UNIVERSITY OF SOUTHAMPTON

SOME ESTIMATION AND BIAS ISSUES IN BUSINESS SURVEYS

Dan Erik Hedlin

Thesis for the degree of

Doctor of Philosophy

Department of Social Statistics

FACULTY OF SOCIAL SCIENCES

June 2003

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES

SOCIAL STATISTICS

Doctor of Philosophy

SOME ESTIMATION AND BIAS ISSUES IN BUSINESS SURVEYS

by Dan Erik Hedlin

This thesis discusses some sources of bias in business surveys, why bias arises and how the bias can be estimated or, in some cases, ameliorated. In practical business statistics there are many bias issues not yet resolved.

The introduction of new businesses on a business frame is subject to reporting delays, that is, there are delays between the time when they have started trading and the time when they appear on the frame. Reporting delays cause undercoverage. The thesis provides methodology to predict the undercoverage, exploiting some links to AIDS research and actuarial science.

Another issue addressed is the bias that will arise if the knowledge of overcoverage gained in sample surveys is mistreated. It is usually discovered in the data collection phase of a survey that some units in the sample are ineligible even if the frame information has indicated otherwise. This information may be fed back to the frame and used in subsequent surveys, thereby making forthcoming samples more efficient by avoiding sampling ineligible units. The thesis investigates what effect on survey estimation the process of feeding back information on ineligibility may have, and derives an expression for the design-bias that can occur as a result of feeding back.

Although asymptotically design-unbiased, widely used GREG estimators may produce bad estimates. The thesis examines the behaviour of GREG estimators when the underlying model is misspecified. A diagnostic for whether a GREG estimate is reasonable is discussed. A common justification for the use of GREG estimators is that, being asymptotically design unbiased, they are relatively robust to model choice. However, this work shows that the property of being asymptotically design unbiased is not a substitute for a careful model specification search.

The thesis raises the question of what the desirable properties of an estimator are and explores several point estimators in a simulation study. Special consideration is given to how prone an estimator is to produce large errors. This property is particularly important in official statistics where the publication of bad estimates may sometimes lead to great losses for society and may also be detrimental to the reputation of the producer.

Contents

List of tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Acknowledgement

This thesis draws on collaborative research, especially with the ONS. I am very grateful to many former and current members of staff in the Methodology Group at the Newport site of the ONS for suggesting interesting problems and providing me with the opportunity to work with these problems. They include Peter Brodie, Pam Davies, Hannah Falvey, Trevor Fenton, Susan Full, Tim Jones, Mark Pont, Paul Smith, Markus Šova, Pam Tate and Ceri Underwood. Hopefully, I have been able to provide useful solutions to the problems I have been asked to consider; and I do hope that I have been able to convey the methods and results of this work in appropriate forums.

I am also very grateful to Chris Skinner and Ray Chambers, who, with a leap of faith, offered me a position at the Department of Social Statistics. Thinking back to the job interview where Ray articulated his conviction that I'd make 'a good carthorse', I'd hope that this prediction of his turned out reasonably correct.

Many thanks to Chris and Ray without whose firm belief that everybody should write a PhD thesis this work would not have been done.

Bengt Rosén has been important to me. It was Bengt's inspiring course on the excellent book Model Assisted Survey Sampling by Carl-Erik Särndal, Bengt Swensson and Jan Wretman (who was my first teacher on survey sampling and in that capacity brought me into this field in the first place) that started a learning process which eight years later on got me to this point. Throughout my writing on the MSc dissertation, Bengt would always patiently reject any suggestion that someone like me might get away with anything less than a meticulous piece of work. On the other hand, I patiently taught him SAS.

I would also like to extend my thanks to Mac McDonald and Suojin Wang who have contributed to part of the work reported in the thesis. Mac and Suojin have been great colleagues to work with.

I am also grateful to Jelke Bethlehem and Danny Pfeffermann, both of whom managed to actually read the submitted thesis and provide many useful suggestions (or requirements if you like) for improvement, all which will be helpful to future readers, although these may well exist only in a conceptual world of objects with no real-world correspondence. I am also truly grateful for their letting me retain some dignity at the viva.

After having adjusted to and gone through the PhD process, I find myself mulling over fees for renting a gown and a hood in the appropriate colours (£40 for one day), and other ceremonial stuff that I am expected to relish. Somehow I'm left with some vague feeling of emptiness rather than pride. What was this all about, really?

Chapter 1

Introduction

1.1 Business surveys

In this chapter a brief overview of business survey methods and a general literature review are given. I will also give an outline of the following chapters. They address some special issues germane to business surveys and each chapter includes a literature review of the particular area that the chapter is about.

Business surveys (often also called establishment surveys) are defined through the population unit. 'Business' is here used as a very broad term. My use of the term business surveys is extended to economic activities of institutions and governmental bodies, as opposed to 'social surveys', which are concerned with people. There are surveys that can be viewed as either 'social' or 'business surveys'. For example, a survey of small family businesses may be concerned with both people and, say, turnover. Surveys that are neither 'business' nor 'social' surveys include 'environmental' surveys, e.g. surveys aiming at estimating carbon dioxide emission and wildlife abundance. I will use the term business loosely. For the purposes of this thesis we do not need a precise definition of the units under study. For a formal definition of business surveys, see Cox and Chinnappa (1995, p. 3), and for business units, see European Union (1993, Sec. III).

Dutka and Frankel (1991) distinguish between two types of business survey:
1. 'Enumerative surveys'. For example, the aim may be to estimate the total sales by industry. 'Descriptive surveys' is an often used synonym.
2. 'Analytical surveys'. Here the aim is market analysis, for example to assess customer satisfaction, or receptivity towards new products.

This thesis pertains to enumerative surveys. The overall goal of enumerative business surveys is to provide information on structure and development of economical activities within industries, regions or in the nation as a whole through estimation of parameters that are closely associated with the finite population of businesses.

Most major business surveys are conducted by National Statistical Institutes (NSIs), for example, the UK Office for National Statistics (ONS). There are typically two general types of business survey carried out by NSIs:

- Annual surveys. These have often very large sample sizes or are censuses above some threshold in terms of business size. The aim is 'structural' information such as business production, employment and finances, often for small domains.
- Subannual surveys. These are far smaller in terms of sample size and number of variables than the annual surveys and are aimed at estimating change and trend as well as population or domain totals. Timeliness is particularly important for subannual surveys.

Business surveys in terms of use and content can broadly be classified into five groups:

- Sample surveys, annual or subannual, that collect information on characteristics such as stocks, turnover, employment and production. The parameters of interest are typically totals, or differences between totals, by industry.
- Price index surveys: consumer price index, producer price index, etc. Price index surveys pose special statistical problems. It is, for example, a non-trivial problem to define the target population parameter (Leaver and Valliant 1995).
- Ongoing register surveys that are used for updating the frame. The term proving refers in this context to the checking of activity status, and sometimes also to industry classification.
- Economic censuses.
- Economic cycle forecasting: surveys that ask about the investment plans of businesses.

The ONS conducts some 100 business surveys on a regular basis, many of which are small. However, 17 of these have sample sizes larger than 10,000 a year (broken down into smaller chunks for subannual surveys). Several submit more than 100,000

questionnaires a year. Table 1 in Smith, Pont and Jones (2003) lists the ten most important ONS business surveys.

I shall discuss business surveys from two different perspectives. Firstly, I outline a general methodology framework for business surveys. Secondly, I discuss differences between business surveys and social surveys. In the last section of Chapter 1 I give short summaries of the following chapters. In both section 1.2 and section 1.3 I highlight the connections to the following chapters. This structure will lead to some repetition but I found it useful to separate methodology from the discussion of special features of business surveys.

1.2 A methodological framework for business surveys

The main reference on business surveys is Cox et al. (1995). There is a rather limited number of other texts that specifically address issues with business survey methodology; they include College (1989), Edwards and Cantor (1991), Hidiroglou and Srinath (1993). There is also the ICES II conference proceedings Kovar (2000). Hidiroglou and Laniel (2001) provide an account of modern business survey methodology, while Smith et al. (2003) describe the development of ONS business survey methodology since 1995.

Sample surveys are the main means for meeting the aims of enumerative surveys. A goal of such a sample survey is to estimate some *target population parameter*, which is a function on the target population U. More specifically, a business survey *target population* consists of a finite number of labelled units (elements) $U = \{1, 2, \ldots, N\}$, where N is usually unknown. There is an associated *study variable vector* $\mathbf{y}_k' = (y_{1k}, y_{2k}, \ldots, y_{qk})$ and usually also an *auxiliary variable vector* $\mathbf{x}_k' = (x_{1k}, x_{2k}, \ldots, x_{pk})$ for each unit $k \in U$. Skinner, Holt, and Smith (1989, pp. 14-15) discuss the nature of auxiliary information.

Like Särndal, Swensson and Wretman (1992), I use the symbol $\sum_s$ to indicate a sum taken over all units in the set s. The terms will usually be indexed by k. The by far most common types of target population parameter to be estimated are totals. The *total* of

one scalar study variable $\mathbf{y}' = (y_1, y_2, \ldots, y_N)$ on U is defined as $t_y = \sum_U y_k$. The *change* is also important, often defined as the difference or ratio of the totals from consecutive periods or two periods twelve months apart.

Most business survey samples are taken from a *frame*, which in this context is a list of enterprises, businesses and in some cases also local units and other units that are part of businesses. Throughout the thesis I assume that the population units, frame units and the sample units are of the same type, in short referred to as businesses. There are, however, business surveys where these units are different; for example, the population units may be kind of activity units and the frame may consist of whole businesses only.

Units belonging to the target population but not present on the frame constitute the *undercoverage*. There is nearly always *overcoverage* as well: businesses that have ceased trading but still are on the frame. While undercoverage may lead to bias, the consequence of overcoverage is a waste of part of the sample and hence loss of precision.

The frame is the hub of a business survey. The sources used to build up and maintain a business frame are typically tax registers. Multiple administrative sources for updating the frame may lead to duplication of units. Alternatively, one source could be used as the main source and the other sources to estimate the undercoverage only (Colledge 1989, 1995). The ONS uses the former approach.

It is reasonable to believe that in most industrialised countries, all businesses, except a small minority operating entirely within the black market, will eventually come on to the administrative records system and hence have their details subsequently passed on to the NSI. Therefore, the main reason for undercoverage of the population as a whole is *reporting delays*. For a subpopulation (for example an industry) misclassification may be an even more serious problem, i.e. the circumstance that some businesses have incorrectly been put in another subpopulation than the one they belong to according to the definition of the subpopulation. There may be a considerable time lag between the detection of births, industry classification of the new units, and their inclusion on the frame. Chapter 2 proposes a way of estimating this type of undercoverage. There is a similar problem for deaths. In some cases, the dead unit is not taken off the frame until

an annual update. In the UK the reporting delays tend to be long, and deaths show even longer reporting delays than the birth delays, mainly due to the fact that a business is formally in existence until all outstanding claims that the state may have are settled.

The sample surveys themselves are another source of information relevant to frame maintenance. If it is found in a sample survey that a certain business has ceased trading, this information may be passed on to the frame. However, this may lead to bias in forthcoming surveys. This issue is discussed in Chapter 3.

The *sampling design* (or just *design*) is a function $p(\cdot)$ that gives the probability $p(s)$ of selecting sample s out of the set S of all possible samples from U. Naturally, I assume that $p(s)$ is a well-defined probability measure. Furthermore, I assume that $p(s)$ may depend on $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$, but not on any $\mathbf{y}_k$, $k \in U$. Having said that, there is a subtle form of association between $p(s)$ and the $\mathbf{y}_k$ in the application discussed in Chapter 3.

Since most business frames used by NSIs are believed to be largely satisfactory as sampling frames, the design of business surveys is usually single-stage. The most common business survey design is *stratified simple random sampling without replacement* (STSI). The population is partitioned into *subpopulations* sometimes called *pre-strata*, for example industries, and in each subpopulation a predetermined number of *size strata* are created. Denote the number of size strata in a subpopulation by H, although this number often varies over subpopulations. The size strata $A_1, A_2, \ldots A_H$ are determined by a scalar auxiliary variable x_k, $k = 1, 2, \ldots, N$, and stratum boundary points $b_1 < b_2 < \ldots < b_{H-1}$:

$$
\begin{aligned}
A_1 &= \{k : x_k \le b_1\}, \\
A_h &= \{k : b_{h-1} < x_k \le b_h\}, \; h = 2, 3, \ldots H-1, \\
A_H &= \{k : b_{H-1} < x_k\},
\end{aligned}
\tag{1.1}
$$

where x_k, $k = 1, 2, \ldots, N$, is the *stratification variable*. At the ONS, employment is used as the single stratification variable. The size strata are at the ONS usually referred to as *sizebands*.

There is a considerable literature on the problem of *univariate stratification*, i.e. how to set the boundary points. Sigman and Monsour (1995) give an overview. Hedlin (1998, 2000) extends classical stratification techniques (Cochran, 1977, Sec. 5A.7).

Under the STSI design a *simple random sample without replacement* (SI) is drawn from each stratum independently of samples of other strata. The design SI is defined as the sample selection scheme that attaches the same selection probability to all samples of some predetermined size. The only sampling designs that feature in the thesis are the STSI and SI designs. Although possibly suboptimal, I have taken the stratification already done for the surveys as given.

Size stratum H is in most business surveys completely enumerated, that is, all frame units in the size stratum are included in the sample. I will use the abbreviation '*CE stratum*'.

A sample is taken and $(\mathbf{x}_k,\ \mathbf{y}_k)$ is ideally observed for all units $k \in s$. The sample may exhaust the known part of the target population in which case the survey is a census. Often the study variable cannot be observed for some selected units. These units are called *nonrespondents*. More often than not the study variable can be observed only with some *measurement error*.

An estimate of the target population parameter is computed through a rule referred to as an *estimator*. The *Horvitz-Thompson estimator* of the population total t_y (HT-estimator; Horvitz and Thompson, 1952) is

$$\hat{t}_{y\pi} = \sum_s w_k y_k \,, \qquad (1.2)$$

where $w_k = \pi_k^{-1}$ and $\pi_k = \Pr(I_k = 1)$ with I_k taking the value 1 if unit k is in the sample and 0 otherwise. I will refer to an estimate obtained with (1.2) as an HT-estimate. The π_k is called the *inclusion probability of unit k*. With an STSI design $\hat{t}_{y\pi}$ specialises to the *expansion estimator* with $w_k = N_h/n_h$ for a unit k that belongs to stratum h, where N_h is the number units in the part of the frame that belongs to stratum h and n_h is the size of the sample taken from stratum h.

The *generalised regression (GREG) estimator* is a wide class of estimators that together with the HT estimator covers the vast majority of all estimators used in practice in business surveys. The GREG is defined as

$$\hat{t}_{yreg} = \sum_s g_{ks} w_k y_k , \qquad (1.3)$$

where the sample-dependent g_{ks} are the *g-weights*, to use a term coined by Särndal, Swensson and Wretman (1989). These weights are defined as

$$g_{ks} = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)' \left(\sum_s w_j q_j \mathbf{x}_j \mathbf{x}_j'\right)^{-1} (q_k \mathbf{x}_k), \qquad (1.4)$$

where $\mathbf{t}_x$ and $\hat{\mathbf{t}}_{x\pi}$ are vectors of the auxiliary variable population totals and HT-estimates of them, and the q_k's are some additional weights. I shall discuss GREG estimation in detail in Chapter 4.

There are two main approaches to inference in finite population sampling. First, the randomisation or *design-based* approach adopts the view that the values of the study variable and the auxiliary variable are fixed (i.e. non-random). The properties of an estimator are evaluated with respect to the set $\mathcal{S}$ of all possible samples under the given sampling design. For example, an estimator $\hat{t}_y$ of t_y is *design-unbiased* if $E(\hat{t}_y) = t_y$, where the expectation is defined with respect to all possible samples under the design, that is, $E(\hat{t}_y) = \sum_{s \in \mathcal{S}} \hat{t}_{ys} p(s)$ where I write $\hat{t}_{ys}$ to emphasise that the estimator is evaluated for the particular sample s.

The other main approach to inference in finite population sampling, the prediction or *model-based* approach, views the values $(\mathbf{x}_k, \mathbf{y}_k)$, $k \in s$, as a result of a stochastic process and the estimators are evaluated with respect to a conceptualised infinite sequence of realisations of the stochastic variables $(\mathbf{X}_k, \mathbf{Y}_k)$, $k \in U$, under a superpopulation model. However, the auxiliary variable vector is typically treated as fixed. The *model-assisted* approach, as described by Särndal et al. (1992), is essentially design-based. A superpopulation model is adopted to guide the choice of estimator but the evaluation is based solely on design-based principles. Other frameworks for survey

sampling have been put forward, see Thompson (1997, p. 156) and Valliant, Dorfman, and Royall (2000, p. 7) for further references.

Finite population sampling differs from many other areas of statistics in that the target population exists physically. Especially for business surveys, there is a compelling notion of study variables representing true characteristics of real-world entities. Therefore, it is natural to make inference about the target population parameters, as opposed to (only) model parameters. This seems to support the design-based approach, which in practice is by far the more common of the two approaches. There is however an ongoing discussion about this topic, see e.g. Valliant et al. (2000) and Smith (1997). Finite population sampling theory has a different tradition than the body of statistical theory (Smith 1994, 2001, and Brewer 1994).

It is convenient to regard the outcome of some survey process as 'random' if our best understanding of the process leaves us with uncertainty that is best explained by some probabilistic model (cf. Dembski's 1998 first explication of the concept of randomness). There are four potential sources of randomness in this sense in business surveys under the framework outlined here:

1. There may be an underlying process assumed to be generating the true values of the population units (a superpopulation model).
2. The mechanism by which population units come onto the frame.
3. The mechanism by which sample units are selected to be observed (often called 'response mechanism').
4. The method used to obtain the measurements for the observed units ('measurement process').

There is also a sample selection procedure:

5. The mechanism by which frame units are included in the sample (sampling design).

The actual drawing of the sample is often a list-sequential drawing, that is, the frame list is gone through with a Bernoulli experiment being performed for each unit which decides whether the unit should be included in the sample or not. In many official business statistics systems, the sample selection procedure involves an algorithm for constructing pseudo-random numbers and another algorithm for the selection of a

sample according to the prescribed sampling design, given the distribution of the pseudo-random numbers. Fan, Muller and Rezucha (1962) give an algorithm for drawing a fixed-size SI sample with only one pass through the frame. Two other examples of sampling procedures are Atmer, Thulin, and Bäcklund (1975) and Ohlsson (1998). It may be problematic to regard this type of sample selection procedure as a random process (cf. Dembski's 1998 second explication of the concept of randomness: 'pseudo-randomness'), but to pursue this interesting topic further would take this introduction to far. Suffice to say that two different conditional approaches, model-based and design-based inference, that address points 1 and 5 have been mentioned above. With few exceptions, the only sources of randomness I recognise in this thesis are the second and the fifth one, although superpopulation models will in a sense be the focus of Chapter 4.

There are many other practical and methodological issues that need to be addressed for a business survey to be conducted successfully. These include 'profiling', to use a term coined at Statistics Canada, which refers to the delineation of units (Pietsch 1995). Also, the target population units must be classified in a number of ways, for example by industry, see Nijhowne (1995). The act of checking and correcting respondent data in surveys is usually referred to as editing. Since business surveys are often heavily edited and the benefit of the editing has often been shown to be meagre (Granquist and Kovar 1997), more effective methods have been sought. With selective editing techniques, responses are prioritised according to believed impact on estimates. Discussions of selective editing include Lawrence and McKenzie (2000) and Hedlin (2002a, 2003). Overviews of editing include Granquist and Kovar (1997) and Bethlehem and van de Pol (1998). Another issue still is the storage and utilisation of the results. The production, documentation and storage of microdata and metadata in databases are increasingly important both for business statistics and other types of official statistics (e.g. Fellegi and Wolfson, 1999, Colledge 1999). The organisation and evaluation of survey processes have attracted the attention of many NSIs in recent years; see e.g. Biemer and Caspar (1994), Bethlehem (1997), Morgenstein and Marker (1997), and Keller (1999). Cox et al. (1995) and Lyberg et al. (1997) provide extensive overviews of practical and theoretical issues that arise in surveys, although only the former is specifically concerned with business surveys.

In any country, business survey data are protected by confidentially regulations and practices. One special circumstance in the UK is that the UK Statistics of Trade Act 1947 prohibits disclosure of any microdata, even if the microdata are anonymised and there is no way of identifying the unit. Since most results in this thesis are supported by real data, gratefully obtained from the ONS, I have been obliged to adhere to the UK Statistics of Trade Act. One consequence is that I cannot attach the real values to the axes of scatter plots if these show real data.

1.3 What's special about business surveys?

There are clear differences between business and social surveys, see Cox and Chinnappa (1995) and Rivière (2002). Cox and Chinnappa (1995, p. 1) note that 'unlike the situation for social surveys, economic surveys [business surveys in my terminology] have too few commonly accepted, practiced, and published methodologies'. However, business and social surveys are also similar in many respects, and sometimes actual or alleged differences are used as an excuse not to go against the rather rigid tradition of business surveys in many countries (Dillman, 2000). The general methodology is the same. An often repeated myth is that business surveys are more 'factual' than social surveys. Apart from the important aim of some social surveys to measure attitudes, social surveys are no less focused on facts that any other surveys. The units in business and social surveys are per definition different, although, as I pointed out above, there is some overlap. Clearly, the interplay between the survey organisation and the respondent is in general different for business and social surveys. This has of course implications for the data collection (Dillman, 2000; Eldridge, Martin, and White, 2000; and Edwards and Cantor, 1991). I will, however, focus on some other differences that are highly relevant to the thesis. It is these special features of business surveys that give rise to the issues addressed in the thesis.

Business populations are volatile in terms of variable values, units and population. Typical for elements in business statistics is that they change fairly rapidly; they "die", and they merge and split. There is a lively stream of new businesses coming onto the frame more or less continuously, albeit with a skewed reporting delay distribution, which can be estimated (Chapter 2). Also, the rapid turnover in the frame combined with the frequent use of samples over time that overlap makes the information on

newly dead units gleaned from the samples especially valuable. However, too simplistic a use of this information can create bias (Chapter 3).

In business surveys, subpopulations are often more interesting than the whole population. The subpopulations are often rather small. The sampling design is often highly stratified, and stratum samples sizes can be very small indeed. Stratum sample sizes of 5 are not uncommon in practice. Estimation theory relies to a large extent on large sample results, but it is not clear whether these are applicable to business surveys (Chapters 4 and 5).

Business variable distributions tend to be extremely skewed. Often a few units dominate. Furthermore, the values of the study variable often contain a large proportion of zeroes. The skewed distributions may make design-based inference more problematic for business surveys than for social surveys, or, as Kalton (2002, p. 130) says: 'The situation is somewhat different for establishment surveys [as opposed to social surveys] because of the highly skewed distributions of many of the variables of interest, leading to small numbers of establishments dominating the survey estimates. Furthermore, the sampling frames for establishment surveys often contain auxiliary variables related to the sizes of the establishments that can play an important role in estimation. Although design based inference is still the generally preferred mode for establishment surveys, the case for model-dependent methods is stronger in this area [than in social surveys]' (Chapters 4 and 5).

One of the most important, or the most important, recipient for official business statistics is the national accounts of the nation (Lewington, 1995). The output from the surveys is combined, adjusted and complemented with output from other sources and goes into the national accounts. Most systems of national accounts cannot use estimates of mean squared errors or confidence intervals because only functions of the estimated totals are inserted into the supply and demand tables. In theory, but probably not in practice, two estimates corresponding to the end-points of a confidence interval rather than the single number that is the point estimate could be inserted to allow for a sensitivity study. However, for the large number of point estimates that are combined to form the national accounts (literary thousands every quarter) the vast number of combinations of end-points will be infeasible to handle. This fact makes properties of

interval estimates less important than those of point estimates, as, for example, design-bias (Chapter 5).

1.4 Aim of the thesis and a summary of Chapters 2 to 5

This thesis is about 'real-world' problem solving. The topics are methodological issues that have arisen in my work as a consultant for some NSIs. The general aim of the thesis is to clarify the nature of a number of these issues to explore the sources and consequences of various problems and to investigate ways of avoiding the negative consequences. I discuss some sources of bias in a general sense in business surveys, why bias arises and how the bias can be estimated or, in some cases, ameliorated. In practical business statistics there are many bias issues not yet resolved.

One of the imperfections of a sampling frame is miscoverage caused by delays in recording real-life events on the frame. New units generally appear on the frame some time after they were 'born' and units that have 'died' are not removed from the frame immediately. Chapter 2 provides methodology for predicting the undercoverage due to reporting delays, whereby links to AIDS research and actuarial science are explored. Generalised linear models are fitted to historical data and used to predict future data. The approach presented here is novel in a business survey context. As a special case, I also predict the number of new-born units per month. The methodology is applied to the business register in the UK, maintained by the Office for National Statistics.

Chapter 3 addresses the bias that will arise if the knowledge of overcoverage gained in sample surveys is mistreated. It is usually discovered in the data collection phase of a survey that some units in the sample are ineligible even if the frame information has indicated otherwise. For example, in many business surveys a nonnegligible proportion of the sampled units will have ceased trading since the latest update of the frame. This information may be fed back to the frame and used in subsequent surveys, thereby making forthcoming samples more efficient by avoiding sampling nonnegligible units. I investigate what effect on survey estimation the process of feeding back information on ineligibility may have, and derive an expression for the design-bias that can occur as a result of feeding back. The focus is on estimation of the total using the expansion

estimator. An estimator that is nearly design-unbiased in the presence of feed back is obtained.

Although asymptotically design-unbiased, GREG estimators may produce bad estimates. Chapter 4 starts with a summary of GREG estimation theory and goes on to examine the behaviour of GREG estimators when the underlying model is misspecified. It shows how an efficient GREG estimator was found for a business survey that posed some problems. The work involved data exploration in several steps, combined with analyses of g-weights, residuals and standard regression diagnostics. An important concept introduced is the 'g-weight function', which also serves as a diagnostic for whether a GREG estimate is reasonable or not. I take an in-depth look at some design-based estimators and ask the question whether model considerations are superfluous in the design-based context once the estimator has been determined. A common justification for the use of GREG estimators is that, being asymptotically design unbiased, they are relatively robust to model choice (cf. Särndal et al. 1992, Sec. 6.7). However, the property of being asymptotically design-unbiased is not a substitute for a careful model specification search, especially when dealing with the highly variable and outlier prone populations that are the focus of many business surveys.

Chapter 5 discusses estimation of the total for some study variables in two business surveys conducted by the ONS. I ask what the desirable properties of an estimator are and explore several point estimators in a simulation study. Special consideration is given to how prone an estimator is to produce large errors. This property is particularly important in official statistics where the publication of bad estimates may sometimes lead to great losses for society and may also be detrimental to the reputation of the NSI. Some widely used design-based estimators (and one or two less widely used ones) are contrasted with a model-based estimator that explicitly draws on the special structure of a business population. Sections 5.1-5.3 are based on Hedlin (2002b).

There is a concluding discussion in Chapter 6 in which I debate the practical use of the main findings of the thesis.

This thesis draws on collaborative research, especially with the ONS, where a substantial part has been my own original work. Chapter 2 is based on research with Trevor Fenton, John W. McDonald, Mark Pont, and Suojin Wang. This research has generated a separate manuscript, with myself as the first author. Part of this work is reported in Hedlin, Pont, and Fenton (2000). It was Mark Pont of the ONS who brought our attention to the problem discussed in this chapter and it was Dr McDonald's idea to look at AIDS research for methods that could be transferred to the business survey problem discussed here. Dr McDonald and Professor Wang assisted me with the log-linear modelling and prediction issues. An ONS report, Fenton, Hedlin, Perry, and Pont (1999), was mainly written by Trevor Fenton, who also produced a CD with SAS files that I have been using. The ONS report contains some of the results presented in Chapter 2. When I noticed that a similar problem is addressed in actuarial science I made use of some results from this area in our research.

Chapter 3 is based on research with Suojin Wang and is closely related to the paper Hedlin and Wang (2002). I have taken the lead in this research, but Professor Wang has assisted me throughout the research. In particular, it is fair to say that we have made equal contributions to the derivation of the inclusion probabilities reported in Chapter 3.

Chapter 4, except for the first section on GREG estimation theory, draws heavily on Hedlin, Falvey, Chambers, and Kokic (2001). Hannah Falvey, then a member of staff at the ONS, did all the rather heavy work of getting the data into a form that we could use. She also computed the vast majority of the estimates that are reported in the paper using the Canadian software GES, results that I have compared with those of the Swedish software CLAN. This has been reported in Falvey and Hedlin (1999). Professor Chambers, who also is one of my supervisors, assisted me throughout the research that eventually produced the paper Hedlin et al. (2001). Dr Kokic also made valuable contributions.

Other research areas that I have contributed to during work on this thesis include selective editing. Appendix 1, which is not formally part of the thesis, provides an introduction to Hedlin (2003).

Chapter 2

Estimating the Undercoverage of a Sampling Frame due to Reporting Delays

2.1 Introduction

Most sample surveys draw their samples from a frame. More often than not, part of the target population is not accessible from the frame: the survey will suffer from undercoverage. A *reporting delay* or, using an equivalent term, a *birth lag* is defined as the time from *birth* (for a frame of businesses, the day when the business began to trade) to *frame introduction* (when the business came onto the frame). Conversely, the *death lag*, causing overcoverage, is the time between cessation of activity (*death*) and the business being removed from the frame. It is believed that for the business surveys run by the ONS, reporting delays are the most important source of undercoverage of the population as a whole. As was mentioned in Chapter 1, for a subpopulation (e.g. industry) misclassification may be an even more serious problem. In this chapter, I estimate the number of businesses that have started trading but have not yet come onto the frame. This number is an important frame quality indicator. I will also provide methodology for estimating the bias that may follow from reporting delays.

Most information on births and deaths is updated as soon as it is received in the ONS. However, some information relating to births and deaths is held back pending further information or investigation. When the size information indicates that the new unit has employment of 20 or more, and the unit cannot be matched against existing frame units, the recording of the unit is further delayed pending proving of the information about the unit. The lengths of birth lags form a highly skewed distribution. Some businesses report to the relevant authority in the UK as soon as they are set up, resulting in short lags. Others may have been operating for years below the level of annual turnover above which registration is compulsory, i.e., before their growth necessitates their registration. In these cases the lag may be very long indeed. Some

businesses report to an administrative body in advance of their launch, sometimes resulting in a negative birth lag. Figure 2.1 shows the distribution of the number of births over non-negative birth lags. The vast majority of new businesses (85%) have come onto the ONS frame within four months of their birth. About 10% have longer birth lags than five months.

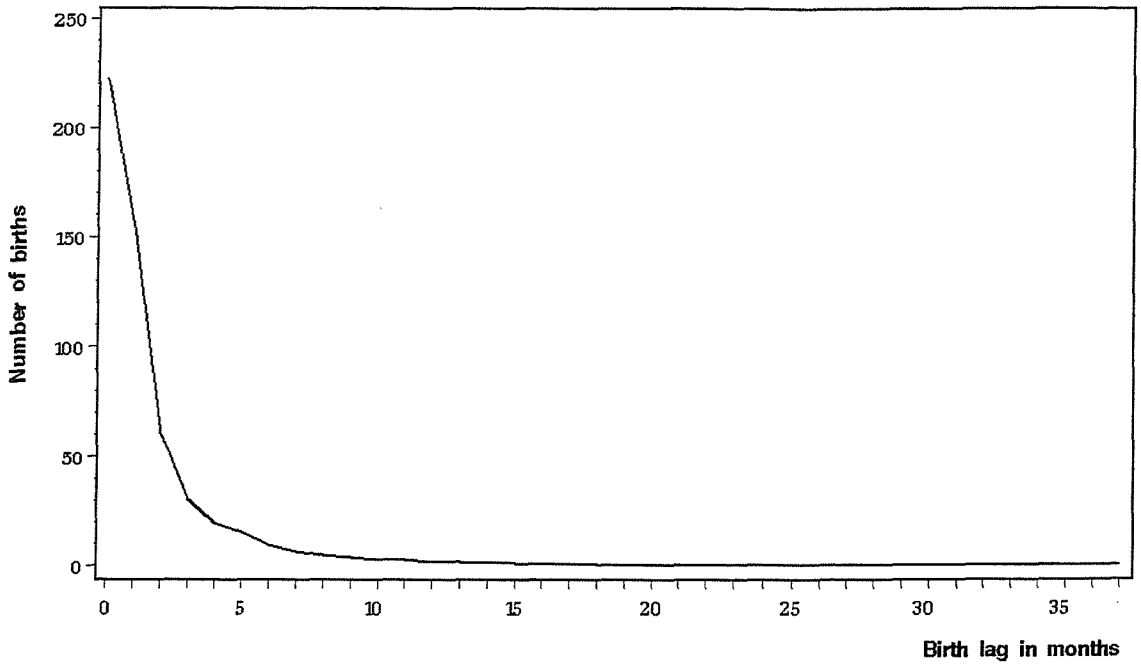


Figure 2.1. Number of observed births (in thousands) against birth lag (months).

The aim of this chapter is to devise a method for estimating the undercoverage that is caused by birth lags. The approach is to fit a generalised linear model (see, e.g., McCullagh and Nelder 1989) to historical frame records for which both birth dates and reporting delays have been recorded. The model will then be used for predicting forthcoming births. I have not attempted to accommodate economic cycles as the usable data go back only to 1995. Businesses that never come onto the frame, for example, very small businesses or businesses operating entirely within the black market, are ignored.

It is not in general possible to tell at the ONS whether a dead business has been closed because of a genuine death, or because it has been part of a merger, takeover etc. Information that precedes the start of a business in legal terms is not recorded. The net number of births may therefore be more interesting than the total number of births.

Deaths are reported through the same administrative bodies and the resulting reporting delays will be similar to birth lags, although they tend to be longer. The net number of births can be estimated as the difference between the predicted gross numbers of births and deaths. While I focus on birth lags, the same methods could be applied to death lags.

Table 2.1 indicates the birth lag distribution for businesses born between January 1, 1995 and March 22, 1998. The rows of the table represent the numbers of businesses that were born in each month. The month a business started operating is referred to as its *birth month*. The columns are birth lags measured in months. For example, 5444 businesses started operating in January 1995 and came on to the frame within one month, which is counted as a zero lag.

Table 2.1. Number of observed births per lag (in months) and birth month. Partially unobservable cell counts are indicated with a ≥ symbol, totally unobserved cell counts with a dash.

	Birth lag						
	0	**1**	**2**	**...**	**38**	**>38**	**Total**
Jan, 95	5,444	4,982	1,910	...	≥6	–	16,054
Feb, 95	5,333	4,069	1,280	...	–	–	13,425
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
Jan, 98	7,783	4,102	≥1,346	...	–	–	13,231
Feb, 98	7,075	≥3,087	–	...	–	–	10,162
Mar, 98	≥5,888	–	–		–	–	5,888
Total	226,582	156,517	61,346	...	6	–	549,386

The business registers of the ONS were merged in 1993 and the Interdepartmental Business Frame (IDBR) was created. Before 1995 the IDBR was in a state of considerable flux as data from the two previous registers were being matched. Hence I only use data from 1 January 1995.

The administrative sources that the IDBR is built upon are mainly HM Customs and Excise and Inland Revenue. HM Customs and Excise provides the ONS with information relating to Value Added Tax (VAT)-registered legal units weekly. These indicate new registrations, and any units that have deregistered. Inland Revenue provides a file of all Pay As You Earn (PAYE) employer records each quarter. In the PAYE scheme employers pay the employees' income tax and national insurance

contributions. From these notifications, new registrations and deregistrations can be detected by comparison with the file from the previous quarter. Because the ONS is not notified continuously, frame introductions tend to be clustered in time. The total number of businesses on the IDBR was around 1998 about 1.8 million (in addition to the data analysed here there were a large number of businesses that went unchanged through a period starting in 1995 and ending in February 1998).

With the observation window spanning the period January 1995 –March 1998 the longest observable birth lag is 38 months. The count of the rightmost cell in the first row of Table 2.1 is unobservable (unless we gain access to data that go beyond the final date in the data currently available). Adhering to common terminology, cells with unknown counts are referred to as *structural zeroes* (see e.g. Agresti 1990); their unknown counts are represented in Table 2.1 with dashes. The term structural zero is conventional but in this case 'unobservable counts' might have been more telling. With structural zeroes, the table is an *incomplete contingency table*. The rightmost diagonal of the upper triangle containing observed counts is partially unobservable. Another way of expressing the fact that we cannot observe new businesses that have not yet been introduced on the sampling frame is to say that the data are *right-truncated* (as distinct from the term censoring, which refers to units whose existence are known but cannot be observed). The problem of estimating the undercoverage due to birth lags is equivalent to estimating the number of businesses that have been subjected to right-truncation. On March 31, 1998, the undercoverage is the sum of the unknown counts in the lower triangle of Table 2.1, ignoring longer birth lags than 38 months. As a special case, the row totals can be predicted; they correspond to the number of births per month. Note that it is the column sums of Table 2.1, excluding partially truncated cells, that are graphed in Figure 2.1.

There is surprisingly little literature on reporting-delay induced undercoverage of a frame used for sample surveys, considering the importance of the problem and the fact that there is research on similar issues in other areas. The approach presented here to estimate the number of unobservable businesses is akin to and was inspired by estimation of the incidence of cases of AIDS in the presence reporting delays, see Wang (1992), Sellero et al. (1996) and references therein. However, my application is

different: the dataset and the contingency tables are very large. There is also a structure to the data that needs to be accommodated.

An extension to the problem of predicting the population size is to predict the population total of some variable. Most businesses in transition between start and frame introduction are part of the target population and hence their absence from the sampling frame will result in a negative bias in estimated totals if these are based solely on samples from the frame. I propose a method of estimating this bias. A similar estimation problem is addressed in actuarial science. Insurance companies need to estimate the net sum of claims that have been made but not yet been settled; see e.g. Haberman and Renshaw (1996).

Section 2 explores the data behind the incomplete contingency table and the table itself. In Sections 3 and 4 Poisson regression models are fitted to the upper triangle of Table 2.1 to predict the unobservable cell counts in the lower triangle. In Section 5 the precision of each model is assessed by a cross-validation type of study. Section 6 addresses the problem of bias in estimates of the total in the presence of reporting delays. Chapter 2 concludes with a discussion.

2.2 Exploring Data

It is useful to start with an in-depth data exploration. In addition to measuring the overall length of birth lags, I have also examined lags by industry classified by the Standard Industrial Classification 1992 (SIC92) and by region. There is little to choose between most of the different industries. However, it is clear that Health and Social Work has longer birth lags than any other industry. Most regions have very similar average lags except for Northern Ireland, which stands out as having greater than average lags. I do not take differential reporting delays in industries and regions into account in this chapter.

As the focus is on undercoverage due to birth lags, the businesses of interest are those which came onto the frame *after* they were born. In addition to this stipulation I selected for further analysis only those businesses with birth between January 1, 1995, and Feb 28, 1998, to exclude the rightmost partly truncated diagonal in Table 2.1.

Table 2.2 and Figure 2.2 show some aggregates of births and the distribution of births. Except for the truncation effect clearly visible from November 1997 in Figure 2.2, the curve is astonishingly regular over time. Note that this curve represents the row sums of Table 2.1 apart from partially truncated cells. Note also that the scales of Figures 2.1 and 2.2 are very different: there is far more variability in counts between lags, especially short lags, than between birth months.

Table 2.2. Number of observed births per year and monthly average.

Born in year	Number	Average per month
1995	174,300	14,500
1996	172,600	14,400
1997	171,300	14,200
1998 (Jan and Feb)	19,000	9,500
Total	537,200	14,100

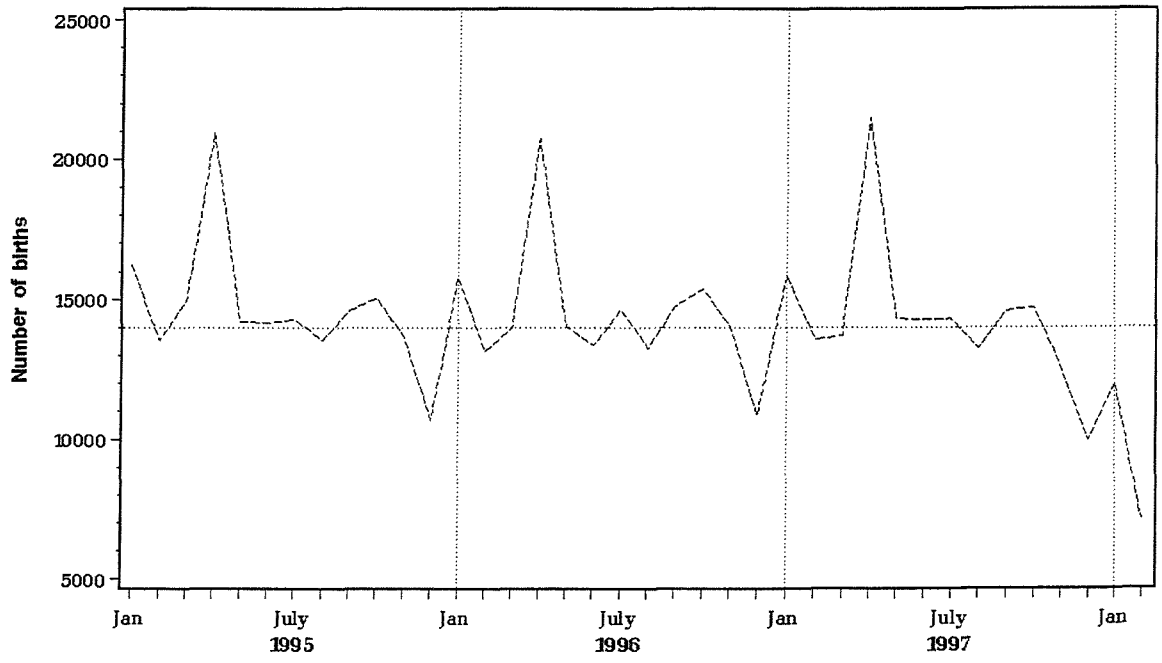


Figure 2.2. Number of observed births per birth month.

The longest birth lag we can fully observe is 37 months. Longer lags are entirely negligible as only 15 out of the 16,000 businesses that were born in January 1995 have 37 months birth lag; only 48 out of 30,000 born in either January or February 1995 have 36 months birth lag or more.

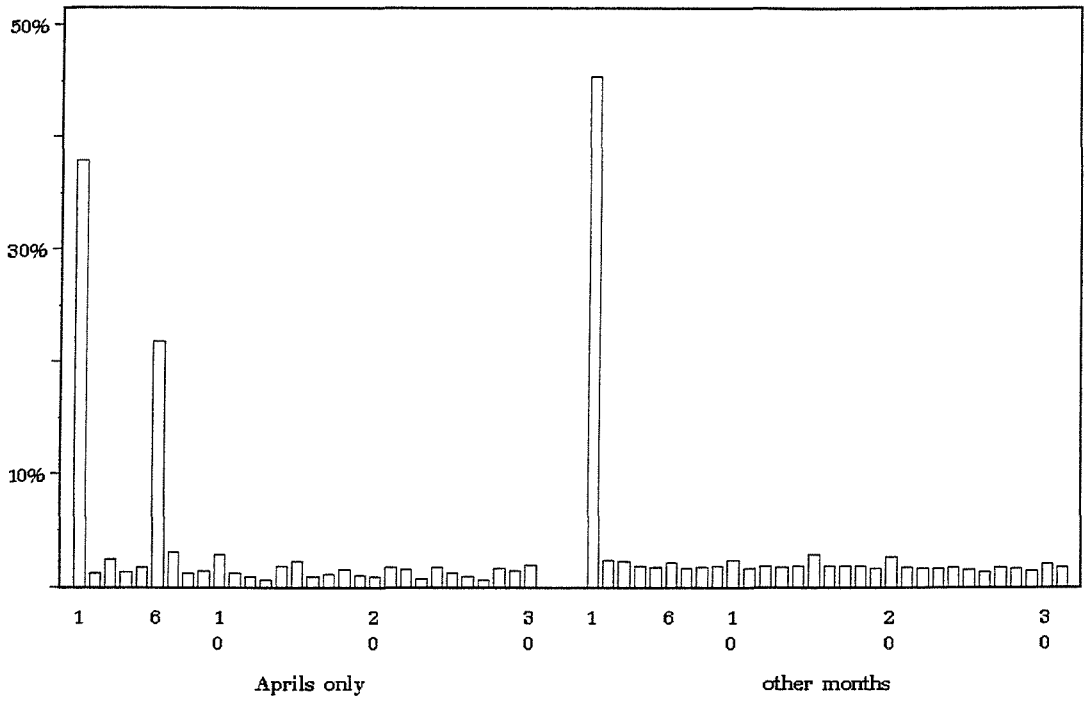


Figure 2.3. Percent of observed births per day of the month.

Figure 2.3 displays number of births by day for births in 1995 - 1997. The two panels contrast the distribution of birthday for Aprils with that of other months. In Aprils 38% of all new businesses started trading on the first of the month, in other months the proportion was even higher. The eye-catching peak at April 6 in Figure 2.3 is due to this day being the start of the taxation year in the UK. In practice, owing to differing interpretation of what constitutes the start of a business, it is frequently hard to fix on one day as the actual birthday for a business. The first of the month is often perceived as a convenient date for administrative purposes, both for the business managers and for the administrative bodies. Moreover, the first month of the taxation year, i.e. April, is a convenient month for administrative purposes, which seems to explain the higher birth rate in those months (see Figure 2.2). Also, there is some heaping visible in Figure 2.3 in that most of the bars for dates like 10, 15 and so forth are slightly taller than most other bars. Therefore, month seems to be the smallest viable unit in the classification of number of births; it does not seem meaningful to split months into smaller units.

The birth lags in months were computed as follows. The average length of a month was set to 30.4 days and the birth lag was set to the integer part of the ratio of the birth lag in days to 30.4. Since the frame introduction dates seem uniformly distributed over the days of the month, there is little risk of non-negligible approximation effects in the computation of the birth lags in months.

Figure 2.4 gives a contour plot of Table 2.1 with partly truncated cells excluded. The area with the largest counts is to the far left, and then the counts fall as we proceed to the right. The contour levels are 3, 21, 148 and 1096 (equal distances on a log scale), so area 1 consists of cells with counts greater than or equal to 1096. A couple of the 'islands' in area 4 are counts smaller than 3. The scarcity of islands in all areas indicates a large degree of homogeneity. The dashed horizontal lines mark Aprils. Areas 2 and 3, in particular, jut out along the dotted lines indicating areas with relatively large counts that are stretched to the right. This is partly due to the fact that there are more births in the month of April, partly due to a more skewed lag distribution for businesses born in April (the average birth lag is 2.5 months for businesses born in April and 1.6 months for businesses born in other months). There is also a diagonal pattern emerging in areas 2 and 3 above the horizontal line that indicates April 1996. The diagonals correspond approximately to frame introduction months; that is, businesses that came onto the frame in the same month are located along one or two diagonals running from right to left in the contingency table. It appears likely that what produces these diagonal ridges visible in Figure 2.4 is the reporting of births from Inland Revenue. Since this is done in a roughly quarterly basis, the notifications of new businesses come in sizable batches.

2.3 Models

The number of businesses in transition between birth and frame introduction can be viewed as a stochastic process over time. The process is not stationary since Figure 2.4 indicates among other things that birth lags tend to be longer for businesses born in April than for businesses born in any other time of year.

In this section I fit models to the upper triangle of the contingency Table 2.1, excluding partially truncated cells. It is convenient to confine the class of models to generalised linear models. A generalised linear model has a random component, which identifies the probability structure of a response variable Y, a link function which specifies the relationship between the expected value μ of the response and the systematic component, which in turn defines a linear function of the explanatory variables (e.g. McCullagh and Nelder 1989). The systematic component can rather easily accommodate the seasonality and the non-stationary structure we have observed. Another advantage with generalised linear models is that they are useful even if the parametric assumption underlying the model is ill-fitting, since the ML estimation of parameters uses only the link function, choice of covariates and the variance function $V(\mu)$, where $V(Y)=\phi V(\mu)$ and ϕ is known as the overdispersion parameter (Davison and Hinkley 1997, Ch. 7). Thus my approach is essentially semi-parametric.

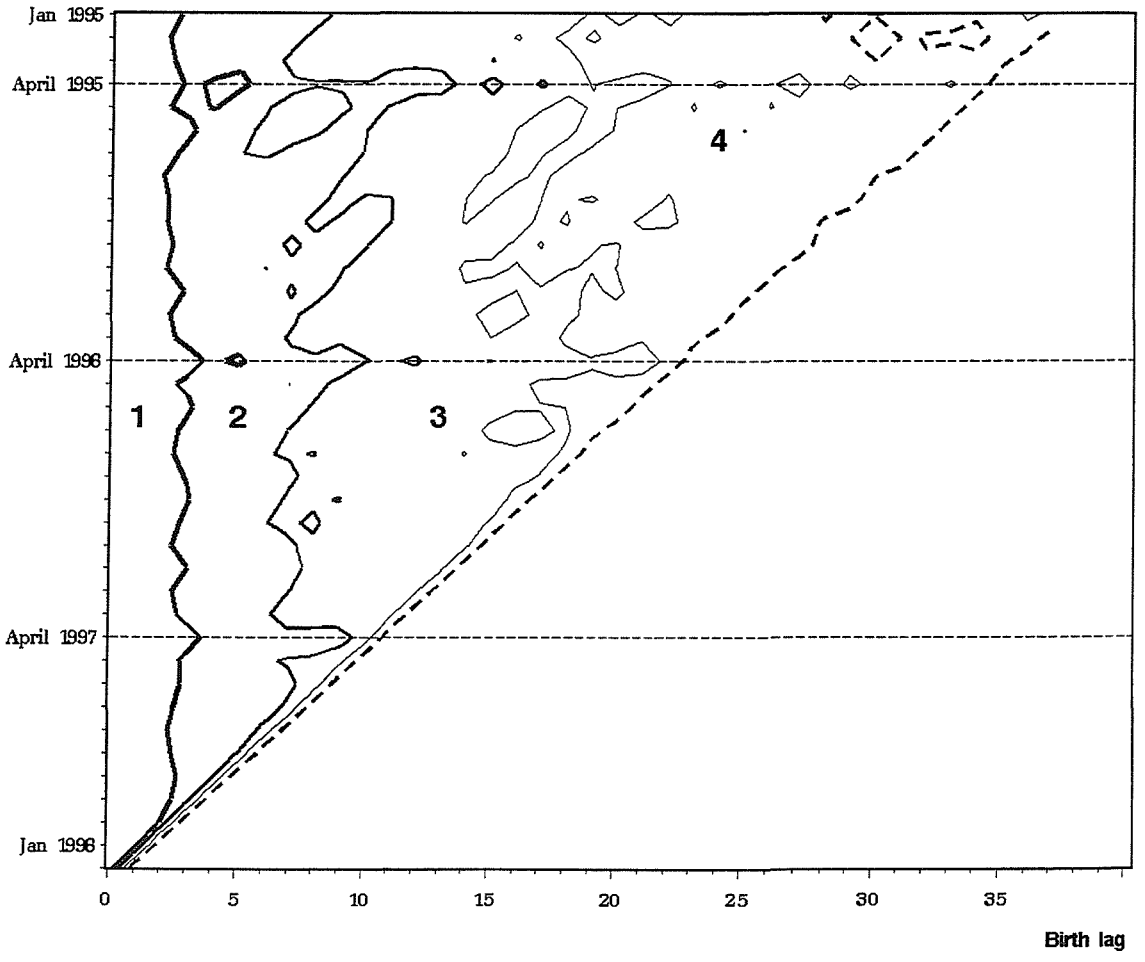


Figure 2.4. A contour plot of the contingency table, Table 2.1. Levels for number of frame introductions.

Let r be the number of rows in the table and let m_{ij} be the expected number of businesses that were born in month i, $i = 1, 2, \ldots, r$, and that were introduced on the frame in month $d = i + j - 1$, that is with a birth lag $j, j = 1, \ldots c$, where c is the maximum birth lag we can observe. For convenience, renumber the index j to start at 1 rather than at 0. We have seen that the birth rate is higher in some months, such as Aprils, than in other months. It seems plausible that a higher (or lower) birth rate for certain months should give roughly *proportionally* larger (or smaller) counts of new businesses for all birth lags, as opposed to an additive structure where the extra amount of births for Aprils, say, would be distributed equally over birth lags. Hence it seems more plausible that birth months, birth lags and other effects that potentially could be part of the systematic component are multiplicative than additive. This leads us to the following type of log-linear model:

$$\log(m_{ij}) = u + u_{(ij)}, \qquad (2.1)$$

for $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c - i + 1$, where u is an intercept and $u_{(ij)}$ is a parameter for cell i and j in the fully observed triangle in Table 2.1, with total number of rows r and columns $c = r$, here $r = 38$. Hence the link function is the logarithmic function, which conveniently converts multiplicative effects on the original scale to additive effects on the log scale. The variance function $V(\mu) = m_{ij}$ is reasonable even if the cell counts are not independent and Poisson distributed, since the overdispersion parameter can account for discrepancies between the variance of the response and the variance function (albeit only discrepancies that are constant over all cells).

One of the most parsimonious models (i.e. with fewest parameters) that we may be interested in is a log-linear model with just birth lag effects with $u_{(ij)} = u_{lag(j)}$, where $u_{lag(j)}$ is a parameter associated with birth lag j only. Considering Figures 2.1 and 2.2, which show that the range of number of births is 0 to 250,000 across birth lags and only 5,000 to 21,000 across birth months, the lag effect should be far more important than a birth month effect. The latter effect may perhaps even be dropped altogether. Although this may be an oversimplification, the model with a lag effect only is interesting as a reference model. Under this model all cells in a column have the same expected value. Another log-linear model arises from the assumption that the expected

cell counts are separable into quasi-independent row effects and column effects with $u_{(ij)} = u_{birthmonth(i)} + u_{lag(j)}$. See McDonald (1998) for a definition of quasi-independence and ML estimation for incomplete tables. Since the underlying stochastic process is not stationary, there is in fact an interaction between birth months and lags, which the quasi-independence model fails to capture.

Another model still is one with a seasonal effect and a lag effect. The underlying assumption is that some of the rows of the contingency table show a repetitive pattern in that their effects are the same and do not depend on year. Figure 2.2 suggests that all Januaries are similar, and so forth. It seems reasonable to examine a model with twelve 'season' parameters, as opposed to 38 birth month parameters. The model is

$$log(m_{ij}) = u + u_{season(k)} + u_{lag(j)}, \qquad (2.2)$$

$i = 1, 2, \ldots, 38,\ j = 1, 2, \ldots, 38 - i + 1,\ k = i \ (\text{modulo}\, 12)$.

When this model is fitted to the fully observed counts in Table 2.1, the residuals show a clear diagonal pattern, a pattern that is visible in Table 2.1 itself. Recall that businesses that came onto the frame in the same month are located along one diagonal and that reports from the PAYE source are obtained by the ONS on a roughly quarterly basis. Hence, the businesses whose births are reported from this source will tend to appear in ridges in the contingency table, from right to left, approximately three months apart. For example, 4,982 + 5,333 businesses in Table 2.1 came onto the frame in February 1995, 4,982 of which were born in January the same year (hence with one month's birth lag) and 5.333 in February 1995 (with no birth lag). Therefore, a diagonal effect can be added to the model to obtain a better fit. Further, an 'April effect' can accommodate part of the observed longer lags for businesses with births in April:

$$\log(m_{ij}) = u + u_{season(k)} + u_{lag(j)} + u_{diag(d)} + \alpha j I(k = 4), \qquad (2.3)$$

with i, j and k defined as for the model in (2.2), $d = i + j - 1$, α is a parameter and $I(\cdot)$ is an indicator function taking value 1 if the argument is true, 0 otherwise.

The models above were fitted to the fully observed upper triangle of Table 2.1 using ML estimation. The usual likelihood ratio test statistic (the 'G^2 statistic') and the Pearson chi-squared test statistic gave very similar results. The estimation of parameters was done with Proc Genmod in the SAS System® version 8.02 for

Windows, see Zelterman (2002). To ensure that the Genmod procedure gives correct results, it was run on some well-known datasets with structural zeroes. To check the numerical stability for the very large table analysed, the order of columns was changed, likewise the order of the rows for the model $u_{(ij)} = u_{birthmonth(i)} + u_{lag(j)}$, but the results remained the same.

Table 2.3 gives the values of test statistics for four models. The p-values are not given in the table below; all are miniscule. The G^2-values in Table 2.3 are extremely large due to the very large cell counts and the large number of cells. It is not meaningful in this application to use G^2-values for significance tests since any useful model would be rejected. We can, however, use G^2-values for the comparison of models without formal tests. Another general strategy for dealing with large counts in a contingency table is to look for non-random patterns among residuals for different models. I will also study how well the models predict future observations.

Table 2.3. Goodness of fit for Models 1–4.

Model	# para-meters	Degrees of freedom	G^2	Decrease in G^2	Knoke-Burke-ratio
1. Lags only	38	703	49,323		
2. Lags and seasons	49	692	38,259	11,064	22%
3. Lags and birth months	75	666	36,888	12,435	25%
4. Lags, seasons, diagonals and April effect	87	654	21,829	27,494	56%

The Knoke-Burke ratio (Knoke and Burke 1980) is $1 - G^2_{alt}/G^2_{ref}$, where G^2_{ref} is the value of the test statistic under a reference model (here Model 1, lag effect only) and G^2_{alt} under an alternative model that includes the reference model as a special case. Note that if the alternative model is the saturated model then the Knoke-Burke ratio attains its maximum, 100%. Knoke and Burke (1980) suggest that this ratio may be used for very large datasets; a large value indicates that the alternative model is satisfactory. I shall refer to the models using the order number in Table 2.3. Clearly, Model 4 gives the best fit. It is the addition of the diagonal effect that accounts for the major part of the reduction in G^2. Adjusted residuals from Model 4 are large but show

no clear pattern. While at least some reduction in G^2 is expected with an increasing number of parameters, it will be shown later on that this reduction is not entirely reflected in extra strength in prediction power when it comes to prediction of the lower part of Table 2.1.

There are other modelling approaches in the AIDS diagnoses literature. Harris (1990) and Wang (1992) discuss parametric and non-parametric methods, respectively, to estimate the size of the population. Davison and Hinkley (1997, examples 7.4 and 7.12) contrast what here is termed Model 3 with a generalised additive model which gives smoother predictions of nonobservable counts in a register of English and Welsh AIDS patients. Generalised additive models is a class of models that includes generalised linear models (Hastie and Tibshirani 1986). The link function in these models is a sum of nonparametric curve components. In my problem I could take $\log(m_{ij}) = u + u_{season(k)} + u(j)$ with $u(j)$ being some nonparametric curve describing the marginal relationship between cell counts and birth lags. Figure 2.1 suggests that the flat part of the curve may not need a different parameter for each birth lag, as they have in Models 1 - 4. I leave this idea for future research.

2.4 Predicting Number of Births

The models fitted to the upper triangle of the contingency table in Table 2.1 are now used for predicting counts in the lower triangle. To fix notation I first give a brief general account of Poisson log-linear models with 'matrix notation'. The contingency table has r rows, c columns and $rc = a$ cells. A general log-linear model is

$$\log(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{1}\mu, \qquad (2.4)$$

where $\mathbf{m} = (m_1, m_2, \ldots m_a)'$ is a vector of the expected cell counts, with the cells labelled from left to right starting with the first row, $\boldsymbol{\beta}$ is a parameter vector and the design matrix $\mathbf{X}$ specifies the model. The quantity μ is a parameter and $\mathbf{1}$ is a vector of ones. In order not to burden the notation I let the dimension of $\mathbf{1}$ be given by the context. Let o be the number of cells that are not structural zeroes (o for 'observed', a for 'all'). I denote the set of the fully observed cells by O and the set of all cells by A. The difference between A and O is denoted by S, which includes both partially

observed cells and cells with structural zeroes. I distinguish quantities that are defined for O only by a star. In the presence of structural zeroes the rows in (2.4) that correspond to them would not be included in the model. What remains of $\mathbf{m}$ and $\mathbf{X}$ after omission of rows that correspond to structural zeroes is denoted by $\mathbf{m}^*$ and $\mathbf{X}^*$. For example, consider a two-way table with $r = c = 2$ and without structural zeroes. Then a model with a row factor and a column factor and no interaction would have

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

If the fourth cell is a structural zero then

$$\mathbf{X}^* = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

If, for example, the cell a in a table with a cells is a structural zero the model is

$$\log(\mathbf{m}^*) = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{1}\mu, \qquad (2.5)$$

with $\mathbf{m}^* = (m_1^*, m_2^*, \ldots, m_{a-1}^*)'$ and $\mathbf{X}$ adjusted accordingly. In general, I have $\mathbf{m}^* = n_o\mathbf{p}^*$, where $\mathbf{p}^* = (p_1, p_2, \ldots, p_o)'$ is the vector of true probabilities under the Poisson distribution and n_o is the sum of the cell counts in O. Note that for a model pertaining to O only, $\mathbf{p}^*$ is not defined outside O. Thus

$$n_o\mathbf{p}^* = \exp(\mathbf{X}^*\boldsymbol{\beta} + \mathbf{1}\mu). \qquad (2.6)$$

Since the elements of $\mathbf{p}^*$ add up to unity, we obtain by summing over the columns of each side of (2.6)

$$n_o = \mathbf{1}'\exp(\mathbf{X}^*\boldsymbol{\beta} + \mathbf{1}\mu). \qquad (2.7)$$

and

$$\mathbf{p}^* = \exp(\mathbf{X}^*\boldsymbol{\beta}) / [\mathbf{1}'\exp(\mathbf{X}^*\boldsymbol{\beta})] \ . \qquad (2.8)$$

The estimator of $\mathbf{p}^*$ is

$$\hat{\mathbf{p}}^* = \exp(\mathbf{X}^*\hat{\boldsymbol{\beta}}) / [\mathbf{1}'\exp(\mathbf{X}^*\hat{\boldsymbol{\beta}})] \ , \qquad (2.9)$$

and that of μ is

$$\hat{\mu} = \log(n_o) - \log\left[\mathbf{1}' \exp\left(\mathbf{X}^* \hat{\boldsymbol{\beta}}\right)\right] , \qquad (2.10)$$

where the parameter vector β is estimated with, e.g., maximum likelihood estimation.

Let $T = T_o + T_s$, where T_o and T_s are the sum of observable and unobservable cell counts, respectively. Then it is natural to predict T by $\hat{T} = T_o + \hat{T}_s$, where $\hat{T}_s$ is a predictor for T_s. Under the natural assumption that (2.5) can for Models 1-3 be extended to model (2.4) by replacing $\mathbf{X}^*$ with $\mathbf{X}$ we have for cell i

$$m_i = \exp(\mathbf{X}_i'\boldsymbol{\beta} + \mu), \qquad (2.11)$$

where and $\mathbf{X}_i'$ is the ith row of $\mathbf{X}$. Thus $\mathbf{X}_i'$ corresponds to the ith cell in the contingency table. The parameters β and μ, which in (2.5) are defined for O only, will for Models 1-3 remain the same for A, with the predictor for cell i in S

$$\hat{m}_i = \exp\left(\mathbf{X}_i'\hat{\boldsymbol{\beta}} + \hat{\mu}\right). \qquad (2.12)$$

Hence the sum of the cell counts in the set S is predicted by

$$\hat{T}_s = \sum_{i \in S} \exp\left(\mathbf{X}_i'\hat{\boldsymbol{\beta}} + \hat{\mu}\right). \qquad (2.13)$$

For Model 4 it is assumed that the diagonal pattern observed for the last 12 months can be extrapolated periodically; that is, to predict cells along a diagonal d' in the part of the lower-right triangle where $c+1 \le d' < c+12$, the parameter associated with diagonal $d'-12$ in the upper-left triangle is used. To predict cells along a diagonal in the next band of twelve consecutive diagonals, $c+13 \le d'' < c+24$, the parameter associated with diagonal $d''-24$ is used, and so on. Thus, only the rightmost band of 12 diagonals in the observed triangle is used for prediction. While this may seem to underutilize the information, there does not seem to exist a periodic model for the diagonal effects that uses all observed diagonals and gives smaller prediction errors than the model just described that only uses the last 12 observed diagonals.

Table 2.4 gives the number of births aggregated to year levels. As seen in the table the observed count in 1997 is about 8-9% less than the predicted count. The difference between the sum of the predicted counts under Model 4 and the observed count is

570,000–542,000 = 28,000. Hence, in terms of number of businesses the undercoverage due to reporting delays is about 1.6% (28,000 on 1.8 million).

Table 2.4. Observed number of births per year and the ratio predicted counts to observed counts.

		Ratio predicted count to observed count			
Year	Observed number of births	Model 1	Model 2	Model 3	Model 4
1995	175,898	1.00	1.00	1.00	1.00
1996	174,013	1.01	1.01	1.01	1.01
1997	172,570	1.09	1.08	1.09	1.08
1998	19,103	1.75	1.74	1.92	1.69

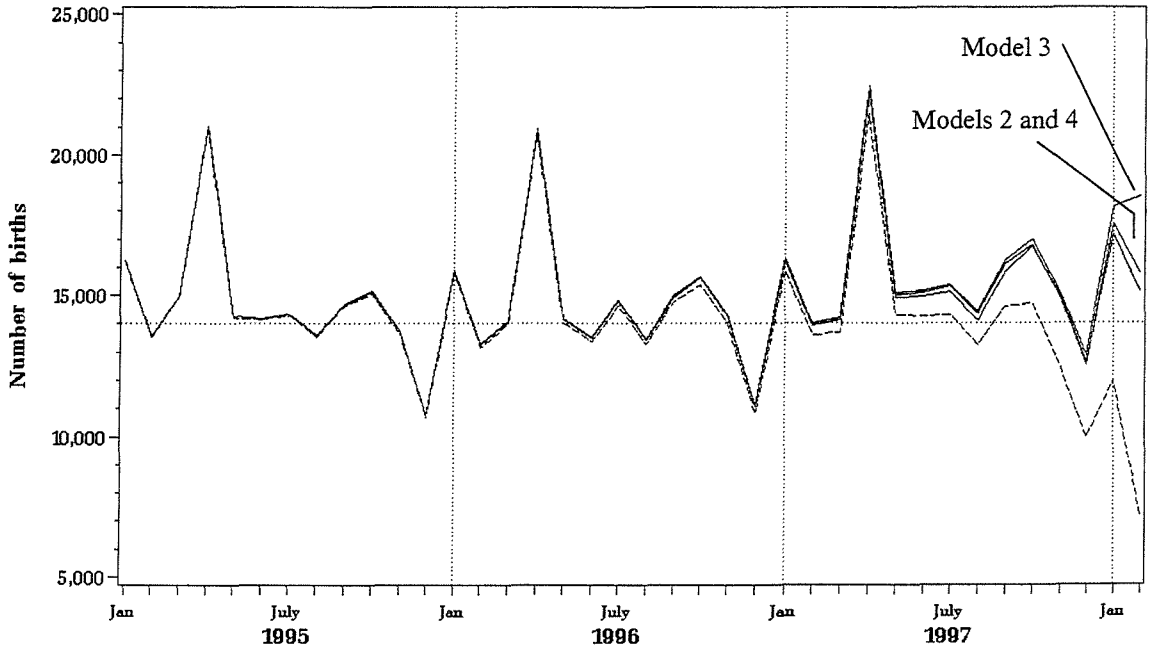


Figure 2.5. Predicted number of births per month under Models 2-4. The observed counts are graphed with a dashed line.

Figure 2.5 shows the observed and predicted number of births per month for Models 2-4, as numbered in Table 2.3. The dashed curve in Figure 2.5 is the same one as in Figure 2.2. Judging from Figure 2.5 there is little to choose between the prediction

methods with only Model 3 being somewhat separated from the others. There is a 1% truncation effect as early as September 1995 that each model captures.

2.5 Prediction error

To assess the prediction error, we can turn the clock backwards, for example to the end of May 1995, and pretend that all observed businesses born afterwards are unknown. Hence there will be a 5x5 square subtable with observed counts in the upper-left triangle and 'missing' counts in the lower-right triangle. A natural estimate of the error is obtained by estimating parameters for the upper triangular subtable and basing the prediction error on the difference between the observed and predicted counts in the lower-right triangle. Using this approach, Figure 2.6 shows the number of births per month for data cut off at the end of April 1997. The dashed curve is the number of births per month obtained from the full original table (that is, it is the same curve as in Figure 2.2). Models 3 and 4 are indistinguishable while Model 2 predicts the rise in births in April rather better than the other models.

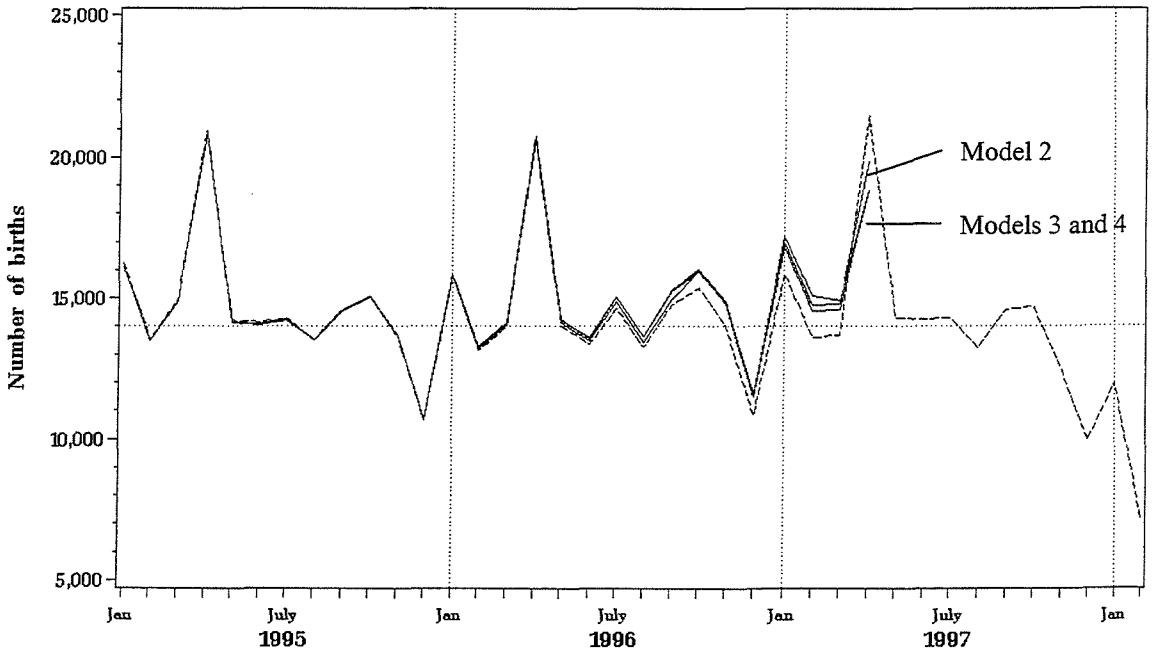


Figure 2.6. Predicted number of births based on data up to 30 April 1997: Models 2-4 and observed counts as at Feb 28 1998 (dashed line).

Thus the ends of the solid curves in Figure 2.6 show the predicted number of births for the month that corresponds to the last row of the particular triangular subtable which has been obtained by cutting the full table off at the end of April 1997. Figures 2.7 and 2.8 exhibit the prediction errors for a series of subtables, from the one obtained by cutting off at the end of December 1995 to the one where data after December 1997 were discarded. In Figure 2.7 the final-month errors are shown, defined as the difference between the predicted number of births in the last month of the subtable and the observed number of births in the same month in the part of the original table covered by the subtable. The part of Figure 2.7 to the right of July 1997 is clearly influenced by the bias resulting from truncation of the original series. In the beginning of the series the error is as expected large due to the fact that in the beginning of the series there is less data for the estimation of parameters. It seems reasonable to forego the prediction errors before July 1996 and after July 1997.

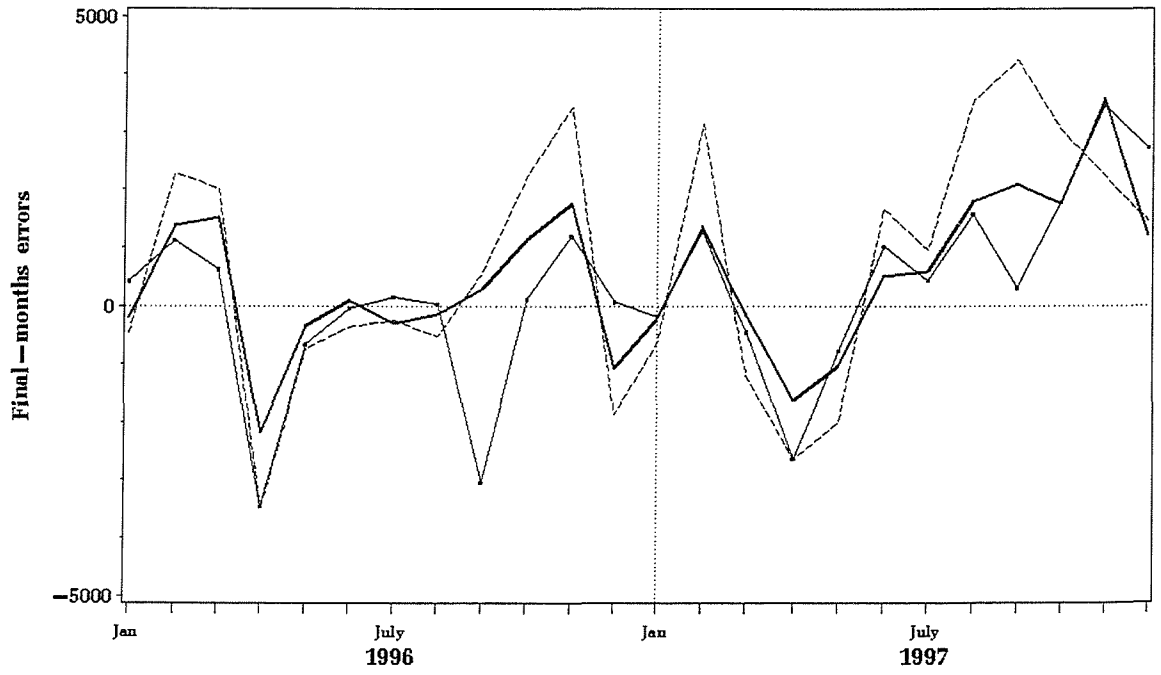


Figure 2.7. Difference between predicted and observed number of births for the final month in successive subtables. The months along the x-axis represent the final month in each subtable. Three models: Model 2 (thick), Model 3 (dashed), and Model 4 (thin).

As seen in Figure 2.7, Model 2 gives smaller final-month errors than Model 3 for each month in this interval. This may seem paradoxical since Model 3 has more parameters

and gave a better fit to the upper triangle of the contingency table (Table 2.3). However, the models play two roles here. One is to fit counts in the upper triangle of the contingency table. The other is to be a tool for prediction. Good performance in one of these roles does not necessarily imply good performance in the other. Model 3 does not draw on the seasonal pattern. Stated somewhat loosely, Model 2 borrows strength from similar months in previous years. With Model 3, the predictions depend completely on single rows of the table and are much more variable. Model 2 has the additional advantage over Model 3 that it allows prediction beyond February 1998. As seen in figure 2.7, Model 4 often gives smaller errors than Model 2, but certainly not always.

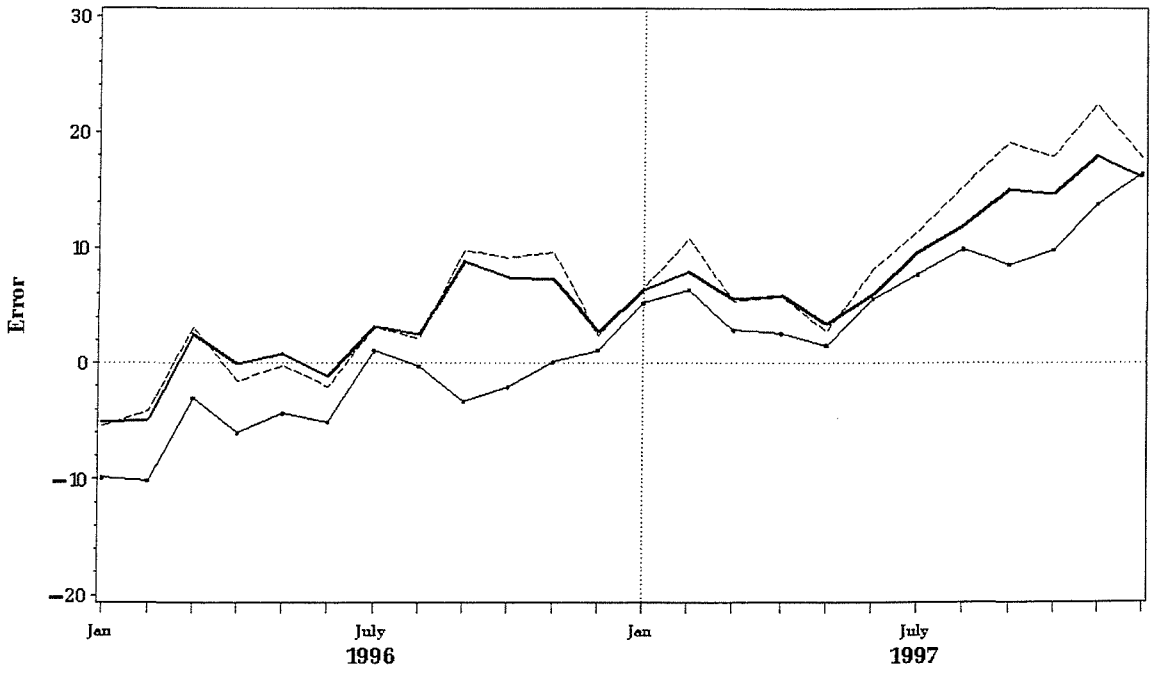


Figure 2.8. Difference in 1000s between the sum of predicted number of births and observed number of births in successive subtables. The months along the *x*-axis represent the final month in each subtable. Three models: Model 2 (thick), Model 3 (dashed), and Model 4 (thin).

The difference between the sum of monthly predictions and observations is a measure of error more directly connected to the estimation of the undercount. These differences for a sequence of subtables are displayed in Figure 2.8. In the beginning of the series the difference is negative because the predictions for 1995 are too low. The difference becomes positive when the truncation effect in the original series becomes pronounced.

Figure 2.8 makes it clear that Model 4 is better than Model 2. As seen in Figure 2.8, the largest prediction error in absolute terms for Model 2 in the interval July 1996 – July 1997 is less than 10,000. For Model 4 the largest error is less than 6,000.

2.6 Bias Resulting from Reporting Delays

The undercoverage will lead to a negative bias in an estimate of the total. Suppose the aim is to estimate the total $t_y = \sum_U y_k$ of a study variable $\mathbf{y}' = (y_1, y_2, \ldots, y_N)$ on a population U with unit labels $\{1,2,\ldots, N\}$. Let U_{ij} be the population of businesses with birth month i and reporting delay j. The total of the unseen part of the population, t_{Us}, is the sum of $t_{yij} = \sum_{Uij} y_k$ over the not fully observed cells (i, j) in Table 2.1, each of which holds the population U_{ij}.

While the available reporting delay data do not contain any study variable, the variable turnover at frame introduction was stored for the businesses whose counts are reported in Table 2.1. I will perform the numerical analysis on this auxiliary variable. To estimate the bias with respect to a study variable, for example capital expenditure, one approach would be to model the relationship between this variable and turnover at frame introduction. I leave this for future research.

I draw on actuarial science to find a method for predicting t_{Us}, which is in that context interpreted as, for example, the sum of incurred but not reported (IBNR) losses for which the clients are insured. The *chain ladder* method is widely used in insurance practice. For this method transferred to the current issue, consider an auxiliary variable x_k, $k = 1, 2, \ldots N$, and let $C_{ij} = \sum_{t=1}^{j} t_{xit}$ be the cumulative totals of the auxiliary variable for businesses with birth month i and birth lag not longer than j. Introduce the *development factors*

$$\hat{\lambda}_j = \left(\sum_{i=1}^{r-j+1} C_{ij} \right) \left(\sum_{i=1}^{r-j+1} C_{i,j-1} \right)^{-1},$$

where $j \leq r$ and $r = c$ is the total number of rows (columns) in the table. The development factors are applied to the largest observed cumulative total in row i, that

is $C_{i,r-i+1}$ to give an estimate of the cumulative total for the subsequent columns in row i:

$$\hat{C}_{i,r-i+2} = C_{i,r-i+1}\hat{\lambda}_{r-i+2},$$

$$\hat{C}_{i,r-i+3} = C_{i,r-i+1}\hat{\lambda}_{r-i+2}\hat{\lambda}_{r-i+3},$$

and so on. Hence the assumption, for simplicity expressed here for unobservable cell $(2,c)$ only, is that

$$\frac{C_{1,c-1}}{C_{2,c-1}} = \frac{C_{1c}}{C_{2c}}.$$

Mack (1991) and Renshaw and Verrall (1998) show that the chain ladder technique necessarily gives the same cell predictions as the quasi-independence model, which is labelled Model 3 in this chapter. An extension of the chain ladder technique is thus to apply my Models 2 and 4 to observed totals of some frame variable to predict non-observed cell totals of this variable.

There are other approaches in actuarial science. In the widely used Bornhuetter-Ferguson technique (Bornhuetter-Ferguson 1972), the C_{ic} are taken as though they were known constants obtained from some external source and the only free parameters are the lag parameters. Using an argument from credibility theory, Mack (2000) discusses the approach where the final predictions are linear combinations of the Bornhuetter-Ferguson predicted values and the predictions obtained through the chain-ladder method. Overviews of the IBNR prediction problem are given by England and Verrall (2002) and De Vylder (1996, Ch. 7). It is usual to assume stationarity for IBNR prediction.

Alternatively, one can fit a model to the frame variable to obtain an estimate of the expected value in each cell and multiply this with the predicted number of units in that cell. Klugman, Panjer, and Willmot (1998, p. 292) argue that modelling counts and the continuous variable separately has some advantages in the IBNR losses context. In my situation it is useful to compare the distribution of the study variable for different birth lags with that of the counts. Also, to investigate the impact of legal and procedural changes (for example if the VAT threshold for mandatory reporting to the relevant UK authority changes or if new proving processes are introduced at the ONS) it is helpful to model the distribution of the counts and the study variable separately to avoid confounding. I will not pursue this approach here.

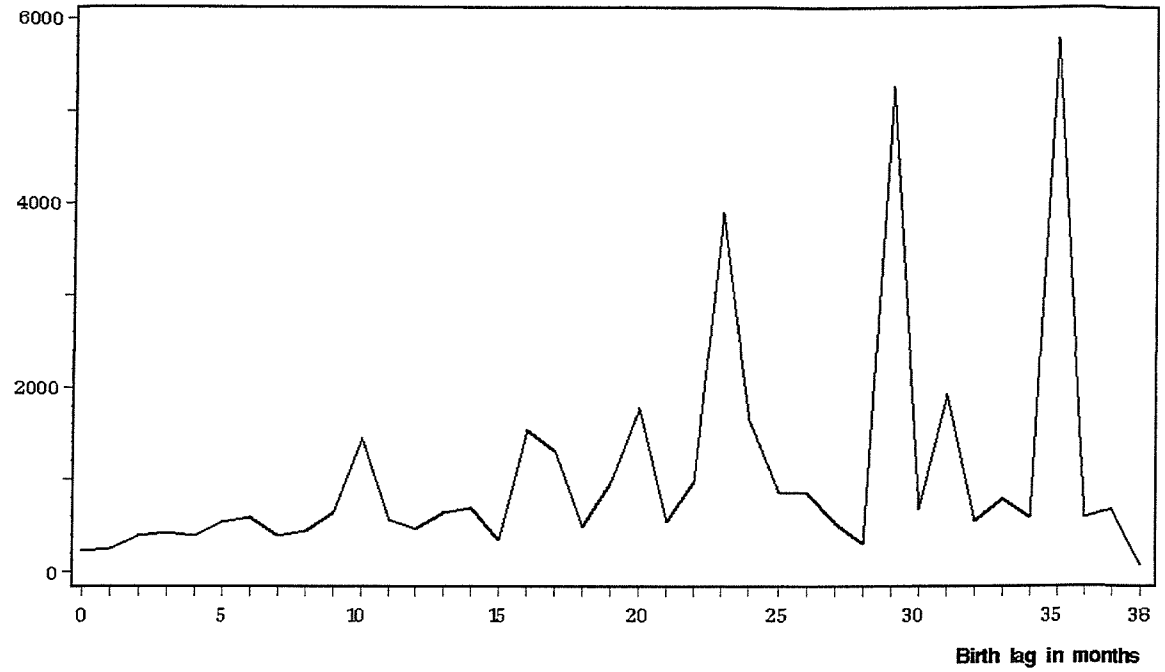


Figure 2.9. Average turnover in £000 at frame introduction against birth lag.

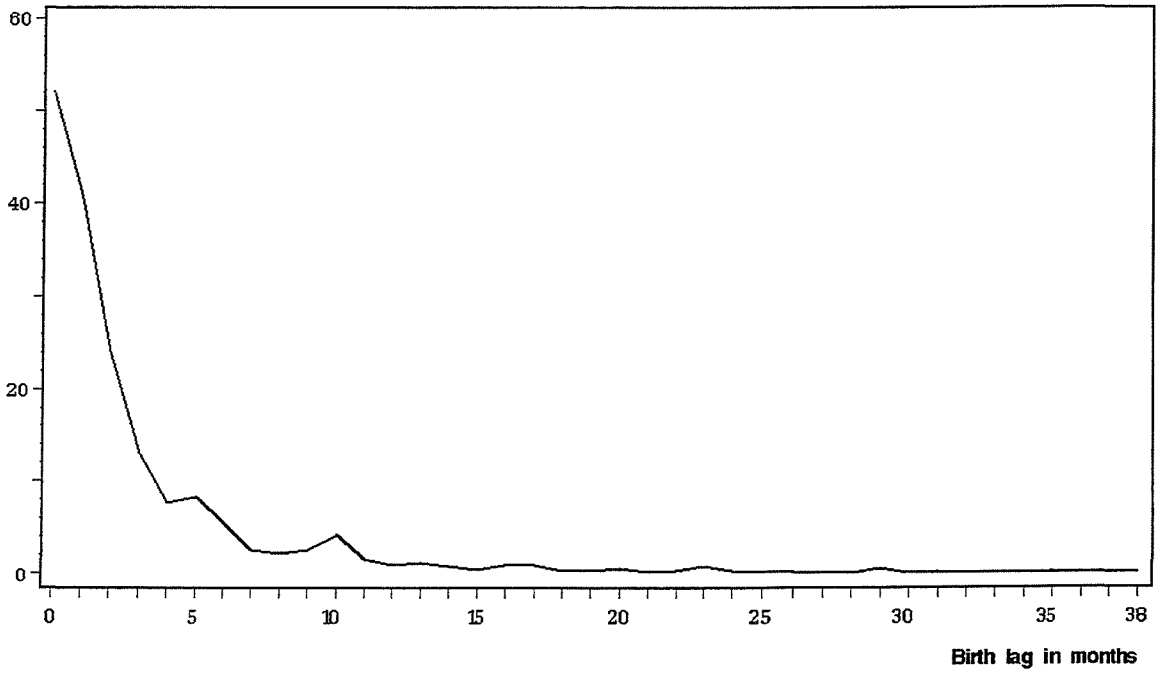


Figure 2.10. Total turnover at frame introduction in £billion against birth lag (months).

Figure 2.9 exhibits the fact that businesses that are very large in terms of turnover when they come onto the frame tend to have long birth lags. It is believed that few of

these large businesses are genuinely new; they are results of mergers and other types of restructuring. To avoid duplication large businesses that are reported as new are subjected to an often lengthy proving process which can not usually be done without the help of the business itself. However, there is little information stored on the frame on the history of a business.

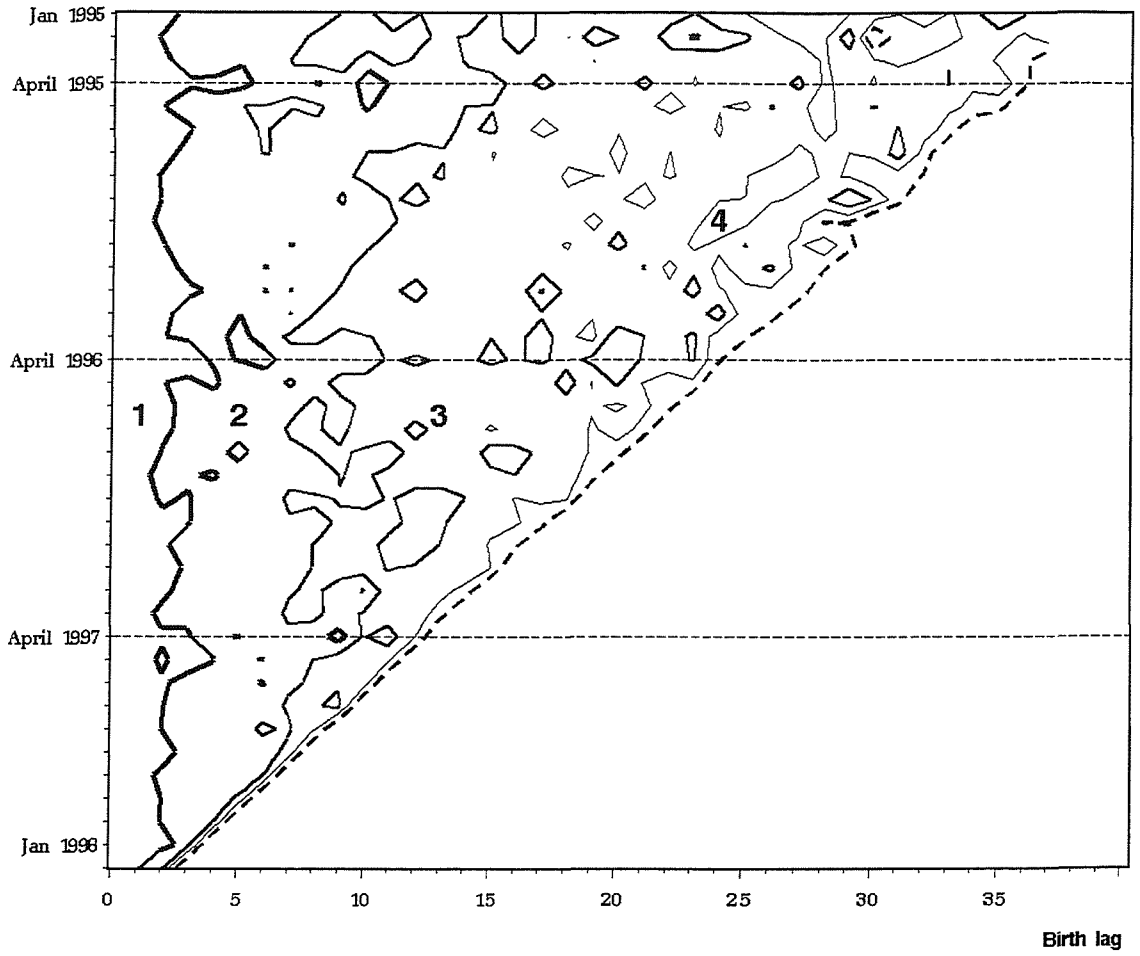


Figure 2.11. A contour plot of levels for total turnover at frame introduction. The levels are 54, 1000, 22000 and 1.2m (all in £1000).

Figures 2.10 and 2.11 show the distribution of total turnover at frame introduction against birth lag and birth month. The similarity of these to Figures 2.1 and 2.4 suggests that it may be possible to model the cell totals of turnover with the methods applied to the counts. Cross-validation errors that parallel those of Figure 2.7 are displayed in Figure 2.12 . The estimated total undercoverage is 2.400 £billion (1000 million). Unfortunately, the prediction errors displayed in Figure 2.12 are of similar size as the point estimate of the error caused by reporting delays. The large businesses with long lags, clearly visible in the contour plot but also in Figure 2.10, make

prediction intrinsically difficult. They enter the frame irregularly and produce large variation in total turnover per birth month.

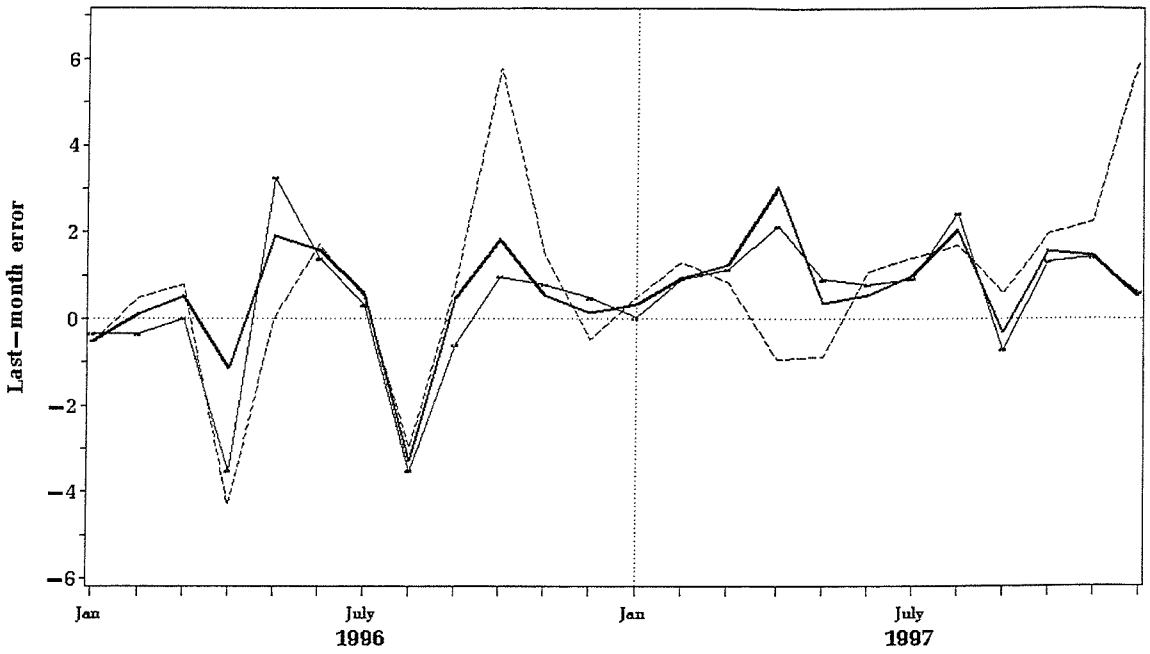


Figure 2.12 . Difference in £bn between predicted and observed number of births for the final month in successive subtables. The months along the *x*-axis represent the final month in each subtable. Three models: Model 2 (thick), Model 3 (dashed), and Model 4 (thin).

2.7 Discussion

Undercoverage is arguably the most important type of frame imperfection. I believe that the work initiated here provides a useful measure of frame quality. A time series of the undercoverage as estimated each month in terms of number of businesses is a useful tool for monitoring frame quality. For example, a long-term increase will spur questions about what developments in the processes causes the changes in the reporting delay distribution.

I have predicted gross totals with a log-linear model. The prediction error was estimated with a non-parametric method that has considerable natural appeal. At the

end of February 1998 the undercount was 28,000 businesses, or 1.6% of all registered businesses. The error of this estimate was predicted to be less than 6,000.

The sum of the turnover of the unobservable businesses was not possible to predict with any accuracy due to a heavy tail in the reporting delay distribution. The heavy tail is due to the fact that many businesses that are very large when they enter the frame are not genuinely new businesses. Since the history of businesses is currently not stored on the business register of the ONS, it has been proposed to create a new life status variable that will store more complete information about changes to businesses. This will be a log of events that have occurred in the life of the business and allow the separation of genuinely new businesses from businesses that are new only in a legal sense.

Chapter 3

Feeding Back Information on Ineligibility From Sample Surveys to the Frame

3.1 Introduction

To facilitate estimation of change, consecutive samples in many repeated survey are overlapping. If several surveys draw samples from the same frame, it is often desirable to spread the response burden by making sure that samples for different surveys are not overlapping to a greater extent than necessary. This is particularly desirable if the frame is moderately large and used for many continuing surveys, which is a situation that many national statistical institutes face when conducting business surveys. Stratified simple random sampling is a very common design for business surveys. The skewed distributions of business study variables call for large sampling fractions in many strata, which aggravate the response burden for medium size and large businesses. Both response burden issues and estimation of change are of paramount importance in official business statistics. Therefore, sampling systems have been constructed that allow the organisation to co-ordinate samples, either positively or negatively (i.e. to create overlap or to make sure that there is little overlap).

For example, the ONS uses the *Permanent Random Number* (PRN) technique, which is a widely used method for drawing samples from lists. A PRN, drawn from the uniform distribution on [0,1], is attached to each frame unit independently of each other and independently of the unit labels and any variables associated with the units. The units can be plotted on a line starting at 0 and ending at 1 and I refer to this line as the *PRN line*. To draw a simple random sample without replacement, an *SI*, with a predetermined sample size n, a point is selected (randomly or purposively) on the PRN line and the n units to the

right, say, are included in the sample. For overviews and further details see Ohlsson (1995) and Ernst, Valliant and Casady (2000). Table 3.1 shows starting points of sampling intervals of some of the business surveys the ONS conducts on a regular basis.

Table 3.1. Starting points of the PRN sampling intervals of some of the business surveys the ONS conducts

Survey	Starting point of sampling interval
The Monthly Inquiry for the Distribution and Services Sector, and other monthly surveys covering other sectors of the business population	0
The Quarterly Capital Expenditure Inquiry	0.125
The UK Survey of Products of the European Community	0.375
The Inquiry of Stocks	0.5
The Annual Business Inquiry	0.625
The Annual Employment Survey	0.75

Samples for repeated surveys can also be selected with a panel technique where the sample at the first wave of the survey is partitioned into a set of rotation groups. At the second wave one, say, of the groups is replaced with a new sample of the same size as the outgoing sample and the other groups are retained in the sample.

There are in principle two main sources of data that are used to maintain a frame: administrative ones and surveys. As was described in Chapter 2, various administrative bodies send tapes to the ONS on a regular basis with information of, e.g., births and deaths of businesses. While these tapes are sent in to the ONS very frequently, the distribution of the time it takes for a new unit or an alteration of one old unit to be registered on the frame is highly skewed. Due to frame maintenance procedures, there is also very often a considerable difference in time between the actual and formal termination of a business. Therefore, most of the business surveys at the ONS share the information on deaths they obtain through their samples with other business surveys to speed up the information

process. In this chapter, I examine the effects of using sample surveys to update a frame that is used for repeated surveys. This is in principle how information of dead units is treated in business surveys at the ONS and some other national statistical institutes.

It would seem natural that this new information should be made available to other sample surveys, which otherwise may include the dead units in their samples and therefore lose precision. However, as has been pointed out by Srinath (1987) among others, such a procedure may cause bias. I refer to this as *feed back bias*, which results whenever the sampling mechanism is not independent of the feed back procedure. For example, consider a situation where all dead units in the sample are deleted at the first wave of a panel survey. If no further deaths have occurred up to the second-wave observation of the panel units, the second-wave sample contains only live units. Without knowledge of the total number of live units in the population at the time of the second wave, an unbiased estimator of the total cannot be constructed. While more information about the population has been gathered when the deaths were recorded at the first wave, there is actually less information in the second wave-sample on the proportion of live units in the population. While the existence of feed back bias has been long recognised, little study has been made of the size of the feed back bias. I show how an estimate of the number of live units in the population can be used to construct an approximately unbiased estimate of the population total.

A safe recommendation would be that no information on deaths from sample surveys, other than from completely enumerated strata, may be used to update the frame when samples are co-ordinated over time (cf. Ohlsson 1995, p. 168, and Colledge 1989, p. 103). However, to prohibit feeding back seems to deny oneself the use of all available information. I obtain an expression for the size of the feed back bias and show that the feed back bias can be estimated and used to adjust conventional estimators. Schiopu-Kratina and Srinath (1991) adjust the sampling weights to counter an expected too low proportion of dead units in the rotating sample of the Survey of Employment, Payroll and Hours conducted by Statistics Canada. Hidiroglou and Laniel (2001) discuss the feed back issue briefly. A general discussion of frame issues is given by Colledge (1995) and overviews of issues associated with continuing business surveys include

College (1989), Hidiroglou and Srinath (1993), Srinath and Carpenter (1995), and Hidiroglou and Laniel (2001).

Instead of the terms 'eligible' and 'ineligible' I use the more emotive words *dead* and *live*, although my reasoning does cover all kinds of ineligibility. The discussion is confined to the estimation of the total

$$t_y = \sum_U y_k \qquad (3.1)$$

of some study variable $\mathbf{y}' = (y_1, y_2, ..., y_N)$ on a population U with unit labels $\{1,2,...,N\}$. When the sampled units are observed, I assume that *all* dead units in the sample are correctly classified as dead and the frame is updated with this information. This may be difficult in practice. In some surveys, however, eligibility of all nonresponding units can be correctly identified.

Section 3.2 introduces the necessary notation and concepts and gives an expression for the feed back bias when estimating a total. Section 3.3 discusses three strategies that may be used in the presence of feed back and compares these in a simulation study. Chapter 3 concludes with a discussion.

3.2 An expression for feed back bias

I assume throughout that a dead unit is always out of scope and that the value of the study variable of a dead unit is always zero. (It is conceivable that dead units are eligible in some surveys; for example, a business survey collecting data on production may have defined businesses that were alive at least a part of the reference period as eligible.) I adopt the design-based view that the study variable values are fixed and non-stochastic at any given point in time. The situation addressed is as follows. One or more samples are drawn from the frame which comprises the *original survey population*, U_{orig}. For convenience it is assumed that the frame units and population units are of the same type. The updated frame, where all dead units that have been included in samples from U_{orig} have been excluded, is referred to as the *current survey population*, $U_{current}$. For example, two surveys may simultaneously work with a sample each, and after they have fed back,

U_{orig} has shrunk to $U_{current}$. In this chapter I disregard births of new units and other deaths than those deleted through samples from U_{orig}. I also disregard undercoverage, nonresponse and measurement errors. In practice, administrative sources will provide information on deaths. They work independently from the sampling procedures employed by the statistical agency and will therefore not contribute to feed back bias. These units are *dead by administrative sources*. We can think of these dead units as being excluded from the population. While the sampling design is here assumed to be SI, it can readily be extended to stratified simple random sampling.

Let U_d and U_l be the two subsets of the current survey population, $U_{current} = U_d \cup U_l$, that consist of dead and live units, respectively. All units in U_d and U_l are assumed to be flagged on the frame as live. Units that are flagged as dead but for which the independence of detection and the sampling mechanism cannot be assured are called *dead by sample survey sources*. In our set-up, these are the dead units detected in samples taken from U_{orig}. Let the set of these units be denoted by U_{sd}, and we have the relationship $U_{orig} = U_{current} \cup U_{sd}$. Let N with a proper subscript be the size of each population, respectively. Then $N_{current} = N_l + N_d$, and $N_{orig} = N_l + N_d + N_{sd}$. At the time when samples are drawn from $U_{current}$, $N_{current}$ and N_{sd} are known numbers, whereas N_l and N_d are unknown. Moreover, N_{sd}, N_d and $N_{current}$ could be viewed as random quantities depending on feed back results, while N_l is fixed. Following principles of Durbin (1969) and more recently in Thompson (1997), we would in many situations prefer to condition on N_{sd}. For example, if it is seen that U_{sd} is in fact empty, then it does not seem appropriate to include in the inference the possibility that N_{sd} could have been large. However, to analyse the development of the feed back bias over a series of waves in a panel survey when planning the survey, unconditional analysis would be preferable. An expression for the unconditional feed back bias is also obtained below.

Denote by $U_{nodeads}$ the part of $U_{current}$ that was covered by the previous sample(s) drawn from U_{orig}; see Figure 3.1. Clearly, $U_{nodeads}$ is a random set depending on previous samples. Since $U_{nodeads}$ is winnowed from dead units we have $U_{nodeads} \subset U_l$. The

complement in $U_{current}$ to $U_{nodeads}$, denoted by $U_{withdeads}$, encompasses all of U_d and a part of U_l: we have

$$U_{nodeads} \cup U_{withdeads} = U_l \cup U_d = U_{current}.$$

The only sets introduced so far that are non-random are U_{orig}, U_l, and $U_{withdeads}$. The latter is viewed as non-random since it is assumed that $U_p = U_{nodeads} \cup U_{sd}$, i.e. the part of U_{orig} covered by previous samples taken from U_{orig} (subscript p for 'previous'), have a total sample size determined by design.

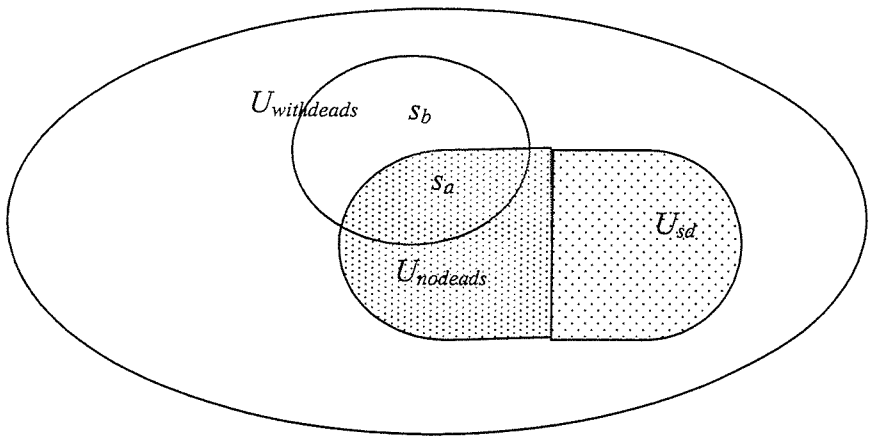


Figure 3.1. The original survey population, U_{orig}, and its subsets $U_{orig} = U_{withdeads} \cup U_{nodeads} \cup U_{sd}$. The sample from $U_{current} = U_{orig} - U_{sd}$ consists of two sample parts, s_a and s_b.

To derive the feed-back bias I will first obtain the inclusion probabilities. To do this, it is useful to consider a sample of size n with a sample part s_a of size n_a taken from $U_{nodeads}$ through PRN sampling or a panel sampling technique, and the remaining part s_b is taken from $U_{withdeads}$. If the sampling is done with a panel technique, the sample parts s_a and s_b are the old and new rotation groups, respectively. If the sample is drawn with PRN sampling, s_a and s_b consist of units with PRN's that fell, or did not fall, in the samples from U_{orig}, respectively. Whether the sample was drawn through PRN sampling or a panel sampling technique, the sample parts can be viewed as two fixed size samples, each drawn with the SI design from their respective subpopulation. I will condition on the outcome of n_a and n_b throughout. With the notation $(k \in s_a)$ I refer to the event that a unit is first

included in the first-wave sample from U_{orig} and then in the second-wave sample taken from what remains of the first-wave sample after dead units have been taken out. The notation $(k \in s_b)$ is analogous. Let $I(k \in s_a) = 1$ when unit k is included in s_a, otherwise $I(k \in s_a) = 0$.

Recall that $y_k = 0$ if k is a dead unit and that $U_{current} = U_d \cup U_l$. Thus we have $\sum_{s_a} y_k = \sum_{U_l} y_k I(k \in s_a) = \sum_{U_{current}} y_k I(k \in s_a)$ and, assuming that $N_l > 0$,

$\Pr[k \in s_a \mid N_{sd}] = \frac{n_a}{N_l}$, since a sample of size n_a is effectively selected from a population of size N_l with the SI design (through an SI sample from U_{orig} followed by an SI sample from U_{nodead}). Note that a unit k in s_a must be alive since $U_{nodeads}$ consists solely of live units and only live units can be included in s_a.

To derive the overall bias it is convenient to analyse the biases from the sample parts s_a and s_b separately. I derive an expression for each of these, and they will be amalgamated in (3.10) below. Denote the bias of an estimator $\hat{\theta}$ for the parameter θ by $B(\hat{\theta}, \theta)$. Then with respect to the population total $t_y = \Sigma_{U_{current}} y_k$, the bias of a general linear estimator $\hat{t}_y^{(s_a)} = \sum_{s_a} w_k y_k$ based on s_a, with any given w_k's, is

$$B(\hat{t}_y^{(s_a)}, t_y | N_{sd}) = \sum_{U_l} \{w_k E[k \in s_a \mid N_{sd}] - 1\} y_k = \Sigma_{U_l} \left(\frac{w_k n_a}{N_l} - 1 \right) y_k$$

$$= \sum_{U_{current}} \left(\frac{w_k n_a}{N_l} - 1 \right) y_k . \qquad (3.2)$$

For the sample part s_a, the naïve expansion estimator that ignores the feed back bias would have weights $w_k = N_{current}/n_a$. From (3.2) we see that the bias of this estimator, $\hat{\tau}_y^{(s_a)} = \frac{N_{current}}{n_a} \sum_{s_a} y_k$, is

$$B(\hat{\tau}_y^{(s_a)}, t_y | N_{sd}) = \frac{N_d}{N_l} t_y . \qquad (3.3)$$

Alternatively, the sampling of s_a can be seen as a two-phase sampling scheme, where the first phase is the set $U_{nodeads}$ being 'selected' from U_{orig}. The first phase inclusion probabilities are

$$\Pr\left[k \in U_{nodeads} \middle| k \text{ alive}, N_{sd}\right] = \Pr\left(k \in U_{nodeads}, k \text{ alive} \middle| N_{sd}\right) / \Pr\left(k \text{ alive} \middle| N_{sd}\right)$$

$$= \frac{N_{nodeads}}{N_{orig}} \Big/ \frac{N_l}{N_{orig}} = \frac{N_{nodeads}}{N_l} . \qquad (3.4)$$

Thus,

$$\Pr\left[k \in s_a \middle| N_{sd}\right] = \frac{N_{nodeads}}{N_l} \frac{n_a}{N_{nodeads}} = \frac{n_a}{N_l} . \qquad (3.5)$$

Note that $N_{nodeads}$ (and thus N_{sd}) cancels out. The probability of $(k \in s_a)$ depends on the feed back process to have taken place but not on the size of U_{sd}.

Next, to derive the bias for the sample part s_b of size n_b taken from $U_{withdeads}$, first note that $(k \in U_{withdeads})$ is the same event as $(k \notin U_p)$, where $U_p = U_{nodeads} \cup U_{sd}$. Then

$$\Pr\left[k \in U_{withdeads} \middle| N_{sd}\right] = \Pr\left[k \notin U_p \middle| N_{sd}\right] = \frac{N_{orig} - N_p}{N_{orig}} = \frac{N_{withdeads}}{N_{orig}} . \qquad (3.6)$$

This conditional probability again does not depend on the relative sizes of $U_{nodeads}$ and U_{sd}. On the other hand, the probability of including a unit in s_b given that feed back has occurred is

$$\Pr\left[k \in s_b \middle| k \in U_{withdeads}, N_{sd}\right] = \frac{n_b}{N_{withdeads}} . \qquad (3.7)$$

From (3.7) we obtain that the conditional expected value of $\hat{t}_y^{(s_b)} = \sum_{s_b} w_k y_k$ is

$$E\left(\hat{t}_y^{(s_b)} \middle| N_{sd}\right) = E\left[\frac{n_b}{N_{withdeads}} \sum_{U_{withdeads}} w_k y_k \middle| N_{sd}\right]$$

$$= \frac{n_b}{N_{withdeads}} \frac{N_l - N_{nodeads}}{N_l} \sum_{U_{orig}} w_k y_k .$$

The second equation above is due to the fact that given N_{sd}, all N_l live units in U_{orig} are equally likely to be in $U_{withdeads}$, which has $N_l - N_{nodeads}$ live units. Therefore, the conditional bias of $\hat{t}_y^{(s_b)}$ is

$$B\left(\hat{t}_y^{(s_b)}, t_y | N_{sd}\right) = \sum_{U_{orig}} \left(\frac{w_k n_b}{N_{withdeads}} \frac{N_l - N_{nodeads}}{N_l} - 1 \right) y_k$$

$$= \sum_{U_{current}} \left(\frac{w_k n_b}{N_{withdeads}} \frac{N_l - N_{nodeads}}{N_l} - 1 \right) y_k \,. \qquad (3.8)$$

For the naive expansion estimator $\hat{\tau}_y^{(s_b)}$ with weights $w_k = N_{current}/n_b$ the bias is

$$B\left(\hat{t}_y^{(s_b)}, t_y | N_{sd}\right) = Bt_y \,, \qquad (3.9)$$

where

$$B = \frac{N_{current}}{N_{withdeads}} \frac{N_l - N_{nodeads}}{N_l} - 1 = \frac{N_{current}\left(N_l - N_{nodeads}\right) - N_l\left(N_{current} - N_{nodeads}\right)}{N_{withdeads} N_l}$$

$$= -\frac{N_d N_{nodeads}}{N_l N_{withdeads}} = -\frac{N_d\left(N_p - N_{sd}\right)}{N_l\left(N_{orig} - N_p\right)} \,.$$

The bias is always non-positive since $B \le 0$. It is easy to see that B is an increasing function of N_{sd} since $N_d = N_{totaldeads} - N_{sd} \ge 0$, where $N_{totaldeads}$ is the fixed number of all dead units in U_{orig}. It is also readily seen that the maximum of B is attained when U_{sd} encompasses all dead units in U_{orig}, that is, when $N_{sd} = N_{totaldeads}$.

Combining (3.9) with (3.3) we obtain the overall bias of $\hat{\tau}_y = \frac{N_{current}}{n} \Sigma_{s_{current}} y_k$ to be

$$E\left(\frac{N_{current}}{n} \Sigma_{s_{current}} y_k - t_y \right) = E\left(\frac{N_{current}}{n} \left(\Sigma_{s_a} y_k + \Sigma_{s_b} y_k \right) \right) - t_y$$

$$= \frac{n_a}{n} E\left(\frac{N_{current}}{n_a} \Sigma_{s_a} y_k \right) + \frac{n_b}{n} E\left(\frac{N_{current}}{n_b} \Sigma_{s_b} y_k \right) - t_y$$

and hence

$$B\left(\hat{\tau}_y, t_y | N_{sd}\right) = E(\hat{\tau}_y \mid N_{sd}) - t_y = \frac{N_d}{N_l} \left(\frac{n_a}{n} - \frac{n_b}{n} \frac{N_{nodeads}}{N_{withdeads}} \right) t_y = \tilde{c} t_y \,. \qquad (3.10)$$

The bias in the expansion estimator is really down to not knowing the correct population size. In (3.3) the bias stems from multiplying the sample average over live units with $N_{current}$ rather than the unknown N_l. The bias from the sample parts s_a and s_b will in

absolute terms be less than (3.3) and (3.9), respectively, if some of the dead units in the samples from U_{orig} have not been identified as dead and therefore have not been weeded out. This would happen, for example, if the status of nonresponding units is difficult to determine.

An unconditional analysis in the presence of feed back can be obtained directly by taking expectation of (3.10) with respect to N_{sd}. Thus, unconditionally, we have

$$E\left(\frac{N_{current}}{n}\sum_{s_{current}} y_k\right) - t_y$$

$$= \left\{\frac{N_{totaldeads} - E(N_{sd})}{N_l}\left(\frac{n_a}{n} - \frac{n_b}{n}\frac{N_p - E(N_{sd})}{N_{withdeads}}\right) - \frac{n_b}{nN_lN_{withdeads}}V(N_{sd})\right\} t_y = ct_y \quad (3.11)$$

where $E(N_{sd}) = N_p\, N_{totaldeads}/N_{orig}$ and $V(N_{sd}) = N_p N_{totaldeads} N_l / N_{orig}^2$.

Lavallée (1996) took an interesting approach to a similar problem with panel survey data. In that paper, the problem of frame update using a panel design with rotation is addressed among other issues. My approach is different from the approach of that paper in that I consider the two conditional probabilities $\Pr[k \in s_a |\, k \text{ alive}, N_{sd}]$ and $\Pr[k \in s_b | N_{sd}]$ separately.

3.3 Three simple strategies

A strategy, which is referred to as Strategy 1 here, is to feed back, delete the set U_{sd} from the frame and accept the feed back bias. However, the size of the bias is seldom known. The estimator for Strategy 1 is $\hat{\tau}_y = \frac{N_{current}}{n}\sum_{s_{current}} y_k$ where $s_{current}$ is a sample taken from $U_{current}$. This estimator can also be viewed as the common estimator of a domain total, where the domain here is the set of live units (Cochran 1977, formula 2.54). To obtain Strategy 2, note that if consistent estimates of N_d and N_l are available these may be plugged into (3.10) or (3.11) and an estimator with favourable properties is obtained:

$$\hat{\tau}'_y = \hat{\tau}_y(1+\hat{c})^{-1}, \quad (3.12)$$

where $\hat{c} = \frac{\hat{N}_d}{\hat{N}_l}\left(\frac{n_a}{n} - \frac{n_b}{n}\frac{N_p - N_{sd}}{N_{orig} - N_p}\right)$ for both the conditional and unconditional cases since the term $n_b V(N_{sd})(nN_l N_{withdeads})^{-1}$ in (3.11) is negligible. The estimates $\hat{N}_d$ and $\hat{N}_l$ of the sizes of the domains U_d and U_l can be obtained from a sample from the original or current survey population with

$$y_k = \begin{cases} 1, \text{ if unit } k \in N_d\ (N_l), \\ 0, \text{ otherwise.} \end{cases}$$

As the following argument shows, we do not expect the bias of (3.12) to be large:

$$E(\tau_y') = E\left[\hat{\tau}_y(1+\hat{c})^{-1}\right] \approx E(\hat{\tau}_y)(1+c)^{-1} = t_y(1+c)(1+c)^{-1} = t_y .$$

Another strategy, here denoted by Strategy 3, is to feed back the information that certain units are dead, but to retain them on the frame and allow them to be sampled. The resulting estimator is unbiased, but the disadvantage of this strategy is that the precision will suffer as part of the sample is lost on ineligible units. The estimator of Strategy 3 is $\hat{\tau}_y'' = \frac{N_{orig}}{n}\sum_{s_{orig}} y_k$. It is shown in Appendix 2 that Strategies 2 and 3 are the same if $n_a/n = 0$, $N_{current} = \hat{N}_d + \hat{N}_l$ and if $\hat{N}_l$ is estimated as N_{orig} times the proportion live units found in the samples covering U_p, that is, $\hat{N}_l = N_{orig}\frac{N_p - N_{sd}}{N_p}$.

3.4 A simulation study

A simulation study may shed some light on which of the Strategies 1-3 is to be preferred. As was mentioned in Chapter 1, in business surveys estimates for subpopulations (industries) are often more interesting than the whole population. To simulate a subpopulation, a frame consisting of 1000 units was created to form the original survey population. A gamma distributed value, Y1, was associated with each unit. I used the same gamma distribution as the one that generated Population 12 in Lee, Rancourt, and Särndal (1994, p. 236). The coefficient of variation (population standard deviation divided by the mean) was 0.57. Another study variable, Y2, was created by performing independent

Bernoulli trials, one for each population unit, which obtained value 1 with probability equal to 0.5 and value 0 otherwise. Unlike in Lee et al. (1994), some of the units were dead. Each unit was independently of other units classified as dead with a probability P_{dead}. All dead units were assigned zero values for both Y1 and Y2. A set of Y1 and Y2 were simulated for each of four values of P_{dead}: 0.03, 0.05, 0.2, and 0.5. These sets contained 29, 54, 201 and 494 dead units, respectively. Having $P_{dead} = 0.50$ is not unrealistic; there are situations where there may be 50% ineligible units or more.

A PRN was attached to each unit and the units were laid out along a PRN line. The first sample, s_1, was drawn by identifying the 500 units with the smallest PRNs. All dead units in s_1 were flagged as 'dead by sample survey sources'. Hence, U_p covered approximately the first half of the PRN line. The frame with the units flagged as dead by sample survey sources excluded made up the current survey population. The estimates of N_d and N_l used in Strategy 2 were based on s_1. A second sample, denoted by $s_{2current}$, was drawn by taking 100 units to the right of a starting point, *start* 2, disregarding units dead by sample survey sources. Another sample of 100 units was selected from *start* 2, but units dead by sample survey sources were this time allowed to be included in this sample. Hence, this sample was drawn from U_{orig}, and we denote it by s_{2orig}. Figure 3.2 shows the PRN intervals and the study variable Y1.

The procedure described in the preceding paragraph was repeated 1000 times. That is, for each of the values of P_{dead} mentioned above and for each of three starting points of s_2, to be defined, 1000 sets of PRNs were generated and attached to the units. The frame was reordered for each new set of PRNs, and three samples were drawn for each reordering (s_1, $s_{2current}$, and s_{2orig}). Two values of *start* 2, 0.0 and 0.7, were chosen so as to make the proportion of $s_{2current}$ that fell in U_{nodead} 100% and 0%, respectively. That is, n_a/n was set to 100% and 0%. Further, to make n_a/n on average 50% under each of the chosen P_{dead}, appropriate values of *start* 2 were derived. They are 0.448, 0.447, 0.438, and 0.4 for the P_{dead} values 0.03, 0.05, 0.2, and 0.5, respectively.

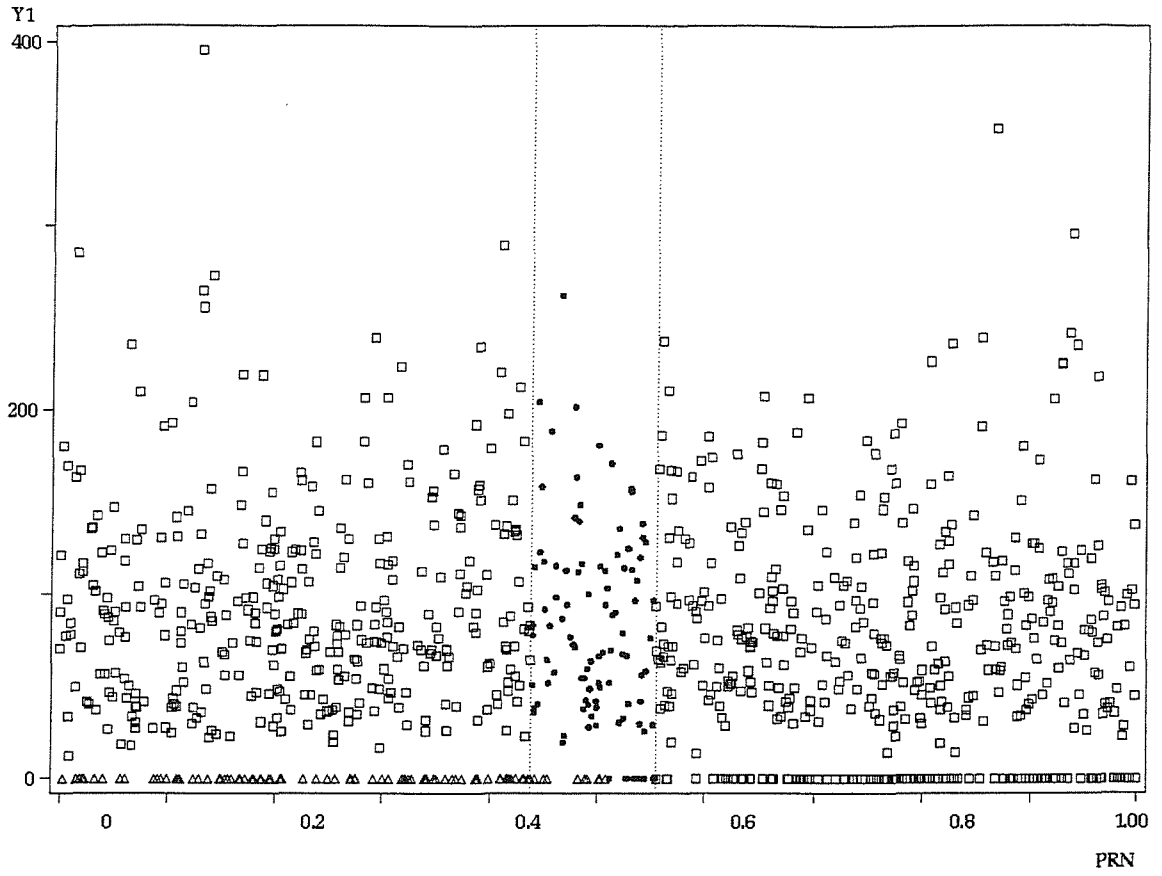


Figure 3.2. A plot of one of the simulated populations, the study variable Y1 against the PRNs, with P_{dead} = 0.20. The dots are units included in $s_{2current}$ (the sample from the current survey population); the triangles are units that are dead by sample survey sources and squares represent units belonging to the current survey population but are not included in the sample from this population. The PRN interval for s_1 (the 500 units in the first sample from the original survey population) is (0, 0.51) and the one for $s_{2current}$ is (0.44, 0.55).

In summary, the population and samples sizes, the study variables Y1 and Y2, and which of the units that were dead were held fixed in this study. For twelve combinations of P_{dead} and n_d/n, the reordering of the units on the PRN line through the simulation of new PRNs made the following factors vary:

- which of the units that were included in s_1, $s_{2current}$, and s_{2orig};
- how many and which of the dead units that were dead by sample survey sources;
- which of the units that belonged to $U_{nodeads}$ and $U_{withdeads}$.

Thus the quantities N_{sd}, N_d and $N_{current}$ vary in the simulations. It seems practical to let them do so rather than to control them in an experiment with more factors than P_{dead} and n_a/n.

Table 3.2. Bias, % of total of Y1. The first entry in each cell is the bias under Strategy 1, the second is the bias under Strategy 2.

	Average of n_a/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	-1.6	-0.1	0.4	0.4	1.5	0.0
0.05	-2.8	0.0	0.4	0.4	2.9	0.0
0.20	-10.2	-0.2	1.5	0.4	12.7	0.1
0.50	-24.6	0.2	12.5	0.3	49.0	0.2

Table 3.2 shows the empirical relative bias of Strategies 1 and 2, computed as the straight average of the 1000 differences between the estimate and the parameter in terms of the percentage of the total obtained in the simulation. Strategy 3 is unbiased and is therefore not included in Table 3.2. The bias of Strategy 3 that nevertheless appeared in the simulations reflects the simulation error; it was at most 0.5%. As seen in Table 3.2, Strategy 2 is virtually unbiased as well. Note that the simulated bias under Strategy 1 is what (3.11) predicts (with allowance for simulation error). This bias is appreciable in nearly all cases and if the proportion of dead (or ineligible) units is high the bias can be very severe indeed. Figure 3.3 shows the conditional bias given N_{sd} for $P_{dead} = 0.50$ and $n_a/n = 0\%$. Note that the bias given by (3.9) is locally well described by the regression line in the figure defined by the OLS fit of the conditional bias on N_{sd}. For example, if $N_{sd} = 220$, then both N_d/N_l and $(N_p - N_{sd})/(N_{orig} - N_p)$ equal 0.56 and $B = -0.31$. With $N_{sd} = 250$, $B = -0.25$.

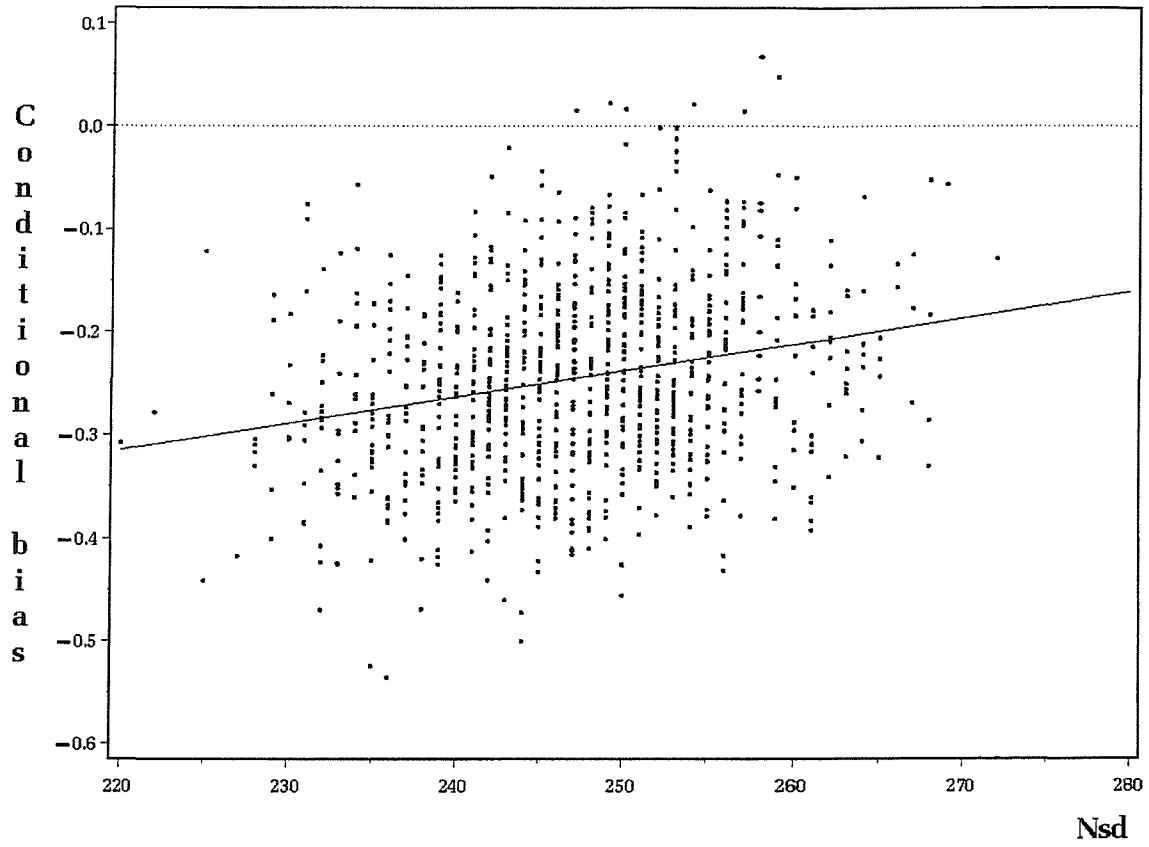


Figure 3.3. The simulated conditional bias plotted against the number of units dead by sample survey sources, N_{sd}, for $P_{dead} = 0.50$ and $n_a/n = 0\%$. An OLS regression line shows the general trend of the conditional bias as a function of N_{sd}.

To assess the bias it helps to look at the coverage probabilities. Table 3.3 shows the empirical coverage probabilities, based on symmetric 'confidence intervals' with a width of two times the simulated empirical variance of each side of the point estimate. While Strategy 2 gives in all cells coverage probabilities close to the targeted 95%, Strategy 1 achieves that in general only for the population with 3% dead units. The coverage probability under Strategy 1 tends also to be acceptable for populations with a larger proportion of dead units, if half of the sample is taken from the part of the PRN line where dead units have been weeded out, and the other half from the part of the PRN line where the original proportion of dead units has been retained, as the negative bias from the first half of the sample tends to cancel out the positive bias from the second half.

Table 3.3. The coverage probability in percentage for estimating the total of Y1. The first entry in each cell is the under Strategy 1, the second is the coverage probability under Strategy 2.

	Average of n_d/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	94.6	94.3	94.6	94.8	94.3	95.1
0.05	93.3	95.2	94.4	93.9	90.8	95.0
0.20	65.9	94.5	93.8	94.8	46.1	94.6
0.50	21.2	95.1	78.4	94.7	0.0	94.8

Tables 3.4 and 3.5 show the variance of the estimated totals of Y1 and Y2, respectively, under Strategies 2 and 3 relative to that of Strategy 1, which in all cases gives a smaller variance than Strategy 3. Hence, considering the extra complexity of Strategy 2, the feed back strategy may be preferable for populations with a small proportion of ineligible units, say 3% or less. If this proportion is larger than, say, 5%, the bias of Strategy 1 may cause poor coverage probabilities and misleading estimates. The variance of Strategy 2 is no worse than that of Strategy 3; in most cases Strategy 2 is superior. The non-monotone variance ratios in the bottom row of Table 3.4 is due to the estimation of N_d and N_l combined with the specific details of the simulation.

3.5 Discussion

I have derived conditional and unconditional expressions for the feed back bias when the total is estimated with the expansion estimator and I have shown that the feed back bias can be large. With as little as 5% ineligible units on the frame, feeding back information of these from sample surveys can result in about 2-3% bias. However, a small-scale simulation study indicates that if the proportion of ineligible units is 3% or less, the feed back strategy does not seem to create problems in terms of bias and variance. Having said that, the bias-variance trade-off is rather more complicated in business surveys than in social surveys for reasons discussed in Chapter 1.

Table 3.4. Variance ratio of the estimator of the total of Y1. The first entry in each cell is the variance under Strategy 2 relative to that of Strategy 1, the second is the variance under Strategy 3 relative to Strategy 1.

	Average of n_a/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	1.04	1.04	1.00	1.06	0.98	1.08
0.05	1.08	1.08	0.98	1.14	0.95	1.15
0.20	1.28	1.28	0.85	1.27	0.83	1.46
0.50	1.85	1.85	0.52	1.34	0.58	2.24

Table 3.5. Variance ratio of the estimator of the total of Y2. The first entry in each cell is the variance under Strategy 2 relative to that of Strategy 1, the second is the variance under Strategy 3 relative to Strategy 1.

	Average of n_a/n					
P_{dead}	**0%**		**50%**		**100%**	
0.03	1.03	1.03	1.00	1.03	0.97	1.03
0.05	1.06	1.06	0.99	1.04	0.95	1.06
0.20	1.25	1.25	0.92	1.15	0.80	1.19
0.50	1.80	1.81	0.65	1.40	0.50	1.36

I have also derived a virtually unbiased estimator. The simulation study shows that this estimator compares favourably in terms of variance with the alternative strategy of retaining ineligible units on the frame and letting them be included in further samples. This estimator relies on the availability of consistent estimates of the number of eligible and ineligible units in the population.

In order to facilitate the theoretical development, I have made simplifying assumptions. The most important of these is the assumption that *all* dead units, or more generally all ineligible units, have been found in earlier sample surveys and have been fed back to the frame. We have envisaged a frame with one 'white' area, where all ineligibles have been flagged as such, and one 'black' area, where no ineligibles have been touched. In practice, this is not likely to happen. If the frame is moderately large and used for many continuing surveys, some of which may feed back to varying intensity, the frame will turn 'grey' rather than 'black and white'. Clearly, the feed back bias will then in general be smaller than in the 'black and white' situation. It has not, however, been in the scope of this thesis to quantify the bias for a 'realistically grey' frame. In this sense, what has been examined here is a worst case scenario.

Chapter 4

Does the Model Matter for GREG Estimation?

Business surveys often pose a variety of data problems that can be very difficult to resolve simultaneously. For example, the study variable(s) may be highly skewed, there may be a large proportion of zero responses, some negative values and there may be several auxiliary variables that can be used to improve estimation but these may include some extreme values.

Till recently, simple survey estimation techniques such as classical ratio or regression estimation have been sufficient for the business surveys carried out by many National Statistical Institutes, such as the ONS. However, the wider use of more sophisticated estimation methods, the growing use of a greater amount of auxiliary information in estimation, and the pressure to substantially reduce sample sizes or to produce accurate estimates for small domains has increased the importance of recognising and dealing with the data issues mentioned above. Chapter 4 illustrates methods for addressing some of these issues in a real business survey, with an emphasis on the importance of model choice in model-assisted GREG estimation. I will illustrate these methodologies using data from the quarterly survey of capital expenditure (the CAPEX survey) carried out by the ONS.

Section 4.1 reviews some GREG theory. Section 4.2 introduces the CAPEX Survey. In Section 4.3 a relationship is shown between the g-weight of a sample unit and its DFBETA, a well-known measure of the influence of a sample unit on the slope of a regression line. In Section 4.4 the result of applying different GREG estimators to the CAPEX survey data is reported. This leads to some rather surprising outcomes, and in Section 4.5 we explore these data to reveal particular features that underpin these outcomes. Section 4.6 offers an explanation of the behaviour of the GREG estimators in the light of this analysis. Section 4.7 reports on an attempt to get around these problems. In Section 4.8 the findings are discussed.

4.1 A summary of the theory for the generalised regression estimator

The aim of many business surveys is to estimate totals and differences between totals. In this section I review the literature on design-based linear methods of estimating the total $t_y = \sum_U y_k$ of a study variable $\mathbf{y}' = (y_1, y_2, \ldots y_N)$ on a population U where the units have the labels $\{1,2,\ldots N\}$. The issue is how to use auxiliary information effectively. My main focus in Section 4.1 is on results that will be referred to in the remainder of Chapter 4 and also in Chapter 5. A far-reaching exposition of sample survey theory and philosophy can be found in Thompson (1997). Valliant et al. (2000) contrast design-based estimation with model-based estimation. Recent overviews of sample survey theory, in particular regression estimation, include Rao (1997) and Fuller (2002). I concentrate on single-study variable estimation. Multivariate study variable estimation issues are discussed, among others, by Bethlehem and Keller (1987) and Chambers (1996).

I assume that there is a known auxiliary vector $\mathbf{x}_k' = (x_{1k} \quad x_{2k} \quad \ldots \quad x_{pk})$ for each element in U. This assumption is unnecessarily strong for most estimators I study, but more often than not, $\mathbf{x}_k$ is indeed available for all units on the frame in actual business survey systems. A sample s of size n is taken and $(\mathbf{x}_k, y_k)$ is assumed to be observed for all units k in the sample. In Chapters 4 and 5, nonsampling errors, that is nonresponse, measurement and coverage errors are disregarded.

Before defining and showing why the GREG is reasonable, I start with the more elementary Horvitz-Thompson estimator.

4.1.1 The Horvitz-Thompson estimator

Consider estimation of the population total t_y. It is easy to see that there is for a general sampling design and a general configuration of $\mathbf{y}' = (y_1, y_2, \ldots y_N)$ only one design-unbiased estimator of the form

$L = \sum_U \omega_k I_k y_k$

where ω_k is a constant and I_k is an indicator function that takes the value 1 if unit k is in the sample and 0 otherwise. The unbiased estimator requires the weights ω_k to equal the inverse of the first-order inclusion probabilities $\pi_k = \Pr(I_k = 1)$. As was mentioned in Chapter 1, this estimator is usually called the Horvitz-Thompson estimator (HT-estimator; Horvitz and Thompson, 1952). Following Särndal et al. (1992), I will denote the HT estimator by $\hat{t}_{y\pi}$ and write

$$\hat{t}_{y\pi} = \sum_s w_k y_k \qquad (4.1)$$

with $w_k = \pi_k^{-1}$.

The variance of the HT-estimator is $V(\hat{t}_{y\pi}) = \sum_U \Delta_{kl} y_k y_l / \pi_k \pi_l$, where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ with $\pi_{kl} = \Pr(I_k = I_l = 1)$ being the second order inclusion probability. The most common variance estimator is $\hat{V}(\hat{t}_{y\pi}) = \sum_s \breve{\Delta}_{kl} y_k y_l / \pi_k \pi_l$, where $\breve{\Delta}_{kl} = (\pi_{kl} - \pi_k \pi_l)/\pi_{kl}$. For a fixed sample size and measurable design (i.e. all $\pi_{kl} > 0$), another unbiased variance estimator is the *Yates-Grundy-Sen variance estimator* (Yates and Grundy 1953, and Sen 1953):

$$\hat{V}(\hat{t}_{y\pi}) = -\frac{1}{2} \sum\sum_s \breve{\Delta}_{kl} (y_k/\pi_k - y_l/\pi_l)^2 .$$

The HT-estimator does not use any auxiliary information other than through the choice of inclusion probabilities, which are controlled by the survey statistician.

The design-unbiased property is regarded at most national statistical institutes as highly desirable. However, the explicit use of auxiliary information in the estimator is also regarded as extremely important. In business statistics, the perceived main reason for this is that auxiliary information will usually increase precision considerably. It may also reduce nonsampling errors. In fact, in household surveys this may be the most important benefit of using auxiliary information, traditionally through post-stratification (Jayasuriya and Valliant, 1996). Bethlehem (1988), Lundström and Särndal (1999, 2001) and Fuller (2002) discuss the use of generalised regression estimation to reduce nonresponse bias. Skinner (1999) discusses calibration as a means of reducing both nonresponse bias and effects from measurement error.

The *linear form* of any estimator, such as (4.1), is also considered very important since it helps computation. Another reason for wanting a linear estimator is the nice interpretation it offers: the form of the estimator reflects the view that a sampled element can be seen as representing $\omega_k - 1$ nonsampled units in addition to itself and thus has a strong intuitive appeal. Brewer (1999, p. 36) calls this the *Representative Principle* and points out that good design-based inference rests on the compliance to this principle. The most common compromise between these conflicting goals is to only require approximate design-unbiasedness and hence allow for a wider class of estimators. I shall continue the discussion of the benefits of the linear form of an estimator in Chapter 5.

4.1.2 The calibration estimator

The *calibration estimator* can be written as

$$\hat{t}_{ycal} = \sum_s w'_{ks} y_k \,, \qquad (4.2)$$

where the weights w'_{ks} are derived as is described in the following paragraphs. The calibration estimator constitutes a class of linear estimators that satisfy the following constraint for any realisation of I_k, $k = 1, 2, \ldots, N$:

$$\sum_U w'_{ks} I_k \mathbf{x}_k = \mathbf{t}_x, \qquad (4.3)$$

where $\mathbf{t}_x = \sum_U \mathbf{x}_k$ is a vector of known auxiliary variable totals (often referred to as benchmarks or control totals) and w'_{ks} are appropriate weights that may be sample dependent. With the terminology of Deville and Särndal (1992), an estimator that satisfies (4.3) is *calibrated* on the known population totals $\mathbf{t}_x$. This class of estimators includes all of the most widely used business survey estimators of the total that explicitly use auxiliary information. The totals $\mathbf{t}_x$ are often published in tables or elsewhere. Often the publications contain also estimates of $\mathbf{t}_x$ either explicitly or available as, for example, a sum of domain estimates. Consistency between these is a concern for both statistics users and producers. Hence, the property of an estimate being calibrated on some population totals is highly desirable in official statistics. Also, it is reasonable to regard a discrepancy between known totals $\mathbf{t}_x$ and corresponding estimates as an indication of survey error in the study variables.

The idea of Särndal's and Deville's calibration estimator technique is to find a linear estimator that satisfies (4.3) *exactly* and makes the weights w'_{ks} 'as close as possible' to the weights $w_k = \pi_k^{-1}$. One way of making the latter requirement operational is to minimise the design-expectation of the sample sum of the distances between w'_{ks} and w_k, where the distance is measured by some reasonable distance function $G_k(w_k, w'_{ks})$. Thus, the w'_{ks} that minimises $E_p\left[\sum_s G_k(w_k, w'_{ks})\right]$ over all samples s under (4.3) is sought. This is, in principle, an easy exercise of the Lagrange multiplier technique; however, if the solution exists and is unique depends on the exact form of the distance function. The resulting estimator is (4.2).

The simple form of (4.2) is attractive from a practical point of view. Of a particular note is that the same set of weights can be used for different study variables. However, the calculation of the weights w'_{ks} may pose numerical problems for some distance functions for which there is no explicit solution. There are many possible distance functions, the only restriction being some general mathematical properties. See Deville and Särndal (1992), Bardsley and Chambers (1984), Singh and Mohl (1996), Thompson (1997) and references in the latter for a discussion of different distance functions. It is not clear why any of them should be preferred to the others. Simulation studies by Singh and Mohl (1996) do not indicate large differences. Estevao and Särndal (2000) argue that the distance functions studied by Deville and Särndal (1992) and Singh and Mohl (1996) are so similar that the estimators should be similar at least for large samples. However, from a theoretical point of view there is a particularly interesting distance function that leads to the GREG estimator.

4.1.3 The GREG estimator

Let

$$G_k(w_k, w'_{ks}) = (w'_{ks} - w_k)^2 / w_k q_k , \qquad (4.4)$$

where q_k is some additional set of weights yet to be specified. With this distance function the general form of the calibration estimator becomes

$$\hat{t}_{yreg} = \sum_s g_{ks} w_k y_k . \qquad (4.5)$$

where $w'_{ks} = g_{ks} w_k$. As was said in Chapter 1, the sample-dependent g_{ks}, $k = 1, 2, \ldots n$, are known as the 'g-weights' and are defined as

$$g_{ks} = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)' \left(\sum_s w_j q_j \mathbf{x}_j \mathbf{x}'_j\right)^{-1} (q_k \mathbf{x}_k). \qquad (4.6)$$

Thus the 'total' weight $w'_{ks} = g_{ks} w_k$ in (4.5) is partitioned into a purely design-dependent weight w_k and a weight g_{ks} that forces the estimator to be calibrated on $\mathbf{t}_x$. Note that we do not need to know values of auxiliary variable for individual non-sample units, only $\mathbf{t}_x$, to compute (4.6) and (4.5).

The estimator (4.5) is widely known as the generalised regression (GREG) estimator. It is a special case of the calibration estimator. From the calibration property it follows that $\sum_s g_{ks} w_k \mathbf{x}_k = \mathbf{t}_x$. Note that (4.5) is reminiscent of a HT-estimator of $\sum_U g_{ks} y_k$, although this is not a proper population parameter since the g_{ks} are sample dependent. Nevertheless, this shows that if the g-weights for a particular sample are far away from 1 then we might be estimating something that is very different from t_y.

Another derivation of the GREG estimator is obtained by starting from the difference estimator,

$$\hat{t}_{ydif} = \sum_U y_k^0 + \sum_s w_k \left(y_k - y_k^0\right), \qquad (4.7)$$

where y_k^0 is a non-random proxy for y_k. It is readily seen that this estimator is design-unbiased no matter the y_k^0. Since $\sum_s w_k \left(y_k - y_k^0\right)$ is the HT-estimator for $\sum_U \left(y_k - y_k^0\right)$, the variance of (4.7) is $V\left(\hat{t}_{ydif}\right) = \sum_U \Delta_{kl} \left(y_k - y_k^0\right)\left(y_l - y_l^0\right)/\pi_k \pi_l$. Thus, the better the proxy, the smaller the variance, and hence a good estimator in terms of variance will be obtained if the proxies are computed via a good prediction rule. Standard linear regression may provide the prediction rule. However, in the model-assisted GREG a version of standard linear regression that uses inclusion probabilities is preferred. To motivate the GREG on the basis of (4.7), a superpopulation model M is assumed:

Model M

1. $y_1, y_2, ..., y_N$ are realisations of the independent random variables $Y_1, Y_2, ..., Y_N$
2. $E_M(Y_k) = \mathbf{x}_k'\boldsymbol{\beta}$
3. $V_M(Y_k) = \sigma_k^2$, $k = 1, 2, \ldots, N$
4. $Cov_M(Y_k, Y_l) = 0$ for $k \neq l$,

where the moments are taken over the model and $\boldsymbol{\beta}$ and σ_k^2, $k = 1, 2, \ldots, N$, are unknown parameters.

Särndal et al. (1992, pp. 226-227 and pp. 238-239) discuss the role of the model in model-assisted theory. They focus on design-bias and design-variance and say that the GREG is asymptotically design-unbiased no matter what model has been chosen but in terms of design-variance there may be a big gain over the HT-estimator in choosing the best model. They go on to say on p. 239 that 'if the population is not well described by the model, the improvement on the π estimator [i.e. the gain of using a GREG rather than the HT estimator] may be modest, but the regression estimator still guarantees approximate unbiasedness'. As we shall see, this statement is rather optimistic.

The quantity q_k is in the context of the general calibration estimator seen as an extra, unspecified weight. In the context of GREG estimation, $\sigma_k^2 = q_k$ and σ_k^2 is seen as a model parameter. The introduction of an explicit model into the reasoning is a fundamental difference between the 'calibration view' and this 'model assisted' view.

A finite population parameter vector $\mathbf{B}$ is envisaged:

$$\mathbf{B} = \left(\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}'\right)^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{Y},$$

using the same notation as in standard regression theory. The parameter $\mathbf{B}$, referred to as the 'census fit' by Särndal (1982), can be viewed a hypothetical weighted least squares estimate of the superpopulation parameter vector $\boldsymbol{\beta}$. The census fit $\mathbf{B}$, in turn, is estimated using HT-estimators for $\left(\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}'\right)^{-1}$ and $\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, respectively. Thus the estimator of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = \left(\mathbf{X}_s\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s'\right)^{-1}\mathbf{X}_s\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{Y}_s,$$

where $\boldsymbol{\Pi}_s^{-1}$ is the diagonal matrix of the first-order inclusion probabilities for the units in s, and subscript s indicates that the quantities are based on the sample only. The

estimator $\hat{\mathbf{B}}$ is approximately design-unbiased for large samples (Särndal 1980). Using $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}$ in (4.7) the following well-known form of the GREG is obtained.

$$\hat{t}_{yreg} = \sum_U \hat{y}_k + \sum_s w_k (y_k - \hat{y}_k). \qquad (4.8)$$

It is not necessary to have HT-estimators in the GREG, as is noted by Särndal et al. (1992, p. 230), although this is in practice the by far most common alternative.

The estimator (4.8) is algebraically equivalent to (4.5). Furthermore, if σ_k^2 can be expressed as linear combination of the components of the auxiliary vector for any unit k, i.e. if the following condition holds for some constant p-vector $\lambda \in \Re^p$,

$$\sigma_k^2 = \boldsymbol{\lambda}'\mathbf{x}_k, \qquad (4.9)$$

for all for all $k \in U$, then $\hat{t}_{yreg} = \sum_U \hat{y}_k$ (proof in Särndal et al, 1992, p. 231). The simplified form $\hat{t}_{yreg} = \sum_U \hat{y}_k$ may be called the *projective form* of the GREG. Assume for a moment that the model contains only one auxiliary variable and no intercept. If $\sigma_k^2 = \sigma^2 > 0$ then (4.9) is not satisfied, since $\sigma^2 \neq \lambda x_k$ for all real-valued λ and for a general configuration of the auxiliary variable x_k, $k = 1, 2, \ldots, N$. If, on the other hand, $\sigma_k^2 = \sigma^2 > 0$ and $\mathbf{x}_k' = (1 \quad x_k)$, we can define $\boldsymbol{\lambda}' = (\sigma^2 \quad 0)$ and we have $\sigma^2 = \boldsymbol{\lambda}'\mathbf{x}_k$.

Even if the GREG is confined to cases where the projective form is valid, it comprises a large class of estimators. As an example, consider a special case of model M. Let the population be partitioned into G groups, $U = \bigcup_{g=1}^{G} U_g$ and let the model be an ANOVA model with a constant expectation and variance of Y_k within the groups (poststrata). Assume that unit k belongs to group g and let $\mathbf{x}_k$ be a vector with 1 in the *gth* position and zeroes elsewhere. Then the model is

$E_M(Y_k) = \beta_g$,

$V_M(Y_k) = \sigma_g^2$, for $k \in U_g$. Condition (4.9) is satisfied with

$\boldsymbol{\lambda} = (\sigma_1^2, \ldots, \sigma_g^2, \ldots, \sigma_G^2)'$. Here $\mathbf{B}$ is constant within groups with the estimator

$\hat{\mathbf{B}}_g = \bar{y}_g = \sum_{s_g} w_k y_k / \sum_{s_g} w_k$ for the sample part $s_g = U_g \cap s$, so

$\hat{t}_{ypost} = \sum_U \hat{y}_k = \sum_{g=1}^{G} \sum_{U_g} \tilde{y}_g = \sum_{g=1}^{G} N_g \tilde{y}_g$, where N_g is the number of units in U_g. The estimator $\hat{t}_{yreg}$ is now called $\hat{t}_{ypost}$ since this is the widely used *poststratified estimator* (e.g. Cochran 1977). Thus the poststratified estimator is a special case of the GREG.

Other special cases of the GREG are discussed by Särndal (1982) and Särndal et al. (1989, 1992).

Note that the g-weights depend on both model and design. Hence the role of the g-weights is to formally bring the survey statistician's beliefs, expressed in a model, into the estimator. For an amusing illustration consider Basu's elephants. In a blatant breach with what Brewer later called the Representative Principle, Basu (1971) gives an example of a design with the worst possible connection between the inverse of the inclusion probabilities and the number of units a sampled unit can be thought of representing. The elephant Sambo (unit i) is known to have a study variable value, y_i, close to the average of the population. Sambo is selected with a design close to a judgement sample (or with what has later been called a balanced sample) and the very reasonable estimator Ny_i of t_y is rejected in favour of the HT-estimator (4.1). In an attempt to impose some inclusion probabilities on the design that in effect dictates that unit i should be selected, this units is given inclusion probability 99/100. Since the inclusion probabilities in Basu's example are silly, the HT weight $w_i = \pi_i^{-1} \approx 1$ attached to the selected unit i makes it represent far from itself plus $N-1$ nonselected units. The GREG (4.5), however, recovers the Representative Principle if $\mathbf{x}_k$ is taken as a scalar that always takes the value 1, and if $E_M(Y_k) = \beta$, and $V_M(Y_k) = \sigma^2$. Then the g-weight is $N\pi_i^{-1}$ and the GREG is Ny_i. Here the 'model-adjustment' that is implicit in the g-weights is drastic, since they make the inclusion probabilities vanish altogether. Another example of the role of the g-weights is the ratio estimator where the g-weights are $t_x / \hat{t}_{x\pi}$, which is a straightforward adjustment for sample imbalance with respect to the auxiliary variable. In both these examples the g-weights are constant over k, which is not in general true.

It is convenient at this stage to recall some asymptotic properties of the GREG. Let subscript t index a sequence of populations and associated parameters and estimators. In all frameworks deployed to explore asymptotic properties of estimators in finite population sampling, the population U is thought of as being embedded in a sequence of populations, U_t, $t = 1, 2, 3, \ldots$. The size N_t grows indefinitely as $t \to \infty$, that is $N_t \to \infty$, while the structure of the population remains the same. In the asymptotic framework of Isaki and Fuller (1982), $n_t \to \infty$ without restriction on the relative pace of the increase of N_t and n_t. A sample s_t is taken from U_t using the sampling design $p_t(s_t)$. The sequence of estimators $\hat{t}_{ty}$ of the U_t-total t_{ty} is said to be *asymptotically design-unbiased* if $\lim_{t\to\infty}\left(E(\hat{t}_{ty}) - t_{ty}\right) = 0$, where the expectation refers to the $p_t(s_t)$ distribution. Similarly, the sequence of estimators $\hat{t}_{ty}$ is said to be *design-consistent* if

$$\lim_{t\to\infty}\left(\Pr\left[\ |\hat{t}_{ty} - t_{ty}| > \varepsilon\ \right]\right) = 0, \quad \forall \varepsilon > 0.$$

What is meant when one says that 'an' estimator has some asymptotic property may not be entirely clear, but for the purposes of the thesis we do not need to go into great detail here.

Under certain regularity conditions, the GREG is design-consistent and asymptotically design-unbiased (Isaki and Fuller, 1982, Robinson and Särndal, 1983, Wright 1983). The former property implies the latter under mild conditions. There are slight differences in the definitions of the asymptotic framework, the regularity conditions and the asymptotic properties, but these differences 'do not seem to be consequential' (Särndal and Wright 1984, p. 148).

Although being asymptotically design-unbiased, the GREG is certainly not (exactly) unbiased. The bias of the GREG is (Särndal 1980)

$$-\sum_{j=1}^{J} Cov\left(\hat{t}_{y\pi}\left(\mathbf{1}_J'\hat{\mathbf{t}}_{x\pi}\right)^{-1}, \hat{t}_{xj\pi}\right),$$

where $\mathbf{1}_J$ is a J-vector of ones and $\hat{t}_{xj\pi}$ is the jth component of $\hat{\mathbf{t}}_{x\pi}$. This expression shows that high-leverage points may cause bias.

Next, I consider the Godambe-Joshi lower bound (Godambe and Joshi 1965). To define it we need the concept of *anticipated variance* (Isaki and Fuller 1982), which is

the variance of $\hat{T}_{Yt} - T_{Yt}$ under the joint distribution of a superpopulation model ξ and the sampling design $p(s)$, i.e. $E_\xi E_p\left(\hat{T}_{ty} - T_{ty}\right)^2 - \left[E_\xi E_p\left(\hat{T}_{ty} - T_{ty}\right)\right]^2$. The Godambe-Joshi lower bound of the anticipated variance for any estimator that is unbiased under the joint distribution, that is,

$E_\xi E_p\left(\hat{T}_{Yt} - T_{Yt}\right) = 0,$

is

$$\sum_U \left(\frac{1}{\pi_k} - 1\right)\sigma_k^2 .$$

Under model M, the variance of the GREG attains the Godambe-Joshi lower bound asymptotically (e.g. Wright 1983, Särndal et al. 1992, p. 453).

4.1.4 Estimating the variance of the GREG

Let $E_k = y_k - \mathbf{x}_k'\mathbf{B}$ be the 'census fit residuals'. It is straightforward to show (e.g. by modifying the proof of Särndal et al. 1992, p. 231) that under (4.9)

$t_y = \sum_U y_k = \sum_U \mathbf{x}_k'\mathbf{B}$. Using the calibration property

$t_y = \mathbf{t}_x'\mathbf{B} = \sum_s g_{ks} w_k \mathbf{x}_k'\mathbf{B}$ is obtained. Hence, under (4.9), the error of the GREG is

$$\hat{t}_{yreg} - t_y = \sum_s g_{ks} w_k E_k . \qquad (4.10)$$

This result suggests that the g-weights should appear in the variance as well as in the variance estimator. Through Taylor linearisation an approximate expression for the variance is obtained: $AV\left(\hat{t}_{yreg}\right) = \sum\sum_U \Delta_{kl} E_k E_l / \pi_k \pi_l$ (Särndal et al. 1992, Sec. 6.6). Hence a natural variance estimator, akin to the HT-variance estimator, is

$$\hat{V}'\left(\hat{t}_{yreg}\right) = \sum\sum_s \breve{\Delta}_{kl} e_k e_l / \pi_k \pi_l , \qquad (4.11)$$

where $e_k = y_k - \mathbf{x}_k\hat{\mathbf{B}}$ are the 'sample fit residuals'.

Since there are no g-weights in (4.11), Särndal (1982) proposes that the g-weights are put back in the variance estimator. Thus, the suggested variance estimator is

$$\hat{V}\left(\hat{t}_{yreg}\right) = \sum\sum_s \breve{\Delta}_{kl} g_{ks} e_k g_{ls} e_l / \pi_k \pi_l . \qquad (4.12)$$

Särndal et al. (1989) show that if (4.11) is design-consistent then so is (4.12). Both of them are approximately design-unbiased for $AV(\hat{t}_{yreg})$, see Särndal (1982), although $AV(\hat{t}_{yreg})$ tends to be smaller than the exact variance. However, if (4.9) holds, then (4.12) is approximately model-unbiased with respect to M, which is in general not true for (4.11), see Särndal et al. (1989). As a consequence of this, (4.12) specialises to more appealing formulae than does (4.11). Some of these have also been shown superior in simulation studies. Let us now take a look at two examples.

Example 1: The poststratified estimator

Recall that the poststratified estimator is defined as $\hat{t}_{ypost} = \sum_{g=1}^{G} N_g \tilde{y}_g$, where $\tilde{y}_g = \sum_{s_g} w_k y_k / \sum_{s_g} w_k$. Under simple random sampling without replacement, SI, $\hat{t}_{ypost}$ is $\hat{t}_{SI,ypost} = \sum_{g=1}^{G} N_g \bar{y}_g$ with $\bar{y}_g = \sum_{s_g} y_k / n_g$ being the straight mean for the *gth* poststratum. Using the variance estimator (4.11), it is shown in Appendix 3 that the variance estimator of the poststratified estimator
is

$$\hat{V}'_{SI}(\hat{t}_{ypost}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{g=1}^{G}\frac{n_g - 1}{n-1}s_g^2, \qquad (4.13)$$

where $s_g^2 = \dfrac{\sum_{s_g}(y_k - \bar{y}_g)^2}{n_g - 1}$ and n_g is the number of units in s_g. This is not a very natural estimator. One problem with (4.13), stemming from the fact that it is not conditional on the realised values of the random variables $n_1, n_2, \ldots, n_H$, is that the weights of the estimated poststratum variances s_g^2 are proportional to n_g. Assuming that s_g^2 are estimated without error, we see that in (4.13) poststrata that happened to get a smaller than expected sample size will give a *smaller* contribution to the variance than expected. Also the variance estimator that follows from Cochran's (1977, p. 135) variance formula suffers from a similar conditional shortcoming: this one is insensitive to the outcome of the poststratum sample sizes. Holt and Smith (1979) and Särndal et al. (1989) argue that this is counterintuitive. Note that the g-weights $N_g/\hat{N}_g = N_g n/Nn_g$ are constant within poststratum g. With these in (4.12) we obtain the variance estimator

$$\hat{V}_{SI}\left(\hat{t}_{ypost}\right)=\left(1-\frac{n}{N}\right)\sum_{g=1}^{G}N_g^2\,s_g^2/n_g\,, \qquad (4.14)$$

where $s_g^2=\dfrac{\sum_{s_g}\left(y_k-\bar{y}_g\right)^2}{n_g-1}$. Unlike (4.13), the estimator (4.14) has the appealing property of inflating the variance for poststrata with small sample sizes and letting large sample poststrata be more stable. A similar argument can be made about Bernoulli sampling, which is a design with a stochastic overall sample size. See Särndal et al. (1989, Ex. 4.2; 1992, Ch. 7.10.1).

Example 2: The ratio estimator

The variance estimator (4.12) for a ratio estimator for simple random sampling is

$$\hat{V}_{SI}\left(\hat{t}_{yrat}\right)=N^2\left(\frac{t_x}{\hat{t}_{x\pi}}\right)^2\left(\frac{1}{n}-\frac{1}{N}\right)\sum_s\left[\left(y_k-(\bar{y}/\bar{x})x_k\right)\right]^2/(n-1), \qquad (4.15)$$

where $\left(t_x/\hat{t}_{x\pi}\right)^2$ is the (constant) g-weight squared. This variance estimator has in simulation studies been shown to be better than the corresponding estimator without g-weights (Wu and Deng, 1983).

Software has been developed for the computation of (4.12). Fuller (2002) gives a list of relevant computer implementations although his list is already slightly outdated. Fuller also discusses other variance estimation methods such as the jack-knife.

4.1.5 Problems with the GREG estimator

The GREG is very flexible in that it comprises a large number of different estimators, some of which are widely used. There is no limit to what auxiliary variables that can be used apart from some mathematical restrictions such as the non-singularity of the matrix $\mathbf{X}_s\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s'$. The auxiliary variables may be qualitative or quantitative; and they may be associated with units of different level, e.g. company and local unit. Furthermore, since the GREG is derived for a general set of inclusion probabilities it can be specialised to any sampling design.

The GREG estimator has received a lot of attention the last decades, in particular after the publication of the book Särndal et al. (1992), which is now widely regarded as part of the standard literature in survey sampling. One reason for the popularity of the GREG is undoubtedly its flexibility in spite of its simple linear form, which is attractive both from a conceptual and computational point of view. Also, the model-assisted approach provides explanation in two ways: the model explains why some estimators work better than others in a particular situation and the models in general show how various estimators are interrelated. In a classical design-based account (e.g. Cochran, 1997) different estimators tend to be jumbled up together. Model-assisted theory has certain elegance. Chalmers (1982, Ch. 6) argues that elegance does play a major role for the survival chances of a theory.

However, the GREG estimator has some serious downsides, some of which have not yet been fully explored. First, one well known drawback is that the GREG can, and often will, give negative weights. This may lead to poor estimates. The estimate may even be negative for a variable that cannot take negative values. 'In practice, negative weights are rare...', Stukel et al (1996, p. 119) write. This may be true outside the realm of business surveys. However, I will give an example based on real business data where the estimate for a very reasonable model is close to zero due to the adjustment term discussed above in conjunction with (4.8) and (4.9). As noted by Chambers (1996) and as will be discussed further below, negative weights is symptomatic of deeper estimation problems and model misspecification. I also discuss a popular method of restricting the g-weights to, for example, positive values.

Second, the fact that design-weights and the calibrated weights are 'as close as possible' does not necessarily mean that they are similar. In fact, as will be shown later, the supremum of the distance between the two sets of weights is arbitrary large (infinite) in some situations. Hence the Representative Principle will be upset.

Third, while it is true that the g-weights tend to be close to 1 in large samples (Särndal 1982), we will see that they can be very far away from 1 in either direction for moderate size samples and for data that are not 'pathological' in any way. It is often mentioned that the calibrated weights for the raking ratio estimator, which is another

class of calibrated estimators, can be very large, but the same behaviour of the GREG weights has not yet been analysed.

Fourth, as we will see in an example based on real data, the variance can be so large that the point estimate is useless, even if the model is the one that fits the data best – that is, within the GREG class of models.

4.2 The CAPEX Survey

The Office for National Statistics is responsible for the lion's share of UK's official statistics output. Many subannual and other surveys are carried out simultaneously in streamlined, repetitive production processes. The surveys are coordinated through strict stratification rules to facilitate estimation of change and trend and, at the same time reduce the response burden on individual businesses. The data, both the auxiliary and the study variables, have passed through an extensive edit and validation process, in which virtually all businesses that fail a validation check are called back. Apart from this substantial editing effort, performed by a separate division, time constraints make it hard for the analysts and methodologists to inspect all data for potential problems, e.g. outliers. For any proposed method or process, the gain of extra complexity will have to be balanced against the cost in terms of implementation resource, production time and human factors such as the learning curve of new staff. In Chapter 4 I work within this context. In particular, I accept the stratification as 'given'; I also accept the correctness of the data values, and the just mentioned requirements for the production process.

At the time of writing, the CAPEX survey collects data from approximately 16,000 private sector businesses. The main study variables are acquisitions and disposals, and their difference is referred to as net capital expenditure. The results from the survey contribute to the estimates of gross domestic fixed capital formation, one of the expenditure components of the gross domestic product in the national accounts. Estimates from the survey account for about one half of the total gross domestic fixed capital formation, which in turn makes up about one sixth of the gross domestic product.

Estimates of totals are required at a fairly fine industry group level. A stratified random sample design is employed, with two levels of stratification. The first level consists of 47 industry groups corresponding to important study domains. At the second level, each domain is divided into size strata, where size is measured by the register variable employment, here abbreviated to EMP. There are typically four size strata within a domain. We refer to a cell within the cross-classification of domains by size strata as a design stratum. A sizeband corresponds to the collection of design strata with the same range of size values, where sizeband 4 ($20 \leq \text{EMP} \leq 49$) and sizeband 1 ($\text{EMP} \geq 300$) comprise the smallest and the largest units respectively. Sizeband 1 is completely enumerated, although some nonresponse occurs.

Currently, estimation in the capital expenditure survey is based on the combined ratio estimator (see e.g. Cochran, 1977) within a domain with register employment EMP as the auxiliary variable. Apart from register employment there is another important potential auxiliary variable, register turnover (TO), which currently is not used in the CAPEX survey. The combined ratio estimator is defined by combining design strata in sizebands 2, 3 and 4 within the domain. Design strata from sizeband 1 are not combined, see Table 4.1. From a model assisted point of view, a regression line with no intercept is fitted through the scatter of points in a plot of the study variable against the auxiliary variable. The line can be fitted separately for each stratum or to data that are combined from several strata. The two types of model give rise to the separate and the combined ratio estimator (see Särndal et al. 1992). If the model is relaxed other types of estimator will result. For example, if an intercept is allowed, the resulting estimator is the regression estimator. If the variance of the population scatter about the regression line is the same no matter what the value of the auxiliary variable, then the model is homoscedastic; as opposed to a heteroscedastic model in which the variance changes with the auxiliary. Different degrees of heteroscedasticity result in different estimators. All of these estimators are collectively called GREG estimators.

Table 4.1. Current sampling and estimation strategy in a domain

Design strata (employment sizebands within the domain)	Strategy
1	A completely enumerated stratum + the separate ratio estimator to account for nonresponse
2	Genuine sampling strata + combined ratio estimator
3	
4	

In what follows, I show both theoretically and with a practical example what may happen if univariate linear regression models are fitted to data that are not readily amenable to linear modelling. The class of univariate linear models I consider covers the vast majority of the models (and associated estimators) that are used for business surveys at national statistical institutes. There may, however, be other models and estimators that may ameliorate model misspecification. For example, the observation that a large proportion of the values of the CAPEX survey study variables are zero is not exploited in the models here. Karlberg (2000) uses a lognormal-logistic mixture model for a scalar variable that can take exact values with nonzero probability and is continuously distributed otherwise, which I utilise in Chapter 5.

When analysing the data we also devote considerably more time to the fitting of the models and their diagnostics than would generally be possible in a production process at a national statistical institute, thereby gaining insight into the properties of GREG estimators, and indeed into the nature of model assisted theory.

4.3 Influence on the GREG Estimator

Let z_k be the value of a scalar auxiliary variable that determines the heteroscedasticity in the regression of y_k on $\mathbf{x}_k$, and let γ be the heteroscedasticity coefficient defined by the specialisation of the variance function in model M:

$V_M(Y_k) = \sigma^2 z_k^\gamma$, $k = 1, 2, \ldots, N$.

For most survey populations, $1 \leq \gamma \leq 2$, and for business survey populations particular, γ is often about 1.5 (Brewer 2002, p. 58).

Let

$$\hat{\mathbf{B}} = \left(\sum_s w_k \mathbf{x}_k \mathbf{x}_k' / z_k^\gamma\right)^{-1} \left(\sum_s w_k \mathbf{x}_k y_k / z_k^\gamma\right)$$

be the GREG estimate of the finite population parameter $\mathbf{B}$. The definition of $\mathbf{x}_k$ will depend on the model; for example $\mathbf{x}_k' = \begin{pmatrix} 1 & x_{1k} & x_{2k} \end{pmatrix}$ if the model contains an intercept and two auxiliary variables. In general, I take the term 'auxiliary variable' to mean a 'proper' variable as opposed to an intercept (which, incidentally, Jean-Claude Deville has called an 'auxiliary variable free of charge').

The second term in (4.8) is a weighted sum of the residuals. This term is necessarily zero under all models we consider, except some of those involving heteroscedasticity. Recall that the general condition for the weighted sum of the residuals to vanish is (4.9). The function of this term is to make the GREG asymptotically design-unbiased (Wright, 1983; see also Särndal et al., 1992, sec. 7.3.4).

Next I derive a relationship between the g-weight of a sample unit and its influence on the value of the GREG estimate of $\mathbf{B}$. The most common measure of the influence of a sample unit is its DFBETA (Cook and Weisberg, 1982). This is defined by the change in the estimate of $\mathbf{B}$ when the unit is excluded from the sample data used to estimate $\mathbf{B}$. In the context of the regression model underlying GREG, the vector DFBETA for unit k is defined as

$$DFBETA_k = \left(\sum_s \mathbf{x}_j \mathbf{x}_j' / z_j^\gamma\right)^{-1} \left(\frac{\mathbf{x}_k}{z_k^{\gamma/2}}\right) \left(\frac{e_k}{1 - h_k}\right), \qquad (4.16)$$

where

$$h_k = \left(\mathbf{x}_k / z_k^{\gamma/2}\right)' \left(\sum_s \mathbf{x}_j \mathbf{x}_j' / z_j^\gamma\right)^{-1} \left(\frac{\mathbf{x}_k}{z_k^{\gamma/2}}\right)$$

is the leverage of unit k, and can be thought of as a measure of the remoteness of that unit from the rest of the sample points.

From (4.6) and (4.16) we see that for equal probability sampling with weights $w_k = a$ for all k,

$$g_{ks} = 1 + \frac{1}{a}\left(\frac{1-h_k}{e_k}\right)\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)' DFBETA_k . \qquad (4.17)$$

Recall that $\mathbf{t}_x$ is a vector whose elements are the population totals of the auxiliary variables defining $\boldsymbol{x}_k$ and $\hat{\mathbf{t}}_{x\pi}$ is the corresponding vector of Horvitz-Thompson estimates of these population totals. For example, under equal probability sampling and with a model consisting of an intercept and one auxiliary variable we have

$$\hat{\mathbf{t}}'_{x\pi} = \begin{pmatrix} N & N\bar{x}_s \end{pmatrix}, \qquad (4.18)$$

where $\bar{x}_s$ is the sample mean of the auxiliary variable. We will use the DFBETA in Section 4.6 to identify influential sample units.

4.4 Comparing estimates based on different models

As part of a review of the methodology used in the CAPEX survey, a study of different estimation methods for the survey was carried out in 1998. The estimators that were considered in this study are listed in Table 4.2. All of them are GREG estimators. Note that C/Rat in this table is the estimator currently used in the CAPEX survey (Table 4.1). Note also that all of them except S/Reg/1.5 and S/Reg/2.0 satisfy condition (4.9). Throughout we assume nonresponse in the survey is ignorable conditional on the stratified design.

The auxiliary variable $\mathbf{x}_k$, $k = 1, 2, \ldots, N$, defining the separate regression estimators does not need to be scalar. We considered situations where $\mathbf{x}_k$ is bivariate, being made up of the two auxiliaries, register employment and turnover, EMP and TO, with or without intercept. Furthermore the heteroscedasticity auxiliary $\mathbf{z}_k$ is often a scalar component of $\mathbf{x}_k$. We considered both TO and EMP as $\mathbf{z}_k = z_k$.

In classical design-based theory the two main properties of an estimator are its design variance and its design bias. The ratio and regression estimators in Table 4.2 are usually believed to be approximately design unbiased if the sample sizes within strata are fairly large, as is the case in the CAPEX survey. This may not be true if the model

does not fit data well. However, going along with the not too uncommon approach of essentially ignoring the risk of model misspecification, we prefer a more parsimonious model (and hence a computationally simpler estimator) to a more complex one, unless the gain in terms of variance from use of the latter is considerable. It would therefore seem reasonable to consider each of the estimation methods listed in Table 4.2, apply it to the CAPEX survey data for a number of quarters, estimate the design variances of the resulting estimates of total net capital expenditure, and choose the method that leads to an appropriate trade-off between low overall estimated design variance and parsimony.

Table 4.2. The estimators considered in the CAPEX survey review

S/E	The stratified expansion estimator.
C/Rat	The combined ratio estimator. Combined ratio estimates based on the scalar auxiliary x_k without intercept, are calculated by combining sizebands 2-4 within each domain, see Table 4.1.
S/Rat	The separate ratio estimator based on x_k without intercept.
S/Reg/0.0	The separate regression estimator. This is based on fitting a separate homoscedastic linear regression model to the study variable in terms of the auxiliary $\mathbf{x}_k = (1 \quad x_k)'$, within each design stratum (Särndal et al., 1992, Sec. 7.8).
S/Reg/1.0	As above, but now based on fitting a separate heteroscedastic linear regression model to the study variable in terms of the auxiliary $\mathbf{x}_k$, within each design stratum, with heteroscedasticity proportional to the unit power of a positive-valued scalar auxiliary variable z_k,.
S/Reg/1.5	As above, but with heteroscedasticity proportional to the 1.5 power of the auxiliary variable z_k,.
S/Reg/2.0	As above, but with heteroscedasticity proportional to the square of the auxiliary variable z_k,.

Table 4.3 shows what happens when this approach is applied within one domain of one of the quarters (waves) of the CAPEX survey. For confidentiality reasons we cannot give details that may help identify units; for this reason we refer to this domain as *domain V* from now on. The columns in this table show the estimate of total net capital

expenditure, its estimated variance, the CV (square root of the estimated variance divided by the estimate of total) and the *variance ratio* (the estimated variance of the total estimate divided by the estimated variance of the stratified expansion estimate). Register turnover and employment were used as auxiliary variables. The variance estimation software packages CLAN (Andersson and Nordberg, 1998) and GES (Estevao, Hidiroglou, and Särndal, 1995) were used for these computations. They gave very similar results. All variance estimation makes use of g-weights as in (4.12).

From the results set out in Table 4.3 the simple regression estimator S/Reg/0.0 with EMP as auxiliary seems preferable. The more complex bivariate regression estimator based on EMP and TO does offer some gain in terms of increased efficiency. It also shows stability over the different versions of S/Reg, and, as we shall see, this may be more important than increased efficiency. However, for the practical reasons indicated in Section 4.1, both in reality and here the attention was focussed on relatively simple univariate GREG estimators. Judging from Table 4.3, EMP seems more efficient than TO as an auxiliary variable. This was rather surprising. In the majority of other domains the reverse situation had been observed. Furthermore, the general experience at the Office for National Statistics is that within design strata there is a higher correlation between net capital expenditure and TO than between net capital expenditure and EMP.

4.5 Exploration of model problems

The strange results obtained in the previous section called for a more in-depth evaluation of the situation in domain V. Standard statistical procedures were therefore used to explore the fit of a variety of models for the study variable net capital expenditure in domain V, with the aim of identifying a "best" model (and hence estimator). This is in line with the ideas underlining the model-assisted approach to survey estimation (Särndal et al., 1992).

As a first step it was necessary to determine whether a linear model was adequate to describe the relationship between net capital expenditure and the auxiliary variable TO. This was assessed using a method proposed by Sen and Srivastava (1990, p. 198). The range of TO was divided into three parts, representing a compromise between having

an equal number of points in each part and dividing the whole range of TO into intervals of equal length. Median net capital expenditure and median TO were then determined in each part. Lines joining these median points, as in Figure 4.1, give an impression of the underlying relationship between net capital expenditure and TO in domain V. The boundaries between the 3 parts shown in Figure 4.1 correspond to the 67^{th} and 90^{th} percentiles of TO in domain V. For confidentiality reasons we are not allowed to show the true scales of the axes. It should be noted that the linearity displayed in this plot is not sensitive to definition of these boundaries – when they were moved around the impression of linearity remained.

Table 4.3. Estimates for domain V

Method	**X**	**Z**	**Estimate**	**Variance** $\div 10^8$	**CV**	**Variance ratio, %**
S/E	—	—	120,84	2.9	0.14	100.0
C/Rat/B	TO	TO	97,62	4.3	0.21	150.5
S/Rat	TO	TO	99,54	4.3	0.21	148.7
C/Rat/B	EMP	EMP	117,92	2.7	0.14	93.3
S/Rat	EMP	EMP	117,25	2.9	0.14	92.0
S/Reg/0.0	TO	—	117,62	2.7	0.14	93.8
S/Reg/1.0	TO	TO	108,32	3.2	0.16	109.3
S/Reg/1.5	TO	TO	96,62	8.5	0.30	295.2
S/Reg/2.0	TO	TO	71,45	35.3	0.83	1220
S/Reg/0.0	EMP	—	118,46	2.5	0.13	87.2
S/Reg/1.0	EMP	TO	122,13	2.8	0.14	98.2
S/Reg/1.5	EMP	TO	124,96	3.1	0.14	107.5
S/Reg/2.0	EMP	TO	126,88	3.3	0.14	114.4
S/Reg/0.0	TO,EMP	—	114,66	2.3	0.13	79.7
S/Reg/1.0	TO,EMP	TO	109,51	3.2	0.16	111.0
S/Reg/1.5	TO,EMP	TO	94,13	12.1	0.37	419.7
S/Reg/2.0	TO,EMP	TO	42,39	97.6	2.33	3381
S/Reg/0.0	TO,EMP	—	114,66	2.3	0.13	79.6
S/Reg/1.0	TO,EMP	EMP	115,17	2.3	0.13	79.5
S/Reg/1.5	TO,EMP	EMP	115,47	2.3	0.13	79.9
S/Reg/2.0	TO,EMP	EMP	115,80	2.3	0.13	80.6

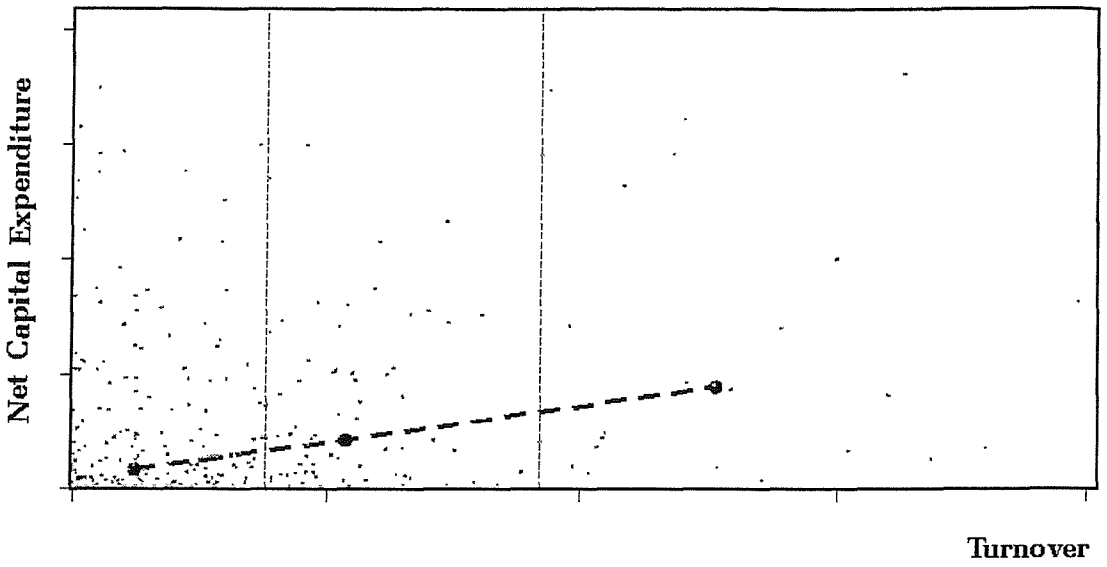


Figure 4.1. Relationship between net capital expenditure and turnover in domain V. Both axes are truncated.

There is a low correlation (0.12) between net capital expenditure and turnover in domain V. A model where the only auxiliary information is the number of units in domain V should therefore be kept in mind. Such a model leads to the expansion estimator.

Visual inspection of scatterplots of net capital expenditure against TO indicated substantial heteroscedasticity. We estimated the degree of this heteroscedasticity by fitting a model of the form

$$Var(y_k) \propto x_k^{\gamma}. \qquad (4.19)$$

Residuals and predicted values were computed, defined by the OLS fit of net capital expenditure against TO. An estimate of γ was then obtained as the slope of the OLS fit of the logarithms of the absolute values of these residuals against the logarithms of the predicted values. This estimate turned out to be about 1.7. It remained about the same even after deleting the five units with the smallest absolute residuals whose logarithms were very small. (Recall Brewer's (2002) remark that $\gamma \approx 1.5$ for most business surveys).

Standard diagnostics tools revealed that the residuals generated by all the models underlying the estimators in Table 4.2 were significantly non-normal. Furthermore, more detailed investigations indicated the presence of a number of influential points in

the sample data. These points were not only associated with very small and very large values of the auxiliary variables but also with moderate values of these variables combined with large values for net capital expenditure. That is, there were a number of outliers, defined both with respect to net capital expenditure and with respect to TO and EMP. Consequently it was not surprising that there were problems with fitting many of the linear models underlying the estimators in Table 4.2 to the data from domain V. In what follows we refer to observations with extreme values of TO or EMP as *outliers in x-space* and observations with large net capital expenditure values as *outliers in y-space.*

It is also worth noting at this stage that the sample from domain V was substantially unbalanced with respect to TO. In particular, the stratified expansion estimate of the total for TO in domain V was 30% larger than the known total of this quantity (taken from the population register). At the population level, over all domains, this estimate was 17% larger than the register value.

Following investigation of the anomalous behaviour of the GREG estimators in Table 4.2, we identified one particular design stratum in domain V, called *sizeband 3* in what follows, as an important contributor to this behaviour. This stratum gave vastly different estimates depending on the assumption of the degree of heteroscedasticity in the model underlying the GREG. In fact, it was this stratum that caused most of the differences between the estimates in Table 4.3.

To start, we note some basic facts about sizeband 3. There were 743 units in this stratum, of which 112 were sample respondents. The structure of the sample data for the stratum is displayed in Figure 4.2. There is one extreme TO value as well as large net capital expenditure values associated with fairly low values of TO. It would have been better if this unit were included in the completely enumerated sizeband 4 rather than in the sampled sizeband 3. In Chapter 5, I suggest a diagnostic that could identify this unit before the sample was drawn.

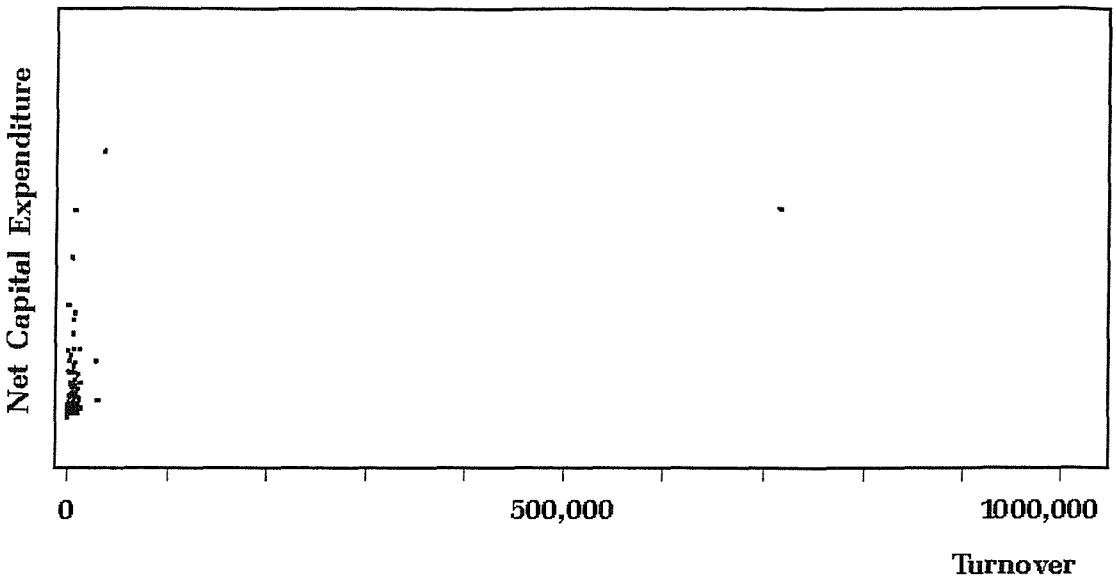


Figure 4.2. Respondents in sizeband 3 in domain V. The scale for turnover is fictitious for confidentiality reasons.

Figures 4.3a and 4.3b show regression lines fitted to the data in sizeband 3. The regression models underlying these lines are denoted E (mean model, $E(y) = \beta$), C (linear regression, homoscedastic), X1 (linear regression, $\gamma = 1.0$, see (4.19)), X3/2 (linear regression, $\gamma = 1.5$) and X2 (linear regression, $\gamma = 2.0$). By definition E is insensitive to the auxiliary variable. The high-leverage point (the extreme TO value) controls the regression line for model C. The outliers in y-space increase in importance as one moves from C to X1 to X3/2 to X2. Table 4.4 shows the GREG estimates of total net capital expenditure for sizeband 3 obtained under the different models. The resulting estimates of total net capital expenditure decreased monotonically from model E, through models C, X1, X3/2 to X2. From experience with other surveys at the ONS it was clear that the estimate produced by model X2 was far too low.

Comparing the estimates for model C with those for model X2 we see that the point estimate is higher for model C and the variance estimate is lower. As we will see, the reason for this is the presence of the outlier in x-space *and* the outliers in y-space. For comparison, the numbers in brackets in Table 4.4 are the estimates obtained after replacing the very large TO value by the median TO value, leaving net capital expenditure unchanged. Note the large differences in the estimates of total net capital expenditure generated by the estimates based on models C, X1, X3/2 and X2. Under

model E, of course, the estimates are unchanged. Table 4.4 also shows the parameters of the regression lines

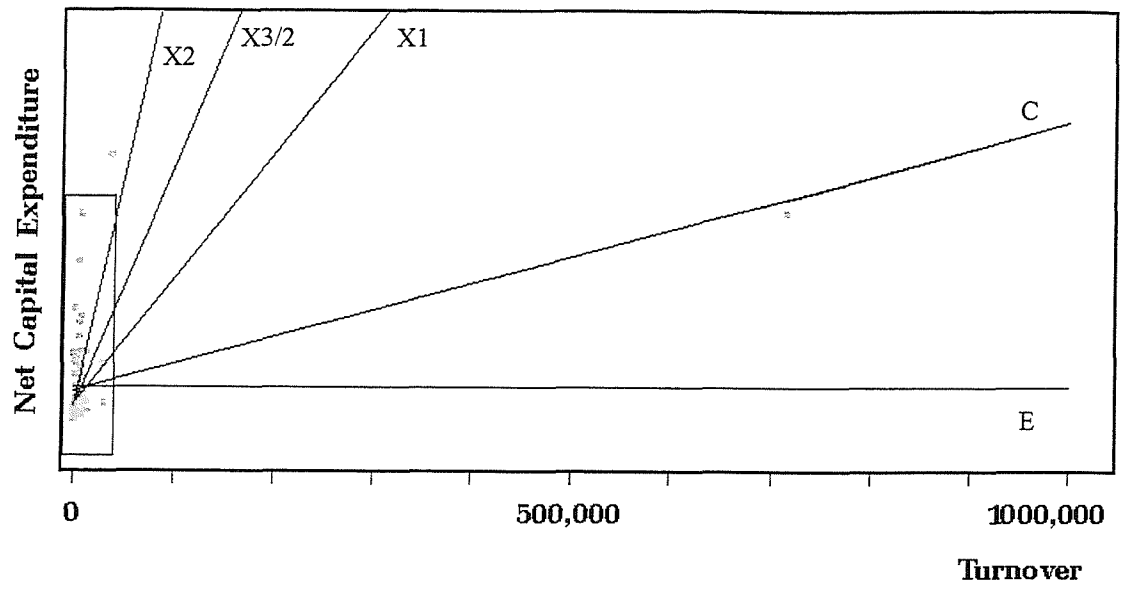


Figure 4.3a. Regression lines fitted to the data in Figure 4.2 under the models E, C, X1, X3/2 and X2. The area within the small rectangle is shown in more detail in Figure 4.3b.

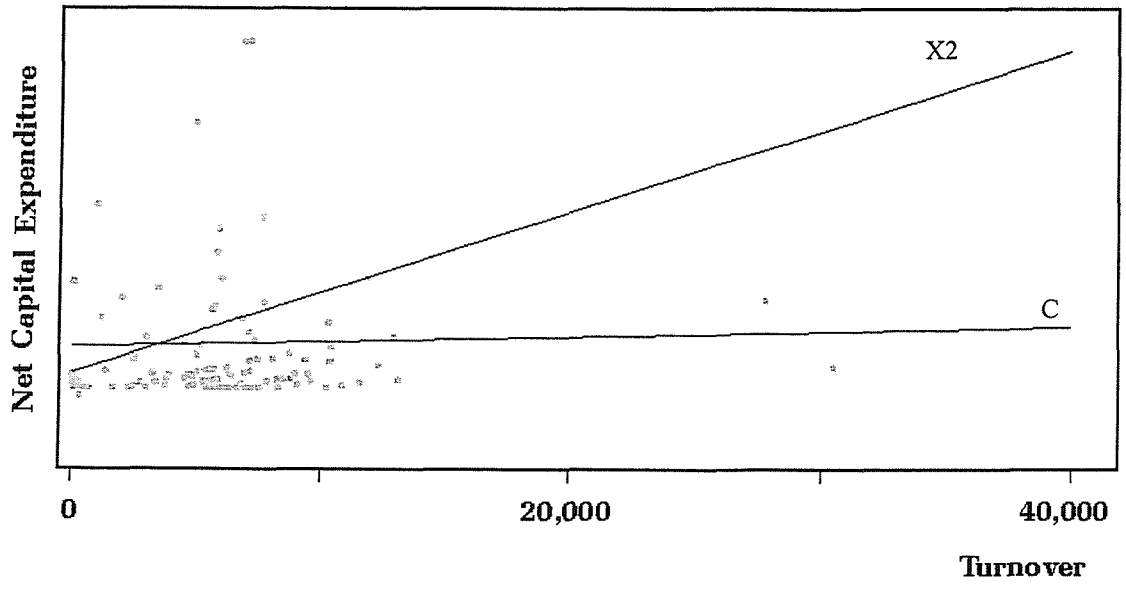


Figure 4.3b. Detail of Figure 4.3a. Models C and X2.

defined by the modified value of TO. Note that the line for X3/2 with original TO values almost coincides with the line for X1 defined by the modified values. This does

not imply that we get the same estimates. In fact, the estimate of total net capital expenditure was more than 100% higher for the new data under model X1 than for the original data under model X3/2. The reason for this is the presence of the residual correction term in the GREG. This will be discussed further below.

Table 4.4. Estimates for sizeband 3 in domain V. Turnover (TO) is the auxiliary variable. The figures in parentheses are obtained when the extreme TO value in sizeband 3 is replaced by the median TO value for the stratum

Estimator	Model	Estimate of total net capital expenditure / 1000	Standard deviation of total estimate / 1000	Intercept	Slope × 1000
S/E	E	55.5	10.4	74.8	0
S/Reg/0.0	C	52.5 (61.3)	9.9 (11.5)	65.4 (16.3)	0.74 (9.1)
S/Reg/1.0	X1	41.6 (60.0)	9.8 (11.3)	31.1 (28.8)	3.42 (7.2)
S/Reg/1.5	X3/2	28.8 (60.5)	24.5 (11.3)	26.6 (26.4)	6.55 (8.0)
S/Reg/2.0	X2	3.2 (64.1)	57.1 (12.3)	23.7 (23.7)	12.8 (13.7)

Table 4.5. Distribution of g-weights for sizeband 3 in domain V (extreme TO value not modified)

Model	g-weights, low – high	The g-weight of the outlier in x-space	Median of g-weights	Proportion of nonpositive g-weights
E	1 – 1	1.00	1.00	0/112
C	0.14 – 1.02	0.14	1.01	0/112
X1	0.54 – 13.9	0.54	0.60	0/112
X3/2	-1.3 – 58.6	0.92	0.12	25/112
X2	-22.8 – 245.9	0.99	0.01	57/112

4.6 A diagnostic for GREG estimation

In an effort to explain why such widely different estimates were obtained in Table 4.4, we computed the g-weights (4.6) generated under the different models. Table 4.5 shows the distribution of these g-weights.

Under all models considered here the unit with the lowest TO value attained the largest positive g-weight. Under model C and X1 the outlier in x-space attained the lowest g-

weight. Under model C all units except the outlier in x-space had g-weights close to unity. The outlier in x-space was the only point with which model C could not cope. Under models X3/2 and X2 on the other hand, the smallest units all had g-weights with large absolute values. These models (which standard diagnostics indicated as the most appropriate regression models for the stratum) effectively moved the estimation problem from the outlier in x-space to the outliers in y-space. The reason for this can be seen when one considers how the g-weights under the different models change as a function of the auxiliary variable.

Table 4.6. The g-weight functions under simple random sampling

Model	g-weight function
C	$1 + A_1 - A_2 x$
X1	$1 - B_1 + B_2/x$
X3/2	$1 - C_1/\sqrt{x} + C_2/(x\sqrt{x})$
X2	$1 - D_1/x + D_2/x^2$

For a given sample we can view the g-weights as function of $x = x_k$. It is shown in Appendix 4 that for models C, X1, X3/2 and X2 under simple random sampling these functions are as in Table 4.6. We refer to these functions as GC, G1, G3/2 and G2 below. The functions G1, G3/2 and G2 are shown in Figure 4.4. The constants B_1, C_1 and D_1 in Table 4.6 are defined by different values of γ, their general form is

$$\frac{n}{N}\frac{\hat{t}_{x\pi} - t_x}{\sum_s (x_k - \tilde{x}_s^{(\gamma)})^2 / z_k^{\gamma}}.$$

Also, $B_2 = B_1\tilde{x}_s^{(1)}$, $C_2 = C_1\tilde{x}_s^{(1.5)}$ and $D_2 = D_1\tilde{x}_s^{(2)}$, where n and N are the sample and population size respectively, and $\tilde{x}_s^{(\gamma)}$ is the weighted average:

$$\tilde{x}_s^{(\gamma)} = \frac{\sum_s x_k / z_k^{\gamma}}{\sum_s 1/z_k^{\gamma}}.$$

In sizeband 3 the values of $\tilde{x}_s^{(\gamma)}$ were approximately 13,000, 800, 140, and 57 for $\gamma = 0$, 1.0, 1.5, and 2.0 (models C, X1, X3/2, and X2), respectively.

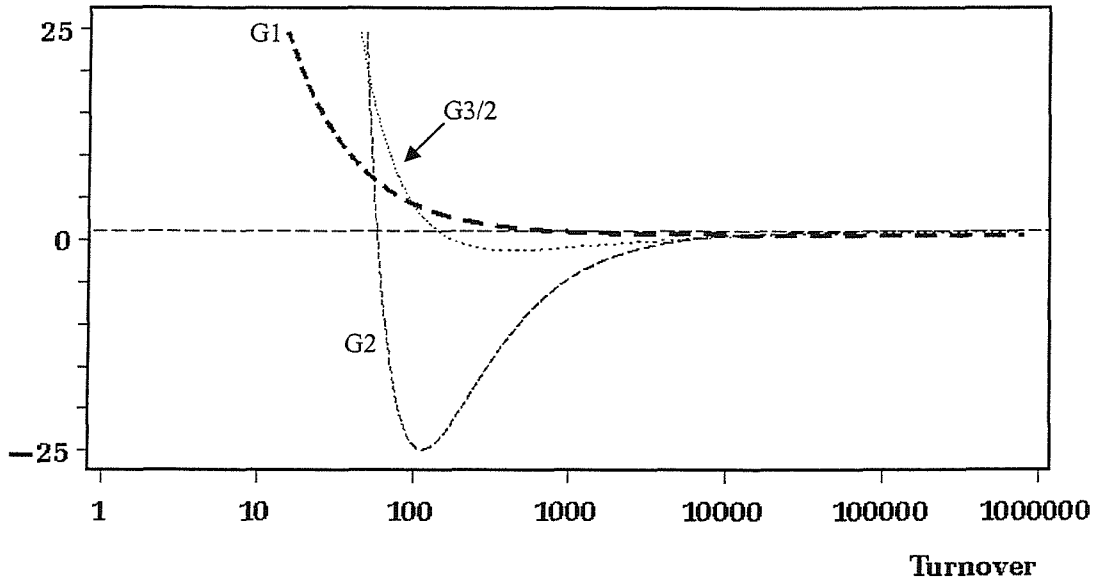


Figure 4.4. The g-weight functions G1, G3/2 and G2 for sizeband 3, as a function of turnover. The horizontal line represents model E.

As can be seen in Figure 4.4 and Table 4.6 the g-weights show undesirable behaviour for some ranges of x_k; i.e., g-weights smaller than 1 or very large g-weights. These ranges are listed in Table 4.7.

Table 4.7. Range of x_k for which the g-weights show undesirable behaviour

g-weight function	Range
GC	Large x_k
G1	Very small x_k and large x_k
G3/2	From the origin to and slightly beyond $3\tilde{x}_s^{(3/2)}$
G2	From the origin to and slightly beyond $2\tilde{x}_s^{(2)}$

The function GC decreases without bound as x increases. The slope of this function is

$$-A_2 = -\frac{n}{N}\frac{\hat{t}_{x\pi} - t_x}{\sum_s (x_k - \bar{x}_s)^2}.$$

Suppose the sample is overbalanced, as is the case in sizeband 3. Then $\hat{t}_{x\pi} - t_x$ is large and positive. In the case of sizeband 3 this term is 9.5 - 5.4 = 4.1 million, which can be compared to the estimated total of net capital expenditure for this stratum which is roughly 50,000 for most of the estimators we consider here. Although the slope might not be very large if the estimated stratum x-variance

$$(n-1)^{-1}\sum_s (x_k - \bar{x}_s)^2$$

is large, it is clear that the g-weights of units with very large values of x_k will be low. This confirms the behaviour of model C g-weights noted in Table 4.5.

The function G1 converges to $1 - \mathbf{B}_1$ as x increases, where $\mathbf{B}_1$ can be large (it is 0.45 in sizeband 3). In contrast, the functions G3/2 and G2 tend to 1 as x increases. That is, under the models X3/2 and X2 sample units with large values of the auxiliary variable tend to be essentially "inverse π-weighted".

The functions G3/2 and G2 have local minima at $3\tilde{x}_s^{(3/2)}$ and $2\tilde{x}_s^{(2)}$ respectively. The latter minimum is closer to the origin if the x_k that define $\tilde{x}_s^{(3/2)}$ and $\tilde{x}_s^{(2)}$ are positive. Note that the range of possible g-weights generated by any of the models C, X1, X3/2 and X2 is determined by the range of their corresponding g-weight functions. All three g-weight functions shown in Figure 4.4 are unbounded at zero. Consequently it is possible to obtain arbitrarily large g-weights for values of x_k close to zero under any of these models. Observe that near zero G2 increases faster than G3/2, which in turn increases faster than G1. Consequently we expect g-weights obtained under X2 to be most sensitive to small values of the auxiliary variable. This confirms the behaviour noted in Table 4.5. Furthermore we can see that negative g-weights are also possible under all three models, but their values are bounded from below. In particular, the smallest g-weight possible under X3/2 is defined by the minimum of G3/2, which is

$$1 - \frac{2C_1}{3\sqrt{3\tilde{x}_s^{(3/2)}}}.$$

In contrast, the smallest g-weight possible under X2 is the minimum value of G2:

$$1 - \frac{D_1}{4\tilde{x}_s^{(2)}}.$$

In the case of sizeband 3, Figure 4.4 shows that the minimum value for G2 is considerably less than that of G3/2 or G1.

Under simple random sampling the usual estimator of the design variance of a GREG is

$$\hat{V}\left(\hat{t}_{reg}\right) = K\sum_s \left(g_{ks}e_{ks} - \bar{g}_s\bar{e}_s\right)^2 , \qquad (4.20)$$

where K is a constant and $\bar{g}_s$ and $\bar{e}_s$ are averages of the g-weights and residuals (Särndal et al., 1992, Ch. 6). That is, the products of g-weights and residuals are the essential ingredients in the estimated variance. Given the sensitivity of the g-weights and the residuals associated with Models X3/2 and X2 to the influential points in the particular stratum we have considered, sizeband 3, it is not surprising that the variance estimates in Table 4.4 for the estimators S/Reg/1.5 and S/Reg/2.0 were extremely large.

Table 4.8a shows the residuals obtained in Sizeband 3. For all models, except model C, the lowest residual is associated with the outlier in x-space. The smallest residual for model C is generated by the sample unit with lowest net capital expenditure. As can be seen in the last two columns of Table 4.8a, the sums of residuals under models X3/2 and X2 are very large compared to either the corresponding sum of predicted values or to the estimates of total net capital expenditure (Table 4.4). Only a minor part of these large residual sums is accounted for by the residual associated with the outlier in x-space. The a-weights (inverse π) for sizeband 3 are all equal to 743/112. The weighting of every data point with $1/x_k^{\gamma}$ implied by models X3/2 and X2 makes the outlier in x-space less important. However, the influence of the outliers in y-space increases, compared to models C and X1, forcing the fitted regression model away from the bulk of the data. This can be seen in Figure 4.3b.

Table 4.8b shows the residuals with the TO value for the outlier in x-space replaced by the median TO. The only model that now differs considerably from the others is X2. The effect of the weighted sum of the residuals is the main reason why, even though the regression lines were similar, the estimate for total net capital expenditure for model X3/2 based on the original data was so much lower than that for model X1 with the outlier in x-space replaced by the median TO (Table 4.4).

Table 4.8a. Distribution of residuals in sizeband 3 in domain V.

Model	Residuals, low – high	Residual of lowest TO	Median	Proportion of positive residuals	Sum of residuals ÷ 1000	Ratio of the absolute value of the sum of residuals to the sum of predicted values
C	-77.6 – 1081.3	-47.4	-49.5	26/112	0	0
X1	-1922.5 – 1085.2	-13.2	-31.0	30/112	0	0
X3/2	-4151.5 – 1054.5	-8.7	-39.1	27/112	-26.2	0.48
X2	-8637.3 – 986.4	-6.1	-68.9	18/112	-83.5	0.96

Table 4.8b. Outlier in x-space replaced by median TO

Model	Residuals, low – high	Residual of lowest TO	Median	Proportion of positive residuals	Sum of residuals ÷ 1000	Ratio of the absolute value of the sum of residuals to the sum of predicted values
C	-260.9 – 1035.8	1.4	-41.3	28/112	0	0
X1	-214.1 – 1045.3	-11.0	-43.1	26/112	0	0
X3/2	-235.9 – 1038.7	-8.6	-44.7	24/112	-2.0	0.03
X2	-408.0 – 976.6	-6.0	-72.6	19/112	-27.3	0.30

The link between a sample point's $DFBETA_k$ value and its g-weight was shown by (4.17). Here we use this relationship to identify influential points in sizeband 3. A standardised value of $DFBETA_k$ is obtained by dividing this quantity by the residual variance computed without unit k. This measure is called $DFBETAS_k$ (Cook and Weisberg, 1982). Note that under stratified random sampling the first co-ordinate of $\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}$ in the regression estimators considered here is necessarily zero. The second co-ordinate of $DFBETAS_k$ therefore serves as a measure of the influence of unit k. For simplicity, we let the term $DFBETAS_k$ refer to the second co-ordinate in what follows. Figure 4.5 shows the values of this second co-ordinate plotted against the logarithm of TO under the model X3/2 with TO as the auxiliary. Belsley, Kuh and Welsch (1980) suggested that $DFBETAS_k$ with absolute values larger than $2n^{-1/2}$ should be marked for

further examination. The value of $2n^{-1/2}$ is about 0.19 in sizeband 3 (the dotted reference lines in Figure 4.5). As might be expected, the sample units in sizeband 3 that fall on or outside these boundaries are all associated with small or large values of TO. We observed the same pattern in all other strata as well.

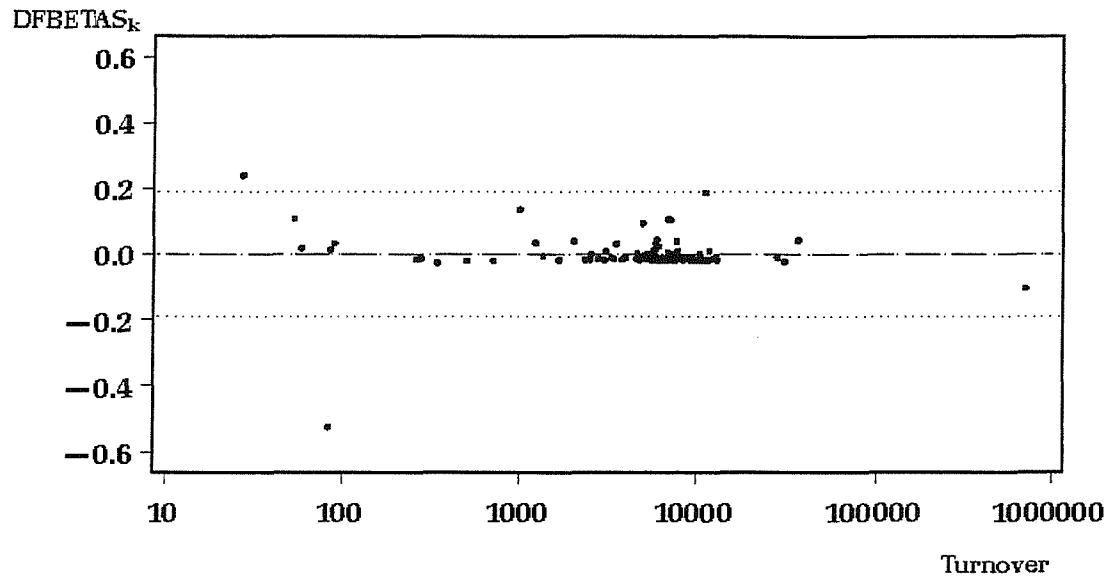


Figure 4.5. Influential points in sizeband 3 in domain V for Model X3/2

4.7 Using poststratification to minimise the impact of influential points

The idea here is to use the regression estimator S/Reg/1.5 (since we found in Section 4.5 that model X3/2 gave a better fit to the data than the other models), but only in that part of the stratum where there is little impact from outliers and influential points. To start, based on the observation that the influential points tended to be those with small or large values of TO, we partitioned sizeband 3 into three poststrata on the basis of TO. Estimation for sizeband 3 was then carried out using a method not influenced by outliers in x-space (expansion estimation) in poststrata 1 and 3, and using regression estimation based on S/Reg/1.5 in poststratum 2. Model X3/2 still gave the best fit in this subset of sizeband 3. The regression estimator generated an estimated variance about two thirds of the estimated variance of the expansion estimator for the poststratum. For sizeband 3 overall the poststratified procedure resulted in an estimated

variance that was 78% of the expansion estimator for the stratum. For comparison, regression estimation based on other models was conducted in poststratum 2. As shown in Table 4.9, the estimates are now very stable over the regression models, as opposed to the estimates given in Table 4.4.

Table 4.9. Poststratified estimates for sizeband 3 in domain V. TO is the auxiliary variable

Estimator	Model	Estimate of total net capital expenditure / 1000	Standard deviation of total estimate / 1000	Intercept in poststratum 2	Slope × 1000 in poststratum 2
S/E	E	53.52	9.30	59.07	0
S/Reg/0.0	C	52.64	9.16	41.11	9.84
S/Reg/1.0	X1	52.63	9.17	41.13	9.96
S/Reg/1.5	X3/2	52.63	9.17	41.14	10.00
S/Reg/2.0	X2	52.63	9.17	41.14	10.03

Choice of poststratum boundaries was subjective under this approach, but is advised by the need to make poststratum 2 as large as possible (to maximise the gains from regression estimation) while at the same time ensuring that the "outlier poststrata" 1 and 3 are not so small that variance estimation becomes problematic. I adhered to the common "rule" that there should be at least 20 units in every poststratum (see for example Särndal et al., 1992, p. 270). In fact, I picked exactly 20 units each for the two extreme poststrata.

An alternative to poststratification is GREG estimation based on restricted g-weights. I therefore computed restricted versions of S/Reg/1.5 and S/Reg/2.0 for sizeband 3 with the g-weights restricted to the interval (0.001, 8). This was done using Statistics Sweden's software CLAN (Andersson and Nordberg, 1998), which uses a method proposed by Deville and Särndal (1992). For the estimators S/Reg/1.5 and S/Reg/2.0 I obtained variance ratios of 314% and 467%, respectively, compared with the simple expansion estimator S/E. Other ranges for the g-weights were explored by trial-and-error, but either the algorithm did not converge, or worse variance ratios were

obtained. Thus, restricted g-weights gave considerably lower variances than unrestricted g-weights, but higher than poststratification. A total of 93 of the 112 observations in sizeband 3 got g-weights equal to the lower limit (0.001) for the estimator S/Reg/2.0.

As discussed in Section 4.1.2, an alternative to the GREG is to use some other distance function $G_k(w_k, w'_{ks})$. The distance function that may be most common in practice apart from (4.4), which leads to the GREG, is probably
$G_k(w_k, w'_{ks}) = w'_{ks} \log(w'_{ks}/w_k) - w'_{ks} + w_k$.
This distance function has especially been discussed in the context of raking ratio estimation, that is, when estimates for a cross-classification are calibrated on marginal totals. See Särndal et al. (1992, Section 7.9.2). With this distance function, positive weights are guaranteed but, as Deville and Särndal (1992) point out, these may be extreme.

4.8 Discussion

In the context of stratified simple random sampling I have explored the behaviour of some GREG estimators when the underlying models are misspecified due either to the presence of outliers in x-space or outliers in y-space (or both) in the sample data. I have shown a diagnostic for whether a GREG estimate is reasonable or not, a diagnostic which draws on the observation that for a given sample the g-weights can be seen as a function of the auxiliary variable. These g-weight functions can be graphed and inspected visually.

The g-weight of a sample unit is connected to its DFBETA, which is the change in the estimate of **B** when the unit is excluded from the sample data used to estimate **B**. Here I have used this measure of influence to identify a strategy which enabled us to keep the outliers away from the sensitive regression estimator. The strategy is to poststratify and use the expansion estimator for poststrata with highly influential units and a more efficient estimator, for example a regression estimator, for other poststrata.

The diagnostic and the following poststratification may seem impractical in official business statistics where large, often highly stratified data sets must be processed quickly. However, it should not be overwhelmingly onerous to produce and inspect g-weight functions and sums of residuals. Instead of graphs of g-weight functions lists of extreme g-weights can be produced. The poststratification may constitute an additional task for the survey statistician since the poststratum boundaries are in our approach determined on an ad-hoc basis. In the following chapter I will suggest a different approach.

The business survey example of Chapter 4 shows, for a set of real data, how important good modelling practice is. Different GREG estimators produced wildly different results. One regression estimator gave an estimated total which was less than 10% of the ordinary expansion estimate. All estimators we have explored are, at the first look, entirely reasonable. The difference between them lies entirely in model choice. The fact that the sample was considerably imbalanced against the auxiliary variable exacerbated the problem.

In conclusion I therefore reiterate the point made earlier on. It is just not true that GREG estimators are relatively robust to model choice. The fact that they are asymptotically design unbiased is *not* a substitute for a careful model specification search, especially when dealing with the highly variable and outlier prone populations that are the focus of many business surveys.

Chapter 5

A Comparison of Some Alternative Estimators of Totals for UK Business Surveys

5.1 Introduction

Since business data are skewed, outlier prone and often contain a large proportion of zeroes, it is not obvious that traditional methods of using auxiliary data, e.g. ratio and regression estimation, have the properties they often are believed to have, such as being virtually free from bias and have competitive variance. In fact, we saw in the previous chapter an example of GREG estimates with large errors. I study in this chapter whether the total can be estimated more accurately *and/or* more robustly by either robustifying these instances of the GREG estimator or by relying more explicitly on a model.

Most surveys at the ONS are multipurpose with customers who use the statistics in different ways. As was pointed out in Chapter 1, the estimated totals for business surveys are particularly important as they are input to the National Accounts.

What properties of an estimator of the total are vital? One could think of, e.g., small variance, negligible bias, good confidence intervals or minimum risk of obtaining estimates with large error; or versatility or ease of implementation. In this chapter, I report on a simulation study in which several GREG estimators are compared with a not widely used local regression estimator and a robust regression estimator that is novel in a design-based context. The former is similar to the GREG but has the ability to accommodate local departures from the underlying linear model.

For many estimators there is a choice of *model groups* to be made (Särndal et al. 1992,

Sec. 7.5). For example, a ratio model can be fitted within strata (leading to the separate ratio estimator) or across strata (the combined ratio estimator), where strata coincide with model groups in the former case while in the latter case the model group comprises the strata across which the model underlying the combined ratio estimation is fitted. There is little research on how to choose model groups. Silva and Skinner (1997) minimise the mean squared error to find the optimal set of auxiliary variables and thereby also model groups. I have simulated three types of model group partition and computed five criteria for each combination of estimator and type of model group partition. Two of the criteria are rather non-traditional.

Many of the business surveys at the ONS use a stratified simple random sampling design with four size strata within industry, three of which are genuine sampling strata and the one with the largest units is a completely enumerated (CE) stratum. There are two interval scaled variables on the frame: register employment and turnover. Industries are important domains of study. There are typically four size strata (sizebands) within a domain, see Figure 4.1. Sizeband 4 comprises the largest units and is completely enumerated, although some nonresponse occurs. I will, however, assume full response and ignore measurement errors and incomplete coverage of the target population.

In Section 5.2 the model groups and estimators used in the simulation study are defined, whose results are reported in Section 5.3. A simple way of ameliorating the effect of outliers in x-space is proposed in Section 5.4. Chapter 5 ends with a discussion.

5.2 Estimators

5.2.1 Aim

The set-up is similar to that of previous chapters. The aim is to estimate the total $t_y = \sum_U y_k$ of a study variable $\mathbf{y}' = (y_1, y_2, \ldots, y_N)$ on a population U. It is assumed that there is a known auxiliary variable $(N \times p)$ matrix $\mathbf{X}_N$ with $\mathbf{x}_k' = (x_{1k} \quad x_{2k} \quad \ldots \quad x_{pk})$ in row k. A sample s of size n is taken and $(\mathbf{x}_k, \ y_k)$ is

observed for all units k in the sample. Let stratum quantities and sets be indexed by h. For example, N_h and s_h refer to stratum size and the sample that is taken from stratum h. The populations of interest are industries. I assume that all units are correctly classified to industries before the sample is drawn. The terms domain (industry) and population, and the terms sizeband and stratum, will be used interchangeably.

5.2.2 Model groups

I use subscript g to index model groups in the partitioning that defines the G model groups (subsets) within each of which the model is to be fitted, $U = \bigcup_{g=1}^{G} U_g$. Three types of model group partition are studied:

a) groups coincide with strata;
b) one group consists of all genuine sampling strata and another group of the CE stratum within the industry (Figure 4.1);
c) all strata within an industry, including the CE stratum, constitute one group.

I refer to Case *b* as '**ONS model groups**', since this is the type of partition the ONS use for many business surveys. Cases *a* and *c* are labelled '**within strata**' and '**over all strata**'.

5.2.3 Point estimators

General form of point estimators

Many estimators used in practice are of the form

$$\hat{t}_y = \sum_{k \in U} \boldsymbol{\omega}_k' \mathbf{y}_s + \sum_{j \in s} \tilde{\omega}_j \left(y_j - \hat{y}_j \right), \qquad (5.1)$$

where $\boldsymbol{\omega}_k$ is a weight vector and $\tilde{\omega}_k$ a scalar, neither dependent on $\mathbf{y}$, and $\mathbf{y}_s' = (y_1, y_2, \ldots, y_n)$. The weights $\boldsymbol{\omega}_k$ and $\tilde{\omega}_k$ may be sample dependent. Often it is natural to interpret $\boldsymbol{\omega}_k' \mathbf{y}_s$ as a predicted value $\hat{y}_k = \boldsymbol{\omega}_k' \mathbf{y}_s$. Then (5.1) consists of a 'model-based' or 'synthetic' term plus a 'bias adjustment' or 'correction' term. We can refer to an estimator that can be written on the form (5.1) as a *projective bias adjusted estimator* ('projective' because the predicted values are projected to non-sample units or all population units). The Horvitz-Thompson estimator is a rather degenerate special

case of (5.1) with $\tilde{\omega}_j = 0$, $\forall j$, and $\boldsymbol{\omega}_1' = \left(\pi_1^{-1}, \ldots, \pi_n^{-1}\right)$, say, and $\boldsymbol{\omega}_k' = \mathbf{0}$ for $k > 1$. Let $\mathbf{X}_s$ be the (n x p) matrix that is the sample version of $\mathbf{X}_N$. For the estimators I study, $\mathbf{x}_k' = \begin{pmatrix} 1 & x_k \end{pmatrix}$ or $\mathbf{x}_k = x_k$. To see that the GREG can be written in the form (5.1), take $\tilde{\omega}_j = \pi_j^{-1}$ and $\boldsymbol{\omega}_k' = \mathbf{x}_k' \left(\mathbf{X}_s \boldsymbol{\Sigma}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s'\right)^{-1} \mathbf{X}_s \boldsymbol{\Sigma}_s^{-1} \boldsymbol{\Pi}_s^{-1}$, where and $\boldsymbol{\Pi}_s$ and $\boldsymbol{\Sigma}_s$ are diagonal matrixes with π_k and the residual variance σ_k^2 in position (k, k), respectively.

It can readily be shown that (5.1) is a linear estimator (Appendix 7), i.e., it can be written as a sample sum of the products of the y_k and some weights that do not depend on $\mathbf{y}$. This property is highly desirable from a national statistical institute's point of view. The main reason is practical: for example, it was mentioned in Chapter 4 that the weights can be thought of as 'grossing factors', stored in one column in a file and be applied in a simple way to all study variables without recomputation (Bethlehem and Keller 1987). Also, a linear estimator is internally consistent in the sense that if $\hat{t}_i$ is an estimator for a variable i, then $\hat{t}_1 + \hat{t}_2 = \hat{t}_{1+2}$ for the sum of the variables. Theoretical arguments do not abound, but one reason put forward by Sugden and Smith (2002), is that if the population parameter to be estimated is a single sum of a function of the population units (most parameters of practical interest are) then the estimator must have the same form if it is going to reduce to the parameter when $n = N$.

What the weight vectors are and how $\hat{\mathbf{y}}_s$ is computed may depend on the sampling design and the way the model is fitted. For example, the predicted values may be obtained with a least squares fit or with a model-assisted approach involving the inverse inclusion probabilities as weights. If $\boldsymbol{\omega}_k$ has zeroes in positions 'far' from position k then only elements in the vicinity of k in $\mathbf{y}_s$ will contribute to the predicted value $\hat{y}_k$. Thus there is a trade-off between using as many observations as possible to predict $\hat{y}_k$ and not letting possibly less relevant observations far away from k play a role. Also, there is a decision to make about the exact impact of units in the vicinity of k and that of those further away.

For some estimators the bias adjustment term always is zero (e.g. S/Rat, C/Rat and the regression estimators S/Reg/0.0 and S/Reg/1.0 in Table 4.2), for some other estimators

it will take whatever value to achieve some overall property. Consider Model M defined in Section 4.1.3. Regression estimators corresponding to a heteroscedastic regression model with $V_M(Y_k) = \sigma_k^2$ proportional to $x_k^{1.5}$ or x_k^2 (those estimators are labelled S/Reg/1.5 and S/Reg/2.0 in Table 4.2) are asymptotically design-unbiased if and only if the bias adjustment is allowed to be unbounded. As we saw in Chapter 4, this can lead to extremely poor performance for these estimators. Also, in a model-based setting the bias adjustment term can explicitly be regarded as an estimate of the bias due to model misspecification (Chambers, Dorfman, and Wehrly 1993). If a model M^*, $y_k = m(\mathbf{x}_k) + \varepsilon_k$, say, is correct and $\hat{t}$ is based on another, working model M^{**} (perhaps a simpler model than M^*) then the bias is $\sum_{k \in U-s} v_k$, where v_k denotes the non-observed residuals $\hat{y}_k - m(\mathbf{x}_k)$ for non-sample points. $E_{M^{**}}(v_k)$ can be estimated from the observed residuals r_j: $\sum_{k \in U-s} \hat{v}_k = \sum_{k \in U-s} \sum_{j \in s} \ddot{\omega}_{jk} r_j = \sum_{j \in s} \ddot{\omega}_{j\bullet} r_j$ with some appropriate weights. Hence $\sum_{j \in s} \ddot{\omega}_{j\bullet} r_j$ can be viewed as an estimate of the bias due to model misspecification and a special case of the bias adjustment term in (5.1). In a design-based framework such as GREG estimation, the second term of (5.1) may be $\sum_{j \in s} \pi_j^{-1}(y_j - \hat{y}_j)$; this term estimates $\left(-\sum_{k \in U} v_k\right) = t_y - \sum_{k \in U} \hat{y}_k$.

In general, there is an interplay between the choice of $\boldsymbol{\omega}_k$ and $\tilde{\omega}_k$. The bias adjustment term should normally be far smaller than the 'model-based' term. If not, there is an indication of model misspecification or a dysfunctional relationship between the structure of the data and what you do with them. As was shown in Chapter 4, the ratio of the bias adjustment term to $\sum_U \hat{y}_k$ is an important diagnostic for some GREG estimators, where a large value indicates model problems. However, not all estimators offer flexibility in the choice of weights $\boldsymbol{\omega}_k$ and $\tilde{\omega}_k$. For those estimators, once the estimator has been chosen one has to accept the weights that the estimator prescribes.

Winsorisation is one way of curbing the influence of outliers that is not included in this study. Winsorisation is a value-modification strategy where the value of a sampled unit is adjusted downwards if it is larger than a predefined cut-off (Kokic and Bell 1994).

Value modification could be viewed as artificial and hence it may run the risk of not gaining public acceptance. Furthermore, the main argument for Winsorisation is that of minimum mean squared error, even if it comes at the expense of a large bias. Minimum MSE may be strong argument for some surveys but less so for others. Many other outlier-robust estimators have been proposed, in particular model-based ones. Overviews include Chambers and Kokic (1993), Lee (1995), Valliant, Dorfman, and Royall (2000, Ch. 11) and Brewer (2002, Ch. 14).

Below follows a detailed description of the estimators used in the simulations.

The Horvitz-Thompson estimator

Let

$$\hat{t}_{yg\pi} = \sum_{s_g} w_k y_k \qquad (5.2)$$

be the expansion estimator for the group total $t_{yg} = \sum_U y_k I(k \in g)$, where $I(k \in g) = 1$ if unit k belongs to group g, and 0 otherwise. Here s_g is the part of the sample that falls in group g and $w_k = \pi_k^{-1}$ is the sampling weight for unit k. Let $\hat{t}_{y\pi}$ be the expansion estimator for the total t_y in U (i.e., the sum of the group estimates $\hat{t}_{yg\pi}$). I shall use the label 'E' for the expansion estimator in what follows.

GREG estimators for model groups

The ratio estimator for some set of model groups is (Särndal et al. 1992, Sec. 7.7):

$$\hat{t}_{yrat} = \sum_{g=1}^{G} t_{xg} \frac{\hat{t}_{yg\pi}}{\hat{t}_{xg\pi}}, \qquad (5.3)$$

where x denotes an auxiliary scalar. The label for this estimator will be '*Rat*'. With a slight change of the notation used in Chapter 4, I do not use the leading characters S and C in this chapter to label the separate and combined ratio estimator since I combine strata in two different ways: 'ONS model groups' and 'over all strata'. The Rat estimator for the ONS type of model group is the estimator the ONS uses for many business surveys, including the CAPEX survey.

The GREG estimator (4.8) can be written for model groups as

$$\hat{t}_{yreg} = \sum_U \hat{y}_k + \sum_s w_k (y_k - \hat{y}_k), \qquad (5.4)$$

where $\hat{y}_k = \mathbf{x}'_{kg}\hat{\mathbf{B}}_g$, $\hat{\mathbf{B}}_g = \left(\mathbf{X}'_{sg}\boldsymbol{\Sigma}^{-1}_{sg}\boldsymbol{\Pi}^{-1}_{sg}\mathbf{X}_{sg}\right)^{-1}\mathbf{X}'_{sg}\boldsymbol{\Sigma}^{-1}_{sg}\boldsymbol{\Pi}^{-1}_{sg}\mathbf{y}_{sg}$, and $\mathbf{X}_{sg}$ is a matrix with $\mathbf{x}'_{kg}$ in the *kth* row (two special cases to be given shortly). The data are assumed to follow a superpopulation model $\dot{M}$ for which $E_{\dot{M}}(Y_k) = \mathbf{x}'_{kg}\boldsymbol{\beta}_g$ and $V_{\dot{M}}(Y_k) = \sigma_k^2$, $k = 1, 2, \ldots, N$, where the moments are taken over the model. The Rat estimator $\hat{t}_{yrat}$ is a special case of (5.4) with $\mathbf{x}'_{kg} = I(k \in g)x_k$ and $V_M(Y_k) = \sigma^2 x_k$, $k = 1, 2, \ldots, N$. The '*Reg*' estimator is another special case with $\mathbf{x}'_{kg} = I(k \in g)(1 \quad x_k)$. For '*Reg/1.0*' we assume $V_{\dot{M}}(Y_k) = \sigma^2 x_k$, and for '*Reg/1.5*' $V_{\dot{M}}(Y_k) = \sigma^2 x_k^{1.5}$.

The choice of variance function that gives the best fit to the data used in simulations reported below is $V_{\dot{M}}(Y_k) = \sigma^2 x_k^{1.5}$, as is the case for many business surveys. Hence, we would expect good performance for Reg/1.5.

Local and robust regression estimators

The predicted values $\hat{y}_k$ in (5.4) can be replaced with some other predicted values that makes the estimator less sensitive to outliers and a nonlinear relationship between the study and auxiliary variable. Breidt and Opsomer (2000) use a local polynomial regression estimator weighted with inverse inclusion probabilities w_k to produce predictions $\hat{m}_k$ that in many cases will be close to $\hat{y}_k$. The estimator, here referred to as *Local*, is

$$\hat{t}_{yloc} = \sum_U \hat{m}_k + \sum_s w_k (y_k - \hat{m}_k). \qquad (5.5)$$

Chambers et al. (1993) and Dorfman (2000) suggest similar but model-based estimators. A *bandwidth* b_k and a smoothing window $(x_k - b_k, x_k + b_k)$ is defined. To predict y_k only observations whose auxiliary variable values are within the smoothing window are used. A weight function, referred to as the *Kernel function*, assigns the largest weights to units with auxiliary variable values close to x_k. A somewhat less general estimator than Breidt's and Opsomer's is

$$\hat{m}_k = \mathbf{e}'_{(2)1}\left(\mathbf{D}'_k\mathbf{W}_k\mathbf{D}_k\right)^{-1}\mathbf{D}'_k\mathbf{W}_k\mathbf{y}_s, \; k = 1, 2, \ldots, N, \qquad (5.6)$$

where $\mathbf{e}'_{(d)j}$ is a d-vector with 1 in the jth position and 0s otherwise, $\mathbf{D}_k$, $k = 1$, 2, ..., N, are n x 2 matrices, each with $\begin{bmatrix} 1 & (x_j - x_k) \end{bmatrix}$ in the jth row, $j = 1, 2, \ldots, n$; $\mathbf{W}_k$, for $k = 1, 2, \ldots, N$, are n x n diagonal matrices with $w_k b_k^{-1} K\left[(x_j - x_k) b_k^{-1}\right]$ in cell (j,j) with $K(\cdot)$ and b_k being the kernel function and the bandwidth, respectively. Apart from the presence of the sample weight $w_k = \pi_k^{-1}$ in $\mathbf{W}_k$, the prediction $\hat{m}_k$ is standard in the literature on local linear regression (e.g. Loader 1999). For a fixed bandwidth, Breidt and Opsomer prove that the sample weights in $\mathbf{W}_k$ and in (5.5) make $\hat{t}_{yloc}$ asymptotically design-unbiased. Their estimator has several other desirable theoretical properties. For example, the anticipated variance attains the Godambe-Joshi lower bound asymptotically. Like Breidt and Opsomer I use the Epanechikov kernel

$$K(u_{jk}) = \max\left[0, \frac{3}{4}\left(1 - u_{jk}^2\right)\right]. \qquad (5.7)$$

Chambers (1996) uses a bandwidth with fixed minimum length. For a fixed minimum length, however, the minimum bandwidth has to be longer than the longest distance between two consecutive x-values, which for skewed populations would prohibit truly local regression. Therefore, I use two types of *nearest neighbour bandwidth*. The first one is

$$b_k^{(s)} = x_{k+20} - x_{k-20}, \qquad (5.8)$$

where x_{k-20} and x_{k+20} are units in the sample file, sorted by x_k in ascending order. Note that $K(u_{jk}) = 0$ if $u_{jk} = (x_j - x_k)/b_k^{(s)} \geq 1$. Hence the kernel defines a window around unit k outside which units will not contribute to the prediction of $\hat{m}_k$. The window slides across stratum boundaries including the CE stratum (sizeband 4). Note that $\hat{m}_k$ will cancel out in the CE stratum in (5.5). Even so, the 20 smallest units in terms of the auxiliary variable in the CE stratum will be used in prediction of units in sizeband 3. If k is so small that x_{k-20} does not exist, x_{k-20} is taken as the minimum x-value and similarly for x_{k+20}. No adjustment has been made for these boundary effects.

For the other type of nearest neighbour bandwidth,

$$b_k^{(f)} = x_{k+40} - x_{k-40}, \qquad (5.9)$$

x_{k+40} and x_{k-40} are taken from the frame sorted by x_k in ascending order. The number of

sample units in the window will vary with the π_k: for parts of the frame with small sample fractions the local regression fit will tend to be more 'wiggly' than in more densely sampled areas. It seems reasonable that a point in a lightly sampled stratum should be given more influence. Care must be taken so that $\mathbf{D}_k'\mathbf{W}_k\mathbf{D}_k$ is not singular. The local regression estimators with bandwidths (5.8) and (5.9) are labelled *Local/s20* and *Local/f40*, respectively.

The prediction (5.6) can be rewritten as

$$\hat{m}_k = \bar{y}_{loc,k} + \left(x_k - \bar{x}_{loc,k}\right)\frac{\sum_{j\in s} q_{jk}\left(x_j - \bar{x}_{loc,k}\right) y_j}{\sum_{j\in s} q_{jk}\left(x_j - \bar{x}_{loc,k}\right)^2} \qquad (5.10)$$

where the q_{jk} are diagonal elements in the $\mathbf{W}_k$, $\bar{y}_{loc,k} = \sum_{j\in s} q_{jk} y_j \left(\sum_{j\in s} q_{jk}\right)^{-1}$, and $\bar{x}_{loc,k} = \sum_{j\in s} q_{jk} x_j \left(\sum_{j\in s} q_{jk}\right)^{-1}$. Note that $\bar{y}_{loc,k}$ is what would have been obtained with local constant prediction without the x-variable in 5.10. Formulation (5.10) shows that the local linear prediction is $\bar{y}_{loc,k}$ plus a term that counteracts effects stemming from the local slope of the data and the conditional bias that the predictor $\hat{m}_k = \bar{y}_{loc,k}$ would have exhibited in some neighbourhood of the boundary point x_1.

Let j and k index sample and population units, respectively. Note that (5.5) can be written

$$\begin{aligned}\hat{t}_{yloc} &= \sum_s w_j y_j + \sum_U \left[1 - I(k\in s) w_k\right]\hat{m}_k \\ &= \sum_{j\in s} w_j y_j + \sum_{j\in s}\left\{\sum_{k\in U}\left[1 - I(k\in s) w_k\right]\mathbf{e}_{(2)1}'\left(\mathbf{D}_k'\mathbf{W}_k\mathbf{D}_k\right)^{-1}\mathbf{D}_k'\mathbf{W}_k\mathbf{e}_{(n)j}\right\} y_j \qquad (5.11)\end{aligned}$$

that is, $\hat{t}_{yloc} = \sum_s w_{loc,js} y_j$ is a linear estimator with weights

$$w_{loc,js} = w_j + \sum_{k\in U}\left[1 - I(k\in s) w_k\right]\mathbf{e}_{(2)1}'\left(\mathbf{D}_k'\mathbf{W}_k\mathbf{D}_k\right)^{-1}\mathbf{D}_k'\mathbf{W}_k\,\mathbf{e}_{(n)j}. \qquad (5.12)$$

The subscript s reminds us that the weights are sample dependent. To continue the analogy with the GREG estimator, the local regression estimator weights can be partitioned into sampling weights w_j and 'local g-weights'

$$g_{loc,js} = 1 + \frac{1}{w_j}\left\{\sum_U \left[1 - I(k\in s) w_k\right]\mathbf{e}_{(2)1}'\left(\mathbf{D}_k'\mathbf{W}_k\mathbf{D}_k\right)^{-1}\mathbf{D}_k'\mathbf{W}_k\right\}\mathbf{e}_{(n)j} \qquad (5.13)$$

The Local/f40, Local/s20 and the E estimators are the only estimators in Chapter 5 that do not depend on the partitioning of the population into model groups. The Local estimators are of the projective bias adjustment form (5.1). They are flexible in that for a long bandwidth they will be similar to the GREG, and for a shorter bandwidth they will capture local model departures. Different kernels will give different distributions of weights ω_k within the window. The main difference between my and Briedt's and Opsomer's versions is my use of variable bandwidths.

Another estimator, here called *RobReg/f40,* was inspired by Welsh and Ronchetti (1998) and Kuk and Welsh (2001). One difference is that I take a design-based approach. In (5.6), $\mathbf{y}_S$ is replaced with $\tilde{\mathbf{r}}' = (\tilde{r}_1, \tilde{r}_2, ..., \tilde{r}_n)$, where the tilde indicates a robust fit obtained with bounded-influence estimation (to be specified shortly), to produce a smoothed value $\hat{m}_k^*$. The advantage of projecting $\tilde{\mathbf{r}}' = (\tilde{r}_1, \tilde{r}_2, ..., \tilde{r}_n)$ to each frame unit k is to allow for an asymmetric distribution of the residuals. Hence I robustify in two dimensions, first horizontally through the bounded-influence regression, then vertically through the smoothing of each $\tilde{\mathbf{y}}_k$ separately. Here the bandwidth $b_k^{(f)}$ in (5.9) was applied. It is conjectured that RobReg is approximately design-unbiased.

The bounded-influence method utilises the $DFFITS_k$ of each observation k, which is a well known measure of how much the prediction for this observation's x-value would change in terms of standard deviations of the predicted value if the regression line is refitted without observation k. Welsch (1980) suggests the use of the inverse $DFFITS_k$ as regression weights, a method analysed by Ryan (1997, Ch. 11). Belsley, Kuh and Welsch (1980) suggest as a rule of thumb for univariate regression that observations with larger absolute value of $DFFITS_k$ than $2n^{-0.5}$, n being the number of observations, should get special attention. The regression weights proposed by Welsch are

$$\delta_k = \begin{cases} 1 \text{ if } |DFFITS_k| \le 2n^{-0.5} \\ 2n^{-0.5}|DFFITS_k|^{-1} \text{ if } |DFFITS_k| > 2n^{-0.5} \end{cases} \qquad (5.14)$$

The regression parameters are estimated with weighted least squares with the weights $\delta_k x_k^{-3/4}$. The residuals are

$$\tilde{r}_k = \frac{y_k - \mathbf{x}_{kg}\tilde{\boldsymbol{\beta}}_g}{x_k^{3/4}} \qquad (5.15)$$

RobReg is robust to outliers. However, it is not of form (5.1). It is not linear and it has not the internal consistency property. Theoretical properties such as bias will be developed elsewhere.

Mixture model estimators

The Karlberg (2000) estimator can be seen as a transformation–retransformation estimator. It is based (model-based) on a mixture model. First I define a model $\ddot{M}$ to which a lognormal assumption will be added later on. Let Z_k be the logarithm of the study variable $Y_k > 0$. Assume that $y_1, y_2, ..., y_N$ are realisations of the random variables $Y_1, Y_2, ..., Y_N$, and, conditional on the auxiliary variable, $E_{\ddot{M}}(Z_k|Y_k > 0) = \mu_g = \mathbf{x}'_{kg}\boldsymbol{\beta}_g$, $V_{\ddot{M}}(Z_k|Y_k > 0) = \sigma_g^2$ where $\mathbf{x}'_{kg} = \mathbf{1}(k \in g)(1 \quad x_{2k})$, with x_{2k} being the logarithm of the auxiliary variable, provided that $x_{2k} > 0$. The parameter $\boldsymbol{\beta}_g$ is estimated through OLS regression applied to the logtransformed data. The model $\ddot{M}$ differs from that of Karlberg (2000) in that I allow for different model groups but not heteroscedasticity in the logscale. Not to burden the notation, subscript g is suppressed from now. Let $\mathbf{X}$ be the matrix with $\mathbf{x}'_k$ in the kth row, and let subscript s indicate the corresponding sample entity. To estimate the total of the nonsampled units on the original scale, the sum of the back-transformed predicted values of the study variable are multiplied by a bias correction factor. Let a_{kk} be the diagonal elements in a matrix $\mathbf{X}(\mathbf{X}'_{s+}\mathbf{X}_{s+})^{-1}\mathbf{X}'$, which is rather similar to the 'hat matrix', with s+ indicating that the matrix is restricted to positive sample values of the study variable. Let $\hat{Z}_k$ be the predicted value for unit k on the logscale, i.e. $\hat{Z}_k = \mathbf{x}'_k\hat{\boldsymbol{\beta}}$. It is reasonable to assume that $\hat{Z}_k$ is approximately normally distributed, and hence that $\exp(\hat{Z}_k)$ follows a lognormal distribution. Then $E_{\ddot{M}}[\exp(\hat{Z}_k)] = \exp(\mu + a_{kk}\sigma^2/2)$, (see e.g. Casella and Berger, 1990, for the mean of a lognormal distribution, and e.g. Sen and Srivastava, 1990, for the variance of $\hat{Z}_k$). Hence $\exp(\hat{Z}_k)$ is a biased estimator of Y_k on the original scale. Under the additional assumption that Y_k follows a lognormal distribution with mean and variance given by $\ddot{M}$, so that $E_{\ddot{M}}(Y_k) = \exp(\mu + \sigma^2/2)$,

Karlberg derives a approximately model-unbiased predictor:

$$\hat{Y}_k = \exp\left(\hat{Z}_k\right)\exp\left[\frac{\hat{\sigma}^2}{2}\left(1-a_{kk}\right)-\frac{\hat{\sigma}^4}{4n_+}\right]. \qquad (5.16)$$

where n_+ is the number of positive elements in the model group and

$$\hat{\sigma}^2 = \frac{\mathbf{Z}'_{s+}\mathbf{Z}_{s+} - \hat{\beta}'\mathbf{X}'_{s+}\mathbf{X}_{s+}\hat{\beta}}{n_+ - 2} = \sum_{k=1}^{n_+} e_k^2/(n_+ - 2), \qquad (5.17)$$

with e_k being the residuals on the logscale. If $n_+ \leq 2$, then the denominator of (5.17) is set to 1; this happened only once in the simulations. The *Logn/pr* estimate of a total for a model group g is

$$\hat{T} = \sum_{k=1}^{N_g - n_g} \hat{p}_{hk}\hat{Y}_k + \sum_{k=1}^{n_g} Y_k \qquad (5.18)$$

where $\hat{p}_{hk} = n_h^{-1}\sum_{k\in h} I\left(Y_k > 0\right)$ is the sample proportion of units with positive value of the study variable in the sizeband h unit k belongs to. Alternatively, a logistic model is fitted within each sizeband to obtain an estimated probability $\hat{p}_h(x)$ for a unit with a certain x-value to have a positive y-value. This estimator is labelled *Logn/log*.

In the simulations it often happened that the two groups defined by whether the study variable is zero or not were completely separated in a sense that is best explained by an example: if all x-values for zero study variable values are smaller than those of the positive study variable values, then the groups are completely separated. Then no ML estimates of the parameters of the logistic model exist. In this case, $\hat{p}_h(x)$ was set to one for x-values greater than the average of the largest and smallest of the sample x-values on either side of the separation point, and zero otherwise. For the rather more unlikely contingency that the groups were completely separated apart from one shared sample x-value ('quasi-complete separation'), $\hat{p}_h(x)$ was set to ½ for the shared point. If the sample x-values overlap the ML estimates exist and are unique. Overlap, complete and quasi-complete separation partition the space of data configurations (Albert and Anderson 1984).

The mixture model estimators are sensitive to errors in $\hat{\sigma}^2$. Therefore, *RLogn/pr* is obtained by replacing (5.17) in Logn/pr with a robust estimate of the variance, $\hat{\sigma}_R^2$.

The beta coefficient $\hat{\beta}_R$ was computed through a regression relationship within model groups of $\log(y_k)$ on $\log(x_k)$, with homoscedastic errors and weights (5.14). The estimate $\hat{\sigma}_R$ was taken as 1.4826 times the median absolute deviation of the residuals $y_k - \hat{\beta}_R x_k$ from their median. The constant 1.4826 is chosen so as to make $\hat{\sigma}_R^2$ consistent if the residuals were standard normal.

The mixture model estimators are attractive in their relative simplicity, but they are not in general design unbiased. They cannot be written on the form (5.1). Transforming to log scale makes many business survey datasets nicely linear, apart from the zero-valued observations. The flipside is the need to estimate the potentially influential parameter σ^2 and, as a consequence of the lognormal model assumption, the need to estimate the propensity for a unit to have a zero value. The partition of the sample data into positives and zeroes makes the effective sample data set smaller.

5.2.4 Variance estimators

Although Chapter 5 focuses on point estimation, I have computed coverage probabilities and hence variance estimates. The variance estimators below account for the original stratification through the inclusion probabilities. It is shown in Appendix 5 that a g-weighted variance estimator (4.12) for $\hat{t}_{yrat}$ for the three types of model group combined with stratified simple random sampling (STSI) is

$$\hat{V}_{STSI}(\hat{t}_{rat}) = \sum_{h=1}^{H} \left(\frac{t_{xg}}{\hat{t}_{xg\pi}} \right)^2 \left[N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{1}{n_h - 1} \sum_{s_h} (e_k - \bar{e}_h)^2 \right], \qquad (5.19)$$

where $e_k = y_k - x_k' \hat{B}_{ratg}$ with $\hat{B}_{ratg} = \dfrac{\hat{t}_{yg\pi}}{\hat{t}_{xg\pi}}$. Here the g-weights are $g_{ks} = t_{xg} / \hat{t}_{xg\pi}$.

For example, for ONS model groups, (5.19) is

$$\hat{V}_{STSI}(\hat{t}_{rat}) = \left(\frac{t_{x1}}{\hat{t}_{x1\pi}} \right)^2 \sum_{h=1}^{3} \left[N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{1}{n_h - 1} \sum_{s_h} (e_k - \bar{e}_h)^2 \right], \qquad (5.20)$$

where the totals in group $g = 1$ (all genuine sampling strata $h = 1, 2$, and 3) are

$$\hat{t}_{x1\pi} = \sum_{h=1}^{3} \sum_{s_h} w_k x_k \text{ and } t_{x1} = \sum_{h=1}^{3} \sum_{U_h} x_k .$$

It is also shown in Appendix 5 that the g-weighted variance estimator for the group regression model is

$$\hat{V}_{STSI}(\hat{t}_{reg}) = \sum_{h=1}^{H} \left[N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{1}{n_h - 1} \sum_{s_h} g_{ks}^2 (e_k - \bar{e}_h)^2 \right], \qquad (5.21)$$

where $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_g$ with $\hat{\mathbf{B}}_g$ defined above, and

$$g_{ks} = 1 + \left(\mathbf{t}_{xg} - \hat{\mathbf{t}}_{xg\pi} \right)' \left(\sum_s w_j \mathbf{x}_{jg} \mathbf{x}_{jg}' / \sigma_j^2 \right)^{-1} \left(\mathbf{x}_{kg} / \sigma_k^2 \right). \qquad (5.22)$$

The g-weights for regression models were studied in great detail in Chapter 4.

The variance estimator used here for Local is

$$\hat{V}_{STSI}(\hat{t}_{yloc}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{S_h^2}{n_h}, \qquad (5.23)$$

where $S_h^2 = \dfrac{\sum_{s_h} (\gamma_k - \bar{\gamma}_h)^2}{n_h - 1}$, $\gamma_k = y_k - \hat{m}_k$, and $\bar{\gamma}_h = n^{-1} \sum_{s_h} \gamma_k$.

Breidt and Opsomer (2000) show that (5.23) is for a fixed bandwidth a consistent estimator (in the finite population sense described in Chapter 4) of an approximate variance

$$AV(\hat{t}_{yloc}) = \sum \sum_U \Delta_{kl} \Gamma_k \Gamma_l / \pi_k \pi_l , \qquad (5.24)$$

where $\Gamma_k = y_k - m_k$, m_k being the smoothed values one would get with (5.6) based on the whole population, and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ with π_{kl} being the probability that both units k and l are included in the sample. The expression (5.24) has the same form as the usual approximate variance of the GREG, see Chapter 4. The local g-weights (5.13) could be inserted into (5.23). The estimator (5.23) was used for RobReg as well with $y_k - \hat{m}_k^*$ replacing γ_k.

I have not computed variance estimates for the mixture model estimators. While Karlberg (2000) suggests a rather complicated variance estimator for her estimator, we shall see that there are bias problems with the mixture model estimators that make them less appealing, whether the variance can be estimated accurately or not.

5.3 Simulations based on MIDSS and CAPEX data

Some domains of the Quarterly Survey of Capital Expenditure (CAPEX) and the Monthly Inquiry for the Distribution And Services Sector (MIDSS), both conducted by the ONS, provided data for a simulation study. The sampling design for both surveys is reported in Figure 5.1. The study variable is turnover for the MIDSS. Here net capital expenditure was used as the CAPEX study variable. For the purposes of this study, the auxiliary variable for both the MIDSS and the CAPEX was turnover as recorded on the frame, which is the frame variable that correlates most strongly with either of the study variables.

Figures 5.2 to 5.6 show scatter plots of three MIDSS and two CAPEX domains on logscale. For confidentiality reasons the scales of the axes are suppressed. Note that the

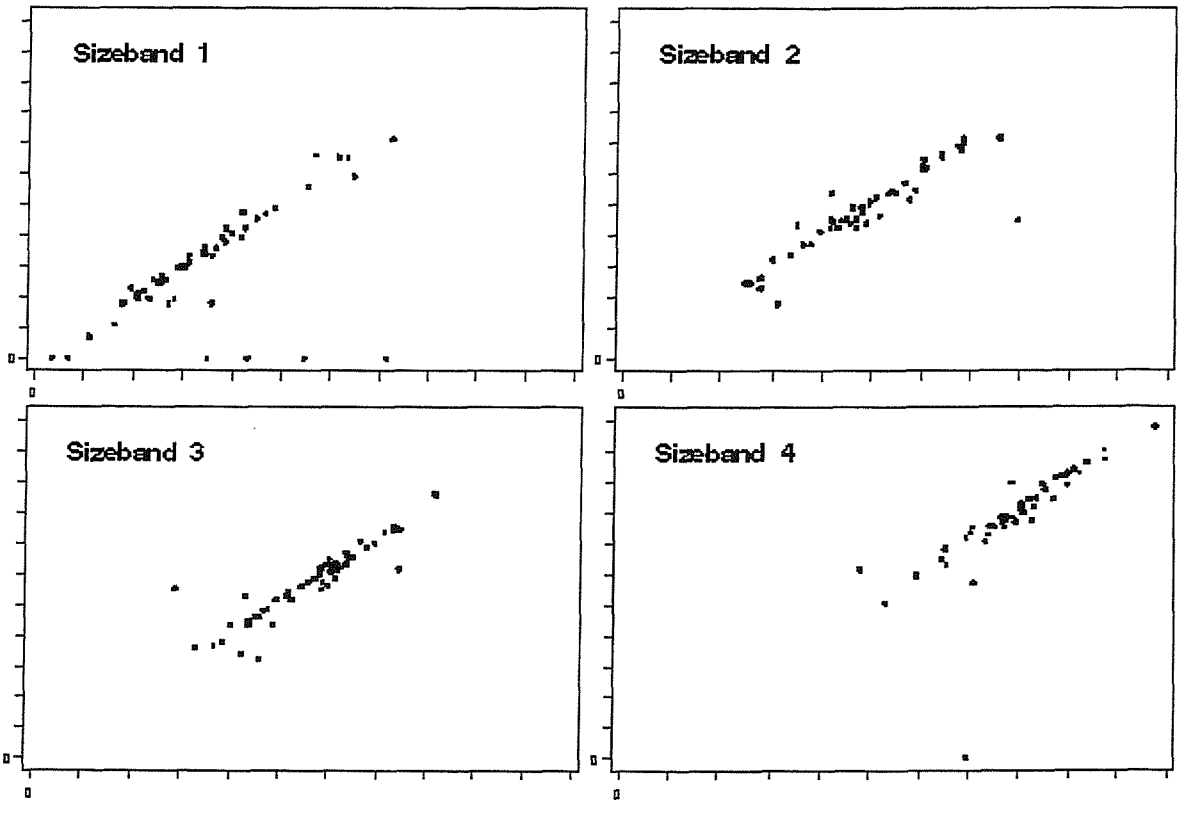


Figure 5.1. MIDSS, domain A. Log of the study variable against log of the auxiliary variable, with unity added to both variables

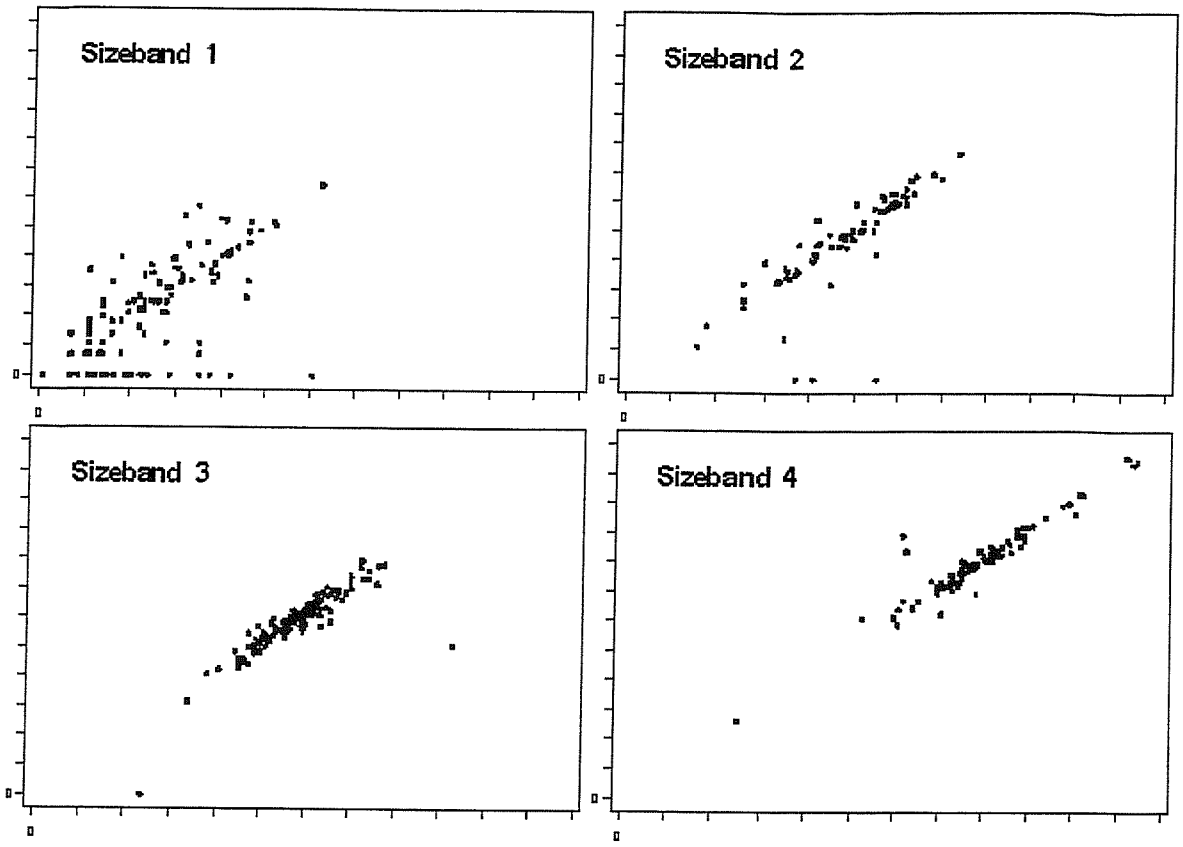


Figure 5.2. MIDSS, domain B. Log of the study variable against log of the auxiliary variable, with unity added to both variables

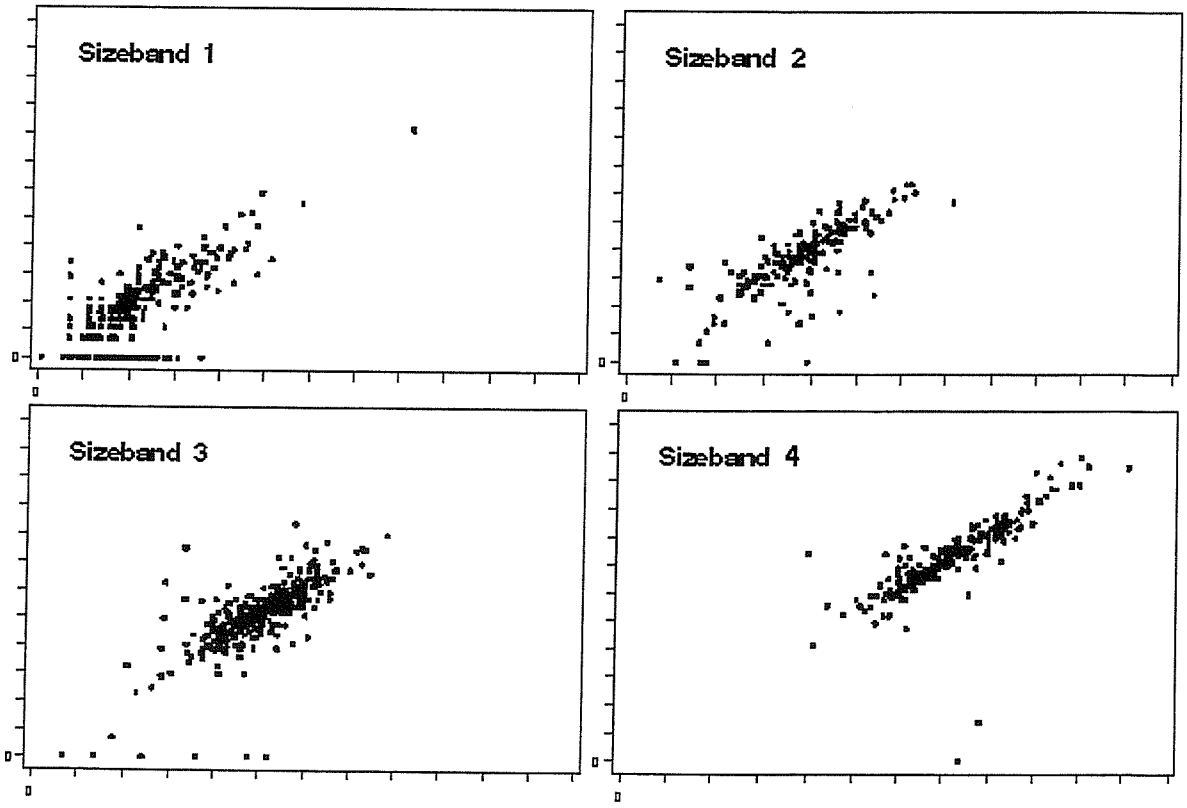


Figure 5.3. MIDSS, domain C. Log of the study variable against log of the auxiliary variable, with unity added to both variables

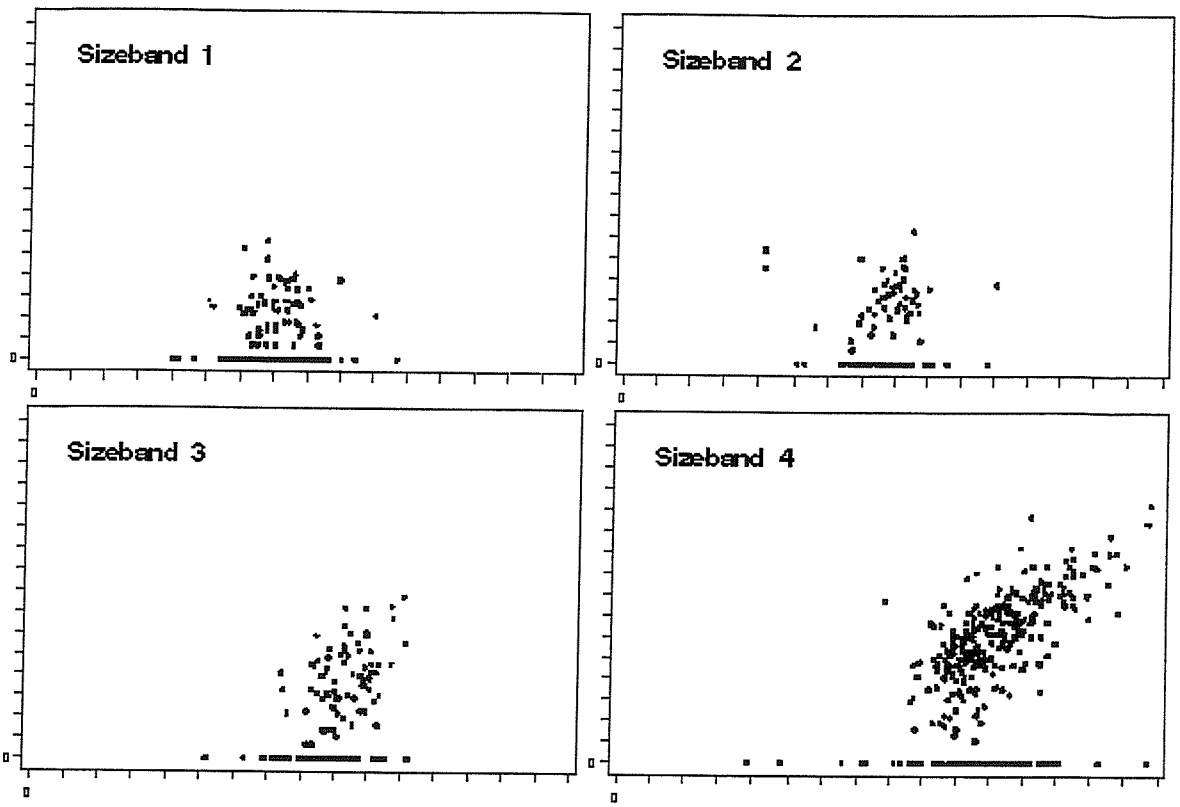


Figure 5.4. CAPEX, domain U. Log of the study variable against log of the auxiliary variable, with unity added to both variables

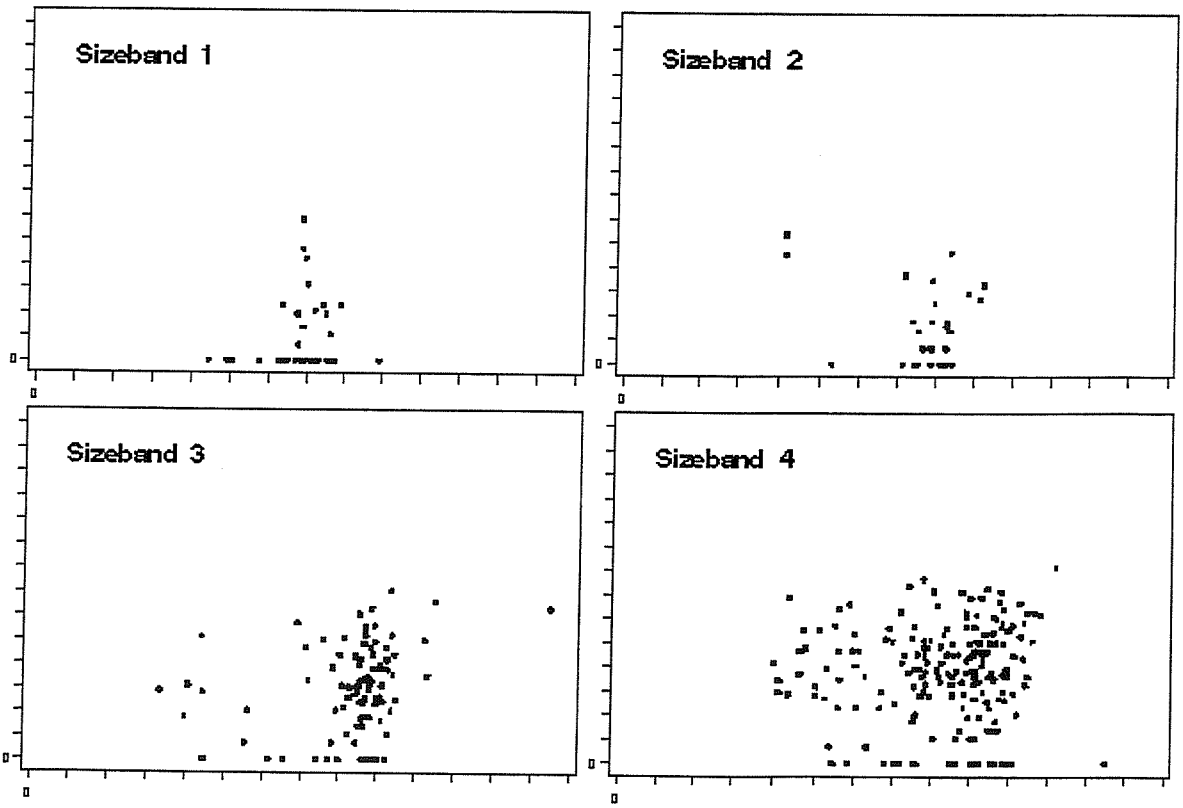


Figure 5.5. CAPEX, domain V. Log of the study variable against log of the auxiliary variable, with unity added to both variables

CAPEX domains U and V are very different from the MIDSS domains A, B and C. Note in particular that the largest value of the auxiliary variable in domain V is in sizeband 3, i.e. a sampled sizeband. The proportion zero values for the study variable is rather small for the MIDSS. For the CAPEX it is about 40% and 20% for domains U and V respectively.

Domain V is the domain studied in Chapter 4. Unlike in Chapter 4, in this chapter I draw subsamples from the original set of respondents to simulate a sample survey.

In the simulations below I have used the existing strata but allocated the sample on the frame variable turnover, with the exception of CAPEX domain V where 'even' sample sizes were chosen. Sample sizes used in simulations are shown in Tables 5.1 to 5.5. The columns labelled by N_h contain the original number of respondents. Here they are considered population sizes. One thousand samples were drawn from each domain.

Table 5.1. Sample sizes for the simulated samples, MIDSS domain A

Size-band	N_h	n_h	n_h/N_h %
1	39	9	23
2	33	19	57
3	52	32	62
4	43	43	100
Sum	167	103	62

Table 5.2. Sample sizes for the simulated samples, MIDSS domain B

Size-band	N_h	n_h	n_h/N_h %
1	73	5	7
2	51	28	54
3	88	67	77
4	74	74	100
Sum	286	174	61

Table 5.3. Sample sizes for the simulated samples, MIDSS domain C

Size-band	N_h	n_h	n_h/N_h %
1	206	59	29
2	129	13	10
3	305	128	42
4	213	213	100
Sum	853	413	48

Table 5.4. Sample sizes for the simulated samples, CAPEX domain U

Size-band	N_h	n_h	n_h/N_h %
1	254	25	10
2	107	24	21
3	133	51	38
4	393	393	100
Sum	887	493	56

Table 5.5. Sample sizes for the simulated samples, CAPEX domain V

Size-band	N_h	n_h	n_h/N_h %
1	40	10	25
2	33	10	30
3	112	30	27
4	202	202	100
Sum	387	252	65

5.3.1 Properties of an estimator

I am interested in the following measures.

1. ***Coefficient of variance (CV)***. The ratio of the standard deviation of the simulated point estimates to the true total.
2. ***Bias***. The mean of the errors of the simulated estimates divided by the true total.
3. ***Coverage probability***. The 95% confidence intervals computed as ± 1.96 times the square root of the variance estimates (5.19) – (5.23).
4. What proportion of the point estimates that are further away from the true total than 0.675 times the standard error of the point estimates. The constant 0.675 is so chosen that if the estimates are normally distributed then 50% will be ***Non-centred.***
5. The maximum of the absolute differences between the 95% and 5% percentile and the true total, divided by the true total. This has the flavour of a minimax criterion with the survey error, i.e. the difference between estimate and population parameter, as loss function. This criterion is labelled ***Large Error***.

Unfortunately, there is no hard and fast rule about which properties to prioritise. The first three measures are the traditional properties that together with the MSE often are taken as the guiding rule. Despite the strong position of the MSE, there is some arbitrariness in using squared error loss as the one and only loss function (see also Robert, Hwang, and Strawderman, 1993, in particular the discussion that follows the

paper). Although there is not likely to be any other loss function that is less arbitrary than squared error loss, this loss function is not sacrosanct in any way. Based on the statistical adage that most sampling distributions are 'normal in the middle', we might expect close to 50% of the estimates to be Non-centred. The fifth measure, Large Error, is particularly important in official statistics where the publication of bad estimates may sometimes lead to great losses for society and may also be detrimental to the reputation of the national statistical institute. I would argue that the criterion Large Error is easier to understand and explain to the public than are the CV or the MSE.

5.3.2 Simulation results

Tables 5.6 to 5.10 report on the CV and the other measures for five domains. Table 5.11 shows the biases of the variance estimators. In the tables, the type of model group is indicated by a number: 1 for 'within strata', 2 for 'ONS model groups' and 3 for 'over all strata'. For example, as seen in Table 5.6, the estimator most widely used in ONS business survey estimation, here called Rat_2, gives poorer CV than does the expansion estimator, E, for four out of five domains. Some other observations are listed in connection to each table. Boxplots of the point estimates are shown in Appendix 8.

Comments to Table 5.6:

1. The Logn and RLogn estimators fitted within strata broke down for several domains. The reason that Logn and RLogn 'within strata' broke down for three domains is that few units are sampled from sizeband 1 (Tables 2, 4 and 5), many of which may be zero. As all three mixture model estimators only use positive values of the study variable to fit a lognormal model, the fit will be very unstable for small samples. However, these estimators fitted over all strata performed well.
2. Among design-based estimators it is E that gives the smallest CV for several domains. The reason is poor correlation between study and auxiliary variables. This lack of correlation arises either through outliers (domain B, Figure 5.2) or through overall weak association (domains U and V, Figures 5.5 and 5.6).
3. For weak-association and outlier-prone domains (such as U and V) larger groups give smaller CV. The opposite is true for the MIDDS domains.
4. In terms of CV, RobReg is among the worst estimators for several domains,

including domain V for which estimation is particularly challenging.

5. Local is among the best design-based estimators for the CAPEX, and not far worse than Rat and Reg/1.0 for the MIDSS (between 5% and 24% higher CV than the best Rat or Reg/1.0).
6. In domain V, Figure 5.5, the extreme-leverage observation in sizeband 3 causes extrapolation far beyond the sample range for all samples without this observation. The result is unstable estimates for most estimators.
7. Reg/1.5 is worse than Reg/1.0 throughout, and far worse for domain V. This is rather surprising considering that the model underlying Reg/1.5 fits data better than that of Reg/1.0.

Table 5.6. Per cent coefficient of variation (CV) for five domains.

	MIDSS			CAPEX	
	A	B	C	U	V
E	2.42	0.92	1.61	1.13	6.62
Rat_1	1.51	1.29	1.05	1.24	15.46
Rat_2	1.74	1.29	1.16	1.15	15.6
Rat_3	1.83	1.34	1.17	1.14	24.83
Reg/1.0_1	1.52	1.28	1.03	1.42	14.01
Reg/1.0_2	1.72	1.28	1.15	1.16	14.2
Reg/1.0_3	1.83	1.34	1.16	1.14	7.94
Reg/1.5_1	1.7	1.38	1.07	1.4	21.74
Reg/1.5_2	1.78	1.36	1.21	1.41	42.11
Reg/1.5_3	1.83	1.41	1.2	1.14	19.35
Local/f40	1.87	1.36	1.19	1.13	6.42
Local/s20	1.83	1.36	1.07	1.14	6.69
RobReg/f40_1	1.87	1.49	1.06	1.19	19.54
RobReg/f40_2	1.83	1.37	1.17	1.27	45.85
RobReg/f40_3	1.82	1.42	1.17	1.13	12.24
Logn/pr_1	1.59	362619	0.96	8E44	..
Logn/pr_2	1.26	1.09	1.02	0.49	6.95
Logn/pr_3	0.77	0.81	0.58	0.33	4.47
Logn/log_1	1.71	379092	0.98	1E45	..
Logn/log_2	1.38	1.1	1.05	0.51	7.01
Logn/log_3	1.03	0.82	0.6	0.38	4.57
RLogn/pr_1	1.67	525230	0.85	8E44	..
RLogn/pr_2	1.45	1.16	0.8	0.58	11.83
RLogn/pr_3	0.86	0.87	0.46	0.26	6.04

Table 5.7. Bias for five domains (per cent of true total)

	MIDSS			CAPEX	
	A	B	C	U	V
E	0.06	-0.05	-0.01	-0.01	0.07
Rat_1	0.42	0.22	0.11	-0.02	9.91
Rat_2	0.12	0.09	0.06	-0.01	9.20
Rat_3	0.04	-0.01	0.01	-0.01	7.41
Reg/1.0_1	0.42	0.27	0.07	-0.22	4.93
Reg/1.0_2	0.13	0.09	0.06	-0.04	-2.80
Reg/1.0_3	0.04	-0.01	0.01	-0.01	0.88
Reg/1.5_1	0.25	0.12	0.05	-0.16	1.62
Reg/1.5_2	0.09	0.01	0.04	-0.10	-2.82
Reg/1.5_3	0.05	-0.01	0.02	-0.02	0.23
Local/f40	0.04	0.01	0.07	-0.01	-1.90
Local/s20	0.06	0.02	0.07	0	-3.04
RobReg/f40_1	0.04	0	0.02	-0.06	0.07
RobReg/f40_2	0.03	-0.02	0.02	-0.05	4.40
RobReg/f40_3	0.03	-0.02	0.01	-0.01	1.04
Logn/pr_1	1.11	14700	-0.04	2E43	3E167
Logn/pr_2	1.27	0.90	1.42	3.39	7.43
Logn/pr_3	1.72	1.19	1.10	3.72	35.2
Logn/log_1	1.02	13300	0.13	4E43	3E167
Logn/log_2	1.19	0.97	1.65	3.46	7.59
Logn/log_3	1.65	1.27	1.32	3.98	35.47
RLogn/pr_1	4.00	19900	-1.07	2E43	3E167
RLogn/pr_2	0.1	0.19	-0.05	3.32	9.75
RLogn/pr_3	0.74	0.51	-0.6	3.32	33.34

Comments to Table 5.7:

1. The bias can be very large for weak-association populations with extreme-leverage points, such as domain V displayed in Figure 5.5. This is particularly true for Rat applied to a population that calls for a positive intercept. For other populations, linear or not, or outlier prone or not, the bias is negligible for the design-based estimators, including RobReg.
2. Logn/pr and Logn/log tend to give positive bias. This is in accordance with Karlberg's (2000) empirical findings. Rlogn/pr seems rather better in this respect. Consequently, Logn does not perform well in terms of root MSE (not shown here).

The reason for the poor performance does not seem to be lack of lognormality, see Appendix 6. Logn breaks down in terms of bias for the same reason as stated for the CV above.

3. The bias is often slightly larger for small model groups but still negligible for all domains but one.

Table 5.8. Coverage probability, in per cent, for five domains.

	MIDSS			CAPEX	
	A	B	C	U	V
E	89.0	92.7	90.6	91.5	87.9
Rat_1	73.8	63.9	90.8	85.6	73.6
Rat_2	80.6	65.1	91.2	89.2	68.8
Rat_3	79.9	64.7	93.2	85.9	31.7
Reg/1.0_1	72.7	63.7	91.1	89.6	84.8
Reg/1.0_2	80.6	65.1	91.1	92.1	86.3
Reg/1.0_3	80.0	64.7	93.3	85.9	90.2
Reg/1.5_1	80.6	65.3	91.2	92.5	90.1
Reg/1.5_2	81.1	65.5	91.9	96.3	85.0
Reg/1.5_3	80.2	64.7	93.1	88.0	87.4
Local/f40	79.4	64.7	93.7	85.3	79.4
Local/s20	78.9	64.7	94.4	85.2	75.5
RobReg/f40_1	79.4	64.6	92.9	84.2	55.2
RobReg/f40_2	80.0	64.7	93.8	81.7	36.9
RobReg/f40_3	80.2	64.7	93.4	85.5	63.7

Comments to Table 5.8:

1. The coverage probability is poor in many cases. No estimator except the E estimator gives acceptable coverage for all domains. The reason is the non-normality of the estimates for many domains, in particular B. The sample distribution is bimodal for this domain. The main reason for this to happen is the high leverage point in sizeband 3 visible in Figure 5.2.
2. If the population is linear (such as the MIDSS domains, Figures 5.2 to 5.4), then 'within stratum' model groups seem worse than larger model groups, in terms of coverage probability. This makes the lower CV for 'within stratum' a moot point.
3. The variance estimator for RobReg seems unreliable.

Table 5.9. Per cent Non-centred estimates for five domains.

	MIDSS			CAPEX	
	A	B	C	U	V
E	52	47	51	36	53
Rat_1	55	62	51	34	61
Rat_2	56	64	51	36	67
Rat_3	57	71	50	36	80
Reg/1.0_1	52	58	52	27	48
Reg/1.0_2	56	63	51	35	47
Reg/1.0_3	57	71	50	36	54
Reg/1.5_1	55	67	53	33	55
Reg/1.5_2	56	70	50	35	40
Reg/1.5_3	56	71	51	36	45
Localf40	56	68	50	36	56
Local/s20	58	68	49	37	57
RobReg/f40_1	54	66	51	36	39
RobReg/f40_2	56	72	51	37	27
RobReg/f40_3	56	72	50	36	46
Logn/pr_1	57	0	49	0	0
Logn/pr_2	68	41	76	100	62
Logn/pr_3	94	74	89	100	100
Logn/log_1	56	0	50	0	0
Logn/log_2	64	42	81	100	63
Logn/log_3	86	79	93	100	100
RLogn/pr_1	51	0	78	0	0
RLogn/pr_2	59	57	51	100	46
RLogn/pr_3	66	39	78	100	100

Note: a point estimate is Non-centred if it is further away from the true total than 0.675 times its standard error.

Comments to Table 5.9:

Percentages far away from 50 in Table 5.9 indicate a sampling distribution that is far from normal. Numbers less than 50% indicate that the estimates are more tightly centred, and hence have smaller error, than what would be expected if they were normally distributed.

1. The distributions of the estimates are clearly non-normal for domains B, U, and V.
2. Most of the design-based estimators are similar in terms of the Non-centred criterion. However, E and RobReg stand out, giving equal or better performance.

Table 5.10. Per cent estimates with Large Error for five domains.

	MIDSS			CAPEX	
	A	B	C	U	V
E	4.1	1.5	2.7	1.2	11.5
Rat_1	2.9	2.3	2	1.2	34.8
Rat_2	2.9	2.2	2.1	1.2	33.7
Rat_3	2.8	2.1	1.9	1.2	33.5
Reg/1.0_1	2.9	2.4	1.9	1.2	29.5
Reg/1.0_2	2.8	2.1	2.1	1.2	17
Reg/1.0_3	2.8	2.1	1.9	1.2	13.7
Reg/1.5_1	3	2.3	1.9	1.6	33.7
Reg/1.5_2	2.8	2.2	2.1	1.7	89.4
Reg/1.5_3	2.8	2.2	2	1.2	36
Local/f40	2.8	2.1	2	1.2	8.8
Local/s20	2.7	2.2	2	1.2	25
RobReg/f40_1	2.8	2.2	1.9	1.2	8.1
RobReg/f40_2	2.8	2.2	1.9	1.2	8.1
RobReg/f40_3	2.8	2.2	1.9	1.2	8.1
Logn/pr_1	3.9	2.7	1.6	3.9	31.9
Logn/pr_2	3.4	2.8	3.2	4.2	19.7
Logn/pr_3	3	2.5	2.1	4.3	42.5
Logn/log_1	3.9	2.7	1.9	4.1	25.7
Logn/log_2	3.4	3	3.4	4.3	20.2
Logn/log_3	3.1	2.6	2.3	4.7	42.5
RLogn/pr_1	3.2	3.4	0.5	4.6	374.7
RLogn/pr_2	2.4	2.1	1.3	4.3	33.1
RLogn/pr_3	2.1	1.9	0.2	3.8	43.8

Note: Large Error is defined as the maximum of the absolute differences between the 95% and 5% percentile and the true total, divided by the true total. Hence, a small value of this measure indicate a small risk for obtaining an estimate with a large error.

Comments to Table 5.10:

1. In terms of the Large Error criterion, RobReg is the best estimator for domain V and no worse than any other estimator for other domains. Local/f40 also performs well.
2. The Large Error and Non-centred criteria combined show that the distribution of 'within stratum' estimates are both more peaked and fat-tailed than the distribution

for larger model groups. Again, this makes the lower CV for 'within stratum' a moot point.

Table 5.11. Bias of variance estimates, in per cent, for five domains.

	MIDSS			CAPEX	
	A	B	C	U	V
E	3.8	4.2	3.4	3.1	2.5
Rat_1	-54.2	-51.6	-16.1	-14.9	-38.9
Rat_2	-18.9	-28.4	-15.7	-9.3	-29.9
Rat_3	-1.4	0.2	-5.8	-6.1	-22.1
Reg/1.0_1	-55.8	-50.9	-14.8	-13.1	-29.4
Reg/1.0_2	-18.3	-27.5	-14.3	-9.4	-29.9
Reg/1.0_3	-1.2	0.5	-5.4	-4.2	-18.4
Reg/1.5_1	-34.7	-16.7	-12.8	-12.8	-13.5
Reg/1.5_2	-3.6	-2.9	-8.7	-3.6	-9.7
Reg/1.5_3	-2.0	1.3	-5.8	-1.1	2.1
Local/f40	1.3	1.8	-7.3	5.3	4.3
Local/s20	5.6	0.8	14.1	2.2	1.5
RobReg/f40_1	1.3	-16.1	7.0	8.2	-5.3
RobReg/f40_2	5.5	-0.7	-8.1	1.7	-6.8
RobReg/f40_3	7.4	-7.2	-8.4	5.7	-3.0

Note: Variance estimates were computed using formulae in Section 5.2.4

Comments to Table 5.11:

1. The bias is negative and with very large absolute value in many cases, in particular for 'within stratum' for GREGs. Wu and Deng (1983) also found that the variance estimator used here for Rat gave large negative bias. While the large biases contribute to an explanation of the poor coverage probabilities, it is not the only reason: note the weak correlation between the coverage probabilities in Table 5.8 and the bias in Table 5.11. The boxplots in Appendix 8 show lack of normality of point estimates.
2. For GREGs, the bias decreases with size of model groups.
3. The variance estimator for the Local estimators is gives reasonable results in terms of bias.

Conditional properties

In classical design-based inference conditional properties are not given much attention. However, most official statistics users would agree that if the sample is severely imbalanced in terms of an auxiliary variable that is believed to have some 'explanatory power', then properties such as design-unbiasedness that hold only as an 'summary measure' over all possible samples are less appealing than they would have been with a balanced sample – unless the estimators have been shown to have good properties conditional on the estimate of the auxiliary variable. To give one simple example: if a properly drawn random sample from a domain turns out to contain mostly larger-than-average businesses in terms of a frame variable and if the estimated total of the study variable for the domain is higher than last year, no informed user would believe in this estimate. This argument is formalised by Thompson (1997, Ch. 5).

Scatter plots of the estimated total of the study variable against the estimated total of the auxiliary variable for MIDSS domain A are shown in Figures 5a-c, with one plot per type of model group. It is reasonable to plot against the estimated auxiliary variable total or the difference between this estimate and the population parameter since either alternative gives a measure of the imbalance in the sample. Here the estimates for the study variable are plotted against the expansion estimate of the auxiliary total, $\hat{t}_{x\pi}$. A loess curve (Cleveland 1979) was fitted to the 1000 pairs of study variable and auxiliary variable estimates for each of the estimators E, Rat, Reg/1.0 and 1.5, Local/s20 and f40, RobReg, Logn/pr and log, and Rlogn/pr. The loess curve was fitted with the SAS procedure Proc Loess with the smoothing parameter set to 0.20 which makes the bandwidth comprise 20% of the units, see SAS Institute (2000). The distance from the dotted horizontal line, which indicates the true total, and the fitted value gives an impression of the conditional bias. This is essentially the same type of plot Valliant (1987) produced for an STSI design, although he focuses on variance estimation and use a different smoother.

As seen in Figures 5.7 a-c, the expansion estimator E has the largest conditional bias, apart from the region $280{,}000 < \hat{t}_{x\pi} < 285{,}000$ where $\hat{t}_{x\pi}$ is close to the population total. We would expect this conditional bias to disappear in the GREG type of estimators since they are designed to cope with this type of imbalance. Indeed, this is

the case for 'within stratum' model groups (Figure 5.6 a), but, interestingly, the other model groups overadjust for the imbalance (Figures 5.7 b and c). For these model groups the GREGs are similar to the Local regression estimators and RobReg. With the unconditional bias deducted, the estimators with the smallest conditional bias for regions outside $280{,}000 < \hat{t}_{x\pi} < 285{,}000$ are the mixture model estimators. In terms of conditional bias (adjusted for the unconditional bias) Logn/pr, Logn/log, and RLogn/pr are all similar, and this for all modelgroups. The difference discernable from Figures 5.7 a-c is that RLogn has smaller unconditional bias.

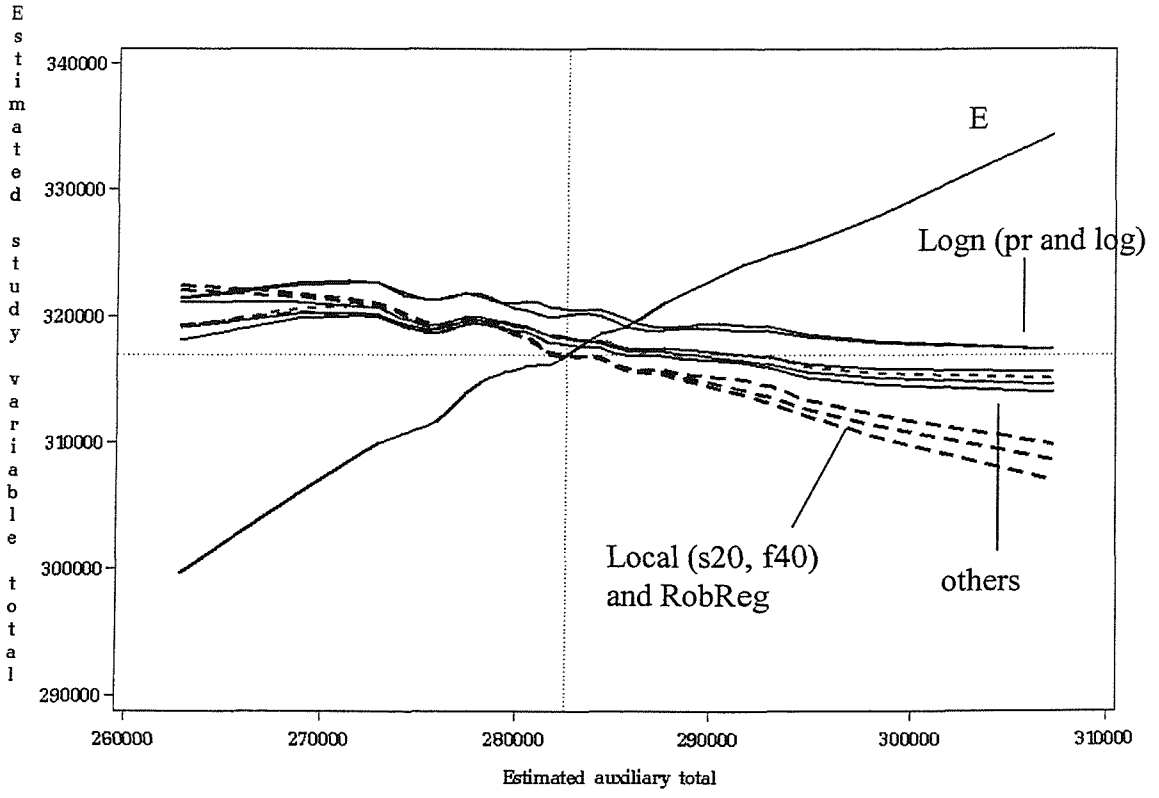


Figure 5.6 a) Domain A, model groups: within strata

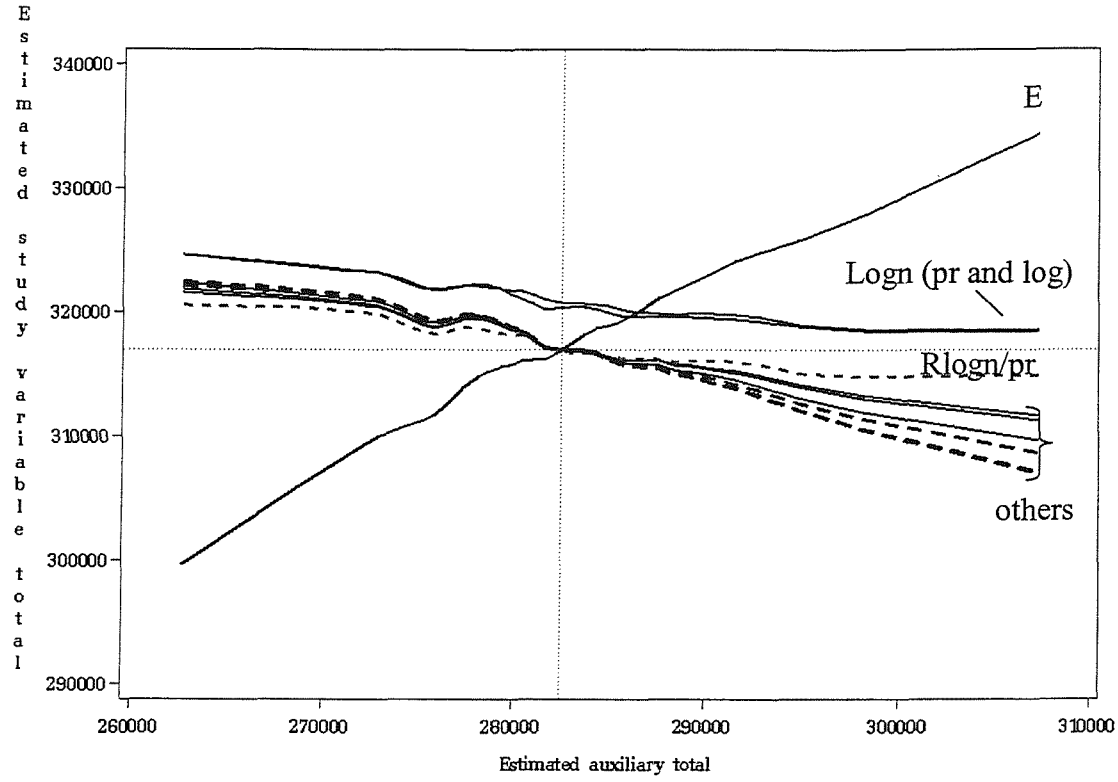
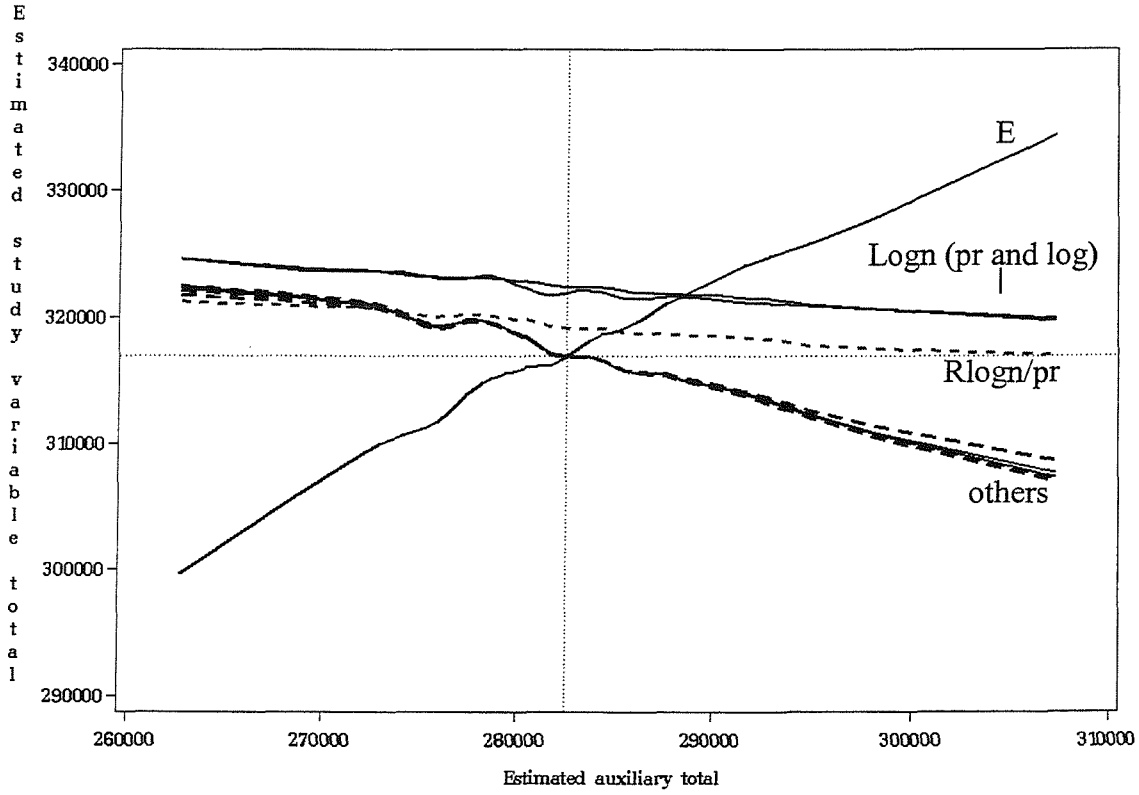


Figure 5.6 b) Domain A, ONS model groups



c) Domain A, model group: over all strata.

Figure 5.6 a-c. The estimated total of the study variable against the estimated total of the auxiliary variable. Loess curves indicate the conditional bias; horizontal and vertical lines indicate the true totals.

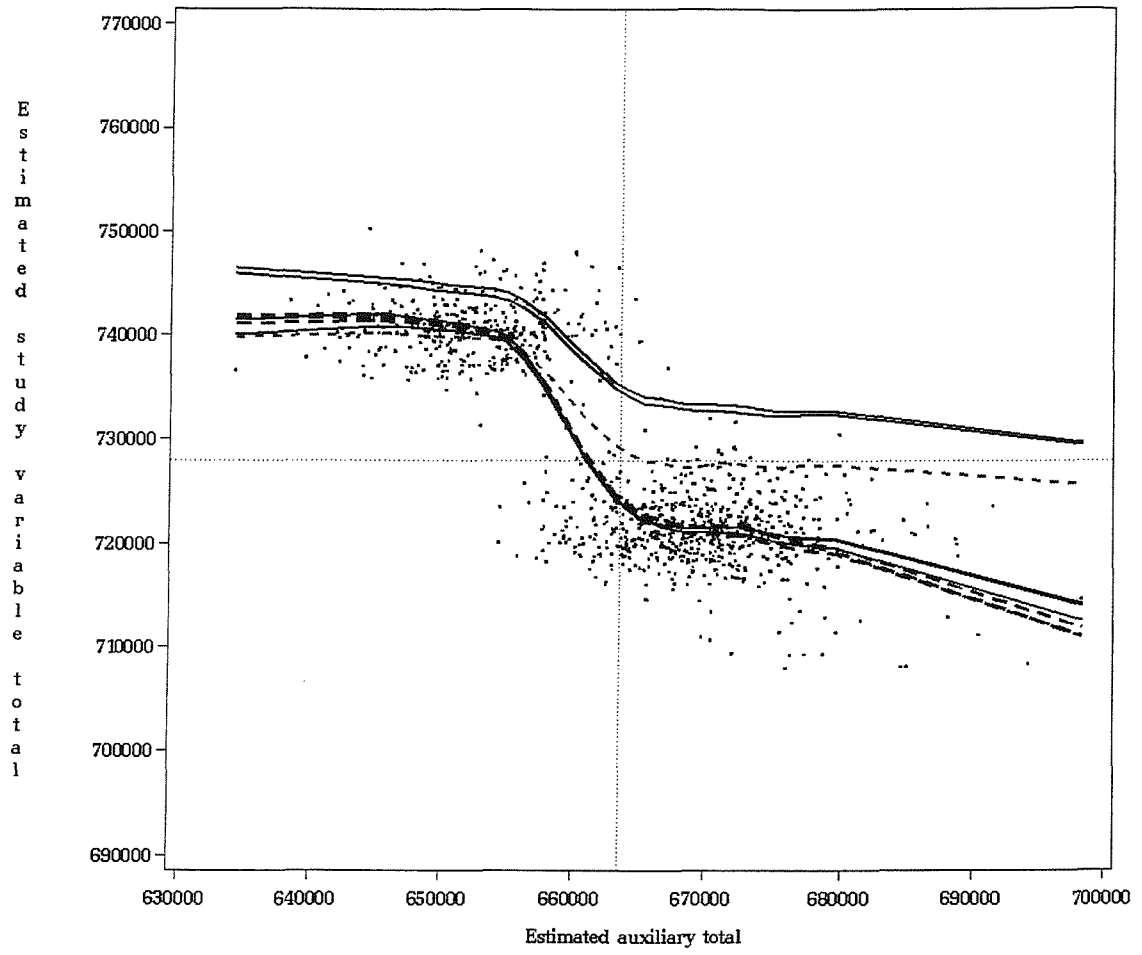


Figure 5.7. Domain B, 'over all strata' model group. The estimated total of the study variable against the estimated total of the auxiliary variable. Loess curves indicate the conditional bias, with one curve per estimator. They are from top to bottom, Log/pr and log (almost indistinguishable), RLogn/pr (dashed), and all other estimators tightly together in one group. The dots represent the outcome for 1000 simulated Reg/1.0 estimates.

In principle, the conditional bias for the other domains, conditional on $\hat{t}_{x\pi}$, showed the same pattern, although there was one conspicuous feature in MIDSS domain B and Capex domain V: the scatter of points are almost entirely separable into two clusters. See Figure 5.7 for domain B. If a sample contains the high-leverage point in sizeband 3 visible in Figure 5.2, the estimate belongs to the cluster towards the lower-right corner in Figure 5.7, otherwise it belongs to the other cluster. Furthermore, as can be deduced from Figure 5.7, the sampling distribution of the estimated study variable total is bimodal. This explains the very poor coverage probabilities in Table 5.8.

5.4 A pre-sample diagnostic

In Chapter 4 the g-weight function was suggested as a diagnostic for a GREG estimator that can be applied after the sample has been drawn. Now the focus is on pre-sample diagnostics. If a 'difficult' frame unit can be identified before that sample is drawn, it can be moved out of the stratum where it first appeared and put in a completely enumerated (CE) stratum and thus obviate the need for 'outlier treatment'. Hence, we need to find some rule that detects all awkward units but ideally no unit that needs no special treatment.

Clearly, a pre-sample diagnostic cannot flag units with frame characteristics that are close to the average in the group they belong to. When such a unit is observed in a sample it may turn out to be an outlier if its value of the study variable or the combination of the study and auxiliary variables is very different from neighbouring units. Once the sample is drawn and observed the estimation process needs to accommodate outliers that cannot be seen prior to the sampling process.

The advantage of detecting outlying units before they are sampled is not only practical; it is also a matter of accuracy. With modern robust estimation techniques outliers may not always be as bad a problem as they once were, but robust methods come with a price in terms of increased variance. Outlier treatment used in the past implicitly removed units from their original stratum and put them in a special poststratum, usually with weight 1.

There are in principle two ways of constructing pre-sample diagnostics: either they are solely based on what is known about the auxiliary variables or on some assumed model for the relationship between the study variable and the auxiliary variables.

My main approach is to look for high-leverage points. One situation when such points will appear with high probability is when a business survey is conducted with one stratification variable but another auxiliary variable is used in estimation. For example, at the ONS, most business surveys use a common stratification variable and common strata boundaries. At estimation stage many of the surveys will use a different auxiliary variable to gain accuracy. In general, with improved registers and the ever increased use of more complex estimation techniques there will often be additional auxiliary

information that can be used in estimation, but it would be impractical or impossible to use all of them at the sampling stage. For example, for a set of business surveys coordinated through a Permanent Random Number system (Ohlsson 1995), if not all or at least the vast majority of surveys do not use the same stratification variable and the same stratum boundaries the system will be hard to implement (Kokic and Brewer 1996, p. 5).

5.4.1 Diagnostics based on the normal distribution

Note that both (5.4) and (5.5) can be written as the expansion estimators plus a term

$$\sum_U [1 - I(k \in s) w_k] \tilde{y}'_k \qquad (5.25)$$

where

$$\tilde{y}'_k = \begin{cases} \mathbf{x}_{kg} \hat{\mathbf{B}}_g & \text{for GREGs} \\ \hat{m}_k & \text{for Local} \end{cases}$$

In the kind of applications we are interested in, $\mathbf{B}_g \approx \begin{pmatrix} 0 & B_{g1} \end{pmatrix}'$, where B_{g1} is some number, and hence (5.25) for GREGs expresses the balance of the sample, $t_{xg} - \hat{t}_{xg\pi}$, times B_{g1}, where t_{xg} is a scalar component of $\mathbf{t}_{xg}$. There is little risk that $t_{xg} - \hat{t}_{xg\pi}$ will turn out large in relation to $V(\hat{t}_{xg\pi})$ if $\mathbf{x}_g$ is such that central limit theorem applies to the estimate $\hat{t}_{xg\pi}$. Therefore, one approach is to base a pre-sample diagnostic on some diagnostic for the goodness of the approximation of the sampling distribution of $\hat{t}_{xg\pi}$ to the normal distribution. Cochran (1977, p. 42) gives a rule of thumb for simple random sampling that relates the smallest sample size necessary to achieve a coverage probability of at least 0.94 to 25 times the square of the skewness coefficient G_1 of the study variable :

$n > 25G_1^2$,

where $G_1 = \dfrac{\sum_U (y_k - t_y/N)^3}{\sigma_y^3}$ and σ_y^3 is the third power of the population standard deviation of the study variable.

Sugden, Smith, and Jones (2000) discuss the precise meaning of this rule and suggest an improved version,

$$n > 28 + 25G_1^2, \qquad (5.26)$$

which adds on 28 as penalty for not knowing the exact variance of $\hat{t}_{xg\pi}$. Their formula (4.4) extends (5.26) to situations with non-negligible finite population correction and positive kurtosis.

For the Local estimator, (5.25) expresses the balance in terms of local regression predictions, $t_{\hat{m}} - \hat{t}_{\hat{m}\pi}$. In view of the robustness of Local, this error should be smaller than $t_{xg} - \hat{t}_{xg\pi}$.

One approach is therefore to approximate G_1 with the skewness of **x** (rather than the unknown **y**) and see if the planned sample size n is indeed larger than the right hand side of (5.26) or the corresponding expression given by the extension proposed by Sugden et al. (2000), which also involves the coefficient of skewness. However, both rules are very conservative. For a business survey population stratum the skewness is often 3 or (far) larger, so the minimum sample would by (5.26) be at least 253, which is a large sample size to take from each stratum or even between the genuine sampling strata in an ONS domain. For example, the lower quartile, median and upper quartile of the skewness coefficient for turnover in 36 MIDSS domains are 5.2, 7.2 and 9.8, respectively. The reason why (5.26) and other similar rules indicate large sample sizes is that they are designed to cover all possible distributions. The conclusion is therefore that this approach is not very useful in a business survey context.

5.4.2 Diagnostics based on measures of influence of observations

Another approach is to base diagnostics on some measure of the influence an observation may have on estimation. Since $\hat{t}_{yreg} = \sum_U \hat{y}_k + \sum_s w_k (y_k - \hat{y}_k)$, with the second term often approximately or exactly equal to zero, GREGs are sensitive to the estimation of $\mathbf{B}_g$ through $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}_g$. Hence a measure of the influence of one or several sample units on $\hat{\mathbf{B}}_g$ indicates also influence on the corresponding GREG. I focus on single-case measures. Again, I will make use of the $DFBETA_i$ (4.16), which for unit i is $\hat{\mathbf{B}} - \hat{\mathbf{B}}_{(i)}$, where $\hat{\mathbf{B}}_{(i)}$ is the estimate obtained with unit i deleted.

An observation whose removal results in a large value of $\hat{\mathbf{B}}_{g'} - \hat{\mathbf{B}}_{g'(i)}$ for a group g' may have a small impact though, if the point is sampled in a stratum with small units. Therefore, the *impact* $I(i)$ of an observation i in group g' can be defined as the approximate relative change:

$$I(i) = \frac{\left(\sum_{U_{g'(i)}} \mathbf{x}_k\right)\left(\hat{\mathbf{B}}_{g'} - \hat{\mathbf{B}}_{g'(i)}\right)}{\sum_{U_{g'(i)}} \mathbf{x}_k \hat{\mathbf{B}}_{g'} + \sum_{g \neq g'} \sum_{U_g} \mathbf{x}_k \hat{\mathbf{B}}_g}, \qquad (5.27)$$

where $U_{g'(i)}$ is population group g' with unit i removed. Alternatively, one could relate $\hat{\mathbf{B}}_{g'} - \hat{\mathbf{B}}_{g'(i)}$ to some knowledge of the variance, for example the variance for some previous period of the survey.

For given x-values, the smallest or largest $DFBETA_i$ we might expect is (4.16) with e_i replaced with $\pm p_a \sigma z_i^{\gamma/2}$, where p_α is a suitable constant. If the distribution of the residuals were normal with zero expectation and variance $\sigma^2 z_i^{\gamma}$, we could choose p_α by looking at a percentile of the standard normal distribution. However, knowing that the tails of the actual distributions are likely to be thicker than those of a normal distribution, I have chosen a large percentile, $\alpha = 99.9$, which makes $p_\alpha = 3.1$. I also put $z_k = x_k$. The maximum and minimum 'feared' $DFBETA_i$, divided by σ, is therefore

$$\mathbf{R}_i = \pm\left(\sum_s \mathbf{x}_k \mathbf{x}_k' / x_k^{\gamma}\right)^{-1}\left(\frac{\mathbf{x}_i}{x_i^{\gamma/2}}\right)\left(\frac{p_\alpha x_i^{\gamma/2}}{1-h_i}\right). \qquad (5.28)$$

As been pointed out a couple of times, γ is often about 1.5 for business survey populations. One approach is therefore to set $\gamma = 1.5$ in (5.28) and list the units with the largest absolute value of the second component of $\mathbf{R}_i$, denoted by R_i. Some of these units will be moved prior to the sampling to either some completely enumerated stratum or to other genuine sampling strata where the units are of a size that matches the units removed. This process can be repeated.

I shall now motivate the definition of *Risk of impact* that will be defined as (5.30) below. Since the impact (5.27) cannot be used as a pre-sample diagnostic, it needs

some modification. In (5.27), $\hat{\mathbf{B}}_{g'} - \hat{\mathbf{B}}_{g'(i)}$ is replaced with $\mathbf{R}_i$, and the $\hat{\mathbf{B}}_g$'s in the denominator are replaced with reasonable numbers based on previous experience. We focus on the second component of $\hat{\mathbf{B}}_{g'} - \hat{\mathbf{B}}_{g'(i)}$ only. Suppose the second components of the $\hat{\mathbf{B}}_g$'s are roughly the same over groups and denote the common value by B_0. Then we have the Risk of impact

$$RI(i) = \frac{R_i\left(\sum_{U_{g'(i)}} x_{gk}\right)}{B_0 \sum_{U_{g'(i)}} x_{gk} + B_0 \sum_{g \neq g'} \sum_{U_g} x_{gk}} \qquad (5.30)$$

Note that (5.30) gives, up to a constant σ, an indication of the relative change in the estimate that can be expected if the observation i is deleted. To put the pre-sample diagnostic (5.30) into practical use, B_0 will have to be assigned a value based on experience or subject-matter considerations. In my experience, this can always be done. As shown in next section, (5.30) can be readily applied as a pre-sample diagnostic.

5.4.3 Application

The diagnostic (5.30) was applied to domain B of the MIDSS survey, plotted in Figure 5.2. The constant B_0 was set to 1 since both the survey variable and the auxiliary variable are measures of turnover, albeit taken from different sources and time periods. Some measures are listed in Table 5.12 for the ten units with the largest risk of impact in each sizeband. There is one unit, the one in sizeband 3, that has a risk of impact that stands out.

I repeated the simulation study to be able to compare sample survey estimates with and without the unit in sizeband 3 with the relatively large risk of impact. Table 5.13 shows the effect of moving this unit to a CE stratum. When it was removed the sample size was reduced with one unit. As can be seen in Table 5.13, all of the measures CV, coverage probability and large error improved dramatically for all estimators except for the HT estimator (E). It is interesting to note that the improvement is as large for the Local/s20 regression estimator as for the others: not even this estimator can cope with the point with extreme leverage.

Table 5.12. Diagnostics for the 10 units in each sizeband 1-3 with largest risk of impact

a) Sizeband 1.

Unit	Leverage	Second coordinate of DFBETA x 100	Second coordinate of (5.28)x 100	Risk of impact
1	0.046	-8.3	45.5	0.02
2	0.042	-75.1	38.4	0.02
3	0.027	-8.3	15.7	0.01
4	0.027	-5.8	15.1	0.01
5	0.023	-3.4	11.1	0.00
6	0.021	3.6	9.0	0.00
7	0.020	-4.0	8.7	0.00
8	0.020	-9.5	8.6	0.00
9	0.020	-0.2	8.6	0.00
10	0.020	-9.8	8.2	0.00

b) Sizeband 2.

Unit	Leverage	Second coordinate of DFBETA x 100	Second coordinate of (5.28)x 100	Risk of impact
1	0.043	11.0	32.6	0.06
2	0.035	-12.9	21.8	0.04
3	0.032	0.2	18.0	0.03
4	0.026	8.7	12.0	0.02
5	0.026	-3.7	11.3	0.02
6	0.025	7.0	10.8	0.02
7	0.024	-5.4	9.6	0.02
8	0.023	-2.0	9.4	0.02
9	0.023	2.2	9.4	0.02
10	0.023	-0.6	9.0	0.02

c). Sizeband 3.

Unit	Leverage	Second coordinate of DFBETA x 100	Second coordinate of (5.28)x 100	Risk of impact
1	0.046	-195.2	55.4	0.42
2	0.021	-0.9	11.7	0.09
3	0.021	-1.6	11.4	0.09
4	0.020	0.5	10.1	0.08
5	0.020	-9.8	9.9	0.08
6	0.018	-3.1	8.2	0.06
7	0.018	1.5	8.1	0.06
8	0.017	-1.2	7.4	0.06
9	0.017	12.5	7.1	0.05
10	0.017	6.6	7.0	0.05

Table 5.13. CV, Coverage Probability (Cov) and Large Error (LE) for MIDSS domain B. The first column for each measure is based on the full sample (Tables 5.6, 5.8 and 5.10), the second on the full sample with the unit in sizeband 3 that has a relatively large risk of impact removed.

	CV		Cov		LE	
	Full	Reduced	Full	Reduced	Full	Reduced
E	0.92	0.89	92.7	92.9	1.5	1.5
Rat_1	1.29	0.58	63.9	88.0	2.3	1
Rat_2	1.29	0.50	65.1	93.8	2.2	0.8
Rat_3	1.34	0.51	64.7	94.2	2.1	0.8
Reg/1.0_1	1.28	0.66	63.7	85.2	2.4	1.1
Reg/1.0_2	1.28	0.50	65.1	93.8	2.1	0.8
Reg/1.0_3	1.34	0.51	64.7	94.2	2.1	0.8
Reg/1.5_1	1.38	0.64	65.3	87.6	2.3	1.1
Reg/1.5_2	1.36	0.51	65.5	93.8	2.2	0.8
Reg/1.5_3	1.41	0.52	64.7	93.8	2.2	0.9
Local/s20	1.36	0.53	64.7	92.4	2.2	0.9

For CAPEX domain V there was again one unit with a higher risk of impact than the others: this is the high-leverage unit in sizeband 3 visible in Figure 5.5. Its risk of impact is 1.55, the second largest in any of sizebands 1-3 is 0.07. I repeated the simulations for domain V without this unit; the CV, coverage probability and the Large Error are reported in Table 5.14. For this domain it also of interest to see the bias is

greatly reduced for all estimators that suffered from bias in the original population, see Table 5.15.

Table 5.14. CV, Coverage Probability (Cov) and Large Error (LE) for CAPEX domain V. The first column for each measure is based on the full sample (Tables 5.6, 5.8 and 5.10), the second on the full sample with the unit in sizeband 3 that has a relatively large risk of impact removed.

	CV		Cov		LE	
	Full	Reduced	Full	Reduced	Full	Reduced
E	6.62	6.32	73.6	83.7	11.5	10.8
Rat_1	15.46	6.25	68.8	86.7	34.8	10.6
Rat_2	15.6	6.15	31.7	85.5	33.7	10.4
Rat_3	24.83	6.21	84.8	87.5	33.5	10.6
Reg/1.0_1	14.01	6.38	86.3	83.6	29.5	11.6
Reg/1.0_2	14.2	6.58	90.2	83.1	17.0	11.5
Reg/1.0_3	7.94	6.22	90.1	84.0	13.7	10.7
Reg/1.5_1	21.74	6.61	85.0	85.5	33.7	12.3
Reg/1.5_2	42.11	7.84	87.4	84.9	89.4	13.1
Reg/1.5_3	19.35	6.57	79.4	84.0	36.0	11.1
Local/f40	6.42	6.18	73.6	84.2	8.8	10.3

Table 5.15. Bias for CAPEX domain V. The first column for each measure is based on the full sample (Table 5.7), the second on the full sample with the unit in sizeband 3 that has a relatively large risk of impact removed.

	Bias	
	Full	Reduced
E	0.07	-1.31
Rat_1	9.91	-0.96
Rat_2	9.20	-1.24
Rat_3	7.41	-1.27
Reg/1.0_1	4.93	-1.8
Reg/1.0_2	-2.80	-1.52
Reg/1.0_3	0.88	-1.27
Reg/1.5_1	1.62	-1.88
Reg/1.5_2	-2.82	-1.34
Reg/1.5_3	0.23	-1.13
Local/f40	-1.90	-1.31

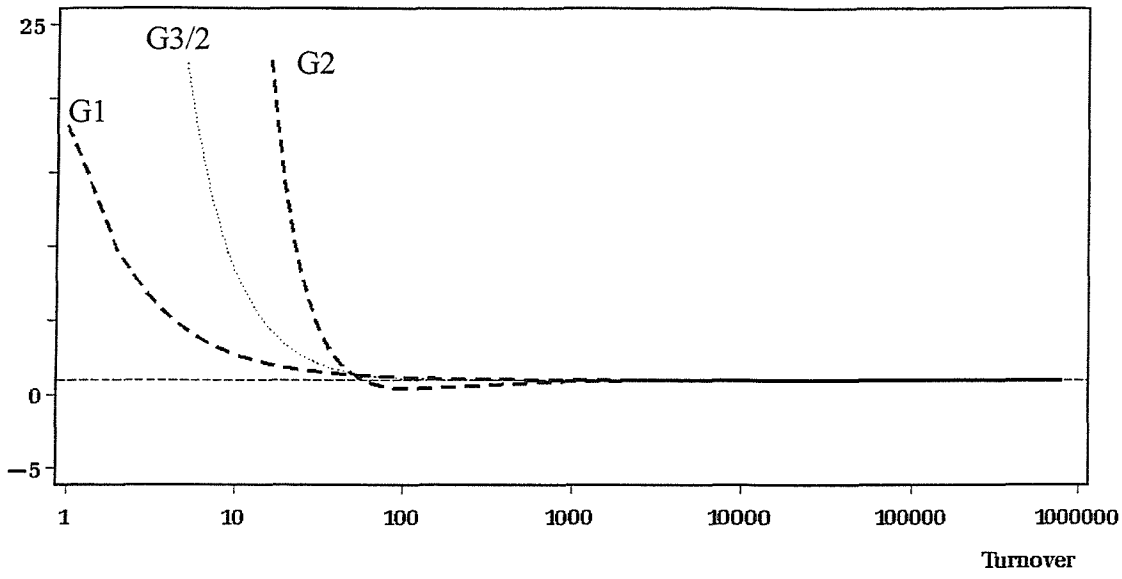


Figure 5.8. The g-weight functions G1, G3/2 and G2 for sizeband 3 in CAPEX domain V as a function of turnover. The horizontal line represents model E.

We now leave the simulation study and return to the g-weight functions studied in Chapter 4. It is interesting to see what effect deleting one large-risk-of-impact point may have on g-weight functions. These functions for sizeband 3 in domain V are shown in Figure 4.4. The unit with the large risk of impact was deleted from the sample, so the original sample of businesses was reduced from 112 units to 111, with the population reduced accordingly. The re-calculated g-weight functions are graphed in Figure 5.8. The shapes are, in principle, the same but the minimums for G3/2 and G2 are larger. This has great practical importance (see Section 4.1.5) since now there are no negative g-weights.

5.5 Discussion

I have conducted a simulation study of estimation in business surveys and contrasted GREGs with a local regression estimator and a robust regression estimator. I evaluated each estimator with three types of model grouping, where relevant, against five criteria. Three of the criteria are conventional (bias, variance and coverage probability), whereas the other two measured aspects of the absolute error: the

proportion of the estimates that were close to the true value and the proportion that where very far from the true value. Instead of using the MSE as one criterion, I have looked at bias and variance separately. In my opinion, an estimator with non-negligible design-bias is not a good estimator, no matter what its variance or MSE is.

Some general conclusions are:

1. There is no estimator that is *the best*. It all depends on the use of the estimates and on the population. Different criteria will be more important for different uses. The users, however, should not decide what estimators are being used. The users may change but the national statistical institute cannot afford to flit (cf Holt's 1998, pp. 17-19, discussion on measuring inflation).
2. The estimators that have the best unconditional properties across all populations are the expansion estimator, Reg/1.0 fitted across all strata and the Local regression estimators. In particular, there seems to be no reason to prefer the ratio estimator to Reg/1.0.
3. The standard way of constructing confidence intervals (1.96 times the standard error, estimated with formulas such as those in Sec 5.2.4) often gives poor coverage. If the main aim is good confidence intervals then the expansion estimator is preferable, although the price to pay will be wide intervals.
4. For design-based estimators, fitting models within strata (leading to estimators such as the separate ratio or regression estimator) tends to give small CVs, but fitting models across strata tends to make estimates more robust.

Other conclusions that concern specific estimators are:

i. The choice of nearest neighbour bandwidth for local regression estimators does not seem overly sensitive.
ii. The robust regression estimator and one of the local regression estimators are superior if the aim is to minimise the proportion of estimates that are very far from the true value in absolute terms. This is particularly important in official statistics. These estimators have reasonably small conditional bias, although GREGs fitted within strata have smaller conditional bias.
iii. The model-based mixture model estimator is bias prone and will give poor estimates for some populations. Robust estimation of the variance parameter seems to be an approach that reduces some of the problems. Robust estimation of the

slope parameter on the logarithmic scale is an option which I leave for future research. Also, if the model is fitted across strata, including the completely enumerated stratum, the parameter estimation tends to be more reliable. However, like the robust regression estimators, this estimator is not linear, nor has it the internal consistency property.

iv. The regression estimator that is associated with the best model (with variance about the regression line proportional to the auxiliary raised to 1.5) is more erratic than the regression estimator modelled on a variance proportional to the auxiliary. The reason was seen to be variance in the bias adjustment term, i.e., the second term in (5.1) and (5.4). This term is non-zero only for the former estimator.

High leverage points need to be addressed. In Chapter 4 I suggest post-stratification where the expansion estimator is applied to the 20 or so units with the largest values of the auxiliary and that e.g. some GREG estimator that does use auxiliary information is applied to post-strata without high leverage points. Here I have proposed a pre-sample diagnostic. For units with particularly large value of the diagnostic action can be taken before the sample is drawn. They can, for example be moved to a CE stratum, or, if misclassification the reason why they appear in a stratum where the auxiliary variable values are in general smaller, they can be re-classified and moved to the stratum where they rightly belong. As the discussion above shows, this can be done in a fully automated way. There will be no bias incurred by the re-classification. Also, if the allocation is based on the variance of the auxiliary variable, moving inordinately large units before the allocation is done will give more reasonable sample sizes.

The units that will obtain a large value of the diagnostic tend to have large leverage and could readily be spotted with a simple graph. However, a numerical measure to guide actions is helpful since the interpretation of a graph relies on subjective decisions and hence to some extent on the experience of the staff making the decisions. Furthermore, at many national statistical institutes there is an ever-increasing pressure to make the survey process more efficient and a numerical pre-sample diagnostic is a key ingredient in a semi-automatic screening of the population. Note that the transparency of this diagnostic is vital: in my experience it is very hard to introduce methodology at an NSI if staff do not feel that the methods proposed in their view carry a natural interpretation.

Chapter 6

Concluding Discussion

In this thesis I have addressed some issues that are of concern in official business statistics. Each of Chapters 2 – 4 ends with a discussion, which will not be repeated here. In this final chapter I shall endeavour to discuss if and in what way any of the results of the research presented here can become part of every-day processes at NSIs. This thesis is about 'real-world' problem solving and it would be disappointing if the main findings were 'interesting' rather than useful.

We have seen that reporting delays of the registration of new businesses can cause a non-negligible bias. How large the bias is depends crucially on the frame maintenance processes at the NSI. The size of this bias is probably unknown at most NSIs, and – given the lack of literature in this area – even the awareness of the consequences of reporting delays may be absent. I have highlighted the existence of some research in other areas that can be used to address the problem caused by reporting delays. The prediction methodology proposed in Chapter 2 is practical and fits reasonably well into existing structures at most NSIs. To use this to predict and monitor the reporting-delay induced undercount should not pose any great practical problems. To predict the negative bias of estimates of the population total caused by reporting delays proved difficult due to very long reporting delays for large businesses. It is not clear at this stage whether the long reporting delays are peculiar to frame maintenance processes at the ONS or if other NSIs that update their business registers from more than one administrative source face the problem of long delays due lengthy cross-checking and proving processes. For the ONS, the solution to the problem of estimating the bias requires new administrative procedures that allow separation of genuinely new businesses from those that are new only in legal terms. With these procedures in place, refitting the models may indicate whether the methods in Chapter 2 are adequate or if there is a need for further research into modelling extreme events (that is, a few large businesses having long reporting delays).

Feeding back information on ineligibility obtained in sample surveys to the frame is generally discouraged. However, as there is no quantification of the feed back bias published in a refereed journal, my theoretical expressions for the bias accompanied with small-scale simulations does constitute a contribution to our shared knowledge of survey sampling methodology. Further research of feed back bias may involve unequal probability sampling designs and assessment of the bias in situations where the eligibility of sampled units may be misclassified.

We have seen that there are problems with GREG estimation, and that the model-dependence of model assisted estimators cannot be ignored. This point may well be one of the most important results of the thesis. The g-weights continue to play a rather mysterious role in model assisted theory. They are clearly useful, but also potentially problematic since they are not bounded to some finite interval in the set of positive real values. Like GREG estimation, calibration has become very popular during the last decade, both as an idea and as part of methodologists' every-day work. Again, this technique is useful but the very strict constraint it imposes on estimates is problematic. There are research results on 'relaxed' calibration constraints, in the sense that some discrepancy between estimates of auxiliary variable parameters and their population counterpart is allowed (Bardsley and Chambers 1984). This idea has not yet been explored in a model assisted context.

I have suggested some diagnostics that can help the practitioner to pin down those of the mass of estimates produced by an NSI that are likely to be very inaccurate. In my experience, *diagnostics* provide a way of thinking that is not often made use of in design-based survey sampling. This is puzzling considering the rather prominent role diagnostics have in other areas of statistics. In my opinion, the pre-sample diagnostic is the one of the diagnostics I have suggested that has the by far best chances of ever being used in practice. This is because it fits readily into common official statistics processes and thinking. The idea of g-weight functions may attract interest as something that sheds light on the nature of g-weights, but it is not likely to become part of the methodologist's toolkit. This is because regression estimation seems to continue to play the second fiddle to ratio estimation (which has constant g-weights) and because the g-weight functions require manual intervention to be assessed. Poststratification, as it is proposed at the end of Chapter 4, also requires manual

intervention. There is also the additional problem of the difficulty of assessing the contribution of the intervention to the overall uncertainty. How do we account for having chosen to poststratify after we have seen the data?

Finally, I have explored some GREGs and other estimators that can be used to estimate a population total. The result was – as expected – rather inconclusive, although the design-based local linear regression estimators came out well. I have also raised the question what we want to achieve when we choose estimators; what are the most important properties of an estimator? I believe that the traditional approach to focus on design-bias and design-variance only will give way for a wider outlook where properties conditional on the realised sample will attract more interest among practitioners at NSIs than has been the case in the past. I believe also that practitioners will include some minimax type of property in their range of quality components when assessing strategies (but not necessarily the Large Error property discussed in Chapter 5).

Although providing some new insights, Chapters 4 and 5 have the limitation of dealing with only one variable at a time. Official statistics involve a large system of many variables and many surveys. How to make use of the information different surveys can provide to each other may be one of the more important future research areas in official statistics.

Appendix 1. Selective editing

The act of checking and correcting respondent data in surveys is usually referred to as *editing*. Many national statistical institutes use a 'micro-editing' approach for business surveys. Micro-editing (also called input editing) focuses on the individual record or questionnaire, as opposed to macro-editing where checks are used on aggregated data. In micro-editing, the respondent data are passed through *edits* (edit rules, checks) that typically aim at detecting unusual item responses. For example, many repetitive business surveys use ratio edits whereby a response from a business is compared to its prior response. If the relative movement is more than, say, a% or less than b%, the incoming datum point fails the edit, and the questionnaire will be inspected manually. The business may be called back. As business data are volatile, a sizeable proportion of the respondents will confirm reported large movements. Thus, micro-editing will lead to many false signals, unless the edits have been carefully designed (Thompson and Sigman, 1999).

Since editing is one of the most time consuming processes in the production of official statistics (Granquist and Kovar, 1997, p. 418), more efficient methods have been discussed and implemented for many surveys. With selective editing the incoming units are prioritised, and those that have been given priority are selected for editing. The prioritisation step often involves the computation of a score for each datum point that reflects the importance of investigating this datum point. Some types of score can be computed for all data, others only for those that have failed at least one edit. A score may be computed for each item on the questionnaire, and the item scores may be combined to a unit score. Questionnaires with unit scores above a predetermined threshold are inspected manually; other units are left unattended or passed on to another process, for example, some other type of editing or automatic imputation. Alternatively, there may be more than two levels of priority, with a high threshold defining the highest priority, and so on. The set of scores for one item can be viewed as the range of a *score function* of the unedited data and the background data, such as past edited data.

Hedlin (2002a, 2003) defines two types of score function: one estimate-related method that prioritises by the predicted impact a suspect value will have on particular estimates, and one edit-related method that is explicitly based on specified edits and prioritises suspect values by the Mahalanobis distance (e.g. Krzanowski 1990) of the magnitude of edit failures. Hedlin (2003) also discusses how to measure the effectiveness of a score

function, and what it is that makes this method effective at reducing manual editing. It is interesting that both methods proved effective, although they are based on very different rationales. Somewhat surprisingly, the former method was seen to work rather better for a wide range of different estimates, apart from the estimates it targeted.

In many situations it is useful to specify a threshold that splits the responses into two groups where manual editing is directed at responses with scores above the threshold. Hedlin (2003) suggests a simple graphical tool that allows the analyst to assess the threshold. The paper also suggests some graphs that will help to understand an editing process.

One of the two types of score function has been implemented for the Monthly Inquiry for the Distribution and Services Sector (MIDSS) conducted by the ONS and been subjected to pilot studies for several other business surveys at the ONS (Underwood 2001). About 50% of the editing effort could be spared for the MIDSS.

Appendix 2. Equivalence of Strategies 2 and 3

Note that $\sum_{s_{current}} y_k = \sum_{s_{orig}} y_k$. Therefore, if $n_a/n = 0$ and

$$\left[1 - \frac{\hat{N}_d}{\hat{N}_l}\frac{N_p - N_{sd}}{N_{orig} - N_p}\right]^{-1} = \frac{N_{orig}}{N_{current}}, \qquad (A2.1)$$

Strategies 2 and 3 are the same. We will find for what estimate $\hat{N}_l$ (A2.1) holds.

Since $N_{current} = \hat{N}_d + \hat{N}_l$ by assumption, the left-hand side of (A2.1) is

$$\frac{\hat{N}_l\left(N_{orig} - N_p\right)}{\hat{N}_l\left(N_{orig} - N_p\right) - \left(N_{current} - \hat{N}_l\right)\left(N_p - N_{sd}\right)}$$

$$= \frac{\hat{N}_l\left(N_{orig} - N_p\right)}{\hat{N}_l N_{orig} - N_p N_{current} + N_{sd}\left(N_{current} - \hat{N}_l\right)}$$

$$= \frac{1}{N_{current}}\frac{\hat{N}_l\left(N_{orig} - N_p\right)}{\hat{N}_l - \left(N_p - N_{sd}\right)}. \qquad (A2.2)$$

Define

$$D = \frac{\hat{N}_l\left(N_{orig} - N_p\right)}{\hat{N}_l - \left(N_p - N_{sd}\right)}. \qquad (A2.3)$$

If $D = N_{orig}$, the right-hand side of (A2.2) equals $N_{orig}/N_{current}$ and Strategies 2 and 3 are the same.

From (A2.3),

$$\hat{N}_l = \frac{D\left(N_p - N_{sd}\right)}{\left(D - N_{orig} + N_p\right)}.$$

Hence $D = N_{orig}$ and $\hat{N}_l = N_{orig}\dfrac{N_p - N_{sd}}{N_p}$ are equivalent statements.

Appendix 3. A variance estimator for the poststratified estimator of the total

Let subscript g refer to poststratum g. The sample residual for a unit k in poststratum g is

$e_k = y_k - \mathbf{x}_k \hat{\mathbf{B}}_g$ with $\hat{\mathbf{B}}_g = \tilde{y}_g = \sum_{s_g} w_k y_k \Big/ \sum_{s_g} w_k$ for the sample part $s_g = U_g \cap s$.

For a simple random sample without replacement (SI),
$w_k = N/n$,
$e_k = y_k - \tilde{y}_g$
$\quad = y_k - \bar{y}_g$,
where $\bar{y}_g$ is the straight mean in s_g.

Recall that $\breve{\Delta}_{kl} = (\pi_{kl} - \pi_k \pi_l)/\pi_{kl}$ and $\pi_{kl} = \Pr(I_k = I_l = 1)$. So for an SI

$$\pi_{kl} - \pi_k \pi_l = -\frac{n(N-n)}{N^2(N-1)},\ k \neq l, \text{ and}$$

$$\frac{\breve{\Delta}_{kl}}{\pi_k \pi_l} = \frac{n(n-N)}{N^2(N-1)} \frac{N(N-1)}{n(n-1)} \left(\frac{N}{n}\right)^2 = -\frac{(N-n)N}{(n-1)n^2},\ k \neq l.$$

Using the Yates-Grundy-Sen variance estimator, we have

$$\hat{V}'(\hat{t}_{yreg}) = \sum\sum_s \breve{\Delta}_{kl}\, e_k e_l / \pi_k \pi_l$$

$$= -\frac{1}{2} \sum\sum_s \breve{\Delta}_{kl} (e_k/\pi_k - e_l/\pi_l)^2.$$

Hence for an SI,

$$\hat{V}'_{SI}(\hat{t}_{ypost}) = \frac{1}{2} \frac{(N-n)N}{(n-1)n^2} \sum\sum_s [e_k - e_l]^2$$

$$= \frac{(N-n)N}{(n-1)n^2} \sum_s n e_k^2$$

$$= \frac{(N-n)N}{n} \sum_{g=1}^{G} \frac{n_g - 1}{n-1} s_g^2$$

where $s_g^2 = \dfrac{\sum_{s_g} (y_k - \bar{y}_g)^2}{n_g - 1}$ and n_g is the number of units in s_g.

Appendix 4. G-weight functions

Let $w_k = c\sigma_k^2$, $a_k' = (w_k \pi_k)^{-1}$ and $\mathbf{x}_k = (1 \quad x_k)'$. With $\sum_s v_k$ we will refer to the sum of the quantities v_k, $i = 1, 2, \ldots, k, \ldots, |s|$, with $|s|$ being the number of elements in s and k the summation index. With the summation sign operating on a series of matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k, \ldots \mathbf{A}_{|s|}$, we refer to $\sum_s \mathbf{A}_k = \{\sum_s A_{ijk}\}_{ij}$, i.e., the sums of the matrix elements A_{ijk}, $k = 1, 2, \ldots, |s|$.

The g-weights are

$$g_{is} = 1 + (N - \hat{N} \quad t_x - \hat{t}_{x\pi})\hat{\mathbf{T}}^{-1}\begin{pmatrix} 1 \\ x_i \end{pmatrix}\frac{1}{w_i},$$

where

$$\hat{\mathbf{T}}^{-1} = \left(\sum_s a_k' \mathbf{x}_k \mathbf{x}_k'\right)^{-1} = \frac{1}{\left(\sum_s a_k' x_k^2\right)\left(\sum_s a_k'\right) - \left(\sum_s a_k' x_k\right)^2}\begin{pmatrix} * & * \\ -\sum_s a_k' x_k & \sum_s a_k' \end{pmatrix}$$

with * representing some numbers. For designs for which $N \equiv \hat{N}$ (e.g. simple random sampling) we have

$$g_{is} = 1 + (t_x - \hat{t}_{x\pi})\frac{x_i \sum_s a_k' - \sum_s a_k' x_k}{w_i\left(\sum_s a_k'\right)\left[\left(\sum_s a_k' x_k^2\right) - \left(\sum_s a_k' x_k\right)^2\left(\sum_s a_k'\right)^{-1}\right]}.$$

Let $\tilde{x}_s = \left(\sum_s a_k' x_k\right)\left(\sum_s a_k'\right)^{-1}$ and note that

$$\sum_s a_k'(x_k - \tilde{x}_s)^2 = \sum_s a_k' x_k^2 - 2\tilde{x}_s \sum_s a_k' x_k + \tilde{x}_s^2 \sum_s a_k' = \sum_s a_k' x_k^2 - \tilde{x}_s^2 \sum_s a_k'.$$

Then

$$g_{is} = 1 + (t_x - \hat{t}_{x\pi})\frac{x_i - \tilde{x}_s}{w_i \sum_s a_k'(x_k - \tilde{x}_s)^2}.$$

For simple random sampling, which is assumed in what follows, we have $a_k' = N(nw_k)^{-1}$.

Let $\tilde{x}_s^{(\gamma)} = \left(\sum_s x_k / w_k\right)\left(\sum_s 1/w_k\right)^{-1}$ with $w_k = cx_k^\gamma$.

For **Model C** we have $\gamma = 0$ and $w_k = c$ and

$$g_{is} = 1 + (t_x - \hat{t}_{x\pi})\frac{n(x_i - \tilde{x}_s^{(0)})}{N\sum_s (x_k - \tilde{x}_s^{(0)})^2} = 1 + A_1 - A_2 x_i,$$

where $A_2 = \dfrac{n(\hat{t}_{x\pi} - t_x)}{N\sum_s (x_k - \tilde{x}_s^{(0)})^2}$ and $A_1 = A_2 \tilde{x}_s^{(0)}$. Note that $\tilde{x}_s^{(0)} = \bar{x}_s$ for Model C.

For **Model X1** we have $\gamma = 1$ and $w_k = cx_k$ and

$$g_{is} = 1 + \left(t_x - \hat{t}_{x\pi}\right)\frac{n\left(1 - \tilde{x}_s^{(1)} x_i^{-1}\right)}{N\sum_s x_k^{-1}\left(x_k - \tilde{x}_s^{(1)}\right)^2} = 1 - B_1 + B_2 x_i^{-1},$$

with

$$B_1 = \frac{n\left(\hat{t}_{x\pi} - t_x\right)}{N\sum_s x_k^{-1}\left(x_k - \tilde{x}_s^{(1)}\right)^2}$$

and $B_2 = B_1 \tilde{x}_s^{(1)}$.

The function $f(x) = 1 - B_1 + B_2 x^{-1}$ does not have a minimum value but its infimum is $1 - B_1$.

Repeating this for **Model X3/2**, we obtain $w_k = cx_k^{1.5}$ and

$$g_{is} = 1 + \left(t_x - \hat{t}_{x\pi}\right)\frac{n\left(x_i^{-0.5} - \tilde{x}_s^{(1.5)} x_i^{-1.5}\right)}{N\sum_s x_k^{-1.5}\left(x_k - \tilde{x}_s^{(1.5)}\right)^2} = 1 - C_1 x_i^{-0.5} + C_2 x_i^{-1.5},$$

$$C_1 = \frac{n\left(\hat{t}_{x\pi} - t_x\right)}{N\sum_s x_k^{-1.5}\left(x_k - \tilde{x}_s^{(1.5)}\right)^2}$$

and $C_2 = C_1 \tilde{x}_s^{(1.5)}$.

The derivative of the function $f(x) = 1 - C_1 x^{-0.5} + C_2 x^{-1.5}$ is $0.5C_1 x^{-1.5}\left(1 - 3\tilde{x}_s^{(1.5)} x^{-1}\right)$ which takes zero value at $x = 3\tilde{x}_s^{(1.5)}$. The extreme value at this point (minimum if $C_1 > 0$) is $f\left(3\tilde{x}_s^{(1.5)}\right) = 1 - \frac{2C_1}{3\sqrt{3\tilde{x}_s^{(1.5)}}}$.

Finally, for **Model X2** we have $\gamma = 2$ and $w_k = cx_k^2$ and

$$g_{is} = 1 + \left(t_x - \hat{t}_{x\pi}\right)\frac{x_i - \tilde{x}_s}{w_i \sum_s a_k' \left(x_k - \tilde{x}_s\right)^2}.$$

$$g_{is} = 1 + \left(t_x - \hat{t}_{x\pi}\right)\frac{n\left(x_i^{-1} - \tilde{x}_s^{(2)} x_i^{-2}\right)}{N\sum_s x_k^{-2}\left(x_k - \tilde{x}_s^{(2)}\right)^2} = 1 - D_1 x_i^{-1} + D_2 x_i^{-2},$$

$$D_1 = \frac{n\left(\hat{t}_{x\pi} - t_x\right)}{N\sum_s x_k^{-2}\left(x_k - \tilde{x}_s^{(2)}\right)^2}$$

and $D_2 = D_1 \tilde{x}_s^{(2)}$.

The derivative of the function $f(x) = 1 - D_1 x^{-1} + D_2 x^{-2}$ is $D_1 x^{-2}\left(1 - 2\tilde{x}_s^{(2)} x^{-1}\right)$ which takes zero value at $x = 2\tilde{x}_s^{(2)}$. The extreme value at this point (minimum if $D_1 > 0$) is

$$f\left(2\tilde{x}_s^{(2)}\right) = 1 - \frac{D_1}{4\tilde{x}_s^{(2)}}$$

Appendix 5. Derivation of ratio and regression estimators for the group ratio model

I will use subscript k unit and g for group. The group ratio model is defined as

$$E_\xi(y_k) = \beta_g x_k$$
$$V_\xi(y_k) = \sigma_g^2 x_k, \; g = 1, 2, \ldots, G, \qquad (A5.1)$$

and the group regression model as

$$E_\xi(y_k) = \alpha_g + \beta_g x_k$$
$$V_\xi(y_k) = \sigma_g^2 x_k^\gamma, \; g = 1, 2, \ldots, G, \qquad (A5.2)$$

where I confine choice of γ to 1 or 1.5.

The design is fixed size stratified simple random sampling with independent samples taken from the strata (STSI).

Let the Horvitz-Thompson estimator for the total of y for group g be

$$\hat{t}_{yg\pi} = \sum_{h=1}^{H} \sum_{sgh} w_h y_k \qquad (A5.3)$$

where s_{gh} is the sample in the intersection of stratum h and group g and $w_h = N_h n_h^{-1}$.

Särndal's et al. (1992) general expression for an approximate variance of a regression estimator for a design with a fixed sample size is

$$AV_{STSI}(\hat{t}_{reg}) = \sum\sum_U w_h^2 \Delta_{kl} E_k E_l = -\frac{1}{2}\sum\sum_U w_h^2 \Delta_{kl}(E_k - E_l)^2, \qquad (A5.4)$$

where, for distinct units k and l no matter what group,

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$$
$$-w_h^2 \Delta_{kl} = \begin{cases} N_h\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{1}{N_h - 1} & \text{if } k \neq l \in \text{stratum } h \\ 0 \;\; \text{if } k \text{ and } l \text{ belong to different strata} \end{cases} \qquad (A5.5)$$

and $E_k = y_k - x_k' \mathbf{B}_g$ and,

$\mathbf{B}_g = B_{ratg} = \frac{t_{yg}}{t_{xg}}$ for the group ratio model.

For the group regression model,
$\mathbf{B}_g = \mathbf{B}_{regg} = (\alpha_g, \beta_g)$ with α_g and β_g defined by the weighted least squares fit of the regression of $\mathbf{y}$ on $\mathbf{x}$ in group g.

From (A5.4) and (A5.5),

$$AV_{STSI}\left(\hat{t}_{reg}\right)=$$

$$\frac{1}{2}\sum_{h=1}^{H}\left[N_h\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{N_h-1}\sum\sum_{U_h}\left[\left(E_k-\bar{E}_h\right)-\left(E_l-\bar{E}_h\right)\right]^2\right]$$

$$=\sum_{h=1}^{H}\left[N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{N_h-1}\sum_{U_h}\left(E_k-\bar{E}_h\right)^2\right] \qquad (A5.6)$$

where

$$\bar{E}_h=\frac{1}{N_h}\sum_{U_h}E_k$$

is not necessarily zero.

The variance estimator with g-weights,

$$\hat{V}\left(\hat{t}_{reg}\right)=\sum\sum_{s}\breve{\Delta}_{kl}\,g_{ks}e_k g_{ls}e_l/\pi_k\pi_l\,, \qquad (A5.7)$$

where $e_k=y_k-x_k'\hat{\mathbf{B}}_g$ and $\breve{\Delta}_{kl}=\Delta_{kl}/\pi_{kl}$. For STSI $\pi_{kl}=\dfrac{n_h(n_h-1)}{N_h(N_h-1)}$ for k and l in stratum h, and for the group ratio model $g_{ks}=t_{xg}/\hat{t}_{xg}$. Comparing (A5.4) with (A5.7) we get from (A5.6)

$$\hat{V}_{STSI}\left(\hat{t}_{rat}\right)=\sum_{h=1}^{H}\left(\frac{t_{xg}}{\hat{t}_{xg}}\right)^2\left[N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\sum_{s_h}\left(e_k-\bar{e}_h\right)^2\right] \qquad (A5.8)$$

where $e_k=y_k-x_k'\hat{B}_{ratg}$, $\bar{e}_h=\dfrac{1}{n_h}\sum_{s_h}e_k$, and $\hat{B}_{ratg}=\dfrac{\hat{t}_{yg\pi}}{\hat{t}_{xg\pi}}$.

For the group regression model,

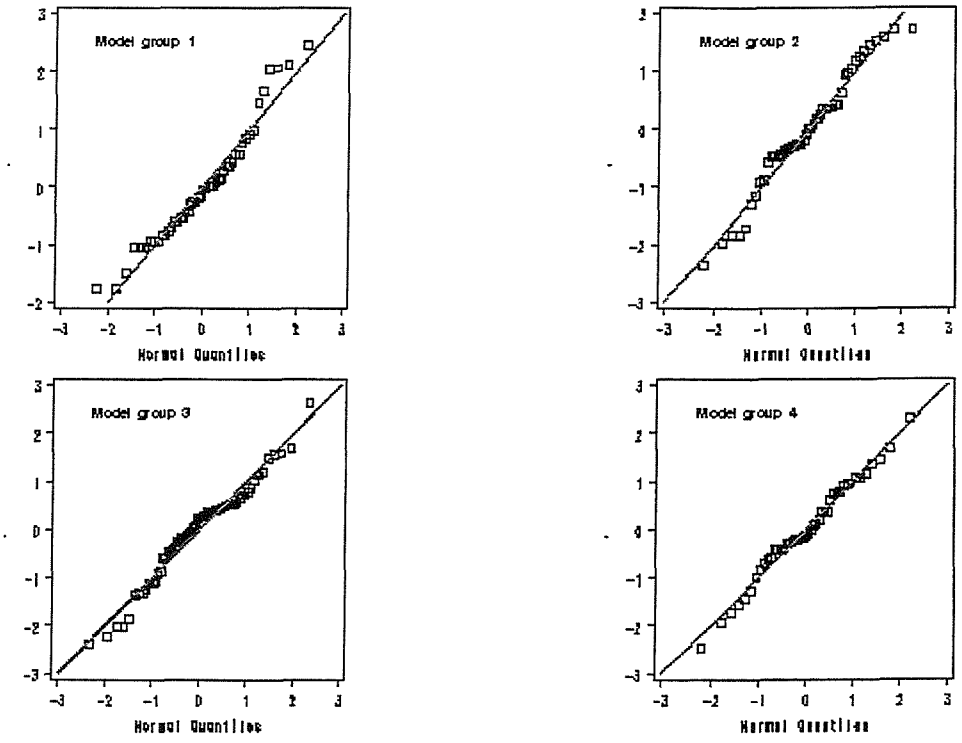
$$\hat{V}_{STSI}\left(\hat{t}_{reg}\right)=\sum_{h=1}^{H}\left[N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\sum_{s_h}g_{ks}^2\left(e_k-\bar{e}_h\right)^2\right] \qquad (A5.9)$$

where $e_k=y_k-\mathbf{x}_k'\hat{\mathbf{B}}_{regg}$ and

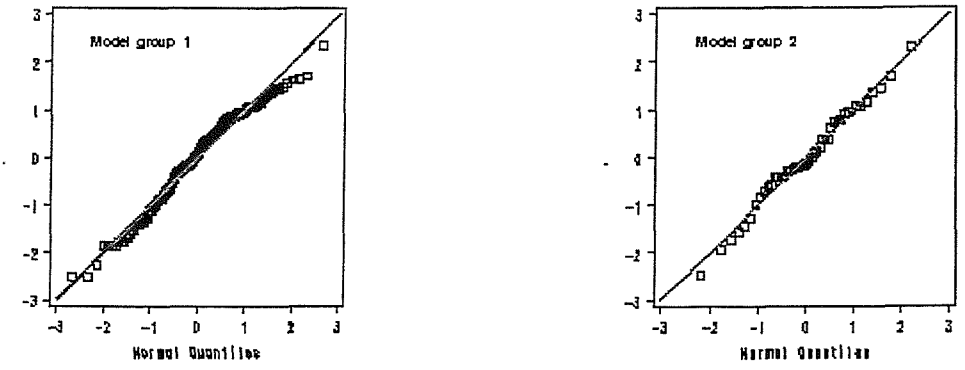
$$g_{ks}=1+\left(\mathbf{t}_{xg}-\hat{\mathbf{t}}_{xg}\right)'\left(\sum_s w_j\mathbf{x}_{jg}\,\mathbf{x}_{jg}'/\sigma_j^2\right)^{-1}\left(\mathbf{x}_{kg}/\sigma_k^2\right). \qquad (A5.10)$$

Appendix 6. Log-normality tests

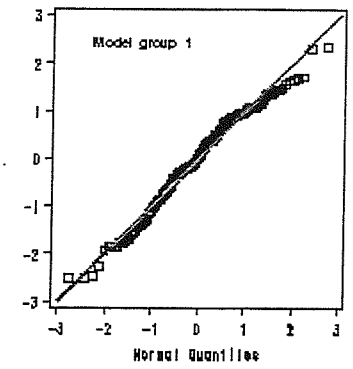
In this section it is investigated whether lognormal models can be fitted to the data in the MIDSS domains A to C. The QQ-plots in Figures A6.1–A6.3 suggest that a logarithmic transformation of both the study and auxiliary variable will help to restore linearity and reduce the residuals. Taking logs can be seen as a special case of Box-Cox transformations (Box and Cox 1964). The transformation in this family that reduced the residuals the most was actually found to be the logarithmic transformation. See Sen and Srivastava (1990, p. 204-208) for the objective function that was minimised to find the optimal transformation.



a) Model groups coincide with strata.

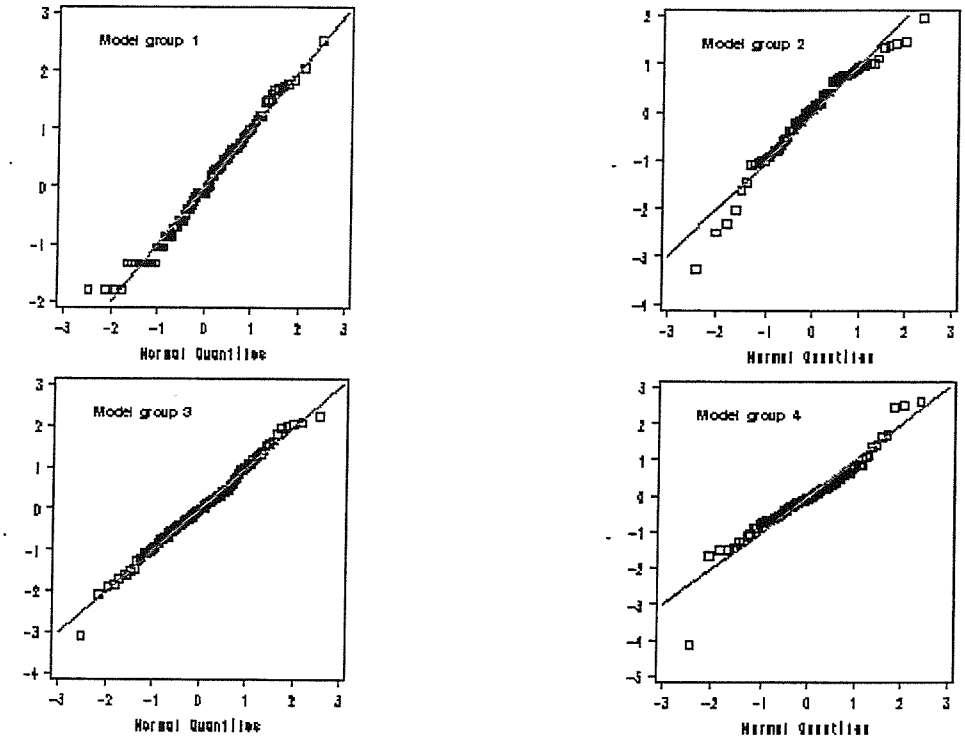


b) Domain divided into two model groups; 'ONS model groups'.

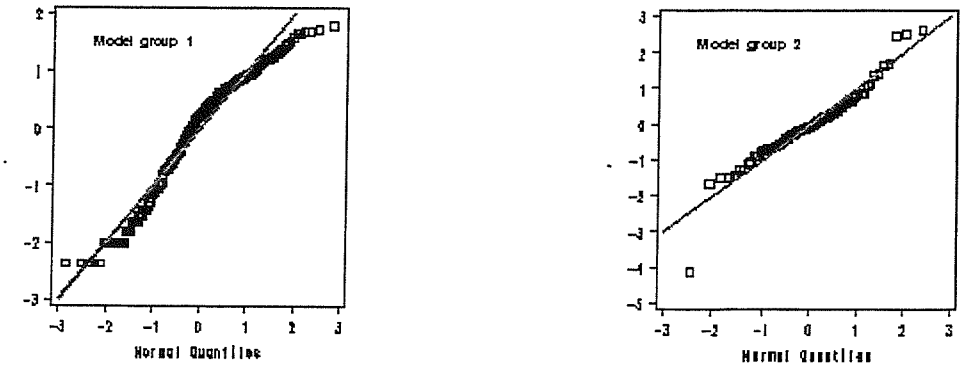


c) The whole domain constitutes one model group.

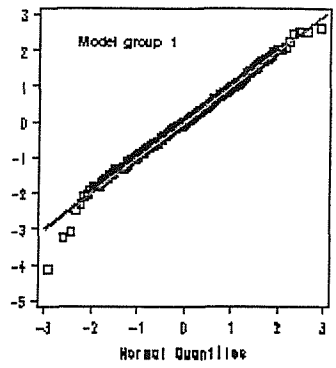
Figure A6.1 a-c. MIDSS, domain A. QQ-plots of the standardised logarithms of the study variable against theoretical quantiles from the standard normal distribution.



a) Model groups coincide with strata.

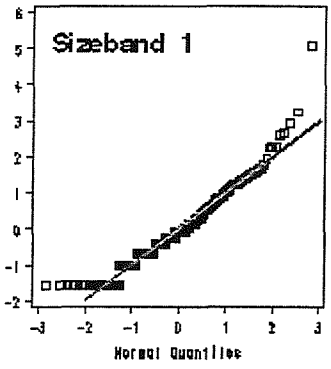


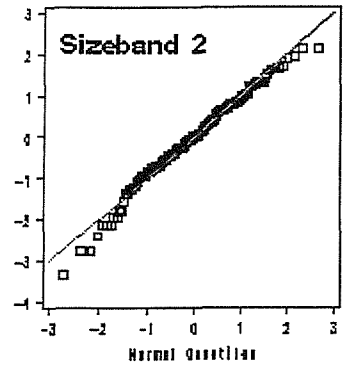
b) Domain divided into two model groups; 'ONS model groups'.

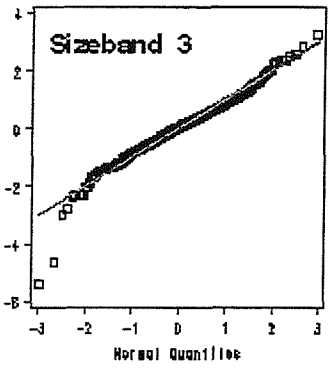


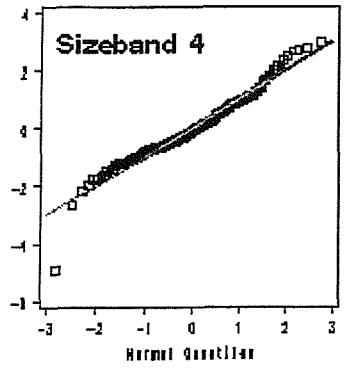
c) The whole domain constitutes one model group.

Figure A6.2 a-c. MIDSS, domain B. QQ-plots of the standardised logarithms of the study variable against theoretical quantiles from the standard normal distribution.

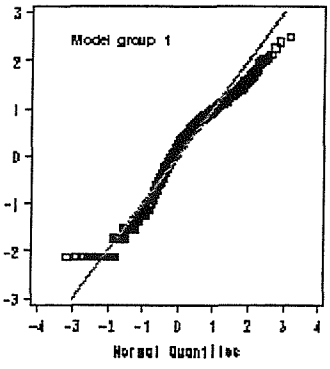


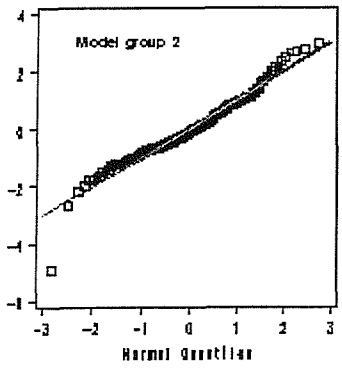




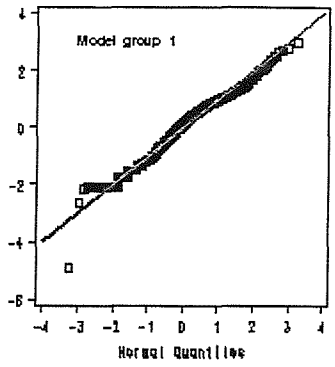


a) Model groups coincide with strata.





b) Domain divided into two model groups; 'ONS model groups'.



c) The whole domain constitutes one model group.
Figure A6.2 a-c. MIDSS, domain C. QQ-plots of the standardised logarithms of the study variable against theoretical quantiles from the standard normal distribution.

Next, I test the lognormality of the population by computing the test statistic

$$\psi = \max_i \left(z_i - \bar{z}\right) / s_z ,$$

where z_i is the log of the positive study variable, $\bar{z}$ and s_z are the mean and the standard deviation of the z_i in the model group (zero values excluded). This statistic measures the maximum distance from the mean, which is of particular interest as the largest values will contribute most to the total (cf Thorburn 1991). For each model group, a set of 1000 values of ψ was generated from 1000 populations of the same size as the model groups, consisting of normally distributed random variables. The observed MIDSS test statistics were compared to the 2.5th and 97.5th percentiles in these lists. Table A6.1 compares means based on the log scale and original scale, as well as test statistics compared with 2.5th and 97.5th normal percentiles. Only Domain C is shown as an example. Table A6.2 summarises the results for Domains A-C. No association between lack of log-normality and simulated bias can be seen in this table. The conclusion is that although there are indications of some model problems, it is not these that cause the bias the mixture model estimation shows in Table 5.7.

Table A6.1. MIDSS, domain C

a) Model groups equal to strata, genuine sampling strata only

Model group	Number of units	Superpop mean based on lognormal model	Arithmetic population mean on original scale	Ratio mean on logscale to orig scale
1	206	14.5	23.5	0.6
2	129	93.0	83.0	1.1
3	305	286.3	279.1	1.0

Model group	Test statistic	2.5 percentile	97.5 percentile
1	5.1	2.3	3.9
2	2.2	2.1	3.6
3	3.3	2.4	3.7

b) ONS model groups, group consisting of genuine sampling strata only

Model group	Number of units	Superpop mean based on lognormal model	Arithmetic population mean on original scale	Ratio mean on logscale to orig scale
1	640	240.6	147.8	1.6

Model group	Test statistic	2.5 percentile	97.5 percentile
1	2.5	2.7	4.0

c) The whole domain constitutes one model group

Model group	Number of units	Superpop mean based on lognormal model	Arithmetic population mean on original scale	Ratio mean on logscale to orig scale
1	853	1106.5	768.9	1.4

Model group	Test statistic	2.5 percentile	97.5 percentile
1	2.8	2.8	4.0

Table A6.2. Results of lognormality tests and simulations, MIDSS domains

Domain	Type of model group	Model mean too large (+), to small (-), or within confidence interval (0)	Simulated bias, %; First number is bias for Logn/pr and Logn/log, second number in brackets is bias for RLogn/pr
A	Within strata, Stratum 1	0	1 (4)
	Stratum 2	0	
	Stratum 3	0	
	ONS groups	0	1 (0)
	Over all strata	0	2 (1)
B	Within strata, Stratum 1	0	Very large
	Stratum 2	0	
	Stratum 3	0	
	ONS groups	+	1 (0)
	Over all strata	0	1 (0.5)
C	Within strata, Stratum 1	-	0 (-1)
	Stratum 2	0	
	Stratum 3	0	
	ONS groups	0	1.5 (0)
	Over all strata	0	1 (-0.5)

Note: the first column of results compares superpopulation mean with observed mean. The second column reproduces some of the simulated bias results reported in Table 5.7.

Appendix 7. The projective bias adjusted estimator is linear

First note that

$$\hat{t}_y = \sum_{k\in U} \hat{y}_k + \sum_{j\in s} \tilde{\omega}_j \left(y_j - \hat{y}_j\right)$$

$$= \sum_{j\in s} \tilde{\omega}_j y_j + \sum_{k\in U} \left[1 - \tilde{\omega}_k I(k \in s)\right] \hat{y}_k .$$

By definition $\hat{y}_k = \boldsymbol{\omega}_k' \mathbf{y}_s$, $\mathbf{y}_s' = (y_1, y_2, \ldots, y_n)$ for some vector $\boldsymbol{\omega}_k$. Hence

$$\hat{t}_y = \sum_{j\in s} \tilde{\omega}_j y_j + \left\{ \sum_{k\in U} \left[1 - \tilde{\omega}_k I(k \in s)\right] \boldsymbol{\omega}_k' \right\} \mathbf{y}_s .$$

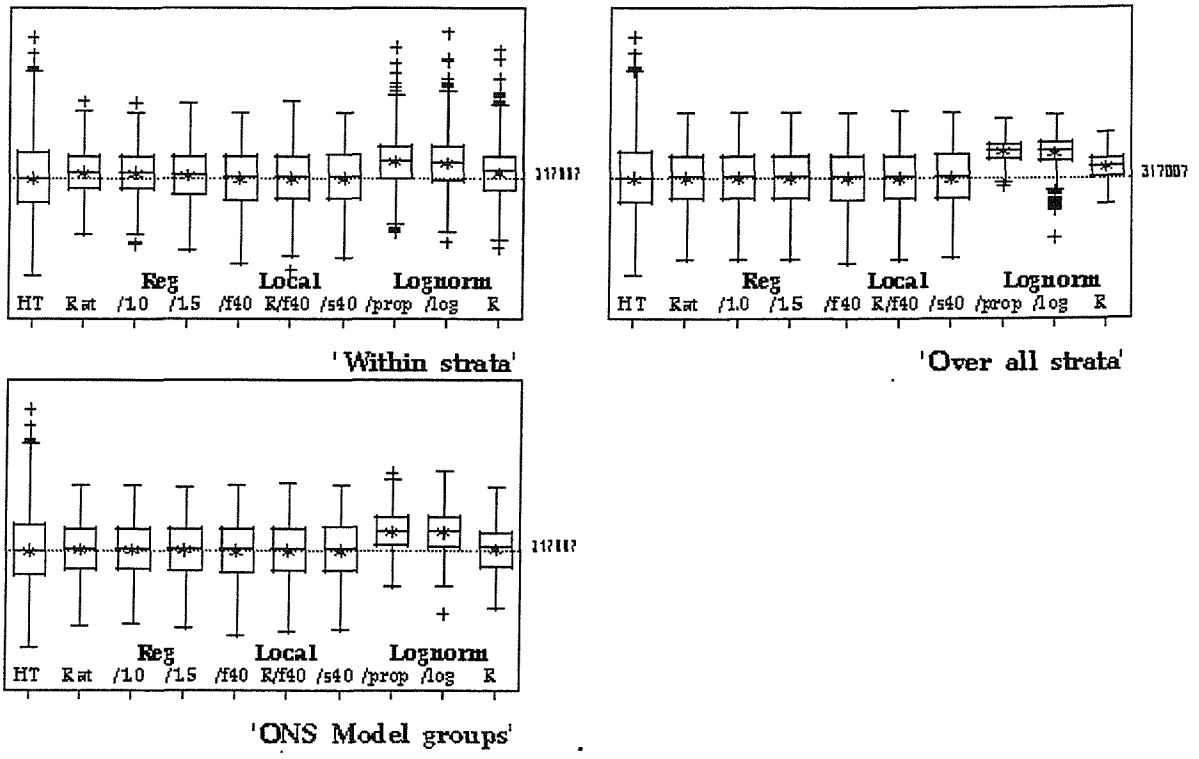
$$= \sum_{j\in s} \tilde{\omega}_j y_j + \sum_{j\in s} \left(\sum_{k\in U} \left[1 - \tilde{\omega}_k I(k \in s)\right] \omega_{kj} \right) y_j ,$$

$$= \sum_{j\in s} \left[\tilde{\omega}_j + \left(t_{\omega j} - \hat{t}_{\omega j}\right)\right] y_j ,$$

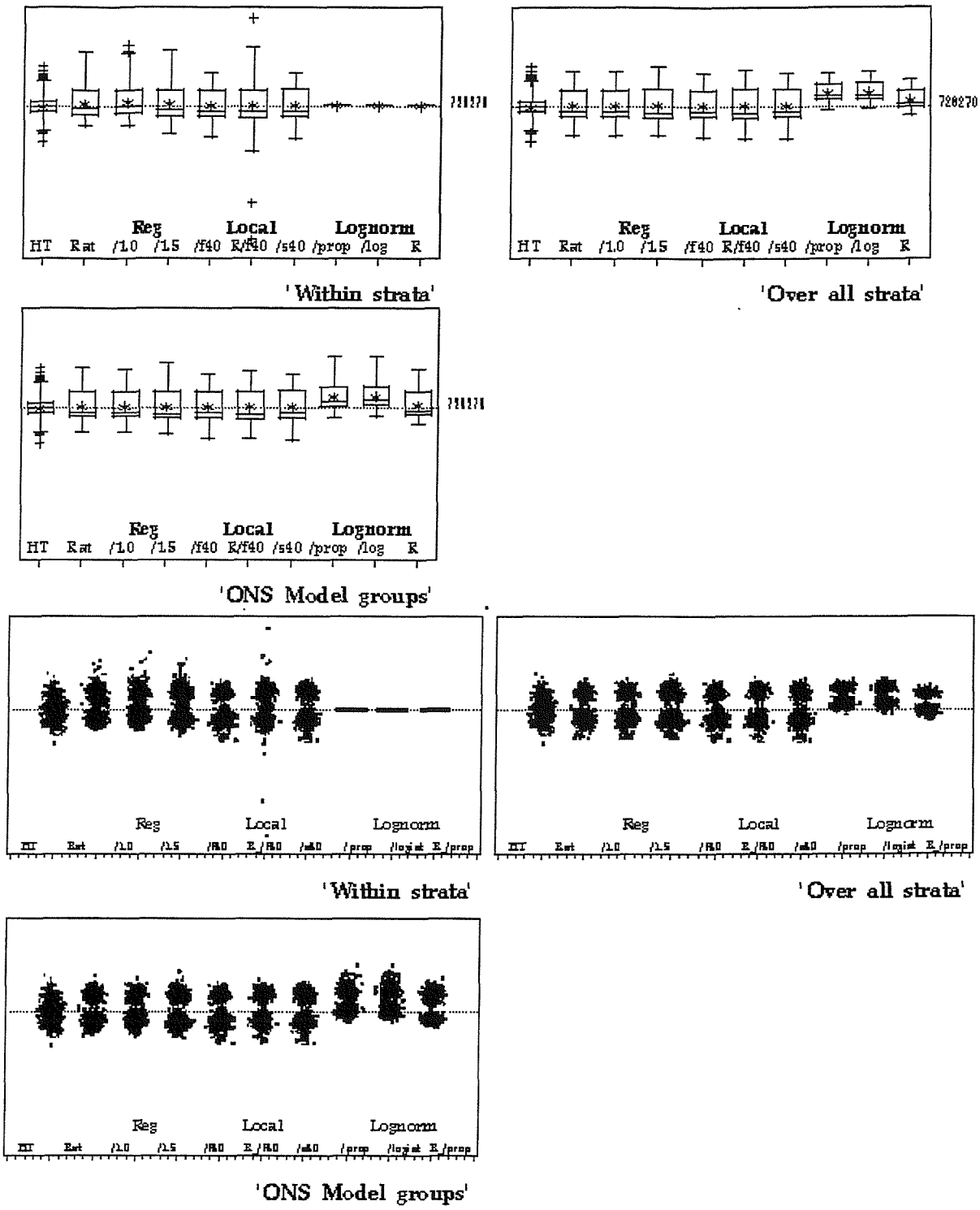
where $t_{\omega j} = \sum_{k\in U} \omega_{kj}$ and $\hat{t}_{\omega j} = \sum_{k\in s} \tilde{\omega}_k \omega_{kj}$.

Appendix 8. Simulation results, boxplots

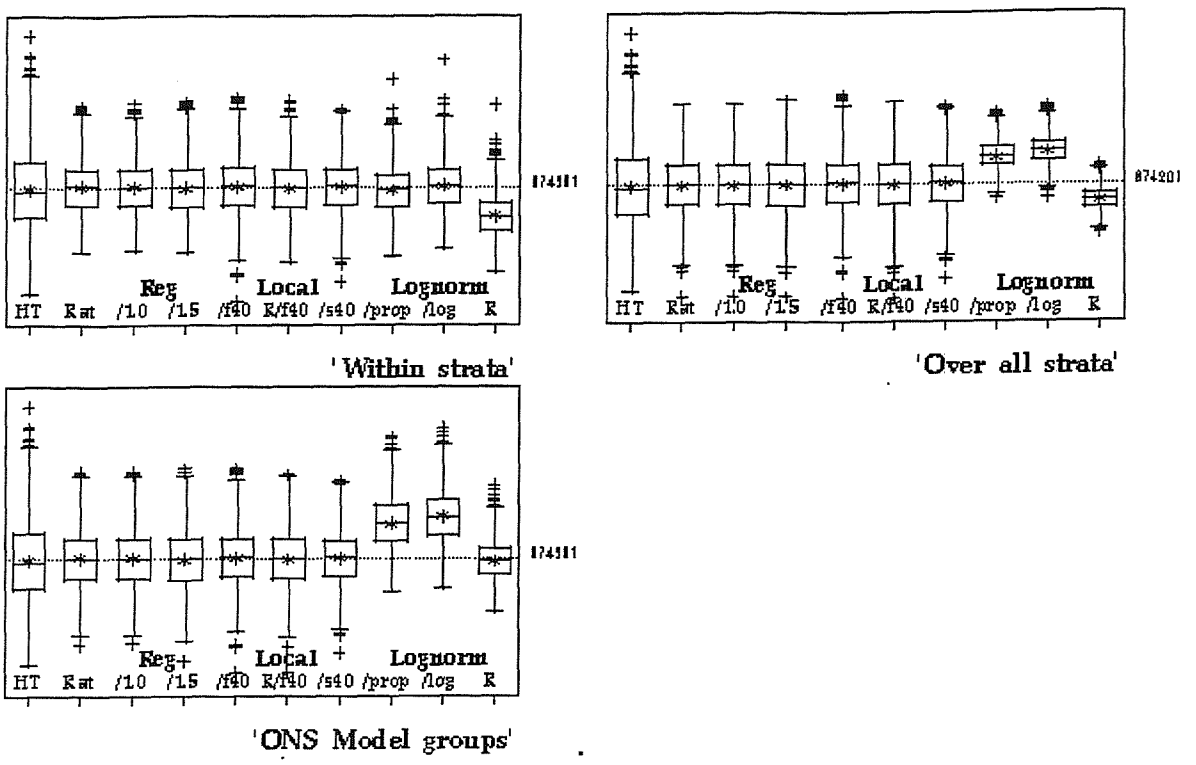
Figure A8.1 and A8.2 show box plots for the point estimates for the MIDSS domains A-C and the CAPEX domains U and V. The estimator RobReg is denoted by R/f40. The scale of the y-axis is the same for the figures in the same panel, but it will not be the same between panels. The design-unbiased estimators produce, as they should, estimates with the arithmetic average (a star) on or near the true total. Dot plots are added to the box plots for MIDSS domain B to highlight the bimodal distribution of the point estimates for this domain. The box plots indicate that the estimators fall into three groups: Lognorms, the HT and the others.



a) MIDSS, domain A.

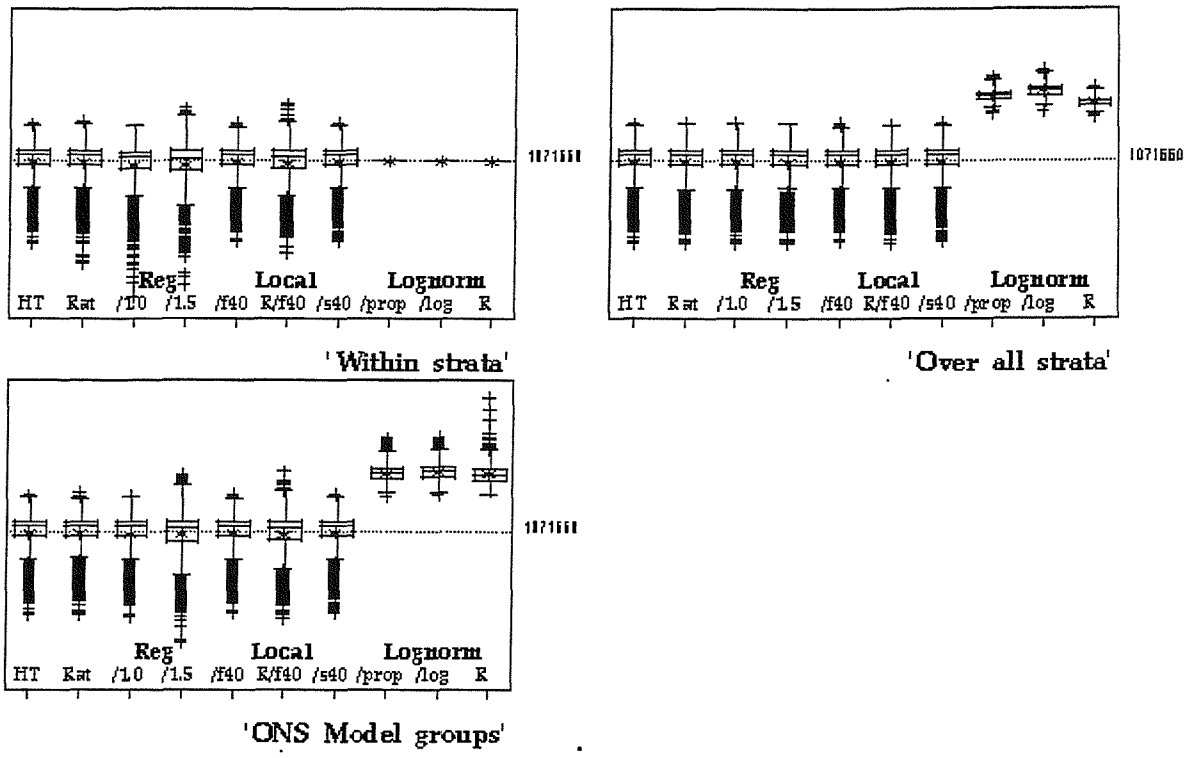


b) MIDSS, domain B. Lognorm 'within strata' are taken out not to swamp the other box plots. The top set of 3 graphs shows box plots, the bottom set jittered dot plots

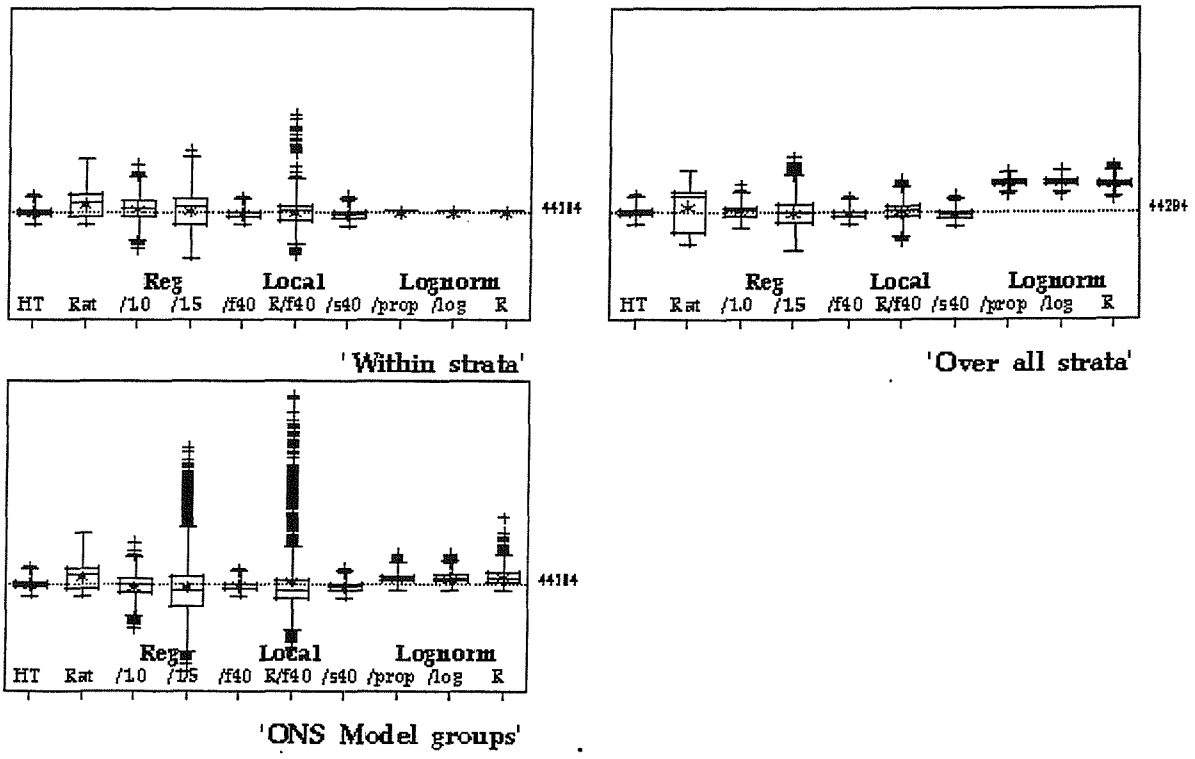


d) MIDSS, domain C.

Figure A8.1. Box plots of point estimates for MIDSS domains A-C. The averages of the estimates are marked with a star. The horizontal dotted lines show the true total of the populations. The scale of the y-axes is the same for all three graphs within a panel.



a) CAPEX, domain U.



b) CAPEX, domain V.

Figure A8.2. Box plots of point estimates for CAPEX domains U and V. The averages of the estimates are marked with a star. The horizontal dotted lines show the true total of the populations. The scale of the y-axes is the same for all three graphs within a panel.

References

Agresti, A. (1990). Categorical Data Analysis. New York: Wiley.

Albert, A. and Anderson, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. Biometrika, 71, 1-10.

Andersson, C. and Nordberg, L. (1998). A User's Guide to CLAN 97. Statistiska Centralbyrån - Statistics Sweden.

Atmer, J., Thulin, G., and Bäcklund, S. (1975). Samordning av urval med JALES metoden, Statistisk Tidskrift, 13, 443-450.

Bardsley, P. and Chambers, R.L. (1984). Multipurpose Estimation from Unbalanced Samples. Applied Statistics, 33, 290-299.

Basu, D. (1971). An Essay on the Logical Foundations of Survey Sampling. In Foundations of Statistical Inference, eds. V.P. Godambe and D.A. Sprott. Toronto: Holt, Rinehart and Winston, 203-242.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias through Regression Estimation. Journal of Official Statistics, 4, 251-260.

Bethlehem, J.G. (1997). Integrated Control Systems for Survey Processing. In Survey Measurement and Process Quality, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: Wiley, 371-392.

Bethlehem, J.G. and Keller, W.J. (1987). Linear Weighting of Sample Survey Data. Journal of Official Statistics, 3, 141-153.

Bethlehem, J.G. and van de Pol, F. (1998). The Future of Data Editing. In Computer Assisted Survey Information Collection, eds. M.P. Couper, R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls II, J.M. O'Reilly. New York: Wiley, 201-222.

Biemer, P. and Caspar, R. (1994). Continuous Improvement for Survey Operations: Some General Principles and Applications, Journal of Official Statistics, 10, 307-326.

Bornhuetter, R.L. and Ferguson, R.E. (1972). The Actuary and IBNR. Proceedings of the Casualty Actuarial Society, LIX, 181-195.

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations, Journal of the Royal Statistical Society, series B, 26, 211-252.

Breidt, F.J. and Opsomer, J.D. (2000). Local Polynomial Regression Estimation in Survey Sampling. The Annals of Statistics, 28, 1026-1053.

Brewer, K.R.W. (1994). Survey Sampling Inference: Some Past Perspectives and Present Prospects. Pakistan Journal of Statistics, 10, 213-233.

Brewer, K.R.W. (1999). Design-Based or Prediction-Based Inference? Stratified Random vs Stratified Balanced Sampling. International Statistical Review, 67, 35-47.

Brewer, K.R.W. (2002). Combined Survey Sampling Inference: Weighing Basu's Elephants. London: Arnold.

Casella, G. and Berger, R.L. (1990). Statistical Inference. Belmont: Duxbury Press.

Chalmers, A.F. (1982). What is This Thing Called Science? 2nd ed. Milton Keynes: Open University Press.

Chambers, R.L. (1996). Robust Case-Weighting for Multipurpose Establishment Surveys. Journal of Official Statistics, 12, 3-32.

Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. Journal of the American Statistical Association, 88, 268-277.

Chambers, R.L. and Kokic, P.N. (1993). Outlier Robust Sample Survey Inference. Bulletin of the International Statistical Institute, Invited Papers, 69-86.

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74, 829-836.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed.. New York: Wiley.

Colledge, M.J. (1989). Coverage and Classification Maintenance Issues in Economic Surveys. In Panel Surveys, eds. D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh. New York: Wiley, 80-107.

Colledge, M.J. (1995). Frames and Business Registers: An Overview. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 21-47.

Colledge, M.J. (1999). Statistical Integration through Metadata Management. International Statistical Review, 67, 79-98.

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall.

Cox, B., Binder, D., Chinnappa, N., Christianson, A., Colledge, M., and Kott, P. (eds) (1995). Business Survey Methods. New York: Wiley.

Cox, B. and Chinnappa, N. (1995). Unique Features of Business Surveys. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 1-17.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and their Application. Cambridge: Cambridge University Press.

De Vylder, F.E. (1996). Advanced Risk Theory. Brussels: Editions de l'Universite de Bruxelles.

Dembski. W.A. (1998). Randomness. In Routledge Encyclopedia of Philosophy, Vol. 8, ed. E. Craig. London: Routledge, 56-59.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376-382.

Dillman, D.A. (2000). Procedures for Conducting Government-Sponsored Establishment Surveys: Comparison of the Total Design Method (TDM), a Traditional Cost-Compensation Model, and Tailored Design. Proceedings of the Second International Conference on Establishment Surveys. American Statistical Association, 343-352.

Dorfman, A.H. (2000). Non-Parametric Regression for Estimating Totals in Finite Populations. Proceedings of the Survey Research Methods. American Statistical Association, 47-54.

Durbin, J. (1969). Inferential Aspects of the Randomness of Sample Size in Survey Sampling. In New Developments in Survey Sampling, eds. N.L. Johnson, and H. Smith. New York: Wiley, 629-651.

Dutka, S. and Frankel L.R. (1991). Measurement Errors in Business Surveys. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N.A. Mathiowetz, and S. Sudman. New York: Wiley, 113-123.

Edwards, W.S. and Cantor, D. (1991). Towards a Response Model in Establishment Surveys. In Measurement Errors in Surveys, eds. P.P. Biemer, R.M. Groves, L. Lyberg, N.A. Mathiowetz, and S. Sudman. New York: Wiley, 211-233.

Eldridge, J., Martin, J. and White, A. (2000). The Use of Cognitive Methods to Improve Establishment Surveys in Britain. Proceedings of the Government Statistical Service Methodological Conference, UK Government Statistical Service, 19-31.

England, P.D. and Verrall, R.J. (2002). Stochastic Claims Reserving in General Insurance. Paper presented to the Institute of Actuaries, London, UK, 28 January 2002.

Ernst, L.R., Valliant, R., and Casady, R.J. (2000). Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths. Journal of Official Statistics, 16, 211-228.

Estevao, V., 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.

Estevao, V.M. and Särndal, C.-E. (2000). A Functional Form Approach to Calibration. Journal of Official Statistics, 16, 379-399.

European Union (1993). Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community.

Falvey, H., and Hedlin, D. (1999). Quarterly capital expenditure inquiry review 1998 (MQ 055). Methods & Quality Division, Office for National Statistics.

Fan, C.T., Muller, M.E., and Rezucha, I. (1962). Development of Sampling Plans by Using Sequential (item by item) Techniques and Digital Computers. Journal of the American Statistical Association, 57, 387-402.

Fellegi, I. and Wolfson, M. (1999). Towards Systems of Social Statistics – Some Principles and Their Application in Statistics Canada. Journal of Official Statistics, 15, 373-393.

Fenton, T., Hedlin, D., Perry, J., and Pont, M. (1999). Births and Deaths of Business Units. The Impact of Reporting Delays on the Processes of Maintaining the UK Business Register and Producing Economic Statistics (MQ 063). Methods & Quality Division, Office for National Statistics.

Fuller, W.A. (2002). Regression Estimation for Survey Sampling. Survey Methodology, 28, 5-23.

Godambe, V.P. and Joshi, V.M. (1986). Admissibility and Bayes Estimation in Sampling Finite Populations, 1. Annals of Mathematical Statistics, 36, 1707-1722.

Granquist, L. and Kovar, J.G. (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: Wiley, 415-435.

Haberman, S. and Renshaw, A.E. (1996). Generalized Linear Models and Actuarial Science. The Statistician, 45, 407-436.

Harris, J.E. (1990). Reporting Delays and the Incidence of AIDS. Journal of the American Statistical Association, 85, 915-924.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. Statistical Science, 1, 297-310.

Hedlin, D. (1998). On the Stratification of Highly Skewed Populations (No. B:41). Research Report, series B, Applied Research. Unpublished manuscript. University of Stockholm, Institute of Actuarial Mathematics and Mathematical Statistics.

Hedlin, D. (2000). A Procedure for Stratification by an Extended Ekman Rule. Journal of Official Statistics, 16, 15-29.

Hedlin, D. (2002a). Score Functions to Reduce Business Survey Editing at the UK Office for National Statistics. Invited paper, UNECE Work Session on Statistical Data Editing, Helsinki, Finland, 27-29 May.

Hedlin, D. (2002b). Estimating Totals in some UK Business Surveys. Statistics in Transition, 5, 943-968.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the UK Office for National Statistics. Journal of Official Statistics, 19, in press.

Hedlin, D., Falvey, H., Chambers, R., and Kokic, P. (2001). Does the Model Matter for GREG Estimation? A Business Survey Example. Journal of Official Statistics, 17, 527-544.

Hedlin, D., Pont, M., and Fenton, T. (2000). Estimating the Effects of Birth and Death Lags on a Business Register [CD-ROM]. Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association.

Hedlin, D. and Wang, S. (2002). Feeding Back Information on Ineligibility from Sample Surveys to the Frame. Revised version submitted for publication.

Hidiroglou, M.A. and Laniel, N. (2001). Sampling and Estimation Issues for Annual and Sub-Annual Canadian Business Surveys. International Statistical Review, 69, 487-504.

Hidiroglou, M.A. and Srinath K.P. (1993). Problems Associated with Designing Subannual Business Surveys. Journal of Business and Economic Statistics, 11, 397-405.

Holt, D. (1998). Statistical Integrity. The Fourth Kenneth Hill Memorial Lecture. University of Southampton.

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

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

Isaki, C.T. and Fuller, W.A. (1982). Survey Design under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, 89-96.

Jayasuriya, B.R. and Valliant, R. (1996). An Application of Restricted Regression Estimation in a Household Survey. Survey Methodology, 22, 127-137.

Kalton, G. (2002). Models in the Practice of Survey Sampling (revisited). Journal of Official Statistics, 18, 129-154.

Karlberg, F. (2000). Survey Estimation for Highly Skewed Populations in the Presence of Zeroes. Journal of Official Statistics, 16, 229-241.

Keller, W. (1999). The Synthetic Census. Proceedings of the Government Statistical Service Methodological Conference, UK Government Statistical Service, 5-16.

Klugman, S.A., Panjer, H.H., and Willmot, G.E. (1998). Loss Models: From Data to Decisions. New York: Wiley.

Knoke, D. and Burke, P. (1980). Log-Linear Models. Sage University Paper Series on Quantitative Applications in the Social Sciences (07-020). Beverly Hills and London: Sage Publications.

Kokic, P.N. and Bell, P.A. (1994). Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator. Journal of Official Statistics, 10, 419-435.

Kokic, P.N. and Brewer, K.R.W. (1996). Operational Aspects of the IDBR. Methods & Quality Division, Office for National Statistics.

Kovar, J.G., ed. (2000). Proceedings of the Second International Conference on Establishment Surveys. American Statistical Association.

Krzanowski, W.J. (1990). Principles of Multivariate Analysis. A User's Perspective. Oxford: Clerendon Press.

Kuk, A.Y.C. and Welsh, A.H. (2001). Robust Estimation for Finite Populations Based on a Working Model. Journal of the Royal Statistical Society, series B, 63, 277-292.

Lavallée, P. (1996). Frame Update Problems with Panel Surveys. Proceedings of Statistical Days '96, Statistical Society of Slovenia, 252-261.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243-253.

Leaver, S. and Valliant, R. (1995). Statistical Problems in Estimating the U.S. Consumer Price Index. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 543-566.

Lee, H. (1995). Outliers in Business Surveys. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 503-526.

Lee, H., Rancourt, E., and Särndal, C.-E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. Journal of Official Statistics, 10, 231-243.

Lewington, R.J. (1995). The Role of National Accounts and Their Impact on Business Surveys. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 655-690.

Loader, C. (1999). Local Regression and Likelihood. New York: Springer-Verlag.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. Journal of Official Statistics, 15, 305-327.

Lundström, S. and Särndal, C.-E. (2001). Estimation in the Presence of Nonresponse and Frame Imperfections. Statistiska Centralbyrån - Statistics Sweden.

Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N, and Trewin, D., eds (1997). Survey Measurement and Process Quality. New York: Wiley.

Mack, T. (1991). A Simple Parametric Model for Rating Automobile Insurance or Estimating IBNR Claims Reserves. ASTIN Bulletin, 21, 9-109.

Mack, T. (2000). Credible Claims Reserves: The Benktander Method. ASTIN Bulletin, 30, 333-347.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd ed. London: Chapman & Hall.

McDonald, J.W. (1998). Quasi-Independence. In Encyclopedia of Biostatistics, eds. P. Armitage and T. Colton. New York: Wiley, 3637-3639.

Morgenstein, D. and Marker, D.A. (1997). Continuous Quality Improvement in Statistical Agencies. In Survey Measurement and Process Quality, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: Wiley, 475-500.

Nijhowne, S. (1995). Defining and Classifying Statistical Units. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 49-64.

Ohlsson, E. (1995). Coordination of Samples Using Permanent Random Numbers. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 153-169.

Ohlsson, E. (1998). Sequential Poisson Sampling. Journal of Official Statistics, 14, 149-162.

Pietsch, L. (1995). Profiling Large Businesses to Define Frame Units. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott. New York: Wiley, 101-114.

Rao, J.N.K. (1997). Developments in Sample Survey Theory: An Appraisal. Canadian Journal of Statistics, 25, 1-21.

Renshaw, A.E. and Verrall, R.J. (1998). A Stochastic Model Underlying the Chain-Ladder Technique. British Actuarial Journal, 4, 903-923.

Rivière, P. (2002). What Makes Business Statistics Special? International Statistical Review, 70, 145-159.

Robert, C.P., Hwang, J.T.G., and Strawderman, W.E. (1993). Is Pitman Closeness a Reasonable Criterion? (with Discussion). Journal of the American Statistical Association, 88, 57-76.

Robinson, P.M. and Särndal, C.-E. (1983). Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling. Sankhyā, series B, 45, 240-248.

Ryan, T.P. (1997). Modern Regression Methods. New York: Wiley.

Särndal, C.-E. (1980). On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling. Biometrika, 67, 639-650.

Särndal, C.-E. (1982). Implications of Survey Design for Generalized Regression Estimation of Linear Functions. Journal of Statistical Planning and Inference, 7, 155-170.

Särndal, C.-E., Swensson, B., and Wretman, J. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 527-537.

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

Särndal, C.-E. and Wright, R.L. (1984). Cosmetic Form of Estimators in Survey Sampling. Scandinavian Journal of Statistics, 11, 146-156.

SAS Institute (2000). SAS/STAT User's Guide, Version 8. Cary, NC, SAS Institute Inc.

Schiopu-Kratina, I. and Srinath, K. P. (1991). Sample Rotation and Estimation in the Survey of Employment, Payrolls and Hours. Survey Methodology, 17, 79-90.

Sellero, C.S., Fernández, E.V., Manteiga, W.G., Otero, X.L., Hervada, X., Fernández, E., and Taboada, X.A. (1996). Reporting Delay: a Review with a Simulation Study and Application to Spanish AIDS Data. Statistics in Medicine, 15, 305-321.

Sen, A.R. (1953). On the Estimate of the Variance in Sampling with Varying Probabilities. Journal of the Indian Society of Agricultural Statistics, 5, 119-127.

Sen, A. and Srivastava, M. (1990). Regression Analysis. Theory, Methods and Applications. New York: Springer-Verlag.

Sigman, R. and Monsour, N. (1995). Selecting Samples from List Frames of Business. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, P. Kott. New York: Wiley, 133-152.

Silva, P.L.D.N. and Skinner C.J. (1997). Variable Selection for Regression Estimation in Finite Populations. Survey Methodology, 23, 23-32.

Singh, A.C. and Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. Survey Methodology, 22, 107-115.

Skinner, C.J. (1999). Calibration Weighting and Nonsampling Errors. Research in Official Statistics, 1, 33-43.

Skinner, C.J., Holt, D., and Smith, T.M.F. (eds) (1989). Analysis of Complex Surveys. New York: Wiley.

Smith, T.M.F. (1994). Sample Surveys 1975-1990. An Age of Reconciliation? International Statistical Review, 62, 5-19.

Smith, T.M.F. (1997). Social Surveys and Social Science. Canadian Journal of Statistics, 25, 23-44.

Smith, T.M.F. (2001). Biometrika Centenary: Sample Surveys. Biometrika, 88, 167-194.

Smith, P., Pont, M., and Jones, T. (2003). Developments in Business Survey Methodology in the Office for National Statistics, 1994–2000. The Statistician, 52, 1-30, in press.

Srinath, K.P. (1987). Methodological Problems in Designing Continuous Business Surveys: Some Canadian Experiences. Journal of Official Statistics, 3, 283-288.

Srinath, K.P. and Carpenter, R.M. (1995). Sampling Methods for Repeated Business Surveys. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott, New York: Wiley, 171-183.

Stukel, D.M., Hidiroglou, M.A., and Särndal, C.-E. (1996). Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization. Survey Methodology, 22, 117-125.

Sugden, R.A. and Smith, T.M.F. (2002). Exact Linear Unbiased Estimation in Survey Sampling. Journal of Statistical Planning and Inference, 102, 25-38.

Sugden, R.A., Smith, T.M.F., and Jones, R.P. (2000). Cochran's Rule for Simple Random Sampling. Journal of the Royal Statistical Society, series B, 62, 787-793.

Thompson, M.E. (1997). Theory of Sample Surveys. London: Chapman & Hall.

Thompson, K.J. and Sigman, R.S. (1999). Statistical Methods for Developing Ratio Edit Tolerances for Economic Data. Journal of Official Statistics, 15, 517-535.

Thorburn, D. (1991). Behandling av Ouliers i Ekonomisk Statistik. Empiriska test av lognormalfördelning (PM nr. 3). Unpublished manuscript, University of Stockholm, Department of Statistics.

Underwood, C. (2001). Implementing Selective Editing in a Monthly Business Survey. Paper given at the Sixth Government Statistical Service Methodological Conference, London, 25 June.

Valliant, R. (1987). Some Prediction Properties of Balanced Half-Sample Variance Estimation in Single-Stage Sampling, series B, 49, 68-81.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley.

Wang, M.-C. (1992). The Analysis of Retrospectively Ascertained Data in the Presence of Reporting Delays. Journal of the American Statistical Association, 87, 397-406.

Welsch, R.E. (1980). Regression Sensitivity Analysis and Bounded-Influence Estimation. In Evaluation of Econometric Models, eds. J. Kmenta and J.B. Ramsey. New York: Academic Press, 153-167.

Welsh, A.H. and Ronchetti, E. (1998). Bias-Calibrated Estimation from Sample Surveys Containing Outliers. Journal of the Royal Statistical Society, series B, 60, 413-428.

Wright, R.L. (1983). Finite Population Sampling with Multivariate Auxiliary Information. Journal of the American Statistical Association, 78, 879-884.

Wu, C.F. and Deng, L.Y. (1983). Estimation of Variance of the Ratio Estimator: An Empirical Study. In Scientific Inference, Data Analysis, and Robustness, eds. G.E.P. Box, T. Leonard, C.F. Wu. New York: Academic Press, 245-277.

Yates, F. and Grundy, P.M. (1953). Selection Without Replacement from Within Strata with Probability Proportional to Size. Journal of the Royal Statistical Society, series B, 15, 235-261.

Zelterman, D. (2002). Advanced Log-Linear Models Using SAS. Cary: SAS Institute Inc.