Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Cosmetic calibration with unequal probability sampling

by K.R.W. Brewer

December 1999

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 205-212
Statistics Canada, Catalogue No. 12-001

Cosmetic calibration with unequal probability sampling

K.R.W. Brewer [1]

Abstract

Cosmetic estimators are by definition interpretable both as design-based and as prediction-based estimators. Formulae for them can be obtained directly by equating these two estimators or indirectly by a simple form of calibration. Since they constitute a subset of Generalized Regression Estimators, their design-variances cannot be estimated without knowing the relevant second order inclusion probabilities, but under the prediction model to which they are calibrated those probabilities do not affect their anticipated variances, so it is more appropriate to estimate these and/or their prediction-variances instead. An unanticipated spin-off of cosmetic calibration is a simple and effective method for eliminating negative and unacceptably small positive sample weights. The empirical performance of cosmetically calibrated estimators is put to the test using Australian farm data.

Key Words: Anticipated variance; Design-based estimation; Non-negative weights; Prediction-based estimation; Regression estimation.

1. Introduction

Cosmetic estimation was introduced by Särndal and Wright (1984). A cosmetic estimator is one that is readily interpretable both as a design-based and as a prediction-based estimator. A procedure for constructing a cosmetic estimator was suggested in Brewer (1995). An improved version of it is presented and discussed in this paper.

Deville and Särndal (1992) described a number of variants on the theme of calibration. The common thread running though them was that for large samples the sample weights had to approximate the Horvitz-Thompson (HT) weights (Horvitz and Thompson 1952); that is to say, the reciprocals, π_j^{-1}, of the first order inclusion probabilities, π_j. Weighted sums of the differences between the calibration weights and the HT weights were minimized to achieve this end. The simplest of Deville and Särndal's calibration estimators was a particular case of the Generalized Regression Estimator or GREG (Cassel, Särndal and Wretman 1976). It requires only a slight modification to become a cosmetic estimator as well. Such estimators will be described here as cosmetically calibrated.

The special case of cosmetic calibration under a stratified simple random sampling design was treated in Brewer (1999). In this paper we consider the generalization to unequal probability sampling. In section 2 the cosmetic and the calibrated approaches to estimation are outlined and shown to be compatible. The design-variance of the cosmetically calibrated estimator is considered in section 3 and its prediction-variance in section 4. It is shown in section 5 that when using cosmetic calibration it is serendipitously easy to overcome the problem of negative and unacceptably small positive weights. Section 6 contains the results of an empirical study based on a somewhat challenging set of Australian farm data. In section 7 the concept is evaluated.

2. The two approaches to cosmetic calibration

Throughout this paper it will be assumed that the population being sampled can be described to a reasonable approximation by the following regression or prediction model:

$$y_j = \mathbf{x}_j'\boldsymbol{\beta} + \varepsilon_j;\ E_\xi\varepsilon_j = 0,$$
$$E_\xi\varepsilon_j^2 = \sigma^2 a_j^2,\ E_\xi(\varepsilon_j\varepsilon_k) = 0 \forall k \neq j, \qquad (1)$$

where $\mathbf{x}_j$ is a p-vector of explanatory variables for unit j and ε_j is a random error with the properties shown, σ^2 is an unknown scalar and the a_j^2 are assumed known. We will also write diag(a_j) as $\mathbf{A}$. Expressions such as "prediction-unbiased" and "prediction-variance" when used in this paper refer to unbiasedness, variance *etc.* in terms of the model (1). It is not uncommonly assumed that the a_j^2 are proportional to some power of a measure of size, say $z_j^{2\gamma}$, where γ lies between 0.5 and 1. When $\gamma = 1$ the coefficient of variation of ε_j is constant. The value $\gamma = 0.5$ corresponds to the situation where the large units behave like random aggregations of small units. Solving the model for the three cases $\gamma = 0.5$, 0.75 and 1 usually gives a realistic range of variance estimates.

The cosmetic approach requires that there be an estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{COS}}$, such that the standard and the predictor forms (Royall 1970) of the GREG estimator are numerically equal, *i.e.*,

$$\hat{T}_{\mathrm{COS}}(\mathbf{y}) = \mathbf{1}_n'\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s + (\mathbf{1}_N'\mathbf{X} - \mathbf{1}_n'\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s)\,\hat{\boldsymbol{\beta}}_{\mathrm{COS}} \qquad (2)$$

$$= \mathbf{1}_n'\mathbf{y}_s + (\mathbf{1}_N'\mathbf{X} - \mathbf{1}_n'\mathbf{X}_s)\hat{\boldsymbol{\beta}}_{\mathrm{COS}}. \qquad (3)$$

1. K.R.W. Brewer, Department of Statistics and Econometrics, Faculty of Economics and Commerce, Australian National University, ACT 0200, Australia.

where $\mathbf{y}$ and $\mathbf{y}_s$ are population and sample vectors of the y values, $\mathbf{X}$ and $\mathbf{X}_s$ are the full-rank $N \times p$ population and $n \times p$ sample matrices respectively of supplementary variables [so that $\mathbf{1}'_N\mathbf{y} = T(\mathbf{y})$ and $\mathbf{1}'_N\mathbf{X} = T(\mathbf{X})$], $\hat{T}_{\text{COS}}(\mathbf{y})$ is the Cosmetic Estimator of $T(\mathbf{y})$ and $\boldsymbol{\Pi}_s$ is the $n \times n$ diagonal matrix of the sample π_j. It is assumed here that these inclusion probabilities are determined entirely by quantities known to the survey designer from non-sample sources, so that the question of possible informativeness arises only for secondary analysis. $\hat{T}_{\text{COS}}(\mathbf{y})$ also possesses the internally bias-calibrated property defined by Firth and Bennett (1998).

Expressions (2) and (3) are equal when $\mathbf{1}'_n(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}_{\text{COS}}) = 0$. Assuming $\hat{\boldsymbol{\beta}}_{\text{COS}}$ is of the projection form $(\mathbf{Q}'_s\mathbf{X}_s)^{-1}\mathbf{Q}'_s\mathbf{y}_s$, this condition is satisfied when $\mathbf{1}'_n(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)$ spans the row space of $\mathbf{Q}'_s$, for then there must be some row p-vector $\boldsymbol{\alpha}'$ such that $\boldsymbol{\alpha}'\mathbf{Q}'_s = \mathbf{1}'_n(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)$, so $\mathbf{1}'_n(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}_{\text{COS}}) = \boldsymbol{\alpha}'\mathbf{Q}'_s[\mathbf{y}_s - \mathbf{X}_s(\mathbf{Q}'_s\mathbf{X}_s)^{-1}\mathbf{Q}'_s\mathbf{y}_s] = 0$ as required.

Brewer (1995) suggested a way of achieving this result using instrumental variables, but subsequent empirical tests (along the lines of those described in section 6) indicated that this approach was not efficient. It is more efficient, and simpler, to take the Best Linear Unbiased Estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\text{BLUE}} = (\mathbf{X}'_s\mathbf{A}_s^{-2}\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{A}_s^{-2}\mathbf{y}_s$, where $\mathbf{A}_s$ contains only the sample values in $\mathbf{A}$, and replace the $\mathbf{A}_s^{-2}$ factor by $\mathbf{Z}_s^{-1}(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)$ where $\mathbf{Z}_s$ is $n \times n$ diagonal and $\mathbf{Z}_s\mathbf{1}_n = \mathbf{X}_s\boldsymbol{\alpha}$ is any linear combination of the columns of $\mathbf{X}_s$. For then

$$\hat{\boldsymbol{\beta}}_{\text{COS}} = [\mathbf{X}'_s\mathbf{Z}_s^{-1}(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{X}_s]^{-1}\mathbf{X}'_s\mathbf{Z}_s^{-1}(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{y}_s, \quad (4)$$

which is of the required projection form with $\mathbf{Q}'_s = \mathbf{X}'_s\mathbf{Z}_s^{-1}(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)$. Also, since $\boldsymbol{\alpha}' = \mathbf{1}'_n\mathbf{Z}'_s\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}$, $\boldsymbol{\alpha}'\mathbf{Q}'_s = \mathbf{1}'_n\mathbf{Z}'_s\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{Z}_s^{-1}(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)$. But $\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{Z}_s\mathbf{1}_n = \mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{X}_s\boldsymbol{\alpha} = \mathbf{X}_s\boldsymbol{\alpha} = \mathbf{Z}_s\mathbf{1}_n$, so $\mathbf{1}'_n\mathbf{Z}'_s\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s = \mathbf{1}'_n\mathbf{Z}'_s$ and $\boldsymbol{\alpha}'\mathbf{Q}'_s = \mathbf{1}'_n(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)$ as required.

The choice of $\mathbf{Z}_s$ is still somewhat arbitrary but, if we aim for the $(\pi_j^{-1} - 1)z_j^{-1}$ to be as closely proportional to the a_j^{-2} as possible, $\hat{\boldsymbol{\beta}}_{\text{COS}}$ can approximate $\hat{\boldsymbol{\beta}}_{\text{BLUE}}$. One case is of particular interest here. If (i) the π_j are chosen to be proportional to the a_j, (aiming to minimize the design-variance of the GREG), (ii) they are all small compared with unity, so that $\pi_j^{-1} - 1 \approx \pi_j^{-1}$, and (iii) the a_j themselves are proportional to the elements $\breve{z}_j$ of $\breve{\mathbf{z}}$, a linear combination of the columns of $\mathbf{X}$, then the choice $z_j = \breve{z}_j$ will achieve the desired close proportionality.

An alternative way to derive the estimator $\hat{\boldsymbol{\beta}}_{\text{COS}}$ is to use the calibration approach described by Deville and Särndal (1992), in which sample weights are made as "close" as possible to the π_j^{-1}, subject to the condition that, for every variable in the columns of $\mathbf{X}$, the sample estimate defined by these weights should be without error. The "closeness" is defined by an arbitrary distance function, but for our present purposes the appropriate function is

$$D = (\mathbf{w}_s - \boldsymbol{\varpi}_s)'[(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}]^{-1}(\mathbf{w}_s - \boldsymbol{\varpi}_s) + 2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{1}_N - \mathbf{X}'_s\mathbf{w}_s),$$

where $\mathbf{w}_s$ is the n-vector of the sample weights w_j, $\boldsymbol{\varpi}_s = \boldsymbol{\Pi}_s^{-1}\mathbf{1}_n$ is the n-vector of the inverse inclusion probabilities π_j^{-1} and $\boldsymbol{\lambda}'$ is a $1 \times p$ row vector of undetermined multipliers. [This is the same as the first variant of Calibration Estimation used in Deville and Särndal (1992), except that $\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_s^{-1}$ is replaced by $(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\,\mathbf{Z}_s^{-1}$.] Differentiating with respect to $\mathbf{w}_s$,

$$\frac{\partial D}{\partial \mathbf{w}_s} = 2[(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}]^{-1}(\mathbf{w}_s - \boldsymbol{\Pi}_s^{-1}\mathbf{1}_n) - 2\mathbf{X}_s\boldsymbol{\lambda}.$$

Solving $\dfrac{\partial D}{\partial \mathbf{w}_s} = 0$ yields

$$\mathbf{w}_s = \boldsymbol{\Pi}_s^{-1}\mathbf{1}_n + (\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}\mathbf{X}_s[\mathbf{X}'_s(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\,\mathbf{Z}_s^{-1}\mathbf{X}_s]^{-1} \times (\mathbf{X}'\mathbf{1}_N - \mathbf{X}'_s\boldsymbol{\Pi}_s^{-1}\mathbf{1}_n), \quad (5)$$

and the corresponding Calibration Estimator, defined as $\mathbf{w}'_s\mathbf{y}_s$, reduces to the formula given for $\hat{T}_{\text{COS}}(\mathbf{y})$ in its GREG form, shown in (2) above. Since (2) and (3) are equivalent, there is also an alternative formula for $\mathbf{w}_s$, namely

$$\mathbf{w}_s = \mathbf{1}_n + (\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}\mathbf{X}_s[\mathbf{X}'_s(\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}\mathbf{X}_s]^{-1} \times (\mathbf{X}'\mathbf{1}_N - \mathbf{X}'_s\mathbf{1}_n). \quad (6)$$

$\hat{T}_{\text{COS}}(\mathbf{y})$ being the intersection of Särndal and Wright's (1984) Cosmetic Estimators with Deville and Särndal's (1992) Calibration Estimators, we will refer to it from now on as the Cosmetic Calibration Estimator and will write it as $\hat{T}_{\text{COSCAL}}(\mathbf{y})$. Similarly we will write $\hat{\boldsymbol{\beta}}_{\text{COS}}$ as $\hat{\boldsymbol{\beta}}_{\text{COSCAL}}$.

3. Design-variance and anticipated variance

We consider first the design-variance of $\hat{T}_{\text{HT}}(\mathbf{y})$ and also that of any GREG estimator that is prediction-unbiased under the model (1). Such an estimator can be written $\hat{T}_{\text{GREG}}(\mathbf{y}) = \hat{T}_{\text{HT}}(\mathbf{y}) + \{T(\mathbf{X}) - \hat{T}_{\text{HT}}(\mathbf{X})\}\hat{\boldsymbol{\beta}}_{\text{GREG}}$ where $\hat{\boldsymbol{\beta}}_{\text{GREG}}$ is any prediction-unbiased and prediction-consistent estimator of $\boldsymbol{\beta}$. If the sample size is fixed at n, the design-variance of $\hat{T}_{\text{HT}}(\mathbf{y})$ is

$$V_p\hat{T}_{\text{HT}}(\mathbf{y}) = \sum_{j=2}^{N}\sum_{k=1}^{j-1}(\pi_j\pi_k - \pi_{jk})(y_j\pi_j^{-1} - y_k\pi_k^{-1})^2 \quad (7)$$

where π_{jk} is the joint probability of the inclusion of units j and k in sample. If (1) holds, $\hat{T}_{\text{GREG}}(\mathbf{y}) = T(\mathbf{X})\boldsymbol{\beta} + \hat{T}_{\text{HT}}(\boldsymbol{\varepsilon})$, where $\boldsymbol{\varepsilon}$ is the vector of the ε_j, so writing ε_j in the place of y_j in (7):

$$V_p\hat{T}_{\mathrm{GREG}}(\mathbf{y})=\sum_{j=2}^{N}\sum_{k=1}^{j-1}(\pi_j\pi_k-\pi_{jk})[\varepsilon_j\pi_j^{-1}-\varepsilon_k\pi_k^{-1}]^2. \quad (8)$$

The design-variances of the HT and the GREG estimators are therefore both functions of the π_{jk}. Important problems that this fact raises have been discussed in some detail by Särndal (1996). Basically they are that the π_{jk} tend to be difficult to evaluate and that their use involves a cumbersome double summation. [To this we may add that the Sen-Yates-Grundy (SYG) variance estimator (Sen 1953, Yates and Grundy 1953), which is usually the most efficient one to use when the sample size is fixed, is easily destabilised by the presence of small values among the π_{jk}, and is biased if any of the π_{jk} are zero. Zero values can occur easily, particularly when sampling systematically. Other relevant references on the π_{jk} include Rao and Bayless (1969), Bayless and Rao (1970) and Brewer and Hanif (1983, 62-68).]

Särndal's proposals were to circumvent the problems, either by relaxing the requirement that the first order inclusion probabilities be exactly proportional to size or the demand that the sample size be fixed. Here we suggest another way to circumvent these problems. It depends on the fact that if the working model (1) holds exactly, then the variations in the π_{jk} from one selection method to another (holding the first-order inclusion probabilities constant) contribute nothing to the prediction-variance of $\hat{T}_{\mathrm{GREG}}(\mathbf{y})$, and hence only trivially to its design-variance. Further, the anticipated variance (AV) of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y})-T(\mathbf{y})\}$, as defined by Isaki and Fuller (1982), which is its variance under both the design and model (1), is asymptotically independent of the π_{jk}. This may be seen as follows. Since $\hat{T}_{\mathrm{GREG}}(\mathbf{y})$ is both prediction-unbiased and asymptotically design-unbiased (Brewer 1979, Särndal and Wright 1984),

$$AV\{\hat{T}_{\mathrm{GREG}}(\mathbf{y})-T(\mathbf{y})\}\cong E_\xi V_p\hat{T}_{\mathrm{GREG}}(\mathbf{y})$$

$$=\sigma^2\sum_{j=2}^{N}\sum_{k=1}^{j-1}(\pi_j\pi_k-\pi_{jk})(\pi_j^{-2}a_j^2+\pi_k^{-2}a_k^2)$$

$$=\frac{1}{2}\sigma^2\sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}(\pi_k\pi_j^{-1}a_j^2+\pi_j\pi_k^{-1}a_k^2-\pi_{jk}\pi_j^{-2}a_j^2-\pi_{jk}\pi_k^{-2}a_k^2)$$

$$=\sigma^2\sum_{j=1}^{N}[(n-\pi_j)\pi_j^{-1}-(n-1)\pi_j^{-1}]a_j^2$$

$$=\sigma^2\sum_{j=1}^{N}(\pi_j^{-1}-1)a_j^2. \quad (9)$$

This is the same expression as was shown by Godambe (1955) to be the minimum possible anticipated variance (given the values of π_j) for any design-unbiased estimator of $T(\mathbf{y})$. (It also provides the justification for the choice of $\pi_j\propto a_j$ when seeking to minimize the design-variance.) It would therefore seem preferable, if (1) is indeed a useful working model, to estimate the AV of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y})-T(\mathbf{y})\}$ rather than the design-variance of $\hat{T}_{\mathrm{GREG}}(\mathbf{y})$. It follows immediately from (9) that a large-sample estimator of this AV is $\hat{\sigma}^2\sum_{j=1}(\pi_j^{-1}-1)\,a_j^2$, where $\hat{\sigma}^2$ can be the estimator of σ^2 obtained from a regular regression package based on the use of $\hat{\boldsymbol{\beta}}_{\mathrm{BLUE}}$, but preferably from one in which $\hat{\boldsymbol{\beta}}_{\mathrm{COSCAL}}$ takes the place of $\hat{\boldsymbol{\beta}}_{\mathrm{BLUE}}$ (*cf.* Fuller 1975). Since the only approximation involved in deriving (9) is the omission of terms arising from the design-bias, the proposed estimator may perform reasonably well in smaller samples where the design-bias is known to be small; as, for example, where a regression of the y_j on a single supplementary variable goes almost through the origin.

If, for each assumed value of γ, the sample a_j^2 are normalized to sum to (say) n, the values of $\hat{\sigma}^2$ will be comparable, but the best choice of γ is not necessarily the one that minimizes $\hat{\sigma}^2$. A robust estimator of γ can be obtained by finding the value of $\hat{\gamma}$ for which the correlation between $(y_j-\mathbf{x}_j\hat{\boldsymbol{\beta}})^2/z_j^{2\hat{\gamma}}$ and the rank of z_j is zero. However, estimates of γ, except where they come from large samples, are typically subject to high variance, and should be treated with caution, especially if they lie outside the range $0.5\le\gamma\le 1$.

When the analysis is secondary (*i.e.*, not carried out by the person or organization responsible for the design or conduct of the survey) the unavailability of certain relevant information can cause the sample selection to be informative. Special precautions are then usually needed when estimating the model (Pfeffermann, Skinner, Holmes, Goldstein and Rabash 1998). However if the sample values of all the $\mathbf{x}_j$ are known, σ^2 can be estimated using standard regression analysis and the only problem lies in the estimation of the expression $\sum_{j=1}^{N}(\pi_j^{-1}-1)\,a_j^2$. The HT estimator $\sum_{j\in s}\pi_j^{-1}(\pi_j^{-1}-1)\,a_j^2$ is always available as a last resort, but if a population total such as that of the z_j or of the π_j is known, or better still both are known, that estimator can be improved upon.

4. Prediction-variance

It is appropriate to estimate the anticipated variance for sample design purposes, but for the analysis of any particular sample the prediction-variance is a more logical choice. That prediction-variance is, by definition,

$$
\begin{aligned}
&E_\xi[\hat{T}_{\mathrm{COSCAL}}(\mathbf{y}) - T(\mathbf{y})]^2 \\
&\quad = E_\xi\left[\sum_{j\in s} w_j y_j - \sum_{j=1}^{N} y_j\right]^2 \\
&\quad = E_\xi\left[\sum_{j\in s} (w_j - 1)\, y_j - \sum_{j\notin s} y_j\right]^2 \\
&\quad = \sigma^2\left[\sum_{j\in s} (w_j - 1)^2 a_j^2 + \sum_{j\notin s} a_j^2\right] \\
&\quad = \sigma^2\left[\sum_{j\in s} w_j (w_j - 1)\, a_j^2 + \left(\sum_{j=1}^{N} a_j^2 - \sum_{j\in s} w_j a_j^2\right)\right] \qquad (10) \\
&\quad = \sigma^2\left[\sum_{j\in s} w_j (w_j - 1)\, a_j^2 + \left(\sum_{j=1}^{N} a_j^2 \;\; - \sum_{j\in s} \pi_j^{-1} a_j^2\right)\right. \\
&\qquad\qquad \left. - \left(\sum_{j\in s} w_j a_j^2 - \sum_{j\in s} \pi_j^{-1} a_j^2\right)\right]. \qquad (11)
\end{aligned}
$$

Assuming the a_j^2 are known or can be satisfactorily imputed, (10)-like (9)-can be estimated prediction-consistently by replacing σ^2 by $\hat{\sigma}^2$. [However the w_j are not defined for the nonsample units, so it is not possible to use either (10) or (11) to obtain a formula or estimator for the AV of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y}) - T(\mathbf{y})\}$.]

The first expression in round brackets in (11) is the difference between the population sum of the a_j^2 and its HT estimator, and has design expectation zero. Further, since w_j tends asymptotically to π_j^{-1}, the second such expression in (11) is asymptotically zero and negligible for large samples. Hence a simpler but still prediction-cum-design-consistent estimator of the prediction variance of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y}) - T(\mathbf{y})\}$ is $\hat{\sigma}^2 \sum_{j\in s} w_j (w_j - 1)\, a_j^2$. Since this does not require knowledge of the non-sample a_j^2, it is an attractive choice for secondary analysis.

Both the suggested estimators conveniently take the value zero when every unit in the population is also in sample with $w_j = 1$ for all j. However if the disparity between the population mean and the sample mean is substantial, it may, as in section 3, be necessary to construct special estimators of the unknown $\sum_{j=1}^{N} a_j^2$ and related population sums by calibrating on whatever relevant population data may be available.

5. The problem of negative and other unacceptably small sample weights

It was pointed out in Brewer (1999) that strong conditions had to be fulfilled before the Representative Principle underlying design-based inference could be regarded as useful. (This Principle required that for every sample unit included with probability π_j there should be approximately $\pi_j^{-1} - 1$ units with reasonably similar properties in the non-sample portion of the population.) Such strong conditions can nevertheless hold when both the population and the sample are large and the inclusion probabilities are an explicit function of a known measure of size, which is usually a linear function of the columns of $\mathbf{X}$.

The manner in which the Cosmetic Calibration weights, w_j, are constructed, however, implies that they are better indexes of the relevant properties than the π_j^{-1} are themselves. So there is a sense in which the inverse weights, w_j^{-1}, can be thought of as analogous to inclusion probabilities. Sample units with large weights (and hence small w_j^{-1}) can be considered as typical in their characteristics in that they "represent" large numbers of population units. Sample units with smaller weights can still be regarded as typical, but they "represent" fewer population units. A sample unit with weight unity is only on the borderline of being typical. It does not represent any other unit. A sample unit with a weight less than one is definitely atypical. It does not even represent itself. A sample unit with a negative weight is perversely atypical and counter-representative. For a small enough domain, it can actually produce negative estimates of total. Its presence in the sample is a "rare event". Yet it must be part of the population, or it could not be in the sample.

The obvious procedure to adopt for a unit with $w_j < 1$ is to delete it both from the sample and the sample frame, to recalculate the w_j so that the remaining sample units are calibrated on the totals of the remaining population units, and then add the deleted unit on as an atypical extra. This, of course, is precisely what many design-oriented survey statisticians have been doing with "outlying observations" for decades. It is also the natural thing to do with sample units that are allocated weights in an unacceptable range.

If, however, we start with a GREG that has not been cosmetically calibrated and attempt to remove the unacceptable weights by setting them at unity and recalculating the remainder, we usually find that many of the newly recalculated weights are themselves unacceptable. If the procedure is taken through further iterations, the number of units whose weights have been set to unity increases steadily, and the larger positive weights that are needed for those that remain leads to a substantial increase in prediction-variance.

This problem can be substantially reduced by using cosmetic calibration. Wherever a sample contains one or more units with such unacceptable weights, each of the corresponding π_j values can (by a convenient fiction) be set equal to one, and the calculation then repeated. The factor $(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}$ in (5) and (6) ensures that wherever π_j is set equal to unity, the corresponding w_j is also unity. The comparable factor for the standard GREG, $\mathbf{\Pi}_s^{-1}\mathbf{Z}_s^{-1}$, does not possess this property.

Removing negative and unacceptably small positive weights in this fashion provides no absolute guarantee that the remaining weights do not include some large ones. There is then the danger of introducing a substantial design-bias, but the results of the empirical study presented in the next section suggest that this danger is less than might be

feared. Where the inclusion probabilities increase only modestly with size, the cosmetically calibrated GREG can be seen to reduce the incidence of unacceptable weights substantially, and for one sample design entirely eliminate it, without materially increasing the design-variance or introducing any appreciable squared bias term into the design-MSE (mean squared error).

6. An empirical study using australian farm data

The actual performance of cosmetic calibration as compared with certain alternative estimation procedures has been studied using data obtained from two farm surveys conducted by the Australian Bureau of Agricultural and Resource Economics using economic and production data collected from a sample of 904 farms in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) and Australian Dairy Industry Survey (ADIS) in the late 1980s (Chambers 1996). The data set includes two variables (incomes from wheat and dairy sales), that follow model (1) reasonably well, and two others (incomes from sheep and beef sales) that do not. These properties make it a useful and exacting data set for testing purposes.

Chambers carried out a comparison of various estimation strategies using three sets of stratified random subsamples from his 904 sample farms. Each set consisted of 500 stratified simple random samples of 100 farms. The size variable used for stratification purposes was Dry Sheep Equivalents (DSEs). For the present study, three additional sets were selected, each again consisting of 500 samples of 100 farms. The inclusion probabilities used in each set were proportional to a fractional power of the DSE. For Set 4 that power was 0.60, for Set 5 it was 0.75, and for Set 6 it was 0.90. In each case there was also a completely enumerated sector. It was smallest for Set 4 and largest for Set 6.

The larger of the two versions of model (1) used by Chambers to construct his estimators had eleven supplementary variables: hectares of wheat, numbers of sheep, beef and dairy cattle, and seven zero-one Industry indicators. This model provided the stronger challenge, and the comparisons presented here relate to that model only.

Chambers calculated sample weights and RMSEs (root mean squared errors) for each of Sets 1-3 using six different estimators. The first of these, "RATIO", was the HT ratio estimator based on each survey variable's natural supplementary variable (such as hectares of wheat for wheat income). He calculated this only as a basis of comparison for other estimators, holding it to be essentially unsatisfactory in that the sample weights differed from one survey variable to another.

The five estimators (other than "RATIO") that were used by Chambers were the standard "GREG" [identical with the first variant of Calibration Estimation used in Deville and Särndal (1992), the variable z in this instance being DSE], "BLUP," (the Best Linear Unbiased Predictor), "RIDGE," (a ridge regression estimator that enabled all weights to fall within the acceptable range) and two estimators, "NWD3" and "NWDAR3," that applied Nadaraya-Watson nonparametric adjustments to the weights for the estimator "RIDGE".

As a supplement to Chambers' study, the cosmetic calibration estimator, "COSCAL," was calculated for Sets 2-6. "COSCAL" and "GREG," having nearly identical formulae, usually had very similar MSEs. Since Chambers' only reason for introducing "RIDGE," "NWD3" and "NWDAR3" was to get rid of unacceptable sample weights, the relevant comparisons in MSE terms are those between these three estimators and "COSCAL."

Except in Set 6, all or nearly all the "COSCAL" weights for any given sample were eventually made greater than or equal to one, but occasionally one or more of the 100 weights could not be found a value in the acceptable range. The most intractable instances occurred where only three farms had been selected for the Dairy Industry and all three of them were of larger than average size. It was therefore logically impossible to calculate any set consisting of all positive weights to specify an estimator calibrated both on the number of dairy farms and on the Dairy Industry's total size measure. (Dairy farms in Australia are typically on the small side, so for a sample of given size there are fewer dairy farms selected when the probability of inclusion increases rapidly with size of farm than when it increases slowly.)

The actual extent of the unacceptable weight problem is indicated in Table 1. The elimination procedure broke down completely only for Set 6. It seems probable that there was less of a problem for Set 3 than for Set 6 because Set 3 had no inclusion probabilities that were close to but not equal to one.

Table 2 shows the initial incidences of unacceptable weights for the estimators "GREG", "BLUP/RIDGE" and "COSCAL". The corresponding final incidence for "BLUP/RIDGE" is uniformly zero, but the estimator is then no longer "BLUP" but "RIDGE". For Set 2 the initial incidence of such small weights is substantially less for "BLUP" than it is for "GREG" or "COSCAL". For Set 3, however, the initial incidence for "COSCAL" is substantially smaller than that for "GREG" or "BLUP". For Set 6 ($\pi_j \propto \mathrm{DSE}_j^{0.90}$) the number of unacceptable weights found and the number of intractable samples discovered were already unacceptably large after only two iterations.

RMSEs were obtained for "COSCAL", both before and after the unacceptable sample weights had been eliminated as far as possible. Table 3 contains a comparison between these RMSEs and those reported in Chambers (1996) for the other estimators.

Most of the RMSEs obtained for the final versions of "COSCAL" are very similar to those obtained from the initial versions, and also from the standard "GREG". The deterioration seen for "COSCAL" in the Dairy Income

estimates is due to the small number of Dairy Industry farms selected (particularly in Set 3) and the consequently rapid rise in RMSE that occurred as more and more farms with unacceptable weights were given unit weight and the effective sample size was consequently decreased. The same is true to a lesser extent for wheat farms.

Table 1
Progressive elimination of unacceptable COSCAL sample weights

Sample Set	Iteration number	Number of samples with sample weights < 1	Intractable samples detected	Number of sample weights < 1 across samples
Set 2	0	277	0	496
	1	85	0	127
"Compromise"	2	18	0	29
allocation	3	7	0	16
	4	2	2	3
Set 3	0	226	0	701
	1	100	1	303
"Optimal"	2	48	1	134
allocation	3	27	4	75
	4	13	7	48
	5	8	7	39
	6	8	8	39
Set 4	0	188	0	322
	1	55	0	80
Allocation	2	11	0	14
$\propto \mathrm{DSE}^{0.60}$	3	0	0	0
Set 5	0	204	0	341
	1	51	0	77
Allocation	2	11	0	16
$\propto \mathrm{DSE}^{0.75}$	3	3	1	4
	4	1	1	2
Set 6	0	187	0	592
	1	96	1	229
Allocation	2	46	6	154
$\propto \mathrm{DSE}^{0.90}$	Further analysis abandoned			

After the elimination of unacceptable weights, the cosmetically calibrated estimator is design-biased. This is on account of atypical farms that were not selected with certainty being given unit weight. Table 4 shows that for all variables other than Dairy Income the change in the bias between Initial and Final/Intermediate was less than one third of a percentage point. For Dairy Income it is 3.21% for the intractable Set 6, 1.25% for the next most intractable Set 3 and 0.53% or less for the remainder. In every case the squared bias is less than 11% of the MSE, being largest for Wheat Income Set 6 (both Initial and Final).

Table 4 also supplements Table 3 in giving data both on the accuracies of the sample estimates obtained using Sets 4, 5 and 6 and on the percentage Mean Average Deviation Errors (% MADE) for all sets. Sets 4 and 5 seem close to having optimal inclusion probabilities, both in terms of MSE and in the ease with which unacceptable weights can be removed (seemingly just a coincidence, but certainly a happy one). By contrast, Set 6 performs rather poorly. The ratios of the MADEs to the MSEs almost all fall in the range from 0.52 to 0.68. The three exceptionally small ratios, all for Dairy Income, appear to be indicative of occasional large deviations from the mean when the number of dairy farms selected was particularly small.

Table 2
Percentages of samples containing unacceptable sample weights

Sample Set	GREG		BLUP/ RIDDGE		COSCAL			
					Initial		Interm/Final*	
Set 1 (srswor)	77	(4.27)	77	(4.27)	77	(4.27)	n.c.	(n.c.)
Set 2 "Compromise"	53	(1.83)	20	(1.83)	55	(1.79)	0.4	(1.50)
Set 3 "Optimal"	94	(9.73)	93	(5.87)	45	(3.10)	1.6	(4.88)
Set 4 ($\propto \mathrm{DSE}^{0.60}$)	n.c.		n.c.		38	(1.71)	0.0	–
Set 5 ($\propto \mathrm{DSE}^{0.75}$)	n.c.		n.c.		41	(1.67)	0.2	(2.00)
Set 6 ($\propto \mathrm{DSE}^{0.90}$)	n.c.		n.c.		37	(3.17)	n.c.	(n.c.)

n.c. not calculated.
Numbers in parentheses are the average numbers of sample weights less than unity in samples containing at least one such unacceptable sample weight.
*Intermediate for Set 6, Final for all other Sets.

Table 3
RMSEs of the estimated population means of survey variables as percentages of the corresponding population values (Chambers' original sets only)

Estimator	Income From:				
	Wheat	Beef	Sheep	Dairy	Total
	Set 1 (srswor)				
RATIO	14.7	28.9	19.1	14.4	16.7
GREG	13.6	26.1	17.0	15.0	17.3
BLUP	13.6	26.1	17.0	15.0	17.3
RIDGE	15.7	23.6	16.0	17.1	15.7
NWD3	15.0	22.1	15.9	17.5	14.6
NWD3AR	14.5	22.4	15.6	17.0	14.7
	Set 2 "Compromise"				
RATIO	10.0	11.6	15.5	19.2	8.3
GREG	9.9	11.9	14.8	20.3	8.4
BLUP	10.8	12.8	14.3	20.5	8.9
RIDGE	13.2	13.0	15.6	23.1	9.8
NWD3	10.5	11.5	14.1	19.8	8.1
NWD3AR	10.5	11.6	14.1	19.7	8.1
COSCAL:					
Initial	9.9	12.1	14.8	20.3	8.4
Final	9.9	12.0	14.8	21.1	8.4
	Set 3 "Optimal"				
RATIO	10.1	10.1	15.9	25.7	7.9
GREG	11.6	11.6	17.4	32.3	8.4
BLUP	11.9	11.1	16.4	32.1	8.0
RIDGE	23.5	9.6	21.3	47.8	11.9
NWD3	12.5	9.1	15.6	30.7	7.3
NWD3AR	12.9	8.9	15.7	31.5	7.3
COSCAL:					
Initial	11.6	11.4	17.6	32.5	8.3
Final	14.6	11.6	18.1	41.4	8.7

Table 4
Performances of initial and final (or intermediate*) COSCAL estimates

Survey Variable	Sample Set	Inital			Intermediate/Final*		
		% Bias	% RMSE	% MADE	% Bias	% RMSE	% MADE
Wheat Income							
	Set 2 "Compromise"	0.22	9.9	6.4	0.13	9.9	6.4
	Set 3 "Optimal"	0.99	11.6	7.7	0.67	14.6	7.7
	Set 4 ($\propto \mathrm{DSE}^{0.60}$)	1.83	8.9	6.0	1.79	8.8	6.0
	Set 5 ($\propto \mathrm{DSE}^{0.75}$)	2.93	9.7	5.7	2.92	9.7	5.7
	Set 6 ($\propto \mathrm{DSE}^{0.90}$)	3.45	11.0	7.0	3.52	10.8	6.9
Beef Income							
	Set 2 "Compromise"	-0.01	12.1	8.1	-0.08	12.0	8.1
	Set 3 "Optimal"	0.50	11.4	7.0	0.25	11.6	7.0
	Set 4 ($\propto \mathrm{DSE}^{0.60}$)	2.22	13.0	7.4	2.49	11.4	7.3
	Set 5 ($\propto \mathrm{DSE}^{0.75}$)	2.00	10.4	6.6	1.97	10.4	6.6
	Set 6 ($\propto \mathrm{DSE}^{0.90}$)	1.55	11.4	6.2	1.71	10.9	6.4
Sheep Income							
	Set 2 "Compromise"	1.05	14.8	9.9	1.09	14.8	9.9
	Set 3 "Optimal"	0.94	17.6	10.8	0.72	18.1	10.9
	Set 4 ($\propto \mathrm{DSE}^{0.60}$)	-0.09	13.5	9.0	-0.04	13.6	9.0
	Set 5 ($\propto \mathrm{DSE}^{0.75}$)	0.27	14.5	9.8	0.35	14.4	9.8
	Set 6 ($\propto \mathrm{DSE}^{0.90}$)	1.04	16.9	9.9	1.11	17.2	10.3
Dairy Income							
	Set 2 "Compromise"	-0.24	20.3	11.4	0.25	21.1	11.4
	Set 3 "Optimal"	1.32	32.5	15.2	2.57	41.4	15.1
	Set 4 ($\propto \mathrm{DSE}^{0.60}$)	-0.52	20.2	11.7	-0.30	20.1	11.7
	Set 5 ($\propto \mathrm{DSE}^{0.75}$)	-2.47	20.4	13.4	-1.94	21.4	13.5
	Set 6 ($\propto \mathrm{DSE}^{0.90}$)	0.01	29.8	16.3	-3.20	57.8	16.6
Total Income							
	Set 2 "Compromise"	0.18	8.4	5.4	0.14	8.4	5.4
	Set 3 "Optimal"	0.69	8.3	5.4	0.46	8.7	5.6
	Set 4 ($\propto \mathrm{DSE}^{0.60}$)	1.75	8.9	4.6	1.92	7.9	4.6
	Set 5 ($\propto \mathrm{DSE}^{0.75}$)	1.83	7.5	5.0	1.84	7.5	4.9
	Set 6 ($\propto \mathrm{DSE}^{0.90}$)	1.83	8.5	4.6	1.87	8.3	4.6

* Intermediate for Set 6, Final for all other Sets.

No single estimator out of "RATIO", "GREG", "BLUP" and "COSCAL" has a consistent edge over any of the others on the sole ground of low RMSE. "RIDGE" is generally inferior, as might be expected on account of its prediction-bias, and the two Nadaraya-Watson estimators are generally superior, as might also be expected on account of their nonparametric calibration. However, their superiority is neither compellingly large nor consistent over all variables.

If the choice is restricted to the three simplest estimators capable of producing the same weights for all variables, namely "BLUP", "GREG" and "COSCAL", then all three are comparable in accuracy but "BLUP" and "GREG" are inferior to "COSCAL" in the elimination of unacceptable sample weights. It is true that "COSCAL" was not uniformly successful in eliminating such weights, but the test it faced was exceptionally severe. Eleven explanatory variables were used for samples of size $n = 100$, the totals of these explanatory variables included several linked pairs (each consisting of a production measure and a count of the number of contributing farms) and for two of the six sample sets the inclusion probabilities increased rapidly with size. Such a stringent combination of requirements should seldom be encountered in normal survey practice. However, especially in circumstances where the explanatory variables include such linked pairs, it would seem prudent to avoid using inclusion probabilities that increase rapidly with size, even at the expense of a moderate departure from the otherwise optimal rule that the π_j should be proportional to the a_j.

7. Evaluation

It takes little effort to change a standard GREG estimator into a cosmetically calibrated estimator. The matrix $\mathbf{\Pi}_s^{-1}$ in one formula must be replaced by $\mathbf{\Pi}_s^{-1} - \mathbf{I}_n$, and it may also be desirable to replace the existing $\mathbf{Z}_s$ matrix by another choice. The efficiency of the estimator seems to be little changed as a result, but there are several unequivocal advantages.

(i) The estimator is then clearly interpretable as prediction-based as well as design-based.

(ii) Its anticipated variance and its prediction-variance can both be estimated more easily and more

efficiently than the design-variance of the standard GREG. [Although these options are also available for any GREG estimator, the most appropriate estimator of $\boldsymbol{\beta}$ for the purpose of estimating σ^2 is one that is equally relevant to design-based and prediction-based inference. The $\hat{\boldsymbol{\beta}}_{COS}$ obtained by equating (2) and (3) is such an estimator.]

(iii) Design-based estimation has a tendency to be more reliable for large samples, and prediction-based estimation for small samples and small domains (Brewer 1999). It is not surprising, therefore, that the estimators used for large domains are typically design-based while those for small domains are often purely prediction-based or "synthetic." If the large-domain estimators are cosmetically calibrated, the estimates for their component small domains automatically sum to them without forcing.

(iv) As an unexpected spin-off, the elimination of negative and other unacceptably small weights is streamlined by the use of cosmetic calibration.

Acknowledgements

I am indebted to Prof. R.L. Chambers, who provided me with the initial impetus to consider this problem, and to Prof. A.H. Welsh and Drs. P.S. Kott and D.A.Binder for their advice and encouragement along the way. An Associate Editor and two referees provided extended comments on the paper when it was first submitted, which greatly improved both its content and its presentation.

References

Bayless, D.L., and Rao, J.N.K. (1970). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling ($n = 3$ or 4). *Journal of the American Statistical Association*, 65, 1645-1667.

Brewer, K.R.W. (1979). A class of robust sample designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

Brewer, K.R.W. (1995). Combining design-based and model-based inference. Chapter 30 in *Business Survey Methods*, (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott.). New York: John Wiley & Sons, Inc., 589-606.

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., and Hanif, M. (1983). *Sampling With Unequal Probabilities, Lecture Notes in Statistics*. New York: Springer-Verlag, 15.

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.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, Series B, 60, 3-21 with discussion on 41-56.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, Series B, 17, 269-278.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rabash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-40 with discussion on pp 41-56.

Rao, J.N.K., and Bayless, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units. *Journal of the American Statistical Association*, 64, 540-549.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Särndal, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

Särndal, C.-E., and Wright, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

Sen, A.R. (1953). On the estimate of the variance when sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 18, 52-56.

Yates, F., and Grundy P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Series B, 15, 253-261.