R&D Report 2004:1
Research – Methods – Development

Business Survey Estimation

by

Dan Hedlin

Statistiska centralbyrån Statistics Sweden

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 2004:1. Business survey estimation / Dan Hedlin.
Digitalt skapad fil, anpassad efter de digitaliserade delarna i serien. Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-2004-X101OP0401

R&D Report 2004:1
Research – Methods – Development

Business Survey Estimation

by

Dan Hedlin

R&D Report 2004:1

Research - Methods - Development

Business Survey Estimation

Från trycket Mars 2004

Producent Statistiska centralbyrån, *Statistics Sweden,* metodenheten
Box 24300, *SE*-104 51 STOCKHOLM

Förfrågningar Dan Hedlin
Dan.Hedlin@scb.se
Telefon 08 – 506 943 34

ISSN 0283-8680

Printed in Sweden
SCB-Tryck, Örebro 2004

February 2004

Abstract

The paper discusses estimation of the total for some study variables in two business surveys conducted by the Office for National Statistics (ONS) in the UK. The MSE cannot be the one and only criterion of estimator quality: other desirable properties of an estimator are proposed. Special consideration is given to the proneness of an estimator 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. Several point estimators are explored in a simulation study. Some widely used design-based estimators for stratified simple random sampling (and two less widely used ones) are contrasted with a model-based estimator that explicitly draws on the special structure of a business population.

Acknowledgements: The main part of this work was done while the author worked for Office for National Statistics. Part of it has appeared in Statistics in Transition, Vol 5.

1. Introduction

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 Statistics Sweden and Office for National Statistics (ONS) in the UK. 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. This paper illustrates methods for addressing some of these issues in a real business survey.

The choice of estimator depends on the foreseen or believed use of the resulting estimates. One of the most important, or the most important, recipients of official business statistics is the national accounts. 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 constitutes 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 in business surveys than those of point estimates, as, for example, design-bias.

2. Pros and cons with the generalised regression estimator

2.1 A refresher

The aim of many business surveys is to estimate totals and differences between or ratios of totals. We explore mainly design-based linear estimators of the total $t_y = \sum_U y_k$ of a study variable $\mathbf{y}' = (y_1, y_2, \ldots y_N)$ in a population U where the units have the labels $\{1, 2, \ldots N\}$. The issue is how to use auxiliary information effectively. We concentrate on estimation of a single study variable. Multivariate study variable estimation issues are discussed, among others, by Bethlehem and Keller (1987) and Chambers (1996). Discussions of multiparameter design issues include Holmberg (2002). 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, but more often than not, $\mathbf{x}_k$ is indeed available for all units on the frame in actual business survey systems. Cassel, Lundquist, and Selén (2002) propose a model-based estimator that only needs auxiliary values for sample units (i.e. not even totals of $\mathbf{x}_k$).

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. Nonsampling errors, that is nonresponse, measurement and coverage errors are disregarded.

The generalised regression (GREG) was introduced by Cassel, Särndal, and Wretman (1976). It can be written

$$\hat{t}_{yreg} = \sum\nolimits_s g_{ks} w_k y_k \ , \qquad (1)$$

where $w_k = \pi_k^{-1}$ is the sampling weight for unit *k* and the sum is taken over sample units. See Särndal, Swensson, and Wretman (1992, p. 232) for a definition of the sample-dependent 'g-weights' g_{ks}, *k* = 1, 2, … *n*. Thus the 'total' weight $w'_{ks} = g_{ks} w_k$ in (1) is partitioned into a purely design-dependent weight w_k and a weight g_{ks}.

The GREG estimator is a special case of the calibration estimator (Deville and Särndal 1992). The calibration estimators have all the property that the auxiliary totals are recovered; in the case of GREG we have $\sum\nolimits_s g_{ks} w_k \mathbf{x}_k = \mathbf{t}_x$, where $\mathbf{t}_x$ is a vector of totals of the components of $\mathbf{x}_k$. The g-weights force the estimator to be 'calibrated' on $\mathbf{t}_x$. Note that $\mathbf{t}_x$ may contain totals on different levels, for example overall totals and domain totals. Note also that (1) is reminiscent of a Horvitz-Thompson-estimator (HT-estimator) $\hat{t}_{y\pi} = \sum\nolimits_s w_k y_k$ of $\sum\nolimits_U g_{ks} y_k$, 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.

2.2 Strong sides of the GREG

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 mild mathematical restrictions. The auxiliary variables may be qualitative or quantitative; and they may be associated with units of different level, e.g. company and local unit. Softwares like CLAN (Andersson and Nordberg 1998) allow the user to specify the marginal sums to be calibrated on without having to work out the exact form of the estimator. Furthermore, since the GREG is derived for a general set of inclusion probabilities it can be specialised to any sampling design. Special cases of the GREG are discussed by, among others, Särndal et al. (1992).

In business statistics, the perceived main advantage of GREGs 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. 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. Lundström (2000) reports on practices at Statistics Sweden.

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.

The GREG offers a nice interpretation: the form of the estimator reflects the view that a sampled element can be seen as representing $g_{ks} w_k - 1$ 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.

Note that the g-weights depend on both model and design. Hence the g-weights allow the survey statistician to bring his or her beliefs, expressed in a model, into the estimator. For an amusing illustration consider Basu's elephants. In a blatant breach with 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. 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 (1), 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. A more prosaic 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 units, which is not true in general.

Under certain regularity conditions, the GREG is design-consistent and asymptotically design-unbiased (Isaki and Fuller, 1982). The former property implies the latter under mild conditions. 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 *j*th component of $\hat{\mathbf{t}}_{x\pi}$. This expression shows that high-leverage points may cause bias, which is highlighted in following sections.

2.3 Problems with the GREG estimator

The GREG estimator has some serious downsides, some of which have not yet been fully explored. The business survey example of Hedlin, Falvey, Chambers, and Kokic (2001) 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 they explore 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. The following four points are taken from that paper.

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, Hedlin et al. (2001) give an example where the estimate for a very reasonable model is close to zero due to negative g-weights. As noted by Chambers (1996), appearance of negative weights is symptomatic of deeper estimation problems and model misspecification.

Second, one motivation of g-weights is that the product of these and the design-weights are made 'close' to design-weights, that is, the g-weights should be close to 1 (Deville and Särndal 1992). The g-weights do approach unity asymptotically (Särndal 1982). However, it does not follow from this fact that the design weights and calibration weights are similar. In fact, 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 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 is less often mentioned.

Fourth, 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.

3. A Comparison of Some Alternative Estimators of Totals

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 having competitive variance. We shall explore some alternative estimators for business surveys.

Most business surveys at an NSI are multipurpose with customers who use the statistics in different ways. 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 interval or minimum risk of obtaining an estimate with large error; or versatility or ease of implementation. We 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. Lundström and Särndal (1999, 2001), and Lundström and Gustafsson (2003) discuss choice of auxiliary variables. Here five properties for each combination of estimator and type of model group partition have been measured. Two of the properties 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, see Figure 1. There are two interval scaled variables on the frame: register employment and turnover. Industries are important domains of study. We assume full response and ignore measurement errors and incomplete coverage of the target population.

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

Figure 1. Most common sampling and estimation strategy in an ONS business survey domain

In Section 3.2 the model groups and estimators used in the simulation study are defined, whose results are reported in Section 4. The paper ends with a discussion. An earlier version of Sections 3.2 through 4.2 was published in Hedlin (2002).

3.2 Estimators

3.2.1 Aim

The aim is to estimate the total $t_y = \sum_U y_k$ of a study variable $\mathbf{y}' = (y_1, y_2, \ldots, y_N)$ in a population *U*. It is assumed that there is a known auxiliary variable (*N* x *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. We 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.

3.2.2 Model groups

Let subscript *g* 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 1);
c) all strata within an industry, including the CE stratum, constitute one group.

Case *b* is referred to as '**within genuine sampling strata**'. 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**'.

3.2.3 Point estimators

A general form of point estimators

Many estimators used in practice are of the form

$$\hat{t}_y = \sum_{k\in U} \boldsymbol{\omega}'_k \mathbf{y}_s + \sum_{j\in s} \tilde{\omega}_j \left(y_j - \hat{y}_j\right), \qquad (2)$$

where $\boldsymbol{\omega}_k$ is a weight vector and $\tilde{\omega}_k$ a scalar, neither dependent on **y**, and $\mathbf{y}'_s = \left(y_1, y_2, \ldots, y_n\right)$. 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 (2) consists of a 'model-based' or 'synthetic' term plus a 'bias adjustment' or 'correction' term. We refer to an estimator that can be written on the form (2) 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 (2) 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 in this study, $\mathbf{x}'_k = \left(1 \quad x_k\right)$ or $\mathbf{x}_k = x_k$. To see that the GREG can be written in the form (2), 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 matrices with π_k and the residual variance σ_k^2 in position (k, k), respectively.

It can readily be shown that (2) is a linear estimator, 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 **y**. This property is highly desirable from a national statistical institute's point of view. The main reason is practical: for example, 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. Also, a linear estimator is internally consistent in the sense that if $\hat{t}_i$ is an estimator of the total of 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 in (2) is always zero (e.g. the ratio estimator), for some other estimators it will take whatever value to achieve some overall property. Regression estimators corresponding to a heteroscedastic regression model with variance function $V_M\left(Y_k\right) = \sigma_k^2$ proportional to $x_k^{1.5}$ or x_k^2 (Särndal et al. 1992, Ch. 6) are asymptotically design-unbiased if and only if the bias adjustment is allowed to be unbounded. As shown by Hedlin et al. (2001), 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 $\sum\limits_{k \in U-s} v_k$, where v_k denotes the unobserved residuals $m(\mathbf{x}_k) - \hat{y}_k$ for non-sample points. The bias can be estimated from the observed residuals r_j: $\sum\limits_{k \in U-s} \hat{v}_k = \sum\limits_{k \in U-s} \sum\limits_{j \in s} \ddot{\omega}_{jk} r_j = \sum\limits_{j \in s} \ddot{\omega}_{j\bullet} r_j$ with some appropriate weights. Hence $\sum\limits_{j \in s} \ddot{\omega}_{j\bullet} r_j$ can be viewed as an estimate of the bias due to model misspecification and a special case of the bias adjustment term in (2). In a design-based framework such as GREG estimation, the second term of (2) may be $\sum\limits_{j \in s} \pi_j^{-1}(y_j - \hat{y}_j)$; this term estimates

$\sum\limits_{k \in U} v_k = t_y - \sum\limits_{k \in U} \hat{y}_k$.

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. 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), Valliant, Dorfman, and Royall (2000, Ch. 11) and Brewer (2002, Ch. 14).

Below follows descriptions of the estimators used in the simulations.

The Horvitz-Thompson estimator

Let

$$\hat{t}_{yg\pi} = \sum\nolimits_{s_g} w_k y_k \qquad (3)$$

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 $s_g = U_g \cap s$. Let $\hat{t}_{y\pi}$ be the expansion estimator of the total t_y in U (i.e., the sum of the group estimates $\hat{t}_{yg\pi}$). We 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):

$$\hat{t}_{yrat} = \sum_{g=1}^{G} t_{xg} \frac{\hat{t}_{yg\pi}}{\hat{t}_{xg\pi}}, \qquad (4)$$

where *x* denotes an auxiliary scalar. The label for this estimator will be ***Rat***. The Rat estimator for the ‘within genuine sampling strata’ type of model group is the estimator the ONS uses for many business surveys (Figure 1).

The GREG estimator (1) can be written as

$$\hat{t}_{yreg} = \sum\nolimits_U \hat{y}_k + \sum\nolimits_s w_k \left(y_k - \hat{y}_k\right), \qquad (5)$$

where $\hat{y}_k = \mathbf{x}'_{kg}\hat{\mathbf{B}}_g$, $\hat{\mathbf{B}}_g = \left(\mathbf{X}'_{sg}\mathbf{\Sigma}^{-1}_{sg}\mathbf{\Pi}^{-1}_{sg}\mathbf{X}_{sg}\right)^{-1}\mathbf{X}'_{sg}\mathbf{\Sigma}^{-1}_{sg}\mathbf{\Pi}^{-1}_{sg}\mathbf{y}_{sg}$, 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 $E_{\dot{M}}\left(Y_k\right) = \mathbf{x}'_{kg}\boldsymbol{\beta}_g$ and $V_{\dot{M}}\left(Y_k\right) = \sigma^2_k$, *k* = 1, 2, …, *N*, where the moments are taken over the model. The Rat estimator $\hat{t}_{yrat}$ is a special case of (5) with $\mathbf{x}'_{kg} = I(k \in g)x_k$ and $V_M\left(Y_k\right) = \sigma^2 x_k$, *k* = 1, 2, …, *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 $V_{\dot{M}}\left(Y_k\right) = \sigma^2 x_k$, and for ***Reg/1.5*** $V_{\dot{M}}\left(Y_k\right) = \sigma^2 x_k^{1.5}$.

The choice of variance function that gives the best fit to the data used in simulations reported below is $V_{\dot{M}}\left(Y_k\right) = \sigma^2 x_k^{1.5}$, as is the case for many business surveys (Brewer 2002, p. 87). Hence, we would expect good performance for Reg/1.5.

Local and robust regression estimators

The predicted values $\hat{y}_k$ in (5) 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

$$\hat{t}_{yloc} = \sum\nolimits_U \hat{m}_k + \sum\nolimits_s w_k \left(y_k - \hat{m}_k\right). \qquad (6)$$

Chambers et al. (1993) and Dorfman (2000) suggest similar but model-based estimators. A *bandwidth* b_k and a smoothing window 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

$$\hat{m}_k = \mathbf{e}'_{(2)1}\left(\mathbf{D}'_k\mathbf{W}_k\mathbf{D}_k\right)^{-1}\mathbf{D}'_k\mathbf{W}_k\mathbf{y}_s, \; k = 1, 2, \ldots, N, \qquad (7)$$

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 $\left[1 \quad \left(x_j - x_k\right)\right]$ in the *jth* row, *j* = 1, 2, …, *n*; $\mathbf{W}_k$, for *k* = 1, 2, …, *N*, are *n* x *n* diagonal matrices with $w_k b_k^{-1} K\left[\left(x_j - x_k\right) b_k^{-1}\right]$ 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 (6) 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 (cf Särndal et al. 1992, p. 453). Like Breidt and Opsomer we use the Epanechikov kernel

$$K(u_{jk}) = \max\left[0, \frac{3}{4}(1-u_{jk}^2)\right]. \qquad (8)$$

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. This motivates two types of *nearest neighbour bandwidth*. The first one is

$$b_k^{(s)} = x_{k+20} - x_{k-20}, \qquad (9)$$

where $x_{k\text{-}20}$ 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. Note that $\hat{m}_k$ will cancel out in the CE stratum in (6). 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 $x_{k\text{-}20}$ does not exist, $x_{k\text{-}20}$ 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,

$$b_k^{(f)} = x_{k+40} - x_{k-40}, \qquad (10)$$

x_{k+40} and $x_{k\text{-}40}$ 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 (9) and (10) are labelled ***Local/s20*** and ***Local/f40***, respectively.

The prediction (7) can be rewritten as

$$\hat{m}_k = \bar{y}_{loc,k} + (x_k - \bar{x}_{loc,k}) \frac{\sum_{j \in s} q_{jk}(x_j - \bar{x}_{loc,k})\, y_j}{\sum_{j \in s} q_{jk}(x_j - \bar{x}_{loc,k})^2} \qquad (11)$$

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 (11). Formulation (11) 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 (6) can with the aid of (7) be written as

$$\hat{t}_{yloc} = \sum\nolimits_s w_j y_j + \sum\nolimits_U [1 - I(k \in s) w_k] \hat{m}_k$$

$$= \sum_{j \in s} w_j y_j + \sum_{j \in s} \left\{ \sum_{k \in U} [1 - I(k \in s) w_k] \mathbf{e}'_{(2)1} (\mathbf{D}'_k \mathbf{W}_k \mathbf{D}_k)^{-1} \mathbf{D}'_k \mathbf{W}_k \mathbf{e}_{(n)j} \right\} y_j \qquad (12)$$

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

$$w_{loc,js} = w_j + \sum_{k \in U} [1 - I(k \in s) w_k]\, \mathbf{e}'_{(2)1} (\mathbf{D}'_k \mathbf{W}_k \mathbf{D}_k)^{-1} \mathbf{D}'_k \mathbf{W}_k\, \mathbf{e}_{(n)j} \,. \qquad (13)$$

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

$$g_{loc,js} = 1 + \frac{1}{w_j} \left\{ \sum\nolimits_U [1 - I(k \in s) w_k]\, \mathbf{e}'_{(2)1} (\mathbf{D}'_k \mathbf{W}_k \mathbf{D}_k)^{-1} \mathbf{D}'_k \mathbf{W}_k \right\} \mathbf{e}_{(n)j} \qquad (14)$$

The Local/f40, Local/s20 and the E estimators are the only estimators studied here that do not depend on the partitioning of the population into model groups. The Local estimators are of the projective bias adjustment form (2). 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 $\boldsymbol{\omega}_k$ within the window. The main difference between this and Briedt's and Opsomer's versions is the 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 the current approach is design-based. In (7), $\mathbf{y}_s$ 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 the estimator is robust 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 (11) 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

$$\delta_k = \begin{cases} 1 \text{ if } |DFFITS_k| \le 2n^{-0.5} \\ 2n^{-0.5} |DFFITS_k|^{-1} \text{ if } |DFFITS_k| > 2n^{-0.5} \end{cases} \qquad (15)$$

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

$$\tilde{r}_k = \frac{y_k - \mathbf{x}_{kg} \tilde{\boldsymbol{\beta}}_g}{x_k^{3/4}} \qquad (16)$$

RobReg is robust to outliers. However, it is not of form (2). It is not linear and it does not have 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, 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, $E_{\ddot{M}}\left(Z_k \middle| Y_k > 0\right) = \mu_g = \mathbf{x}'_{kg}\boldsymbol{\beta}_g$, $V_{\ddot{M}}\left(Z_k \middle| Y_k > 0\right) = \sigma_g^2$ where $\mathbf{x}'_{kg} = I\left(k \in g\right)\left(1 \quad x_{2k}\right)$, 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 different model groups are allowed but not heteroscedasticity in the logscale. Not to burden the notation, subscript *g* is suppressed from now. Let **X** be the matrix with $\mathbf{x}'_k$ in the *k*th 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 $\mathbf{X}\left(\mathbf{X}'_{s+}\mathbf{X}_{s+}\right)^{-1}\mathbf{X}'$, which, incidentally, 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 $\exp\left(\hat{Z}_k\right)$ follows a lognormal distribution. Then $E_{\ddot{M}}\left[\exp\left(\hat{Z}_k\right)\right] = \exp\left(\mu + a_{kk}\sigma^2/2\right)$,
(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 $\exp\left(\hat{Z}_k\right)$ 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 $E_{\ddot{M}}\left(Y_k\right) = \exp\left(\mu + \sigma^2/2\right)$, Karlberg derives an approximately model-unbiased predictor:

$$\hat{Y}_k = \exp\left(\hat{Z}_k\right)\exp\left[\frac{\hat{\sigma}^2}{2}\left(1 - a_{kk}\right) - \frac{\hat{\sigma}^4}{4n_+}\right]. \qquad (\ 17\ )$$

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

$$\hat{\sigma}^2 = \frac{\mathbf{Z}'_{s+}\mathbf{Z}_{s+} - \hat{\beta}'\mathbf{X}'_{s+}\mathbf{X}_{s+}\hat{\beta}}{n_+ - 2} = \sum_{k=1}^{n_+} e_k^2 \Big/ \left(n_+ - 2\right), \qquad (\ 18\ )$$

with e_k being the residuals on the logscale. If $n_+ \le 2$, then the denominator of (18) is set to 1; this happened only once in the simulations. The *Logn/pr* estimate of a total for a model group *g* is

$$\hat{T} = \sum_{k=1}^{N_g - n_g} \hat{p}_{hk}\hat{Y}_k + \sum_{k=1}^{n_g} Y_k \qquad (\ 19\ )$$

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* to which a unit *k* belongs. Alternatively, a logistic model is fitted within each sizeband to obtain an estimated probability $\hat{p}_h\left(x\right)$ 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 (18) 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 (15). 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 (2). 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.

3.2.4 Variance estimators

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

$$\hat{V}_{STSI}\left(\hat{t}_{rat}\right) = \sum_{h=1}^{H}\left(\frac{t_{xg}}{\hat{t}_{xg\pi}}\right)^2\left[N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\sum\nolimits_{s_h}\left(e_k-\bar{e}_h\right)^2\right], \qquad (20)$$

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 within genuine sampling strata model groups, (20) is

$$\hat{V}_{STSI}\left(\hat{t}_{rat}\right) = \left(\frac{t_{x1}}{\hat{t}_{x1\pi}}\right)^2\sum_{h=1}^{3}\left[N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\sum\nolimits_{s_h}\left(e_k-\bar{e}_h\right)^2\right], \qquad (21)$$

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

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

It can also be shown that the g-weighted variance estimator for the group regression model is

$$\hat{V}_{STSI}\left(\hat{t}_{reg}\right) = \sum_{h=1}^{H}\left[N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\sum\nolimits_{s_h} g_{ks}^2\left(e_k-\bar{e}_h\right)^2\right], \qquad (22)$$

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

$$g_{ks} = 1 + \left(\mathbf{t}_{xg} - \hat{\mathbf{t}}_{xg\pi}\right)'\left(\sum\nolimits_s w_j \mathbf{x}_{jg}\mathbf{x}_{jg}'/\sigma_j^2\right)^{-1}\left(\mathbf{x}_{kg}/\sigma_k^2\right). \qquad (23)$$

The variance estimator used here for Local is

$$\hat{V}_{STSI}\left(\hat{t}_{yloc}\right)=\sum_{h=1}^{H}N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{S_h^2}{n_h}, \qquad (24)$$

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

Breidt and Opsomer (2000) show that (24) is for a fixed bandwidth a consistent estimator of an approximate variance

$$AV\left(\hat{t}_{yloc}\right)=\sum\sum_U\Delta_{kl}\,\Gamma_k\Gamma_l/\pi_k\pi_l\ , \qquad (25)$$

where $\Gamma_k=y_k-m_k$, m_k being the smoothed values one would get with (7) 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 (25) has the same form as the usual approximate variance of the GREG. The local g-weights (14) could be inserted into (24). The estimator (24) 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.

4. 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 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 2 to 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

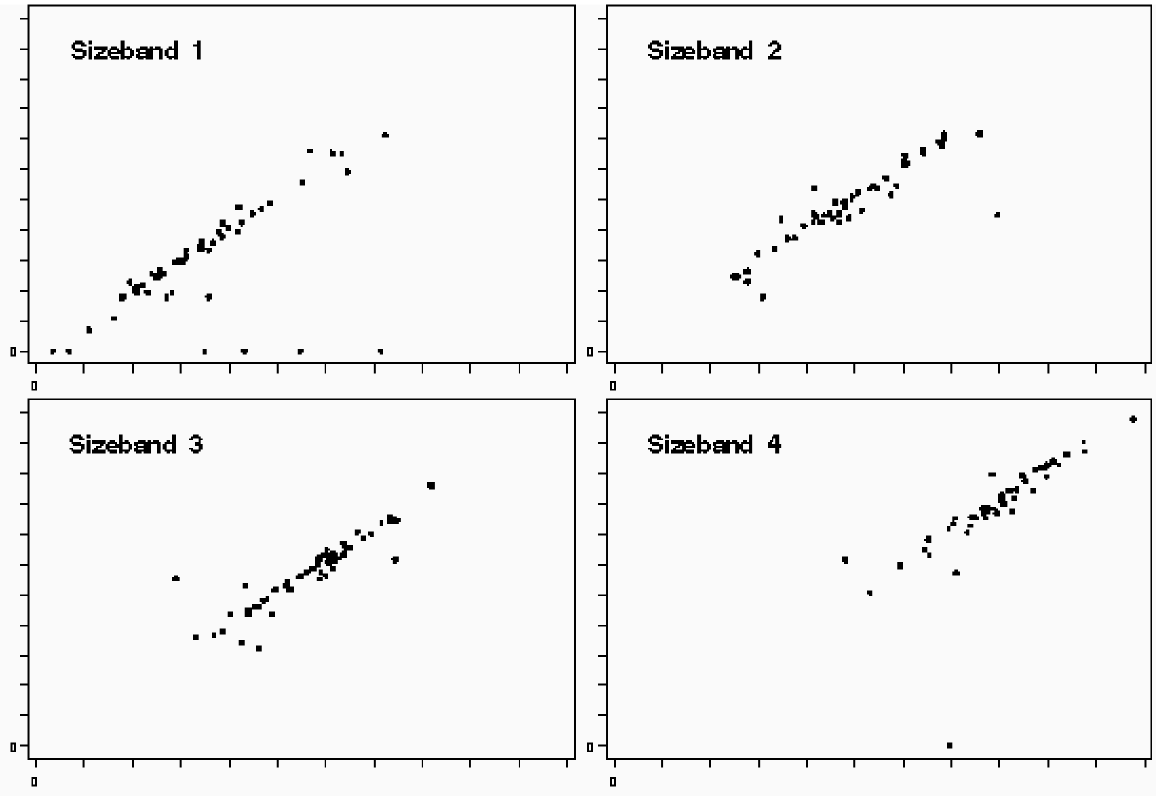


Figure 2. MIDSS, domain A. Log of the study variable against log of the auxiliary variable, with unity added to both variables

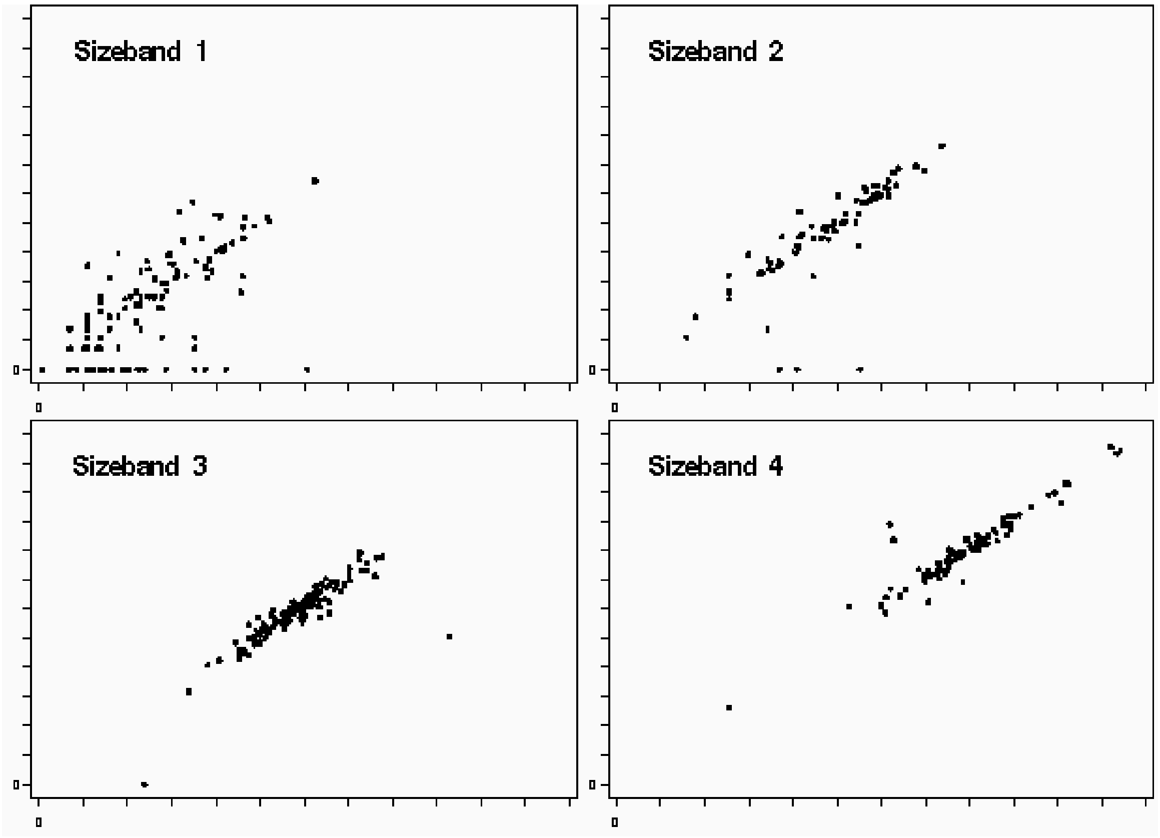


Figure 3. MIDSS, domain B. Log of the study variable against log of the auxiliary variable, with unity added to both variables

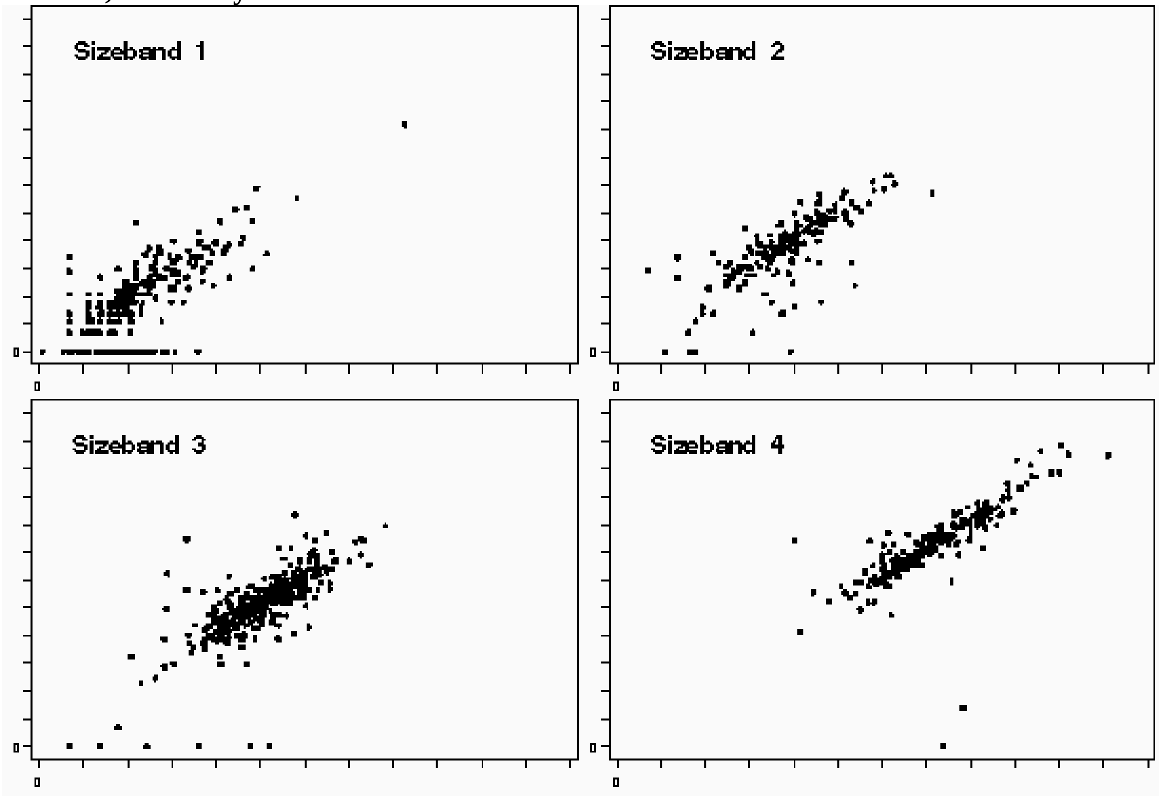


Figure 4. MIDSS, domain C. Log of the study variable against log of the auxiliary variable, with unity added to both variables

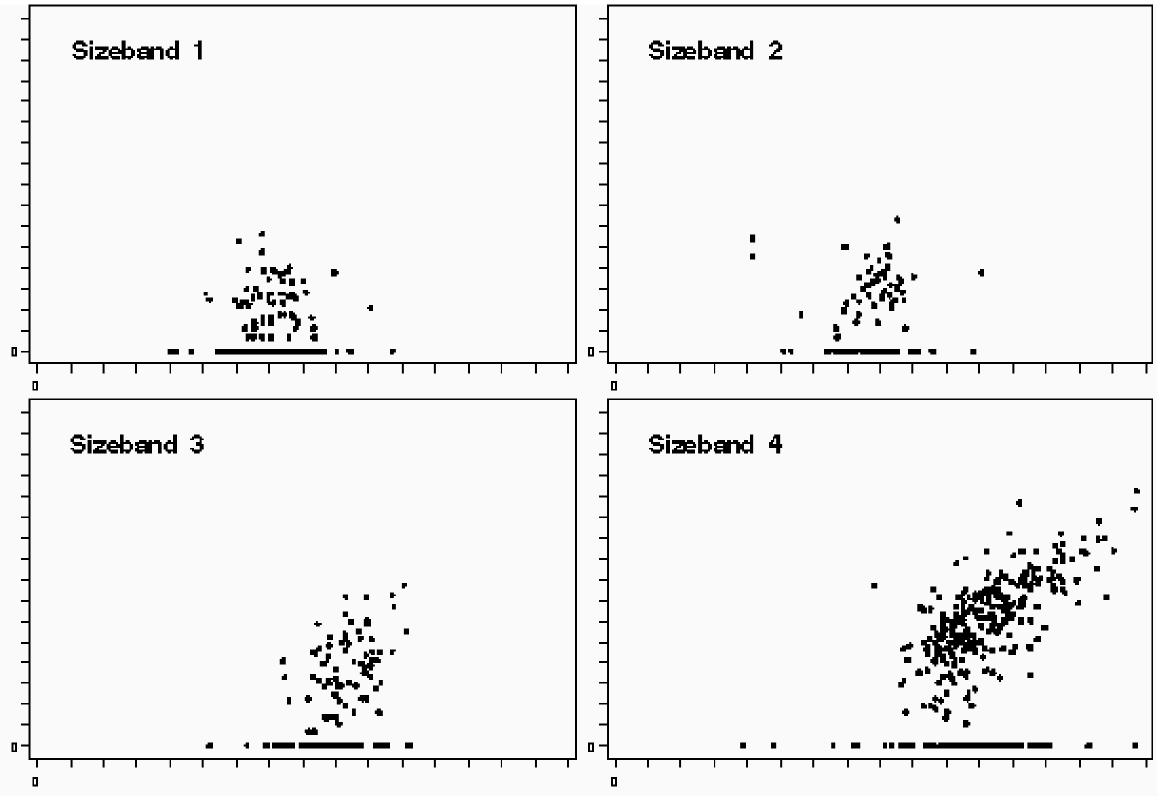


Figure 5. CAPEX, domain U. Log of the study variable against log of the auxiliary variable, with unity added to both variables

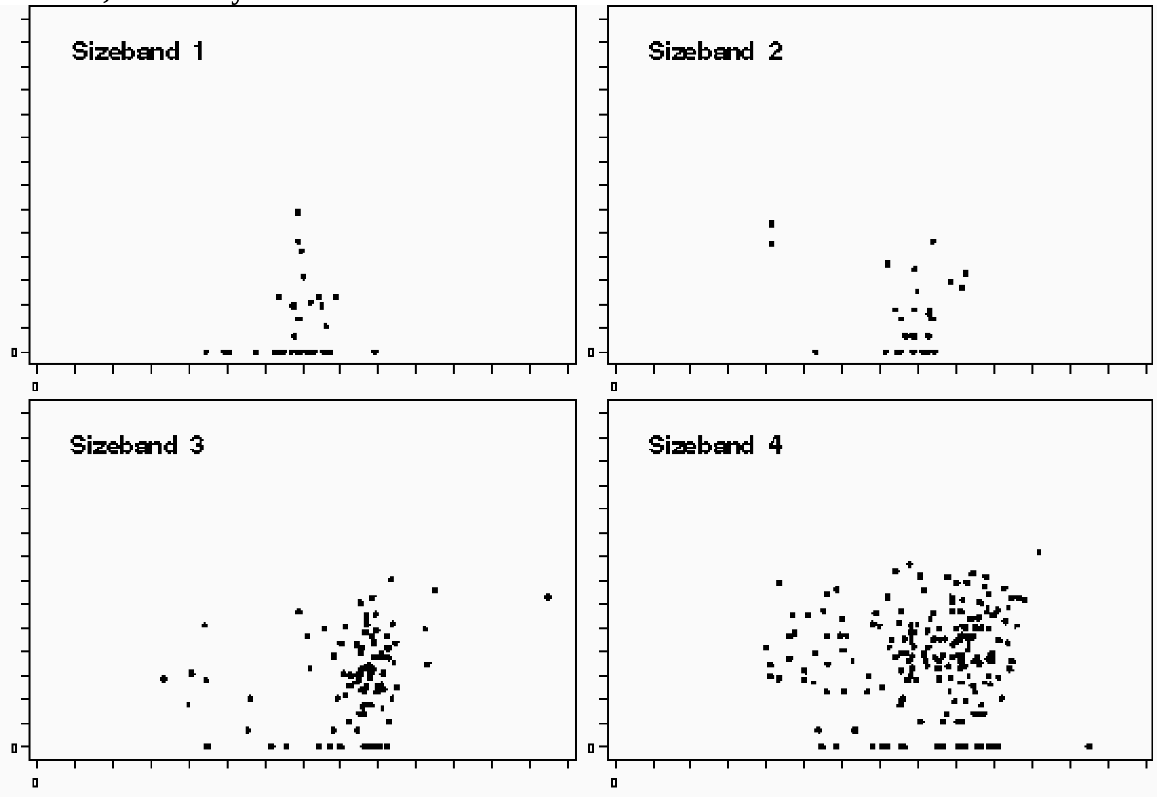


Figure 6. CAPEX, domain V. Log of the study variable against log of the auxiliary variable, with unity added to both variables

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.

In the simulations below the existing strata have been used but the sample is re-allocated 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 1 to 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 1. Sample sizes for the simulated samples, MIDSS domain A

Sizeband	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 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 3. Sample sizes for the simulated samples, MIDSS domain C

Sizeband	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 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. Sample sizes for the simulated samples, CAPEX domain V

Sizeband	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

4.1 Properties of an estimator

Consider 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 $\pm$ 1.96 times the square root of the variance estimates in Section 3.2.4.
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. Turning to the fourth measure, 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.

4.2 Simulation results

Tables 6 to 10 report on the CV and the other measures for five domains. Table 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 ‘within genuine sampling strata’ and 3 for ‘over all strata’. For example, as seen in Table 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 the Appendix.

Table 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	13.46
Rat_2	1.74	1.29	1.16	1.15	13.6
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	43.85
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

Comments to Table 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 use only 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 3) or through overall weak association (domains U and V, Figures 5 and 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, 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 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	33.2
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	33.47
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 7:

1. The bias can be very large for weak-association populations with extreme-leverage points, such as domain V displayed in Figure 3. 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; some analyses not shown here do not indicate a poor fit to the lognormal model. 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 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	83.6	73.6
Rat_2	80.6	63.1	91.2	89.2	68.8
Rat_3	79.9	64.7	93.2	83.9	31.7
Reg/1.0_1	72.7	63.7	91.1	89.6	84.8
Reg/1.0_2	80.6	63.1	91.1	92.1	86.3
Reg/1.0_3	80.0	64.7	93.3	83.9	90.2
Reg/1.5_1	80.6	63.3	91.2	92.5	90.1
Reg/1.5_2	81.1	63.5	91.9	96.3	83.0
Reg/1.5_3	80.2	64.7	93.1	88.0	87.4
Local/f40	79.4	64.7	93.7	83.3	79.4
Local/s20	78.9	64.7	94.4	83.2	73.5
RobReg/f40_1	79.4	64.6	92.9	84.2	53.2
RobReg/f40_2	80.0	64.7	93.8	81.7	36.9
RobReg/f40_3	80.2	64.7	93.4	83.5	63.7

Comments to Table 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 2.
2. If the population is linear (such as the MIDSS domains, Figures 2 to 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 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 9:

Percentages far away from 50 in Table 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 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	23.7
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 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 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	-13.7	-9.3	-29.9
Rat_3	-1.4	0.2	-3.8	-6.1	-22.1
Reg/1.0_1	-53.8	-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	-3.4	-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	-3.8	-1.1	2.1
Local/f40	1.3	1.8	-7.3	3.3	4.3
Local/s20	3.6	0.8	14.1	2.2	1.5
RobReg/f40_1	1.3	-16.1	7.0	8.2	-3.3
RobReg/f40_2	3.5	-0.7	-8.1	1.7	-6.8
RobReg/f40_3	7.4	-7.2	-8.4	3.7	-3.0

Note: Variance estimates were computed using formulae in Section 3.2.4

Comments to Table 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 the poor coverage probabilities, it is not the only reason: note the weak correlation between the coverage probabilities in Table 8 and the bias in Table 11. The boxplots in Appendix 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.

4.3 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. Consider a 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 7a-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 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. The distance from the dotted horizontal line, which indicates the true total, and the fitted value gives an impression of the conditional bias.

As seen in Figures 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 7 a), but, interestingly, the other model groups overadjust for the imbalance (Figures 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 7 a-c is that RLogn has smaller unconditional bias.

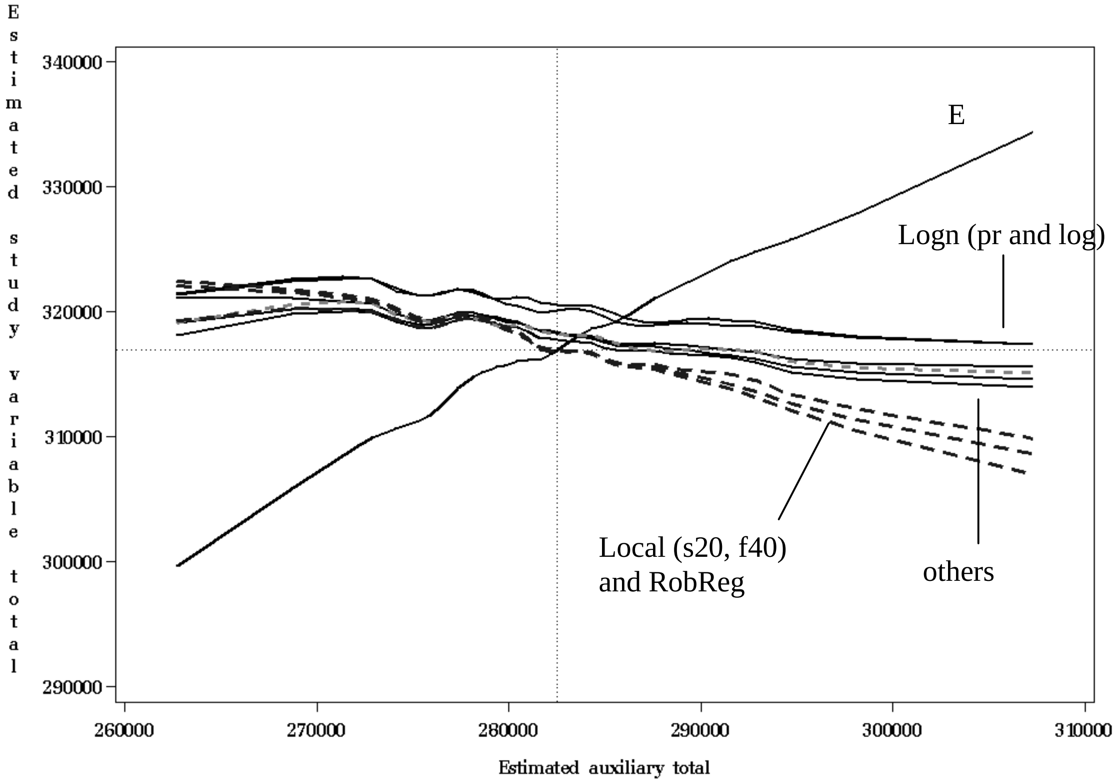


Figure 7 a) Domain A, model groups: within strata

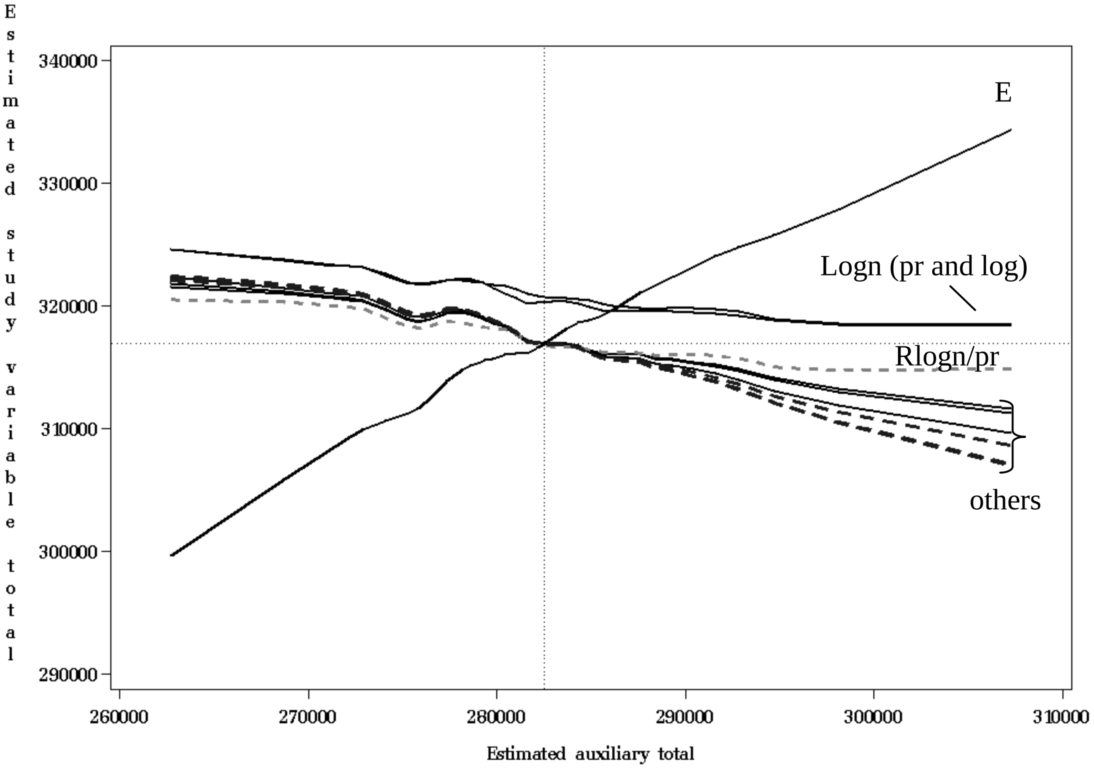


Figure 7 b) Domain A, within genuine sampling strata model groups

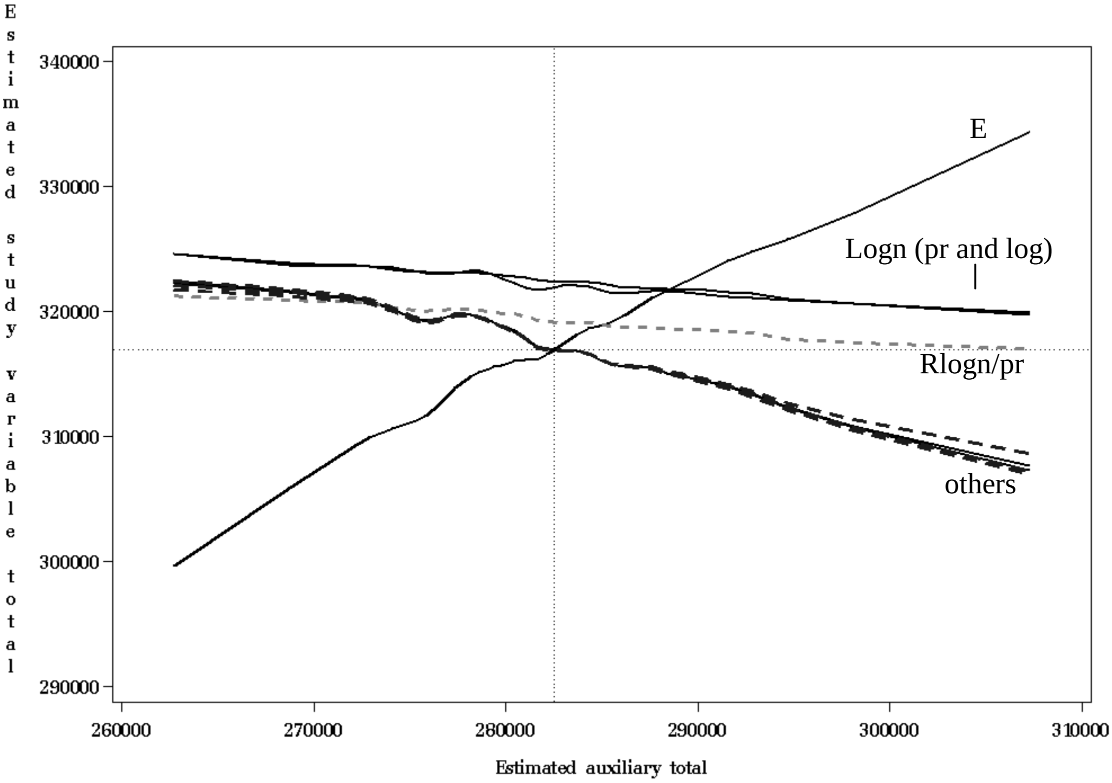


Figure 7 c) Domain A, model group: over all strata.
The estimated total of the study variable against the estimated total of the auxiliary variable. Loess curves indicate the conditional bias; horizontal and vertical lines indicate true totals.

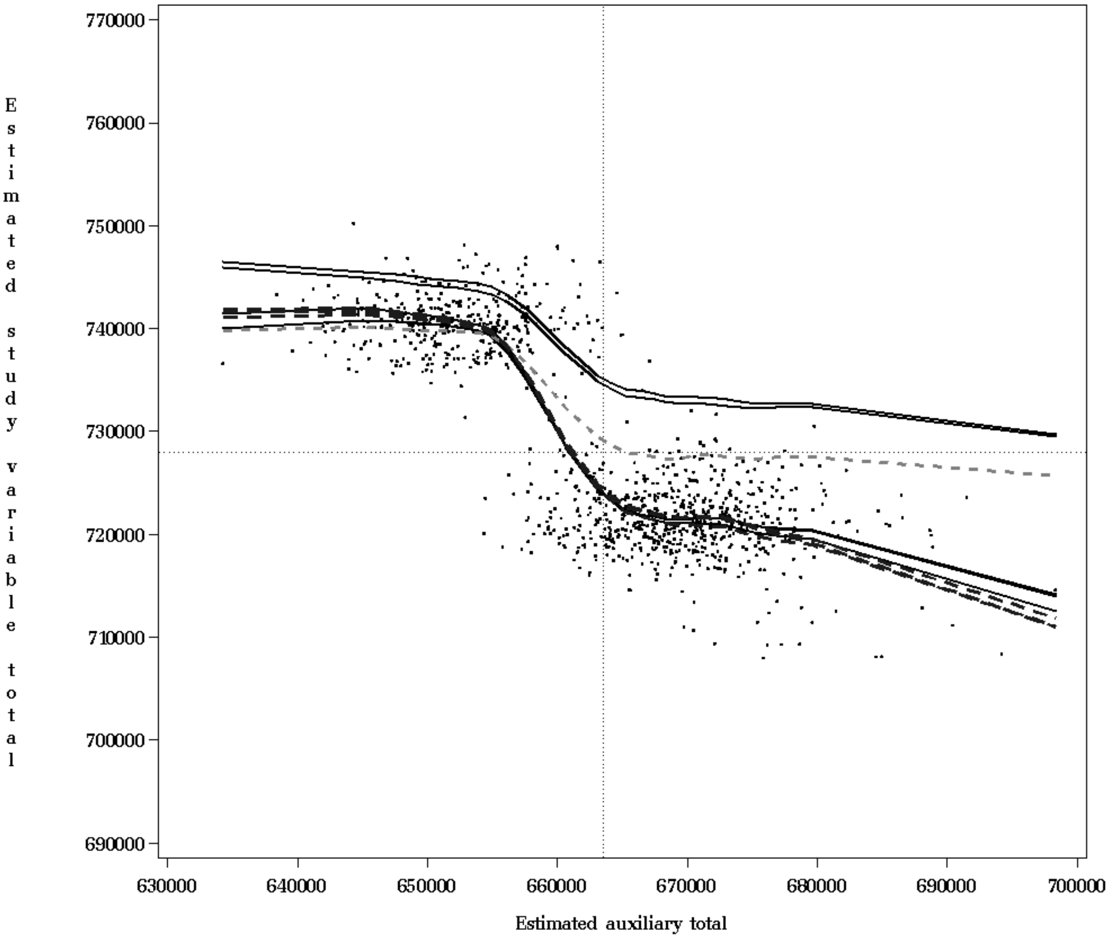


Figure 8. Domain B, 'over all strata' model group. The estimated total of the study variable against the estimated total of the auxiliary variable. Loess curves indicate the conditional bias, with one curve per estimator. They are from top to bottom, Log/pr and log (almost indistinguishable), RLogn/pr (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 8 for domain B. If a sample contains the high-leverage point in sizeband 3 visible in Figure 3, the estimate belongs to the cluster towards the lower-right corner in Figure 8, otherwise it belongs to the other cluster. Furthermore, as can be deduced from Figure 8, the sampling distribution of the estimated study variable total is bimodal. This explains the very poor coverage probabilities in Table 8.

5 Discussion

The GREG is a very flexible and powerful estimator that has enjoyed increasing popularity, in particular since the publication of the book Särndal, Swensson, and Wretman (1992). Yet there are some downsides with the GREG. These are mainly associated with the variability of the g-weights.

A simulation study of estimation in business surveys was conducted, in which some forms of the GREG were contrasted with a local regression estimator and a robust regression estimator. Each estimator was evaluated with three types of model grouping, where relevant, against five criteria. Three of the criteria were 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.

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.
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 3.2.4) 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.

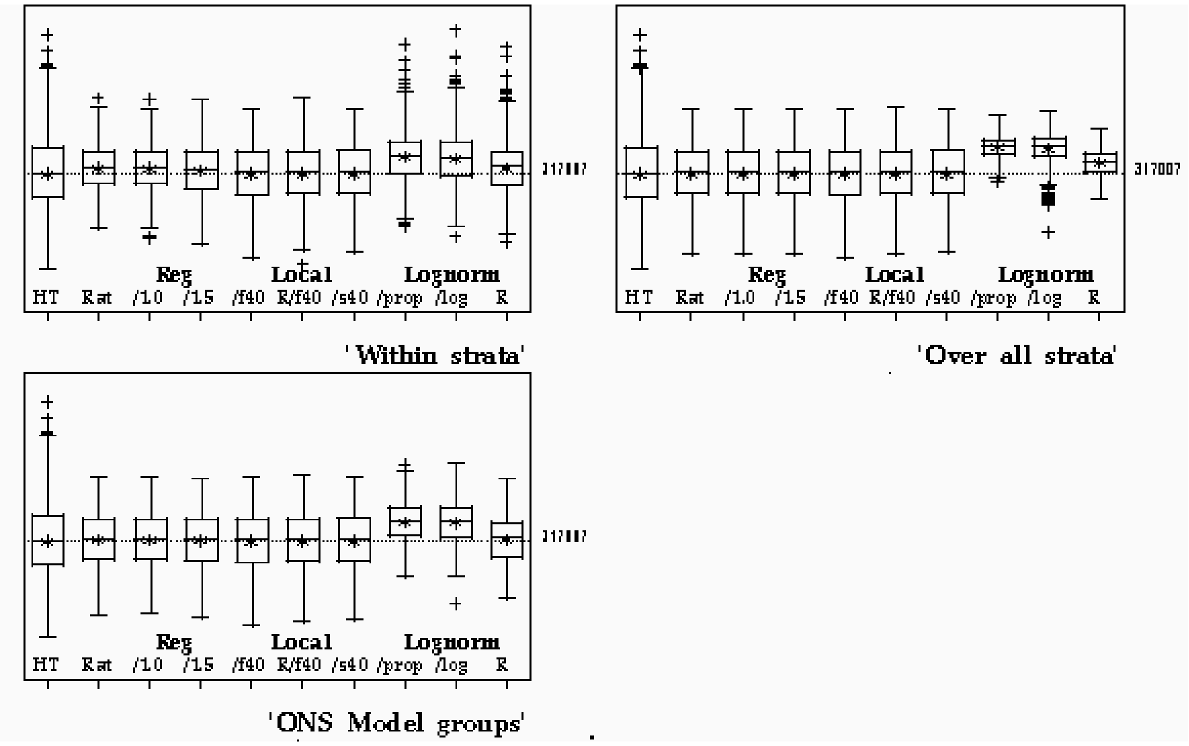
iii. The model-based mixture model estimator is bias prone and will give poor estimates for some populations. Robust estimation of the variance parameter seems to be an approach that reduces some of the problems. Robust estimation of the slope parameter on the logarithmic scale is an option left for 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 (2) and (5). This term is non-zero only for the former estimator.

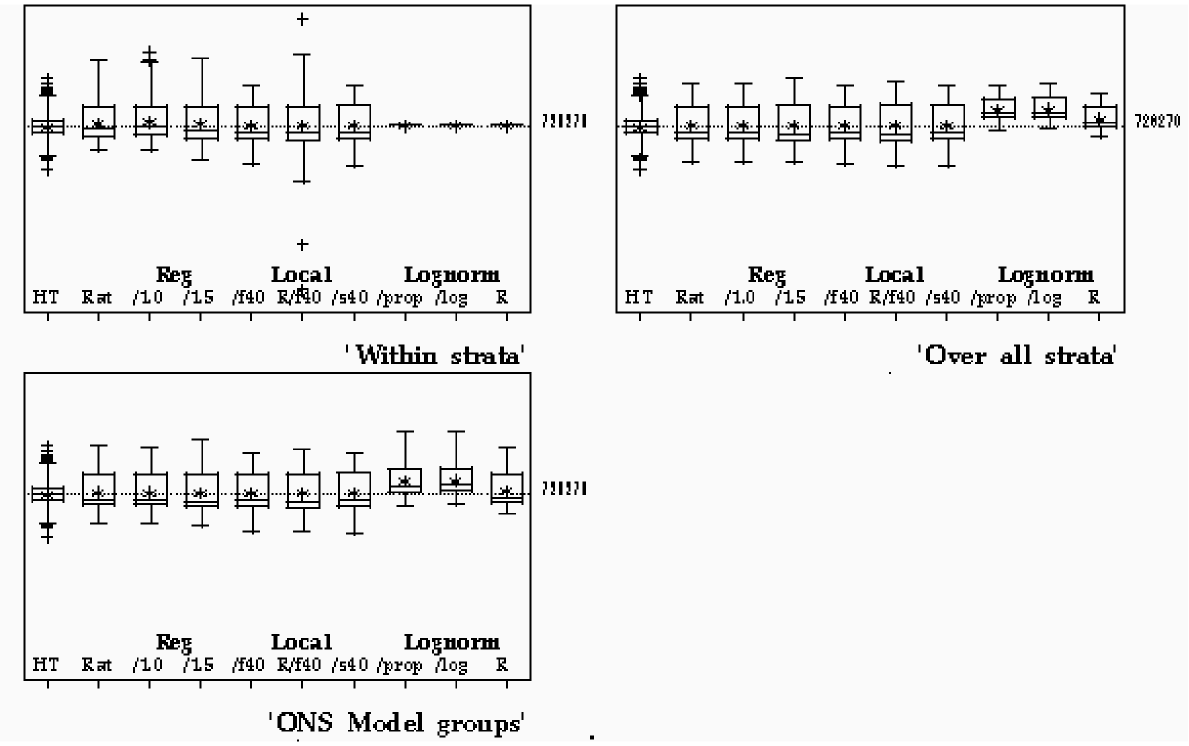
High leverage points need to be addressed. Some other research of the author includes a pre-sample diagnostic, which will assign a value to all units in the sample. For units with particularly large value of the diagnostic, action can be taken before the sample is drawn. They can, for example be moved to a CE stratum, or, if misclassification is 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.

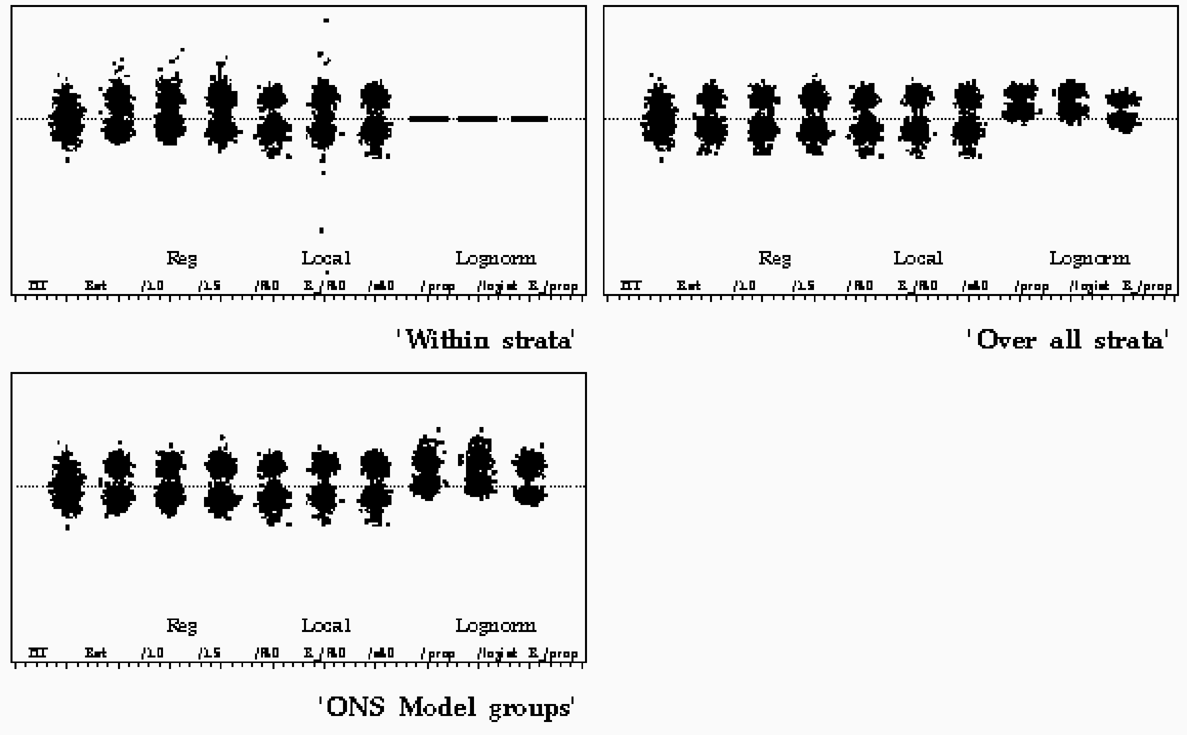
Appendix. Simulation results, boxplots

Figures A.1 and A.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.

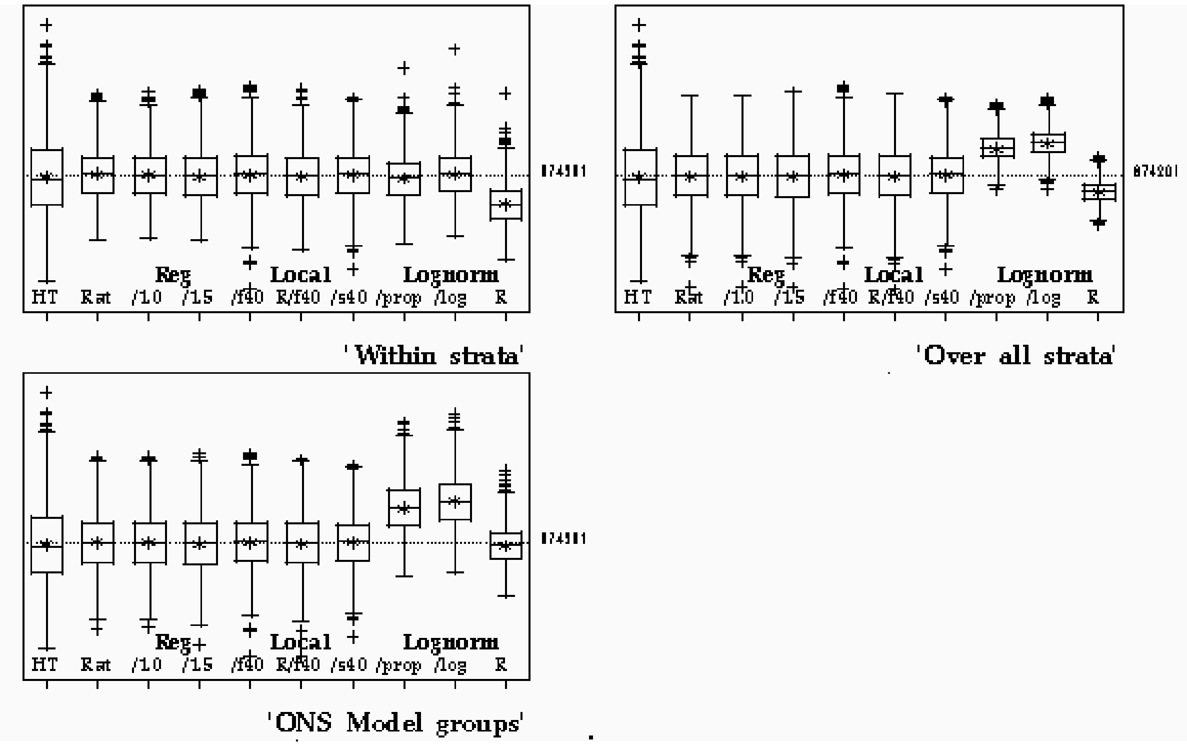


a) MIDSS, domain A.



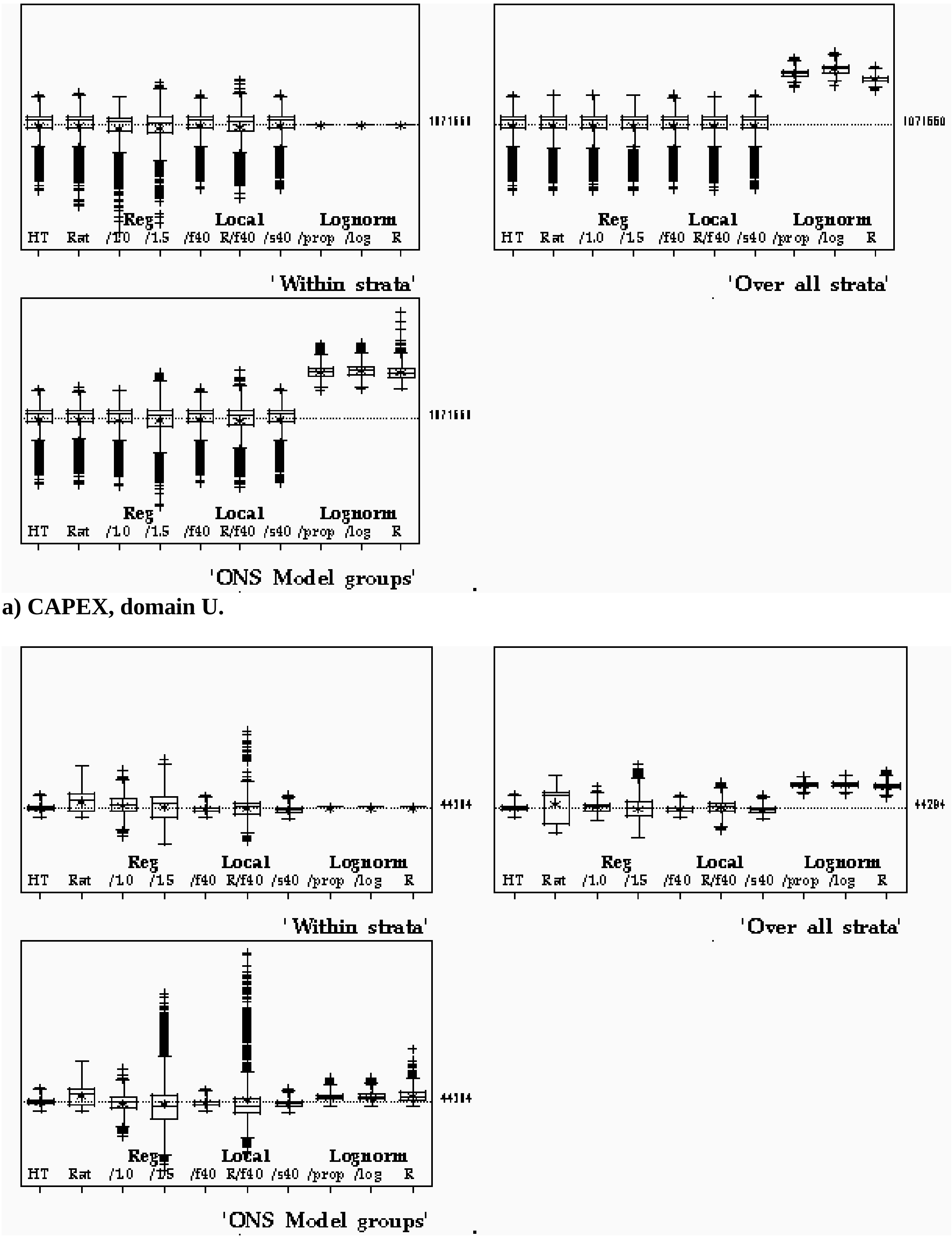


b) MIDSS, domain B. Lognorm 'within strata' have been taken out not to swamp the other box plots. The top set of 3 graphs shows box plots, the bottom set jittered dot plots



c) MIDSS, domain C.

Figure A.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 A.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

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. Statistics Sweden.

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. and Keller, W.J. (1987). Linear Weighting of Sample Survey Data. Journal of Official Statistics, 3, 141-153.

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. (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.

Cassel, C.M., Lundquist, P., and Selén, J. (2002). Model-Based Calibration for Survey Estimation, with an Example from Expenditure Analysis. R&D Report 2002:2, Statistics Sweden.

Cassel, C.M., Särndal, C.-E., and Wretman, J.H.(1976). Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. Biometrika, 63, 615-620.

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.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376-382.

Dorfman, A.H. (2000). Non-Parametric Regression for Estimating Totals in Finite Populations. Proceedings of the Survey Research Methods. American Statistical Association, 47-54.

Fuller, W.A. (2002). Regression Estimation for Survey Sampling. Survey Methodology, 28, 5-23.

Hedlin, D. (2002). Estimating Totals in some UK Business Surveys. Statistics in Transition, 5, 943-968.

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.

Holmberg, A. (2002). On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys. R&D Report 2002:1, Statistics Sweden.

Isaki, C.T. and Fuller, W.A. (1982). Survey Design under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, 89-96.

Karlberg, F. (2000). Survey Estimation for Highly Skewed Populations in the Presence of Zeroes. Journal of Official Statistics, 16, 229-241.

Kokic, P.N. and Bell, P.A. (1994). Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator. Journal of Official Statistics, 10, 419-435.

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.

Loader, C. (1999). Local Regression and Likelihood. New York: Springer-Verlag.

Lundström, S. (2000). Kalibrering av vikter – beskrivning av tekniken och de SCB-fall den prövats i. R&D Report 2000:1, Statistics Sweden.

Lundström, S. and Gustafsson, H. (2003). Estimation vid förekomst av bortfall och rambrister i undersökningen Gymnasieungdomars studieintresse. R&D Report 2003:2, Statistics Sweden.

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.

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.

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. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Sen, A. and Srivastava, M. (1990). Regression Analysis. Theory, Methods and Applications. New York: Springer-Verlag.

Silva, P.L.D.N. and Skinner C.J. (1997). Variable Selection for Regression Estimation in Finite Populations. Survey Methodology, 23, 23-32.

Skinner, C.J. (1999). Calibration Weighting and Nonsampling Errors. Research in Official Statistics, 1, 33-43.

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.

Thompson, M.E. (1997). Theory of Sample Surveys. London: Chapman & Hall.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley.

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.

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.

Förteckning över utkomna R&D Reports

R&D Reports är en för IT-enheten och Metodenheten gemensam publikationsserie, som 1988-01-01 ersatte de tidigare ”gula” och ”gröna” serierna.

Reports published during 2000 and onwards:

2000:1 Kalibrering av vikter – beskrivning av tekniken och de SCB-fall den prövats i *(Sixten Lundström et al)*

2000:2 On Inclusion Probabilities and Estimator Bias for Pareto πps Sampling *(Nibia Aires and Bengt Rosén)*

2000:3 Bortfallsbarometer nr 15 *(Per Nilsson, Ann-Louise Engstrand, Sara Tångdahl, Stefan Berg, Tomas Garås och Arne Holmqvist)*

2000:4 Bortfallsanalys av SCB-undersökningarna HINK och ULF *(Jan Qvist)*

2000:5 Generalized Regression Estimation and Pareto πps *(Bengt Rosén)*

2000:6 A User's Guide to Pareto πps Sampling *(Bengt Rosén)*

2001:1 Det statistiska registersystemet. Utvecklingsmöjligheter och förslag *(SCB, Registerprojektet)*

2001:2 Order πps Inclusion Probabilities Are Asymptotically correct (*Bengt Rosén)*

2002:1 On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys *(Anders Holmberg)*

2002:2 Model-based calibration for survey estimation, with an example from expenditure analysis Surveys *(Claes Cassel, Peter Lundquist and Jan Selén)*

2002:3 On the Choice of Sampling Design in Business Surveys with Several Important Study Variables *(Anders Holmberg, Patrik Flisberg and Mikael Rönnqvist)*

2002:4 The Sampling- and the Estimation Procedure in the Swedish Labour Force Survey *(Hassan Mirza and Jan Hörngren)*

2003:1 Översyn av undersökningen Inrikes och utrikes trafik med svenska lastbilar *(Johan Eriksson, Per Anders Paulson och Bengt Rosén)*

2003:2 Estimation vid förekomst av bortfall och rambrister i undersökningen Gymnasieungdomars studieintresse *(Henrik Gustafsson och Sixten Lundström)*

2004:1 Business Survey Estimation *(Dan Hedlin)*

ISSN 0283-8680

Tidigare utgivna **R&D Reports** kan beställas genom Anna-Lena Carlström, SCB, U/MET, Box 24 300, 104 51 STOCKHOLM (telefon 08-506 943 03, fax 08-506 945 99, e-post annalena.carlström@scb.se).
R&D Reports from 1988-1999 can - in case they are still in stock - be ordered from Statistics Sweden, attn. Anna-Lena Carlström, U/MET, Box 24 300, SE-104 51 STOCKHOLM (telephone +46 8 506 943 03, fax +46 8 506 945 99, e-mail annalena.carlström@scb.se).