Journal of Official Statistics, Vol. 17, No. 4, 2001, pp. 527–544

# Does the Model Matter for GREG Estimation? A Business Survey Example

*Dan Hedlin[1], Hannah Falvey[2], Ray Chambers[1], and Philip Kokic[3]*

Although asymptotically design-unbiased, GREG estimators may produce bad estimates. The article examines the behaviour of GREG estimators when the underlying models are 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. We discuss two diagnostics for whether a GREG estimate is reasonable or not. A common justification for the use of GREG estimators is that, being asymptotically design unbiased, they are relatively robust to model choice. However, we show that 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.

*Key words:* Generalised regression estimator; $g$-weight function; outliers; model misspecification

## 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 official statistical organisations. However, the wider use of more sophisticated estimation methods such as generalised regression (GREG) estimation, the growing use of a larger 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 article 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. We will illustrate

[1] Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K. Emails: deh@socsci.soton.ac.uk and R.Chambers@soton.ac.uk
[2] Higher Education Funding Council for England (HEFCE), Coldharbour Lane, Bristol BS16 1QD, U.K.
[3] QANTARIS Gmbh, Hosteatostr. 25, 65929 Höchst, Frankfurt, Germany.

**Acknowledgments:** This work was carried out while Hannah Falvey was at the Office for National Statistics, U.K., and Philip Kokic was at the University of Southampton. The authors thank Bob Watson for providing data and the associate editor and two anonymous referees for many very useful comments.

these methodologies using data from the quarterly survey of capital expenditure (the CAPEX survey) carried out by the Office for National Statistics (ONS) in the U.K.

The Office for National Statistics is responsible for the lion's share of U.K. 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 resources, production time and human factors such as the learning curve of new staff. In this article we work within this context. In particular, we accept the stratification as ''given''; we also accept the correctness of the data values, and the just mentioned requirements for the production process.

The CAPEX survey currently 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 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, Swensson, and Wretman 1992). If the model is relaxed other types of estimator will result. For example, if an intercept is allowed, the resulting estimator

*Table 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 | |

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 homoskedastic; as opposed to a heteroskedastic model in which the variance changes with the auxiliary. Different degrees of heteroskedasticity result in different estimators. All of these estimators are collectively called GREG estimators.

In what follows, we 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 we 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.

When analysing the data in this article 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.

In Section 2 the definition of the GREG estimator and $g$-weights for the GREG estimator are recalled. 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 3 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 we explore these data to reveal particular features that underpin these outcomes. Section 5 offers an explanation of the behaviour of the GREG estimators in the light of this analysis. Section 6 reports on an attempt to get around these problems. In Section 7 the findings of this article are discussed.

## 2. The GREG Estimator

It is convenient to briefly derive some important properties of GREG estimators, and also to introduce the concept of $g$-weight functions. We will draw on the theory of Särndal et al. (1992). Let $y_k$ and $\mathbf{x}_k$ be the study variable and an auxiliary variable, respectively, for unit $k = 1, 2, \ldots N$. The definition of $\mathbf{x}_k$ will depend on the model; for example $\mathbf{x}_k' = (1\, x_{1k}\, x_{2k})$ if the model contains an intercept and two auxiliary variables. The GREG estimator of the population total of the $y_k$ is then

$$\hat{t}_{reg} = \sum_{k \in \text{population}} \hat{y}_k + \sum_{k \in \text{sample}} a_k e_k \tag{1}$$

where $\hat{y}_k = \mathbf{x}_k' \hat{\beta}$ is the predicted value under the model, $e_k$ is the residual and $a_k$ is the sample weight for unit $k$ (i.e., its inverse inclusion probability). Here

$$\hat{\beta} = \left( \sum_{k \in \text{sample}} a_k \mathbf{x}_k \mathbf{x}_k' / z_k^{\gamma} \right)^{-1} \left( \sum_{k \in \text{sample}} a_k \mathbf{x}_k \, y_k / z_k^{\gamma} \right)$$

is the GREG estimate of the regression parameter $\beta$, where $z_k$ is the value of a scalar auxiliary variable that determines the heteroskedasticity in the regression of $y_k$ on $\mathbf{x}_k$.

The weighted sum of the residuals in (1) is necessarily zero under all models we consider, except some of those involving heteroskedasticity. 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). Alternatively, it can be viewed as a nonparametric adjustment for bias caused by potential model misspecification in the first term on the right hand side of (1) (Chambers, Dorfman and Wehrly 1993). Thus, the size of the absolute value of this sum of residuals relative to the size of the first ''model-based'' term in (1) can be thought of as a measure of model misspecification for the GREG. As we shall see below, this is an important diagnostic for whether a GREG estimate is reasonable or not.

The following form of $\hat{t}_{reg}$ is equivalent to (1):

$$\hat{t}_{reg} = \sum_{k \in \text{sample}} a_k g_{ks} y_k \tag{2}$$

where $g_{ks}$ is the $g$-weight for unit $k = 1, 2, \ldots n$, defined as

$$g_{ks} = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{ax})' \left( \sum_{j \in \text{sample}} a_j \mathbf{x}_j \mathbf{x}_j' / z_j^{\gamma} \right)^{-1} (\mathbf{x}_k / z_k^{\gamma}) \tag{3}$$

The index $s$ in $g_{ks}$ reminds us that the $g$-weights are sample dependent. In (3), $\mathbf{t}_x$ is a vector whose elements are the population totals of the auxiliary variables defining $\mathbf{x}_k$ and $\hat{\mathbf{t}}_{ax}$ 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

$$\hat{\mathbf{t}}_{ax}' = (N \;\; N\bar{x}_s) \tag{4}$$

where $\bar{x}_s$ is the sample mean of the auxiliary variable.

It is generally not considered desirable to have $g$-weights that are far from unity. One reason is that large $g$-weights tend to increase the estimator variance. Negative $g$-weights are particularly undesirable, as they may lead to negative estimates for quantities that are intrinsically positive (Chambers 1996). Another, more philosophical reason for not wanting nonpositive weights is that in a GREG a sampled element can be seen as representing $u_k = a_k g_k - 1$ nonsampled units in addition to itself. Therefore, methods for restricting $g$-weights to positive values have been developed (see e.g., Deville and Särndal 1992; Chambers 1996; Singh and Mohl 1996).

There is a close relationship between the $g$-weight of a sample unit and its influence on the value of the GREG estimate of β. 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 β when the unit is excluded from the sample data used to estimate β. In the context of the regression model underlying GREG, the DFBETA for unit $k$ is defined as

$$\mathrm{DFBETA}_k = \left( \sum_{j \in \mathrm{sample}} \mathbf{x}_j \mathbf{x}_j' / z_j^{\gamma} \right)^{-1} \left( \frac{\mathbf{x}_k}{z_k^{\gamma}} \right) \left( \frac{e_k}{1 - h_k} \right) \tag{5}$$

where

$$h_k = (\mathbf{x}_k / z_k^{\gamma})' \left( \sum_{j \in \mathrm{sample}} \mathbf{x}_j \mathbf{x}_j' / z_j^{\gamma} \right)^{-1} \left( \frac{\mathbf{x}_k}{z_k^{\gamma}} \right)$$

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 (3) and (5) we see that for designs with all $a_k = a$

$$g_{ks} = 1 + \frac{1}{a} \left( \frac{1 - h_k}{e_k} \right) (\mathbf{t}_x - \hat{\mathbf{t}}_{ax})' \mathrm{DFBETA}_k \tag{6}$$

We will use the DFBETA in Section 5 to identify influential sample units.

## 3. 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 2. All of them are GREG estimators. Note that C/Rat in this table is the estimator currently used in the CAPEX survey (Table 1). Throughout we assume that nonresponse in the survey is ignorable conditional on the stratified design.

The auxiliary variable $X$ defining the separate regression estimators does not need to be scalar. We considered situations where $X$ is bivariate, being made up of the two auxiliaries, register employment and turnover, EMP and TO. Furthermore the heteroskedasticity auxiliary $Z$ is often a component of $X$. We considered both $Z = \mathrm{TO}$ and $Z = \mathrm{EMP}$.

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 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 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 3 shows what happens when this approach is applied within one domain of one

*Table 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 $X$ are calculated by combining Sizebands 2-4 within each domain (see Table 1). |
| S/Rat | The separate ratio estimator based on $X$. |
| S/Reg/0.0 | The separate regression estimator. This is based on fitting a separate homoskedastic linear regression model to the study variable in terms of the auxiliary $X$ within each design stratum (Särndal et al. 1992, sec. 7.8). |
| S/Reg/1.0 | As above, but now based on fitting a separate heteroskedastic linear regression model to the study variable in terms of the auxiliary $X$ within each design stratum, with heteroskedasticity proportional to the unit power of a positive-valued scalar auxiliary variable $Z$. |
| S/Reg/1.5 | As above, but with heteroskedasticity proportional to the 1.5 power of the auxiliary variable $Z$. |
| S/Reg/2.0 | As above, but with heteroskedasticity proportional to the square of the auxiliary variable $Z$. |

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 D* from now on. The columns in this table show the estimate of total net capital expenditure, its estimated variance, the CV (squared 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)

*Table 3. Estimates for Domain D*

| Method | $X$ | $Z$ | Estimate | Variance $\div 10^8$ | CV | Variance ratio percent |
|---|---|---|---|---|---|---|
| S/E | – | – | 120,844 | 2.9 | 0.14 | 100.0 |
| C/Rat/B | TO | TO | 97,618 | 4.3 | 0.21 | 150.5 |
| S/Rat | TO | TO | 99,545 | 4.3 | 0.21 | 148.7 |
| C/Rat/B | EMP | EMP | 117,923 | 2.7 | 0.14 | 93.3 |
| S/Rat | EMP | EMP | 117,253 | 2.9 | 0.14 | 92.0 |
| S/Reg/0.0 | TO | – | 117,619 | 2.7 | 0.14 | 93.8 |
| S/Reg/1.0 | TO | TO | 108,325 | 3.2 | 0.16 | 109.3 |
| S/Reg/1.5 | TO | TO | 96,621 | 8.5 | 0.30 | 295.2 |
| S/Reg/2.0 | TO | TO | 71,449 | 35.3 | 0.83 | 1220.6 |
| S/Reg/0.0 | EMP | – | 118,464 | 2.5 | 0.13 | 87.2 |
| S/Reg/1.0 | EMP | TO | 122,133 | 2.8 | 0.14 | 98.2 |
| S/Reg/1.5 | EMP | TO | 124,956 | 3.1 | 0.14 | 107.5 |
| S/Reg/2.0 | EMP | TO | 126,875 | 3.3 | 0.14 | 114.4 |
| S/Reg/0.0 | TO,EMP | – | 114,660 | 2.3 | 0.13 | 79.7 |
| S/Reg/1.0 | TO,EMP | TO | 109,510 | 3.2 | 0.16 | 111.0 |
| S/Reg/1.5 | TO,EMP | TO | 94,127 | 12.1 | 0.37 | 419.7 |
| S/Reg/2.0 | TO,EMP | TO | 42,386 | 97.6 | 2.33 | 3381.0 |
| S/Reg/0.0 | TO,EMP | – | 114,660 | 2.3 | 0.13 | 79.6 |
| S/Reg/1.0 | TO,EMP | EMP | 115,174 | 2.3 | 0.13 | 79.5 |
| S/Reg/1.5 | TO,EMP | EMP | 115,470 | 2.3 | 0.13 | 79.9 |
| S/Reg/2.0 | TO,EMP | EMP | 115,798 | 2.3 | 0.13 | 80.6 |

were used for these computations. They gave very similar results. All variance estimation makes use of *g*-weights (Särndal et al. 1992, ch. 6).

From the results set out in Table 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 1, both in reality and in this article the attention was focussed on relatively simple univariate GREG estimators. Judging from Table 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 was that within design strata there was a higher correlation between net capital expenditure and TO than between net capital expenditure and EMP.

## 4. Exploration of Model Problems

The inconsistencies in the results obtained in the previous section suggested a more in-depth evaluation of the situation in Domain D. Standard statistical procedures were therefore used to explore the fit of a variety of models for the study variable net capital expenditure in Domain D, 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 Srivasta (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 1, give an impression of the underlying relationship between net capital expenditure and TO in domain D. The boundaries between the three parts shown in Figure 1 correspond to the 67th and 90th percentiles of TO in Domain D. 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.

There is a low correlation (0.12) between net capital expenditure and turnover in Domain D. A model where the only auxiliary information is the number of units in Domain D 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 heteroskedasticity. We estimated the degree of this heteroskedasticity by fitting a model of the form

$$\mathrm{Var}(y_k) \propto x_k^{\gamma} \tag{7}$$

Residuals and predicted values were computed, defined by the OLS fit of net capital expenditure against TO. An estimate of $\gamma$ was then obtained as the slope of the OLS fit

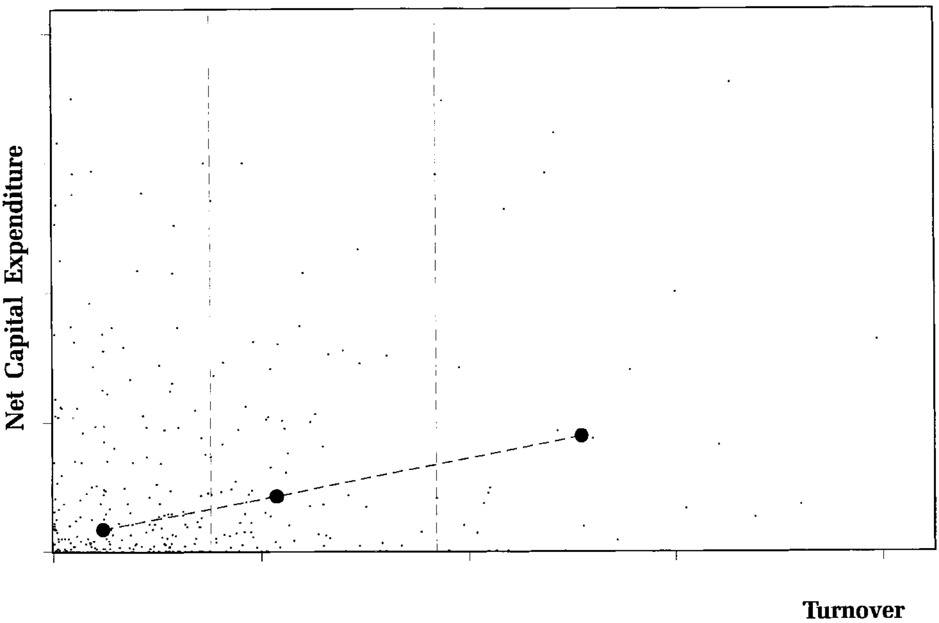

*Fig. 1. Relationship between net capital expenditure and turnover in Domain D. Both axes are truncated*

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.

Standard diagnostics tools revealed that the residuals generated by all the models underlying the estimators in Table 2 were significantly nonnormal. 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 2 to the data from Domain D. 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 D was substantially unbalanced with respect to TO. In particular, the stratified expansion estimate of the total for TO in Domain D was 30 percent larger than the known total of this quantity (taken from the population register). At the population level, over all domains, this estimate was 17 percent larger than the register value.

Following investigation of the anomalous behaviour of the GREG estimators in Table 2, we identified one particular design stratum in Domain D, called *Stratum A* in what follows, as an important contributor to this behaviour. This stratum gave vastly different estimates depending on the assumption of the degree of heteroskedasticity in the model underlying the GREG. In fact, it was this stratum that caused most of the differences between the estimates in Table 3.

To start, we note some basic facts about Stratum A. 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 2. There is one extreme TO value as well as large net capital expenditure values associated with fairly low values of TO.

Figures 3a and 3b show regression lines fitted to the data in Stratum A. The regression models underlying these lines are denoted E (mean model, $E(y) = \beta$), C (linear regression, homoskedastic), X1 (linear regression, $\gamma = 1.0$, see (7)), 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 shows the GREG estimates of total net capital expenditure for Stratum A 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 Office for National Statistics 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. The numbers in brackets in Table 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

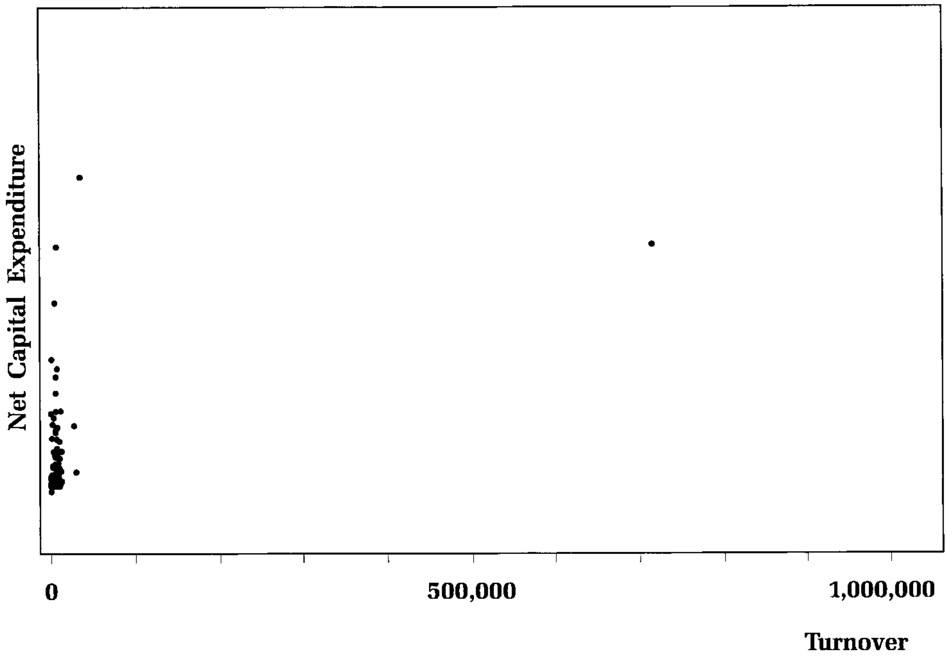

*Fig. 2. Respondents in Stratum A. The scale for turnover is fictitious*

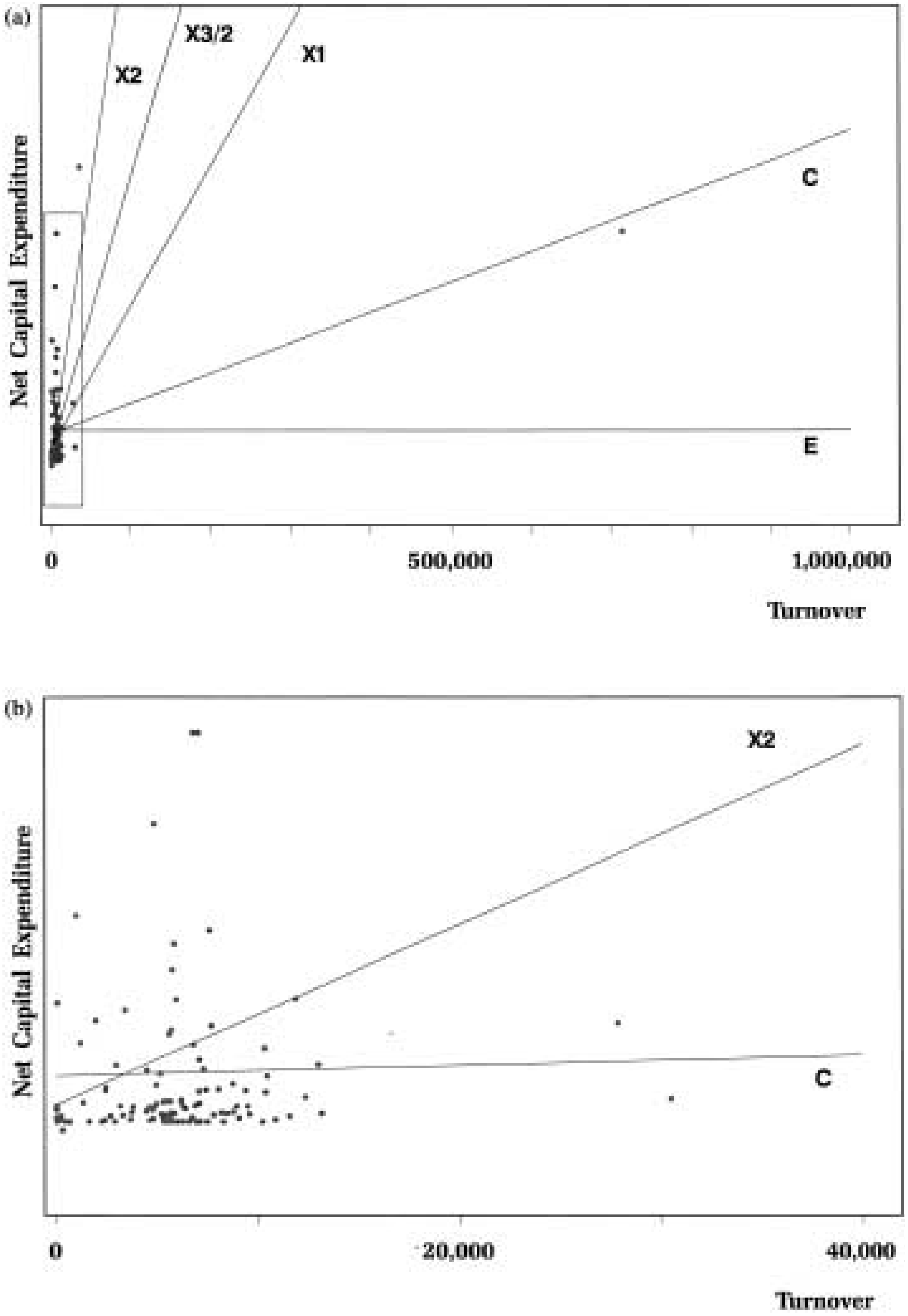

*Fig. 3. a. Regression lines fitted to the data in Figure 2 under the Models E, C, X1, X3/2 and X2. The area within the small rectangle is shown in more detail in Figure 3b. b. Detail of Figure 3a. Models C and X2*

*Table 4. Estimates for Stratum A. TO is the auxiliary variable. The figures in parentheses are obtained when the extreme TO value in Stratum A is replaced by the median TO value for the stratum*

| Estimator | Model | Estimate of total net capital expenditure/ 1,000 | Standard deviation of total estimate/ 1,000 | Intercept | Slope × 1,000 |
|---|---|---|---|---|---|
| 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) |

the estimates based on Models C, X1, X3/2 and X2. Under Model E, of course, the estimates are unchanged. Table 4 shows the parameters of the regression lines 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, however, imply that we get the same estimates. In fact, the estimate of total net capital expenditure was more than 100 percent 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.

## 5. Diagnostics for GREG Estimation

In an effort to explain why such widely different estimates were obtained in Table 4, we computed the $g$-weights (3) generated under the different models. Table 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 Models 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.

For a given sample we can view the $g$-weights as a function of $x = x_k$. It is straight-

*Table 5. Distribution of g-weights for Stratum A (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 | 1 | 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 |

*Table 6. The g-weight functions under simple random sampling*

| Model | $g$-weight function |
|---|---|
| C | $1 + \mathrm{A}_1 - \mathrm{A}_2 x$ |
| X1 | $1 - \mathrm{B}_1 + \mathrm{B}_2/x$ |
| X3/2 | $1 - \mathrm{C}_1/\sqrt{x} + \mathrm{C}_2/(x\sqrt{x})$ |
| X2 | $1 - \mathrm{D}_1/x + \mathrm{D}_2/x^2$ |

forward to show that for Models C, X1, X3/2, and X2 under simple random sampling these functions are as in Table 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. The general form of the constants $\mathrm{B}_1$, $\mathrm{C}_1$ and $\mathrm{D}_1$ in Table 6 is

$$\frac{n}{N}\frac{\hat{t}_{ax} - t_x}{\sum\limits_{k \in \text{sample}} (x_k - \tilde{x}^{(\gamma)})^2/z_k^\gamma}$$

and $B_2 = B_1\tilde{x}_s^{(1)}$, $C_2 = C_1\tilde{x}_s^{(1.5)}$ 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:

$$\tilde{x}_s^{(\gamma)} = \frac{\sum\limits_{k \in \text{sample}} x_k/z_k^\gamma}{\sum\limits_{k \in \text{sample}} 1/z_k^\gamma}$$

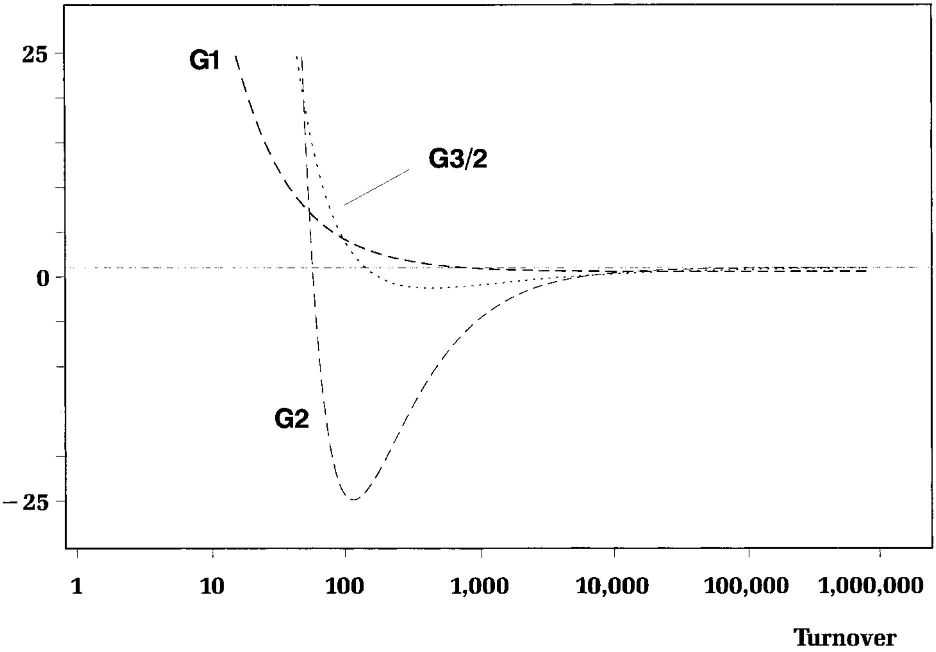

*Fig. 4. The g-weight functions G1, G3/2 and G2 for Stratum A, as a function of turnover. The horizontal line represents Model E*

*Table 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 $2\tilde{x}_s^{(2)}$ |

In Stratum A 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.

As can be seen in Figure 4 and Table 6, the $g$-weights show undesirable behaviour for some ranges of $x_k$. These ranges are shown in Table 7.

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

$$-\mathrm{A}_2 = -\frac{n}{N}\frac{\hat{t}_{ax} - t_x}{\sum\limits_{k \in \text{sample}} (x_k - \bar{x}_s)^2}$$

Suppose the sample is overbalanced, as is the case in Stratum A. Then $\hat{t}_{ax} - t_x$ is large and positive. In the case of Stratum A 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

$$(n-1)^{-1} \sum_{k \in \text{sample}} (x_k - \bar{x}_s)^2$$

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 5.

The function G1 converges to $1 - \mathrm{B}_1$ as $x$ increases, where $\mathrm{B}_1$ can be large (it is 0.45 in Stratum A). 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 ‘‘$a$-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 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 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 Stratum A, Figure 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

$$\hat{V}(\hat{t}_{\text{reg}}) = K \sum_{k \in \text{sample}} \left(g_{ks}e_{ks} - \overline{g_s e_s}\right)^2 \tag{8}$$

where $K$ is a constant (Särndal et al. 1992, Chapter 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, Stratum A, it is not surprising that the variance estimates in Table 4 for the estimators S/Reg/1.5 and S/Reg/2.0 which set $Z = \text{TO}$ and $X = \text{TO}$ were extremely large.

Table 8a shows the residuals obtained in Stratum A. 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 the lowest net capital expenditure. As can be seen in the last two columns of Table 8a, the sums of residuals under Models X3/2 and X2 are very large compared to either the corresponding sum of predicted values or the estimates of total net capital expenditure (Table 4). Recall the role of the weighted sum of the residuals in (1). These large residual sums indicate poor model fit. 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 for Stratum A 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 3b.

Table 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

*Table 8a. Distribution of residuals in Stratum A*

| Model | Residuals, low–high | Residual of lowest TO | Median | Proportion of positive residuals | Sum of residuals ÷ 1,000 | 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 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 ÷ 1,000 | 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 |

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).

The link between a sample point's DFBETA$_k$ value and its $g$-weight was shown by (6). Here we use this relationship to identify influential points in Stratum A. 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 sampling the first co-ordinate of $\mathbf{t}_x - \hat{\mathbf{t}}_{ax}$ 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 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 Stratum A (the dashed reference lines in Figure 5). As might be expected, the sample units in Stratum A 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.

## 6. Using Poststratification to Minimise the Effect of Influential Points

The idea here is to use the regression estimator S/Reg/1.5 (since we found in Section 4 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 effect 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 Stratum A into three poststrata on the basis of TO. Estimation for Stratum A 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 Stratum A. The regression estimator generated an estimated variance about two thirds of the estimated variance of the expansion estimator for the poststratum. For Stratum A overall the poststratified procedure resulted in an estimated variance that was 78 percent of the expansion estimator for the stratum. For comparison, regression estimation based on other models was conducted in Poststratum 2. As shown in Table 9, the estimates are now very stable over the regression models, as opposed to the estimates given in Table 4.

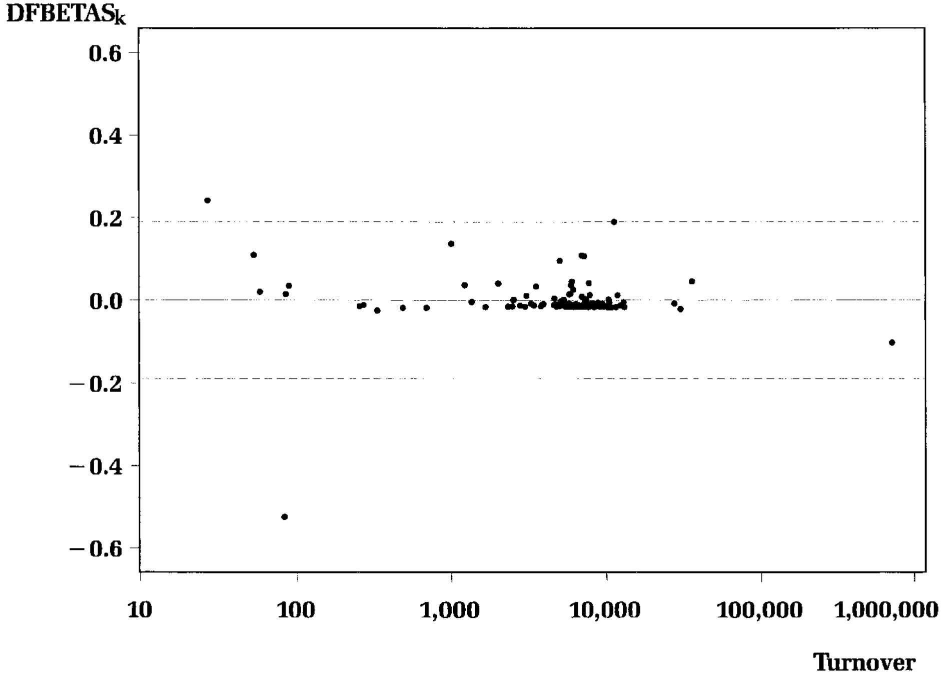

*Fig. 5. Influential points in Stratum A for Model X3/2*

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. We followed 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, we picked exactly 20 units each for the two extreme poststrata.

An alternative to poststratification is GREG estimation based on restricted $g$-weights. We therefore computed restricted versions of S/Reg/1.5 and S/Reg/2.0, for Stratum A 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 we obtained variance ratios of 314 percent and 467 percent, respectively, compared with the simple expansion estimator S/E. Other ranges for the $g$-weights were tried, but either

*Table 9. Poststratified estimates for Stratum A. TO is the auxiliary variable*

| Estimator | Model | Estimate of total net capital expenditure/1,000 | Standard deviation of total estimate/ 1,000 | Intercept in Poststratum 2 | Slope $\times 1,000$ 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 |

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 Stratum A got $g$-weights equal to the lower limit (0.001) for the estimator S/Reg/2.0.

## 7. Discussion

In the context of stratified simple random sampling we 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. We have shown two diagnostics for whether a GREG estimate is reasonable or not. The first diagnostic 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 second diagnostic is the ratio of the absolute value of the sum of sample weighted residuals to the population sum of predicted values (that is, the ratio of the second term to the first term in (1)). The sum of weighted residuals is an adjustment term whose function is to make the GREG asymptotically design-unbiased. This ratio, which should be close to zero, tends to get large in absolute terms when the model is misspecified. For one GREG estimator in our study this ratio was 84 to 87, which is a clear indication of serious model problems.

The $g$-weight of a sample unit is connected to its DFBETA, which is the change in the estimate of $\beta$ when the unit is excluded from the sample data used to estimate $\beta$. In this article we 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 diagnostics 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.

The business survey example of this article 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 percent of the ordinary expansion estimate. However, 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 we therefore reiterate the point made initially (p. 527). 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.

## 8. References

Andersson, C. and Nordberg, L. (1998). A User's Guide to CLAN 97. Statistics Sweden.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

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.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed., New York: Wiley.

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall.

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

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.

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

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

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

Singh, A.C. and Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. Survey Methodology, 22, 107–115.

Wright, R.L. (1983). Finite Population Sampling with Multivariate Auxiliary Information. Journal of the American Statistical Association, 78, 879–884.

Received August 1999
Revised July 2001