar{Y}_p$ are all equal to zero when condition (ii) holds, that is, when the y_k-values are constant within groups. Now, most surveys are designed to measure several (or even many) study variables. Thus, it is difficult to find groups such that the y_k-values are constant not only for every group but also for every one of the y-variables. In trying to meet condition (ii) one would have to rely on judgement and earlier experience concerning the most important study variables.

10.2.4. A further tool for reducing the nonresponse bias

Two choices influence the nonresponse bias of $\hat{Y}_W$, namely the auxiliary vector $\mathbf{x}_k$ and the factor c_k. Until now we have looked at several well-known estimators. These have fixed specifications of the c_k. However, we can choose the factors c_k as we like. The choice of the c_k becomes a tool

for keeping the nonresponse bias small. By way of illustration, let us see how the formulation of c_k may influence the nonresponse bias in the case where the auxiliary vector is uni-variate, $\mathbf{x}_k = x_k$.

Tables 10.2.1 and 10.2.2 show that the RA estimator uses $c_k = x_k^{-1}$, and the bias is eliminated when there exists a constant a such that $\theta_k^{-1} \approx a$ for all $k \in U$. Thus, if we believe that the response probabilities are roughly constant throughout the population, then $c_k = x_k^{-1}$ is an appropriate choice.

However, let us look at the more general specification $c_k = x_k^{-\nu}$. This specification inserted in expression (10.2.2) gives us the condition for a zero nonresponse bias, namely, $\theta_k^{-1} \approx 1 + a x_k^{1-\nu}$. If we believe that the response probability θ_k decreases as x_k increases we could choose $\nu < 1$ and if we believe that the response probability θ_k is increasing as x_k increases we could choose $\nu > 1$.

10.2.5. More extensive auxiliary information

The better we succeed in incorporating relevant auxiliary information into the $\mathbf{x}_k$-vector, the better, generally speaking, are the chances of realising a low nonresponse bias. Therefore, building a potent $\mathbf{x}_k$-vector is of paramount importance. The statistician must first make an inventory of potential auxiliary variables. This process may reveal a surprisingly large quantity of potential auxiliary information. This step should be followed by a selection of the most pertinent variables. The principles (i) and (ii) in Section 10.2.1 should guide this effort. The reasoning was illustrated in Sections 10.2.2 to 10.2.4 and in Example 3.2.4.

10.3. Literature review

There exists a large literature on the selection of auxiliary information and on the resulting specification of the auxiliary vector. Different aspects of the selection are discussed in the literature. Since some of the recommendations made in these articles have relevance also for topics in this CBM, we now present a brief literature review.

A variety of opinions have been expressed. They cannot easily be summarised in a few firm recommendations. Instead we concentrate on five important themes relating to the selection of auxiliary variables, (a) to (e) below, and review the literature under these themes. The review is not exhaustive.

(a) ***Reduction of the variance***

Nascimento Silva and Skinner (1997) focus on a reduction of sampling error. They do not consider nonresponse. Their intention is to select the "optimal" set of auxiliary variables from a rather large set of potentially useful variables. They compare different ways to carry out this selection and finally recommend a sample-based selection technique. This technique relies on an examination of the relationship found in the sample between the study variable and the auxiliary variables. By contrast, Bankier, Rathwell, and Majkowski (1992) consider a selection method based strictly on the auxiliary variables and their interrelationships.

The selection of auxiliary information is also an issue for the construction of a sampling design. Although this question falls outside the scope of this CBM, we note that several authors have addressed the selection problem from this angle.

For example, Kish and Anderson (1978) discuss ways to stratify the population for a survey with many study variables and with many purposes. They stress that it is important to use many stratifiers (perhaps with relatively few categories for each stratifier), rather than many categories for each of a few stratifiers:

> ... the advantages of several stratifiers are much greater for multipurpose surveys... For any stratifier, the gains in reducing the variance within strata show rapidly diminishing returns with few strata...

A similar principle is likely to work well for the construction of a set of poststrata, when the purpose is to reduce the sampling variance.

(b) ***Reduction of the nonresponse bias***

As pointed out in Section 10.2, the nonresponse bias of the calibration estimator $\hat{Y}_W$ is likely to be small when this estimator builds on powerful auxiliary information. Principle (i) in Section 10.2 requires that the variation

of the response probabilities is explained by the auxiliary vector. Therefore, we ought to use all available knowledge and experience about the correlation pattern existing between the response probabilities and the auxiliary variables. There is a vast literature in this area, based mostly on empirical evidence. For example, in surveys on individuals, experience gathered from many studies tells us that lower response rates are usually expected for the following categories of respondents: metropolitan residents; single people; members of childless households; older people; divorced/widowed people; persons with lower educational attainment; self-employed people; see Holt and Elliot (1991) and Lindström (1983). Similarly, a number of articles on business surveys analyse response rates within subgroups of a population or sample of enterprises. A reference in this area is Groves and Couper (1993).

(c) ***Estimation for domains***

Section 10.2.1 stressed the importance of having access to auxiliary information that comes close to identifying the domains (principle (iii)). Such information is not always available. The simultaneous estimation for several sets of domains creates special problems, as we now illustrate.

Assume that estimates are needed for two sets of domains, defined, respectively, by the groups $p = 1, \ldots, P$ and the groups $h = 1, \ldots, H$ referred to in the discussion of *Two-way classification* in Section 6.6. The number of responding elements may be very small in many of the $P \times H$ cells arising from crossing the two groupings. We could then base the calibration on the one-way classification vector $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk})'$ for the first set of domains and $\mathbf{x}_k = (\delta_{1k}, \ldots, \delta_{hk}, \ldots, \delta_{Hk})'$ for the second set of domains. However, a disadvantage is that different weights are then obtained for the two sets of domains. To eliminate this disadvantage we can instead use a compromise vector of the two-way classification type $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk}, \delta_{1k}, \ldots, \delta_{hk}, \ldots, \delta_{Hk})'$. A unique set of weights is then obtained. Lundström (1996) has shown, in connection with a specific survey at Statistics Sweden, that the variance of the domain estimates tends to be only slightly larger for the two-way classification vector than for the alternative with two separate vectors. Similar conclusions are found in Andersson (1996), for a different survey.

(d) ***Interaction between the objective of variance reduction and the objective of nonresponse bias reduction***

Relatively few articles discuss the choice of auxiliary information for the dual purpose of sampling variance reduction and nonresponse bias reduction. However, many statisticians are aware that a judicious choice of auxiliary information can and should serve such a dual purpose.

Little (1986) discusses the choice of suitable groups for the poststratified estimator when a large amount of information is available. In the *predicted mean stratification* method, the study variable y is regressed on $\mathbf{x}$ in the response set r and then the strata are constructed by grouping the predicted means. In the *response propensity stratification* method, the strata are based on intervals of the estimated response probabilities, called response propensity scores. The estimation can be carried out by logistic or probit regression fitting. Little (1986) notes that:

…. predicted mean stratification has the virtue of controlling both the bias and variance…; response propensity stratification controls … bias, but yields estimates … that may have large variance. The latter is particularly true when the response propensity is largely determined by variables that are associated with y.

Oh and Scheuren (1983) and Särndal and Swensson (1987) express the following views on the question of the dual purpose. Oh and Scheuren (1983) state:

A seemingly robust approach is to choose the subgroups such that for the variable(s) to be analyzed, the within-group variation for nonrespondents is small (and the between-group mean differences are large); then, even if the response mechanism is postulated incorrectly, the bias impact will be small.

Särndal and Swensson (1987) react to this statement as follows:

In our opinion, one must separate the role of the RHGs from that of other information … recorded for $k \in s$. Two different concepts are involved. The sole criterion for the RHGs should be that they eliminate bias as far as possible. Every effort should be made, and a prior knowledge used, to settle on groups likely to display response homogeneity. But in addition it is imperative to measure, for $k \in s$, a concomitant vector $\mathbf{x}_k$, that will reduce variance and give added protection against bias. Groups that eliminate or reduce bias are not necessarily variance reducing, and, contrary to what the quotation seems to suggest, the criterion of maximizing between-to-within variation in y does not necessarily create groups that work well for removing bias.

Särndal and Swensson (1987) argue in principle for the use of two different "concepts", but they admit that practical problems can arise:

… in order to eliminate bias due to nonresponse, it is vital to identify the true response model; as this is usually impossible, bias can be greatly reduced if powerful explanatory *x*-variables can be found and incorporated in a regression-type estimator.

In a similar vein, Bethlehem (1988) states:

...it is very important to look for good stratification variables that will reduce both variance and bias.

Some authors warn that the simultaneous reduction of the variance and the nonresponse bias may represent a conflict of ambitions. Kalton and Maligalig (1991) express this in the following way:

In general, a price paid for adjustment cell weighting is a loss of precision in the survey estimators. There is a trade-off to be made between bias reduction and an increase in variance arising from the variation in the weights. The increase is not great when the variation in weights is modest, but it rises rapidly as the variation increases.... Common techniques for restricting the variation in weights are to collapse cells and to trim the weights... Since cells with small sample sizes often give rise to large variation in weights, minimum sample sizes in the cells are often specified (e.g., 25, 30 or 50).

(e) *Problems generated by the random nature of the sample*

Another question discussed in the literature is whether the current sample information should be allowed to direct the choice of auxiliary information. It has long been known that sample-based selection of auxiliary information may affect the properties of the point estimator, particularly its variance. For example, Bethlehem (1988) points to the importance of using information other than the current sample observations:

The choice of stratification variables cannot be made solely on the basis of the available observations. Over or underrepresentation of some groups can mislead us about the relationship between the target and the stratification variable. There has to be additional information about the homogeneity of the target variable.

A favourable situation arises in a regularly repeated survey. Historical data then exist, in addition to the current sample information.

However, in some surveys there is no additional or historical information, and one must base the selection of the auxiliary variables strictly on the data from the current occasion. Nascimento Silva and Skinner (1997) show that estimators that use "best possible" auxiliary information selection for each realised sample can be effective. That is, there is no a priori decision on the auxiliary variables to be used; instead the decision is made upon inspection of the realised sample.

A simple and commonly used sample-based technique is the collapsing of groups, with a collapsing rule based on the number of respondents in the groups of the realised sample.

The choice of groups also affects the performance of the variance estimator. Lundström (1996) notes that, in rare instances, the variance estimator (4.4.1) can degenerate when the number of respondents in a group is extremely small.

Kalton and Kasprzyk (1986) discuss the negative effects that an excessive variability in the weights can have on the variance. They discuss an estimator defined in terms of weights that are products of two sets of sub-weights, where the first set "... compensates for unequal response rates in different sample weighting classes..." and the second set "... makes the weighted sample distribution for certain characteristics ... conform to the known population distribution for those characteristics...". This can be described as a reweighting step followed by a poststratification step. They recommend inspecting the final weights and, if some are too large, collapsing groups or "trimming the weights" in order to avoid unacceptably large variance.

Inspection of the distribution of the final weights is also recommended for the calibration approach to reweighting, as discussed in the earlier chapters. That is, the variation of the weights g_k, v_k and v_{sk} should be analysed. When extreme weights occur, the first question to examine is whether two or more auxiliary variables measure essentially the same thing so that they are collinear. Collinearity, if it exists, should be eliminated.

The problem with highly variable weights can be treated by one of several available methods for restricting weights so that they lie within a prespecified interval. Procedures of this kind are available in CLAN97.

11. Estimation in the presence of nonresponse and frame imperfections

11.1. Introduction

Many surveys are affected by frame imperfections, often referred to as coverage errors. The frame can have undercoverage, overcoverage or both. Thus the estimation procedure needs to deal simultaneously and effectively with three types of error: sampling error, nonresponse error and coverage error. In earlier chapters we have discussed the first two of these.

Incorporating the third type is not a simple step. The theory in regard to coverage errors is not yet well developed. In this chapter we provide a formal structure for a survey affected by the three types of error and use it to indicate a systematic approach to estimation under these conditions. We proceed by expanding the theory for reweighting and imputation presented in earlier chapters. There are few "conventional" methods in this area, but we obtain some of these "benchmarks" as special cases of a general approach.

A feature of the treatment of coverage errors is that one is obliged to rely on assumptions whose validity is hard or impossible to verify. Many surveys display "special problems" for which "special solutions" have been proposed. This chapter is a step towards a more systematic outlook.

As in earlier chapters we wish to estimate the *target population* total

$$Y_U = \sum_U y_k \qquad (11.1.1)$$

where y_k is the value of the study variable, y, for the kth element of the target population $U = \{1,...,k,...,N\}$, which may differ from the *frame population* from which sampling is carried out. The situation that we address is shown in Figure 11.1.1.

Let s_F be a sample of size n_F drawn from the frame population U_F (of size N_F) with the probability $p(s_F)$. The inclusion probabilities, known

for all $k,l \in U_F$, are then $\pi_k = \sum_{s_F \ni k} p(s_F)$ and $\pi_{kl} = \sum_{s_F \supset \{k,l\}} p(s_F)$. Let $d_k = 1/\pi_k$ denote the *design weight* of element k and let $d_{kl} = 1/\pi_{kl}$.

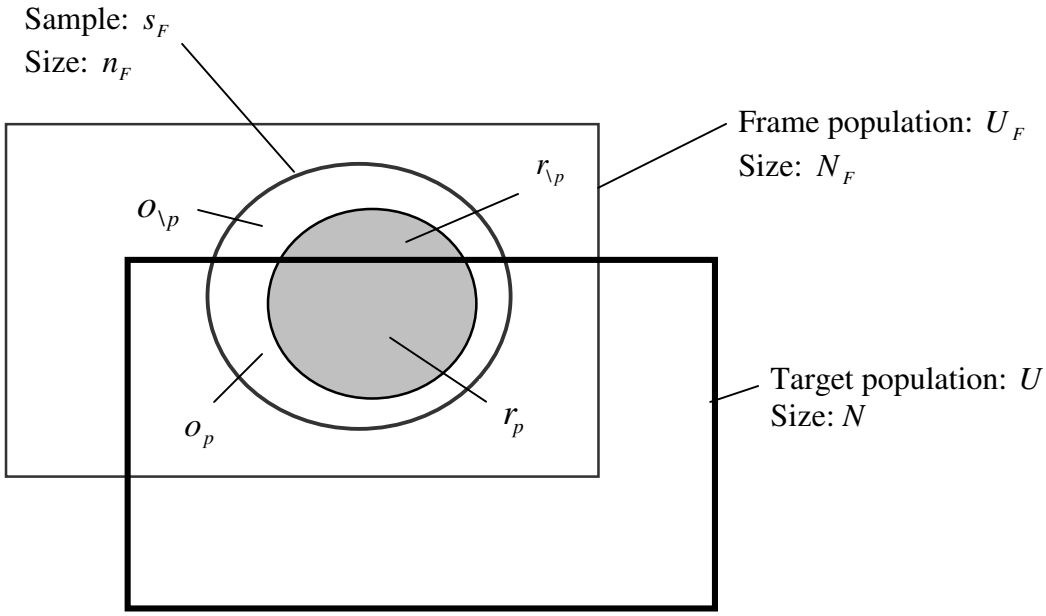


Figure 11.1.1.

The frame U_F in Figure 11.1.1, has both undercoverage and overcoverage. By the *overcoverage set* we mean $U_F - (U \cap U_F)$ and by the *undercoverage set* $U - (U \cap U_F)$. The set of elements that respond *and* belong to the target population is denoted by r_p and its size by m_p. We have $r_p \subseteq s_F$. The subscript *p* is used here to suggest the word "persistors", that is, elements that continue to be in scope for the survey. We denote by $r_{\setminus p}$ the set of elements that respond *and* belong to the overcoverage. Let $m_{\setminus p}$ be the size of $r_{\setminus p}$. The subscript $\setminus p$ is to be interpreted as "non-persistors".

The nonresponding set is $o_p \cup o_{\setminus p}$, where o_p is the part that belongs to the target population and $o_{\setminus p}$ the part that belongs to the overcoverage. The sample s_F is the union of the four nonoverlapping sets r_p, $r_{\setminus p}$, o_p and $o_{\setminus p}$.

We assume that every responding element, $k \in r_p \cup r_{\setminus p}$, can be identified as belonging to either r_p or $r_{\setminus p}$, as the case may be. This identification is usually straightforward in practice. Much more problematic in practice is the separation of the nonresponding elements into their two subsets, o_p and $o_{\setminus p}$.

The observed y-data are $\{y_k : k \in r_p \cup r_{\setminus p}\}$; y-data are missing for $k \in o_p \cup o_{\setminus p}$. We assume that the auxiliary vector value $\mathbf{x}_k$ is available for every $k \in U_F$.

The estimation of the target population parameter $Y_U = \sum_U y_k$ faces the following problems:

(i) the absence of observed y-data from the undercoverage set;

(ii) the absence of a correct auxiliary vector total for the target population;

(iii) the difficulty of separating the nonresponse elements into their two subsets o_p and $o_{\setminus p}$.

We shall examine two different procedures, (1) and (2), for estimating the target population total Y_U. Both procedures are hampered, in different ways, by problems (i), (ii) and (iii).

(1) Estimation of Y_U by adding two terms, considered in Section 11.2. The two terms are: (1A) an estimate of the "persistor total" $Y_{U \cap U_F}$, and (1B) a term to compensate for the undercoverage total $Y_{U - U \cap U_F}$. In Section 11.2 we assume that the compensation in (1B) has been carried out, and we address the simpler problem of estimating $Y_{U \cap U_F}$.

(2) "Direct estimation" of Y_U, considered in Section 11.3.

(The notation used in (1) and (2) and in the rest of this chapter is that if *A* is a set of elements, then Y_A denotes the total $\sum_A y_k$.)

The choice between procedures (1) and (2) depends on the auxiliary information available in the survey. The compensation in step (1B) is problematic. In some surveys, it may be possible to carry out a complementary sample selection from the undercoverage set and to base the compensation on the information thus gathered. In other cases, one must rely on model assumptions and/or the professional judgement of the statistician. Auxiliary information may be used both from the current occasion and from earlier survey occasions. One can expect some bias in the estimates. A systematic approach to this step is presently lacking. If it is possible to arrive at a satisfactory compensation for $Y_{U-U\cap U_F}$, it seems reasonable to choose procedure (1). In other cases, the auxiliary information may be such that procedure (2) is considered a good solution. It is carried out with the aid of the calibration approach, as Section 11.3 shows.

Problem (iii) has particularly grave consequences when an imputation approach is used. We must then impute for the elements $k \in o_p$, but this objective can only be achieved if there is a correct identification of the set o_p. Problem (iii) also affects the variance estimation for the reweighting procedure, as will be discussed in Section 11.2.2.

11.2. Estimation of the persistor total

11.2.1. Point estimation

The persistor set $U \cap U_F$ defines a domain of the target population. This set, which we denote by d_p, also defines a domain of the frame population. The frame usually contains auxiliary information and we can form an auxiliary vector total for U_F, $\sum_{U_F} \mathbf{x}_k$. Since we are assuming that the sampled elements from d_p can be readily identified, the estimation of $Y_{d_p} = Y_{U\cap U_F}$ can follow the principles for domain estimation developed in Section 6.3.

We work with the domain specific variable y_{d_p} such that

$$y_{d_p k} = \begin{cases} y_k & \text{if } k \in d_p = U \cap U_F \\ 0 & \text{otherwise} \end{cases} \qquad (11.2.1)$$

In the reweighting approach, the estimator of Y_{d_p} becomes

$$\hat{Y}_{d_p W} = \sum_{r_p \cup r_{\setminus p}} w_k y_{d_p k} = \sum_{r_p} w_k y_k \qquad (11.2.2)$$

where $w_k = d_k v_k$ with

$$v_k = 1 + c_k \left(\sum_{U_F} \mathbf{x}_k - \sum_{r_p \cup r_{\setminus p}} d_k \mathbf{x}_k\right)' \left(\sum_{r_p \cup r_{\setminus p}} d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \qquad (11.2.3)$$

for $k \in r_p \cup r_{\setminus p}$.

EXAMPLE 11.2.1. *An estimator of $Y_{U \cap U_F}$ commonly used at Statistics Sweden.*

As is common in survey designs at Statistics Sweden, the frame population U_F is divided into strata, U_{Fh}, $h = 1,...,H$, and the sample s_{Fh} is drawn from U_{Fh} by SRS. (Whenever necessary in order to identify a stratum, we add the index h to the notation specified in Figure 11.1.1.) The design weight is then $d_k = N_{Fh} / n_{Fh}$ for $k \in U_{Fh}$. An estimator of $Y_{U \cap U_F}$ commonly used at Statistics Sweden is

$$\hat{Y}_{U \cap U_F} = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{ph} + m_{\setminus ph}} \sum_{r_{ph}} y_k \qquad (11.2.4)$$

It should be emphasised that (11.2.4) estimates the persistor set total only and would lead to a perhaps considerable underestimation if used for the whole target population total $Y_U = \sum_U y_k$. A visual inspection of the weights reveals that they are too small on the average for estimating Y_U. A compensation term, as discussed in step (1B), must be added.

The estimator (11.2.4) can be derived as a special case of the general estimator (11.2.2). This requires defining the auxiliary vector as the stratum identifier $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{hk},...,\gamma_{Hk})'$, where, for $h = 1,...,H$,

$$\gamma_{hk} = \begin{cases} 1 & \text{if } k \in U_{Fh} \\ 0 & \text{otherwise} \end{cases}$$

Then $\sum_{U_F} \mathbf{x}_k = \left(N_{F1},...,N_{Fh},...,N_{FH}\right)'$. Let $c_k = 1$ for all *k*. The matrix to invert in (11.2.3) is diagonal, so the derivation of the weights is straightforward. We obtain

$$v_k = \frac{n_{Fh}}{m_{ph} + m_{\backslash ph}} \quad \text{for} \quad k \in r_{ph} \cup r_{\backslash ph}$$

so that $w_k = \frac{N_{Fh}}{n_{Fh}} \frac{n_{Fh}}{m_{ph} + m_{\backslash ph}} = \frac{N_{Fh}}{m_{ph} + m_{\backslash ph}}$ for $k \in r_{ph} \cup r_{\backslash ph}$.

We have thus obtained the weights of (11.2.4).

□

Alternatively, we can consider an UNIMP-imputation approach. We form an imputed estimator of Y_{d_p} in the manner of (7.2.1) and obtain

$$\hat{Y}_{d_p I} = \sum_{s_F} d_k g_k y_{\bullet d_p k} = \sum_{r_p \cup o_p} d_k g_k y_{\bullet k} \tag{11.2.5}$$

where

$$g_k = 1 + c_k \left(\sum_{U_F} \mathbf{x}_k - \sum_{s_F} d_k \mathbf{x}_k\right)' \left(\sum_{s_F} d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \tag{11.2.6}$$

and

$$y_{\bullet k} = \begin{cases} y_k & \text{for} \quad k \in r_p \\ \hat{y}_k & \text{for} \quad k \in o_p \end{cases} \tag{11.2.7}$$

The estimator $\hat{Y}_{d_pI}$ requires that o_p can be identified, which may or may not be possible in the survey.

Most of Statistics Sweden's surveys on individuals and households rely on sampling from the TPR; see Example 2.2.1. Although the TPR system receives new information on a daily basis, the actual updating of the TPR presently takes place only once a month. Therefore, in a matching of the sample s_F with a very recently updated TPR, the determination of the set o_p will be close to perfect. Even in cases where a few weeks have elapsed since the last updating, the coverage errors can be considered small and inconsequential.

In surveys on enterprises the situation is less favourable. The BR register (see Example 2.2.2) is only updated a few times per year, so at most time points, an accurate determination of o_p is not possible.

11.2.2. Variance estimation

By a straightforward modification of the results of Section 6.4, we derive a variance estimator for $\hat{Y}_{d_pW}$. In the present context with frame imperfections, the weight v_{sk} in Section 6.4 is replaced by

$$v_{sk}^* = 1 + c_k \left(\sum_{s_F} d_k \mathbf{x}_k - \sum_{r_p \cup r_{\setminus p}} d_k \mathbf{x}_k\right)' \left(\sum_{r_p \cup r_{\setminus p}} d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \qquad (11.2.8)$$

for $k \in r_p \cup r_{\setminus p}$. The following variance estimator is obtained:

$$\hat{V}(\hat{Y}_{d_pW}) = \hat{V}_{SAM} + \hat{V}_{NR} \qquad (11.2.9)$$

where

$$\hat{V}_{SAM} = \sum\sum_{r_p \cup r_{\setminus p}} \left(d_k d_l - d_{kl}\right)\left(g_k v_{sk}^* e_{d_p k}\right)\left(g_l v_{sl}^* e_{d_p l}\right) -$$

$$- \sum_{r_p \cup r_{\setminus p}} d_k \left(d_k - 1\right) v_{sk}^* \left(v_{sk}^* - 1\right)\left(g_k e_{d_p k}\right)^2 \qquad (11.2.10)$$

and

$$\hat{V}_{NR} = \sum_{r_p \cup r_{\setminus p}} d_k^2 v_{sk}^* \left(v_{sk}^* - 1\right) e_{d_p k}^2 \qquad (11.2.11)$$

where g_k is given by (11.2.6), and

$$e_{d_p k} = y_{d_p k} - \mathbf{x}_k' \hat{\mathbf{B}}_{d_p v} \qquad (11.2.12)$$

with

$$\hat{\mathbf{B}}_{d_p v} = \left(\sum_{r_p \cup r_{\setminus p}} d_k v_{sk}^* c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_{r_p \cup r_{\setminus p}} d_k v_{sk}^* c_k \mathbf{x}_k y_{d_p k} \qquad (11.2.13)$$

11.3. Direct estimation of the target population total

11.3.1. Introduction

Both the sampling error and the nonresponse error can be substantially reduced when powerful auxiliary information is available and used in reweighting by a calibration approach. The methods recommended in Chapters 4-8 rest on firm ground. In this chapter, we have the additional problem of coverage errors. It is likely that coverage errors, too, can be reduced by a calibration approach. One problem is that a strict adherence to the principles of calibration requires that the target population total $\sum_U \mathbf{x}_k$ be exactly known. Since this condition is unlikely to be met in the presence of coverage error, we must find a good approximation. We denote the approximation $\tilde{\mathbf{X}}$. When the auxiliary vector $\mathbf{x}_k$ is made up of variables present in the frame, the frame auxiliary total $\sum_{U_F} \mathbf{x}_k$ is easily derived. If the coverage deficiencies are deemed inconsequential, it is realistic to take $\tilde{\mathbf{X}} = \sum_{U_F} \mathbf{x}_k$. However, if the coverage deficiencies are extensive, it is not self-evident how to obtain a good approximation of the total $\sum_U \mathbf{x}_k$.

As noted in Section 11.2.1, Statistics Sweden's register on individuals, TPR, is updated frequently enough to ensure that the total $\tilde{\mathbf{X}} = \sum_{U_F} \mathbf{x}_k$ will always be a rather good approximation of $\sum_U \mathbf{x}_k$. The situation is less satisfactory in many other surveys.

11.3.2. Point estimation

We discuss first reweighting by the calibration approach, then imputation. In the reweighting case, item nonresponse is first treated by imputation. We assume that reweighting is applied for the unit nonresponse.

Estimating $Y_U = \sum_U y_k$ by the calibration approach leads to

$$\hat{Y}_{UW} = \sum_{r_p} w_k y_k \qquad (11.3.1)$$

with $w_k = d_k v_k$ and

$$v_k = 1 + c_k \left(\tilde{\mathbf{X}} - \sum_{r_p} d_k \mathbf{x}_k\right)' \left(\sum_{r_p} d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \quad \text{for } k \in r_p \qquad (11.3.2)$$

It is easily seen that the weights have the desired calibration property $\sum_{r_p} w_k \mathbf{x}_k = \tilde{\mathbf{X}}$. A judicious use of this approach can lead to an estimator $\hat{Y}_W$ that effectively controls all three types of error (sampling error, nonresponse error, coverage error), provided that there is a strong correlation between the study variables and the auxiliary vector.

EXAMPLE 11.3.1. *An estimator of Y_U commonly used at Statistics Sweden.*

Assume that the sample s_F is drawn by STSRS as described in Example 11.2.1. An estimator of Y_U commonly used at Statistics Sweden is

$$\hat{Y}_U = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{ph}} \sum_{r_{ph}} y_k \qquad (11.3.3)$$

As we now show, it is the special case obtained from the general formula (11.3.1), when the auxiliary vector is defined by the stratum identifier vector $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{hk}, ..., \gamma_{Hk})'$ as in Example 11.2.1. Assume that the population stratum sizes are unchanged between the time of sampling and the time of estimation. In other words, in each stratum, the overcoverage is assumed to

compensate exactly for the undercoverage. Thus, the required total $\tilde{\mathbf{X}} = (N_{F1},...,N_{Fh},...,N_{FH})'$ is known. Let $c_k = 1$ for all *k*.

Some algebra shows that v_k, given by (11.3.2), takes the form $v_k = \frac{n_{Fh}}{m_{ph}}$ for all $k \in r_{ph}$, so that the total weight of element k becomes $w_k = \frac{N_{Fh}}{n_{Fh}} \frac{n_{Fh}}{m_{ph}} = \frac{N_{Fh}}{m_{ph}}$ for all $k \in r_{ph}$. This is the weight of y_k in (11.3.3).

□

In the UNIMP-imputation approach, we suggest the estimator

$$\hat{Y}_{UI} = \sum_{r_p \cup o_p} w_k y_{\bullet k} \qquad (11.3.4)$$

where $y_{\bullet k}$ is given by (11.2.7), and $w_k = d_k v_k$ with

$$v_k = 1 + c_k (\tilde{\mathbf{X}} - \sum_{r_p \cup o_p} d_k \mathbf{x}_k)' (\sum_{r_p \cup o_p} d_k c_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k \qquad (11.3.5)$$

for $k \in r_p \cup o_p$. A requirement is that o_p is identifiable.

11.3.3. Variance estimation

When there are no coverage errors, Section 6.4 suggests a variance estimator for $\hat{Y}_W$ given by (6.4.1). We now modify this variance estimator to so that it can be used for $\hat{Y}_{UW}$ given by (11.3.1).

Recall that the variance estimator (6.4.1) was derived from a two-phase sampling argument in which the response probabilities play the role of second phase inclusion probabilities. Because these probabilities are unknown, they were replaced by proxies. To obtain (6.4.1), the proxies were $1/v_{sk}$, where v_{sk} is given by (6.3.7). But in the present case of coverage imperfections the weights v_{sk} are not defined, so they must be replaced. We present two alternatives for this.

Supposing for the moment that the v_{sk} are known quantities, we change the notation in (6.4.1) to better represent the current situation with coverage errors. For simplicity we set $g_k = 1$ for all *k*. We obtain

$$\hat{V}(\hat{Y}_{UW}) = \hat{V}_{SAM} + \hat{V}_{NR} \tag{11.3.6}$$

where

$$\hat{V}_{SAM} = \sum\sum_{r_p} (d_k d_l - d_{kl})(v_{sk} e_k)(v_{sl} e_l) -$$

$$- \sum_{r_p} d_k (d_k - 1) v_{sk} (v_{sk} - 1) e_k^2 \tag{11.3.7}$$

and

$$\hat{V}_{NR} = \sum_{r_p} d_k^2 v_{sk} (v_{sk} - 1) e_k^2 \tag{11.3.8}$$

where

$$e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_v \tag{11.3.9}$$

and

$$\hat{\mathbf{B}}_v = (\sum_{r_p} d_k v_{sk} c_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{r_p} d_k v_{sk} c_k \mathbf{x}_k y_k \tag{11.3.10}$$

However, the v_{sk} given by (6.3.7) cannot be used in the present situation. They must be replaced by more suitable quantities, such that their inverse values can be considered as proxies for the unknown response probabilities. We suggest two alternatives for this.

Alternative 1

In formulas (11.3.6) to (11.3.10), replace v_{sk} by v_{sk}^* given by (11.2.8). This produces a variance estimator for $\hat{Y}_{UW}$ which does not require an identification of the set o_p.

Alternative 2

A potential weakness with the v_{sk}^{*} in Alternative 1 is that the elements in the overcoverage do not belong to the target population and may therefore be less prone to respond than those that truly belong. Since the overcoverage elements are not targeted, they may experience some or all items on the questionnaire as being irrelevant. To involve the set $r_{\setminus p}$ in the calibration, as is the case in (11.2.8), is then questionable. Instead, Alternative 2 is as follows: In formulas (11.3.6) to (11.3.10), replace v_{sk} by v_{sk}^{**} given by

$$v_{sk}^{**} = 1 + c_k (\sum_{r_p \cup o_p} d_k \mathbf{x}_k - \sum_{r_p} d_k \mathbf{x}_k)' (\sum_{r_p} d_k c_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k \qquad (11.3.11)$$

for $k \in r_p$. This produces a variance estimator for $\hat{Y}_{UW}$ which does require an identification of o_p. The calibration in (11.3.1) from the set r_p to the set $r_p \cup o_p$ seems reasonable.

APPENDIX A. Components of the total variance: Sampling variance and nonresponse variance

In this appendix we consider a survey with some nonresponse, but with no frame imperfections. That is, the frame and the target population are assumed to be identical. We derive the bias, the variance and the mean squared error (MSE) of the estimators $\hat{Y}_W$ and $\hat{Y}_I$ presented in Chapters 6 and 7. We find that the variance can be represented as the sum of a sampling variance component and a nonresponse variance component.

In what follows the *nonresponse estimator* $\hat{Y}_{NR}$ represents both $\hat{Y}_W$ and $\hat{Y}_I$. Further, we denote by $\hat{Y}$ the expression taken by $\hat{Y}_{NR}$ for the case of full response, when $r = s$. We call $\hat{Y}$ the *full response estimator*.

The discussion in this appendix is sufficiently general that we need not specify whether calibration or imputation or a combination is the chosen approach. The total error of $\hat{Y}_{NR}$ can be written as

$$\hat{Y}_{NR} - Y = (\hat{Y} - Y) + (\hat{Y}_{NR} - \hat{Y}) \qquad \text{(A.1)}$$

or, in words,

Total error = Sampling error + Nonresponse error

We are interested in the usual statistical properties - bias, variance and MSE - of $\hat{Y}_{NR}$. This raises the question of the most appropriate probabilistic setup for deriving these properties. The derivation requires that several expected values be evaluated. For this, we use a probabilistic set-up with two phases. An expected value is interpreted as a double averaging process, first over all possible samples s that can be drawn, and secondly over all the possible response sets r than can occur for any fixed sample s. The two probability distributions involved are the known *sampling design* $p(s)$, and the unknown *response mechanism* $q(r \mid s)$. Note that for any realised sample s, there are many possible outcomes of the response set r, so it is necessary to

average over all r, for the given s. The expected value and variance operators are written as E_{pq} and V_{pq} when taken simultaneously over $p(s)$ and $q(r|s)$; when the evaluation is with respect to $p(s)$ alone, the notation will be E_p and V_p; with respect to $q(r|s)$ alone, the notation will be E_q and V_q.

For more detail on the development that follows, the reader is referred to Lundström (1997) and to Lundström and Särndal (1999).

Bias

We assume that the full response estimator is unbiased (or nearly so), that is, $E_p(\hat{Y}) - Y$ is 0, or very nearly 0, as is the case for the GREG estimator. By definition, the bias is given by $B_{pq}(\hat{Y}_{NR}) = E_{pq}(\hat{Y}_{NR}) - Y$, which we can express as $B_{pq}(\hat{Y}_{NR}) = E_p(B_c)$, where B_c is called the *conditional nonresponse bias*, given a realised sample s; it is given by

$$B_c = E_q(\hat{Y}_{NR}|s) - \hat{Y} \qquad \text{(A.2)}$$

The magnitude of this conditional bias depends on the approach chosen for treating the nonresponse, *and* on the response mechanism $q(r|s)$. It is in general non-zero. The conditional bias is 0 if the nonresponse estimator is equal, on average, to the full response estimate for that sample. This is a good property. But otherwise, as in essentially all surveys with nonresponse, there is a (perhaps substantial) conditional bias.

Variance and mean squared error

Under the probabilistic set-up with two phases, the variance of a random quantity is obtained by the well-known rule "the variance of the conditional expectation plus the expectation of the conditional variance". Applied to $\hat{Y}_{NR}$, the two probability distributions being $p(s)$ and $q(r|s)$, this rule gives the "*pq*-variance"

$$V_{pq}(\hat{Y}_{NR}) = V_p(\hat{Y} + B_c) + E_p V_q(\hat{Y}_{NR}|s) \qquad \text{(A.3)}$$

We know that even if a powerful treatment is used, the nonresponse will inevitably cause some bias. We hope that the bias will be small. But it follows that a more interesting indicator than the variance would be the MSE of $\hat{Y}_{NR}$. The MSE is obtained by adding the squared bias, $\left[E_p(B_c)\right]^2$, to the variance $V_{pq}(\hat{Y}_{NR})$. After simplification, the MSE can be written as

$$MSE_{pq}(\hat{Y}_{NR}) = V_p(\hat{Y}) + E_p V_q(\hat{Y}_{NR}|s) + E_p(B_c^2) + 2Cov_p(\hat{Y}, B_c) \quad \text{(A.4)}$$

Here, $V_p(\hat{Y})$ is the variance of full response estimator. The sum of the other three terms on the right hand side, $E_p V_q(\hat{Y}_{NR}|s) + E_p(B_c^2) + 2Cov_p(\hat{Y}, B_c)$, is the addition to the MSE caused by nonresponse, despite a hopefully efficient approach for treating this nonresponse. The covariance term is not likely to be numerically important, but the term $E_p(B_c^2)$ may represent a considerable and undesired addition to the MSE.

Ideally, the nonresponse approach will have succeeded in eliminating the bias, so that $B_c = 0$ for every possible sample s. Although this is unlikely to occur in practice, we would then have

$$MSE_{pq}(\hat{Y}_{NR}) = V_{pq}(\hat{Y}_{NR}) = V_p(\hat{Y}) + E_p V_q(\hat{Y}_{NR}|s) \quad \text{(A.5)}$$

It is fitting to call $V_{NR} = E_p V_q(\hat{Y}_{NR}|s)$ the *nonresponse variance*, because it does not involve the nonresponse bias. Letting $V_{TOT} = V_{pq}(\hat{Y}_{NR})$, $V_{SAM} = V_p(\hat{Y})$ and $V_{NR} = E_p V_q(\hat{Y}_{NR}|s)$, we can also write (A.5) as

$$V_{TOT} = V_{SAM} + V_{NR} \quad \text{(A.6)}$$

or, in words,

Total variance = Sampling variance + Nonresponse variance

APPENDIX B. Proxies of the unknown response probabilities to be used in the variance estimator

Section 6.1 describes the two-phase approach for treating nonresponse. The description is based on an estimator suggested by Särndal, Swensson and Wretman (1992), which uses auxiliary population totals; see (6.1.2). However, they also suggest an estimator in the case when only HT-estimates of these totals are known. We will use this estimator in what follows. The estimator is

$$\hat{Y}_{SSW,s} = \sum_r d_k g_{sk\theta} y_k / \theta_k \qquad \text{(B.1)}$$

where

$$g_{sk\theta} = 1 + c_k (\sum_s d_k \mathbf{x}_k - \sum_r d_k \mathbf{x}_k / \theta_k)' (\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k' / \theta_k)^{-1} \mathbf{x}_k \qquad \text{(B.2)}$$

The corresponding estimator in the calibration approach is the estimator $\hat{Y}_{Ws}$, given by (6.3.6). To derive a relevant variance estimator for $\hat{Y}_{Ws}$ we can argue as follows: The expression (B.1) is based on two-phase sampling theory and cannot be used as it stands in the nonresponse situation, because the second-phase inclusion probabilities are then unknown. They must therefore first be estimated in some way. We raise the following questions: What "hypothetical" response probabilities will make $\hat{Y}_{SSW,s}$, given by (B.1), identical to $\hat{Y}_{Ws}$? If we replace the unknown θ_k in the variance estimator for $\hat{Y}_{SSW,s}$ suggested by Särndal, Swensson and Wretman (1992), Chapter 9, by such "estimated" response probabilities, will the resulting expression provide a good variance estimator for $\hat{Y}_{Ws}$?

The first question is answered by the following proposition.

Proposition B.1. Let v_{sk} be given by (6.3.7). When θ_k is replaced by $\hat{\theta}_k = v_{sk}^{-1}$, then $\hat{Y}_{SSW,s}$ becomes identical to $\hat{Y}_{Ws}$ given by (6.3.6). Moreover,

the values $\hat{\theta}_k = v_{sk}^{-1}$ satisfy the reasonable condition $\sum_r \frac{d_k \mathbf{x}_k}{\hat{\theta}_k} = \sum_s d_k \mathbf{x}_k$.

When θ_k is replaced by $\hat{\theta}_k$ in the weights $g_{sk\theta}$, given by (B.2), then these weights are equal to unity for all k.

□

The proof of the proposition is given in Lundström (1997).

The equivalence of $\hat{Y}_{SSW,s}$ and $\hat{Y}_{Ws}$ stated in Proposition B.1 suggests the following procedure: For $\hat{Y}_{Ws}$, we propose to use the variance estimator given by Särndal, Swensson and Wretman (1992), formula (9.7.28), where we replace $\pi_{k|s_a}$ by θ_k and then θ_k by the proxy value $\hat{\theta}_k = 1/v_{sk}$, where v_{sk} is given by (6.3.7). We also assume that elements respond independently.

To derive an estimator of the variance of $\hat{Y}_W$, given by (5.2.3), we propose a similar approach: In the variance estimator given by Särndal, Swensson and Wretman (1992), formula (9.7.22), replace $\pi_{k|s_a}$ by θ_k and then θ_k by the proxy value $\hat{\theta}_k = 1/v_{sk}$, where v_{sk} is given by (6.3.7). We also assume that elements respond independently. Several simulation studies reported in Lundström (1997) show that the suggested variance estimators for $\hat{Y}_{Ws}$ and $\hat{Y}_W$ work well.

APPENDIX C. A general expression for the nonresponse bias for the calibration estimator

In this section a general expression of the nonresponse bias is developed. Special cases of this general expression, corresponding to specific $\mathbf{x}_k$-vectors, are derived and discussed in Sections 10.2.2 and 10.2.3. We assume that the response probabilities, $\theta_k = \Pr(k \in r \mid s)$, are independent of the realized sample s.

Proposition C.1. For large response sets the nonresponse bias of $\hat{Y}_W$ given by (6.3.2) is

$$B_{pq}(\hat{Y}_W) \approx -\sum_U (1-\theta_k) E_{\theta k} \tag{C.1}$$

where $E_{\theta k} = y_k - \mathbf{x}_k' \mathbf{B}_\theta$ and $\mathbf{B}_\theta = \left(\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k c_k \mathbf{x}_k y_k$.

For large response sets, the right hand side of (C.1) also represents the approximate nonresponse bias of $\hat{Y}_{Ws}$ given by (6.3.6).

□

PROOF. It is easily seen that the estimator $\hat{Y}_W$ given by (6.3.2) can be written

$$\hat{Y}_W = \sum_U \mathbf{x}_k' \hat{\mathbf{B}}_r + \sum_r d_k \left(y_k - \mathbf{x}_k' \hat{\mathbf{B}}_r\right) \tag{C.2}$$

where $\hat{\mathbf{B}}_r = \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k c_k \mathbf{x}_k y_k$.

Thus, the error $\hat{Y}_W - Y$ can be written

$$\hat{Y}_W - Y = \sum_U \mathbf{x}_k' \hat{\mathbf{B}}_r + \sum_r d_k \left(y_k - \mathbf{x}_k' \hat{\mathbf{B}}_r\right) - \sum_U y_k =$$

$$= \sum_r d_k \left(y_k - \mathbf{x}_k' \mathbf{B}\right) - \sum_U \left(y_k - \mathbf{x}_k' \mathbf{B}\right) +$$

$$+ \left(\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k\right)' \left(\hat{\mathbf{B}}_r - \mathbf{B}\right) \qquad \text{(C.3)}$$

where $\mathbf{B} = \left(\sum_U c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U c_k \mathbf{x}_k y_k$.

However,

$$\hat{\mathbf{B}}_r - \mathbf{B} = \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k c_k \mathbf{x}_k E_k \qquad \text{(C.4)}$$

where $E_k = y_k - \mathbf{x}_k' \mathbf{B}$.

For large response sets

$$E_{pq}\left[\left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k c_k \mathbf{x}_k E_k\right] \approx \mathbf{B}_{\theta E}$$

where

$$\mathbf{B}_{\theta E} = \left(\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k c_k \mathbf{x}_k E_k \ .$$

However, it is easily seen that

$$\mathbf{B}_{\theta E} = \mathbf{B}_\theta - \mathbf{B} \qquad \text{(C.5)}$$

where

$$\mathbf{B}_\theta = \left(\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k c_k \mathbf{x}_k y_k \ .$$

Consequently, the nonresponse bias $B_{pq}(\hat{Y}_W)$ can be approximated by

$$B_{pq}(\hat{Y}_W) = E_{pq}(\hat{Y}_W) - Y \approx$$

$$\approx \sum_U \theta_k E_k - \sum_U E_k - \left(\sum_U \theta_k \mathbf{x}_k - \sum_U \mathbf{x}_k\right)' (\mathbf{B}_\theta - \mathbf{B}) =$$

$$= -\sum_U (1-\theta_k) E_k + \sum_U (1-\theta_k) \mathbf{x}_k' (\mathbf{B}_\theta - \mathbf{B}) =$$

$$= -\sum_U (1-\theta_k) y_k + \sum_U (1-\theta_k) \mathbf{x}_k' \mathbf{B}_\theta = -\sum_U (1-\theta_k) E_{\theta k} \qquad \text{(C.6)}$$

where $E_{\theta k} = y_k - \mathbf{x}_k' \mathbf{B}_\theta$.

It is easily seen that the estimator $\hat{Y}_{Ws}$ can be written

$$\hat{Y}_{Ws} = \sum_s \mathbf{x}_k' \hat{\mathbf{B}}_r + \sum_r d_k \left(y_k - \mathbf{x}_k' \hat{\mathbf{B}}_r\right)$$

and then the expression corresponding to (C.6) will be

$$\hat{Y}_{Ws} - Y = \sum_r d_k (y_k - \mathbf{x}_k' \mathbf{B}) - \sum_s d_k (y_k - \mathbf{x}_k' \mathbf{B}) +$$

$$+ \left(\sum_s d_k \mathbf{x}_k - \sum_r d_k \mathbf{x}_k\right)' (\hat{\mathbf{B}}_r - \mathbf{B}) + \sum_s d_k y_k - \sum_U y_k \qquad \text{(C.7)}$$

It is easy to repeat the steps from (C.1) to (C.6) of the proof of Proposition C.1 and conclude that the same approximate bias expression holds for the estimator $\hat{Y}_{Ws}$. The following corollary is easily derived.

□

Corollary. Suppose that $c_k = 1/\boldsymbol{\mu}' \mathbf{x}_k$ for all $k \in U$, where $\boldsymbol{\mu}'$ is a constant column vector of the same dimension as $\mathbf{x}_k$ and not dependent on *k*. Then, for large response sets,

$$B_{pq}(\hat{Y}_W) \approx -\sum_U E_{\theta k} \tag{C.8}$$

or equivalently,

$$B_{pq}(\hat{Y}_W) \approx \sum_U \mathbf{x}_k' \mathbf{B}_{\theta E} \tag{C.9}$$

where

$$\mathbf{B}_{\theta E} = \left(\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k c_k \mathbf{x}_k E_k \tag{C.10}$$

□

Proposition C.2. If there exists a constant column vector $\boldsymbol{\lambda}$ such that $\theta_k^{-1} = 1 + c_k \boldsymbol{\lambda}' \mathbf{x}_k$ for $k \in U$ then $B_{pq}(\hat{Y}_W) \approx 0$, where $B_{pq}(\hat{Y}_W)$ is given by either (C.1) or (C.8).

□

PROOF. After some algebraic manipulations of expression (C.1) we will have

$$B_{pq}(\hat{Y}_W) \approx \sum_U \mathbf{x}_k' \mathbf{B}_\theta - \sum_U \theta_k \mathbf{x}_k' \mathbf{B}_\theta - \sum_U (1-\theta_k) y_k \tag{C.11}$$

Multiply the term for element k in the first sum on the right hand side of (C.11) by $\theta_k (1 + c_k \boldsymbol{\lambda}' \mathbf{x}_k) = 1$ and it becomes

$$\sum_U \mathbf{x}_k' \mathbf{B}_\theta = \sum_U \theta_k \mathbf{x}_k' \mathbf{B}_\theta + \boldsymbol{\lambda}' (\sum_U \theta_k c_k \mathbf{x}_k' \mathbf{x}_k) \mathbf{B}_\theta =$$

$$= \sum_U \theta_k \mathbf{x}_k' \mathbf{B}_\theta + \boldsymbol{\lambda}' (\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k') (\sum_U \theta_k c_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U \theta_k c_k \mathbf{x}_k y_k) =$$

$$= \sum_U \theta_k \mathbf{x}_k' \mathbf{B}_\theta + \sum_U \theta_k c_k \boldsymbol{\lambda}' \mathbf{x}_k y_k$$

However, since $\theta_k c_k \boldsymbol{\lambda}' \mathbf{x}_k = 1 - \theta_k$,

$$\sum_U \mathbf{x}_k' \mathbf{B}_\theta = \sum_U \theta_k \mathbf{x}_k' \mathbf{B}_\theta + \sum_U (1 - \theta_k) y_k$$

It follows that (C.11) equals zero.

□

Remark C.1. It is easily seen that the component $\sum_U \theta_k c_k \mathbf{x}_k E_k$ in expression (C.10) will be a vector of zeros, when the population residual $E_k = 0$ for all k, and consequently the nonresponse bias will be zero.

□

Remark C.2. Assume that each element responds with the same probability $\theta_k = \theta_0$ for all k. Then the expression (C.1) becomes $B_{pq}(\hat{Y}_W) \approx -(1-\theta_0)\sum_U E_k$. When the factors c_k are completely general there is no guarantee that $\sum_U E_k = 0$ and thus, that the bias will be zero. However, if $c_k = 1/\boldsymbol{\mu}' \mathbf{x}_k$ for all $k \in U$, then $\sum_U E_k = 0$.

□

Bethlehem (1988) and Fuller, Loughin and Baker (1994) also discuss expressions of the nonresponse bias.

APPENDIX D. Cases where imputation and reweighting result in the same estimator

Proposition D.1. Consider the imputed GREG estimator $\hat{Y}_I$ given by (7.2.1), where g_k is given by (4.3.4), and assume that GREG-conformable multiple regression imputation is used so that, for $k \in o = s - r$, $\hat{y}_k = \mathbf{z}_k' \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = \left(\sum_r q_k \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \sum_r q_k \mathbf{z}_k y_k$, with $q_k = d_k c_k$ and $\mathbf{z}_k = \mathbf{x}_k$. Then the imputed GREG estimator is identical to the reweighted estimator $\hat{Y}_W$ given by (6.3.2), that is, $\hat{Y}_I = \hat{Y}_W$ for every possible response set r.

□

PROOF. The imputed GREG estimator $\hat{Y}_I = \sum_s d_k g_k y_{\bullet k}$, where $g_k = 1 + c_k (\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)' \left(\sum_s d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$, can be written

$$\hat{Y}_I = \sum_s d_k g_k y_{\bullet k} = \sum_r d_k g_k y_k + \sum_o d_k g_k \hat{y}_k =$$

$$= \sum_r d_k g_k y_k + (\sum_o d_k g_k \mathbf{z}_k') \hat{\boldsymbol{\beta}} \qquad \text{(D.1)}$$

Let us examine the two components (i) $\sum_r d_k g_k y_k$ and (ii) $(\sum_o d_k g_k \mathbf{z}_k') \hat{\boldsymbol{\beta}}$ at the right hand side of (D.1).

The component (i) can be written

$$\sum_r d_k g_k y_k = \sum_r d_k y_k + C \qquad \text{(D.2)}$$

where $C = (\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)' \left(\sum_s d_k c_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k c_k \mathbf{x}_k y_k$.

Moreover, the component (ii) can be written:

$$(\sum_o d_k g_k \mathbf{z}'_k)\hat{\boldsymbol{\beta}} = (\sum_s d_k g_k \mathbf{z}_k - \sum_r d_k g_k \mathbf{z}_k)'\hat{\boldsymbol{\beta}} = (\sum_U \mathbf{z}_k - \sum_r d_k g_k \mathbf{z}_k)'\hat{\boldsymbol{\beta}} =$$

$$= (\sum_U \mathbf{z}_k - \sum_r d_k \mathbf{z}_k)'\hat{\boldsymbol{\beta}} -$$

$$- (\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)'\left(\sum_s d_k c_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}\left(\sum_r d_k c_k \mathbf{x}_k \mathbf{z}'_k\right)\hat{\boldsymbol{\beta}} \qquad \text{(D.3)}$$

Since $q_k = d_k c_k$ and $\mathbf{z}_k = \mathbf{x}_k$, the vector $\hat{\boldsymbol{\beta}}$ has the form

$$\hat{\boldsymbol{\beta}} = \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \sum_r d_k c_k \mathbf{x}_k y_k \qquad \text{(D.4)}$$

Insering expression (D.4) into expression (D.3), component (ii) becomes

$$(\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k)' \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}\left(\sum_r d_k c_k \mathbf{x}_k y_k\right) -$$

$$- (\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k)'\left(\sum_s d_k c_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}\left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}'_k\right)\times$$

$$\times \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}\left(\sum_r d_k c_k \mathbf{x}_k y_k\right) =$$

$$= (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k)' \left(\sum_r d_k c_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}\left(\sum_r d_k c_k \mathbf{x}_k y_k\right) - \mathrm{C} \qquad \text{(D.5)}$$

By adding (D.2) and (D.5) it follows that

$$\hat{Y}_I = \sum_r w_k y_k \qquad \text{(D.6)}$$

where $w_k = d_k v_k$ and v_k is given by (6.3.3).

This is exactly the expression for $\hat{Y}_W$ given by (6.3.2).

□

Proposition D.2. Consider the imputed HT estimator (7.2.3) and use, for $k \in o = s - r$, GREG-conformable multiple regression imputation according to $\hat{y}_k = \mathbf{z}_k' \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = \left(\sum_r q_k \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \sum_r q_k \mathbf{z}_k y_k$, with $q_k = d_k c_k$ and $\mathbf{z}_k = \mathbf{x}_k$. Then the resulting imputed HT-estimator is identical to the reweighted estimator $\hat{Y}_{Ws}$ given by (6.3.6), that is, $\hat{Y}_I = \hat{Y}_{Ws}$ for every possible response set r.

□

PROOF. The imputed estimator $\hat{Y}_I = \sum_s d_k y_{\bullet k}$ can be written

$$\hat{Y}_I = \sum_s d_k y_{\bullet k} = \sum_r d_k y_k + \sum_o d_k \hat{y}_k = \sum_r d_k y_k + (\sum_o d_k \mathbf{z}_k')\hat{\boldsymbol{\beta}} =$$

$$= \sum_r d_k y_k + (\sum_s d_k \mathbf{z}_k - \sum_r d_k \mathbf{z}_k)'\hat{\boldsymbol{\beta}} =$$

$$= \sum_r d_k y_k + (\sum_s d_k \mathbf{z}_k - \sum_r d_k \mathbf{z}_k)' \left(\sum_r q_k \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \left(\sum_r q_k \mathbf{z}_k y_k\right) \quad \text{(D.7)}$$

Insert $q_k = d_k c_k$ and $\mathbf{z}_k = \mathbf{x}_k$ into (D.7) and it follows that $\hat{Y}_I = \sum_r d_k v_{sk} y_k$, where v_{sk} is given by (6.3.7). Thus, $\hat{Y}_I = \hat{Y}_{Ws}$ for every possible response set r.

□

Gabler and Häder (1999) present alternative proofs of Propositions D.1 and D.2 based on the conditional minimax principle.

References

Andersson, C. and Nordberg, L. (1998). CLAN97 - a SAS-program for computation of point- and standard error estimates in sample surveys. Statistics Sweden.

Andersson, S. (1996). Användning av administrativt datamaterial som hjälpinformation vid estimationen - en studie av användningen av SRU-materialet som hjälpinformation till finansstatistiken. ES-Metod 1996:1, Statistics Sweden.

Atmer, J., Thulin, G. and Bäcklund, S. (1975). Coordination of Samples with the JALES Technique. *Statistisk Tidskrift*, **13**, 343-350.

Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. *Proceedings of the Section on Survey Research Methods,* American Statistical Association*„ 764-769.*

Bethlehem, J.G. and Kersten, H.M.P. (1985). On the treatment of nonresponse in sample surveys. *Journal of Official Statistics* **1**, 287-300.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* **4**, 251-260.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376-382.

Deville, J.C., Särndal, C.E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013-1020.

Djerf, K. (1997). Effects of post-stratification on the estimates of the Finnish Labour Force Survey. *Journal of Official Statistics*, **13**, 29-39.

Djerf, K. (2000). Properties of some estimators under unit nonresponse. Statistics Finland, Research Report no. 231.

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics* **3**, 325-337.

Estevao, V.M., Hidiroglou, M.A. and Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, **11**, 181-204.

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide Food Consumption Survey. *Survey Methodology* **20**, 75-85.

Gabler, S. and Häder, S. (1999). Representive Weights and Imputation for the 1997 German ISSP: An Application of the Conditional Minimax Principle. Paper presented at the International Conference on Survey Nonresponse in Portland, Oregon, U.S.A.

Groves, R.M. and Couper, M.P. (1993). Unit nonresponse in demographic surveys. *Proceedings of the Bureau of the Census Annual Research Conference,* 593-619.

Holt, D. and Elliot, D. (1991). Methods of weighting for unit non-response. *The Statistician* **40**, 333-342.

Hörngren, J. (1992). The use of registers as auxiliary information in the Swedish Labour Force Survey. Statistics Sweden, R&D Report no. 1992:13.

Jagers, P. (1986). Post-stratification against bias in sampling. *International Statistical Review* **54**, 159-167.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing data. *Survey Methodology* **12**, 1-16.

Kalton, G. and Maligalig, D.S. (1991). A comparison of weighting adjustment for nonresponse. *Proceedings of the Bureau of the Census Annual Research Conference,* 409-428.

Kish, L. and Anderson, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association* **73**, 24-34.

Kish, L. (1979). Samples and censuses. *International Statistical Review* **47**, 99-110.

Lee, H., Rancourt, E. and Särndal, C.E. (2000). Variance estimation from survey data under single value imputation. Ottawa: Statistics Canada, Technical report HSMD - 2000 - 006E.

Lee, H., Rancourt, E. and Särndal, C.E. (2001). Variance estimation from survey data under single imputation. To appear as Chapter 21, invited papers volume, *International Conference on Survey Nonresponse*, Portland, Oregon, 1999.

Lindström, H. (1983). Nonresponse errors in sample surveys. *Urval 16*, Statistics Sweden.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* **54**, 139-157.

Lundström, S. (1996). Kalibrering av vikter i Kortperiodisk sysselsättningsstatistik för privat sektor. *Proceedings of the 20th Conference of Nordic Statisticians*, Copenhagen, 442-451.

Lundström, S. (1997). Calibration as a standard method for treatment of nonresponse. Doctoral dissertation, Stockholm University.

Lundström, S. and Särndal, C.E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics* **15**, 305-327.

Nascimento Silva, P.L.D. and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology* **23**, 23-32.

Oh, H.L. and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In: W.G. Madow, I. Olkin and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 143-184.

Rubin, D.B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20-34.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Särndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* **55**, 279-294.

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

SOS (1998). Newly started enterprises in Sweden 1996 and 1997. Statistical Report Nv 12 SM 9801 in the series Official Statistics of Sweden. Örebro, Sweden: Statistics Sweden.

Statistics Sweden (1980). Räkna med bortfall. Statistics Sweden.

Statistics Sweden (1997). Minska bortfallet. Statistics Sweden.

Swedish Data Inspection Board (1974). Beslut N:741203 angående Arbetskraftsundersökningarna.

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of non-response when analysing survey data. *Statistisk Tidskrift* **11**, 278-285.

Thomsen, I. (1978). A second note on the efficiency of weihgting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift* **16**, 191-196.

INDEX of important terms

For many general and frequently used terms only a single page indication is given, usually the number of the page where a term is explained or defined. These terms are indicated by an asterisk (*).